Data science and usefulness in domains of human action

Anton Andreacchio (1132675)

April 7, 2023

Thesis submitted for the degree of
Master of Philosophy
in
Applied Mathematics
at The University of Adelaide
Faculty of Sciences, Engineering and Technology
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Contents

Signed Statement **xi**

Acknowledgements **xiii**

Abstract **xv**

1 Introduction **1**
1.1 Selecting decision makers and consultation . . . 3
1.2 Usefulness, explainability and interpretability . . . 4
1.2.1 Performance . . . 5
1.2.2 Scalability . . . 5
1.2.3 Comprehensibility . . . 6
1.2.4 Actionability . . . 6
1.2.5 Justifiability . . . 7
1.2.6 Importance of characteristics in each domain . . . 7
1.3 Outline of thesis . . . 8
1.4 Publication of work . . . 9

2 Background **11**
2.1 Markov model . . . 11
2.1.1 Discrete-time Markov chains . . . 11
2.1.2 Absorbing states and absorbing probabilities . . . 12
2.2 Pairwise performance comparisons . . . 12
2.2.1 Elo rating system . . . 13
2.3 Natural language processing . . . 14
2.3.1 Tokenisation . . . 14
2.3.2 Sentiment analysis . . . 14
2.3.3 Sentiment arcs . . . 15
2.3.4 Screenplay analysis . . . 15
2.4 Forgetting curves . . . 16
2.5 Statistical modelling techniques . . . 16

2.5.1 Regression analysis . . . 16
2.5.2 Decision tree learning . . . 17
2.5.3 k-nearest neighbours . . . 17
2.5.4 Confusion matrix . . . 17

3 Modelling Australian Rules Football as spatial systems with pairwise comparisons 21
3.1 Introduction . . . 21
3.1.1 Decision makers and usefulness . . . 21
3.1.2 Proposed approach . . . 22
3.2 Defining spatial systems in AFL . . . 23
3.3 Determining relative team performance of match systems . . . 25
3.3.1 Describing AFL as a Markov model . . . 27
3.4 Defining and tuning model parameters . . . 29
3.5 Predictive accuracy of the system model . . . 31
3.6 Line ratings as a description of team strategy . . . 34
3.7 Discussion and usefulness . . . 39

4 Capitalisation pathways of South Australian startups 41
4.1 Introduction . . . 41
4.1.1 Decision makers and available data in startup capitalisation pathways 42
4.1.2 Model requirements for usefulness . . . 43
4.1.3 Markov chain approach . . . 43
4.2 Startups in South Australia . . . 44
4.2.1 Recoding capital events . . . 45
4.2.2 Adjusting capital events for inflation . . . 46
4.2.3 Temporal normalisation . . . 46
4.2.4 Cumulative amounts and state allocations . . . 47
4.3 Mapping capitalisation pathways . . . 48
4.3.1 Startup transformation count matrix . . . 48
4.3.2 New startups entering the ecosystem . . . 49
4.3.3 Startup ecosystem transition matrices . . . 50
4.3.4 Application of the transition matrix . . . 51
4.3.5 Multi-year projections using the startup transition matrix . . . 52
4.3.6 Defining additional states . . . 53
4.4 Startup capital transformation pathways . . . 55
4.4.1 Model 1 - Total capitalisation pathways . . . 56
4.4.2 Model 2 - Grant funding pathways . . . 58
4.4.3 Model 3 - Private capitalisation pathways . . . 60
4.5 Does grant funding unlock private capital in South Australia? . . . 63

4.5.1 Probability of private capital transformation from different grant states . . . 64
4.5.2 Absorbing probabilities of Stable and Death states . . . 65
4.5.3 Does grant funding unlock private capital? . . . 66
4.5.4 Probability of mobility from each state . . . 68
4.6 Application as a forecasting tool . . . 69
4.6.1 Forecast 1: "Current trajectory" scenario and forecast . . . 70
4.6.2 Forecast 2 - "More startups" scenario and forecast . . . 73
4.6.3 Forecast 3 - No grant funding scenario . . . 75
4.7 Discussion and further research . . . 80
4.7.1 Insights in the South Australian startup ecosystem . . . 80
4.7.2 Evaluation of model usefulness . . . 81

5 Natural language processing in screenplay development 83
5.1 Introduction . . . 83
5.1.1 Usefulness of story arc analysis . . . 84
5.1.2 Proposed methodology . . . 85
5.2 Natural language processing and understanding story arcs . . . 85
5.2.1 Analysing episode scripts as text corpora . . . 86
5.2.2 Attributing sentiment . . . 87
5.2.3 Generating story arcs using the window method . . . 88
5.3 Introducing forgetting curves . . . 89
5.3.1 Alternative models . . . 90
5.3.2 Application to text corpora . . . 90
5.4 The story arcs of *Aftertaste* . . . 91
5.4.1 Analysing Episode 1 . . . 92
5.4.2 Episode and scenario comparisons . . . 96
5.4.3 Reading an episode in context . . . 98
5.5 Discussion and further research . . . 100
5.5.1 Usefulness of the experience arc approach . . . 101
5.5.2 Comparison with the window approach . . . 102
5.5.3 Alternative dictionaries . . . 103
5.5.4 Forgetting distribution . . . 103
5.5.5 Further work . . . 104

6 Conclusion 105
6.1 Summary of results . . . 105
6.2 Governance and developing useful metrics . . . 106
6.3 Future research . . . 107

Bibliography 109

List of Tables

1.1 Primary stakeholders in each of the domains of interest. 7
1.2 Relative importance of each characteristic in each domain. 8

3.1 Defined system states in Australian Rules Football 24
3.2 Predictive accuracy of optimised spatial system model based on sampling across 2015, 2016 and 2017 AFL Premiership Seasons 32
3.3 Selected variables for spatial system model based on distribution of tuned parameters. 33
3.4 Confusion matrix of spatial state model across 2015, 2016 and 2017 AFL Premiership Seasons 33
3.5 Predictive accuracy across 2015, 2016 and 2017 AFL Premiership Seasons . 34
3.6 Comparison of logistic regression models using different variable combinations 36

4.1 Quartiles of cumulative total capital received by startups in South Australia from 2015 to 2019. 47
4.2 Selected states for total capitalization of individual startups in South Australia 47
4.3 Number of South Australian startups in selected total capitalisation states from 2015 to 2019 48
4.4 State space profile of South Australian startup ecosystem in 2017 52
4.5 Absorbing probabilities using Model 1 (Total Capitalisation) 2015-2019 state transition matrix 57
4.6 Quartiles of cumulative grant funding received by startups in South Australia from 2015 to 2019. 58
4.7 Selected states for total grant funding of individual startups in South Australia 58
4.8 Absorbing probabilities using Model 2 (Grant funding) 2015-2019 transition matrix 60
4.9 Quartiles of cumulative private capital received by startups in South Australia from 2015 to 2019. 60
4.10 Selected states for total private capitalization of individual startups in South Australia 61

4.11 Absorbing probabilities using Model 3 (Private capitalisation) 2015-2019 transition matrix 62
4.12 Selected states for total private capitalization and total grant funding of individual startups in South Australia 64
4.13 Absorbing probabilities using Model 4: 2-dimensional 2015-2019 transition matrix 66
4.14 Pathways from startup capitalisation state $(p_0, g_0) \to (p_1, g_1)$ 67
4.15 Comparison of grant funding transformation probabilities from private capitalization states 67
4.16 Comparison of private capitalization transformation probabilities from grant funding states 67
4.17 Probabilities of general startup transformation in South Australia in a given year 69
4.18 Forecast of startup ecosystem profiles - Model 1: Current trajectory . . . 72
4.19 Forecast of startup ecosystem profiles - Model 2: More startups scenario . 74
4.20 Forecast of startup ecosystem profiles - Model 3A: No grant funding . . . 77
4.21 Forecast of startup ecosystem profiles - Model 3B: No grant funding . . . 79

5.1 Table of tokenised data from excerpt of *Aftertaste* - Season 1, Episode 1 . . 87
5.2 Token valence using AFINN dictionary from excerpt of *Aftertaste* - Season 1, Episode 1 88
5.3 Decay parameters of token valence from using $\lambda = 1, \beta = \psi = 0.5$ from excerpt of *Aftertaste* - Season 1, Episode 1 91
5.4 First non-zero valence words from *Aftertaste* - Season 1, Episode 1 92
5.5 Credited writers for each episode of *Aftertaste* Season 1 96

List of Figures

2.1 Confusion matrix example 18

3.1 State space areas overlaid over a typical Australian Rules Football ground 25
3.2 Distribution of projected home team goal scoring against actual match outcome 34
3.3 Comparison of team system ratings for Premiership winners across 2015, 2016 and 2017 AFL Premiership seasons. 35
3.4 Classification and regression tree of system ratings and home team absorbing probability across the 2015, 2016 and 2017 AFL Premiership Season. . . . 38
3.5 Classification and regression tree of system ratings and home team absorbing probability across the 2015, 2016 and 2017 AFL Premiership Season, optimised for complexity parameter. 38

4.1 Startup capital events distribution in South Australia from 2012 to 2019 . 46
4.2 State diagram of startup capitalisation in 2015 financial year 49
4.3 State diagram of startup capitalisation in 2015 financial year including no capitalisation state 50
4.4 State diagram of startup capitalisation in 2015 financial year including exit states 55
4.5 Plot of startup ecosystem states - Model 1: Current trajectory 73
4.6 Plot of startup ecosystem states - Model 2: More startups scenario 75
4.7 Plot of startup ecosystem states - Model 3A: No grant funding 77
4.8 Plot of startup ecosystem states - Model 3B: No grant funding 79

5.1 Excerpt from screenplay of *Aftertaste* - Season 1, Episode 1 86
5.2 Decay profiles across five sets of forgetting curve coefficients 93
5.3 Decayed cumulative sentiment arcs for *Aftertaste* - Season 1, Episode 1 . . 95
5.4 Decayed cumulative sentiment arcs across forgetting scenarios of *Aftertaste* - Season 1 97
5.5 Perfect memory sentiment arcs comparing independent consumption, binge-watching and window method: *Aftertaste* Season 1 99

5.6 Fast and shallow decayed cumulative sentiment arcs comparing independent consumption, binge-watching and window method: *Aftertaste* Season 1 . . 100

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: ..... Date: .....

Acknowledgements

Sincere thanks to Professor Nigel Bean and Professor Lewis Mitchell for their encouragement, support, patience and dedication. Returning to study after a decade was a significant personal and professional challenge, and it was an honour and privilege to study under your guidance.

To my brother Carlo, who has kept our companies steady, thank you for affording me the time and space to pursue this work and find my path.

Finally, thank-you to Emily for all your support, and for tolerating all the lost evenings and weekends.

Abstract

Rapid advancements in computational science and artificial intelligence are transforming virtually every industry. In many situations however, uninterpretable modelling techniques are challenging to implement, and raise significant governance, operational and ethical challenges.

In this thesis, we focus on three domains where there is limited data availability, significant complexity, and human decision makers. Rather than focusing on increasing data collection in these domains, we focus on developing useful models based on readily available data, describing a model's usefulness as being based on five characteristics of interest: *performance, scalability, comprehensibility, justifiability* and *actionability*.

In Chapter 3, we model Australian Rules Football as spatial systems rather than individual possession events. Several methods are introduced to disentangle relative team performance and the functioning of sub-systems to evaluate historical games and predict future performance.

Next, Chapter 4 explores startup transformation pathways in South Australia. Working with limited data in the South Australian startup ecosystem to map startup capitalisation, we follow 151 startup journeys over an eight year period to develop an approach to support policy-makers to understand ecosystem transformation, with a focus on grant interventions and private capitalisation events.

In Chapter 5, we explore creativity and the writers room, working alongside the TV series *Aftertaste* to evaluate the limits and potential for natural language processing to support the creative process. By approaching the intersection of creativity and data analysis from the direction of usefulness, we are able to evaluate an existing method for story arc generation and rethink the approach to make it a more useful tool to support creative development.

Finally, we conclude in Chapter 6 by discussing our results and the role of data science modelling in these three domains. This includes a summary of results across the three domains, the relationship between governance and data science projects, and areas for further research in each direct domain.

This work presents advances in each of the three domains explored, presenting new practical approaches as well as revealing significant new areas for further research. In addition, the work demonstrates the viability of usefulness characteristics for data sci-

ence research, with positive implications for governance, research, and development of complementary techniques to uninterpretable artificial intelligence and machines learning methods.

Chapter 1

Introduction

The advent and rise of information technologies has ushered in a new social and technological era, often called the Information Age. Exponential advancements in computation, digital information storage, data capture, and data transmission over the past fifty years have radically transformed economies, society, and culture (Castells (1996)), and enabled new frontiers of value creation through productivity gains, automation, and the formation of new industries.

Mathematical and computer sciences have been central to this evolution, as the digital foundations on which these new technologies are all built are conceptualised, developed, governed and understood through the lens of mathematical models and processes. This includes discrete mathematics working with binary information systems, software engineering, computational algorithms, data structures and data based theory (Aho & Ullman (1992)), and the critical importance of logic in system design (Huth & Ryan (2004)).

These advances have spawns important new fields of research that are adjacent to traditional applied mathematics, such as data science, computer science, machine learning, and cryptography, which have in turn transformed virtually all industries and disciplines by unlocking new tools for modelling and analysis (Prabhu et al. (2011)).

As these new fields have become more powerful, a range of challenges have emerged. The ethics of artificial intelligence are of growing interest, as nations navigate concerns around emerging and rapidly escalating ethical issues (Barocas & Boyd (2017), Hagendorff (2020), Jobin et al. (2019)). In an effort to guide this development, Australia has launched an Artificial Intelligence (AI) Ethics Framework to evaluate how automated decision making and prediction of human behaviour can be placed within an ethical framework, and how development in AI can be refocused on improving well-being (Dawson et al. (2019)).

Another challenge is the trade-off in machine learning between the predictive power of models and interpretability (Gilpin et al. (2018)). As deep learning techniques have become increasingly powerful, their "black-box" nature has resulted in a prioritisation of model performance and outcome prediction over the traditional statistical focus of

explaining relationships in data (Kelleher & Tierney (2018)). The lack of interpretability is a source of growing distrust (Zhang et al. (2021)) though technological advancements show no signs of slowing (Theis & Wong (2017)).

Furthermore, as research and technological capability has rapidly advanced, the rapid implementation of artificial intelligence and data science in many domains has led to a range of biases (Brown et al. (1998)). The "golden hammer" bias, for instance, refers to an over-reliance on tools that are familiar and available, based on the observation by Abraham Maslow's that "I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail" (Maslow (1966)). This has become increasingly concerning as artificial intelligence has been deployed across the social services, where predictive modelling on biased data sets have critical implications in criminal justice, healthcare, public education, and welfare. Flawed algorithms can amplify biases through feedback loops (Zou & Schiebinger (2018)), and whilst regulation is starting to occur (Regulation (2018)), oversight continues to lag behind implementation.

Whilst there is no doubt big-data approaches can be tremendously powerful, not all problems have sufficient data for these techniques to be effective, and can result in tools either being inappropriately used, or inconvenient problems being deprioritised in favour of domains more suited to big-data analysis.

These and other concerns continue to mount. To generalise the problem, the American sociobiologist Edward O. Wilson noted during a debate at the Harvard Museum of Natural History in 2009, that: "The real problem of humanity is the following, we have Paleolithic emotions, medieval institutions, and god-like technology" (Wilson (2009)).

In response to these concerns, a new trend is emerging in data science research to grapple with the responsibility of these new tools, with a focus on the purpose of modelling as well as a rethink of the measures of success beyond predictive power.

Interpretability and *usefulness* are new terms that are fast becoming utilised to determine whether tools are suitable for the problems to which they are being applied. These measures include predictive power and performance, but also additional underlying characteristics such as *scalability*, *comprehensibility*, *justifiability* and *actionability* (Coussement & Benoit (2021)).

This approach is congruent with traditional data analysis, suggesting that this is part of a cycle that the mathematical and computer sciences have navigated before. In John Tukey's 1962 article "The Future of Data Analysis", data analysis was defined as a science because it passed three tests (Tukey (1962)):

1. intellectual content;
2. organisation into an understandable form;
3. reliance upon the test of experience as the ultimate standard of validity.

In this thesis, we explore the notion of *usefulness* in data science, focusing on regimes

that are heavily reliant on human action, complex decision making, and limited data availability.

For our analysis, we explore and develop novel modelling approaches in areas where there is limited data availability and significant complexity. These problems are commonly known as "small n large p" problems, or systems that suffer from the curse of dimensionality (Bellman (1966)) where there are relatively few data points compared to the many features and variables.

Our objective in this analysis is to develop approaches that have increased *usefulness*, including measures of *performance, scalability, comprehensibility, justifiability* and *actionability*. Drawing on Tukey's third condition, we have selected domains that are not readily automatable and where human experience is the standard of validity.

The three disparate domains selected for this analysis are:

- high performance sport, in particular, complex systems analysis in the Australian Football League (AFL);
- startup capital transformation pathways in early stage entrepreneurship;
- story arc analysis in screenwriting, with a focus on supporting the creative development process.

These three areas align with the author's own experience working professionally across these fields, which enables a sufficient understanding of practical problems that exist at the intersection of applied data science and industry challenges.

With an increasing focus on how digital technologies and AI-based solutions support human decisions and actions (Nahavandi (2019)), this research is timely and relevant.

1.1 Selecting decision makers and consultation

Given the importance of usefulness characteristics of actionability and justifiability, to develop useful approaches, the decision making stakeholders in each domain must to be identified. Direct industry consultation in each domain was undertaken to identify critical decision makers, as well as their availability for direct dialogue and interest for data science support. The author's relationships in each domain enabled models and their usefulness to be actively evaluated throughout the research process.

For Australian Football League analysis, our objective is to improving decision making processes of coaching departments, and hence the decision makers are coaches and football departments at each club. A significant amount of the research in the Australian Football League is concerned with "beating the odds" (Leushuis (2018)), however, this appears solely interested in outcome prediction rather than systems improvement. The author's direct experience implementing technology and training programs at AFL clubs enabled direct consultation to occur with senior and assistance coaching staff.

For startup capital transformation pathways, governmental policy decision makers are the primary decision makers given the extent of government support for incubators, accelerators, grant funding programs and other startup transformation entities. The author's role on the Government of South Australia's Entrepreneurship Advisory Board, experience with multiple startup endeavours and direct relationships with both startups and policymakers provided exposure to decision making pressures and requirements.

Lastly, application of natural language processing to story arc analysis resulted in creative producers and screenwriters being identified as the primary decision makers. The author has an established post-production company in the film industry and explored multiple decision making interfaces. In 2021, the opportunity for speaking engagements at national screen industry conferences Screen Forever and Screenmakers provided direct opportunity for the author to engage with broad audiences, as well as direct consultation with the Australian Government's screen funding agency Screen Australia and the South Australian Film Corporation. Screenwriters were selected as the primary decision makers given their fundamental role in crafting the underlying texts, and participation in the development of active television episodes and streaming shows provided a direct opportunity to actively test insights during the creative process.

1.2 Usefulness, explainability and interpretability

In his 1933 book *Science and Sanity*, Alfred Korzybski described the need to recognise that models are only representations of the system that they describe. He writes: "A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness." (Korzybski (1933)).

The relationship between a model's usefulness to the world that it seeks to explain is more recently defined in data science using the interpretable Decision Support System (iDSS) (Coussement & Benoit (2021)). This system proposes five characteristics of interest to describe a model's usefulness:

1. performance,
2. scalability,
3. comprehensibility,
4. justifiability,
5. actionability.

This approach enables the evaluation of relationships between the human decision makers with the data, the identified problem, and the proposed modelling approach. In machine learning and data science, terms such as *interpretability* and *explainability* are

defined as "the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model" (Dam et al. (2018)). We include these terms within the concept of comprehensibility, using comprehension to include both how to interpret and explain the applied models.

The five characteristics have very different meanings in our three domains of interest, and so each was evaluated independently.

1.2.1 Performance

The performance characteristic of modelling Australian Football League games to coaches and football departments describes whether the model accurately predicts game, or sub-game, outcomes. Sub-game outcomes may refer to smaller in-game systems that might unfold within the broader match, at an appropriate resolution for decision making and implementation. Australian Rules Football is very different to many other sports such as American gridiron or basketball as there are fewer opportunities for set plays to be coordinated, and the continuous nature of the game makes deterministic analysis challenging. Some sub-game systems were identified, such as center bounces following a goal event, kick-ins following a point event, or throw-ins after an out-of-bounds event, however, there was demand for a more generalised approach to modelling game outcomes.

For policymakers evaluating startup ecosystems, performance of a model is based on its ability to predict startup pathways and likelihood of private capitalisation events. Given the objectives of encouraging high-risk, high-reward startup endeavours suitable to attracting early stage capital investment, the focus of decision makers was less in producing stable companies, and more in whether startups were able to raise private capital through venture capital or alternative routes.

Performance in screenwriting was interpreted very differently as a characteristic, as the creative process is not necessarily a problem to be solved. Performance instead was defined by screenwriters as the ability of the model to be able to accurately describe the experience that is being created, and was considered of less importance relative to other characteristics such as actionability.

1.2.2 Scalability

The scalability characteristic examines a given models ability to perform and fulfill the expectations of other characteristics as the size of the application increases. For the football environment, the ability to describe both individual systems, the performance of subsystems, and the performance over a set of matches was critical.

For modelling startup ecosystems, a model is considered scalable if it effectively applies to small startup ecosystems, such as South Australia or industry sectors, as well as large ecosystems.

Lastly in screenwriting, the ability to apply a technique to an individual scene, an act, an entire screenplay, or a season of episodes was critical for screenwriters to be able to comparatively assess elements in different contexts. Furthermore, the experience of the arc should be the same for the viewer or reader at a given point in time regardless of the resolution of the analysis.

1.2.3 Comprehensibility

Comprehensibility is another very important characteristic that describes a model's ability to be understood. The characteristic has different meanings in different scenarios, and is often referred to as *explainability*, *intelligibility* or *interpretability* depending on the domain and application (Gadzinski & Castello (2022)).

The value of this characteristic has become rapidly increasing, as the use of black box machine learning models in high-stakes decision making is causing significant problems in a range of applications (Rudin (2019)). Simply "dumping data into "smart" algorithms is not the silver bullet" (Flath & Stein (2018)), and the need to understand modelling processes and outcomes has both moral and governance implications.

In medicine, ceding decision-making to black box systems without clear reasoning or rationale is seen by many as contravening the profound moral responsibility of clinicians (London (2019). The automation of the criminal justice system also creates a range of serious concerns, as predictive computer systems play increasing roles in every stage of the system, from policing to parole (Wexler (2017)).

There is a growing acceptance that there is a trade-off relationship between a model's comprehensibility and its performance, and research continues to strive to understand this relationship, identify practical challenges and develop best practice approaches (Caruana et al. (2020)).

In each of our domains of interest, the decision makers that were consulted had little interest in "black-box" models, particularly given that they would be personally accountable for actions that would potentially be taken. This lens of accountability will significantly shape the application of artificial intelligence into the future, and has implications for not just individual decision makers, but also the structure of corporate governance (Hilb (2020)).

1.2.4 Actionability

There was significant demand from decision makers in each domain for any developed model to produce results that were actionable.

Actionability in the domain of football relates to a coach's ability to direct players and staff either through strategic direction to players on game style, development of targeted set-play configurations, directing the on-field positioning of players, or team selection. It can also extend to personnel list selection, popularised in baseball by the 2003 book,

Moneyball: The Art of Winning an Unfair Game (Lewis (2004)). From consultation with coaching departments of three different AFL teams, a consistent sentiment was communicated about statistical observations of their sport. Readily available player possession data was useful for evaluating player performance, however, broader application to strategic game-plan development was limited. In short: "it's all well and good to predict whether we win or lose, but what can I do about it?".

In the startup ecosystem domain, as the decision makers for our analysis are those developing and implementing public policy regarding support for startups and ecosystem transformation entities, actionability relates directly to available "policy levers". This can include grant funding programs, investment into physical infrastructure, or providing services to support startup transformation.

For screenwriting, actionability relates to a writer's ability to interpret the results of the natural language processing model, in our case story arc visualisation, to enable discussion and to potentially lead to changes to the script.

1.2.5 Justifiability

Lastly, we look at justifiability. A model that fulfills the other criteria of being scalable, comprehensible, actionable, and has sufficiently high performance, is considered to be useless if it cannot be justified.

This speaks less to the construction of the model, but rather to its relationship with established heuristics from both the decision makers, and how they would typically justify their reasoning to their subsequent stakeholders. To evaluate this, an understanding of the stakeholder environment in each domain was critical, but given the multiplicity of stakeholders within each industry of interest, priority was given to the primary relationships. For example, in Australian Rules Football, the fans are obvious stakeholders of each club, but the coaching stakeholders are primarily interested in being able to justify strategy to those executing the strategy.

Domain	Decision maker of interest	Primary stakeholders
Australian Rules Football	Coach	Coaching staff and players
Startup ecosystems	Policymakers	Startups and ecosystem actors
Film and television industry	Screenwriters	Screen industry

Table 1.1: Primary stakeholders in each of the domains of interest.

1.2.6 Importance of characteristics in each domain

Our five characteristics are not necessarily of equal importance in each domain, and the iDSS approach does not combine the five characteristics into a single metric of usefulness. In many domains, a characteristic might be linearly correlated with usefulness, and in

others it might be a step function. Rather than attempt to unify the characteristics, our analysis evaluates each domain as qualitatively different. The relative importance of each of the domains is summarised below to highlight these differences:

Characteristic	Australian Rules Football	Startup Ecosystem	Film industry
Performance	Medium	Medium	Medium
Scalability	Low	High	High
Comprehensibility	Medium	Medium	Medium
Justifiability	High	High	Low
Actionability	High	High	High

Table 1.2: Relative importance of each characteristic in each domain.

1.3 Outline of thesis

This thesis is separated into six chapters. We begin Chapter 2 by providing a background of research into model usefulness in human decision making, in particular the intersection of data science and machine learning techniques with model interpretability and actionability.

This is followed by relevant research into small-data modelling of complex systems where humans perform critical tasks and cannot be readily replaced by automation. We introduce several important concepts for disentangling data, particularly pairwise performance metrics, Markov chain modelling, and natural language processing.

In Chapter 3, we explore our first modelling approach, modelling Australian Rules Football as spatial systems rather than individual possession events. Several methods are introduced to disentangle relative team performance and the functioning of sub-systems to evaluate historical games. This approach is optimised and proposed as a methodology for supporting coaching decision making regarding team selection and evaluation, as well as a potential future framework for setting fairer fixtures. The 2015, 2016 and 2017 Australian Football League seasons are analysed to develop the model.

In Chapter 4, we evaluate startup transformation pathways, working with limited data in the South Australian startup ecosystem to map startup capitalisation. We follow 151 startup journeys over an eight year period to develop an approach to support policymakers to understand ecosystem transformation, with a focus on grant interventions and private capitalisation events.

In Chapter 5, we explore creativity and the writer's room, working with the writers of the television program *Aftertaste* to evaluate the limits and potential for natural language processing to support writers rooms. By approaching the intersection of creativity and data analysis from the direction of usefulness, we are able to evaluate an existing method for story arc generation and rethink the approach to make it more suitable for use on individual projects.

Finally, we conclude in Chapter 6 by reflecting on the role of data science modelling in these three domains of human action with a discussion and outlook for further work. This includes a summary of results across the three domains, the relationship between governance and data science projects, and areas for further research in each domain, as well as more broadly with regards to usefulness.

Literature reviews and connections to academic work in each of three domains are contained within each chapter.

1.4 Publication of work

Methods developed in our domains of interest demonstrate significant progress in each field and the author intends to publish each of the works independently.

The research into new modelling techniques for Australian Rules Football, which features in Chapter 3, has been published in the Journal for Quantitative Analysis in Sport under the title "Modelling Australian Rules Football as spatial systems with pairwise comparisons" (Andreacchio et al. (2022)).

Two additional papers are currently in development, focusing on the research into modelling startup transformation pathways, shown in Chapter 4, and new natural language processing techniques for story arc generation and analysis, from Chapter 5.

Chapter 2

Background

In this chapter, we introduce core concepts and methods that are used in our exploration, including Markov chain analysis, pairwise performance metrics, natural language processing and machine learning techniques for model testing. These concepts will be used to develop new analyses, methods and results in each of our three domains, and enable discussion about model usefulness characteristics.

Further domain-specific background will be introduced in each chapter, including targeted literature reviews.

2.1 Markov model

A Markov model is a stochastic model used to describe changing systems (Markov (1971)). Markov models require a central assumption that future states depend only on the current state, and not previous states that have led up to a current point in time. This assumption enables us to simplify complex system analyses that might otherwise have been intractable.

2.1.1 Discrete-time Markov chains

The simplest type of Markov model, a discrete-time Markov chain, describes a chain of events that move states at discrete time steps.

This can be represented as a sequence of random variables $X_1, X_2, X_3, \ldots$ with the Markov property that the future state only depends on the present state and not on the previous states. This can be represented as:

$$\Pr(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x | X_n = x_n)$$

if both conditional probabilities are well defined. The values of X_i are taken from a countable set, defined as the state space of the Markov chain.

