6-1-2022

Scenario Acceleration Through Automated Modelling: A Method and System for Creating Traceable Quantitative Future Scenarios Based on FCM System Modeling and Natural Language Processing

Christopher W.H. Davis
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Part of the Computer Sciences Commons

Let us know how access to this document benefits you.

Recommended Citation

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Scenario Acceleration Through Automated Modelling: A Method and System for Creating Traceable Quantitative Future Scenarios Based on FCM System Modeling and Natural Language Processing

by

Christopher W.H. Davis

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Technology Management

Dissertation Committee:
Antonie M. Jetter, Chair
Ameeta Agrawal
Charles Weber
Philippe Giabbanelli

Portland State University
2022
Abstract

Scenario planning is used extensively in strategic planning because it helps leaders broaden their perspectives and make better decisions by presenting possible futures in story form. Some of the benefits of using scenarios include breaking away from groupthink, creating better products, acceleration of organization learning and reducing bias. Product development teams, particularly for digital products, are gaining more autonomy in organizations and tend to manage risk by undergoing very short development iterations on their products while leaning on their consumers for feedback – a process known as agile development. This method tends to limit the perspective of the team and foster groupthink, two side effects which could potentially be addressed using scenarios. However, the time-consuming and expensive processes used to create scenarios are inaccessible to agile product development teams, and even teams that use scenarios for strategic direction typically use them at the beginning of product development and do not keep them up to date over time, eventually making them irrelevant to decision making. This research explores automating the bottlenecks of the scenario process so they can be incorporated into autonomous agile teams by creating and rigorously tests an artifact that combines Natural Language Processing (NLP) to understand data, Interactive Machine Learning (IML) to combine automation with human expertise, Fuzzy Cognitive Maps (FCM) for quantitative scenario modeling, and Horizon Scanning (HS) to keep models up to date; a system I call Scenario Acceleration through Automated Modelling (SAAM). Using Design Science Research (DSR), I
demonstrate how these technologies can be used together to speed up the scenario creation process while keeping people in the loop, and how they can be kept up to date over time. This research lays the foundation for product development teams to use scenarios in agile processes, with the goal of creating better products and avoiding disruption.

This work makes several contributions: Firstly, it furthers the body of knowledge on scenario development by showing how to create scenarios with automation and how scenarios could be used by agile teams. Secondly, it demonstrates a novel method of creating FCM with NLP and human collaboration, and how to use Horizon Scanning to keep models up to date over time. Finally, I leave an artifact that can be used by other teams who want to continue this vein of research, or for product teams that want to utilize this method.
Dedication

To my wife, Heather – thank you for encouraging me to go down this path and standing by me throughout this process. This publication would not exist if it were not for you!
Acknowledgements

I first must express my eternal gratitude towards my parents, who always emphasized the importance of education and worked so hard to make sure I had the space to learn.

I am extremely grateful and appreciative of my advisor and committee chair, Dr. Antonie Jetter, for her mentorship and guidance over the last several years. Dr. Jetter has been an amazing teacher and has always helped me find how to lean on my strengths and experience in my research.

I would like to express my deep appreciation to my dissertation committee members, Dr. Philippe Giabbanelli, Dr. Charles Weber, and Dr. Ameeta Agrawal for their guidance, feedback, and support through this work. In particular Dr. Giabbanelli for your mentorship and guidance in the field of simulation and helping me become a better academic writer.
Table of Contents

Abstract .......................................................................................................................... i
Dedication ................................................................................................................... iii
Acknowledgements ................................................................................................... iv
List of Tables ............................................................................................................ viii
List of Figures .......................................................................................................... ix
List of Abbreviations .............................................................................................. xi

1 Introduction ........................................................................................................... 1
  1.1 Design Science Framework: Approach to proposed work ......................... 8

2 The Potential for Automating FCM Based Scenarios ........................................ 12
  2.1 Introduction to Scenario Techniques ...................................................... 14
  2.2 Scenarios for Product Development ..................................................... 18
      2.2.1 Benefits of Scenarios in Product Development ......................... 20
      2.2.2 Scenarios in Agile Projects: Potentials and Problems ............... 24
  2.3 Requirements for Scenario Techniques in Support of Agile Teams ........ 27
  2.4 Foundations of a future solution: FCM Scenarios and Interactive Machine Learning .......................................................... 31
      2.4.1 Fuzzy Cognitive Maps Explained ............................................. 34
      2.4.2 Interactive Machine Learning Explained .................................. 42

3 Systematic Review of the State-of-the-Art ......................................................... 47
  3.1 Systematic Review of the use of FCM-based Scenarios in Product Development .......................................................... 48
  3.2 Systematic Review on the Use of Text Mining for the Creation of Scenarios .......................................................... 50
  3.3 Concluding Discussion ............................................................................. 60

4 Research Gaps, Objective, and Questions ......................................................... 63

5 SAAM: Scenario Acceleration through Automated Modelling – A Conceptual System ......................................................................................... 66
  5.1 Motivation ............................................................................................... 66
  5.2 Foundations: Scenario Planning and IML ............................................. 73
      5.2.1 Scenario Planning ...................................................................... 73
5.2.2 Interactive Machine Learning (IML) ................................................... 75
5.2.3 Big Data in Futures Studies ............................................................ 82
5.2.4 Modeling Scenarios with Fuzzy Cognitive Maps ......................... 84
5.2.5 Building Fuzzy Cognitive Maps with Big Data and NLP ............... 87
5.2.6 Recent Advances in NLP with BERT and Question and Answering
Algorithms ............................................................................................. 88
5.2.7 Horizon Scanning Background ....................................................... 92
5.3 An Overview of the Conceptual System .......................................... 94
5.3.1 IML Flow - Integrating People and AI ........................................... 96
5.3.2 Phase 1 – Setting Up; Collecting Data and Defining Questions ...... 96
5.3.3 Phase 2 – Building the Model ....................................................... 100
5.3.4 Phase 3 – Using the Model ............................................................ 109
5.3.5 Phase 3 – Using HS to Inform Model Inputs .................................. 114
5.4 Concluding Summary ..................................................................... 117
6 Evaluation .......................................................................................... 119
6.1 Overview of the Evaluation Approach ............................................. 120
6.2 Experiment 1 – SAAM vs Humans .................................................. 125
  6.2.1 Method Overview ...................................................................... 126
  6.2.2 Model Building with SAAM ....................................................... 128
  6.2.3 Static Analysis ........................................................................... 132
  6.2.4 Dynamic Analysis ..................................................................... 138
  6.2.5 Discussion and Limitations ....................................................... 141
6.3 Experiment 2 – SAAM vs LSA/FARM ............................................. 142
  6.3.1 Method Overview ...................................................................... 143
  6.3.2 Data Collection ......................................................................... 145
  6.3.3 Formulating Questions ............................................................... 146
  6.3.4 Comparing Outputs ................................................................... 148
  6.3.5 Discussion and Limitations ....................................................... 155
6.4 Experiment 3: integrating Horizon Scanning ................................... 157
  6.4.1 Horizon Scanning Data ............................................................... 157
List of Tables

Table 1 - DSR Guidelines adapted from (A. R. Hevner et al. 2004) ........................................ 9
Table 2 - Benefits of scenario planning for product development from the literature ... 21
Table 3 - Scenario Planning limitations from the literature ............................................. 23
Table 4 - Requirements for integrating scenarios into product development teams with number labels ................................................................................................................... 30
Table 5 - Example Concepts .............................................................................................. 37
Table 6 - Matrix representation of example FCM ............................................................. 39
Table 7 - Initial literature review search results (1) .......................................................... 48
Table 8 - Initial literature review search results (2) .......................................................... 52
Table 9 - Data sources used by studies that automate parts of the scenario creation process ................................................................. 54
Table 10 - Futures Studies using text analysis and big data ............................................. 83
Table 11 - Example output of NLP model given the question, "what causes pollution to increase?" ......................................................................................................................... 91
Table 12 - Questions used in the illustrative use case for SAAM ................................... 100
Table 13 - Answer filters modelers can set used by SAAM ............................................. 103
Table 14 - Full results from scenarios ............................................................................. 113
Table 15 - DSR evaluation methods, adapted from (A. R. Hevner et al. 2004) .............. 121
Table 16 - First set of questions used in validation experiment 1 .................................. 129
Table 17 - Example results from experiment 1, first set of questions............................ 129
Table 18 – The answers that were filtered after the first round of questions in experiment 1 ................................................................................................................... 130
Table 19 – How headlines were rephrased to questions from the corpus ................. 131
Table 20 – Follow up questions used in experiment 1 .................................................. 131
Table 21 – Network metrics across all models in validation experiment 1 ................. 133
Table 22 - Consistency results from validation experiment 1 ....................................... 139
Table 23 - Full dynamic comparison results from SAAM ............................................. 140
Table 24 – Questions used in Validation Experiment 2 based on PESTEL for Electric Vehicles .................................................................................................................... 146
Table 25 - Concepts of original study mapped to identified concepts from SAAM ...... 149
Table 26 - Final results of all scenarios in experiment 2 ................................................. 155
Table 27 - Scenario validation criteria found in the literature ....................................... 172
Table 28 - Likert scale ratings ......................................................................................... 175
Table 29 - Ranking question ............................................................................................ 176
Table 30 - Post-observation interview questions for expert team............................... 176
Table 31 - Agile team exit interview questions .............................................................. 180
List of Figures

Figure 1 - Example Scenario Cross, adapted from (Duin et al. 2015) ................................. 2
Figure 2 - Dissertation Outline .......................................................................................... 11
Figure 3 - General steps and participants in quantitative, model-based scenario techniques......................................................................................................................... 17
Figure 4 – Where a product team sits in the larger environment, adapted from (van Notten 2006) ........................................................................................................................................... 18
Figure 5 - People involved in the scenario planning process ........................................... 20
Figure 6 - Example Agile development cycle .................................................................... 25
Figure 7 – How Agile processes lose sight of scenario planning over time ..................... 27
Figure 8 - Static vs dynamic scenarios for product development teams ......................... 29
Figure 9 - Directed graph example ................................................................................... 32
Figure 10 - Levels of scenarios, adapted from (van Notten 2006), modified from Figure 4 ........................................................................................................................................... 34
Figure 11 - An example FCM ............................................................................................. 38
Figure 12 – Results from running FCM simulation in our example charted until reaching an end state ........................................................................................................................................... 40
Figure 13 - Example FCM simulation results from a second scenario using the example model ........................................................................................................................................... 41
Figure 14 - Literature review stage results (1) .................................................................. 49
Figure 15 - How automation can help scenario generation ............................................. 52
Figure 16 - Literature review stages (2) ............................................................................ 53
Figure 17 - Visualized methodologies of studies that use TM for scenarios .................... 56
Figure 18 – Gaps in the literature, research objective, and research questions addressed in this dissertation ........................................................................................................................................... 63
Figure 19 - Aspects of the scenario process that can benefit from automation, Figure 3 revisited ........................................................................................................................................... 70
Figure 20 - Each step in SAAM shows where people are in the loop ............................... 72
Figure 21 - Example IML interface, using Microsoft’s LUIS to create a language model . 77
Figure 22 - Example word cloud output from TM ............................................................ 79
Figure 23 – IML System blueprint of SAAM ...................................................................... 81
Figure 24 - An Example FCM ............................................................................................. 85
Figure 25 - Example BERT base encoding where Tokens represent the words, Segments represent the sentences, and Position represents the position of a word within a chunk of text........................................................................................................................................... 89
Figure 26 - Signals found in HS categorized in quadrants, adapted from (Gokhberg et al. 2020) ........................................................................................................................................... 93
List of Abbreviations

AI: Artificial Intelligence
BERT: Bidirectional Encoder Representations from Transformers
DoD: Degree of Diffusion
DoV: Degree of Variability
DSR: Design Science Research
FARM: Fuzzy Association Rule Mining
IML: Interactive Machine Learning
LDA: Latent Dirichlet Allocation
LSA: Latent Semantic Analysis
LUIS: Language Understanding Intelligence Service
ML: Machine Learning
PESTEL: Political, Economic, Social, Technological, Environmental, and Legal
POS: Part of Speech
Q&A: Question and Answer
SAAM: Scenario Acceleration through Automated Modelling
SEC: Securities and Exchange Commission
SPA: Scenario Planning Assistant
TAR: Title and Abstract Review
TM: Text Mining
WoS: Web of Science
1 Introduction

The speed of technology development is accelerating while markets and the competitive landscape can shift quickly (Weber and Tarba 2014). At a strategic level companies attempt to prepare for these changes through scenario planning, which results in narratives that help people see and think through potential futures (Schoemaker 1993; Goodwin and Wright 2001; Bodwell and Chermack 2010), so that they can develop strategies that are robust under multiple future conditions or can be adapted quickly. In addition, scenarios help decision-makers identify aspects of the business environments that they should regularly monitor to stay on top of changes.