To describe the transition probabilities between states of a Markov chain for a given temporal unit, we use a transition matrix to represent the state transition probabilities. If the probability of moving from state i to state j in one discrete time step is given by: $Pr(X_{n+1} = j|X_n = i) = P_{ij}$, the matrix T is constructed by placing P_{ij} in the i-th row and the j-th column. This transition matrix allows us to describe the stepwise transformation of the Markov chain as follows:

$$T = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1n} \\ \vdots & & & \\ P_{n1} & P_{n2} & \dots & P_{nn} \end{bmatrix}.$$

Since the sum of transition probabilities from a state i to all other states must be 1,

$$\sum_{j=1}^{S} P_{ij} = 1.$$

The Markov chain model can be expanded to include decision processes with the addition of actions and rewards. This is referred to as a Markov decision process and is important for future applications of this research.

2.1.2 Absorbing states and absorbing probabilities

In Markov chain models, a state can be defined as an absorbing state if it only communicates with itself. In other words, state i is an absorbing state if:

$$\Pr(X_{n+1} = i|X_n = i) = 1.$$

We can then define the absorbing probability a_{ik} as the probability of being absorbed in the absorbing state k when starting from the initial state i.

$$a_{ik} = \lim_{n\to\infty} \Pr(X_n = k|X_1 = i).$$

2.2 Pairwise performance comparisons

Pairwise comparison is generally any process where pairs are compared to determine which entity is preferred, or successful. In our research, we are interested in the use of pairwise comparisons to disentangle data where relative participant skill is measured. We will use this in Chapter 3 for our analysis of Australian Rules football.

A simplified model for sports is that the probability that Team A beats Team B is defined by a function of each teams' strength or proficiency (Aldous (2017)). This can be described further, such that if each team i has some strength x_i, when teams A and B play:

$$P(A \text{ beats } B) = W(x_A - x_B)$$

for a specified win-probability function W. W must satisfy the following conditions (which we regard as the minimal natural conditions):

$$W : R \to (0, 1) \text{ is continuous and strictly increasing,} \tag{2.1}$$

$$W(-x) + W(x) = 1 \text{ and } \lim_{x \to \infty} W(x) = 1. \tag{2.2}$$

Condition (2.1) defines that the win-probability function maps from the set of real numbers onto $(0, 1)$ enables the result of a game, where the difference or relative difference in team score is in the set of real numbers, to be mapped to a standardised measure for match outcome, where 0 is defined as a win for the away team and 1 a win for the home team. The function has to be continuous and strictly increasing as a team's strength occurs within a continuous distribution, and a better team is always expected to have a higher probability of winning.

Condition (2.2) explains that a win is only expected to be 100% certain if a team is infinitely stronger than the opposition, and as the difference between two opponents broadens, the expected outcome tends towards 100%. The sum of the likelihoods of each team winning is equal to 100%, which means that there are no other outcomes to the pairwise model.

There are several techniques for evaluating W, including Elo, Glicko, and Glicko-2 (Glickman (1995)).

2.2.1 Elo rating system

The Elo rating system is a method for calculating the relative skill levels of players in zero-sum games (Elo (1978)). Named after its creator, Arpad Elo, the rating system is commonly used in chess, but also in other sports (Aldous (2017)).

The rating systems has two components: estimation of outcome based on ratings of opposing players, and an update mechanism for updating the ratings after new information unfolds.

Supposing Player A has a rating of R_A and Player B a rating of R_B, using the logistic curve for expected outcome, we say that the probability of Player A winning is given by:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}.$$

Similarly, the expected score for Player B is

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}.$$

The selection of 10 as the basis for the exponential, and 400 for the denominator, are arbitrary, and are values that are typically chosen for Elo models. This selection indicates that a difference of 400 points between two players corresponds to the stronger player having a 90% probability of winning.

Once a game has concluded, the outcome of the game can be used to update the ratings, with R' denoting the updated ratings, S_A the outcome of the game and K the update factor:

$$R'_A = R_A + K(S_A - E_A).$$

The K-factor is the upper bound of the rating change from a single game, since when the predicted outcome tends towards being completely certain, yet the actual outcome is the opposite, the ratings for each player are updated by K.

Additional updating variables can be introduced, including an additive variable to the rating to represent home-ground advantage in the expected outcome. Defining this as HGA, we can say:

$$E_A = \frac{1}{1 + 10^{(R_B - (R_A + HGA))/400}}.$$

2.3 Natural language processing

To complement our use of Markov chain analysis and pairwise performance metrics, natural language processing is utilised in Chapter 5 for screenplay analysis. Natural language is the type of language used in every day conversation and writing, and is important for our evaluation of creative writing in story analysis.

Natural speech, however, is not designed for machine analysis, and the way that a human conveys information is not easily understood by computers. The field of natural language processing (NLP) studies interactions between humans and computers and is a subfield of linguistics and computer science (Chowdhary (2020)).

2.3.1 Tokenisation

The foundation of natural language processing is tokenisation, a process where individual words or phrases are interpreted as data points for analysis. For screenplays, each word can be interpreted as action or dialogue, and attributed to a scene and a character if the text is dialogue.

2.3.2 Sentiment analysis

Sentiment analysis, or opinion mining, is the contextual analysis of text to extract subjective information (Feldman (2013)). Often used for analysing social media streams, sentiment analysis is enabled by natural language processing to study affective states.

Different dictionaries can be used based on the objectives of the modelling, such as *AFINN*-lexicon or *Senti-Strength* methods for Twitter Sentiment Analysis (Islam & Zibran (2017)). These dictionaries contain information attributed to each word, such as the emotional valence or polarity, which enables each word to be mapped as a sentiment data point.

Dictionary approaches can be very powerful, but they are often limited in their ability to handle the complexities of language, as they do not account for context, sarcasm, or relationships with adjacent words.

2.3.3 Sentiment arcs

The trajectory of sentiment indicators over time can provide insight into narrative arcs for stories, movies and other written works by taking a moving average of sentiment over the course of a text. A big data analysis of Project Gutenberg's 1,327 stories demonstrated that these arcs can be clustered to show that there are six core emotional arcs to stories (Reagan et al. (2016)), demonstrating the ability for natural language processing to provide quantitative insights to story structure.

Sentiment analysis of movie dialogues has also been utilised to create measures such as *Utterance Emotion Dynamics* to measure dialogue valence and arousal. This enables measures such as emotional variability, rise and recovery rate, peak distance and displacement count and length of dialogue emotion (Hipson & Mohammad (2021)).

We will use this in Chapter 5, evaluating the usefulness of the sentiment arc approach for supporting story development, and exploring new methods to increase the usefulness to screenwriters.

2.3.4 Screenplay analysis

Screenplays have a unique structure, with different formats for action and dialogue, and conventions around scene headers and numbering (Field (1982)). When interpreted as text corpora for natural language processing, the structure of the text makes examination of the narrative more complex.

A range of techniques can be used to extract insights from a screenplay, such as parsing dialogue to understand social networks in movies (Agarwal et al. (2014)), evaluating the presence and participation of characters of different genders (Agarwal et al. (2015)) (Selisker (2015)), and analysing narrative turning points (Papalampidi et al. (2019)).

Researchers have strived to create new works using natural language processing and artificial intelligence. Recent advances have yielded results, with a proposed AI model writing under the pseudonym "Alyce Garner Peterson" able to effectively generate a screenplay from an underlying story (Eldhose et al. (2021)).

2.4 Forgetting curves

Forgetting curves model the decline of information retention over time. We evaluate this in Chapter 5 as we review the methodology for story arc generation, and explore methods for modelling the retention of sentiment information.

The rate and nature of this curve has been the subject of significant debate. The Ebbinghaus savings function proposes that forgetting is produced by two factors, time and interference and can be modelled by the strength and fragility of memory (Ebbinghaus (1885)). This can be reduced under typical conditions to the Wickelgren Power Law, a power law where m is considered the memory coefficient over time t, λ is the state of long term memory at $t = 0$, and $\beta, \psi > 0$ (Wixted et al. (2007)):

$$m = \frac{\lambda}{(1 + \beta t)^{\psi}}.$$

Choosing variables $\psi = 0$ or $\beta = 0$ produces a curve with no forgetting or memory decay.

Other proposed distributions approximate forgetting as an exponential curve, measuring R as retrievability, against stability of memory S over time t (Woniak et al. (1995)).

$$R = e^{-\frac{t}{S}}.$$

For our analysis, we use the simplified Wickelgren Power Law. Bayesian analysis of the Wickelgren Law has shown to favor a power law distribution rather than an exponential distribution (Averell & Heathcote (2011)).

2.5 Statistical modelling techniques

Lying at the intersection of computer science and statistics, statistical modelling techniques and machine learning can be used for building predictive models. Techniques have different degrees of interpretability, and different approaches are used to assist with understanding the importance of predictors and variables, as well as the precision and accuracy of our models.

2.5.1 Regression analysis

Regression analysis is utilised to estimate the relationship between predictors and a response variable. Most regression models take the form where response variable $Y_i = f(X_i, \beta) + e_i$, where X_i is an independent variable, Y_i is a dependent variable, f is a function, β is an unknown parameter, and e_i are error terms to account for statistical noise. The general linear model is used for p independent variables, such that:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$$

where x_{ij} is the i-th observation of the j-th independent variable.

To model this, we utilise the least squares technique that finds the value of β that minimises the sum of squared errors, where the error $e_i = \sum_i (Y_i - f(X_i, \beta))^2$ or the more general expression above.

2.5.2 Decision tree learning

Decision tree learning is a method of hierarchical supervised learning that uses decision trees to move from observations about an item (branches) to conclusions about the item's target value (leaves) (Suthaharan (2016)).

There are two types of categories: classification trees (Gupta et al. (2012)) and regression trees (Loh (2011)). Both produce simple visual aids that assist users with understanding relationships between variables and data, and are popular for supporting decision-making processes.

In Chapter 3, we use classification and regression trees to analyse multiple team ratings and measures, utilising the R package *rpart* (Therneau et al. (2019)).

2.5.3 k-nearest neighbours

The k-nearest neighbours (k-*NN*) algorithm is a machine learning technique that determines the probability of an outcome based on comparison of similar events across a given number of variables (Fix & Hodges (1989)).

The method is a supervised learning classifier, where for each test point, the k nearest training points are identified from a given training set. The output of the model is dependent on whether the model is being used for classification or regression.

For k-*NN* regression, the output is the average of the values of the k nearest neighbours, whereas for k-*NN* classification, the output is class membership. The class is assigned based on the most common among its k nearest neighbours.

We use these for our Australian Rules Football analysis, evaluating the performance of our new system metrics by exploring the predictive power of games using the k-*NN* method.

2.5.4 Confusion matrix

To understand the effectiveness of our models, we are interested in more than just predictive accuracy. We utilise receiver operating characteristics (ROC) graphs to organise classifiers and visualise performance (Fawcett (2006)).

To explain ROC, we first define the confusion matrix (fixed threshold) which summarises a model's predictions. The confusion matrix can be visually represented as shown below:

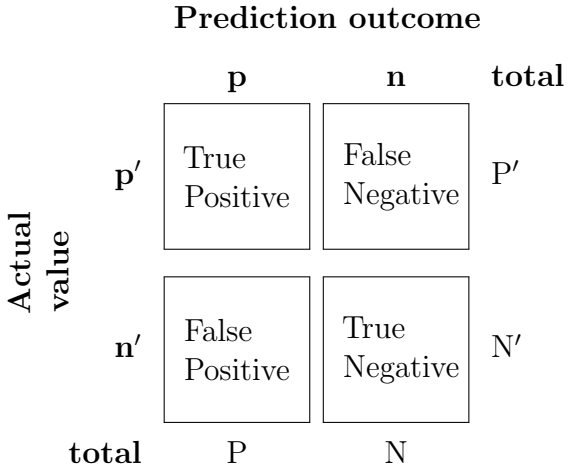

Figure 2.1: Confusion matrix example

The results in each box show the relationship between the the prediction outcome from a model and the actual value.

- True Positive, or TP, is defined as an outcome where the model correctly predicts the positive class.
- True Negative, or TN, is defined as an outcome where the model correctly predicts the negative class.
- False Positive, FP, is defined as an outcome where the model incorrectly predicts the positive class.
- False Negative, or FN, is defined as an outcome where the model incorrectly predicts the negative class.

These test results can be further evaluated to provide insight into the model performance.

The True Positive Rate, or TPR, is the probability that the actual positive will test positive. This is also defined at the recall or sensitivity.

$$\text{True Positive Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The False Positive Rate, or FPR, is the probability that the actual negative will test positive.

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}.$$

Similarly, the False Negative Rate and True Negative Rate can be calculated.

$$\text{True Negative Rate} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{False Negative Rate} = \frac{\text{FN}}{\text{TP} + \text{FN}}.$$

The Accuracy, Precision and Prevalance of the model can be calculated based on these values such as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{Prevalence} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

The F_1 Score is defined as the harmonic mean of precision and sensitivity, and is calculated by:

$$\text{F}_1 \text{ Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

A range of other measures can be derived from these indicators.

Chapter 3

Modelling Australian Rules Football as spatial systems with pairwise comparisons

3.1 Introduction

Statistical analysis is used widely across Australian Rules Football to develop insights into player performance (McIntosh et al. (2018)), possession value (O'Shaughnessy (2006)), team play style (Greenham et al. (2017)), and passing networks (Braham & Small (2018)). The majority of statistical modelling techniques however, focus on the analysis of possession event data, which has three significant limitations that limit the applicability and usefulness for developing coaching strategy.

First, the opposition matters. Teams have strengths and weaknesses, and event data for a given team needs to be interpreted based on the proficiency or strength of that team's opponent.

Second, possession event data does not describe off-the-ball play. Positioning and activity of players that do not have the ball are critical factors to understanding the context of a possession or chain of events.

Third, given the complex and multi-variate nature of Australian Rules Football, there is insufficient possession event data available for meaningful analysis, resulting in multivariate approaches having relatively low statistical power. This is a common problem for many complex team sports (Atkinson & Nevill (2001)).

3.1.1 Decision makers and usefulness

As a result of these limitations, research has struggled to be directly applied to football strategy and tends to find most of its value in predicting match outcomes, which has

potential application for gambling or 'tipping' competitions.

Our interest is in supporting the decision makers in the domain, namely coaching staff and football departments. The author had direct experience and relationships with professional clubs in the national competition, the Australian Football League (AFL), through the implementation of trial virtual reality training programs. Relationships were developed with senior coaching staff at multiple clubs through these projects, and discussions regarding existing and proposed statistical modelling techniques consistently revealed a frustration with many existing approaches. Player possession event data is considered useful for analysing player performance, however overall descriptions of the game are limited and do not systematically support strategic decision making.

To expand on our definition of usefulness, we can evaluate the expectations of coaching decision makers.

The model is required to have a high degree of performance, as if the model does not accurately predict game outcomes, it cannot be considered to be useful.

There is also a requirement that the model be actionable, so that insights can be obtained and action can be taken to improve competitiveness or create advantage. For coaches, this may take the form of player positioning, level of aggression of game style, list management decision making, or analysis of opposition strategies.

Scalability is important for the model so that insights can be obtained on an individual game basis, whilst also being applicable for modelling changes over the course of a season.

Comprehensibility and justifiability are also critical to the model development given the complexity of the stakeholder environment. Coaching staff and players are all required to implement strategic directions, often in high pressure environments, and the buy-in of the playing group is directly impacted by its ability to be understood.

3.1.2 Proposed approach

To support coaching personnel and football departments in strategic decision making, and to overcome the limitations described above, we propose an approach for interpreting event data from AFL matches into a model composed of spatial systems. The modelling approach has three components:

- categorisation of possession events into spatial systems that is consistent with player role designations and coaching departments;
- interpretation of events as a function of both a team's proficiency and its opposition's proficiency, using pairwise performance metrics to extract relative team performance that can be updated on a match-to-match basis;
- reconstitution of team ratings for each match-up to estimate the probability of transition between each system.

This approach produces a model that is more interpretable and useful for coaching decision making by providing a clear methodology for analysing upcoming match-ups, and enabling performance analysis to be undertaken on systems rather than players.

Possession event data from official AFL data provider *Champion Data* is used to develop and optimise the model, with a focus on data from the 2015, 2016 and 2017 AFL Premiership Seasons. This source is not publicly available, however it is the industry standard for teams across the AFL. We find that the geometry of different spatial states, home ground advantage, rating sensitivity and seasonal change are all important factors in optimising a model that is useful for coaching departments.

3.2 Defining spatial systems in AFL

Our model first defines spatial systems that can provide a framework for pairwise performance comparison.

From Chapter 2, we described pairwise performance metrics where the probability that Team A beats Team B in a given sport is defined by a function of each teams' strength or proficiency (Aldous (2017)). We expanded this further such that each team i has some strength x_i, when teams A and B play:

$$P(A \text{ beats } B) = W(x_A - x_B)$$

for a specified win-probability function W. W must satisfy the following conditions (which we regard as the minimal natural conditions):

$$W : R \to (0, 1) \text{ is continuous and strictly increasing,} \tag{3.1}$$

$$W(-x) + W(x) = 1 \text{ and } \lim_{x \to \infty} W(x) = 1. \tag{3.2}$$

Our interest is in the functioning of systems within an AFL game, so rather than determining the probability that Team A beats Team B in a given match, we look at the probability of outcomes whenever the ball enters a spatial state. This requires a clear definition of a win or loss for each spatial state, so that the win-probability function has a meaningful interpretation.

AFL teams separate their player roles into Forward, Midfield and Defensive players (Dawson et al. (2004)), with each category having a specialised coach, corresponding trade-craft, defined objectives and general spatial positioning on an AFL ground. Players will move between roles over the course of a game or season, and GPS tracking enables meaningful analysis of individual player behaviour (Brewer et al. (2010)). Our interest however, is in defining state spaces based on the spatial positioning of the ball, and a team's objective in each state. This enables us to contextualise individual player actions into outcomes produced by the entire team. Drawing from research in other football codes,

particularly soccer (Bialkowski et al. (2014)), different segments of the AFL ground are partitioned to enable possession data to be mapped to transitions between segments.

The forward state is defined as the area where goals can be scored from. The AFL ground has arcs marked approximately 50 metres out from goal, which roughly corresponds with a forward line, however it is not uncommon for players to have the ability to score from further away. The distribution of possession location for each goal scored shows that there is a significant decline in goals scored from further out than 50 metres and we select 56m from goal as our limit. Across the 2015, 2016 and 2017 AFL Premiership seasons, 98.9% of goals were scored from within this limit, and goals from further away are considered outliers.

The corresponding defensive state is the reflection of the forward state at the opposing end of the ground. The midfield state is defined as the area between a team's defensive and forward state, with the general objective of moving the ball into their forward state or stopping the concession of the ball into their defensive area. These states enable clear motivational assumptions to be made:

- when a team is in their forward line, the objective is to score a goal.
- when a team is in the midfield, the objective is to move the ball into a position where they can score from, ie. the forward line.
- when a team is in defense, the objective is to move the ball out of the opponent's scoring range into the midfield.

These assumptions enable transitions between states to be classified as clear win or loss outcomes, which is critical for separating transitions between systems into relative team proficiency ratings using pairwise performance metrics. These assumptions are inline with research into team performance indicators, in particular midfield performance and number of forward system entries (Woods (2016)) and offensive or defensive power measurements (Azhari et al. (2018)).

In addition to the spatial systems, goal states are also added to the model as absorbing states. The absorbing probability can then be interpreted as the probability the ball will end up in a particular team's goal from a centre bounce reset.

The five states are shown in Table 1 and Figure 1 below:

State	Description
1	Away team goal state
2	Away team forward line
3	Midfield
4	Home team forward line
5	Home team goal state

Table 3.1: Defined system states in Australian Rules Football

Figure 1 shows a state space diagram overlaid over an Australian Rules Football ground, where $S_{x \to y}$ is the number of transitions from state x to state y.

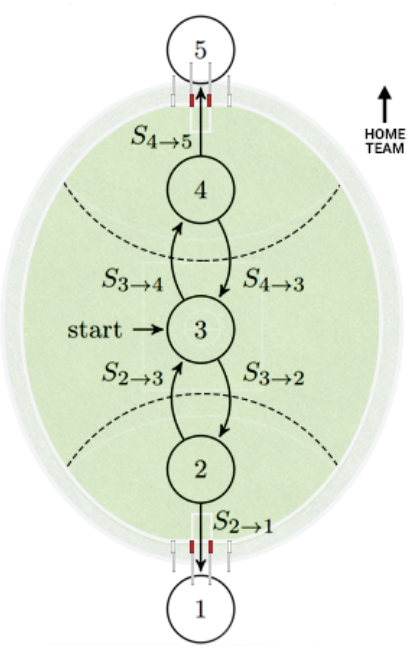

Figure 3.1: State space areas overlaid over a typical Australian Rules Football ground

We are now able to determine the probability that a team scores a goal from a centre bounce by looking at all of the potential transition paths from the midfield state to either goal state. It is appropriate for each goal state to be the end of each transition change as the game is effectively reset after a goal is scored, and the ball is returned to the centre spatial state during a formal break in the game.

3.3 Determining relative team performance of match systems

By interpreting possession event data into spatial state data, we can now separate teams based on team proficiency. Several approaches have been researched in Australian Rules Football to interpret outcomes based on relative team proficiency, including Kalman filters (Leushuis (2018)) and pairwise performance metrics (Ryall (2011), Stefani & Clarke (1992)), however, the focus has been on match outcomes rather than the outcome of spatial states within each match.

We use the Elo rating system (Elo (1978)) as it enables our state space data to be decomposed into team ratings, which can then be utilised to determine expected outcomes

against different opponents.

Recalling from Chapter 2, the expected probability of an outcome, in this case Team A winning (P_A), is based on the ratings of opposing teams A, R_A, and B, R_B:

$$P_A = \frac{1}{1 + 10^{(R_B - R_A)/s}}.$$

.

Here, s is an arbitrary factor and is chosen as 400 based on the original Elo model, which corresponds to a winning probability of 90% for a player that has a rating of 400 greater than their opponent. When the teams have the same rating ($R_A = R_B$), the probability of winning is 0.5. Choosing 400 for our scaling factor is appropriate for each of our systems as it provides a consistent definition and is readily interpretable. Whilst the choice is arbitrary, numerical exploration showed that the overall results are insensitive to the choice of this parameter.

As each match is undertaken, ratings are updated to include the new information as a team's relative performance evolves over time. The updated rating of Team A is given by:

$$R'_A = R_A + K(S_A - P_A),$$

where K is a constant defined as the update factor and S_A is the relative score, or result, for the game. To avoid inflation or deflation of ratings across the league, a team's change in rating is equal to the opposite of their opposing team's change in rating. S_A is defined as the ratio of Team A's final score, Score_A, over the total scores of the game, such that:

$$S_A = \frac{\text{Score}_A}{\text{Score}_A + \text{Score}_B}.$$

This approach can be adapted for our spatial system approach to determine a rating for each team's forward, midfield and defensive system: R_{AFwd}, R_{AMid} and R_{ADef}. The motivational assumptions that have been defined for each state allow us to calculate the expected probability of transition between states, defined as $P_{x \to y}$ where x is the current state, y is the potential future state and $R_{A\text{x}}$ is the rating of Team A in state x:

$$P_{x \to y} = \frac{1}{1 + 10^{(R_{B\text{x}} - R_{A\text{x}})/s}}.$$

The midfield game has two potential successful outcomes for each team: either an entry into the forward line, or a goal from the midfield. The forward line was defined as being within 56 metres from goal to capture the significant majority of goals scored, with goals from the midfield to be considered outliers. This enables us to assume that the probability of transition from the midfield state to each goal state directly is 0.

It is also assumed that the number of events that occur within a state is not of interest in this case, only the transition, such that the probabilities can therefore be calculated directly from the teams' system ratings:

$$P_{2\to3} = \frac{1}{1+10^{(R_{B\text{Fwd}}-R_{A\text{Def}})/s}} = 1 - P_{2\to1}$$

$$P_{3\to4} = \frac{1}{1+10^{(R_{B\text{Mid}}-R_{A\text{Mid}})/s}} = 1 - P_{3\to2}$$

$$P_{4\to5} = \frac{1}{1+10^{(R_{B\text{Def}}-R_{A\text{Fwd}})/s}} = 1 - P_{4\to3}$$

Rather than update the ratings after every individual transition, an AFL match can be processed using the tournament approach for processing Elo ratings. This requires all outcomes for a spatial state in a match to be assessed as a ratio, rather than a binary win or loss, and is utilised in other sports when a tournament or multiple games are played. As an example, assessing the outcomes of the midfield spatial state for a given match, where the number of times Team A wins the system is given by $S_{3\to4}$ and the number of times that Team B wins the system is given by $S_{3\to2}$, the spatial state score for Team A is defined as $S_{3\text{A}}$, such that:

$$S_{3\text{A}} = \frac{S_{3\to4}}{S_{3\to4} + S_{3\to2}}.$$

The update function is therefore:

$$R'_{A\text{Mid}} = R_{A\text{Mid}} + K_{\text{Mid}} \left(\frac{S_{3\to4}}{S_{3\to4} + S_{3\to2}} - P_{3\to4} \right).$$

Corresponding update calculations can be performed to update the other system ratings.

3.3.1 Describing AFL as a Markov model

So far, we have constructed a methodology that enables us to describe a team's forward, midfield and defensive system, and expectations of these systems against a given opposition, however to observe the functioning of each system relative to match outcome, we must define a method for combining these systems together.

To do this, we must observe the Markov condition that the outcome of each system is memoryless, and not impacted by what happened prior to entry into that system. This fits intuitively with our Elo approach, as the outcome of each system is based only on team ratings and Elo model parameters.

We must evaluate whether this is a justifiable assumption to make. A midfield system for instance, may function differently depending on whether it is a centre bounce following a goal being score, or a transition play where the ball is quickly moved from defense through the midfield into the forward line. This is a particularly relevant case as in 2019, the AFL implemented a rule to increase the flow of the game, where for each centre bounce,

each team must have six players in their forward 50 arc, six players in their defensive 50 arc, and six players between the arcs.

As our ratings are a broad representation of the relative proficiency of a team's system, the memoryless assumption is a fair one to make, though development of a second-order Markov chain is potentially an area for further exploration.

We also define each transition from one system to another as a discrete step. The model can therefore be defined as a discrete-time Markov chain and the expected probability of transition between each of the state spaces can then be used to populate our model.

Previous applications of Markov chain modelling in the Australian Football League (AFL) have focused on transitions between possession types (Forbes (2006)), however whilst the Markovian property was demonstrated, the focus was on individual possessions as states rather than the spatial system states.

The system ratings of two teams can be used to construct a Markov chain model for an upcoming match. The transition matrix T for the model can be defined as:

$$T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ P_{2\to1} & 0 & P_{2\to3} & 0 & 0 \\ 0 & P_{3\to2} & 0 & P_{3\to4} & 0 \\ 0 & 0 & P_{4\to3} & 0 & P_{4\to5} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

The absorbing probability of Team A's goal state, or equivalently the probability that Team A will score a goal after a centre bounce is:

$$\pi_A = \lim_{n\to\infty} P(X_n = 5 | X_0 = 3).$$

We can derive an equation for this by using the Markov property and the time-homogeneous assumption, such that:

$$\lim_{n\to\infty} P(X_n = 5 | X_0 = 3) = \lim_{n\to\infty} P(X_n = 5 | X_k = 3) = \pi_A, \text{ for all } k.$$

By expanding the absorbing probability across the first two discrete transition steps after a centre bounce, we can use standard arguments to solve for absorbing probablity using our transition probabilities, such that:

$$\pi_A = \lim_{n\to\infty} P(X_n = 5 | X_0 = 3) = \frac{P_{3\to4}P_{4\to5}}{P_{3\to2}P_{2\to1} + P_{3\to4}P_{4\to5}}$$

This result demonstrates that the probability of a goal being scored by either team in two steps, or after a direct forward system entry after a centre bounce, is equal to the probability of a goal being scored by either team in any number of steps.

Similarly, the probability of Team B scoring a goal after a centre bounce is:

$$\pi_B = \frac{P_{3\to2}P_{2\to1}}{P_{3\to2}P_{2\to1} + P_{3\to4}P_{4\to5}}$$

We can then use these goal-scoring probabilities as estimates of the win probabilities for each team.

3.4 Defining and tuning model parameters

To tune the model, match event data from the 2015, 2016 and 2017 AFL Premiership Seasons was utilised. The available data is focused on individual player possession events, so classification of each event into state-space enabled transitions is required.

The parameters that require tuning are:

- the initial team ratings R_0 for each system, since the ball is returned from a forward system to the centre system more often than a goal is scored;
- the update factor K for each system, which changes the sensitivity of ratings updates based on each match outcome;
- home ground advantage HGA, and whether it is different for each system, and;
- seasonal decay factor j, to describe the off-season changes to a team.

The state space boundaries are a key variable to determine when the ball is in a midfield state as opposed to a forward, or defensive state. This is defined as the radius of the circle from each goal position, and from observation of goal scoring locations in our data, the radius was selected as 56m. Manual selection of this distance was appropriate given the motivational assumption of forward states identified earlier.

We found that win prediction is an inappropriate measure to tune these parameters, as the number of accurately predicted wins produced a noisy surface that made optimisation processes ineffective. This was particularly pronounced when predicting close games, as small changes to the tuning parameters produced different win outcomes in inconsistent ways. Instead we optimise the Pearson correlation coefficient between the absorbing probability and the relative score outcome as it produces a smooth surface for optimisation and enables the magnitude of a score difference to be considered.

This was considered a justifiable approach as the effectiveness of the team strategy can be evaluated as the probability to score relative to the opposition.

The optimisation approach selected is the L-BFGS-B algorithm (Zhu et al. (1997)), enabling simple bounds to be placed on each variable.

We also choose the initial values for state space ratings for each team, $R_{0\,\mathrm{Fwd}}$, $R_{0\,\mathrm{Mid}}$ and $R_{0\,\mathrm{Def}}$. For the midfield, 1500 is selected as the initial rating, inline with the default

for Elo ratings. From observation, a forward win of scoring a goal is less likely than a defensive win where the ball is cleared into the midfield. The average defensive rating is therefore higher than the average forward rating and these variables are selected to be symmetrical above and below 1500. Tuning is therefore only undertaken on the initial defensive and forward ratings.

Home Ground Advantage has been demonstrated to have a non-trivial impact in the AFL. This has been quantified using the average margin of home teams (Clarke (2005)), as well as least squares method and exponential smoothing (Stefani & Clarke (1992)). As our approach uses a pairwise comparison, the home ground advantage (HGA) is defined as an additive factor to the home team's rating (Ryall (2011)). This can be applied to each of the system ratings in our model and included as separate parameters in the optimisation. Therefore

$$P_{2\to 3} = \frac{1}{1+10^{(R_{BFwd}-(R_{ADef}+HGA_{ADef}))/400}} = 1 - P_{2\to 1}$$

$$P_{3\to 4} = \frac{1}{1+10^{(R_{BMid}-(R_{AMid}+HGA_{AMid}))/400}} = 1 - P_{3\to 2}$$

$$P_{4\to 5} = \frac{1}{1+10^{(R_{BDef}-(R_{AFwd}+HGA_{AFwd}))/400}} = 1 - P_{4\to 3}.$$

The strategic environment in the Australian Football League evolves significantly over time, with new innovations unlocking changes to how teams approach the game with regards to possession, zone positioning and other tactics (Woods (2016)). These major innovations tend to occur between seasons, where teams have the time and flexibility to redesign their game-plan and approaches to team strategy. Distinct periods of change have been identified across 2001 to 2015 (Woods et al. (2017)).

Furthermore, between seasons, teams often undertake significant personnel and coaching changes, and the off-season period often enables time for player injuries to be overcome or managed.

As a result of these changes to team structure, composition and strategy, the ratings from a previous season might not be appropriate at the start of the new season, and so a seasonal decay factor is introduced (Ryall & Bedford (2010)). This factor reverts a team's rating towards the initial rating by a constant proportion, to account for some, but not all, information from previous seasons factoring through. Defining for Team A, j_A as the seasonal decay factor for Team A, R_A as the ratings at the end of a season, $\tilde{R}_A$ as the ratings at the start of the following season, and R_0 as the initial calibrated rating of the system being measured (e.g. 1500 for the midfield ratings):

$$\tilde{R}_A = (1-j_A)(R_A - R_0) + R_0 = j_A R_0 + (1-j_A)R_A.$$

If the seasonal decay factor is at 1, then $\tilde{R}_A = R_0$ and when there is no ratings decay across the season change, $j_A = 0$, then $\tilde{R}_A = R_A$.

3.5 Predictive accuracy of the system model

To avoid over-fitting, random samples of 80% of games were selected for optimisation against the Pearson correlation coefficient. Each optimised model was then tested on the remaining 20% of games to determine the accuracy of the model in predicting the outcome. This was repeated for 100 samples using the full set of games across the 2015-2017 AFL Premiership Seasons.

Analysis was undertaken using R 3.6.2 (R Core Team (2019)), the tidyverse (v1.3.0; Wickham et al. (2019)), rpart (v4.1.15; Therneau et al. (2019)), stats (v3.6.2; R Core Team (2013)) and rattle (v5.3.0; Williams (2011)) packages.

The seasonal decay factor was introduced to model teams' reversion towards the initial rating at the end of each season. However, to further evaluate the impact of the off-season changes, the sampling approach was also undertaken on a subset of the matches. The first five rounds of each season were were still used to update ratings, but they were excluded from the training and test sets. This enabled the distribution of each optimised variable to be analysed across two sample sets to evaluate whether the model needs time for new information in each season to update ratings appropriately.

A comparison of the model results is shown in Table 3.2. The median values of each variable are also displayed as several of the variables exhibited a skewed distribution. The distribution of true positive, true negative, false positive and false negative rates were also analysed for each optimised model.

The mean accuracy of the models in predicting the outcomes of games across the full season data was 66.38%, and 67.81% when excluding the first five rounds of each season. This is in the range of other statistical modelling approaches but is significantly more interpretable and useful given the ability to separate relative team performance of spatial systems.

Whilst the average accuracy was 66.38%, the model was significantly better at predicting home team wins at 74.19%, compared to away team wins at a rate of 58.58% accuracy. This result is shown in the confusion matrix in Table 3.

The difference between these two rates of win prediction is significant, however it is a comparable difference to the the natural home team win rate of 56.9% of games and the natural away team win rate 43.1% across the 2015 to 2017 period.

We can also evaluate the home ground advantage factor, which was negligible for forward and midfield lines, but material for defensive lines. This can also be interpreted as a disadvantage for away team forward lines, which is plausible given unfamiliar ground conditions, wind, and potentially hostile crowds.

Evaluating the variables for the models optimised across the full season data, the update factor K for the forward and defensive states had a lower standard deviation than the midfield update factor. This indicated that the model was better at optimising the update factor of this system, however the midfield update factor had less variation when excluding the first five rounds of the season.

Dataset	**Full Season**			**Excluding First 5 Rounds**		
Variable	**Mean**	**Median**	**Std. Dev**	**Mean**	**Median**	**Std. Dev**
K_{Fwd} / K_{Def}	95.19	95.53	15.06	68.50	70.78	17.15
K_{Mid}	59.62	60.66	22.75	23.16	20.98	10.30
HGA_{Fwd}	0.36	0	1.88	7.45	0	10.40
HGA_{Mid}	2.03	0	3.95	0.57	0	2.24
HGA_{Def}	41.94	37.77	28.32	29.76	25.71	21.13
j_{Fwd} / j_{Def}	0.12	0	0.17	0.58	0.58	0.19
j_{Mid}	0.68	0.69	0.14	0.84	0.84	0.11
$R_{0\text{Fwd}}$	1326.20	1316.40	47.63	1325.30	1312.85	49.26
Accuracy	0.6638	0.6694	0.0387	0.6781	0.6804	0.0450
TP	37.06%	37.10%	3.91%	38.75%	39.18%	4.32%
TN	29.31%	29.03%	3.69%	29.06%	28.87%	4.21%
FP	12.90%	12.90%	2.53%	13.95%	13.40%	2.72%
FN	20.73%	20.16%	3.18%	18.24%	17.53%	2.57%

Table 3.2: Predictive accuracy of optimised spatial system model based on sampling across 2015, 2016 and 2017 AFL Premiership Seasons

The seasonal decay factor showed significant difference between forward and defensive systems compared to midfield systems. When the first five rounds were excluded, the seasonal decay factor for each system was significantly higher. This is due to our optimisation not requiring games to be accurately predicted in the first five rounds, giving time for the model to update to newer information without requiring accurate predictions.

The seasonal decay factor appeared to be significantly more relevant for a team's midfield, particularly when optimised across the full season of games. This is potentially due to the impact of strategic planning across the offseason affecting the midfield system more than forward or defensive systems, or greater sensitivity to personnel changes, whether returning from injury or through transition between clubs.

When the first five rounds were excluded, the update factors changed considerably. The home ground advantage for the midfield and forward systems were mostly zero, however the home ground advantage for the defensive system was again significant. The seasonal decay factors were more significant, which is expected given the games had the first five rounds of each system to resolve.

Based on the results of the optimisation process, parameters can be selected to produce a model for further analysis, shown in Table 3.3. As the improvement in accuracy was only marginal when the first five rounds were excluded, the values for the full season are utilised.

Variable	Selected Value
K Fwd/Def	95
K Mid	60
HGA Fwd	0
HGA Mid	0
HGA Def	40
Decay Fwd/Def	0
Decay Mid	0.68
Initial FWD Elo	1325

Table 3.3: Selected variables for spatial system model based on distribution of tuned parameters.

The confusion matrix in Table 3.4 demonstrates that the model is significantly better at predicting home team wins than away team wins.

		Actual Outcome		
		Home Win	**Away Win**	**Total**
Predicted Outcome	**Home Win**	222	78	300
	Away Win	131	189	320
	Total	353	267	620

Table 3.4: Confusion matrix of spatial state model across 2015, 2016 and 2017 AFL Premiership Seasons

The accuracy of the model generated by the selected values is summarised in Table 3.5. The periods of highest accuracy are rounds 6 through 23, with the Finals and Grand Final significantly lower. This is potentially due to the small sample size, however, it indicates that there are potentially other factors in play. The year 2016 had the highest accuracy and 2017 the lowest level of accuracy.

The relationship of projected absorbing probability of the home team and the score outcome for matches is shown in Figure 3.2. Using greater projected absorbing probability as the indicator of likely game winner, the top right and bottom left quadrants show accurately predicted game outcomes.

	Win prediction accuracy			
Rounds	**All**	**2015**	**2016**	**2017**
All	66.29%	65.05%	71.01%	62.80%
1 to 5	60.00%	51.11%	66.67%	62.96%
6 to 23	68.99%	69.74%	73.20%	64.05%
Finals	51.85%	55.56%	44.44%	55.56%
Grand Final	33.33%	0.00%	100.00%	0.00%

Table 3.5: Predictive accuracy across 2015, 2016 and 2017 AFL Premiership Seasons

Figure 3.2: Distribution of projected home team goal scoring against actual match outcome

3.6 Line ratings as a description of team strategy

Using our model, the evolution of teams' forward, midfield and defensive ratings can be examined to understand winning strategies relative to the competition. Team profiles are often analysed for clustering (Spencer et al. (2016)) including factors such as forward line entries and goal conversion rates, and we can now compare the effect of individual line performance against the goal absorbing probability, which is a factor of an entire team's system proficiency.

Comparing the three teams that won the AFL Premiership in 2015, 2016 and 2017 reveals that different game-styles were successful in different years. Figure 3.3 shows the evolution of each team's ratings across the three seasons, with the grey area representing the range across all teams in the league.

Figure 3.3: Comparison of team system ratings for Premiership winners across 2015, 2016 and 2017 AFL Premiership seasons.

In 2015, the Hawthorn Football Club won the AFL Premiership, and had a notably higher forward system rating, indicating that forward system conversion of forward entries to goals was a winning strategy in that year. In the following year, the Footscray Football Club (known as the Western Bulldogs) won with a strong midfield rating compared to the other competitors, though had limited advantage with regards to their forward or defensive systems. Notably in 2017, the Richmond Football Club won the Premiership with a defensive rating that was significantly higher than the other two clubs in that year.

To understand the extent that individual system ratings and absorbing probability are useful measures for predicting match outcome, regression and machine learning methods can be employed (Rosli et al. (2018)). Analysis is undertaken on all matches from the 2015, 2016 and 2017 AFL Premiership seasons, utilising the values previously selected in Table 2.

Logistic regression provided a useful insight into the relative importance of factors. This has been demonstrated to be a useful approach in other football codes, where logistic

regression of forward and defensive lines have been used successfully for outcome prediction (Prasetio et al. (2016)).

Three logistic regression models are analysed, with results shown in Table 3,6. The first model consists of both teams' system ratings and absorbing probability as variables, the second contains only the system ratings, and the final model is solely the absorbing probability.

Parameter	**Model 1**	**Model 2**	**Model 3**
Intercept	43.57	25.84	23.70
Home Team Forward Rating	0.02103	0.008533	
Home Team Midfield Rating	0.006986	0.02067	
Home Team Defensive Rating	0.01749	0.005093	
Away Team Forward Rating	-0.02132	-0.008624	
Away Team Midfield Rating	-0.08037	-0.03109	
Away Team Defensive Rating	-0.02367	-0.01118	
Absorbing Probability	-34.37		-11.52
AIC	744.8	743.2	738.9

Table 3.6: Comparison of logistic regression models using different variable combinations

Model 3 has the lowest AIC value, indicating that it is the best model for the data. To examine the coefficients of Model 3, we can assess a team entering a given match with home team absorbing probability of 0.5, indicating that the teams are approximately equal after the defensive home ground advantage rating has been included. There is an expected probability of home team victory of 58.17%, indicating that the home ground advantage factor in the spatial system models does not completely summarise the general home ground advantage.

Using Model 3, for there to be an equal expected probability of each team winning, the home team absorbing probability is 48.42%.

Whilst Model 3 had the lowest AIC value, Model 2 is insightful in understanding the relative importance of home and away team line ratings in predicting match outcome. Midfield ratings have significantly higher coefficients than the forward or defensive ratings, indicating that the midfield proficiency has a greater impact on forecasting match outcome than the proficiency of other systems. Model 1 coefficients do not provide similarly clear insights, given the absorbing probability is a non-linear function of both team's line ratings.

Exploring the system ratings data using classification and regression trees also yields insightful results. Unpruned trees show interesting sub-classifications, as seen in Figure 4 below. For a home team absorbing probability above or equal to 52%, a home team victory is predicted with an accuracy of 83%. For an absorbing probability of less than 52%, there are further branches including individual line ratings and further segments of the home absorbing probability.

Using the *rpart* package (Therneau et al. (2019)), we can optimise the complexity parameter to prune the tree and avoid overfitting. This parameter is specified by how much the cost of a tree is penalized by the number of terminal nodes.

Figure 3.5 shows that the once optimised, the only variable that is useful is the home team goal state absorbing probability, and the individual system ratings do not add additional value. If the home team has a goal absorbing probability of greater than or equal to 52%, they have an 83% probability that the home team will win the game. If the home team absorbing probability is less than 52%, then there is a 57% chance of the away team winning the game, and 43% chance of the home team winning the game.

From a coaching strategy perspective, this shows that the functioning of the team as a whole, particularly when interpreted through the goal state absorbing probability, is more important in predicting and understanding the impact on match outcome than each individual system in isolation.

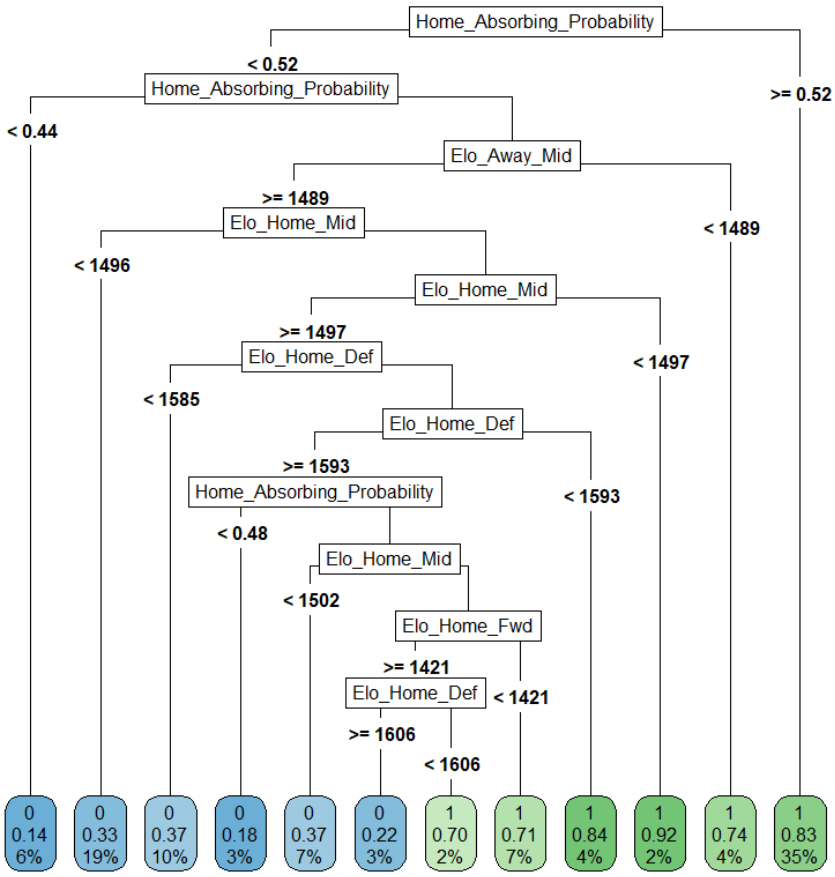

Figure 3.4: Classification and regression tree of system ratings and home team absorbing probability across the 2015, 2016 and 2017 AFL Premiership Season.

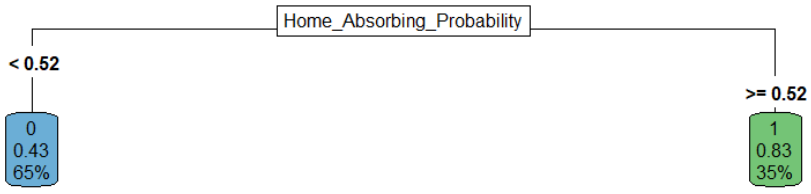

Figure 3.5: Classification and regression tree of system ratings and home team absorbing probability across the 2015, 2016 and 2017 AFL Premiership Season, optimised for complexity parameter.

3.7 Discussion and usefulness

The proposed modelling approach enables event data in Australian Rules Football to be reinterpreted to understand the relative impact of each team's forward, midfield and defensive systems, as well as factor in a given opponent's system proficiency. In addition to predictive value, the model enables a clear understanding of the value of each system and allows for the simulation of incremental change on each system and the resultant change in absorbing probability.

The results of our model indicated that over our period of interest, the midfield performance of teams was less variable than their forward and defensive systems, demonstrated by the difference in optimised update factor. Home ground advantage seemed to be limited to the defensive system, which can also be interpreted as a disadvantage for the forward lines of away teams.

This fulfills our criteria for usefulness, and after engagement with AFL clubs and coaching departments, shows significant promise for implementation.

This also creates a significant opportunity for further research, such as analysing the effects of personnel changes on system ratings to support list management and team selection. As a player is substituted either into the side, or into the game off the bench, the ratings can be analysed to understand a player's relative impact on a given system.

Additional states could potentially be included such as centre-bounce states or wing states but systems require clear objectives for either team to enable the win-loss outcomes to be measured using pairwise performance measures. A potential centre-bounce state would separate the scenario from the general midfield state, given the 2019 rule change requiring six players from each team to be in each of the three ground states at each centre bounce.

Currently, our model does not use the history of how the ball moved into each space, for example whether a forward entry was a fast transition from a team's defensive system through the midfield into the forward line, or whether the opposing team exited their defensive line only for a team to cause a turnover and re-enter their forward line. The Markov chain model can also potentially be reconstructed as a second-order Markov chain, to explore an additional memory to each transition.

Alternative pairwise comparison approaches can also be explored further, including methods that account for uncertainty in a team's skill level such as Glicko-2 (Glickman (1995)), more sophisticated Elo models that have variable update factors dependant on the scoring outcome, and more generalised Bradley-Terry models which have proven effective in other sports such as basketball and soccer (Cattelan et al. (2013)).

Additional states could potentially be included such as time-dependent centre-bounce states or wing states. As discussed, however, this would require clear objectives for either team to enable the win-loss outcomes to be measured using pairwise performance measures. A potential centre-bounce state would separate the scenario from the general midfield state as in 2019, a rule was added to require six players from each team to be in each of

the three ground states.

Despite these limitations, the approach is successful in translating possession event data into system analysis, and is a valid and powerful tool for coaching departments and match analysts.

The approach also has significant potential in its application to other sports, particularly soccer given the similar importance of player spatial positioning and identifiable forward, midfield and defensive systems. Research has been undertaken into soccer as a Markov process to estimate zonal variation of team strengths (Hirotsu et al. (2022)) and understand coordinative structures (Shafizadeh et al. (2013)), however, our pairwise comparison approach has the potential to contribute to further research.

Chapter 4

Capitalisation pathways of South Australian startups

4.1 Introduction

Startups play an important role in modern economies and are increasingly recognized as drivers of potential future economic growth (Amezcua (2010)). Whilst there is a broad and well-recognised historical relationship between government-financed innovation and new startup creation (Fabrizio et al. (2007)), policy approaches in Australia have been largely focused on attempts to emulate the business environment in Silicon Valley that has grown since the 1970s (Mattar (2008)). This includes strategic support for a range of entities and programs designed to support ecosystem development, such as incubators, accelerators, grant funding programs and a variety of subsidies.

Little is known about the value of these programs as their effectiveness on both the individual startups and the broader ecosystem is difficult to measure (Cohen & Hochberg (2014), Hallen et al. (2014), Amezcua (2010)). This poses a significant challenge, not just for policymakers and startups, but also the broader economic environment as corporations seek to engage with startups to enhance corporate innovation (Weiblen & Chesbrough (2015)).

The objective of our analysis is to develop an approach that is useful for supporting policy development, particularly for regions that have relatively new and evolving startup ecosystems. We do this by identifying policy decision makers in South Australia to understand the data that they have available, challenges that they face, and insight into the usefulness of model outcomes.

4.1.1 Decision makers and available data in startup capitalisation pathways

In August 2019, the South Australian Department for Innovation and Skills issued a call for Expressions of Interest to provide a "System of Systems Analysis" of the South Australian entrepreneurship ecosystem. The aim of the project was to obtain "deep understanding and ongoing evaluation of the South Australian entrepreneurial ecosystem" and the document described an interest in understanding the ecosystem as a "complex system" to help inform policy development, improve on existing interventions, and explore new areas of opportunity.

To explore the problem, the author engaged with the Department for Innovation and Skills, in particular the Office of the Chief Entrepreneur of South Australia, to understand the scope of the problem, the objectives of decision makers, and the available data.

Data was available on 152 startups in South Australia over the period from 2012 to 2019, with a focus on tracking grants received and capital raised. The data set commissioned reports including the South Australian Early Commercialisation Fund, Axant SA Startup Reports (2018, 2019) and Techboard (2019).

Consultation with the Department revealed that there were several weaknesses with the dataset, namely:

- inconsistent definition of startups, as opposed to small businesses;
- a perceived lack of feasible methods for systematic data capture and reporting;
- the variety of startups and their different circumstances make them difficult to compare;
- skepticism of some data points given the reliance on startup self-reporting;
- the prevalence of startup "theatre" and the projection of success.

Given that there is limited data available, a significant number of variables to be accounted for, and presence of complex human decision making, new approaches are needed that don't rely on big data analysis. We look to our usefulness characteristics to evaluate the startup ecosystem in South Australia and develop a useful model.

Based on the weaknesses in our data set, we define three research questions to provide insight for policy makers. First, what is the effectiveness of grant funding in unlocking private capital? One of the primary objectives of grant funding was to encourage venture capital or angel investment, but the impact of previous and current interventions was challenging to understand.

Second, is it possible to empirically understand where the system needs the most support? This was important given the range of recommendations and requests directed at policymakers from the startup ecosystem, from focusing on encouraging more early

stage startups to providing targeted support to more advanced startup that have already raised capital.

Last, can we measure the effectiveness of different programs, such as incubators and accelerators? There are a variety of programs that appear critical to the ecosystem, however, it is challenging to disentangle the actual impact of the support from outcomes in raising private capital or long-term system sustainability.

4.1.2 Model requirements for usefulness

Having established the need for the model and an understanding of the limitations of available data, usefulness characteristics can be evaluated. Recall that these include: *performance*, *scalability*, *actionability*, *justifiability* and *comprehensibility*.

For the first characteristic, *performance*, there is a need for the proposed model to accurately describe the startup ecosystem, enable scenario modelling to test the effectiveness of interventions, and provide sufficient insight to answer and explore the critical questions posed by the key decision makers.

Scalability in this scenario can be interpreted as meaning that the modelling approach needs to be relevant for subsets of data, and maintain effectiveness as the ecosystem potentially grows, changes or shrinks. It also means that the models needs to describe the South Australian ecosystem so that it can be compared to other systems that are at different stages of maturity, such as interstate or overseas.

To define *actionability*, the model must be able to be provide insights that enable actions to be taken by policymakers, particularly the Department for Innovation and Skills. Potential actions may include analysis of existing programs to understand the effectiveness, projections of the future ecosystem to support planning, or scenario analysis based on different intervention options.

Comprehensibility and *justifiability* are critical to the model's usefulness, given the interventions are using government funds and require significant accountability to stakeholders, in particular the taxpayer. A "black-box" approach would not be satisfactory to stakeholders across the ecosystem in understanding why some actions are taken and not others, and the ability to clearly justify decisions is critical to underpinning programs and maintaining support.

4.1.3 Markov chain approach

The proposed approach utilises a Markov chain analysis on the startup ecosystem in South Australia, where there is limited data availability and a small-data approach is required. The approach is based on describing the startup ecosystems based on capitalisation events.

Startup capital-raising states are often used to describe the stage of a startup's financing, such as "Pre-seed", "Seed", "Series A" and "Series B" investment. These definitions signal to capital markets and other stakeholders the stage of a startup's development or maturity

(Islam et al. (2018)), and whilst they are often self-reported, it is useful as a descriptive tool. These states, however, are not standardised measures and further quantification is required.

To do this, the proposed approach utilises this concept of capitalisation states based on observations of the distribution of startup capitalisation in South Australia. This enables startups and their capitalisation journeys to be objectively compared, overcoming the stakeholder concerns regarding the validity and objectivity of startups self-reporting their stage of development.

This state-based approach is suitable for regions with relatively new ecosystems with limited available data, as it enables companies and transformations to be generalised into groups that are suitable for the size of the ecosystem. Furthermore, it aligns with our motivation to develop modelling techniques that are useful, with a particular focus on interpretability and actionability to support policy decision making.

We consider the model to be a Markov chain as we assume a memoryless property to startup capitalisation. Once a startup reaches a given capitalisation state, we make the assumption that no previous information influences its next step.

The scope of this analysis is limited to startups that have capitalisation events, and it is acknowledged that other modes of company growth are missed in this analysis, in particular self-funding or "bootstrap" strategies (Vanacker et al. (2010)). This limits the modelling approach to a subset of the startup ecosystem, and after consultation with the identified decision makers, the usefulness of this approach appears based on its ability to understand and interpret the impacts of interventions on stimulating private investment into new startups.

The analysis will be undertaken using three different variables from the dataset: total capitalisation of companies, cumulative grants received, and cumulative private investment received on an annual basis. A hybrid, 2-dimensional model of cumulative grant and cumulative private investment received will then be developed to understand pathways and the effectiveness of grant intervention.

4.2 Startups in South Australia

South Australia is a relatively small startup and entrepreneurship state, accounting for 2% of startup and young company funding in Australia in the 2021 financial year (Techboard (2021)). The Government of South Australia has implemented significant policies to grow the South Australian startup ecosystem and encourage entrepreneurship, including the establishment of an Office of the Chief Entrepreneur, appointment of a voluntary Chief Entrepreneur, and significant support for startup capitalisation and ecosystem development (OSACE (2019)).

The dataset used for this analysis has been provided by the South Australian Government (Department for Innovation and Skills) and is composed of information from several

different sources, including:

- Axant SA Startup Report 2018 2019 - report on survey of SA Entrepreneurship ecosystem including employee counts and funding amounts;
- South Australian Early Commercialisation Fund - data from Department for Innovation and Skills grant programs including employee counts, revenue and funding amounts;
- Startup surveys conducted by the Office of the Chief Entrepreneur.

Information about 152 companies is contained within the data, including longitudinal details about capitalisation events with different types of classifications, such as Seed Investment, Angel Investment, Series A, Series B, Grant, Private Investment and Initial Public Offering. The labelling of private investment appears to be inconsistent, and "Series A", "Seed" and "Preseed" appear to be self-defined labels.

The date range for the dataset is from 2012 to 2019, though information from the first three years of this range is sparse. This is potentially attributable to a rapid increase in new startups in this period, but it is more likely that there is more complete data available from more recent events. To manage this in our modelling, all data is included for the model development, but the period from 2015 to 2019 will be the focus of our analysis.

To process the data, the following approach was used to process the events into company transformation pathways:

- Recode the type of capital event to private investment or grant funding;
- Normalise events for inflation by adjusting against the Australian Consumer Price Index;
- Temporal normalisation, sampling companies on an annual basis rather than a per event basis;
- Allocate to capital states rather than events;
- Define how startups enter the "high growth" system;
- Define how startups exit the "high growth" system.

4.2.1 Recoding capital events

The dataset contains a range of information on startup capital events including the size, date, and type of finance received. Given the self-reporting nature of much of the data, the classification of each capital event varies significantly, with variants of "Seed", "Angel Investment", "Private Investment" and "Series A" all used to describe similar events.

As our interest is in understanding the relationship between grant funding and private capital events, the dataset was simplified to size, date, and the type of capital event, reinterpreted to a binary classification of either grant funding or private capital events.

4.2.2 Adjusting capital events for inflation

Given the startup activity was tracked over an 8 year period, it is necessary to adjust capitalisation events to account for inflation. To do this, we used historical data of the Consumer Price Index from the Australian Bureau of Statistics to normalise each grant and private capitalisation event, so that all amounts are expressed in 2019 Australian Dollars.

Figure 4.1 below shows the distribution of events using a log scale.