The use of scenarios can best be illustrated with an example, such as the example below, which is adapted from Duin et al. (2015). In this study, scenarios were built to better understand the potential for labor in ports, given rapid changes in technology and a total of four alternative futures were identified.

Often when scenarios are completed, they are communicated on a “Scenario Cross”, which uses the two most impactful concepts in the scenario as axes on a graph,
then maps the scenarios to the four quadrants of the graph.

**Figure 1 - Example Scenario Cross, adapted from (Duin et al. 2015)**

Scenarios can also be turned into narratives, where the concepts on the cross are told in story form such as the four scenarios below relating to the cross above.

- **Scenario 1- People first:** People and demography develop positively meaning that no problems with shortage of skills are faced. Enough and well-skilled staff is available for port operation. Potential problems of the demographic change can be overcome, e.g., by recruitment of foreign staff or by extensive training of non-specialist workers. Processes and the application of technology in the port of the future remain unchanged compared to today. Container handling is still
semi-automated and the handling of offshore components still contains much handiwork and is executed project by project (Duin et al. 2015).

- Scenario 2 - Full force ahead: People and demography develop positively meaning that no problems with shortage of skills are faced. Enough and well-skilled staff is available for port operation. Potential problems of the demographic change can be overcome by reducing the needed amount of skilled staff. Processes and the application of technology (e.g. automation) play the key role in overcoming demographic change resulting in efficient re-engineered processes controlled and supported by powerful integrated information systems (ICT) (Duin et al. 2015).

- Scenario 3 - Stagnation: People and demography develop negatively. The average age in workforce is very high resulting in big problems after the next retirement wave. Ports do face the problem of shortage of skilled workforce. Processes and the application of technology in the port of the future remain unchanged compared to today. Container handling is still semi-automated and the handling of offshore components still contains much handiwork and is executed project by project. Effective port operation suffers from non-availability of enough and skilled staff. This scenario represents the worst case for the future of ports (Duin et al. 2015).
- **Scenario 4- Smart Port:** People and demography develop negatively. The average age in workforce is very high resulting which may results in big problems after the next retirement wave. Ports do face the problem of shortage of skilled workforce. This threat is answered by re-engineering of processes and innovation in technology application. Processes are efficient because they have been digitized supported by powerful integrated information systems (ICT) (Duin et al. 2015).

As shown in the example above once scenarios are created they are often communicated in a story format, which makes them more memorable and assists in maintaining a broader future vision during planning and decision making (Bowman et al. 2013). Scenarios are used to enhance the perception of the surrounding business environment, provide a way to deal with uncertainty, act as a communication tool, and assist in organizational learning (Wright, Bradfield, and Cairns 2013). Scenario planning has been shown to have several cognitive benefits including reducing bias, helping frame and reframe issues, and improving imagination (Meissner and Wulf 2013; Ramírez, Österman, and Grönquist 2013; Amer, Daim, and Jetter 2013). However, to date, scenario planning is largely limited to long-term strategic plans or product development projects of strategic importance, such as multi-year platform development. Alternative future scenarios are thus not available to all product development teams – instead, they develop products based on the product and business requirements that were determined by others (e.g., marketing) at the onset of
the project and usually for a single possible future, that was determined in the planning phase, where scenarios may or may not have been used. As a result, product development teams do not have access to the cognitive benefits of thinking about the future as a range of possibilities, which is afforded by scenario planning.

Increasingly, technology products are developed in flat hierarchical structures and with agile methodologies that give development teams growing autonomy (Cao et al. 2009). Rather than defining products once, in alignment with overall strategies, and at the onset of development projects, product features may thus be determined iteratively, based on customer feedback (Cao et al. 2009; Papadopoulos 2015). This decoupling of strategic product planning and development carries risk: when product development teams across an organization dominantly focus on current customer feedback, rather than broadening their horizons to look for neglected areas that consumers may care about in the future, resulting products become more likely to be disrupted by newcomers or other innovators which could end in catastrophic failure for a company (Christensen 1997). There consequently is a need to ensure ongoing strategic perspectives in product development decisions. Kazmim, Naaranoja, and Kytola (2015) have proposed to achieve this by connecting all of the teams responsible for product development and putting them through strategic thinking exercises and Derbyshire and Giovannetti (2017) and Randt (2015) propose scenario planning to mitigate uncertainty; However, such practices are still not common.
One reason for the limited use of scenario planning among product teams is the lack of efficient scenario planning techniques. Existing methods are usually comprised of long workshops with expert practitioners that require the participation of leaders from across an organization (Schoemaker 2020; Rafael Ramirez et al. 2017; Hodgkinson et al. 2006; Franco, Meadows, and Armstrong 2013; O’Brien 2004). The time-consuming process and the reliance on scarce experts make scenario planning so cumbersome that it is usually not feasible to apply the technique to the specifics of product development projects unless they have long development timelines and are of primary strategic importance.

Moreover, scenarios that come out of the strategic planning process are not granular enough to guide a product development team’s decision-making, so some decisions cannot be supported such as what features to prioritize, what type of emerging technology to consider adopting, what types of partners would fit well in a platform ecosystem, or what social shifts will affect target segments. All these future changes require decisions on the tactical level of single product projects, where they are costly to have to revisit or revise multiple times. To date, existing techniques for scenario planning result in scenarios that are (1) not granular and specific enough to support tactical decisions and (2) take too long to create to use in short and iterative development cycles (Leau et al. 2012; Sahoo and Pattnaik 2020; Gabriel et al. 2021).

Recent advances in Artificial Intelligence (AI) and its subfield of Machine Learning (ML) (Mitchell 1997), however, make information on market, social, political, technology
trends, and competitor actions that are available on the web much more accessible and hold the potential to automate scenario development steps that are currently entirely manual and expert-based. For example, there is research that uses conversational AI in the role of a facilitator to allow teams to build causal mental models through conversations with a virtual assistant (Reddy, Giabbanelli, and Mago 2019; Reddy, Srivastava, and Mago 2020; Anjum et al. 2021) which could help accelerate the scenario process. In particular, Text Mining (TM) and Natural Language Processing (NLP) can help to convert unstructured text from various sources including journals, news blogs, etc. into structured representations of the driving forces that shape future scenarios and the interdependencies between them as a computational model, based on Fuzzy Cognitive Maps (FCM). The FCM model serves as a scenario engine that computationally creates scenarios (C. W. H. Davis, Jetter, and Giabbanelli 2020), and that represents the expert understanding of driving forces, as articulated in reports, journal papers, news feeds, blogs, etc. This makes the process of creating scenarios less time-consuming and less challenging and would allow product teams to use scenarios as an “everyday” tool for making sense of the uncertainty they face. In addition, it is likely possible to continuously update the FCM models and run new simulations, as new data becomes available. With this document, I explore this potential and develop and test a method and software system to combine NLP, FCM modeling and simulation, and scenario planning to create future scenarios from data in mostly automated ways. For brevity, I will call the proposed new approach Scenario Acceleration through Automated Modelling (SAAM).
SAAM constitutes design science research because it creates an artifact (i.e., a novel combination of IT-based methods) and evaluates its performance. To ensure that SAAM is rigorously designed and evaluated and delivers generalizable insights, I have employed the Design Science Framework (DSR) by A. R. Hevner et al. (2004) and organized my thesis work accordingly. I organize this thesis, wherever possible, in the form of self-contained chapters so that they can be easily converted into journal publications. This is the reason for the overall structure of this work, which is presented in the following section.

1.1 Design Science Framework: Approach to proposed work

The Design Science Framework, used in Design Science Research (DSR), describes the process used by researchers to refine a theory through the creation, testing, and study of a novel implementation of information technology artifacts. DSR is a popular research method in the management of information systems that have evolved through the contributions of several studies over the past few decades (Gregor and Hevner 2013; A. Hevner 2007; Peffers et al. 2007; A. R. Hevner et al. 2004; Kuechler, Vaishnavi, and Kuechler 2007). DSR is well suited for research in the computer science field because it involves creating something tangible (the artifact) and studying how it solves a practical problem in a specific context. The artifact can be anything practical that can be used in the real world, which included a wide range of things including smartwatches, (Zenker and Hobert 2019), weather monitoring systems (Bertrand, Brusset, and Chabot 2021), medical education systems (Mdletshe, Oliveira, and Twala...
2021), and software (Stol and Fitzgerald 2020). DSR ensures that these artifacts are not only evaluated rigorously but that the research results in insights that are relevant and generalizable beyond the single artifact.

Referring to design in general, not DSR, Walls, Widmeyer, and El Sawy (1992) described design as a process and a product. The process focuses on an artifact that is refined and evaluated until the best possible product for the specific setting under study is created. DSR goes beyond this focus on one artifact in one setting. This is concisely summed up by Romme when he writes that design involves creating a new kind of knowledge, and “validation serves to codify, fine-tune, generalize, justify and (possibly) falsify it” (Romme and Reymen 2018). The guiding principles of DSR are outlined by Hevner et al. (2004) and shown in Table 1. To implement these principles, A. R. Hevner (2004) presents a process for conducting and publishing DSR that maps these principles to specific process steps. For example, DSR work begins by identifying and describing a problem to ensure principle 2 (problem relevance) is implemented before it moves on to designing an artifact (principle 2).

Table 1 - DSR Guidelines adapted from (A. R. Hevner et al. 2004)

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Design as an artifact</td>
<td>Design science research must produce a viable artifact in the form of a construct, model, method, or instantiation</td>
</tr>
<tr>
<td>2 Problem relevance</td>
<td>The objective of DSR is to develop technology-based solutions to important and relevant business problems</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>3</td>
<td>Design Evaluation</td>
</tr>
<tr>
<td>4</td>
<td>Research contributions</td>
</tr>
<tr>
<td>5</td>
<td>Research rigor</td>
</tr>
<tr>
<td>6</td>
<td>Design as a search process</td>
</tr>
<tr>
<td>7</td>
<td>Communication of research</td>
</tr>
</tbody>
</table>

I use these guidelines to frame my research and each chapter of this thesis corresponds to part of the DSR methodology, resulting in the outline presented below in Figure 2. (Chapters that are intended to be published as journal articles are highlighted as such).
Figure 2 - Dissertation Outline
The Potential for Automating FCM Based Scenarios

I plan to submit the following two chapters as a review paper to a journal on Future Studies, such as Futures or Technological Forecasting and Social Change.

Future scenarios can inform product development decisions by helping teams understand driving forces and emerging trends in the market, find opportunities, avoid dangers, and prioritize the highest value work (van Notten 2006). In today's turbulent business environment and rapidly evolving technological landscape, these characteristics could help teams drive more innovative products by getting ahead of market trends and prioritizing features that exploit potential opportunities.

Unfortunately, scenario planning is a relatively time-consuming and demanding process, which makes it unavailable to product development teams for several reasons: Firstly, to cover the different but intersecting aspects of the business environment, domain experts are brought in who all contribute their distinct knowledge about the trends and driving forces relating to their field of expertise. However, today's business environments are so complex that bringing in enough experts to expand a team's peripheral vision may not be organizationally feasible because of scheduling, unmanageable large groups, or workshop processes that are too time-consuming.