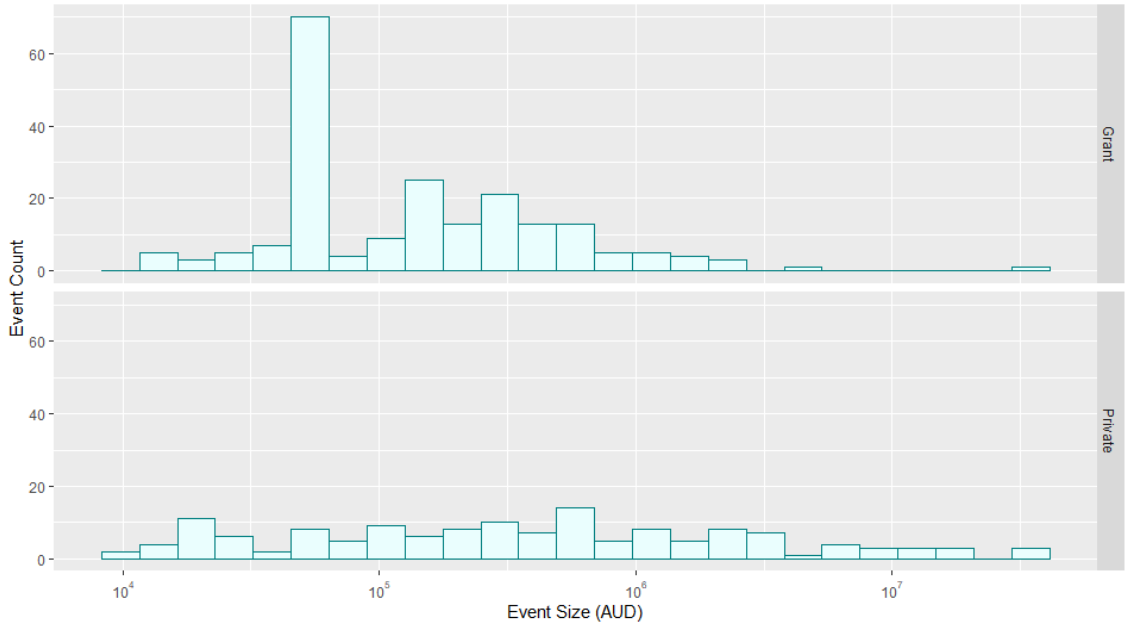

Figure 4.1: Startup capital events distribution in South Australia from 2012 to 2019

We observe that a significant number of grants were issued between $50,000 and $100,000, and the private capital events appear to be more evenly distributed.

4.2.3 Temporal normalisation

The frequency of events differs significantly for each company with some companies registering multiple capital events in a given year, and others receiving single events once every few years. To account for this, we define the temporal unit for our model as one

year, meaning if a company receives multiple events in a given year, they are combined as a single transformation.

To continue our processing, events were aggregated so that each startup has an annual profile that includes total capital received, cumulative grants received, and cumulative private capital raised.

4.2.4 Cumulative amounts and state allocations

Next we examine the cumulative capitalization of startups, which will be the foundation for understanding their capitalization "path".

To do this, we observe the quartiles of cumulative capital received for the period of 2015 to 2019, shown in Table 4.1 below.

0%	**25%**	**50%**	**75%**	**100%**
10,093	100,154	253,516	772,242	70,667,285

Table 4.1: Quartiles of cumulative total capital received by startups in South Australia from 2015 to 2019.

Interpreting these quartiles, we say that the lowest capitalization quartile is less than \$100,000, which is intuitively aligned with companies raising "Pre-Seed" or "Seed" funding. In addition, the median of $\$253,516$ appears plausible given the size of the South Australian economy and entrepreneurship ecosystem, and we use these numbers to select capitalization ranges as the basis for our state spaces.

State	**Lower Bound**	**Upper Bound**
1	10,000	101,000
2	101,001	255,000
3	255,001	775,000
4	775,001	unbounded

Table 4.2: Selected states for total capitalization of individual startups in South Australia

We choose values that are slightly above the quartile number, to act inclusively on round numbers of similar domains. For instance, we observe that there is a significant number of 250 000 events over the period of interest, however inflation adjustment makes this slightly more every year. As these are qualitatively similar, we want to include these within the same states.

There is an increasing number of companies in our data over time, from 42 at the start of the 2015 financial year to 152 in 2019, as shown in Table 4.3. It can also be seen that the number in each state increases, aside from State 2.

	Quartile				
Financial year	**1st**	**2nd**	**3rd**	**4th**	**Total**
Lower Bound ($,000)	10	101	255	775	
2015	14	13	7	8	42
2016	18	19	13	11	61
2017	32	22	16	23	95
2018	41	27	28	35	131
2019	44	30	32	46	152

Table 4.3: Number of South Australian startups in selected total capitalisation states from 2015 to 2019

As we are measuring only capital events, companies that ceased operations, were acquired or stopped increasing capitalisation remain in their existing state. We address this in an upcoming section.

4.3 Mapping capitalisation pathways

To understand how startups develop, the transformation of startups between states is examined over time. We now look to understand these pathways.

4.3.1 Startup transformation count matrix

To do this, we first observe the number of companies that move from one state to another, which can in turn enable a matrix to be constructed that counts the number of transitions.

As an example, there were 2 companies that raised sufficient capital to move from State 1 capitalisation to State 2 in the 2015 financial year. By designating the count matrix row as the current year, and the column as the subsequent year, we say that for 2016, the value at row 1 and column 2 is 2. In other words, for a count matrix C where c_{ij} equals the count of companies that move between State i and State j, $c_{12} = 2$.

Using this method, we construct the following count matrix:

$$C_{2015} = \begin{array}{c} \\ \text{State 1} \\ \text{State 2} \\ \text{State 3} \\ \text{State 4} \end{array} \begin{array}{c} \begin{array}{cccc} \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \end{array} \\ \begin{bmatrix} 12 & 2 & 0 & 0 \\ 0 & 12 & 1 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix} \end{array}.$$

Interpreting this count matrix, we can say that 12 companies in State 1 remained in this state and 2 companies that started the year in State 1 moved to State 2. Similarly, we interpret each row to understand the end state of companies at the end of the year.

This can be visualised as a state diagram:

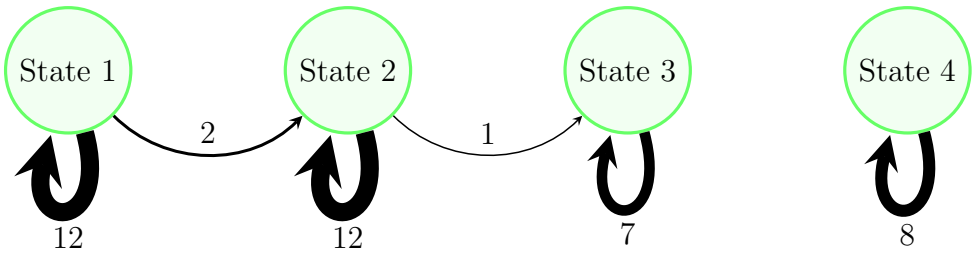

Figure 4.2: State diagram of startup capitalisation in 2015 financial year

The columns can also be interpreted as the number of startups in that state for the next year. These values however, do not completely align with the values in Table 4.3 as whilst we follow the transformations of existing companies, new companies also enter the system.

We can make the further observations that the diagonal values can be referred to as the immobile states, where companies do not change capitalisation in a given year and the values below the diagonal are all zero as companies are unable to reduce capitalisation over time.

4.3.2 New startups entering the ecosystem

For our model to be complete, we need to incorporate new startups entering the system. From our dataset, in 2015, there were:

- 6 new companies in State 1;
- 5 new companies in State 2;
- 5 new companies in State 3;
- 3 new companies in State 4.

To capture this, we add a State 0 for companies that have received no capitalisation. This then enables a row to be added to our matrix to capture this information.

$$C_{2015} = \begin{array}{c} \\ \text{State 0} \\ \text{State 1} \\ \text{State 2} \\ \text{State 3} \\ \text{State 4} \\ \text{Column Sum} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \end{array} \\ \left[\begin{array}{ccccc} 0 & 6 & 5 & 5 & 3 \\ 0 & 12 & 2 & 0 & 0 \\ 0 & 0 & 12 & 1 & 0 \\ 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{array}\right] \\ \begin{array}{ccccc} & 18 & 19 & 13 & 11 \end{array} \end{array} \begin{array}{c} \text{Row Sum} \\ \left[\begin{array}{c} 19 \\ 14 \\ 13 \\ 7 \\ 8 \end{array}\right] \\ \end{array}$$

The sum of each row and the sum of each column are noted to show that the amounts for the 2015 and 2016 financial year in Table 4.3 are exactly the same. We revisit our state space diagram to include these additional transformations, as shown in Figure 4.3 below:

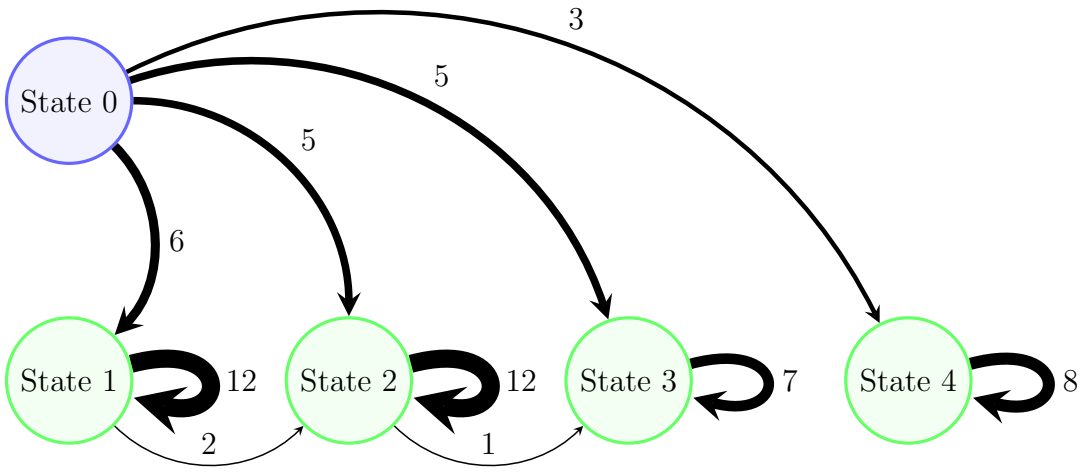

Figure 4.3: State diagram of startup capitalisation in 2015 financial year including no capitalisation state

4.3.3 Startup ecosystem transition matrices

The count matrix provides a simple summary of the number of transformations over a given year, yet for us to make projections and understand the probability of companies moving between states, the matrix needs to be processed into a transition matrix.

To do this, each row can be normalised by dividing by the sum of the row. More sophisticated approaches to calculating these probabilities can potentially be developed, but for the purposes of developing our model, we focus on interpretable methods that are appropriate for our small dataset.

Using C_{2015} again as our example we calculate the transition matrix below:

$$T_{2015} = \begin{array}{c} \\ \text{State 0} \\ \text{State 1} \\ \text{State 2} \\ \text{State 3} \\ \text{State 4} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \end{array} \\ \begin{bmatrix} 0 & 0.316 & 0.263 & 0.263 & 0.158 \\ 0 & 0.857 & 0.143 & 0 & 0 \\ 0 & 0 & 0.923 & 0.077 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

The transition matrix allows us to make observations about the probability of each

potential transformation between states and enables a systematic means of describing the ecosystem and its components. This satisfies our usefulness criteria of comprehensibility and justifiability, and creates a transition matrix that by construction takes advantage of the memoryless assumption required for our Markov chain.

The transition matrix also allows us to describe transformations using a matrix multiplication, enabling direct comparisons between years, and projections across multiple years.

4.3.4 Application of the transition matrix

Suppose we describe the profile of startups at the start of a given year, i, as a vector S_i, where each value is the number of startups in each state. We also include a variable n_i as the number of new startups in the year i.

For 2015, this would be:

$$S_{2015} = \begin{matrix} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \end{matrix}$$
$$\begin{bmatrix} n_{2015} & 14 & 13 & 7 & 8 \end{bmatrix}.$$

Using this vector, we apply the transition matrix to project the startup profile one year into the future. To validate this, we take the 2015 startup vector S_{2015} and the number of new startups entering the system as $n_{2015} = 19$, to show that the S_{2016} profile can be generated.

$$S_{2016} = S_{2015} \times T_{2015}$$
$$= \begin{bmatrix} 19 & 14 & 13 & 7 & 8 \end{bmatrix} \times \begin{bmatrix} 0 & 0.316 & 0.263 & 0.263 & 0.158 \\ 0 & 0.857 & 0.143 & 0 & 0 \\ 0 & 0 & 0.923 & 0.077 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{matrix} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \end{matrix}$$
$$\begin{bmatrix} 0 & 18 & 19 & 13 & 11 \end{bmatrix}.$$

This result can be confirmed comparing the startup profile with Table 4.3. We observe that the resultant vector is zeroed at State 0, and if we are to multiply to project the following year 2017, we would need to add a further value n_{2016}.

The matrix calculation can be rearranged to accommodate this, defining N as the new-startup vector so that:

$$S_{2016} = (S_{2015} + N) \times T_{2015}.$$

Based on our previous calculation N will take the form of a vector with first value representing the new startups, n_{2015}. This also allows us to accommodate startups that might enter the ecosystem with a non-zero accelerated capital state.

4.3.5 Multi-year projections using the startup transition matrix

This approach can be used to project multiple years into the future. Using the transition matrix generated from 2015 as a consistent transition matrix T, and assuming a consistent new-startup vector, we predict two years into the future, rearranging our equation to produce:

$$\begin{aligned} S_{2017} &= (S_{2016} + N) \times T \\ &= ((S_{2015} + N) \times T + N) \times T \\ &= ((S_{2015} \times T + N \times T + N) \times T \\ &= (S_{2015} \times T + N \times T + N) \times T \\ &= S_{2015} \times T^2 + N \times T^2 + N \times T \end{aligned}$$

Assuming the number of new startups is the same as the previous year, we write

$$N = [19 \quad 0 \quad 0 \quad 0 \quad 0]$$

and can therefore calculate that:

$$S_{2017} = \begin{matrix} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \end{matrix} \\ [\quad 0 \qquad 12.43 \qquad 25.10 \qquad 19.46 \qquad 14.00 \ ] \cdot$$

We round these variables as a company cannot partially transform between a state and these values must be integers. The result can also be compared against the actual observed number of startups in each state in 2017.

	Quartile				
Financial year	**1st**	**2nd**	**3rd**	**4th**	**Total**
Lower Bound ($,000)	10	101	255	775	
Projected (2017)	12	25	19	14	70
Actual (2017)	32	22	16	23	95

Table 4.4: State space profile of South Australian startup ecosystem in 2017

We see the actual profile is quite different from the projected startup profile, indicating that the transition matrix and number of new startups weren't consistent in 2017. This is

potentially due to a change in economic factors, the advent of new government programs or a recency bias in the dataset.

We limit the interpretation at this stage until our method is developed further, but this demonstrates the ability to use transition matrices to be able to not only analyse the startup ecosystem, but make projections about startup transformations into the future.

This can be generalised so that for a transition matrix T with consistent startup generation N, we can predict the startup profile S at a time k years from a given year y as:

$$S_{y+k} = S_y \times T^k + N \sum_{i=1}^{k} T^i$$

Using this approach as a foundation, we now develop the model further by adding additional states, evaluating different types of capitalisation, and developing better ways of constructing transition matrices.

4.3.6 Defining additional states

In our model so far, startups either remain in their existing capitalisation state, or increase to a state of greater capitalisation. Over time, this would lead to an accumulation of startups in the upper quartile, yet in practice, startups exit the "high growth" system.

To capture this information, exit states can be defined as additional states. Startups can exit the system for a range of reasons, however, for this analysis we define two categories: "Stable" and "Death" states. Further categories such as merger, acquisition, Initial Public Offering, or others could be defined, but the limited data available means it is appropriate for us to start with as low a resolution as possible.

Firstly, a "stable" state is defined when a startup is no longer participating in the startup capitalisation pathways. There could be a multitude of reasons for this, including reaching a desired size, market interest, or finding an upper limit to growth in a given domain.

By analysing the distribution of time between capital events for each startup in our dataset, a significant majority of capital events occurred within one to two years of a previous event. Hence companies that do not have a capital event for three consecutive years are considered to be "stable".

Secondly, "death" states are defined as occurring when a startup ceases operations. This is defined as when companies formally cease operations or all founders have concluded their employment with the startups. This information is not included within the data provided, so a manual scan of the 152 companies was undertaken.

We explore this by reviewing the 2015 financial year count matrix with the two additional states. We include these as columns to represent transitions when a company in a given row transitions to a stable state or death state. It is possible to not include

Stable or Death states as a row, but for our projection approach, we need the ability to multiply transition matrices together, so adding two zeroed rows enables this transition matrix to be square.

$C_{2015} =$

	State 0	State 1	State 2	State 3	State 4	Stable	Death	Row Sum
State 0	0	6	5	5	3	0	0	19
State 1	0	10	2	0	0	1	1	14
State 2	0	0	10	1	0	1	1	13
State 3	0	0	0	6	0	1	0	7
State 4	0	0	0	0	7	0	0	7
Stable	0	0	0	0	0	0	0	0
Death	0	0	0	0	0	0	0	0
Col Sum		18	19	13	11	3	2	

We note that the row sum is the same for all states except for State 4, where it is one less than the original model. This is due to a company in a previous year exiting the system, rather than accumulating in State 4.

The transition matrix generated from this count matrix can be used to make projections, though our startup profile vector for a given year needs two additional zero states to make the matrix calculation appropriate.

We visualise this expanded model for the 2015 financial year below in Figure 4.4. The count values for each transformation have been removed given the increasing number of paths, and instead the line width represents the relative number of startups moving through the system.

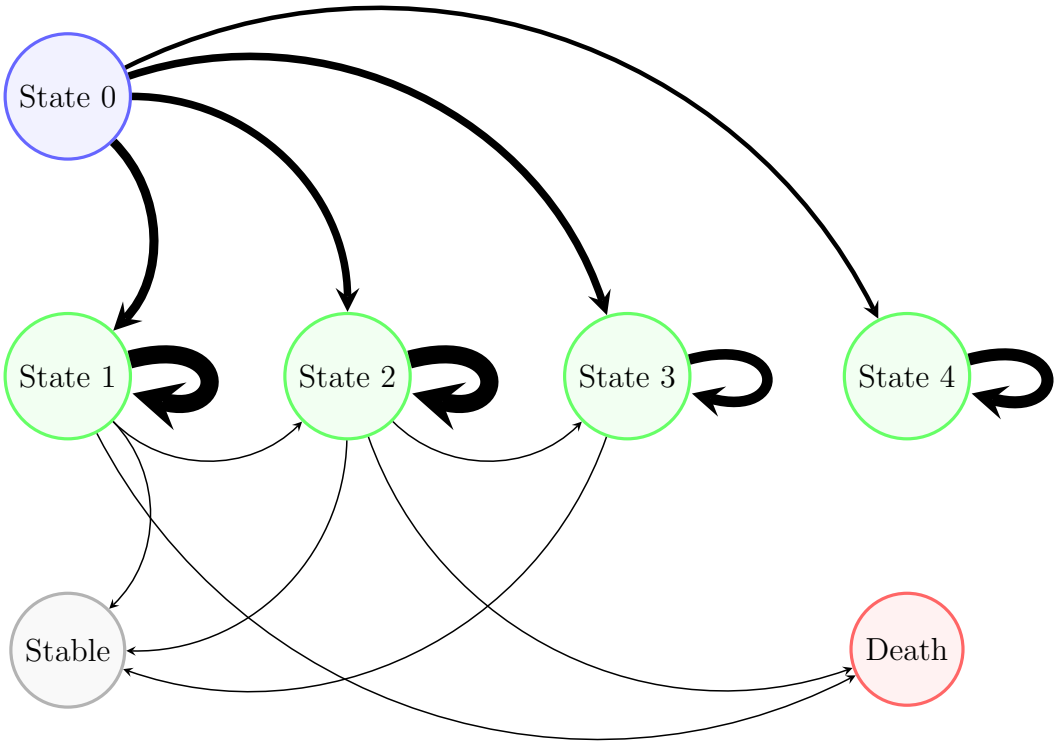

Figure 4.4: State diagram of startup capitalisation in 2015 financial year including exit states

Whilst this approach allows us to describe the startup ecosystem as transformation pathways, the model has limited actionability for stakeholders and does not answer the key questions around the impact of grant funding on private capitalisation.

To explore this further, we evaluate different subsets of capitalisation based on type over a greater time period from our data.

4.4 Startup capital transformation pathways

The model can now be used to evaluate the transformation of startups through the system based on different types of capitalisation. This will enable the comparison of startup journeys from three perspectives: total startup capitalisation, cumulative grant funding received and cumulative private capital raised.

To do this, we take transformations across a multi-year period to create a transition

matrix that can be analysed. The pathways to stable state or death state are also evaluated to enable us to look at the likely exit paths of startups moving through the ecosystem. We do this by looking at the absorbing probabilities, or the probability that a startup in a given state will enter the Stable or Death state, regardless of the path or amount of time.

Examining our dataset, it appears that every year there is an increasing number of startups in the ecosystem, and as noted previously, it is challenging to interpret whether this is the result of a growing ecosystem, or a limitation of the dataset due to recency bias. To manage this, we focus on the period from 2015 to 2019, though data from 2012 is included so there are companies that have been operating for several years included within the analysis.

4.4.1 Model 1 - Total capitalisation pathways

For our first model, we examine the total capitalisation of startups over time to understand their trajectory. As above, our period of interest is 2015 through to 2019, in which case there are four transformations.

To construct our transition matrix over the period, we create a transition matrix for each year and then take the average over the four periods. As we've defined T_i as the transition matrix for a given financial year i, we therefore take the average for total capitalisation as:

$$\bar{T}_{\text{total}} = \frac{1}{4}(T_{2015} + T_{2016} + T_{2017} + T_{2018}).$$

Processing the four periods, this produces:

$$\bar{T}_{\text{total}} = \begin{array}{c} \\ \text{State 0} \\ \text{State 1} \\ \text{State 2} \\ \text{State 3} \\ \text{State 4} \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.385 & 0.174 & 0.230 & 0.191 & 0 & 0.020 \\ 0 & 0.642 & 0.106 & 0.026 & 0.062 & 0.104 & 0.059 \\ 0 & 0 & 0.683 & 0.088 & 0.027 & 0.155 & 0.046 \\ 0 & 0 & 0 & 0.742 & 0.141 & 0.108 & 0.010 \\ 0 & 0 & 0 & 0 & 0.971 & 0.020 & 0.008 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

The transition matrix can be analysed to make observations about the startup ecosystem. These include:

- The state where startups are least likely to move is from State 4;
- The state where startups are most likely to move from is State 1. This can be seen by the immobile probability for State 1 as being 64.2% which means the probability of transformation is 35.8%;

- Startups are most likely to enter a stable state from State 2;
- Startups have the highest probability of entering the death state from State 1;
- New startups enter at a rate of 38.5 % to state 1, 17.4% to state 2, 24.4% at state 3, and 17.6 % at state 4. A small number also ceased activity in their first year.

We also look at the absorbing probabilities from each state into either a stable state company or a death state. It is noted that the objective of many startups is not necessarily to reach a stable state if the total capitalisation is small, however, it is of interest from a policy perspective given the range of startups and the impacts of grant funding on stimulating private investment.

The absorbing probability is defined as the probability that a startup will enter a given state k from any starting state i, regardless of the path or time, and is given by:

$$a_{ik} = \lim_{n \to \infty} \Pr(X_n = k | X_1 = i).$$

Table 4.5 shows the calculated absorbing probabilities from each state.

Initial State	**Stable State Abs. Probability**	**Death State Abs. Probability**
State 0	72.69%	27.31%
State 1	70.13%	29.87%
State 2	77.39%	22.61%
State 3	80.88%	19.11%
State 4	70.97%	29.03%

Table 4.5: Absorbing probabilities using Model 1 (Total Capitalisation) 2015-2019 state transition matrix

An immediate observation is that no matter the starting state, the probability of a startup entering a stable state is much higher than the probability that it will enter a death state. This result is counter to expectations, as whilst there are considerable efforts in precincts around the world to reduce the failure of startups (Cho et al. (2014)), the average industry failure rate around the world is approximately 90% (Krishna et al. (2016)). This result either demonstrates that South Australia has an exceptionally successful startup support ecosystem, or more likely, it is a reflection of its lack of maturity as an ecosystem. High rates of startup failure is not necessary a negative indicator, as the domain is considered a high-risk, high-reward environment. Rather, this high success rate could suggest that the South Australian ecosystem is not tackling high-reward environments.

An evaluation can also be made of specific state absorbing probabilities. The startups with the highest probability of achieving a stable state are those that make it to State 3.

The lowest probability is interestingly, those who have the smallest capitalisation alongside those with the largest.

This is curious as companies with capitalisation momentum appear to have a higher chance of stability, but the drop off at the upper quartile reflects that some companies raise more aggressively looking for a more significant payoff. This approach does not qualitatively look at the types of stable state exits, which is potentially an area for further research.

Three further considerations should be made to contextualise the result. Firstly, the largest capitalisation state has a relatively small sample size. Secondly, the time required to reach a higher capitalisation state or stability means that our data set is limited. And thirdly, the lack of data capture for businesses that fail, or perhaps never raise any capital.

4.4.2 Model 2 - Grant funding pathways

As one of the key objectives is to understand the impact of grant funding on startup transformation, we can re-evaluate the startups and their transformations through the perspective of cumulative grant funding received, rather than total capitalisation.

To do this, the state spaces can be redefined based on an analysis of cumulative grant funding that startups have received. Again using the quartile approach, we examine the distribution of cumulative grant funding states, as shown in Table 4.6, to enable appropriately dense states to be selected.

0%	**25%**	**50%**	**75%**	**100%**
13,040	83,560	151,631	414,481	33,750,000

Table 4.6: Quartiles of cumulative grant funding received by startups in South Australia from 2015 to 2019.

The states are designated g_0 through g_4 to differentiate between our total capitalisation states.

	Grant Funding	
State	**Lower Bound**	**Upper Bound**
State g_0	0	10,000
State g_1	10,000	85,000
State g_2	85,001	155,000
State g_3	155,001	425,000
State g_4	425,001	unbounded

Table 4.7: Selected states for total grant funding of individual startups in South Australia

To calculate the transition matrix, the same method is used of measuring the number of transformations every year from the 2015 financial year through to 2019, aggregation into a series of count matrices, normalisation into a transition matrix for each year and then the calculation of the average probability for each transformation pathway. It should be noted that the definition of whether a company is Stable is independent of the analysis being used. Previously, companies that didn't receive further capitalisation events for a period of 3 years, but continued operations were considered stable. If a company receives a grant, but continues raising private capital for more than three years, it is still considered growing within our ecosystem model.

The mean transition matrix, this time denoted $\bar{T}_{\text{grant}}$, is shown below:

$$\bar{T}_{\text{grant}} = \begin{array}{c} \\ \text{State } g_0 \\ \text{State } g_1 \\ \text{State } g_2 \\ \text{State } g_3 \\ \text{State } g_4 \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{State } g_0 & \text{State } g_1 & \text{State } g_2 & \text{State } g_3 & \text{State } g_4 & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.186 & 0.228 & 0.237 & 0.170 & 0.137 & 0.043 \\ 0 & 0.720 & 0.128 & 0.016 & 0.016 & 0.067 & 0.054 \\ 0 & 0 & 0.683 & 0.085 & 0.040 & 0.168 & 0.025 \\ 0 & 0 & 0 & 0.702 & 0.202 & 0.048 & 0.048 \\ 0 & 0 & 0 & 0 & 0.991 & 0 & 0.009 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

Analysing the transition matrix, we make further observations about the grant funding capitalisation pathways:

- Startups that received grant funding for the first time, received a range of amounts, with probability of entering each state within a similar range;
- The most immobile state was the lowest quartile, between $10,000 and $85,000. This was partly due to fewer companies exiting the ecosystem from this state compared to other states;
- Companies that receive a small amount of grant funding, have a higher probability of receiving more grants rather than exiting the system. Once companies reach state 2, this likelihood of further funding is reduced;
- Startups were most likely to enter a stable state from State g_2 and most likely to enter the death state from State g_1;
- Companies that receive the highest quartile of funding do not have evidence of stable outcomes, though the sample size is small.

The absorbing probabilities into the stable state or the death state can also be evaluated.

Initial State	Stable State Abs. Probability	Death State Abs. Probability
State g_0	39.98%	60.02%
State g_1	50.55%	49.45%
State g_2	57.00%	43.00%
State g_3	16.11%	83.89%
State g_4	0%	100%

Table 4.8: Absorbing probabilities using Model 2 (Grant funding) 2015-2019 transition matrix

Comparing the stable state and death state absorbing probabilities in Model 1 (Table 4.5) and Model 2 (Table 4.8) shows that companies that enter grant funding states have a higher probability of entering the death state than companies through private capitalisation states. This indicates that grant funding is less effective in producing companies that enter a stable state.

To undertake the comparison more directly, a third model looking solely at private capitalisation pathways can be evaluated.

4.4.3 Model 3 - Private capitalisation pathways

We now explore private capitalisation pathways through cumulative private capital raised states. As in Model 1 and Model 2, we define states based on the distribution of cumulative private capitalisation states, using the quartiles to select appropriate states and generate a transition matrix that reflects the probability of transformation between states in a given year.