Secondly, uncertainty often exists at the intersection of trends and fields and might still be missed if domain experts are too focused on their field. Third, in commonly used scenario techniques, scenario teams focus on the big picture, looking at the futures of
entire markets, industries, or even societies, which is useful for high-level strategy, but not for teams that are closer to immediate products and customers and who need more granularity to help with tactical direction. Finally, to build out robust pictures of an environment workshops with deliberation are common, but these settings within an organization have been shown to foster groupthink (Solomon 2006).

Specialized knowledge, expert workshops, time-consuming processes, and very long-term and high-level strategic perspectives are at odds with product development practice, which increasingly utilizes short iterative cycles, and makes it difficult to incorporate scenario thinking into the product development lifecycle. However, with the need for products to continually innovate or die scenarios could be a valuable part of the product development process, though this requires innovative scenario approaches.

In this work, I bridge these gaps by combining FCM-based scenarios and Machine Learning (ML), two newer developments in the field of foresight or future studies to which scenario techniques belong. My technique (i.e., SAAM) reduces or eliminates the need for costly experts in the scenario process, dramatically shortens the amount of time needed to create scenarios, and enables broadly and narrowly focused scenarios, so that they can be adapted for use at multiple levels of organizations. For future research I contribute a list of requirements for integrating these scenarios into the development lifecycle of more agile teams that use short, iterative cycles.

In the remainder of this chapter, I will first introduce background on scenario planning including participants and their roles (section 2.1). I then explain the use of
scenarios in product development and discuss limitations arising for agile product
development teams (2.3). In section 3.4 I outline FCM and ML as two newer
developments in scenario planning that can be combined to provide a solution to the
limited usability of scenario techniques in agile teams. I provide more detail on both
approaches in sections 2.4.1 (FCM) and 2.4.2 (ML) before outlining systematic reviews in
Chapter 4.

2.1 Introduction to Scenario Techniques

Scenarios identify potential futures in story form intending to broaden minds
and break biases (Bowman et al. 2013; van Notten 2006). Originally developed by
industry practitioners to manage future uncertainty, the field of scenario planning has
not only grown in popularity over time, but has remained focused on practice so that
over half of the publications on scenarios are about methodology (Tiberius, Siglow, and
Sendra-García 2020).

There are several schools of thought when it comes to scenario planning and
their methods span a continuum from purely qualitative techniques that rely on
intuition and experience to formal, quantitative techniques that use, among others,
statistical analysis (Amer, Daim, and Jetter 2013; Varho and Tapio 2013). At the
beginning of any scenario study, a topic or domain is chosen and research is done to
identify and analyze influencing factors that shape the future. The further analysis
focuses on a subset of factors that are understood to have a strong influence on future
developments and for which the future states are uncertain, i.e., a scenario planner
estimates that the factor might have a future value ranging from “low” to “medium-high”, or “500K” to “2 million”, or “1% p.a. to 15% p.a.” but cannot confidently forecast a specific value. The goal of scenario planning is to identify plausible combinations of outcome states of scenario drivers, based on the assumption that the drivers either impact each other directly or because they are impacted by the same underlying trends. For example, a scenario team might agree that the combination of factor states “low, 800K, 1.5% p.a.” could occur together and decide to investigate this possible future further, but reject the combination of factor states “medium, 800K, 10% p.a.” because they cannot think of any way in which such a future could occur. The goal of scenario planning is to identify a small number (typically 3-8) of plausible, alternative futures that are quite different from each other and, together, cover much of the “cone of uncertainty” about the future (Amer, Daim, and Jetter 2013). These alternative futures are subsequently described as so-called “narratives” to make them accessible to planning teams (Lindgren and Bandhold 2009). Qualitative and quantitative scenario techniques (and the many mixed approaches in between) differ in how knowledge about factors, their future states, and plausible combinations of factor states is integrated. Purely qualitative techniques trust planning experts to integrate their knowledge into plausible futures without a formal process, though mapping of connections between factors (Wilson et al. 2000; Chakraborty and McMillan 2015) and drawing and other literal visualizations (Godet 2000; Buckman, Arquero de Alarcon, and Maigret 2019) might occur. Fully quantitative techniques might estimate the range of uncertainty for each factor state through statistical techniques and build models that
explicitly represent the relationship between the factors, for example as conditional probabilities (Brauers and Weber 1988; Gordon 1994; Vilkkumaa et al. 2018) or through system models (Enzer 1981; Bishop, Hines, and Collins 2007). Because the future is fundamentally unknowable (and changes in response to a successful scenario exercise, because people aim to influence the future), there is no way of validating scenarios at the time the scenario study is conducted (Thomas Chermack, Susan A. Lynham, and Wendy EA Ruona 2001). Instead, the aim is to ensure that each step of the scenario process is done rigorously, that all drivers are identified, that their future states are carefully investigated, that factor combinations are plausible, and that the resulting scenarios are meaningful and relevant for the type of decisions they are supposed to support (Heijden 2011; Alcamo and Henrichs 2008; Firmansyah et al. 2019). When quantitative models are used, additional steps ensure that models represent what is known and that model results are sufficiently (but not overly) sensitive to different model inputs. Some of the most common concepts for evaluating scenarios include plausibility, where there is a conceivable path to a certain potential scenario from the present state (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007; Spaniol and Rowland 2019), internal consistency to ensure scenarios do not contradict themselves (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007) and relevance of the scenarios to the team (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Durance and Godet 2010). Other evaluation criteria include differentiation between a set of selected scenarios (Wilson 1998), the creativity of the
scenarios (Alcamo and Henrichs 2008), whether they reflect uncertainty and help people create new perspectives (Heijden 2011), and how transparent they are (Durance and Godet 2010).

The many different process steps of a scenario study make it necessary to involve several different roles in the scenario planning process: Domain experts, who I will refer to as “experts,” are people with specialized knowledge (based on research and practice) in a particular domain such as technology trends, changes to the marketplace, or competitors. They can be internal or external to an organization. Stakeholders represent various roles within an organization that have a professional interest in the generated scenarios. Stakeholders can include managers, team leaders, executives, strategic marketing, but also product developers. Figure 3 represents their involvement in the different stages of the scenario process.

Figure 3 - General steps and participants in quantitative, model-based scenario techniques
2.2 Scenarios for Product Development

While scenario techniques are mostly used in strategic planning, they are occasionally used in product development, where they are used to integrate knowledge about the contextual environment (van Notten 2006) including Political, Economic, Social, Technological, Environmental, and Legal (PESTEL) factors and about the transactional environment (van Notten 2006) which includes competition, clients, partners, suppliers, etc. The improved understanding of how the business environment might evolve in the future is used to define the strategy of the organization (van Notten 2006), down to the decision of what type of product development project to engage in as shown in Figure 4. Thus, strategy is set by leaders and project execution and implementation of strategy occur in product development.

*Figure 4 – Where a product team sits in the larger environment, adapted from (van Notten 2006)*
Scenarios have been shown to contribute to product development by improving opportunity identification, idea generation, and idea enrichment (Postma, Broekhuizen, and van den Bosch 2012). In this context, they are useful in identifying varying customer needs over time, especially in uncertain environments (Randt 2015). This is in contrast to traditional forecasting methods typically used in product development which are better suited for mitigating probabilistic risk and do not help product teams think creatively (Derbyshire and Giovannetti 2017). The product development team (referred to as “the team”) is a group of individuals working on and delivering a product to customers in some way, specifically software teams working on digital products or digital product components as they have the most flexibility in the scope of product development. Different people are involved at different points in the scenario process, where experts who are from outside the organization are utilized heavily upfront in preparing for scenario planning, and people from within the organization such as stakeholders are utilized to help build and verify the scenarios (Peterson, Cumming, and Carpenter 2003; Meissner and Wulf 2013). This is exemplified in Figure 5.
2.2.1 Benefits of Scenarios in Product Development

Scenario techniques have been shown to have many benefits for product development. Firstly there is a host of cognitive benefits such as breaking away from groupthink, raising awareness of uncertainty, reducing bias, and accelerating organizational learning (Postma, Broekhuizen, and van den Bosch 2012; Derbyshire and Giovannetti 2017; Meissner and Wulf 2013). Secondly, there are planning benefits associated with scenarios such as mitigating product planning risk in uncertain environments (Collier et al. 2018), connecting strategy with foresight, evaluating strategy selection (Postma, Broekhuizen, and van den Bosch 2012), and aiding product designers in exploring potential futures to shape necessary decisions for open-ended design questions (Anggreeni and van der Voort 2007). Thirdly scenarios can aid in problem-solving because they help people come up with new ideas, explore future possibilities, raise awareness of environmental uncertainty (Postma, Broekhuizen, and
van den Bosch 2012), and help find alternative paths to completion when environments change, or when original schedules can no longer be met (Relich et al. 2020; Wheatcroft et al. 2019). Fourth and finally as all these benefits would imply scenarios have been shown to improve the market success of products (Açikgöz, Latham, and Acikgoz 2020).

These advantages are shown below in Table 2.

**Table 2 - Benefits of scenario planning for product development from the literature**

<table>
<thead>
<tr>
<th>Category</th>
<th>Benefit</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive benefits</td>
<td>Raise awareness to and manage uncertainty.</td>
<td>(Collier et al. 2018; Postma, Broekhuizen, and van den Bosch 2012; Derbyshire and Giovannetti 2017)</td>
</tr>
<tr>
<td></td>
<td>Accelerate organizational learning.</td>
<td>(Postma, Broekhuizen, and van den Bosch 2012)</td>
</tr>
<tr>
<td></td>
<td>Breakaway from groupthink.</td>
<td>(Postma, Broekhuizen, and van den Bosch 2012)</td>
</tr>
<tr>
<td></td>
<td>Reduce framing bias.</td>
<td>(Meissner and Wulf 2013)</td>
</tr>
<tr>
<td>Problem-solving</td>
<td>Come up with new ideas and help explore future possibilities.</td>
<td>(Postma, Broekhuizen, and van den Bosch 2012; Derbyshire and Giovannetti 2017)</td>
</tr>
<tr>
<td></td>
<td>Find alternative paths to completion when environments change, or when original schedules can no longer be met.</td>
<td>(Relich et al. 2020; Wheatcroft et al. 2019)</td>
</tr>
<tr>
<td>Planning advantage</td>
<td>Connect strategy with foresight.</td>
<td>(Postma, Broekhuizen, and van den Bosch 2012)</td>
</tr>
<tr>
<td>Help evaluate strategy selection.</td>
<td>(Postma, Broekhuizen, and van den Bosch 2012; Meissner and Wulf 2013)</td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Market success</td>
<td>Improve market success of products.</td>
<td>(Açikgöz, Latham, and Acikgoz 2020)</td>
</tr>
</tbody>
</table>

All these benefits have been documented in the literature, yet scenario planning is not being adopted into the development lifecycle for many of today’s products, rather they are often used upfront for strategic direction (Gausemeier, Fink, and Schlake 1998; Derbyshire and Giovannetti 2017) to determine the potential risk of a project (Nissen, Pretorius, and Klerk 2017; Galli 2017), for early assessment of risk and feasibility (Fasterholdt et al. 2017), to help create robust architectures (Schuh et al. 2014), and to help developers understand how people might use the product early in the development cycle (Betten et al. 2019). Thus, scenarios are done only once and not updated throughout the project.

Because scenarios are generated once, at the onset of the actual product development, when risk and uncertainty are at their highest (Postma, Broekhuizen, and van den Bosch 2012; A. J. M. Jetter 2003), the methods used for generating product development scenarios do not differ from those traditionally used for strategic planning. They also suffer from the same limitations: scenario planning is a cumbersome process that usually comprises of long workshops with expert practitioners that require the participation of leaders from across an organization (Schoemaker 2020; Rafael Ramirez et al. 2017; Hodgkinson et al. 2006; Franco, Meadows, and Armstrong 2013; O’Brien
Because scenario development utilizes expert opinion there always is a risk of introducing significant bias (Bonaccorsi, Apreda, and Fantoni 2020), particularly if too few experts and too little diversity are introduced to the project. Though many recognize that there are benefits to scenario planning, the process of creating scenarios is recognized as being inefficient (Kwon and Park 2018). These limitations are shown in Table 3.

Table 3 - Scenario Planning limitations from the literature

<table>
<thead>
<tr>
<th>#</th>
<th>Limitation</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cumbersome process involving many people across an organization</td>
<td>(Schoemaker 2020; Rafael Ramirez et al. 2017; Hodgkinson et al. 2006; Franco, Meadows, and Armstrong 2013; O’Brien 2004)</td>
</tr>
<tr>
<td>2</td>
<td>Risk of introducing bias from experts</td>
<td>(Bonaccorsi, Apreda, and Fantoni 2020)</td>
</tr>
<tr>
<td>3</td>
<td>Inefficient process</td>
<td>(Kwon and Park 2018)</td>
</tr>
</tbody>
</table>

Typically, scenarios are used for products with long development cycles because they allow product teams to think critically about a relatively distant future their product will be used in and how it is different from the present (Randt 2015; Postma, Broekhuizen, and van den Bosch 2012). For example, scenarios have been used to help development teams target future goals such as sustainability targets that may not exist today (Gaziulusoy, Boyle, and McDowall 2013), or build strategies based on the future
availability of materials (Weiser et al. 2015). Moreover, in the case of aircraft design, scenario planning is used to attempt to understand the potential future when the product will be delivered (Randt 2015). It has been shown that scenarios can be used in the forecasting process for products with long development cycles in a market with rapid change such as mobile networks by helping explore optimistic, neutral, and pessimistic potential outcomes (Sungjoo Lee et al. 2016). Thus, for projects with long development cycles, during which the business environment can change drastically, upfront scenario planning is worth the effort and is effective (Sungjoo Lee et al. 2016). The picture is much less clear in the case of digital products (or products with digital components) with short and iterative development cycles, which will be discussed in the subsequent section.

2.2.2 Scenarios in Agile Projects: Potentials and Problems

There are millions of mobile applications available for various platforms that are completely digital (“How Many Apps Are There in the World?” 2021) and many physical products are incorporating digital components through the Internet of Things (IoT) or their own digital experience (I. Lee and Lee 2015). Even products with typically long development cycles such as automobiles have incorporated digital components that are updated regularly after a customer has already purchased it (“Software Update Frequency and Substance?” 2016).

In such an environment, agile methodologies are normally used for product development which organizes work in short cycles called “sprints” that can have very
short durations, from weeks to even days in some cases. These cycles typically include reviewing requested features, prioritizing them, then building and testing them before releasing features to customers and starting again (Williams and Cockburn 2003; Christopher W H Davis 2015) as demonstrated in Figure 6.

![Diagram](image)

**Figure 6 - Example Agile development cycle**

There are several reasons for the adoption of agile development frameworks: frequent feedback ensures that the team only works on things that customers value or quickly figures out what the customers do not value, rather than to take off in an unproductive direction for weeks and months on end. Frequent feedback also reduces the need to painstakingly specify all product aspects upfront, which is particularly challenging for software products, where knowledge tends to be sticky and implicit.

Finally, if customer needs shift throughout the project, the development team becomes aware of them and can respond (Sharma, Sarkar, and Gupta 2012; Mohammed and Rauf 2015). Overall, this shortens the time it takes to deliver features to the market and ensures that products are adapted to market needs.
However, there are also drawbacks: Agile teams do not just implement the product requirements given to them by leadership after strategic product planning has been completed but are empowered to adjust throughout the project. Product definition thus is increasingly being handled by autonomous teams that sit firmly within a department of an organization (see the inner oval in Figure 4, partially adopted from (van Notten 2006)). However, the product team’s view of the world will naturally include a small slice of all the factors that affect their product and the environment it is delivered into. Moreover, as development cycles shorten teams gain the ability to change direction faster, but they also get less time to plan and gather requirements (Leau et al. 2012). Ironically, though a team can change direction frequently the shortening of cycles forces them to focus on their micro-environment rather than to take a step back and better understand their macro-environment as demonstrated in Figure 4, a phenomenon known as “the build trap” (Perri 2018). It has been shown that in this agile environment teams are susceptible to ‘group process loss’, where teams suffer from groupthink and focus on current knowledge rather than broadening their horizons (Coyle, Conboy, and Acton 2013; Przybilla, Wiesche, and Krcmar 2019).

This has implications for the use of scenario planning. Scenario planning is typically done to set a strategic direction to define a new product before development begins. Agile teams, however, have the autonomy to make any changes through the development process under the strategic radar, which could cause them to become disconnected from the strategic work done upfront. Also, agile teams typically do not
use the strategic scenarios develop upfront during their development process because they only help set direction but do not describe what needs to be done now and are therefore not easily applicable to the task at hand. In theory, this problem could be alleviated in two ways: by sharing the scenarios with the agile team as an ongoing point of reference or by involving the agile teams in the scenario creation. Unless either of these approaches succeeds, any benefits from using scenario planning at the beginning of a project will be lost over time as the team produces short-term plans and iterates on their product to deliver into rapidly changing environments. This is demonstrated in Figure 7 below.

![Figure 7 – How Agile processes lose sight of scenario planning over time](image)

**2.3 Requirements for Scenario Techniques in Support of Agile Teams**

The first potential solution described in the earlier section – to share the scenarios that were created upfront with the agile teams to improve their strategic mindset so that they can make better tactical decisions – is problematic because scenarios can be difficult to interpret: in the scenario and strategic planning workshops,
the experts and stakeholders who helped create alternative futures can interpret them but if these products were simply handed off to a development team they would have to be self-explanatory and readily applicable to the task at hand (Leffingwell 2010). I posit that for this to happen scenarios need to use enough data to provide visibility into the broader environment but communicate it in a way where teams can relate them directly to their product and the environment their product is being used within. This requires finding and distilling data that is granular enough for the specific project and presenting it back to people in a way where it helps broaden their perspectives.

However, the use of static scenarios at the onset of a development project is unlikely to be sufficient in a dynamic business environment because none of the scenarios will cover all aspects of the change the agile teams are forced to respond to. After all, one of the core tenants of agile teams is to quickly respond to change rather than follow a well-defined plan (Williams and Cockburn 2003), which makes understanding the nature of coming change an important factor in their planning. This makes the second potential solution – to have agile teams create and update their own scenarios and consult them as they undergo their planning in every iteration – more feasible. Such an approach would allow teams to combine the benefits of strategic futures thinking with short term agile planning: at the beginning of a project, the team could create project-specific scenarios based on the strategic direction set upfront, which would help them better understand the potential market their product will be released in. Then, at the beginning of each development iteration, the team could refresh their scenarios with current data to help them understand how their target environment is evolving.
Generating multiple scenarios to think through alternative future situations would also combat groupthink and help the teams keep a more diverse perspective which could result in more relevant and successful products. These two different processes are shown visually in Figure 8 where the current way of using static scenarios is represented by Process A and the proposed method of using dynamic scenarios is represented as Process B.

**Process A: Static scenarios used for product development**

Broad, static scenarios designed up front and handed to the development team

<table>
<thead>
<tr>
<th>Ideation</th>
<th>Initial New Product Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decide what to build based on strategy</td>
<td></td>
</tr>
</tbody>
</table>

Iteration 1 ... Iteration n ... Iteration x

Product Development

Original scenarios describe potential futures that eventually become irrelevant

Benefits of scenario planning are not carried forward

**Process B: Dynamic scenarios used in product development**

Broad, static scenarios designed up front, *parameters are used to inform scenarios for the product team.*

<table>
<thead>
<tr>
<th>Ideation</th>
<th>Initial New Product Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decide what to build based on strategy</td>
<td></td>
</tr>
</tbody>
</table>

Time

Scenarios refreshed and reviewed every iteration

Iteration 1 ... Iteration n ... Iteration x

Product Development

Benefits of scenario planning realized over time

*Figure 8 - Static vs dynamic scenarios for product development teams*

To integrate scenario planning into the shorter development cycles of an agile team and use them effectively as an approach for developing a strategic mindset and for
supporting tactical decisions, a future scenario approach has to address five main requirements: First (R1) scenarios need to reflect the diversity of knowledge sources used in product development to ensure that teams are widening their peripheral vision and considering more than just what is directly in front of them. Secondly, (R2) experts bring in rich outside knowledge and the process and participation help facilitate strategic thinking, so anything used to speed up these processes will need to ensure these qualities of scenario planning are not lost. To keep scenarios up to date at the rate of development large amounts of data would need to be processed to find emerging concepts, also known as weak signals (Sutherland and Woodroof 2009; Gokhberg et al. 2020) the team should be aware of. Third, (R3) scenarios need to be easy to update once they are built so they can be integrated into the continuous lifecycle of modern development. Fourth, (R4) scenarios need to be interpretable by non-expert scenario planners because the team must interact with the scenarios regularly. Fifth, (R5) and finally these scenarios should use quantitative methods to allow for integration through automation and to allow measurement and accuracy. These requirements are outlined in Table 4.

*Table 4 - Requirements for integrating scenarios into product development teams with number labels*

<table>
<thead>
<tr>
<th>#</th>
<th>Requirement</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Need to reflect diversity of knowledge sources used in product development.</td>
<td></td>
</tr>
</tbody>
</table>
2.4 Foundations of a future solution: FCM Scenarios and Interactive Machine Learning

To meet the requirements identified in Table 4 above, this thesis develops a method and software for Scenario Acceleration through Automated Modelling (SAAM). It combines two relatively recent additions to the practice of scenario planning, namely scenarios based on Fuzzy Cognitive Map (FCM) and Interactive Machine Learning (IML), which will be explained in the following.

Fuzzy Cognitive Maps (FCM) are used in scenario planning because they result in quantitative system models of the business environment that represent the driving forces that shape the future (e.g., technology, economy, social trends) and their interdependencies (A. Jetter and Schweinfort 2011; Amer, Daim, and Jetter 2016; Pei Zhang and Jetter 2016). The models can be used for simulation so that the same model with varying inputs and assumptions produces multiple scenarios (outcomes) that are different and internally consistent. The core of FCM models are directed graphs and causal cognitive maps that non-technical people can interact with (Nápoles et al. 2017).
An example directed graph is shown below in Figure 9, where each node directionally relates to another as indicated by the arrows between them.

![Directed graph example](image)

**Figure 9 - Directed graph example**

Cognitive maps and their relatives, concept maps, have been shown to improve learning and broaden perspectives (Lourdel et al. 2007; Villalon and Calvo 2011; Novak, Markey, and Allen 2007), to help people integrate new ideas into their current way of thinking and to construct meaning with visual models (Vanides et al. 2005). FCM thus provide not only to be a powerful modeling method but also a powerful communication tool that helps people understand how experts understand a specific domain and how models represent this insight (E.I. Papageorgiou, Stylios, and Groumpos 2003; Pei Zhang and Jetter 2016; A. Jetter and Schweinfort 2011). This allows non-expert modelers to contribute their knowledge during model building, update models as new information becomes available, and interpret scenario results. FCM and their close relatives, concept maps, have been created as an output from text mining (Son, Kim, and Kim 2019; Kwon and Park 2018; Sandhu et al. 2019) which opens the possibility to efficiently create scenarios by extracting and distilling information from text.

Machine Learning (ML) – and its sub form, Interactive ML (IML) are at the core of the phenomenon of Artificial Intelligence (AI), the ability of smart machines to perform
tasks that typically require human cognition, experience, and reasoning and to thus appear intelligent. ML is enabled by the availability of Big Data: In 2020, it has been estimated that there are already 44 zettabytes of data that come from the increased use of digital experiences such as Facebook, Twitter, Google, YouTube, etc. and by 2025 it is estimated that over 460 exabytes will be created every day from online activity (“How Much Data Is Created Every Day? [27 Powerful Stats]” 2020). Among others, this data includes human language: ML uses this data to learn what words are used synonymously, which words belong to the same topic, what words are used to express emotions, what words indicate a relationship between the concepts represented by words, etc. These capabilities are attractive for supporting and speeding up scenario generation because they make it possible to sort through massive amounts of text, find topics, relations, and causality, and build models that could be used for what-if analysis. Product development teams could use ML to scan through large volumes of text that are specifically selected to reflect their development project and create FCMs as the engine for scenario construction, and automatically run simulations to develop alternative futures or update them when new information becomes available. Because ML can distill large amounts of data, this approach could reduce the need for domain experts to conduct and present specialized research. This process may further be aided by digital tools that allow virtual participation which has recently been adopted by many companies (Savić 2020, 1). Virtual, rather than in-person interactions could make participation more efficient, so that scenario teams can benefit from the knowledge of other teams and internal and external experts without the need of adding many people.
to a workshop. Cognitive maps, which can be represented and collaboratively worked on by such tools, have been shown to help people understand complex dependencies and might also decrease the need for lengthy workshops. Taken together this combined approach could result in a scenario process that is suitable for agile teams: they utilize ML to collect and process data from the contextual (external) environment and use FCM to apply scenarios to the relevant context. Figure 4 could thus be updated as shown below in Figure 10. In the following section, I will introduce the two foundations of this proposed approach in more detail and discuss FCM and ML in general, as well as IML, which is at the core of this dissertation.