0%	25%	50%	75%	100%
10,050	101,244	268,900	1,606,000	70,667,285

Table 4.9: Quartiles of cumulative private capital received by startups in South Australia from 2015 to 2019.

Based on these quartiles, we again designate capital states by including the lower bounds of each quartile for a given state. The cumulative private capital states are denoted p_n to differentiate from the total capitalisation and grant funding states. These smaller states have the potential to be assigned more colloquial capital raising states such as angel, pre-seed or seed funding rounds, with state p_4 and perhaps state p_3 being more likely to be assigned to Series capital raising rounds.

State	**Lower Bound**	**Upper Bound**
State p_0	0	10,000
State p_1	10,001	105,000
State p_2	105,001	275,000
State p_3	275,001	1,650,000
State p_4	1,650,001	unbounded

Table 4.10: Selected states for total private capitalization of individual startups in South Australia

As with Models 1 and 2, we construct an annual transition matrix over the 2015 to 2019 period

$$\bar{T}_{\text{private}} = \begin{array}{c} \\ \text{State } p_0 \\ \text{State } p_1 \\ \text{State } p_2 \\ \text{State } p_3 \\ \text{State } p_4 \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{State } p_0 & \text{State } p_1 & \text{State } p_2 & \text{State } p_3 & \text{State } p_4 & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.222 & 0.119 & 0.268 & 0.182 & 0.099 & 0.110 \\ 0 & 0.650 & 0.033 & 0.069 & 0 & 0.187 & 0.060 \\ 0 & 0 & 0.677 & 0.136 & 0 & 0.188 & 0 \\ 0 & 0 & 0 & 0.856 & 0.047 & 0.097 & 0 \\ 0 & 0 & 0 & 0 & 0.956 & 0.023 & 0.021 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

We make a series of observations about the transformation of startups based on the Model 3 transition matrix.

- For new startups, the most likely amount raised is between \$275,000 and \$1,650,000 at 26.8% or less than \$105,000 at a probability of 22.2%;
- Startups that enter state p_1 are unlikely to then subsequently directly enter p_4 as based on the available data, there is no precedent. There is a probability that a startup can enter the state indirectly through State p_3.
- By observing the probabilities along the diagonal, we see that startups are likely to raise at a slower rate the bigger they get, which is an expected observation.
- The highest probability that a startup enters a stable state is from p_2 and p_3 as all observed companies appeared to become stable or raise further capital.
- Companies that raised the most amount of private capital, in the fourth quartile, appeared to have similar stable and death rates. This is potentially due to companies moving from strategies that prioritise high growth to lower risk configurations with different priorities. Additionally, this could be a result of organisational

transformation as the role of the entrepreneurial founder changes more to conventional management structures.

The exit transformations of startups can be evaluated by exploring the absorbing probabilities from each state.

Initial State	Stable State Abs. Probability	Death State Abs. Probability
State p_0	70.71%	29.29%
State p_1	78.91%	21.09%
State p_2	95.75%	4.25%
State p_3	84.42%	15.58%
State p_4	52.27%	47.73%

Table 4.11: Absorbing probabilities using Model 3 (Private capitalisation) 2015-2019 transition matrix

These absorbing probabilities show that companies in South Australia that enter State p_2 have the highest probability of entering a stable state at 95.75%. Notably, those that have raised in the upper quartile, State p_4, have the lowest probability of entering the Stable State at 52.27%.

We compare the transition matrices between Models 2 and 3. As one of our key questions of interest is the relationship between grant funding and private capitalisation events, we subtract the transition matrix from Model 2, which evaluated grant funding quartile states, from the transition matrix from this model. It is noted again that the quartiles are calibrated independently for the grant funding and private capitalisation states, and the difference in the transition matrices is:

$$\bar{T}_{\text{private}} - \bar{T}_{\text{grant}} = \begin{array}{c} \\ \text{State 0} \\ \\ \\ \\ \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{State 0} & & & & & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.036 & -0.109 & 0.031 & 0.012 & -0.038 & 0.067 \\ 0 & -0.070 & -0.095 & 0.053 & -0.016 & 0.120 & 0.006 \\ 0 & 0 & -0.006 & 0.051 & -0.040 & 0.020 & -0.025 \\ 0 & 0 & 0 & 0.154 & -0.155 & 0.049 & -0.048 \\ 0 & 0 & 0 & 0 & -0.035 & 0.023 & 0.012 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

Whilst the quartile states differ, the difference in matrices allow us to test the hypothesis that the transformation of startups between grant funding quartiles and private funding quartiles are the same.

Aside from the new startup State 0, we observe that all of the stable state column are positive, indicating that the probability of a company entering the stable state from

each quartile in the private funding model is greater than the probability of a company entering the stable state from the corresponding grant funding quartile.

A weakness of this approach is that we are comparing two different subsets of companies. Companies that raise no capital, but receive grant funding, are included in the grant model, and companies that receive no grants and only raise private capital exist in the private model. This explains why the column sums do not equate to zero, as if all startups were included in both systems, the difference in transformations entering the stable state should equal zero.

This comparison yields some insight, however, in the next section we propose a 2-dimensional model of states to further explore the direct relationship between grant funding and private capital events.

4.5 Does grant funding unlock private capital in South Australia?

Whilst the transition matrix approach is useful in describing startup capital transformation pathways, it does not directly help us with understanding whether grant funding unlocks private capital events for startups.

To do this, we create a Markov model but construct our state spaces based on two dimensions, cumulative grant funding received and private capital raised. The allows us to look at the relationship between startups as they move through the state spaces in either or both dimensions.

In the previous sections, we utilised the quartiles to create states with sufficient density within our dataset.

We see that if the quartiles in each dimension were utilised for constructing a count or transition matrix, there would be twenty seven states: our stable and death states, and then an array of five by five states including our generative state at (0,0). Furthermore, these matrices will be relatively sparse, given the inability of companies to move backward in capitalisation in either dimension.

As a result, we take the second quartile, the median, to define three states in each dimension. For private capital, we define this no capital (p_0), lower 50% of total private capitalisation (p_1), and upper 50% of total private capitalisation (p_2). Similarly for grant funding, this is defined as no grants (g_0), lower 50% of total grant funding received (g_1), and upper 50% of total grant funding received (g_2).

This enables us to create the subsequent states as a combination of grant funding and private funding (p_i, g_j), shown in Table 4.10.

			Private Capitalisation		
			p_0	p_1	p_2
			$p = 0$	$p \in (0, 275,000)$	$p > 275,001$
Grants	g_0	$g = 0$	(p_0, g_0)	(p_1, g_0)	(p_2, g_0)
	g_1	$g \in (0, 155,000)$	(p_0, g_1)	(p_1, g_1)	(p_2, g_1)
	g_2	$g > 155,000$	(p_0, g_2)	(p_1, g_2)	(p_2, g_2)

Table 4.12: Selected states for total private capitalization and total grant funding of individual startups in South Australia

Using the same methodology as the previous models, we construct a transition matrix to describe startup transformation pathways across the period from 2015 to 2019. As a result, we let $T_{4[2015,2019]}$ equal:

	(p_0,g_0)	(p_1,g_0)	(p_2,g_0)	(p_0,g_1)	(p_1,g_1)	(p_2,g_1)	(p_0,g_2)	(p_1,g_2)	(p_2,g_2)	Stable	Death
(p_0,g_0)	0	0.128	0.128	0.367	0.020	0.013	0.238	0.027	0.059	0	0.020
(p_1,g_0)	0	0.646	0.016	0	0.046	0	0	0.016	0	0.244	0.033
(p_2,g_0)	0	0	0.718	0	0	0.048	0	0	0.127	0.108	0
(p_0,g_1)	0	0	0	0.666	0.075	0.062	0.049	0	0.019	0.079	0.049
(p_1,g_1)	0	0	0	0	0.515	0.061	0	0.030	0.030	0.333	0.030
(p_2,g_1)	0	0	0	0	0	0.811	0	0	0.056	0.133	0
(p_0,g_2)	0	0	0	0	0	0	0.895	0	0.044	0.025	0.035
(p_1,g_2)	0	0	0	0	0	0	0	0.667	0.250	0.083	0
(p_2,g_2)	0	0	0	0	0	0	0	0	0.978	0	0.022
Stable	0	0	0	0	0	0	0	0	0	0	0
Death	0	0	0	0	0	0	0	0	0	0	0

4.5.1 Probability of private capital transformation from different grant states

This transition matrix can be analysed to make observations about the startup ecosystem, and the relationships between grant funding and private capital funding.

To test whether companies that raise grant funding are likely to raise private capital we compare $\Pr((p_0, g_0) \to (p_1, g_0))$ against $\Pr((p_0, g_1) \to (p_1, g_1))$ and $\Pr((p_0, g_2) \to (p_1, g_2))$.

$$\Pr((p_0, g_0) \to (p_1, g_0)) = 0.128$$
$$\Pr((p_0, g_1) \to (p_1, g_1)) = 0.075$$
$$\Pr((p_0, g_2) \to (p_1, g_2)) = 0.$$

This shows that for companies undertaking their first private capital raise up to the median private capitalisation state, the probability of transformation decreases the more grant funding they've received. For new startups, the probability of raising this private capital is 12.8% per annum compared to startups that have received some grant funding, who transform at 7.5% per annum. There was no instance in our data of companies that

had solely received grant funding to place them in the upper half of grant funding amounts raising private capital into this state, and hence the probability of this transformation is 0.

We test this across the upper half of cumulative private capital states, $\Pr((p_0, g_0) \to (p_2, g_0))$ against $\Pr((p_0, g_1) \to (p_2, g_1))$ and $\Pr((p_0, g_2) \to (p_2, g_2))$.

$$\begin{aligned}\Pr((p_0, g_0) \to (p_2, g_0)) &= 0.128\\ \Pr((p_0, g_1) \to (p_2, g_1)) &= 0.062\\ \Pr((p_0, g_2) \to (p_2, g_2)) &= 0.044.\end{aligned}$$

Similarly, the probability of private capital being raised decreases when companies have already received grant funding.

We also test whether companies that raise private capital are more likely to then receive grant funding by considering

$$\begin{aligned}\Pr((p_0, g_0) \to (p_0, g_1)) &= 0.367\\ \Pr((p_1, g_0) \to (p_1, g_1) \text{ or } (p_1, g_2)) &= 0.052\ .\\ \Pr((p_2, g_0) \to (p_2, g_1) \text{ or } (p_2, g_2)) &= 0.056.\end{aligned}$$

To do this, we observe that $\Pr((p_0, g_0) \to (p_0, g_1)) = 0.367$, and $\Pr((p_0, g_0) \to (p_0, g_2)) = 0.238$. This cumulatively shows us that new companies with no funding of any type have a 60.5% probability that they will receive only grant funding over the next year.

4.5.2 Absorbing probabilities of Stable and Death states

The indicators described provide a useful means of comparing transition probabilities to develop insights into the functioning of the startup ecosystem. The primary challenge with this approach is that they represent only direct transformations and do not capture paths where companies move between states before entering a final state.

To address this, we look at the absorbing probability of stable state and death states based on the transition matrix.

Initial State	Stable State Abs. Probability	Death State Abs. Probability
(p_0, g_0)	47.22%	52.78%
(p_1, g_0)	82.35 %	17.65%
(p_2, g_0)	50.16 %	49.84%
(p_0, g_1)	58.16%	41.84%
(p_1, g_1)	79.14%	20.86%
(p_2, g_1)	70.59%	29.41%
(p_0, g_2)	24.14%	75.86%
(p_1, g_2)	25.00%	75.00%
(p_2, g_2)	0 %	100.00%

Table 4.13: Absorbing probabilities using Model 4: 2-dimensional 2015-2019 transition matrix

These absorbing probabilities reveal that as grant funding increases, the likelihood of companies eventually entering the death state increases significantly.

Of all new startups, the probability that companies will end up in the stable state is 47.22%. The highest probability states for companies to end up being stable is those that enter (p_1, g_0) at 82.35% and (p_1, g_1) at 79.14%. Interestingly, as that grant amount increases to state (p_1, g_2) we see this drop sharply to 25.00%

It is noted that the density of data in states (p_1, g_2) and (p_2, g_2) is limited, which is largely due to companies that raise private capital being less likely to receive grant funding, particularly in the upper half of cumulative grant funding.

4.5.3 Does grant funding unlock private capital?

Whilst absorbing probabilities are useful for understanding the probability that startups will end up in exit states, the different transformation pathways of startups are critical to understanding the effectiveness of interventions and support structures.

For instance, for a company to reach state (p_1, g_1) they have three different pathways. They could raise both grant funding and private capital in one year, or start with grant funding followed by private capital in a subsequent year, or conversely, raise private capital before receiving public support.

This becomes additionally complex when the temporal component is considered. A company for instance could receive grant funding, remain in state (p_0, g_1) for a further year, and then raise capital. To account for this, we normalise the probabilities against the probability of the startup moving at all from each state, as well as excluding the stable and death state transformations.

The three pathways are shown below, and can be interpreted as a comparison of conditional probabilities.

Pathway	Probability
$(p_0, g_0) \to (p_1, g_1)$ in one step	2.0%
$(p_0, g_0) \to (p_0, g_1) \to (p_1, g_1)$	$\frac{36.7\% \times 7.5\%}{1-66.6\%} = 8.1\%$
$(p_0, g_0) \to (p_1, g_0) \to (p_1, g_1)$	$\frac{12.8\% \times 4.6\%}{1-64.6\%} = 1.7\%$
Total Probability $(p_0, g_0) \to (p_1, g_1)$	11.8%

Table 4.14: Pathways from startup capitalisation state $(p_0, g_0) \to (p_1, g_1)$

This difference in probabilities confirms that the capital transformation pathways matter. The probability of raising both grant funding and private capital in a given year is similar to the probability that a company will raise private funds first before subsequently receiving grant funding. This is compared to the probability of raising grant funding first which is approximately four times more likely.

This difference can be interpreted in two different ways. Either companies that receive grant funding are more likely to receive private capital, meaning the grant funding is potentially helpful in unlocking private investment and accelerating new startups, or that companies that have already raised capital find it harder to receive grant funding.

To test this, we also compare the probability of obtaining only private capitalisation in the following year for companies that have had no grants, a small amount of grants (in the lower half), and large amounts of grants.

Pathway	Probability
$(p_0, g_0) \to ((p_1, g_0)$ or $(p_2, g_0))$	35.6%
$(p_0, g_1) \to ((p_1, g_1)$ or $(p_2, g_1))$	13.7%
$(p_0, g_2) \to ((p_1, g_2)$ or $(p_2, g_2))$	4.4%

Table 4.15: Comparison of grant funding transformation probabilities from private capitalization states

Similarly we look at the rates of private capital raised once a small amount p_1 has already been raised.

Pathway	Probability
$(p_1, g_0) \to (p_2, g_0)$	$\frac{1.6\%}{35.4\%} = 4.5\%$
$(p_1, g_1) \to (p_2, g_1)$	$\frac{6.1\%}{48.5\%} = 12.6\%$
$(p_1, g_2) \to (p_2, g_2)$	$\frac{25\%}{33.3\%} = 75.8\%$

Table 4.16: Comparison of private capitalization transformation probabilities from grant funding states

These results allow us to make observations about the South Australian startup ecosystem:

- Table 4.11 shows us that significantly more companies receive grant funding before raising private capital compared to the reverse.

- Table 4.12 reveals that the probability of new companies raising private capital decreases the more grant funding they receive.

- For companies that have raised a small amount of private capital, grant funding increases the probability of further private capital being raised.

Based on this analysis, if the intention for policymakers is to increase capitalisation outcomes, the order of grant funding and private capitalisation is important. The most effective means of support for companies to raise capital is to direct grant funding to companies that have already raised a small amount of private capital, not new companies that have only received grant support.

4.5.4 Probability of mobility from each state

We generalise this analysis to evaluate the probability of mobility from each state, as well as decomposing to the probability of transformation in either grant funding or private capitalisation. Using the probability of receiving grant funding as an example:

$$\Pr(\text{Receiving grant funding from } (p_0, g_0)) = \sum_{i=0}^{2} \sum_{j=1}^{2} ((p_0, g_0) \to (p_i, g_j)) = 72.4\%.$$

We apply this methodology to each state to produce the following table of probabilities that a company will transform in a given year. Values are listed as NA when they are in their highest state, as raising private capital when in the top half of capitalised companies means the company will remain in the top half.

It should also be noted that the starting state (p_0, g_0) probabilities can be discounted given that there is no immobile state for new companies, as they are only registered in this analysis once their first capital event occurs.

Starting state	Any Event	Grant	Private Capital
(p_0, g_0)	100%	72.4%	37.5%
(p_1, g_0)	7.8%	6.2%	1.6%
(p_2, g_0)	17.5%	17.5%	NA
(p_0, g_1)	20.5%	6.8%	15.6%
(p_1, g_1)	12.1%	6%	9.1%
(p_2, g_1)	5.6%	5.6%	NA
(p_0, g_2)	4.4%	NA	4.4%
(p_1, g_2)	25%	NA	25%
(p_2, g_2)	NA	NA	NA

Table 4.17: Probabilities of general startup transformation in South Australia in a given year

Each of these probabilities tells an interesting story about start ups in each of these states.

As companies raise more capital, the probability of further raising goes down, which is to be expected given the failure rate of startups. Similarly to previous analysis, we see that for companies who have raised no capital, the probability of raising private capital decreases as they receive more grants.

Companies that have raised private capital have an increasing probability of raising further private capital as they receive more grant funding. From receiving no grant funding, the probability of 1.6% increases to 9.1% after receiving a small grant, and then to 25% if further grants are received.

This leads to the observation that the most effective grant funding goes to companies that have already raised private capital, and has a significant impact on the probability of further private capital being raised.

4.6 Application as a forecasting tool

To understand this approach's usefulness, the comprehensibility, justifiability, performance and scalability have all been demonstrated in the analysis so far. Actionability requires that the model provides not just insights into the current ecosystem and transformation probabilities, but also the ability to model and make projections about future ecosystem profiles given different interventions.

In Section 4.3.5, we defined the following formula for making projections about future ecosystem states:

$$S_{y+k} = S_y \times T^k + N \sum_{i=1}^{k} T^i,$$

where S_n is the startup profile vector at time n, T is the transition matrix, and N is the vector of new startups in each year. y is defined as the starting year, and k is the number of years projected into the future.

This model assumes that the transition matrix and new-startup vector does not change in a given year, which is satisfactory for our analysis.

Three scenarios are implemented for analysis:

- The "Current trajectory" Scenario, where the ecosystem grows using the same transition matrix and new-startup vector as for the 2015 to 2019 perid;
- The "More startups" Scenario, where the profile of new startups entering the ecosystem increases linearly over time;
- "No grant funding" Scenario, where all grant funding is ceased, and hence transformations through non-zero grant funding states are removed.

Each of these scenarios is evaluated for a forecast period that commences following the 2020 financial year, for a period of ten years.

For the purposes of this analysis, we evaluate the ecosystem through the lens of total startup capitalisation to evaluate the overall ecosystem profile. The total capitalisation matrix T_{total} is utilised for the analysis.

4.6.1 Forecast 1: "Current trajectory" scenario and forecast

To understand the "Current trajectory" Scenario, we first recall the transition matrix of total capitalisation from 2015 to 2019:

$$\bar{T}_{\text{total}} = \begin{array}{c} \\ \text{State 0} \\ \text{State 1} \\ \text{State 2} \\ \text{State 3} \\ \text{State 4} \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.385 & 0.174 & 0.230 & 0.191 & 0 & 0.020 \\ 0 & 0.642 & 0.106 & 0.026 & 0.062 & 0.104 & 0.059 \\ 0 & 0 & 0.683 & 0.088 & 0.027 & 0.155 & 0.046 \\ 0 & 0 & 0 & 0.742 & 0.141 & 0.108 & 0.010 \\ 0 & 0 & 0 & 0 & 0.971 & 0.020 & 0.008 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

From observation, the immobile probability of State 4 is 97.1%, which suggests that a company will on average remain in this state for over 30 years before entering the Stable or Death states. This is not justifiable, and is likely a result of a dataset that is too small and a time frame that is too short to capture large scale capital exits. As a result, we review the State 4 row by extrapolating the immobile probability of each state linearly, and proportionally increasing the Stable and Death states.

As the probability of immobility in State 1 is 64.2%, 68.3% in State 2, and 74.2% in State 3, the linear extrapolation of this is 78.9%. This is justifiable as it is plausible that the amount of time between capital events increases as the size of the events increase. The remaining probability of 21.1% is distributed as 15.1% chance of entering a stable state and 6.03% chance of entering a death state.

We therefore use the our transition matrix:

$$\tilde{T}_{\text{total}} = \begin{array}{c} \\ \text{State 0} \\ \text{State 1} \\ \text{State 2} \\ \text{State 3} \\ \text{State 4} \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.385 & 0.174 & 0.230 & 0.191 & 0 & 0.020 \\ 0 & 0.642 & 0.106 & 0.026 & 0.062 & 0.104 & 0.059 \\ 0 & 0 & 0.683 & 0.088 & 0.027 & 0.155 & 0.046 \\ 0 & 0 & 0 & 0.742 & 0.141 & 0.108 & 0.010 \\ 0 & 0 & 0 & 0 & 0.789 & 0.151 & 0.060 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

At the conclusion of the 2019 financial year, the profile of startups in South Australian is given in our dataset as:

$$S_{2019} = \begin{array}{c} \begin{array}{ccccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 32 & 18 & 25 & 31 & 0 & 0 \end{bmatrix} \end{array}$$

We set the number of companies in the stable state or death state as being 0 so the projected number in each state can accumulate over time.

Lastly, the new-startup vector is required. To calculate this, we observe the average number of new startups over the period, with 21 new startups in 2019, 38 in 2018, 31 in 2017 and 19 in 2016. The average of 27 new startups can be assumed moving forward as a baseline, such that:

$$N = \begin{array}{c} \begin{array}{ccccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 27 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

Recalling our projection equation, we now calculate the startup ecosystem profile at the end of the 2030 financial year as:

$$S_{2030} = S_{2019} \times T^{11} + N \sum_{i=1}^{11} T^{i}.$$

This produces the following profile:

$$S_{2030} = \begin{array}{c} \begin{array}{ccccccc} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 29 & 24 & 35 & 56 & 187 & 71 \end{bmatrix} \end{array}.$$

As the Stable and Death states are absorbing, these numbers represent the total accumulated amount of companies that reached these states. The other states represent the number currently in that state at the conclusion of the 2030 financial year.

We track this on an annual basis to examine the changes year on year.

Startup Profile	**State 0**	**State 1**	**State 2**	**State 3**	**State 4**	**Stable**	**Death**
S_{2019}	0	32	18	25	31	0	0
S_{2020}	0	31	20	27	36	13	5
S_{2021}	0	30	22	29	40	28	11
S_{2022}	0	30	23	30	43	44	17
S_{2023}	0	30	23	32	46	60	23
S_{2024}	0	29	24	32	48	77	30
S_{2025}	0	29	24	33	50	95	36
S_{2026}	0	29	24	34	52	113	43
S_{2027}	0	29	24	34	53	131	50
S_{2028}	0	29	24	34	55	150	57
S_{2029}	0	29	24	35	55	168	64
S_{2030}	0	29	24	35	56	187	71

Table 4.18: Forecast of startup ecosystem profiles - Model 1: Current trajectory

We can plot this to show the projected number of companies in each state over our period of interest.

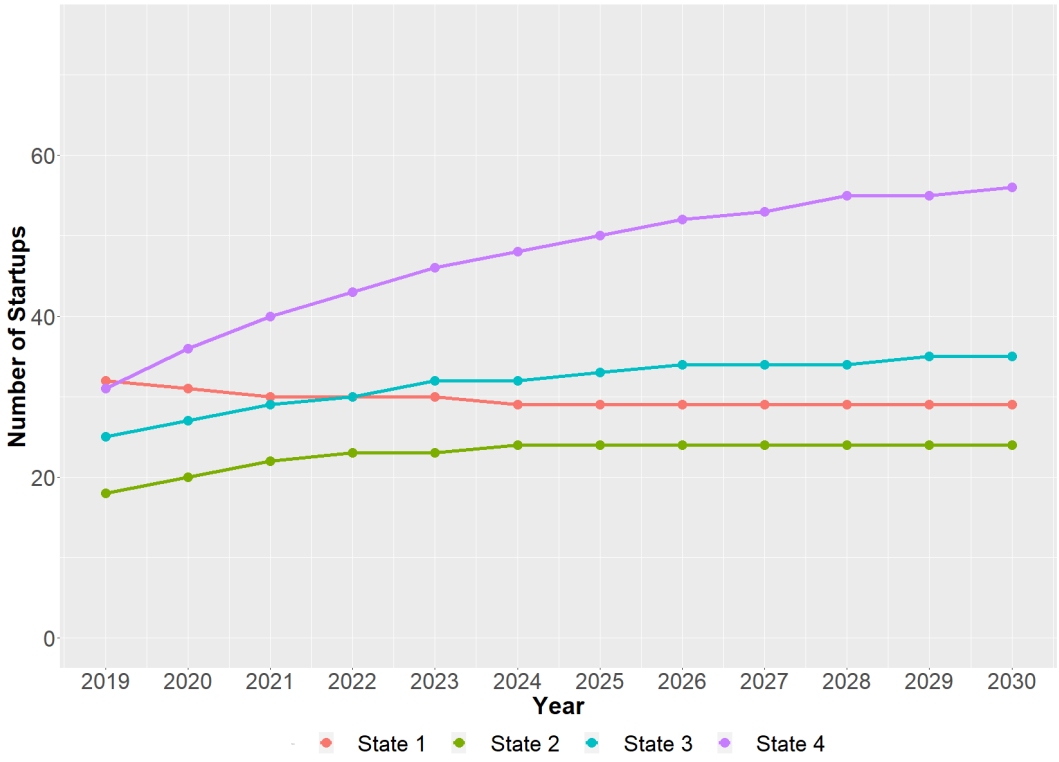

Figure 4.5: Plot of startup ecosystem states - Model 1: Current trajectory

The number of companies in State 1 declines slightly over the 11 year period to a limit of 29. States 2 and 3 increase a similar amount over the same period. State 4 increases rapidly, from 31 to 56, showing an accumulation in the upper quadrant.

4.6.2 Forecast 2 - "More startups" scenario and forecast

For our second forecast, we examine the scenario where there is an increasing number of startups entering the ecosystem. This is aligned with a policy option of focusing all resources on increasing the amount of startup creation.

To determine this increased rate of startup creation, a linear regression is fitted to our data from 2015 to 2019. Over this period, there were approximately 1.3 more startups created each year, so for our model we increase the number of new startups by 1.3 each year. As a result, we use the following new-startup matrix, where n is the number of years after 2019 and:

$$N_n = \begin{matrix} \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} & \text{Stable} & \text{Death} \end{matrix}$$
$$\begin{bmatrix} 27 + 1.3 \times n & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

As our new-startup vector is no longer constant, we revise our equation so that:

$$S_{2030} = S_{2019} \times T^{11} + \sum_{i=1}^{11} N_i \times T^{(11-i)}.$$

Utilising the same transition matrix as Model 1, and the same startup profile in 2019, we forecast the number of startups in each state to 2030. This is shown in the table below:

Startup Profile	**State 0**	**State 1**	**State 2**	**State 3**	**State 4**	**Stable**	**Death**
S_{2019}	0	32	18	25	31	0	0
S_{2020}	0	31	21	27	36	13	5
S_{2021}	0	32	23	30	40	28	11
S_{2022}	0	32	24	32	44	44	17
S_{2023}	0	33	25	34	48	62	24
S_{2024}	0	34	27	36	52	80	31
S_{2025}	0	35	28	38	56	100	39
S_{2026}	0	37	29	40	59	120	46
S_{2027}	0	38	30	42	63	142	55
S_{2028}	0	39	31	44	66	164	63
S_{2029}	0	41	33	45	69	188	72
S_{2030}	0	42	34	47	72	212	81

Table 4.19: Forecast of startup ecosystem profiles - Model 2: More startups scenario

We can again plot this result to show the number of startups in each state over time.