**Figure 10 - Levels of scenarios, adapted from (van Notten 2006), modified from Figure 4**

2.4.1 Fuzzy Cognitive Maps Explained

Among existing scenario planning techniques, FCM-based scenarios have become increasingly popular. FCM act as a middle ground between qualitative and
quantitative methods because they can be built using qualitative methods and work well in a workshop setting but they can also be run as quantitative models (Amer, Daim, and Jetter 2013; van Vliet, Kok, and Veldkamp 2010). FCM, originally developed by Kosko (1986), are graphical models made up of concepts with edges between them that demonstrate causality. Each edge is assigned a weight that represents the strength of causality so FCM can be run as mathematical models as well as visualized as graphical models. This duality allows FCM to visually communicate different actors in a scenario and be used to generate what-if scenarios by changing parameters and running simulations (P. Zhang and Jetter 2018; van Vliet, Kok, and Veldkamp 2010).

FCM has been used to demonstrate scenarios in several domains from autonomous vehicles, natural resource production and allocation, energy development, and new product development to name a few (Elpiniki I. Papageorgiou and Salmeron 2013; Felix Benjamín et al. 2019). All these studies combine qualitative analysis to build models that can then be used to demonstrate what the future could look like in the prospective field to influence policy or help people make better decisions that have major impacts on the f

different ways all studies that use FCM will start by identifying relevant concepts ($C_i$) and connecting them together with weighted edges. Each weight represents the causal influence between two concepts in the M. For example, considering concepts $C_i$ and $C_j$ with a weighted edge of $W_{ij} > 0$, then an increase or decrease in $C_i$ will result in the same effect on $C_j$. If $W_{ij} < 0$, an increase or decrease in $C_i$ will result in the opposite
effect on $C_j$. Finally, $W_{ij} = 0$ when there is no causal relationship between $C_i$ and $C_j$.

The amount of influence is indicated by the absolute value of $W_{ij}$.

When the model is run as a simulation the value of each concept $A_i$ is updated according to its relations to other concepts. The FCM formula has several variations but the one more commonly used originated from Kosko (1986) (Felix Benjamín et al. 2019) is demonstrated in Formula 1.

$$A_i^{(k+1)} = f\left(\sum_{j=1,j\neq i}^{n} W_{ij}A_j^{(k)}\right)$$

*Formula 1 – FCM calculation formula.*

Where $k$ is the iteration, $n$ is the number of concepts and $f$ is the squashing function which keeps the resulting value in the range of [-1,1]. There are different options for the squashing function, two that are commonly used are the sigmoid and hyperbolic tangent, shown in Formulas 2 and 3.

$$\frac{1}{1 + e^{-\lambda x}}$$

*Formula 2 – The Sigmoid*

$$\frac{e^x - e^{-x}}{e^x + e^{-x}}$$

*Formula 3 – Hyperbolic Tangent (TANH)*
Concepts of an FCM typically include all factors that experts consider relevant to a problem domain because they directly or indirectly affect an outcome of interest. In a very simple example, we can look at online retail or eCommerce. A retail expert may determine that some of the relevant concepts that drive eCommerce are the availability of products in a local market, the ability to deliver to homes, and the quality of internet connectivity. These may affect if retailers are going to invest in eCommerce or increase the square footage of their retail locations to accommodate more foot traffic. Concepts can also be labeled as input, state, or output concepts where input concepts represent the inputs to the system, state factors describe the operation of the system, and output concepts are the final values the modelers are concerned about (Mpelogianni and Groumpos 2018; Harmati and Kóczy 2020). The concepts in our example are labeled and cataloged as shown in Table 5.

Table 5 - Example Concepts

<table>
<thead>
<tr>
<th>Number</th>
<th>Concept</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Global Product Availability</td>
<td>Input</td>
</tr>
<tr>
<td>C2</td>
<td>Home Delivery</td>
<td>Input</td>
</tr>
<tr>
<td>C3</td>
<td>Internet Prevalence</td>
<td>Input</td>
</tr>
<tr>
<td>C4</td>
<td>Shop from Home</td>
<td>State</td>
</tr>
<tr>
<td>C5</td>
<td>Retail square footage</td>
<td>Output</td>
</tr>
<tr>
<td>C6</td>
<td>Ecommerce</td>
<td>Output</td>
</tr>
</tbody>
</table>
Once concepts are identified causality is inferred between them (P. Zhang and Jetter 2018; Aminpour et al. 2020; Kosko 1986). For example, if home delivery improves that should have a positive influence on shopping from home. If shopping from home increases, then that should increase the focus on eCommerce and perhaps decrease the need for larger retail locations. The causality is associated with linguistic values, so something may increase “a lot” or decrease by “a little”. Linguistic values are then turned into values in the range of [-1,1] through algorithms that are specific to the use case (Groumpos 2010). This can result in a graphical representation of the FCM as shown in Figure 11 and can be displayed as a matrix representation shown in Table 5. In many cases, FCM concepts are mapped to concept numbers such as C1, C2, etc. to make it easier for labeling, as shown in Table 6.

Figure 11 - An example FCM
### Table 6 - Matrix representation of example FCM

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C4</td>
<td>0.2</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>-0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>C5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The model is run by setting inputs performing the calculation in Formula 1 until the model reaches a stable state where the values no longer change significantly (Giabbanelli, Fattoruso, and Norman 2019). In this example setting the input of the model to \([1,1,1,0.5,0,0]\) results in C1, C2, and C3 starting with values of 1, which means there is plenty of Global product availability, home delivery is logistically possible, and people have strong access to the internet. Running the model with these inputs results is shown graphically in Figure 12.
Figure 12 – Results from running FCM simulation in our example charted until reaching an end state

We can see from Figure 12 that the results stabilize after 6 iterations and given the initial values of the input concepts the output shows that e-commerce will boom and the need for retail square footage should decrease, as expected by looking at the graphical model in Figure 11. Conversely, if we set all the inputs to low values such as [-1,-1,-1,-1,.5,0,0] indicating that products are not available home delivery is not feasible, and people do not have good internet bandwidth, we can see that the opposite happens as shown in Figure 13.
This output aligns with our understanding of the model as graphically represented above in Figure 13 however when models start to become more complex the output may not seem as obvious as our simple example.

Simulations run through this quantitative model result in numeric values that need to be translated back to natural language. This is done by creating a narrative using the output of the model. In our example first scenario we can say that in a scenario where there is no reliable internet, products are not globally available, and home delivery is not possible people will likely not use e-commerce but instead will go to physical retail locations. Conversely, if the three inputs are true eCommerce would boom and fewer people would visit physical retail. Decision-makers can use this input to decide their future strategy, in this case, whether to invest in eCommerce or increase their physical retail footprint.
FCM are a tool used to bridge the gap between objective and subjective techniques and can be calculated with little data and low compute making them fairly easy to construct and run simulations with. Their ability to be represented graphically makes them accessible to broad audiences which fosters transparency in the model and more inclusivity on the team that builds and uses them. However, for teams to be able to use FCM as a viable tool for continuous, evolving product development these models need to change to reflect the environment the product is being developed for throughout the development cycle. This means these models need regular access to new data, must be updated frequently, and new simulations should be run to help teams stay up to date with rapidly evolving landscapes.

Such a vision of easy and frequent updates of FCM for scenarios in light of new qualitative data is currently only in its infancy, though there have been attempts to automate the creation of cognitive maps from summaries (Villalon and Calvo 2011) and large numbers of documents (Hajek, Prochazka, and Pachura 2017; Son, Kim, and Kim 2019; Pillutla and Giabbanelli 2019). These methods end up being semi-automatic which requires data labeling or some type of expert verification.

2.4.2 Interactive Machine Learning Explained

Machine Learning (ML) encompasses a variety of algorithmic approaches, ranging from statistics to artificial neural networks, to identify patterns, rules, or trends in large data sets to generate previously unknown information. Programming ML is so different than previous development practice that has been termed “Software 2.0”;
rather than writing code that logically evaluates inputs ML uses pre-defined algorithms that create models from large amounts of data which are then used to take action on new data (Johannes Gehrke 2019). ML is a large and growing field used in a variety of different applications and this dissertation focuses on how ML is used to find insights in large amounts of semi-structured or unstructured text and how to improve the process by combining ML with human interaction.

The process of extracting insights from data is characterized as “mining” and, accordingly, the term data mining is used for any type of data, web mining is used when the data exists in the world wide web, and text mining (or text data mining) is used when the data in question are written resources with unstructured text (Michael W. Berry and Jacob Kogan 2010; Lin, Hsieh, and Chuang 2009). To make data accessible for mining, it needs to be collected and made available in a format that can be mined. Techniques that support or automate this process are characterized as “scraping”. For example, web scraping, web harvesting, or web data extraction are all used for techniques that extract data from websites. People use different vocabularies and speech patterns in different contexts, so the language used on social media posts differs from, for example, the language used in scientific journals, where words might also take on a different meaning. Importantly, spoken language also differs greatly from written texts. Accordingly, there are many approaches to making sense of language by understanding the context in which it is used. The associated field of research and practice, Natural Language Processing (NLP) broadly refers to the study and
development of computer systems that can interpret speech and text as humans naturally speak and type it. One method used to achieve this goal is Topic Modelling (TM), which creates a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. This is made possible using distributional semantics, where co-occurrence, similarity, and frequency of words are used to determine their meaning (Lenci 2008). One popular implementation is called Latent Semantic Analysis (LSA) (Thomas K. Landauer, Peter W. Foltz, and Darrell Laham 1998) which can produce a set of concepts related to the documents and terms.

The statistical models used to identify topics, the relationships between concepts (terms) within a topic, and the type of relationship are created with the help of very large data sets and are broadly applicable in many different applications – sometimes after some customization or training. Currently, all major tech companies that own the big data needed to create powerful models heavily invest in model creation (e.g., Microsoft, Google, Amazon) and make the resulting models available to researchers and practitioners at very low cost or free of charge, as long as they use their web services. They also strive to make the models easy-to-use, organizing them as modules that can be called in standard formats and with relatively few lines of code. These modules can contain sophisticated neural networks, such as BERT (Devlin et al. 2019). BERT builds general language models by masking random words in large amounts of data such as Wikipedia or the Google News Database, then uses predictive neural networks to determine the meaning of the masked words. This creates pre-trained
foundational models that can be “fine-tuned” on a specific dataset to answer questions about it or derive specific insights. The idea is that BERT contains everything you need to understand language in general, and you can use it to analyze your specific data to provide insights from it or answer questions about it. The availability of modularized ML solutions makes it possible to create a highly automated workflow that combines multiple, sophisticated techniques simply by selecting and combining modules. In my dissertation, I mainly use the Microsoft suite of products such as Azure Machine Learning for data hosting and compute (Blackmist 2021), Machine Learning Pipelines for coordination of tasks (lobrien 2021), and Power Automate Desktop to automate data collection (“Power Automate Desktop” 2021).

The high level of automation, and particularly the “black-box” character of artificial neural networks used for ML have resulted in the concern that resulting solutions may not be interpretable and therefore not meaningful and trustworthy to human decision-makers (Stephanie M. Merritt 2011). In addition, even the best ML today is so highly specialized it has become generally accepted by researchers and practitioners that the best way to harness the potential of AI today is through collaboration with people (Dellermann et al. 2019). The field of Interactive Machine Learning (IML) addresses this by using people in the training process with the hope that they can fill the gaps in data used for training or help create better models faster (Fails and Olsen Jr 2003). By adding a human to help train these algorithms IML has been shown to facilitate co-learning between people and computers (Fiebrink, Cook, and
Trueman 2011), speed up the learning process for ML, and give the people in the loop a better overall understanding of the model (Robert et al. 2016).

Within the field of ML, TM and NLP are the most promising for addressing the bottlenecks of the scenario creation process because they are used for finding insights in large amounts of text data which is perhaps the costliest part of the scenario process managed by expert practitioners today. In addition, by weaving TM and NLP together with IML it could be possible to create a socio-technical system that can identify topics in large bodies of text, create connections between them, and facilitate co-learning between the ML and people who help it train. Throughout this dissertation, I will focus on these specific components, namely TM for identifying concepts, NLP for deriving connections between them, and binding the process together in well-defined IML workflows.
3 Systematic Review of the State-of-the-Art

The prior chapter identifies the need for product scenarios for agile teams and the potential of combining FCM and data-driven approaches to meet this need. Before embarking on proposing and developing the solution at the center of this dissertation – SAAM, I reviewed the literature to understand if similar solutions have been proposed elsewhere and if their design can contribute to my work. To this end, I did several literature reviews that built on each other:

- I searched for work on FCM scenarios in agile product development, using the search terms “FCM” AND “Scenario planning” AND “agile” AND “product”. Despite including grey literature and Google Scholar (rather than an indexed academic database) this search only yielded nine results and none of these papers were relevant to my topic: if FCM modeling is used for scenarios in agile projects at all, nobody is publishing about it.

- I searched to identify papers that use FCM-based scenarios in product development (including non-agile teams and processes).

- I searched for publications about the use of data-driven approaches (ML, AI, NLP) in scenario techniques (without a specific focus on FCM or product development).

In contrast to the first search, the latter two identified multiple sources. Accordingly, I was able to conduct a systematic literature review, following the PRISMA guidelines to identify and screen articles for inclusion (Moher et al. 2009). This method
starts with a broad search across multiple databases, removes duplicates, then screens articles at the title and abstract level to ensure their relevance. The papers identified in this process are subsequently reviewed in full. I am reporting on both reviews separately in sections 3.1 and 3.2.

3.1 Systematic Review of the use of FCM-based Scenarios in Product Development

I started with a search for ("scenario" AND "product development" AND "fuzzy cognitive map") across five different knowledge databases: ACM, Web of Science, Emerald Insight, IEEE, and Science Direct. The initial search results are shown in Table 7.

*Table 7 - Initial literature review search results (1)*

<table>
<thead>
<tr>
<th>Database</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM</td>
<td>0</td>
</tr>
<tr>
<td>Web of Science</td>
<td>0</td>
</tr>
<tr>
<td>Emerald Insight</td>
<td>32</td>
</tr>
<tr>
<td>IEEE</td>
<td>0</td>
</tr>
<tr>
<td>Science Direct</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
</tr>
</tbody>
</table>

The second step in the PRISMA process was to find and remove duplicates, but there were no duplicates across all results. In the third step, I conducted a Title and Abstract Review (TAR) to ensure the results were on topic. This paired the list down to 19 papers. I then read through the papers to ensure that they utilized FCM for the use of
scenario planning and as a result paired the list down to 12. Finally, I reviewed these 12 papers in detail to find which ones use FCM based scenarios for product development, this brought me to only 7 relevant papers. The results of this process are shown below in Figure 14.

Based on this work, the state of the art can be characterized as follows: Research that uses FCM scenarios in the context of product development is predominantly focused on early development stages, where it is used to provide strategic direction and to inform a technology roadmap (Amer, Jetter, and Daim 2011) or to manage uncertainty at the fuzzy front end (A. J. M. Jetter 2003). One study builds FCM by combining knowledge from customers and marketing insights to determine what type of fashion to manufacture based on a combination of views from customers and marketing insights (Choi 2016). Second, there is a strong emphasis on the role of FCM as a tool for capturing and communicating knowledge: three closely related papers with overlapping authors analyze how knowledge is used in product teams (Y. S. Kim and Kim 2012; K.-Y. Kim and Kim 2011; Cheah et al. 2011). They do not use FCM to generate scenarios but they do show that causal knowledge such as what is represented by FCM helps product
teams more than procedural knowledge, that is knowledge gleaned from text searching (Y. S. Kim and Kim 2012; K.-Y. Kim and Kim 2011; Cheah et al. 2011). These studies specifically support product development teams by incorporating knowledge over time and show how causal maps can be used to build such knowledge, particularly for iterative development which improves the knowledge of the team as the product develops since lessons learned are incorporated into the model (Y. S. Kim and Kim 2012). This insight shows that FCM could be used as an effective communication tool for iterative agile projects which could incorporate their own knowledge into the FCM. Another study shows similar results, where data is incorporated from the lifecycle of a product as an FCM and used as a communication tool for the product team showing improvements to the product over time thanks to the accumulation of knowledge (Udoka, Uzochukwu, and Balogun 2016).

This review shows there is a significant opportunity to explore using FCM based scenarios for product development. Though FCM has been used for creating strategic scenarios and has been proposed for scenarios in product development the use of FCM in this field has yet to be thoroughly explored. Furthermore, it is of interest that FCM has been shown to help product teams retain and improve knowledge which improves their products.

3.2 Systematic Review on the Use of Text Mining for the Creation of Scenarios

The systematic review of text mining in scenario development, based on the search terms (“text mining" OR "data mining" OR "web mining" OR "NLP") AND
"scenario development” yielded a larger and richer data set and is, consequently, described in more detail.

The automation of scenarios as a form of foresight is recognized as a valuable possibility (Raford 2015) and some have explored Foresight Support Systems to aid in the creation of scenarios or other types of foresight methodologies (von der Gracht et al. 2015). Because scenarios are so impactful yet the process is so cumbersome Big Data from Web 2.0 has been recognized as a way to rapidly gather data to potentially be used for scenario development (Raford 2015). Specifically, text mining has been linked to scenario development because the benefits of text mining closely align with the stages of scenario development (Kayser and Blind 2017). In scenario planning typically teams start off preparing with interviews before moving into a desk-research phase where interviews are analyzed and combined with other data sources. These steps could potentially benefit from automation in the form of text mining from big data which would allow us to find insights from large volumes of text by finding topics and keywords, document similarity, and how much a certain topic plays a role in a certain text. The next step in scenario planning is to do key-factor analysis with stakeholders and experts in a workshop setting which leads to creating scenarios. This can benefit from automation through the use of NLP and FCM, where NLP is used to analyze the text gathered in the previous step, and FCM are used to build quantitative, graphical scenarios. The techniques could be used to augment or dramatically speed up efforts at the beginning of the scenario development process and allow us to generate scenarios
as frequently as we want, so they could be used in nearly every planning exercise, not just high-level strategy. The potential for automation in the scenario process is shown visually below in Figure 15.

![Figure 15 - How automation can help scenario generation](image)

Given the potential of text mining to speed up the process of scenario building, I was interested in the start of the art. I started by searching with the search term ("text mining" OR "data mining" OR "web mining" OR "NLP") AND "scenario development" in Science Direct, Web of Science, IEEE Xplore, ACM, and Emerald Insight databases. The terms “text mining”, “web mining”, and “data mining” are frequently interchanged in the literature because typically data in the form of text is mined or scraped from the internet, or web. The initial results are shown in Table 8.

![Table 8 - Initial literature review search results (2)](table)

52
Secondly, I removed duplicates which narrowed the results to 152. Thirdly, I conducted a title and abstract review (TAR) to remove any papers that met the search criteria but did not reference text mining in the context of scenarios. This honed the results down to 71 papers. In step four I skimmed through the papers to ensure they use text mining in some way to help create scenarios to ensure that each paper fit the objectives of this review. This weeded out papers that only referenced text mining or scenarios but did not use them together, bringing the list further down to 30 papers. I reviewed these documents in detail to understand the type of data and methodologies they used and how they combined automated approaches with scenario planning. Some of the papers attempted to speed up the scenario process using automation but used none of the text-based approaches I am interested in. Other techniques included using statistical measures from data collected through simulation (Bryant and Lempert 2010), storing manually created scenarios in databases for reuse (Kishita et al. 2020), and distributing scenarios created by executives across the rest of an organization (Ramírez, Österman, and Grönquist 2013). After excluding these papers, five papers were left.

<table>
<thead>
<tr>
<th>1) Initial Search</th>
<th>2) Remove Duplicates</th>
<th>3) TAR</th>
<th>4) Level 5 Analysis</th>
<th>5) Detailed Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>152</td>
<td>71</td>
<td>30</td>
<td>5</td>
</tr>
</tbody>
</table>

*Figure 16 - Literature review stages (2)*
A subsequent search through gray literature identified a significant paper published in the proceedings of a business conference. Detailed analysis of this paper and the five papers from the systematic review yielded the following insights:

The described methods were far from standardized concerning data sources and with regard to the text-based methods used. Sources range from sites that have aggregated opinions on the future (e.g. MIT Technology) to Twitter feeds that look for specific terms, to journals, patents, and industry reports. These sources differ greatly in content, length, and structure. They are shown below in Table 9.

*Table 9 - Data sources used by studies that automate parts of the scenario creation process*

<table>
<thead>
<tr>
<th>Paper</th>
<th>Data Sources Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Kwon and Park 2018)</td>
<td>MIT Technology Review, Wired, the guardian, io9, and Business Insider</td>
</tr>
<tr>
<td>(Kayser and Shala 2020)</td>
<td>Twitter</td>
</tr>
<tr>
<td>(Gokhberg et al. 2020)</td>
<td>Patents, journals, news feeds, forecast reports</td>
</tr>
<tr>
<td>(Feblowitz et al. 2021)</td>
<td>Newsfeeds</td>
</tr>
</tbody>
</table>

As a result of different data sources and decisions made by the authors, methods also varied greatly. Most used LSA for topic modeling but they had nearly nothing else in common. None of the studies demonstrated full automation of the scenario creation
process, rather they all had manual intervention at different stages. The closest to full automation starts with web scraping, uses LSA for topic modeling, but manually tags the macro-topics before doing a basket analysis to determine causality among concepts (J. Kim et al. 2016). Two studies from the same primary author use web scraping and topic modeling, but then use the topics discovered as input to the manual scenario creation process (Kwon, Kim, and Park 2017; Kwon and Park 2018). Another study mines Twitter scrapes referenced websites, then uses morphological analysis, a method within the discipline of linguistics that analyzes words based on the meaningful parts contained within (Fritz Zwicky 1948), to explore all potential options, as the basis to manually create scenarios (Kayser and Shala 2020). Gokhberk et al. (2020) utilize text mining to enhance desk research and enrich the data used by experts to create scenarios. Finally, the last study that meets my criteria asks users to identify risks they want to track, then uses NLP to identify connections between forces that create such risks and present results back as graphically based scenarios that users can interact with (Feblowitz et al. 2021). The methods used in these papers are shown visually in Figure 17, which indicates the steps in the process and where automation is used.
As a result of different data sources and decisions made by the authors, methods also varied greatly. As Figure 17 shows, the three papers by a Korean research team (J. Kim et al. 2016; Kwon and Park 2018; Kwon, Kim, and Park 2017) and work by Kayser & Shala collect data through web scraping and, in the case of Kayser & Shala, also scraping...
Twitter data. Web scraping is the process of using automation to navigate to various web pages, copy text from them, and move it to a database for further processing. Where four of the studies used automation upfront to collect data, Gokhberg et al. (2020) started the process with desk research, the process of collecting data manually, to identify the most useful sources of data across news feeds, journals, patents, and forecast reports before processing it. The five papers that do not do desk research upfront get around it by targeting their automation at pre-aggregated text in a specific field (J. Kim et al. 2016; Kwon and Park 2018; Kwon, Kim, and Park 2017), by searching Twitter for specific terms and following the associated links (Kayser and Shala 2020), or simply by asking users what they should be looking for (Feblowitz et al. 2021). The reason for pre-aggregated text is that they have already been compiled by experts and therefore should be dense and reliable sources. Twitter is used to represent the “wisdom of the crowd”, and in theory is used to find scenarios that are relevant right now for a given topic.