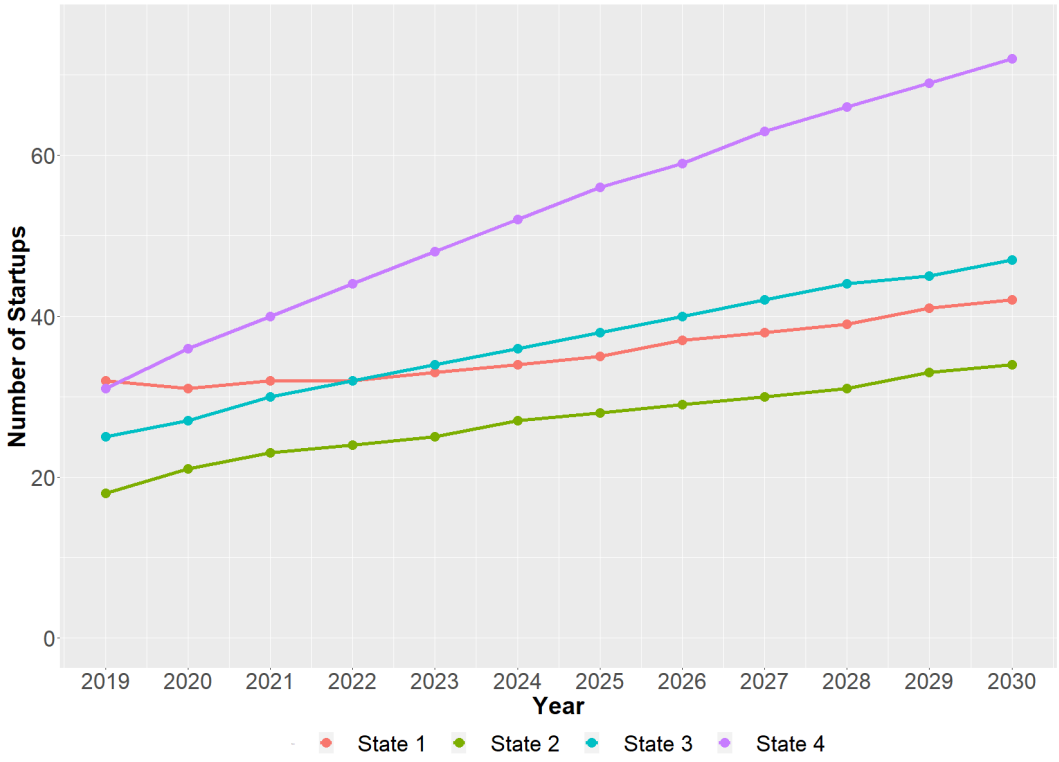

Figure 4.6: Plot of startup ecosystem states - Model 2: More startups scenario

As to be expected, an increasing number of new startups entering the ecosystem results in a similar increase in the number of startups in State 1 over the 11 year period. The number of startups in State 2 and State 3 almost doubled over the period, and the volume in State 4 more than doubled, rising from 31 in 2019 financial year to 72 in 2030 financial year.

This projection is potentially useful for programs that are anticipated to increase the number of startups in the ecosystem over time, which can now be tested based on outcomes over time. In addition, if the number of startups entering the ecosystem did increase, but the startup profiles were not as expected, this would indicate that the intervention is also impacting the transition matrix and various transformation probabilities between states.

4.6.3 Forecast 3 - No grant funding scenario

For our third scenario, we examine the case where all grant funding is ceased and startup capitalisation only increases through private capital events.

There are two options for this analysis, we evaluate only transformations of companies that have not received any grant funding, or we take the private capitalisation approach used in Model 3.

For the first option, we focus on companies that have only received private capital and haven't engaged with grant funding at all in their journey. To do this, we evaluate the 2D transition matrix and select only states where there is no grant funding, being states (p_0, g_0), (p_1, g_0) and (p_2, g_0). The transition matrix can be reduced to the following transition matrix $T_{\text{Private Only}}$, which after renormalisation looks like:

$$T_{\text{Private Only}} = \begin{array}{c} \\ (p_0,g_0) \\ (p_1,g_0) \\ (p_2,g_0) \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccc} (p_0,g_0) & (p_1,g_0) & (p_2,g_0) & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.464 & 0.464 & 0 & 0.072 \\ 0 & 0.688 & 0.017 & 0.260 & 0.035 \\ 0 & 0 & 0.869 & 0.131 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

The transition matrix reveals that the highest stable state transition probabilities and Stable to Death ratios occur in the model, with $\Pr((p_1, g_0) \to S) = 0.260$ and $\Pr((p_2, g_0) \to S) = 0.131$. The matrix also infers that companies in South Australia that only raise private capital, very rarely enter a Death state. This is potentially due to the small sample size, or potentially reveals that companies that are most likely to succeed, avoid the grant funding pathways.

To make projections we must also make an assumption of the number of new startups entering the ecosystem. Keeping our observation that 27 startups enter the ecosystem, and there is a 12.8% chance of entering state (p_1, g_0) and 12.8% chance of entering state (p_2, g_0), we assume that 7 startups enter the ecosystem into these domains each year, such that:

$$N = \begin{array}{c} \begin{array}{ccccc} (p_0,g_0) & (p_1,g_0) & (p_2,g_0) & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 7 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

This number of new startups is significantly fewer than the amount in the other models, as we are solely focusing on those that are raising capital without grant funding.

In 2019, there were also 12 companies in each of the sole private capital states, such that:

$$S_{2019} = \begin{array}{c} \begin{array}{ccccc} (p_0,g_0) & (p_1,g_0) & (p_2,g_0) & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 12 & 12 & 0 & 0 \end{bmatrix} \end{array}.$$

This enables us to project the subset of companies that will not receive grant funding over time.

Startup Profile	State (p_0, g_0)	State (p_1, g_0)	State (p_2, g_0)	Stable	Death
S_{2019}	0	12	12	0	0
S_{2020}	0	11	14	5	1
S_{2021}	0	11	16	9	2
S_{2022}	0	10	17	14	3
S_{2023}	0	10	18	19	4
S_{2024}	0	10	19	24	4
S_{2025}	0	10	20	29	5
S_{2026}	0	10	21	35	6
S_{2027}	0	10	22	40	7
S_{2028}	0	10	22	45	8
S_{2029}	0	10	23	51	9
S_{2030}	0	10	23	56	10

Table 4.20: Forecast of startup ecosystem profiles - Model 3A: No grant funding

We plot this result over time to show the number of startups in each of the capitalisation states.

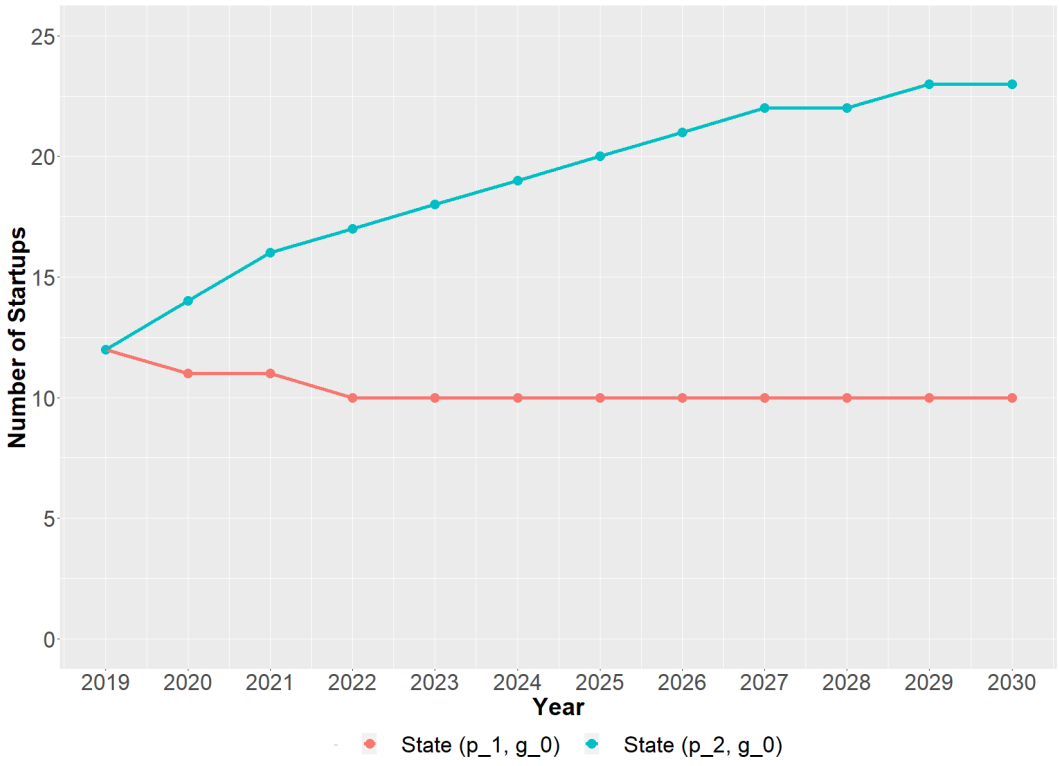

Figure 4.7: Plot of startup ecosystem states - Model 3A: No grant funding

We observe an accumulation in the upper state as the annual probability of exiting

the state is 13.1%, compared to the lower capitalisation state which remains relatively constant at 10.

Whilst this yields some insight, the assumption that companies that receive grant funding won't be appropriate for receiving private capital is not justifiable.

To compare this, we take the Private Capitalisation approach used earlier in Model 3. Recall the transition matrix for this model is:

$$\bar{T}_{\text{private}} = \begin{array}{c} \\ \text{State } p_0 \\ \text{State } p_1 \\ \text{State } p_2 \\ \text{State } p_3 \\ \text{State } p_4 \\ \text{Stable} \\ \text{Death} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{State } p_0 & \text{State } p_1 & \text{State } p_2 & \text{State } p_3 & \text{State } p_4 & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 0.222 & 0.119 & 0.268 & 0.182 & 0.099 & 0.110 \\ 0 & 0.650 & 0.033 & 0.069 & 0 & 0.187 & 0.060 \\ 0 & 0 & 0.677 & 0.136 & 0 & 0.188 & 0 \\ 0 & 0 & 0 & 0.856 & 0.047 & 0.097 & 0 \\ 0 & 0 & 0 & 0 & 0.956 & 0.023 & 0.021 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

In this system we find that

$$S_{2019} = \begin{array}{c} \begin{array}{ccccccc} \text{State } p_0 & \text{State } p_1 & \text{State } p_2 & \text{State } p_3 & \text{State } p_4 & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 0 & 15 & 8 & 16 & 12 & 0 & 0 \end{bmatrix} \end{array}.$$

Over the four years prior, the average number of new startups entering this system is approximately 18. This amount is lower than the total capitalisation model as it doesn't include companies entering the system using a grant funding pathway. As a result, we let:

$$N = \begin{array}{c} \begin{array}{ccccccc} \text{State } p_0 & \text{State } p_1 & \text{State } p_2 & \text{State } p_3 & \text{State } p_4 & \text{Stable} & \text{Death} \end{array} \\ \begin{bmatrix} 18 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

We then produce the result:

Startup Profile	State p_0	State p_1	State p_2	State p_3	State p_4	Stable	Death
S_{2019}	0	15	8	16	12	0	0
S_{2020}	0	14	8	21	16	8	3
S_{2021}	0	13	8	25	19	16	6
S_{2022}	0	12	8	28	23	25	9
S_{2023}	0	12	8	31	26	34	13
S_{2024}	0	12	8	33	30	43	16
S_{2025}	0	12	8	35	33	52	19
S_{2026}	0	12	8	37	37	62	23
S_{2027}	0	12	8	38	40	71	26
S_{2028}	0	11	8	39	43	81	30
S_{2029}	0	11	8	40	47	92	33
S_{2030}	0	11	8	41	50	102	37

Table 4.21: Forecast of startup ecosystem profiles - Model 3B: No grant funding

Plotting this over time shows a divergence in the number of early stage startups in lower private capitalisation states and upper capitalisation states.

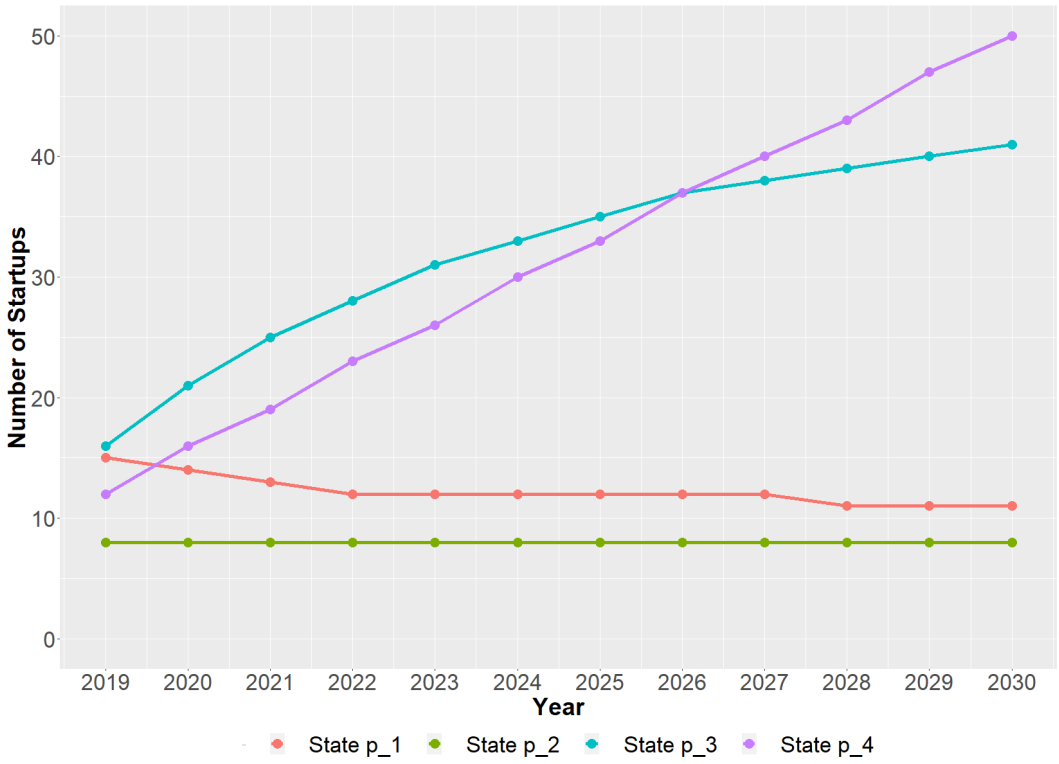

Figure 4.8: Plot of startup ecosystem states - Model 3B: No grant funding

This projection is useful in understanding the private capitalisation states, but for us

to utilise it to analyse a scenario where there is no grant funding, we have to make the assumption that grant funding makes no difference to private capital outcomes. As we've seen in our 2-dimensional model, this isn't the case, so we have two alternative approaches that provide limit cases.

The model is therefore useful in establishing a range, where the two models can provide an upper limit by including all startups, and a lower limit, by focusing only on those that would have not participated in grant funding regardless of whether it was available.

4.7 Discussion and further research

Based on this preliminary analysis, the modelling approach has proven useful in exploring the research questions.

Firstly, grant funding has different effectiveness in stimulating private capital depending on previous capitalisation events.

Secondly, we have demonstrated that the ecosystem can be evaluated empirically based on capitalisation states and transformations.

Lastly, the method can be used to measure the effectiveness of different programs based on the probabilities of transformation between states as well as survival and exit states.

We demonstrated applications for understanding startup pathways using total capitalisation, cumulative grant funding received, and cumulative private capital raised, as well as a 2-dimensional approach linking both grants and private capital.

The approach enables us to make projections around individual startups and the probabilities of different outcomes, as well as evaluating the system as a whole to understand the likely pathways, areas of stagnation, and probability of different exit outcomes.

The relatively small dataset that was available limited the resolution of our capitalisation state spaces, with only 152 companies tracked over an 8 year period. This dataset had questionable completeness in the early years, which potentially resulted in a recency bias, and so conclusions regarding the overall growth or decline of the ecosystem were not possible.

We avoided further data collection in this analysis, as part of our objective was to develop a modelling approach that was useful using the existing data available, but the approach is expected to be enhanced by better data availability. For a more robust analysis, further work into data collection by the Department of Innovation and Skills is strongly recommended.

4.7.1 Insights in the South Australian startup ecosystem

Preliminary insights were obtained in our analysis regarding the South Australian startup ecosystem.

Firstly, the conversion rate of startups from the ecosystem to stable state is very high relative to the number that fail.

As startup and venture capital ecosystems are considered to be a high-risk, high-reward activity, this demonstrates that either South Australia has a conspicuously effective startup ecosystem, or startups have a lower appetite for risk than their interstate and international counterparts. The latter is more plausible, and can be confirmed by the limited companies that have received significant private capitalisation. This also suggests that the data collection process does not sufficiently filter companies that are engaged in high-risk, high-reward activity from companies that could otherwise be classified as small businesses.

Secondly, the most effective grant funding for unlocking private capitalisation is directed towards companies that have already raised some amount of private capital.

Third, startups are taking longer between capitalisation events as they grow in size. This is potentially a natural occurrence as a company scales in size, or the resultant size of funding rounds, or alternatively it suggests that it is easier to raise small amounts of funding in South Australia.

The preliminary insights enable us to make the following recommendations regarding the startup ecosystem and potential policy decisions:

- grant funding should be prioritised to companies that have already raised private capital;
- programs should focus on encouraging startups towards higher-risk, higher-reward domains;
- increase the data collection of capitalisation events of startups over time;
- compare transformation probabilities for different industries to understand which domains are successfully transforming companies.

4.7.2 Evaluation of model usefulness

The relationship of grant funding to private capitalisation was able to be explained and explored and we fulfilled the criteria we have established for usefulness.

The model demonstrated predictive power and performance, is very scalable but still able to operate effectively with small datasets, and by drawing influences from the startup language of capital raising states, is understandable and justifiable.

The model also appeared to be actionable, and developing the model further would potentially assist with this criteria. Whilst the approach has been demonstrated to provide insights into the startup ecosystem, it can also be extended to look at the functioning of subsystems.

By breaking the ecosystem down into different components, transition matrices can potentially be decomposed to pose questions about the different pathways of companies

depending on their attributes. These attributes for instance might be location, industry, participation in programs, or other factors.

This would enable subsystems to have their own transition matrices, which would enable direct comparison between incubators and accelerators, or reveal the underlying contribution of ecosystem participants to the overall picture.

For larger ecosystems, or as more data becomes available over time, the model could potentially be expanded to include additional states. This could include greater resolution for the cumulative grant capitalisation or cumulative private capitalisation states, or additional states such as entry into programs or locations.

The results were presented to a subset of key stakeholders, who responded very positively about the intuitive structure of the transition matrix and potential for scenario planning. Interest was shown in engaging with this thesis after its submission, as well as the potential for further research in the space and potential application in ecosystem analysis.

The approach, whilst relatively simple, is an effective method for analysing and describing the startup ecosystem in South Australia, and is demonstrably useful to those trying to cultivate it.

Chapter 5

Natural language processing in screenplay development

5.1 Introduction

Innovation has significantly transformed the creative industries across the Information Age. Digital technologies have revolutionised almost every part of the film and screen production sectors, including the transformation of celluloid cinematography to digital cameras (Korris & Macedonia (2002)), the rise of visual effects and computer generated imagery (Prince (2011)), and new tools and pipelines for editing and post production (Case (2013)).

Advances in the sector have not been limited to production, as the evolution of the internet has also given audiences direct access to film and series content from around the world through online streaming services (Burroughs (2019)). This has unlocked unprecedented new data sources that enable greater insight into audience demand (Carey (2003)), trends, and viewing habits (Matrix (2014)).

The creative development process, however, is an area of significant interest and debate, with research exploring the role of artificial intelligence in the creative process (Boden (1998)) and many believing that its dominance in the screen industries is inevitable (Datta & Goswami (2021)).

Given the field of natural language processing (NLP) is positioned at the intersection of artificial intelligence, data science and linguistics (Chowdhary (2020)), research into screenplays using NLP has become the primary analytical approach.

A range of script analysis techniques have been introduced, including story plot generation (Gervás et al. (2004)), character personality analysis (Leitch (2013)), power and agency modelling (Sap et al. (2017)), and semantic progression in narrative analysis (Laurino Dos Santos & Berger (2022)).

The author of this thesis has over a decade of experience working with the film and

screen industries, founding a post-production company, producing several feature film and virtual reality projects, as well as a range of other collaborations and governance roles. From consultation with screen practitioners, there is significant skepticism around the potential for technology to support the creative process. Despite this, conference presentations by the author about the promise of natural language processing as a tool in the story development process (vNAB (2017), Screen Forever (2021), Screenmakers (2021) and Southstart (2021)) have generated constructive dialogue with screenwriters, producers and other screen professionals.

The area of most interest from these stakeholders was the application of emotional story arc analysis (Reagan et al. (2016)) to screenplays, and its potential for supporting novel adaptations. This was due to a recognition of the concept of story arcs in traditional story analysis, which is generally defined as a described "line of action or events" that pushes a story plot forward from conflict to resolution (Marks (2015)).

Though feedback was constructive, there was skepticism as to the justifiability and comprehensibility of the approach for generating story arcs from a text corpus, motivating our interest in reviewing the usefulness of the approach and barriers to its application.

5.1.1 Usefulness of story arc analysis

Story arc analysis has been effectively used across large data sets to uncover fundamental arcs of stories (Reagan et al. (2016)) and test the relationships between different arcs and box office success (Del Vecchio et al. (2021)).

As we are interested in its application through the creative development process, we focus on how the analysis can support individual story development rather than big data analysis. To address the stakeholder skepticism, we again turn to our framework of usefulness to evaluate the approach.

The standard approach for generating a story arc utilises a "window" method, where we analyse the sentiment of a fixed-length number of consecutive words that we slide across the text corpus.

To do this we define the following:

- a text corpus of length N words;
- a fixed window size of N_w words;
- n as the number of points in our time series analysis.

The segment length is then calculated by $N_s = \frac{(N-(N_w+1))}{n}$.

The total sentiment over each window segment is calculated and used as a representative of that segment. This process enables a transformation of a string of sentiment values associated with each word in the text corpus, to a time series of generalised sentiment at sampled moments across the story (Reagan et al. (2016)).

The role of window size is important as it determines the number of samples in the time series.

By standardising the number of points in a time series, the arcs of stories that have different lengths can be compared, enabling large numbers of stories to be analysed to evaluate similarities and differences. This raises an important question: Does the arc of a story change if we're sampling it in different ways?

Many of the selected works choose arbitrary window lengths and sampling increments for the time series, which if varied would produce different arcs. The motivation for choosing the right window length does not appear to be rigorously evaluated in existing research and is a significant weakness in the approach's justifiability.

Furthermore, the approach for calculating the sentiment of the window sample does not appear to be congruent with how a human would experience reading a text. As the window length is fixed, the approach can use either the average or cumulative sentiment over the window of words. A human, however, reads one word at a time, not as a block of text, and doesn't recall exactly the last N_s words.

The challenge does not impact performance or actionability, as the arcs reveal objective measures of a story's sentiment arc, which enables screenwriters or producers to take action by addressing sentiment and tone within different sections of the screenplay. The method also appears to be scalable, as the size of the "window" can be altered depending on the size of the text corpus, though as noted, this approach raises challenges, as if the the story arc is intrinsic, it should not change depending on the window size.

This leaves us with an opportunity to evaluate and develop an improved approach for processing a sentiment arc that is more interpretable and justifiable.

5.1.2 Proposed methodology

To do this, we challenge the "window" method and develop an alternative approach that is a more intuitive representation of how human's read.

In our analysis, we apply story arc analysis to episodes from Season 1 of the Series *Aftertaste*, a six-episode series by South Australian based production company Closer Productions. Practical application of the techniques enabled insights to be tested and shared with the "writers room" during development of the Season 2 Series, and enables us to test the scalability to answer further questions, such as: does continuous episode viewing produce different sentiment arcs to episodes watched independently?

5.2 Natural language processing and understanding story arcs

The field of natural language processing bridges human language with computation, and has a variety of applications. In the screen industries, there are two main data sources

that can be utilised for analysis, namely the project's input, the script, or the output, through the text used for subtitles.

For our criteria of usefulness, we are interested in working with a text corpus when there is the ability to impact the screenplay development process, and to enable this actionability, we focus on the script, or screenplay, as a text foundation.

5.2.1 Analysing episode scripts as text corpora

Story arc analysis focuses on interpretation of prose as a text corpus, however, the formatting conventions of screenplays means that pre-processing of the documents is required. This formatting includes different indentation and spacing for action, dialogue and description. Additional information is also included, such as scene headers to tag the location and character names whenever dialogue occur. This is necessary as screenplays are used as foundations that will eventually be transformed into production.

Using Episode 1 of *Aftertaste* as an example, we can demonstrate the structure and formatting of the screenplay. Figure 5.1 below shows an excerpt from the start of the episode.

1 **INT. SHANGHAI/WEST RESTAURANT/ELEVATOR - NIGHT** 1
(EASTON, YOUNG ELEVATOR GUY (50), N/S YOUNG ELEVATOR GIRL)

The quiet hum of an elevator. No muzak. A MAN IN CHEF WHITES is facing the closed doors. There's something about his stillness that cuts a commanding, intimidating presence.

In the back corner of the elevator a YOUNG COUPLE are looking at him and whispering. They recognise him. The guy discreetly holds up his phone to take a photo. The YOUNG WOMAN slides into the shot, smiling, with two fingers up. The GUY takes the pic and the flash goes off. The chef doesn't react. The lift finally comes to a stop. Doors open.

The chef doesn't budge -- making it difficult for the couple to get past him with their luggage.

YOUNG ELEVATOR GUY
...Sorry...excuse...can we--

It seems like the chef is about to move, then -- he hits the CLOSE DOORS BUTTON. The couple look at each other, stunned.

Figure 5.1: Excerpt from screenplay of *Aftertaste* - Season 1, Episode 1

The first refers to the scene description, with the number representing the scene number, "INT." indicating that it is an interior scene, the scene location and the time of day. Following this, the characters in the scene are listen in brackets.

This sort of information is unique to screenplays, and as it does not contribute to the sentiment "arc" of the story, it can be separated in our processing.

This information may be useful for other analyses, such as evaluation of the distribution of information in each location, or the social network of the characters using shared participation in a scene to define adjacency.

Each word can be separated as a data point, or token, shown in the Table 5.1. This is a key aspect of working with text data, and enables us to separate each word and attribute to it the corresponding scene number, screenplay line, and word number. Additionally, if it is dialogue, we can also attribute to it a character.

Data Point	Scene ID	Line ID	Word ID	Character (dialogue only)
The	1	1	1	
quiet	1	1	2	
hum	1	1	3	
of	1	1	4	
an	1	1	5	
elevator	1	1	6	
...	...	...	...	
Sorry	1	12	112	Young Elevator Guy
excuse	1	12	113	Young Elevator Guy
can	1	12	114	Young Elevator Guy
we	1	12	115	Young Elevator Guy
...	...	...	...	

Table 5.1: Table of tokenised data from excerpt of *Aftertaste* - Season 1, Episode 1

The scene headings and character lists are not included within this tokenisation as they are considered to be attributes and not part of the story arc.

5.2.2 Attributing sentiment

The tokenisation of the script now enables us to attribute a sentiment value to each data point. This is a technique frequently used in story arc analysis, and we use the AFINN Sentiment Lexicon to ascribe a corresponding sentiment value to each token.

Developed by Finn Årup Nielsen, the AFINN lexicon is a list of English language terms, each with a sentiment valence between -5 (negative) and +5 (positive) (Nielsen (2011)). For example, the word "pleasure" has a valence of 3, indicating that it is associated with relatively positive sentiment. The word "inadequate" on the other hand has a valence of -2, showing that it has a negative sentiment. Many words, such as "the", "and" and other determiners or connectives do not have a valence associated with them as they do not have meaningful sentiment.

We can update our table to include this valence, and observe that most of the words in our selection do not have a valence associated with them.

Data Point	Scene ID	Line ID	Word ID	Character (dialogue only)	AFINN valence
The	1	1	1		0
quiet	1	1	2		0
hum	1	1	3		0
of	1	1	4		0
an	1	1	5		0
elevator	1	1	6		0
...	...	...	...		
Sorry	1	12	112	Young Elevator Guy	-1
excuse	1	12	113	Young Elevator Guy	-1
can	1	12	114	Young Elevator Guy	0
we	1	12	115	Young Elevator Guy	0
...	...	...	...		

Table 5.2: Token valence using AFINN dictionary from excerpt of *Aftertaste* - Season 1, Episode 1

5.2.3 Generating story arcs using the window method

To generate the story arc from a text corpus, the standard method used is a window method (Reagan et al. (2016)). This approach works by:

- selecting a window size, such as N_w number of words;
- calculating the net valence score over this subset of words by summing the sentiment scores;
- sampling n times across the text corpus by moving the window a fixed proportion through the text.

This creates a time series with a net sentiment valence at each point, which is used to generate the story arc.

As discussed, this approach has created significant interest from writers, producers, and the stakeholder network that was engaged. The ability to compare stories, particularly for adaptations, appears to have significant potential and a range of use cases, but evaluating against our usefulness criteria, the approach appears to be limited.

The justifiability, comprehensibility, and resultant actionability of the method has limitations since whilst the arc is potentially insightful, it is not how humans read. The

approach effectively assumes that the sentiment of a moment in the story is based on an instantaneous aggregation of sentiment across N_w words.

This creates two questions: what is the "right" window size to capture how humans read?; and what is the right distribution of importance given to the valence of words across the window? The window approach is potentially useful for comparative analysis, but if the intent is for the model to be useful in supporting creative intellectual property development, it requires reform to enable it to be interpretable.

To explore this further, we can reframe the approach based on processing the sentiment in a manner that a human might if they were reading the text.

To develop the approach, we can make two observations:

1. Humans read one word at a time.

2. The impact of each word is forgotten over time.

One interpretation of the sliding window approach is that it simulates a reader reading N_w words simultaneously with no memory of previous words. Alternatively, the window approach could be interpreted as a reader reading one word at a time, but with perfect memory for N_w words before completely forgetting the rest of the text.

Neither interpretation appears plausible as a quantification of reader experience, so whilst the window approach has been effective for big data analysis, we now look at an alternative approach that is more useful for representing how a person reads.

5.3 Introducing forgetting curves

Human memory and forgetting is a significant area of research in psychology and neuroscience, and has evolved significantly over the past century. There are a range of reasons why humans forget (Connerton (2008)), but whilst traditional theories are focused on how forgetting assists with navigating everyday life, recent approaches postulate that memory development is a complex process of formation and consolidation, which can be susceptible to interference (Wixted (2004)).

As the objective of our approach is to understand and simulate the human experience of reading, we focus on memory as information retention.

Ebbinghaus Savings Function

One of the first expressions of memory as a mathematical function was produced by Hermann Ebbinghaus (Ebbinghaus (1885)). In an incomplete study, Ebbinghaus produced a "savings" function that modelled forgetting as a function of the strength and fragility of memory.

The model is represented as a "savings" expression b which decays over time t based on parameters c and k, such that:

$$b = \frac{100k}{(log(t))^c + k}.$$

It is widely acknowledged that forgetting and memory formation is impacted by other factors, such as spaced repetition (Hintzman (1976)), though this preliminary approximation is an adequate starting point.

5.3.1 Alternative models

Following this analysis, several, more generalised models have been proposed including a direct exponential function (Woniak et al. (1995)). The exponential approach is represented by:

$$R = e^{-\frac{t}{s}},$$

where R is the retrievability of information, S is the stability of memory, and t is time.

An alternative approach to modelling the distribution of memory retention over time is the Wickelgren Power Law. Derived from the Ebbinghaus savings function, the Law proposes that forgetting is again produced by two factors, time and interference, and can be modelled by the strength and fragility of memory (Wickelgren (1974)(Wixted et al. (2007))) such that:

$$m = \frac{\lambda}{(1 + \beta t)^{\psi}},$$

where $\beta, \psi > 0$ and govern the nature of the forgetting curve.

5.3.2 Application to text corpora

We can utilise the forgetting curve to produce an alternative methodology for calculating the sentiment valence at any point in time. The Wickelgren Power Law is selected as the model for information retention as it enables us to grapple with two clear parameters that govern the speed and extent of memory decay. This provides us with the flexibility to be able to analyse different types of readers and different modes of consumption.

Using the Power Law, we can construct a function that generates a coefficient for the valence of each word based on the location of the reader within the text.

To do this, we let p be the current word being read, and q be the word being analysed.

As the difference between p and q is our unit of time t, we can say that for all $p > q$:

$$m_{p,q} = \frac{\lambda}{(1 + \beta * (p - q))^{\psi}}$$

Applying this to the excerpt from *Aftertaste* and letting $\lambda = 1, \psi = \beta = 0.5$, we present a sample of such values in Table 5.3.

Data Point	Word ID	AFINN Valence	$m_{1,q}$	$m_{2,q}$	$m_{3,q}$	$m_{114,q}$
The	1	0	1.000	0.816	0.707	0.132
quiet	2	0	0	1	0.816	0.132
hum	3	0	0	0	1	0.133
of	4	0	0	0	1	0.134
an	5	0	0	0	0	0.135
elevator	6	0	0	0	0	0.135
...	...	...	...	...	...	...
Sorry	112	-1	0	0	0	0.707
excuse	113	-1	0	0	0	0.816
can	114	0	0	0	0	1
we	115	0	0	0	0	0
...	...	...	...	...	...	...

Table 5.3: Decay parameters of token valence from using $\lambda = 1, \beta = \psi = 0.5$ from excerpt of *Aftertaste* - Season 1, Episode 1

We define the sentiment at a given word as the cumulative sentiment of all preceding words, with each valence decayed using the forgetting curve. Expressed as a function, we can say that the sentiment s_t at time t is given by:

$$\begin{aligned} s_t &= \sum_{i=1}^{t} v_i \times m_{t,i} \\ &= \sum_{i=1}^{t} \frac{v_i}{(1 + \beta \times (t - i))^{\psi}}. \end{aligned}$$

Here we have a method that is comprehensible and justifiable for generating story arcs, where sentiment is accumulated based on historical sentiment, but decayed over time.

5.4 The story arcs of *Aftertaste*

We can now apply this approach to episodes from the first season of the series *Aftertaste*. We explore different variables for β and ψ, compare each episode from the series, and then evaluate the difference between continuous watching of episodes and independent viewing.

This analysis provided the opportunity to present interim results to the writers and producers during the development of the second season, in particular the writers room, to obtain direct feedback on usefulness.

5.4.1 Analysing Episode 1

Analysing Episode 1, we start by processing the episode as described. Filtering out zero valency words, we can provide a summary of the first 10 words, identifying whether they are dialogue or action.

Word	Word ID	valence	type
no	17	-1	action
cuts	35	-1	action
intimidating	38	-2	action
smiling	77	2	action
stop	102	-1	action
difficult	111	-1	action
sorry	122	-1	dialogue
excuse	123	-1	dialogue
like	128	2	action
stunned	148	-2	action
...	...	...	...

Table 5.4: First non-zero valence words from *Aftertaste* - Season 1, Episode 1

Forgetting Scenario Parameters

We can analyse the episode by selecting four cases for the forgetting curve. This enables us to compare the different shapes of the curves based on different selected values for β and ψ. We also include an additional scenario of window decay for comparison, using a step function to replicate the concept of the sliding window. A window length of 250 words is selected for the analysis given the relative decay in other scenarios.

The five scenarios are:

1. Perfect memory where $\beta = 0$ and $\psi = 0$
2. Fast and shallow decay, with $\beta = 2$ and $\psi = 0.1$
3. Consistent decay, where $\beta = 0.005$ and $\psi = 1$
4. Fast and complete decay, where $\beta = 0.5$ and $\psi = 0.5$
5. Window decay, as a step function where the coefficient is 1 for the most recent 250 words, and 0 for the remainder.

The window decay case is slightly different to the standard window approach as the time series typically begins at the first point where a full window length is measured. In

our case, we apply the coefficient from the first word, so for the first 250 words, the window approach is augmented for the available number of words.

We can visualise the five forgetting scenarios below by charting the different coefficients against the number of preceding words.

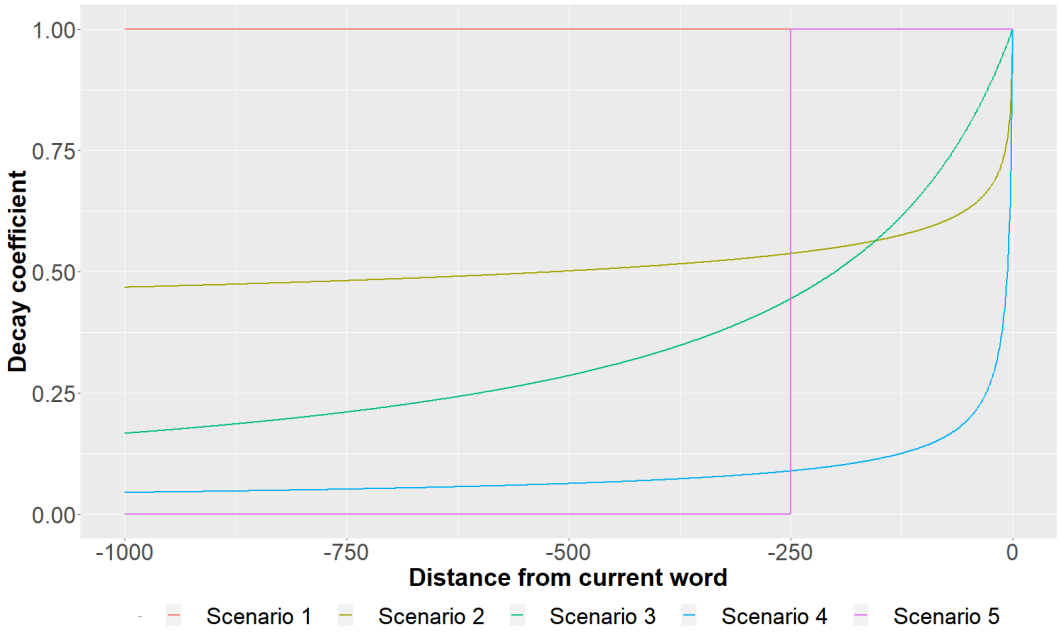

Figure 5.2: Decay profiles across five sets of forgetting curve coefficients

The different coefficients create very different decay rates for the sentiment over time, with the each action or dialogue word being a temporal unit.

The Perfect Memory Scenario, or hyper-attentive viewer, where $\beta = 0$ and $\psi = 0$, is constant at 1, as the zero variables remove any decay. This can be represented by:

$$m_{p,q} = \frac{1}{(1 + 0 \times (p - q))^0} = 1.$$

If either variable for the forgetting curve distribution is zero, the denominator will be equal to 1 and the coefficient will hence be constant at 1. This scenario is effectively the cumulative sentiment arc, as with no decay, the Perfect Memory reader will accumulate all the sentiment valence as they proceed through the story.

The second scenario is described as fast and shallow, as there is an immediate decay which then slows significantly. Figure 5.2 shows this decay at approximately 0.50 over the first 1000 words whereas other scenarios decay faster towards 0. This scenario describes a reader who forgets sentiment relatively quickly, but not completely.

The third scenario describes a more consistent decay, shown by the teal curve in Figure 5.2. Compared to the fast and shallow scenario, the reader forgets at a slower rate initially, but the decay continues at a slower rate of decelaration.

The fourth scenario for comparison has a very fast decay that approaches 0 rapidly. Unlike the second scenario, this scenario describes a reader that forgets more completely and there is very little retention of sentiment valence over time. This could also be considered the experience of a completely distracted viewer.

The last scenario, as defined previously, is the window approach. This is represented as a step function with coefficient at 1 for the length of the window, in this case 250 words, and then 0 for all other words.

Which curve is correct?

Whilst engaging with writers, producers, and other stakeholders, a common question arises: "Which curve is the right one?".

This question assumes that there is an objective underlying curve that's being created from the text corpus. Whether this is a reasonable assumption links back to our original criticism of justifiability using the window approach. Rather than attempt to capture an intrinsic story curve, we are trying to understand the reader experience as they are consuming the text. Some forgetting parameters or alternative functions of emotional valence retention over time might be more appropriate than others, nevertheless, this approach allows us to produce targeted curves for different experience parameters.

As a result, this approach is focused on simulating reader experiences, rather than discovering intrinsic underlying story properties. Based on this, methods might be utilised to attempt to correlate properties across different experiential arcs, but the scope of this investigation is to propose an approach that is more justifiable and hence more useful.

This idea is discussed further later, though is addressed now in the context of comparing *Aftertaste* sentiment arcs.

Forgetting scenarios of Episode 1

Utilising our five different forgetting scenarios, the first episode of *Aftertaste* can be analysed:

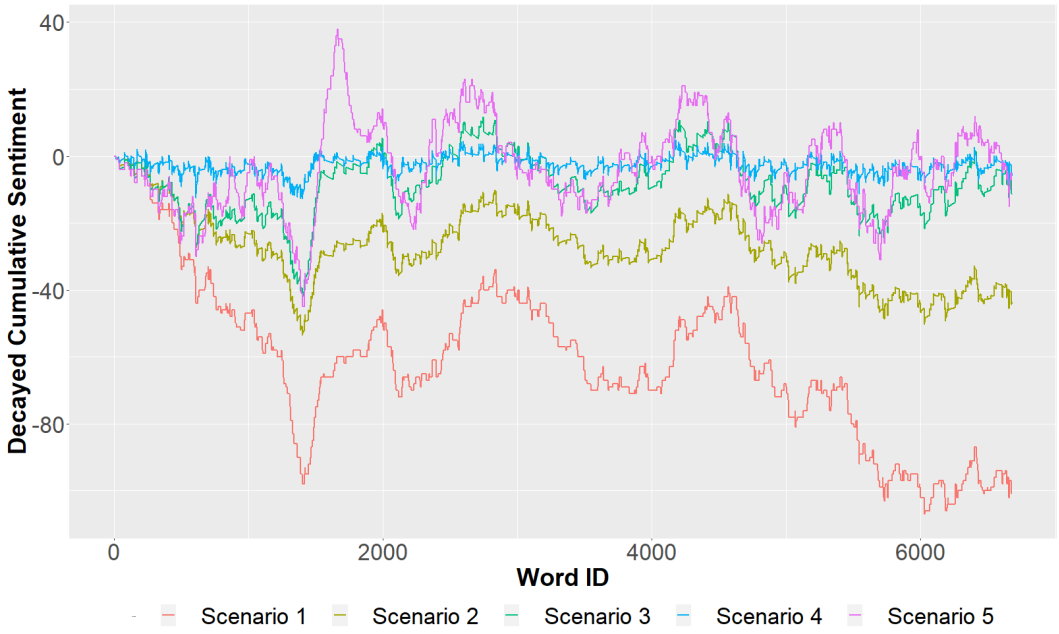

Figure 5.3: Decayed cumulative sentiment arcs for *Aftertaste* - Season 1, Episode 1

The five different arcs show the different experiences that viewers with different memory characteristics might experience.

Scenario 1, the red line, represents the cumulative sentiment with no memory decay over time. Scenario 4, which describes the fast and deep forgetting scenario, more closely describes the instantaneous sentiment of a scene or interaction in the script. Both scenarios represent different insights.

Local peaks and troughs can be identified at different points throughout the story, indicating features to be coherently identified between the different scenarios.

The arcs for scenarios 2 and 3 exhibit similar structure, whereas scenario 3 appears to move around the 0 axis, whilst scenario 2 appears displaced over time. This is due to the completeness of the forgetting curve, as incomplete decay results in a component of the sentiment accumulating over time.

The curve for scenario 4, the fast and complete decay scenario, has limited structure for analysis, and appears relatively noisy. This is intuitive as the scenario describes an inattentive viewer whose engagement with information is fleeting.

Scenario 5, the window approach scenario, shows the peaks and troughs clearly at the same moments as the other scenarios, though they appear to be clearer and more pronounced. This shows the dampening effect that the cumulative memory approach can have, as whilst Scenario 5 is focused on the most recent 250 words, the other scenarios factor in all of the previous words with some degree of decay.

This is evident in the first major peak at approximately the 1,700th word. The window

approach is focused only on a 250 word window, which is very positive leading to a cumulative sentiment of nearly 40, whilst the other approaches factor in all of the negative sentiment to that point and are all close to or less than 0.

5.4.2 Episode and scenario comparisons

We can now compare the decay scenarios across all six episodes of the first season of *Aftertaste*. The word count of each episode script varies significantly, which is potentially due to different content requirements in each episode or the writing styles of each episode's author. A comparison table below shows the different writers for each episode, as well as the word count.

Episode	**Writer(s)**	**Word Count**
Episode 1	Julie de Fina	6,679
Episode 2	Matt Vesely	5,484
Episode 3	Matthew Bate & Jodie Molloy	5,279
Episode 4	Matthew Bate	6,679
Episode 5	Julie de Fina & Mathew Bate	5,590
Episode 6	Julie de Fina	6,122

Table 5.5: Credited writers for each episode of *Aftertaste* Season 1

For each of our five scenarios, we plot the generated Decayed Cumulative Sentiment arc using the parameters previously described.

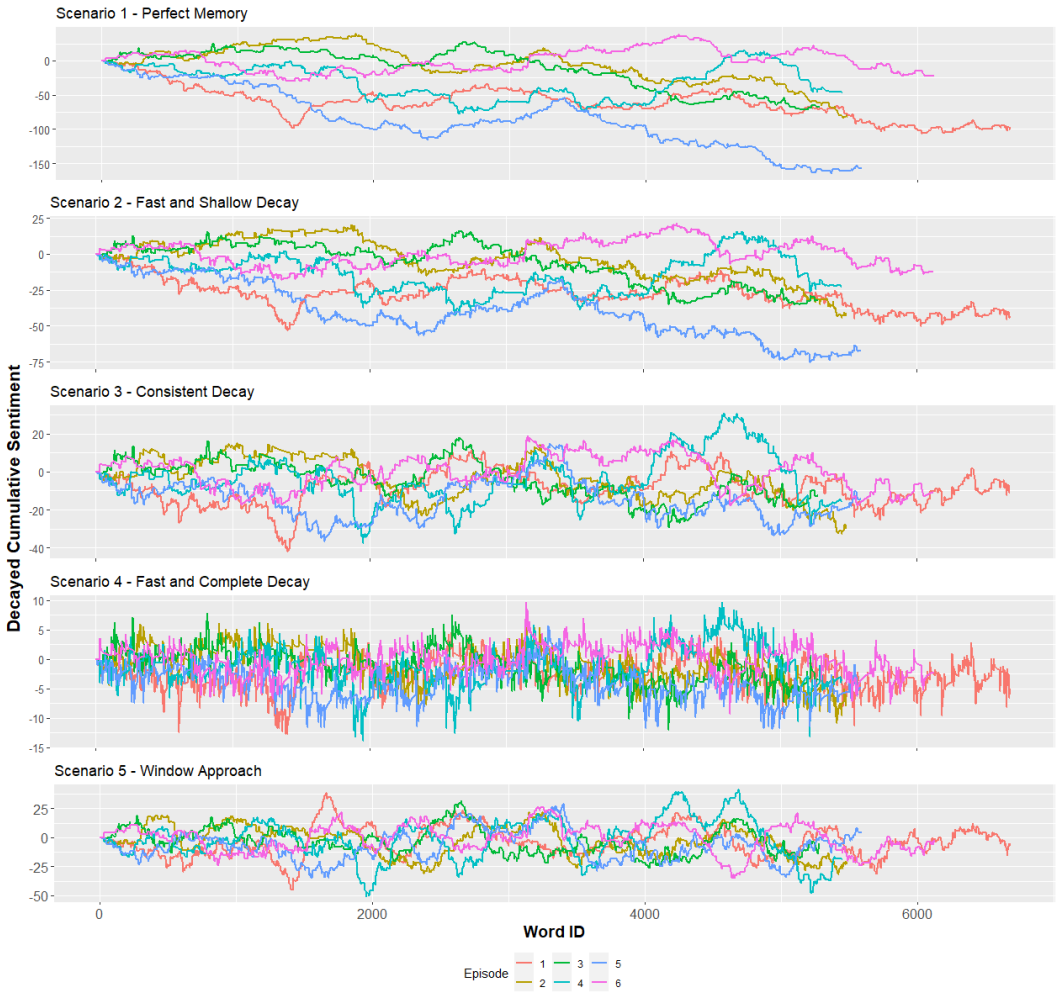

Figure 5.4: Decayed cumulative sentiment arcs across forgetting scenarios of *Aftertaste* - Season 1

The five scenarios show very different types of behaviour.

Scenario 1, the Perfect Memory arc, represents the case where the viewer accumulates all sentiment as they read or consume the media over time. The arcs for each of the six areas can be clearly compared, showing local maxima and minima at different stages.

Scenario 2, which factors in Fast and Shallow Decay, has similar features to each arc visualised using Scenario 1, with local maxima and minima easily distinguishable at the same points in time. They are incredibly similar, on different axis scales, though Scenario 2 has increased variability.

The third scenario, Consistent Decay, exhibits different relative features from one episode to the next. There is a significant peak in episode 4, which is more pronounced than in other scenarios. As noted in this scenario and Scenario 4, Fast and Complete

Decay, the sentiment arc signal starts to appear noisier. This is to be expected as a viewer with limited memory would be responding to immediate events, to the point where a viewer with no memory will observe each word independently as a data point.

As identified in Figure 5.3, the amplitude of the Window Approach is closest to the Consistent Decay scenario.

Rather than attempt to select the ultimate arc methodology, we can observe that each scenario describes a different viewer, or mode, of consuming the content.

Each of these approaches, such as the hyper-attentive viewer in Scenario 1 or the completely distracted viewer in Scenario 4, enables us to compare the story using different lenses.

5.4.3 Reading an episode in context

This new methodology for visualising the emotional experience arc of a reader can also be used to test the proposition that binge-watching, where a viewer watches multiple episodes consecutively, creates a very different experience for the viewer than if they consumed the episodes were consumed independently.

To do this, we compare three methodologies for sentiment arc calculation:

- Episode based - resetting the sentiment at the end of each episode as if watched independently;
- Season based - continuing the cumulative decayed sentiment across consecutive episodes as if watched continuously;
- Window-based Sentiment Arc approach for comparison.

We undertake this using two sets of forgetting parameters that were introduced earlier, Scenario 1 which demonstrated perfect memory and no decay of sentiment, and Scenario 2, fast and shallow decay with $\beta = 2$ and $\psi = 0.1$. The two scenarios are interesting for comparison given the perfect memory scenario is an objective measure of each arc, whereas the second scenario enables our concept of sentiment forgetting to be evaluated.

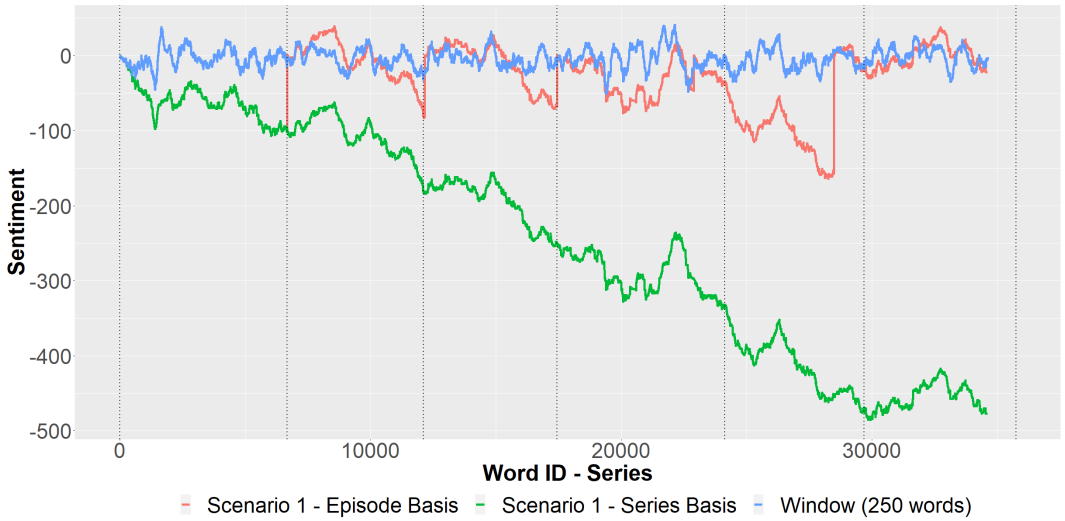

Figure 5.5: Perfect memory sentiment arcs comparing independent consumption, binge-watching and window method: *Aftertaste* Season 1

The story arc for each episode when viewed on an independent episode basis resets to zero at the end of each episode, whereas the cumulative sentiment arc using a series basis to represent continuous binge-watching steadily accumulates. The first episode is understandably identical between the two, as the same methodology is being used across the first episode.

Local minima and maxima can be identified between the three different arcs, but the curves appear to be very different in nature.

The series based curve is of significant interest for stakeholders, as the planning of episodes in the writers room requires episodes to be assessed in the context of its place within the season and relative to other episodes.

We can now add forgetting or memory decay to the methodology for curve generation, using parameters from Scenario 2 previously analysed, where $\beta = 2$ and $\psi = 0.1$. This approach adds an additional level of complexity as the decay function is not scaled as it is applied and by analysing more episodes, we are able to observe the behaviour of the decay curve being applied over a greater data set.

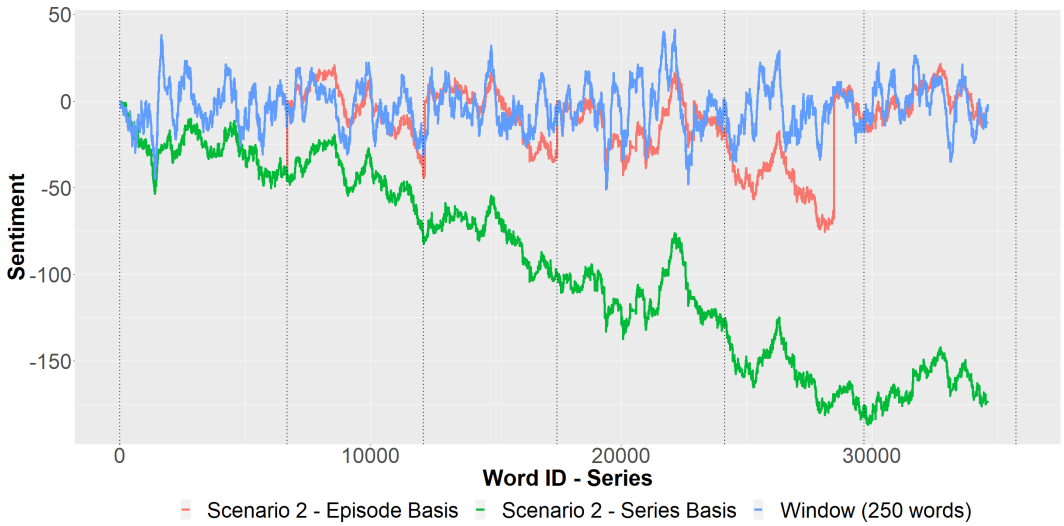

Figure 5.6: Fast and shallow decayed cumulative sentiment arcs comparing independent consumption, binge-watching and window method: *Aftertaste* Season 1

The second comparison shows a similar contrast between the episode based and the series based curves, though with a smaller range on the y-axis due to the introduction of decay.

If the decay is increased, so that there is faster decay to a complete forgetting, the arc will tend towards a more instantaneous measure at any given point in time. This fits with our alternative interpretations of the forgetting curve that describe different types of consumption, such as a viewer with limited attention span or "low stakes" consumption.

This comparison validates that our proposed approach has the flexibility to describe different modes of consuming the media, which has received positive feedback from our engagement and increases the justifiability of using sentiment curves for story analysis.

5.5 Discussion and further research

The proposed new approach for story arc analysis fulfills our objective of finding a methodology with an increased justifiability, and hence usefulness, for application in screenplay development

Testing the approach with this author's stakeholder network yielded a consistently positive response, and the approach of using a model of memory and information retention to decay sentiment valence was considered to be reasonable and intuitive.

Whilst the approach has yielded insights into the first season of the series *Aftertaste*, there are several areas that warrant further discussion, including the overall usefulness of the approach, reflection on the comparison with the "window" method, and the viability of the forgetting curve.

Discussion on the motivation of data science in this domain is also undertaken, given the unique nature of creative industries and the role of data science in creative development as an area for investigation and further research.

5.5.1 Usefulness of the experience arc approach

A model's usefulness was defined earlier as a set of five characteristics of interest: performance, scalability, comprehensibility, justifiability, and actionability.

Our analysis was motivated by a perceived need from stakeholders to increase the justifiability of the methodology for constructing story arcs from a corpus of text using natural language processing. In addition to increasing the justifiability, our proposed new approach utilising memory and valence decay has also impacted several of the other characteristics.

Firstly, the scalability of the model is significantly improved using our methodology. The window approach, requires an arbitrary window to construct the model which creates a series of challenges. For large data sets, this may require a larger window for the curve to appear meaningful, and similarly if the data set is too small, the approach would result in insufficient data being available.

For instance, if a scene that had approximately 250 words was to be analysed and the window size was selected at 250 words, this would only produce a single data point. Alternatively, our memory-based approach treats each word as a data point of reader experience, it is still able to produce a coherent cumulative sentiment arc by analysing the corpus one word at a time with memory decay based on chosen parameters. This means that the methodology can be applied on a very small set of words or a large collection of texts with similar effectiveness.

In addition, the scalability characteristic can be evaluated through our comparison of binge-watching and individual episode viewing. The window approach did not meaningfully capture the difference between the two modes of consumption, however, the memory-based curve naturally described the difference in experience between binge-watching and individual episode consumption.

This result also demonstrates that another characteristic, the performance of the model, is also significantly improved, as it describes the binge-watching experience more appropriately. To assess the performance of the model requires a qualitative link between the screenwriter's interpretation of what they have created, compared to the emotional arc plotted from their text. As observed in our analysis, many of the local maxima and minima can be observed across the different modelling approaches, but the memory-decayed model appears to capture greater trajectories that represent the cumulative experience.

Comprehensibility can be said to be improved based on the ability of the stakeholders to understand the method for story arc generation.

The actionability of the model does not appear to be significantly improved, as whilst the story arc generated is noticeably different, it is a similar output for implementation. It

is recommended that interested screenwriters engaged for this analysis could continue to participate or stay adjacent to further research to understand the potential actionability of this approach in future writers rooms or creative development processes.

Overall, the proposed model has a significantly increased usefulness based on the analysis provided.

5.5.2 Comparison with the window approach

Given the conclusion about the usefulness of the memory-based arc construction approach, it is necessary to review the comparison with the window approach.

To reiterate the concept, the memory-based arc treats sentiment valence as having a cumulative impact on the reader or viewer, which decays over time based on memory capacity or attention. This decay rate is modelled using a forgetting curve distribution, subject to variables that describe the speed and extent of the decay. Using this description, we can rethink the window approach, as it appears to represent perfect memory for a given amount of words that form the width of the window, followed by no memory of anything prior to that point.

Whilst this is an apt comparison, this suggestion is in some ways unfair to the users of this method, as proponents of the window approach do not claim to represent the user experience. It instead measures something more akin to an aggregated sentiment impulse at a given point in a story.

This differentiation is further exhibited by considering the parameters for each method. The memory-based method requires two parameters to describe the speed and extent of a reader or viewer's memory decay, which is useful for describing different types of audiences and also different modes of consuming the media. The window approach is a single parameter, the size of the window, which in turn represents the size of the impulse.