The results from the data collection in all cases are unstructured text data, consisting of nouns, adjectives, verbs, names, dates, Twitter hashtags, etc. which need to be processed. All these studies use common methods of pre-processing such as stop-word removal which removes the most common words in the English language to avoid focusing on things like ‘and’ or ‘the’, lemmatization which groups inflected forms of the same word, and word vectorization which turns the words to numbered vectors that can be used by ML. All the studies then use automation to identify constructs (elements)
of interest to the scenario study, such as scenario driving forces (economy, technology, social trends, etc.), which is done through Topic Modeling. Topic Modeling is a form of unsupervised Machine Learning (ML) that can take all this unstructured data and identify themes based on the words in the documents. The most common method used in topic modeling is Latent Dirichlet Allocation (LDA), where words are clustered together based on how they are used in the text. Kayser and Shala (2020) automate concept mapping at this point in the process using commercial software. Because the software is a black box they use LDA to supplement the findings where all the other studies use LDA as the primary concept identification automation. Feblowitz et al. (2021) use commercial APIs that collect and aggregate data from newsfeeds.

The next step in the process across all studies diverges from automation to analyze the findings from the topic identification. (Kwon, Kim, and Park 2017) simply go directly to scenario analysis, and (Kwon and Park 2018; Kwon, Kim, and Park 2017) manually review results to label concepts. (Kayser and Shala 2020) supplement the automation by undergoing desk research based on their results. (Gokhberg et al. 2020) uses algorithms to classify terms but uses a manual review process to aggregate and verify their results whereas (Feblowitz et al. 2021) uses NLP to find connections between concepts and define causality.

From this point, only one study continues to use automation to hone their scenarios. (Kwon and Park 2018; Gokhberg et al. 2020) use the results from the previous step to manually create scenarios and (Kayser and Shala 2020) manually build a
morphology matrix, which is a diagram showing terms and relations, as a basis for their scenarios. Only (J. Kim et al. 2016) continues processing text, firstly by using Fuzzy Association Rule Mining (FARM) analysis, a fuzzy method typically used in shopping basket analysis, to find causality between concepts, and secondly by constructing an FCM to run simulations for scenario generation. Feblowitz et al. (2021) create mind maps but do not go as far as to create FCM from them.

(Kayser and Shala 2020) use significant visualization during the process, and rely on their morphology matrix to paint a visual story of their results before outputting narratives in plain text at the end of the method. (J. Kim et al. 2016) is the only method that visualizes the output once the scenarios are built. This is because using FCM creates a visual quantitative model that is used to generate the final scenarios before writing them out as a narrative. All other studies use the analysis from previous steps to write out scenarios in a narrative form. Feblowitz et al. (2021) use a custom-built web interface to show the causality between topics.

In a way, SPA (Feblowitz et al. 2021) is fundamentally different than the rest of the studies in this review because where the other studies create scenarios for strategic direction, SPA creates scenarios to find drivers of potential risks. In addition, SPA never translates their scenarios into narrative form, but rather creates a graphical interface where users can see the risks and their driving forces to help them potentially make better decisions.
3.3 Concluding Discussion

Scenario planning has been shown to help people remove blinders, reduce bias, and make better strategic decisions. In this chapter, I conducted three different systematic reviews that showed that (1) scenarios are not used in agile development, (2) FCM is used for scenarios in the context of product development but is predominantly focused on early development stages rather than for ongoing product development and (3) product teams create better products when incorporating knowledge through FCM and (4) agile teams suffer from many limitations that scenario techniques can alleviate. Among existing scenario development techniques, only a handful of studies utilize NLP, and of those that exist, there are few similarities. This all points to the potential for new research that can speed up scenario development so that scenarios can be incorporated throughout the product development lifecycle and in agile teams to improve product success.

The existing methods that use text mining to help in the scenario process use different methods, sources, and a different mix of automation and manual work. In all of them, automation was used for data collection and aggregation and manual intervention was used to make sense of the results, add additional labels and information, and verify the work of the automation. Based on this we can conclude that text mining holds potential for speeding up the scenario process but creating scenarios still needs a fair bit of human interpretation and manual effort. The “human in the loop” is needed to distill results, make final interpretations, and convert models and text.
clusters into the narrative forms that are typically used in scenario planning. After all, it is the narrative form of scenarios that speak to us as humans; stories capture our imagination, help us better absorb information, and help us frame potential futures and how to manage them. Defining causality is also still a manual task in all studies, not surprisingly as causal detection is still an active area of research in the ML community. TM may therefore never completely automate scenarios but can help create them much faster and overcome the main bottleneck in the scenario planning process, namely the need for experts and lengthy research. Different data sources were used across all the studies that used TM. Each has well-defined reasons for choosing its sources, but we cannot currently determine if particular data sources are better suited for different types of scenarios. Moreover, the current processes are far from seamless: only one of these methods generates a quantitative model that can be used to generate scenarios where all the others rely on people to build them once the data has been gathered and processed. The reliance on people for the final scenario steps also makes it difficult to review and update scenarios on an ongoing basis. Accordingly, all the studies create scenarios that look to the future of industry which makes them good candidates for creating strategic scenarios but does not show the ability to use them for product development. Moreover, none of the studies provide guidance or process to integrate them into a product development lifecycle. Of the requirements detailed earlier in this chapter in Table 2, there are still several open requirements that need to be explored before scenario planning can be incorporated into continuously developed product lifecycles. For the first requirement, there is still an open question of how non-expert
scenario planners can interpret the results of scenarios. Perhaps FCM as quantitative visual models can be used to bridge this gap, but more research needs to be done to demonstrate this. For the second requirement, TM can reflect a diversity of knowledge across a broad set of data sources, however, which data sources create the best scenarios for product teams remains an open question. Thirdly none of these studies go through a process of updating scenarios once they are built. Perhaps by creating FCM with data-driven methods this could be possible but is an area that does not show up in the literature. Fourthly TM can process large amounts of data potentially containing weak signals as shown in this review. Finally, because there is a lack of research in how to incorporate scenarios into the product development process understanding how to align with the broader organizational strategy while retaining focus on product development is an area that would benefit from further research.

In conclusion, automation methods have been demonstrated to speed up the scenario process by processing large amounts of text data, but the ability to use it in rapidly moving product teams has yet to be demonstrated. FCM shows promise to address some of the gaps and TM can help with research and data gathering. Scenarios have the potential to address issues that have been documented in autonomous development teams but to translate the benefits of scenarios to ongoing product development more research is needed to understand how to fill the gaps. In the next chapter, I highlight these gaps, align them to my research objectives, and detail the questions this thesis aims to answer.
Research Gaps, Objective, and Questions

Figure 18 outlines the overview of this dissertation, showing the gaps in the literature, the objectives of my research, and the specific questions I will set out to answer.
In the chapters above, I have identified an overarching research gap, namely the inability of existing scenario planning techniques to support agile product development teams in evaluating future change and maintaining a strategic perspective. This problem is owed to two main limitations of existing techniques:

Existing scenario planning techniques require high levels of expertise in the diverse knowledge fields that are to be represented in the scenario, as well as in the practice of scenario generation and interpretation. For the most part, they integrate this expertise in face-to-face interaction of experts and stakeholders during workshops. **Because of limited expert availability, time, resources, and the expertise needed to interpret and use scenarios, current techniques for scenario planning are inaccessible for most product development teams (Gap 1).**

Existing scenario planning techniques are designed for a long-term strategic horizon and, accordingly, are only used once, at the onset of a new product development project. They do not integrate approaches for monitoring and scanning the business environment or for reassessing and adapting scenarios in the light of the large amounts of newly available information on an almost daily basis, that characterizes dynamic business environments. **Existing scenario techniques are poorly adapted to the needs of agile teams because they do not provide solutions for updating, so resulting scenarios can easily become outdated and irrelevant to decision making (Gap 2).**
To close these gaps, my objective is to develop and evaluate Scenario Acceleration through Automated Modelling (SAAM), a method and supporting software system for scenario planning that supports product planning decisions. In section 3.3, I have identified high-level requirements for this approach (see Table 2) and in section 3.4, I have selected its underlying technologies - FCM modeling and simulation and IML.

To achieve this objective, I plan to answer the following research questions:

1. How can NLP be used to identify scenario drivers and their interdependencies from diverse knowledge sources using a Q&A method?

2. How can scenario drivers identified in RQ1 be modeled in an FCM and used to generate alternative futures that are plausible, decision-relevant, and cover the range of uncertainty?

3. How can information about possible changes to scenario drivers – so-called weak signals - be collected through text-based automation and used to update scenarios?

Questions 1-3 pertain to the design of SAAM, i.e., in the language of DSR, the creation of the artifact. I will introduce my plans for creating SAAM in chapter 6 and comment on how I will address these questions. Chapter 7 contains a detailed verification of SAAM through three different experiments, each testing the system from a different perspective.
5 SAAM: Scenario Acceleration through Automated Modelling – A Conceptual System

_I plan to submit this chapter as a paper to a journal on Future Studies. As a self-contained manuscript, it includes some of the background that was covered in earlier chapters._

5.1 Motivation

The goal of this work is to introduce a conceptual system. Scenario planning helps product teams to think strategically and proactively address changes in the business environment, resulting in better products (Postma, Broekhuizen, and van den Bosch 2012; Açikgöz, Latham, and Acikgoz 2020). The scenario process, however, is time-consuming and requires expert resources (Schoemaker 2020; Hodgkinson et al. 2006; Franco, Meadows, and Armstrong 2013), which is why product scenarios are only developed at the beginning of long-running projects of strategic importance (Randt 2015; Derbyshire and Giovannetti 2017). Moreover, they are rarely revisited and updated after the initial product planning stage (Mietzner and Reger 2005; Phaal, Farrukh, and Probert 2005), even though business environments are changing quite rapidly. As a result, agile product teams have little or no scenario-based guidance, despite having more freedom to define the product during project execution and to respond to changes in the business environment than traditional teams. These teams could greatly benefit from lightweight, automated methods for scenario construction
and updates that they can use during product development, rather than having to rely on massive, one-time efforts by strategic planning teams.

There have been scattered “big data” attempts in the literature to speed up scenario creation, using various combinations of online documents as a mechanism for employing the “wisdom of the crowd” for horizon scanning, Text Mining (TM) to distill quantitative models as Fuzzy Cognitive Maps (FCM), and FCM simulation (Kwon, Kim, and Park 2017; J. Kim et al. 2016; Kwon and Park 2018; Kayser and Shala 2020; Gokhberg et al. 2020). Quantitative methods can help accelerate the scenario process by creating models that can be run many times with different inputs to better understand the most relevant possible scenarios (Amer, Daim, and Jetter 2013). In particular, FCM is an emerging method used in scenario planning that uses qualitative data to build quantitative models (Amer, Jetter, and Daim 2011; Alibage and Jetter 2021; P. Zhang and Jetter 2018) and can be built using automated or manual methods (Son, Kim, and Kim 2019; Pillutla and Giabbanelli 2019; Alibage and Jetter 2021).

However, though these studies speed up the scenario building process they do not attempt to integrate these scenarios as a tool for product development teams. Furthermore, though these studies integrate automation and human expertise in the form of scenario planning experts, none of them define a socio-technical system that can be reproduced, scaled up, and continuously integrated into product development teams or any other fast-moving team that needs to think strategically. In response, this research defines a socio-technical system for Scenario Acceleration through Automated
Modelling (SAAM) that combines human expertise with AI-based automation to bring the benefits of scenarios to any team that needs to broaden their perspective and think more strategically.