This demonstrates a qualitative difference between the models. The model that incorporates memory as a representation of information decay allows the story arc to be generated based on memory decay parameters. This reveals that there isn't one objective story arc, but rather a range of experience arcs based on the memory or attentiveness of the viewer.

In comparison, the window approach produces an impulse at a moment in time in the text corpus that isn't subject to interpretation or a function of human experience. This is markedly different, as it infers an objective measure of the underlying story.

Rather than judge the benefits or weaknesses of each, it's appropriate to acknowledge that they are measuring different things. The window approach has shown to be remarkably efficient at comparing different story arcs across large sets text corpora, which whilst not as useful to our screenwriting stakeholders, is a valid and exciting research frontier in the domain as a big data problem.

As a result, we take this opportunity to differentiate the two approaches as three different story arcs for analysing screenplays:

- Sentiment Impulse Arc - the traditional method using the sliding window approach;
- Experience Arc - the decayed cumulative sentiment arc;
- Cumulative Sentiment Arc - cumulative sentiment arc over time with no decay.

We include the third classification as a type of Experience Arc, as the Cumulative Sentiment Arc is a fixed measure where parameter selection is not required. This enables objective comparison, and is the measure of a viewer with perfect memory and engagement.

All three of the curves communicate different information about the underlying text, and are useful for different applications.

5.5.3 Alternative dictionaries

For our analysis, we have used the AFINN Dictionary to attribute sentiment valence to each word in the text corpus.

Alternative dictionaries may generate different arcs with similar or differing structure. Drawing on *SentiBench*, the benchmark comparison of "state-of-the-practice sentiment analysis methods", alternative dictionaries can be proposed using 3-class (positive, negative and neutral) dictionaries (Ribeiro et al. (2016)). 2-class dictionaries (positive and negative) are not recommended at this stage given they require removal of neutral messages prior to analysis, which given our need to capture neutral moments in our time series, is not appropriate for our analysis.

For 3-class dictionaries, only the VADER and LIWC15 dictionaries are ranked higher than AFINN in the *SentiBench* rankings, and are hence recommended for further analysis.

Despite these rankings, dictionary performance is situation dependent (Reagan et al. (2015)), and development of tailored dictionaries for screenplay analysis that differentiates between dialogue text and action text is also appropriate.

Further research may include utilisation of higher-dimensional dictionaries, such as the *nrc* dictionary that categorises words into not just positive or negative sentiments, but also eight emotion categories such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

The scalability of the proposed Experience Arc approach makes this multi-dimensional analysis practical, as each emotion category will not require a threshold of words to construct a meaningful arc.

5.5.4 Forgetting distribution

The Experience Arc generation method relies on an assumption of reader or viewer memory decay over time, shown by a forgetting curve coefficient. Whilst this is a fair assumption for the purposes of information retention, the previous name of the arc as an *emotional* story arc, means that our approach is not necessarily completely appropriate.

The forgetting curve approach assumes a continual and consistent decay of word and its sentiment valence over time, however, we can hypothesise that emotional moments may be more memorable than others.

To test this, direct research could be undertaken into understanding memory and information retention of readers of screenplays, and to understand whether the forgetting curve approach is an appropriate model of reader experience.

5.5.5 Further work

Whilst our focus has been on screenplays and the development of useful models for the writers room, the development of the Experience Arc approach presents an exciting opportunity for further research and development. The paper *The emotion arcs of stories are dominated by six basic shapes* (Reagan et al. (2016)) strongly influenced this exploration, though the work relied on the assumption that a story has an intrinsic underlying arc that can be generated systematically for comparative analysis.

In our approach, we have begun to look at the additional interface of human subjectivity and the role that different modes of consumption play. This leads to a range of questions, such as "Is there a perfect way to read a screenplay", or "What is similar / different in how we interpret a text". These questions are central to the field of natural language processing, as the field continues to grapple with the role of human interpretation (Chowdhary (2020)).

Chapter 6

Conclusion

6.1 Summary of results

The purpose of the thesis is to explore the potential for increasing the usefulness of modelling approaches, with a focus on "small data" regimes where data science and human action intersect. As defined earlier, the concept of usefulness challenges the value of many modelling approaches that are only interested in predictive performance, and requires models to be evaluated based on additional characteristics such as comprehensibility, justifiability, actionability, and scalability. Through this analysis, we have found significant demand in our domains of interest for this research, particularly due to a skepticism from stakeholders of the ability for data science to "solve" complex situations, rather than support human decision makers as they navigate situations with limited data and a large range of variables with substantial uncertainty.

This approach has led to contributions to the literature in each domain. In Chapter 3, we explored the foundations of statistics in the Australian Football League. Previous research and the majority of data science applications focused on counting possession events, however by engaging with coaching and football departments at multiple clubs, it was found that this missed two critical features:

- The majority of the game is played "off-the-ball", and is not captured by possession events;
- It matters who you play against.

Combining these two observations with the acknowledgement that the complexity of the game means that traditional models have low statistical power, our first principles analysis led to a new and promising approach. By evaluating the functioning of subsystems, and integrating pairwise performance metrics, we were able to produce a foundational approach that has the potential to rethink the application of data science in the Australian Football League.

Chapter 4 focused on another unique field of human action, entrepreneurship. We explored research into start-up transformation pathways, modelling start-up transformation, and growth of companies through private capitalisation and grant funding routes. Through engagement with government stakeholders and policy makers, similar frustrations were found to those in Chapter 3, where models appear to have low statistical power and usefulness, and analysis was focused on measuring inputs rather than outputs.

Using our usefulness rubric, an approach was formulated to disentangle startup trajectories based on capitalisation and in turn create a method for understanding the impacts of startup interventions and grant support on private capitalisation outcomes. This is a significant contribution to the literature associated with entrepreneurship and startup transformation as it enables a practical methodology for understanding the trajectory of startups that is useful for ecosystems with limited data, and gauging the effectiveness of ecosystem interventions.

Lastly, we evaluated the use of data science in creative development in Chapter 5, specifically the use of natural language processing to produce story arcs from screenplays. Previous approaches have yielded significant insights, comparing and clustering story arcs as a big data problem, however the usefulness in the "writers room" during the screenwriting stages appeared to be limited. Through engagement with screenwriters and producers, and participation in a writers room for the series *Aftertaste*, an alternative approach for generating story arcs was developed that increased justifiability, scalability, performance and comprehensibility.

This new approach significantly contributes to the field of natural language processing and its intersection with creative development, introducing the role of memory in story analysis through techniques that model information retention. This enables us to review how computers process information to better emulate the human process of reading or consuming content. Rather than replace the existing window-based approach, the proposed methodology creates a new way of generating story arcs that include acknowledgement of human variation in memory or attentiveness.

The work in each domain received strong positive engagement from the participating stakeholder networks, and the appetite for further engagement increased as results and discussions progressed.

6.2 Governance and developing useful metrics

Through this analysis, progress was made in each of these domains by improving the usefulness of modelling approaches. Engagement with a broad range of stakeholders revealed consistent themes, such as an eagerness to engage with new tools and the data landscape, but a frustration with data analysis projects that were uninterpretable or lacked pathways for meaningful action to be taken. The focus on usefulness provided an avenue of promise as it acknowledged several key factors:

- Not every problem has sufficient data to be treated as a big-data problem;
- There are limits to automation, particularly in domains of human action, and a focus on the human interface requires comprehensibility, actionability and justifiability;
- Black-box approaches may be powerful, but the role of risk and accountability in action require greater insight.

These insights reveal a significant new realm of opportunity. From a corporate governance perspective, there is tension between an appetite and drive to engage with metrics, data science, and artificial intelligence, and the risk of poor implementation and false promise. The relatively recent focus on reflexive quantification (Muller (2019)), captured by the adage "If You Can't Measure It, You Can't Manage It" has resulted in a rapid expansion of data capture and analysis projects (Berenson (2016)). Instead, quantifying human performance has been shown to often produce inverse results in education, medical care, businesses and government (Muller (2019)), and leading to greater interest in alternative pathways and more rigour in analysing the strategic implementation of data science projects.

The approach taken in this research provides a framework that is potentially more acceptable to corporate governance expectations, focusing on stakeholder need, strategic objectives, available data and useful modelling.

6.3 Future research

Artificial intelligence continues to play an increasingly large role in all parts of society, and its prevalence in the future is the subject of much debate.

The question of what humans will be doing in this future is not a new one. John Maynard Keynes postulated that productivity gains would mean that his grandchildren would only need to work 15 hours a week (Keynes (1931)), but rather than productivity replacing human effort, the nature of work has instead evolved, sometimes with questionable value (Graeber (2013)).

The rise of big data has also created a false assumption that every problem can be solved with more data and computation, leading to an over-metrification across a range of domains (Muller (2019)).

This heuristic, when taken to its limit, appears as a modern variant of Laplace's Demon (Laplace (1951))(Van Strien (2014)). According to Laplace, if an entity, or demon, knows the exact location and momentum of every atom in the universe, their past and future values for any given time can be calculated using classical mechanics (Marquis de Laplace (1902)). There have been a range of refutations of this over time (Sommer (2013)) (Ulanowicz (2012)), but even though modern big data ambitions cannot be considered interchangeable with computing the future states of the universe, recent arguments are

focused on the compounding complexity when there are multiple entities attempting to compute big data systems (Wolpert (2008)) (Rukavicka (2014)). This reminds us of our focus on human domains of action, where agents are entities that are acting on limited input data in complex environments, rather than mechanical, deterministic participants.

This leaves us with an intersection of human domains of action that have insufficient data sets for big data analysis, which from our broad analysis, can be seen to be found in diverse areas of life and work. The usefulness characteristics provide us with a new set of tools for approaching these challenging domains, showing promise not just in our domains of interest, but many other domains that would benefit from this angle of approach.

Usefulness characteristics might be a new framework in modern data science (Coussement & Benoit (2021)), however, we are again reminded of the quote from Korzybski that this has long been a challenge for modelling.

> A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its *usefulness*. (Korzybski (1933))

Digital computation may have increased exponentially during this time, but the insight of this statement endures as we navigate an increasingly technological future.

Bibliography

Agarwal, A., Balasubramanian, S., Zheng, J. & Dash, S. (2014), Parsing screenplays for extracting social networks from movies, *in* 'Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)', pp. 50–58.

Agarwal, A., Zheng, J., Kamath, S., Balasubramanian, S. & Dey, S. A. (2015), Key female characters in film have more to talk about besides men: Automating the bechdel test, *in* 'Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', pp. 830–840.

Aho, A. V. & Ullman, J. D. (1992), *Foundations of computer science*, Computer Science Press, Inc.

Aldous, D. (2017), 'Elo ratings and the sports model: A neglected topic in applied probability?', *Statistical Science* **32**, 616–629.

Amezcua, A. S. (2010), *Boon or Boondoggle? Business incubation as entrepreneurship policy*, Syracuse University.

Andreacchio, A., Bean, N. & Mitchell, L. (2022), 'Modelling australian rules football as spatial systems with pairwise comparisons', *Journal of Quantitative Analysis in Sports* **18**(4), 215–226.

Atkinson, G. & Nevill, A. M. (2001), 'Selected issues in the design and analysis of sport performance research', *Journal of Sports Sciences* **19**(10), 811–827.

Averell, L. & Heathcote, A. (2011), 'The form of the forgetting curve and the fate of memories', *Journal of Mathematical Psychology* **55**(1), 25–35.

Azhari, H., Widyaningsih, Y. & Lestari, D. (2018), Predicting final result of football match using Poisson regression model, *in* 'Journal of Physics: Conference Series', Vol. 1108, IOP Publishing, p. 012066.

Barocas, S. & Boyd, D. (2017), 'Engaging the ethics of data science in practice', *Communications of the ACM* **60**(11), 23–25.

Bellman, R. (1966), 'Dynamic programming', *Science* **153**(3731), 34–37.

Berenson, R. A. (2016), 'If you can't measure performance, can you improve it?', *Journal of the American Medical Association* **315**(7), 645–646.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S. & Matthews, I. (2014), Large-scale analysis of soccer matches using spatiotemporal tracking data, *in* '2014 IEEE International Conference on Data Mining', IEEE, pp. 725–730.

Boden, M. A. (1998), 'Creativity and artificial intelligence', *Artificial intelligence* **103**(1-2), 347–356.

Braham, C. & Small, M. (2018), 'Complex networks untangle competitive advantage in Australian football', *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28**(5), 053105.

Brewer, C., Dawson, B., Heasman, J., Stewart, G. & Cormack, S. (2010), 'Movement pattern comparisons in elite (AFL) and sub-elite (WAFL) Australian football games using GPS', *Journal of Science and Medicine in Sport / Sports Medicine Australia* **13**, 618–23.

Brown, W. H., Malveau, R. C., McCormick, H. W. S. & Mowbray, T. J. (1998), *AntiPatterns: refactoring software, architectures, and projects in crisis*, John Wiley & Sons, Inc.

Burroughs, B. (2019), 'House of netflix: Streaming media and digital lore', *Popular Communication* **17**(1), 1–17.

Carey, J. (2003), Audience demand for TV over the Internet, *in* 'Internet television', Routledge, pp. 223–240.

Caruana, R., Lundberg, S., Ribeiro, M. T., Nori, H. & Jenkins, S. (2020), Intelligible and explainable machine learning: best practices and practical challenges, *in* 'Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', pp. 3511–3512.

Case, D. (2013), *Film technology in post production*, Taylor & Francis.

Castells, M. (1996), 'The information age: Economy, society and culture (3 volumes)', *Blackwell, Oxford* **1997**, 1998.

Cattelan, M., Varin, C. & Firth, D. (2013), 'Dynamic bradley–terry modelling of sports tournaments', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**(1), 135–150.

Cho, S., Lee, S.-M. & Park, B.-J. (2014), 'Lean startup: The way to reduce the failure rate of startups', *Asia-Pacific Journal of Business Venturing and Entrepreneurship* **9**(4), 41–53.

Chowdhary, K. (2020), 'Natural language processing', *Fundamentals of artificial intelligence* pp. 603–649.

Clarke, S. R. (2005), 'Home advantage in the Australian Football League', *Journal of Sports Sciences* **23**(4), 375–385.

Cohen, S. & Hochberg, Y. V. (2014), 'Accelerating startups: The seed accelerator phenomenon', *ERPN: Governance & Organization (Sub-Topic)* .

Connerton, P. (2008), 'Seven types of forgetting', *Memory studies* **1**(1), 59–71.

Coussement, K. & Benoit, D. F. (2021), 'Interpretable data science for decision making', *Decision Support Systems* **150**, 113664.

Dam, H. K., Tran, T. & Ghose, A. (2018), Explainable software analytics, *in* 'Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results', pp. 53–56.

Datta, A. & Goswami, R. (2021), The film industry leaps into artificial intelligence: Scope and challenges by the filmmakers, *in* 'Rising Threats in Expert Applications and Solutions', Springer, pp. 665–670.

Dawson, B., Hopkinson, R., Appleby, B., Stewart, G. & Roberts, C. (2004), 'Player movement patterns and game activities in the Australian Football League', *Journal of Science and Medicine in Sport* **7**(3), 278–291.

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J. & Hajkowicz, S. (2019), 'Artificial intelligence: Australia's ethics framework. Data 61 CSIRO, australia'.

Del Vecchio, M., Kharlamov, A., Parry, G. & Pogrebna, G. (2021), 'Improving productivity in Hollywood with data science: Using emotional arcs of movies to drive product and service innovation in entertainment industries', *Journal of the Operational Research Society* **72**(5), 1110–1137.

Ebbinghaus, H. (1885), *Memory: A Contribution To Experimental Psychology*, Teachers College, Columbia University, New York City.

Eldhose, K., Jose, C., Siddharth, S., Geejo, S. S. & Sreedevi, S. (2021), Alyce: An artificial intelligence fine-tuned screenplay writer, *in* 'Innovative Data Communication Technologies and Application', Springer, pp. 627–636.

Elo, A. E. (1978), *The rating of chessplayers, past and present*, Arco Pub.

Fabrizio, K., Mowery, D. C., Lamoreaux, N. R. & Sokoloff, K. L. (2007), 'Financing innovation in the United States, 1870 to the present', *The Federal Role in Financing Major Innovations: Information Technology During the Postwar Period* .

Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern Recognition Letters* **27**(8), 861–874.

Feldman, R. (2013), 'Techniques and applications for sentiment analysis', *Communications of the ACM* **56**(4), 82–89.

Field, S. (1982), *Screenplay*, Delacorte New York.

Fix, E. & Hodges, J. L. (1989), 'Discriminatory analysis. nonparametric discrimination: Consistency properties', *International Statistical Review/Revue Internationale de Statistique* **57**(3), 238–247.

Flath, C. M. & Stein, N. (2018), 'Towards a data science toolbox for industrial analytics applications', *Computers in Industry* **94**, 16–25.

Forbes, D. (2006), Dynamic prediction of Australian Rules Football using real time performance statistics, PhD thesis.

Gadzinski, G. & Castello, A. (2022), 'Combining white box models, black box machines and human interventions for interpretable decision strategies', *Judgment and Decision Making* **17**(3), 598.

Gervás, P., Díaz-Agudo, B., Peinado, F. & Hervás, R. (2004), Story plot generation based on CBR, *in* 'International Conference on Innovative Techniques and Applications of Artificial Intelligence', Springer, pp. 33–46.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. & Kagal, L. (2018), Explaining explanations: An overview of interpretability of machine learning, *in* '2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)', IEEE, pp. 80–89.

Glickman, M. E. (1995), 'The glicko system', *Boston University* **16**, 16–17.

Graeber, D. (2013), 'On the phenomenon of bullshit jobs: A work rant', *Strike Magazine* **3**, 1–5.

Greenham, G., Hewitt, A. & Norton, K. (2017), 'A pilot study to measure game style within Australian football', *International Journal of Performance Analysis in Sport* **17**(4), 576–585.

Gupta, D., Malviya, A. & Singh, S. (2012), 'Performance analysis of classification tree learning algorithms', *International Journal of Computer Applications* **55**(6).

Hagendorff, T. (2020), 'The ethics of AI ethics: An evaluation of guidelines', *Minds and Machines* **30**(1), 99–120.

Hallen, B. L., Bingham, C. B. & Cohen, S. (2014), Do accelerators accelerate? A study of venture accelerators as a path to success?, *in* 'Academy of management proceedings', Vol. 2014, Academy of Management Briarcliff Manor, NY 10510, p. 12955.

Hilb, M. (2020), 'Toward artificial governance? the role of artificial intelligence in shaping the future of corporate governance', *Journal of Management and Governance* **24**(4), 851–870.

Hintzman, D. L. (1976), 'Repetition and memory', *Psychology of learning and motivation* **10**, 47–91.

Hipson, W. E. & Mohammad, S. M. (2021), 'Emotion dynamics in movie dialogues', *arXiv preprint arXiv:2103.01345* .

Hirotsu, N., Inoue, K., Yamamoto, K. & Yoshimura, M. (2022), 'Soccer as a markov process: modelling and estimation of the zonal variation of team strengths', *IMA Journal of Management Mathematics* .

Huth, M. & Ryan, M. (2004), *Logic in Computer Science: Modelling and reasoning about systems*, Cambridge University Press.

Islam, M., Fremeth, A. & Marcus, A. (2018), 'Signaling by early stage startups: US government research grants and venture capital funding', *Journal of Business Venturing* **33**(1), 35–51.

Islam, M. R. & Zibran, M. F. (2017), A comparison of dictionary building methods for sentiment analysis in software engineering text, *in* '2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)', IEEE, pp. 478–479.

Jobin, A., Ienca, M. & Vayena, E. (2019), 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence* **1**(9), 389–399.

Kelleher, J. D. & Tierney, B. (2018), *Data science*, MIT Press.

Keynes, J. M. (1931), Economic possibilities for our grandchildren, *in* 'Essays in persuasion', Springer, pp. 321–332.

Korris, J. & Macedonia, M. (2002), 'The end of celluloid: Digital cinema emerges', *Computer* **35**(4), 96–98.

Korzybski, A. (1933), *Science and sanity: An introduction to non-Aristotelian systems and general semantics*, Institute of GS.

Krishna, A., Agrawal, A. & Choudhary, A. (2016), Predicting the outcome of startups: less failure, more success, *in* '2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)', IEEE, pp. 798–805.

Laplace, P.-S. (1951), 'A philosophical essay on probabilities, 1819', *English translation, Dover* **6**.

Laurino Dos Santos, H. & Berger, J. (2022), 'The speed of stories: Semantic progression and narrative success.', *Journal of Experimental Psychology: General* .

Leitch, T. (2013), 'You talk like a character in a book: dialogue and film adaptation', *Film Dialogue* pp. 85–100.

Leushuis, C. (2018), 'Beating the odds - a state space model for predicting match results in the Australian Football League', *Quantitative Methods in Business and Economics* **2**.

Lewis, M. (2004), *Moneyball: The art of winning an unfair game*, WW Norton & Company.

Loh, W.-Y. (2011), 'Classification and regression trees', *Wiley interdisciplinary reviews: data mining and knowledge discovery* **1**(1), 14–23.

London, A. J. (2019), 'Artificial intelligence and black-box medical decisions: accuracy versus explainability', *Hastings Center Report* **49**(1), 15–21.

Markov, A. A. (1971), 'Extension of the limit theorems of probability theory to a sum of variables connected in a chain', *Dynamic Probabilistic Systems* **1**, 552–577.

Marks, D. (2015), *Inside story: The power of the transformational arc*, Bloomsbury Publishing.

Marquis de Laplace, P. S. (1902), *A philosophical essay on probabilities*, Wiley.

Maslow, A. H. (1966), *The Psychology of Science*, New York: Harper & Row.

Matrix, S. (2014), 'The Netflix effect: Teens, binge watching, and on-demand digital media trends', *Jeunesse: Young People, Texts, Cultures* **6**(1), 119–138.

Mattar, Y. (2008), 'Post-industrialism and Silicon Valley as models of industrial governance in Australian public policy', *Telematics and Informatics* **25**(4), 246–261.

McIntosh, S., Kovalchik, S. & Robertson, S. (2018), 'Validation of the Australian Football League player ratings', *International Journal of Sports Science & Coaching* **13**(6), 1064–1071.

Muller, J. Z. (2019), The tyranny of metrics, *in* 'The Tyranny of Metrics', Princeton University Press.

Nahavandi, S. (2019), 'Industry 5.0—a human-centric solution', *Sustainability* **11**(16), 4371.

Nielsen, F. Å. (2011), 'Afinn'.
URL: *http://www2.compute.dtu.dk/pubdb/pubs/6010-full.html*

OSACE (2019), Future industries exchange for entrepreneurship: Entrepreneurship and startup strategy, Technical report, Department for Innovation and Skills, Government of South Australia.

O'Shaughnessy, D. M. (2006), 'Possession versus position: strategic evaluation in AFL', *Journal of Sports Science & Medicine* **5**(4), 533.

Papalampidi, P., Keller, F. & Lapata, M. (2019), 'Movie plot analysis via turning point identification', *arXiv preprint arXiv:1908.10328* .

Prabhu, P., Kim, H., Oh, T., Jablin, T. B., Johnson, N. P., Zoufaly, M., Raman, A., Liu, F., Walker, D., Zhang, Y. et al. (2011), A survey of the practice of computational science, *in* 'SC'11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis', IEEE, pp. 1–12.

Prasetio, D. et al. (2016), Predicting football match results with logistic regression, *in* '2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)', IEEE, pp. 1–5.

Prince, S. (2011), *Digital visual effects in cinema: The seduction of reality*, Rutgers University Press.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M. & Dodds, P. S. (2016), 'The emotional arcs of stories are dominated by six basic shapes', *EPJ Data Science* **5**(1), 1–12.

Reagan, A. J., Tivnan, B., Williams, J. R., Danforth, C. M. & Dodds, P. S. (2015), 'Benchmarking sentiment analysis methods for large-scale texts: a case for using continuum-scored words and word shift graphs', *arXiv preprint arXiv:1512.00531* .

Regulation, P. (2018), 'General data protection regulation', *Intouch* **25**.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A. & Benevenuto, F. (2016), 'Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods', *EPJ Data Science* **5**(1), 1–29.

Rosli, C. M. F. C. M., Saringat, M. Z., Razali, N. & Mustapha, A. (2018), A comparative study of data mining techniques on football match prediction, *in* 'Journal of Physics: Conference Series', Vol. 1020, IOP Publishing, p. 012003.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.

Rukavicka, J. (2014), 'Rejection of Laplace's demon', *The American Mathematical Monthly* **121**(6), 498–498.

Ryall, R. (2011), Predicting Outcomes in Australian Rules Football, PhD thesis, RMIT University.

Ryall, R. & Bedford, A. (2010), 'An optimized ratings-based model for forecasting Australian Rules Football', *International Journal of Forecasting* **26**(3), 511–517.

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H. & Choi, Y. (2017), Connotation frames of agency and power in modern films, *in* 'Conference on Empirical Methods in Natural Language Processing'.

Selisker, S. (2015), 'The Bechdel test and the social form of character networks', *New Literary History* **46**(3), 505–523.

Shafizadeh, M., Sproule, J. & Gray, S. (2013), 'The emergence of coordinative structures during offensive movement for goal-scoring in soccer', *International Journal of Performance Analysis in Sport* **13**(3), 612–623.

Sommer, C. (2013), 'Another survey of foundational attitudes towards quantum mechanics', *arXiv preprint arXiv:1303.2719* .

Spencer, B., Morgan, S., Zeleznikow, J. & Robertson, S. (2016), Clustering team profiles in the Australian Football League using performance indicators, *in* 'Proceedings of the 13th Australasian Conference on Mathematics and Computers in Sport, Melbourne', pp. 11–13.

Stefani, R. & Clarke, S. (1992), 'Predictions and home advantage for Australian rules football', *Journal of Applied Statistics* **19**(2), 251–261.

Suthaharan, S. (2016), Decision tree learning, *in* 'Machine Learning Models and Algorithms for Big Data Classification', Springer, pp. 237–269.

Techboard (2021), Techboard annual funding report fy2021, Technical report, Acceleration venture Catalysts Pty Ltd.

Theis, T. N. & Wong, H.-S. P. (2017), 'The end of Moore's law: A new beginning for information technology', *Computing in Science & Engineering* **19**(2), 41–50.

Therneau, T., Atkinson, B. & Ripley, B. (2019), *rpart: Recursive Partitioning and Regression Trees*. R package version 2.15.0.
URL: *https://CRAN.R-project.org/package=rpart*

Tukey, J. W. (1962), 'The future of data analysis', *Annals of Mathematical Statistics* **33**, 1–67.

Ulanowicz, R. E. (2012), *Growth and development: ecosystems phenomenology*, Springer Science & Business Media.

Van Strien, M. (2014), 'On the origins and foundations of Laplacian determinism', *Studies in History and Philosophy of Science Part A* **45**, 24–31.

Vanacker, T., Manigart, S., Meuleman, M., Sels, L. et al. (2010), The impact of bootstrap strategies on new venture development: A longitudinal study, Technical report, Ghent University, Faculty of Economics and Business Administration.

Weiblen, T. & Chesbrough, H. W. (2015), 'Engaging with startups to enhance corporate innovation', *California Management Review* **57**(2), 66–90.

Wexler, R. (2017), 'When a computer program keeps you in jail', *The New York Times* **13**.

Wickelgren, W. A. (1974), 'Single-trace fragility theory of memory dynamics', *Memory & Cognition* **2**(4), 775–780.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686.

Williams, G. (2011), *Data mining with Rattle and R: The art of excavating data for knowledge discovery*, Springer Science & Business Media.

Wilson, E. (2009), 'Debate at the Harvard Museum of Natural History', *Harvard Museum of Natural History: Cambridge, MA, USA* .

Wixted, J. T. (2004), 'The psychology and neuroscience of forgetting', *Annual Review of Psychology* **55**, 235–269.

Wixted, J. T., Carpenter, S. K. et al. (2007), 'The Wickelgren power law and the Ebbinghaus savings function', *Psychological Science* **18**(2), 133.

Wolpert, D. H. (2008), 'Physical limits of inference', *Physica D: Nonlinear Phenomena* **237**(9), 1257–1281.

Woniak, P. A., Gorzelaczyk, E. J. & Murakowski, J. A. (1995), 'Two components of long-term memory.', *Acta neurobiologiae experimentalis* **55**(4), 301–305.

Woods, C. T., Robertson, S. & Collier, N. F. (2017), 'Evolution of game-play in the Australian Football League from 2001 to 2015', *Journal of Sports Sciences* **35**(19), 1879–1887.

Woods, T. C. (2016), 'The use of team performance indicator characteristics to explain ladder position at the conclusion of the Australian Football League home and away season', *International Journal of Performance Analysis in Sport* **16**(3), 837–847.

Zhang, Y., Tiňo, P., Leonardis, A. & Tang, K. (2021), 'A survey on neural network interpretability', *IEEE Transactions on Emerging Topics in Computational Intelligence* **5**(5), 726–742.

Zhu, C., Byrd, R. H., Lu, P. & Nocedal, J. (1997), 'Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization', *ACM Transactions on Mathematical Software* **23**(4), 550–560.
URL: *https://doi.org/10.1145/279232.279236*

Zou, J. & Schiebinger, L. (2018), 'AI can be sexist and racist — it's time to make it fair', *Nature* **559**(7714), 324–326.