Through the design of SAAM, I will explore three overarching questions about how “big data” automation (specifically NLP) and scenario planning can be integrated. Firstly, how can NLP be used to identify scenario drivers and their interdependencies from diverse knowledge sources using a Q&A method? To be useful, SAAM needs to identify relevant scenario drivers, exclude irrelevant ones, and not miss important driving forces. To achieve this, SAAM should be able to handle diverse knowledge sources. Moreover, because business environments are complex systems in which scenario drivers are interdependent, SAAM needs to represent connections between meaningful drivers, in particular causal relationships, rather than simply linguistic proximity. Secondly, how can scenario drivers identified in RQ1 be modeled in an FCM and used to generate alternative futures that are plausible, decision-relevant, and cover the range of uncertainty? After the first step of identifying the driving forces the next step is to represent their causal connections in a quantitative model, based on FCM, to “run” simulations to gain insights into possible future system states. This model should be able to create diverse scenarios that cover alternative futures that are each plausible and that, together, cover a wide range of uncertainty. Third and finally, how can information about possible changes to scenario drivers – so-called weak signals - be collected through text-based automation and used to update scenarios? With
quantitative models built that can generate a wide range of scenarios, the final step is helping the users of the model to select the most relevant ones. SAAM will need a way to recognize weak signals in large amounts of text if weak signals are embedded in the text at all. The design of SAAM will address these questions by incorporating elements of Big Data, Natural Language Processing (NLP), Fuzzy Cognitive Maps (FCM), Horizon Scanning (HS), and Interactive Machine Learning (IML).

The goal of this thesis is to bring these innovations together: Rather than relying on experts throughout the entire scenario process, big data could be analyzed with NLP. Specifically, NLP can handle the initial research phase by doing two important things: (1) to identify scenario drivers that are important for how the future unfolds and inherently uncertain and (2) to determine the relationships between these drivers to build a quantitative simulation model. During this process, IML can integrate human participation to ensure quantitative models that represent what is known about the business environments in ways that are understandable by decision-makers. The use of quantitative models, which can simulate diverse outcomes in response to input variations, further enables the identification of scenarios that cover the range of existing uncertainty. Finally, HS can be used to determine what combination of input values should be fed into the simulation model to generate scenarios. This occurs by identifying signals in contemporary texts that point to a change of the strength of scenario drivers and that should prompt a reevaluation of existing scenarios by re-running the simulation model. This socio-technical system could integrate scenarios into the fabric of fast-
moving product development teams, allowing them to gain the cognitive benefits of scenarios that have been well documented without having to slow down for planning.

The typical scenario creation process with areas that can benefit from automation is outlined in Figure 19 below.

![Figure 19 - Aspects of the scenario process that can benefit from automation, Figure 3 revisited](image)

The overall flow will be defined as IML, where people and automation interact to get to the end results. People will start the flow by defining a topic and a set of questions. Automation will take over, collecting data and using Q&A NLP to find the concepts that represent driving forces and the causal relationships between them, however, human experts influence this process by defining filtering criteria and informing the creation of an FCM model. Human experts also determine how they want to use the FCM by determining what simulations to run. To keep scenario models up to date one possibility is the practice of HS, where large amounts of data are scanned and
processed to find weak or emerging signals (Yoon 2012; Gokhberg et al. 2020; Konnola et al. 2012). HS categorizes different terms based on how often they are found and how many sources they are found in to help people understand if they are weak signals, growing significantly, or are stable areas. SAAM will incorporate HS to scan real-time data sources for concepts in the model to associate them with a signal strength, informing the model users how model inputs are changing in real-time. This system is described in detail throughout this chapter and the data flow is visually represented in Figure 20.
Figure 20 - Each step in SAAM shows where people are in the loop
A defining characteristic of SAAM is that it follows the goals and practices of interactive machine learning (IML). All defining characteristics of SAAM – Scenario Planning, NLP automation, FCM modeling, Horizon Scanning, and IML will be briefly explained in the subsequent section.

5.2 Foundations: Scenario Planning and IML

SAAM, in essence, constitutes a novel scenario planning approach that combines multiple techniques into a workflow that follows, wherever possible, the principles of IML.

5.2.1 Scenario Planning

**Scenarios** identify potential futures in story form to broaden minds and break biases (Bowman et al. 2013; van Notten 2006). Scenarios have been shown to have many benefits, such as raising awareness for dynamics of the business environment, managing uncertainty, accelerating organizational learning, evaluating different strategies, creatively defining alternative paths for reaching future goals, creating better products, and breaking away from groupthink (Postma, Broekhuizen, and van den Bosch 2012; Derbyshire and Giovannetti 2017; Collier et al. 2018; Relich et al. 2020; Açıkgöz, Latham, and Acikgoz 2020; Meissner and Wulf 2013). Scenarios are increasingly popular, however, there is no standard method in creating them (Tiberius, Siglow, and Sendra-García 2020). Typically, many people are involved in the scenario process, from experts who help do research and facilitate workshops to stakeholders in organizations who use scenarios to inform strategy, to new product development teams who need to
understand initial direction before beginning work. Though there are many variants of scenario planning, their process, in general, follows four main stages. First, is the preparation needed to create scenarios that involve setting the scope of the scenario, defining the domain to be modeled, research in that domain, and engaging stakeholders to set up participatory modeling. Secondly, experts work with stakeholders to create and evaluate the models, often through workshops where the research from the previous step is analyzed. Third a handful of diverse scenarios are selected based on their plausibility and internal consistency (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007). Finally, once models are built and understood, the most relevant scenarios from the model are selected to help stakeholders inform their strategic direction (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Durance and Godet 2010).

These processes take a lot of work and time and as a result, some researchers have tried to automate the up-front research using big data and text-based automation (i.e., forms of machine learning), specifically Natural Language Processing (NLP), to identify scenario drivers and their interdependencies (Kwon, Kim, and Park 2017; J. Kim et al. 2016; Kwon and Park 2018; Kayser and Shala 2020; Gokhberg et al. 2020). In these studies, people are involved in the process to add context to automated research and communicate the results. Outside of the field of scenario planning, such techniques are called Interactive Machine Learning (IML) and are being used to define workflows that
create models while keeping people in the loop by occasionally asking for help in situations where ML still struggles (Amershi et al. 2014).

5.2.2 Interactive Machine Learning (IML)

IML has been shown to create better models than ML alone while at the same time giving people a better understanding of the process being modeled (Robert et al. 2016). Using IML could speed up the scenario process by integrating advances in automation with human participation to create more transparent quantitative scenario models without the workshops. One benefit of creating scenarios from workshops is the participation of the scenario team which provides domain knowledge to the model and facilitates communication and knowledge sharing between people. By completely automating the scenario process we would be endangering that benefit. However, if we were to integrate people into the automated scenario process they could still have input and rely on automation to do most of the heavy lifting. In the field of ML, there is a practice known as Interactive Machine Learning (IML) where people are used as part of the training process to create better models (Fails and Olsen Jr 2003; Amershi et al. 2014). IML attempts to use people in the training process for ML with the hope that they can fill the gaps in data used for training or help create better models faster. This is different than traditional ML modeling because there is not always an expert in data science in the middle analyzing data or tweaking models, instead, people of varying technical ability and in some cases the target users of the system train the model before it is put into production. IML has been used with people that have non-technical
backgrounds to help train models (Fails and Olsen Jr 2003) and is used in a range of applications today such as image recognition (Sanghoon Lee et al. 2020) or sentiment analysis (Wu, Weld, and Heer 2019) to name a few. In a typical IML process, a person is used to dynamically tag data which is then used for training. The interfaces used for IML are customized based on the nature of the algorithm being trained. For example, in an application where IML is being used to improve sentiment analysis, a user may be presented with a screen that has a phrase and asks them to select what they think the sentiment is. IML used to help categorize phrases may show several different utterances and ask the person to categorize them or select keywords. Because humans are so important for the process, a recent IML review focuses on how to design user interfaces to train ML models and determines key usability aspects including engaging the user, exploiting interactivity, effectively representing data, and supporting user understanding (Dudley and Kristensson 2018). An example of an implementation of these principles is the interface for Microsoft’s Language Understanding Service (LUIS) shown below in Figure 21, where people tag words and categorize sentences that are used to train a model. The interface is designed in such a way that non-technical people should be able to understand it and use it to train the specifics of their model whereas ML uses Big Data in the background to do the training. LUIS is designed to identify custom elements from text people enter into a chat; in this example, the trainer of the model enters the text they expect the chat to receive, then labels the elements of the text they want the AI to recognize and make available for downstream programs. Figure 23 shows a user training a model that will recognize different ways someone can ask a
chatbot to book a flight and what the keywords are they need to work with such as where they are leaving from, where they are going to, and when.

Figure 21 - Example IML interface, using Microsoft’s LUIS to create a language model

By adding a human to help train these algorithms models can be trained faster with shorter learning iterations, and the people that interact with the models have a better overall understanding of them (Robert et al. 2016). Because data science experts are not needed to do all the data analysis, modeling, programming, and deployment IML allows faster learning through more rapid and frequent iterations on the model with the user of the model. Additionally, it has been shown that the best way to harness AI is to combine it with human interaction (Dellermann et al. 2019), and adding a person in the
loop increases transparency in models which leads people to trust them more (Kulesza et al. 2012).

5.2.2.1 A Demonstration of IML with Text Mining

In this work, IML is used so people can define parameters that help SAAM reach the most relevant model. In TM people are frequently in the loop; where TM can identify what words are often used together and group them into topics, a person will label or tag these groups of concepts using their judgment. An example output of a TM process is shown in figure 24, where I used text from several books about retail to identify topic groups throughout the texts. In this case, I (the person in the loop) had to view the results of the TM process and update parameters to get to the most relevant model, then tag the outputs of the model to give each topic group a label. Using TM one possible way to do communicate the different grouped outputs is to present a user with a word cloud and ask them to give a one-word response to define the overarching concept that can serve as a header to all the words in the cloud. Example word clouds generated from this TM process are shown in Figure 22 with the original topics. In this case one could label Topic 6 as “team-building”, Topic 18 as “mobile payments”, and Topic 28 as “marketing”. Though none of the labels appear in the text the person in the loop adds a layer of knowledge TM alone cannot identify.
Based on their knowledge of the subject matter and context, humans can choose to interpret Topic 6 as representing team improvement, Topic 18 could represent mobile banking, and Topic 28 could represent marketing. In an IML workflow, a person would be presented with visualization and asked to give them names or tags that are most meaningful to them.

Defining workflows with IML helps us minimize the shortcomings of ML alone by using people to help build models. Defining a workflow that combines human interaction with ML SAAM could create better models than ML alone while increasing the transparency and trust in the models that people interact with.

5.2.2.2 How SAAM Uses IML

IML represents the workflow of how people and AI can be integrated to create better models. SAAM incorporates three overarching steps in an IML workflow: setup, model building, and model use. In the setup stage, people define the domain, data sources, and the questions to start with before handing off to automation. The automation collects the data, runs Q&A algorithms to find traceable answers from the
text, and builds the initial FCM model. People can inspect the answers, define filters, and potentially ask more questions to build out the model further. Once the model is fully built people can craft scenarios as they see fit using the model and guided by the HS capabilities of SAAM to help identify additional plausible inputs to the model based on signals in current events. This interplay of automation, human interaction, and AI helps create better models than people or AI alone and brings people along through the process to create more explainable models. Figure 23 outlines a system blueprint indicating the data flow, points in the system that can be modified or extended, and where automation or manual work is used.
**Model Use:**
Monitor environment, run simulations

**Model Building:**
Ask questions, filter answers

**Setup:**
Get data, define questions

- People define domain, start process
- People define the questions for the Q&A
- Formulate questions
- Scrape data from knowledge bases
- Automation scrapes data from specified domain
- Automation scrapes data from specified domain

- Semi-structured text
- Q&A is run, collecting causal pairs. Follow up questions can be asked based on the size of the text, breadth of the model needed
- Causal pairs are used to construct model as FCM

- Build model

- Recommendations from HS are used to inform people on plausible scenarios
- Horison Scanning: Collect current data
- Simulations & Scenarios
- Recommendations

**Data sources:**

- Automation
- Person/Team

Figure 23 – IMI System blueprint of SAAM
5.2.3 Big Data in Futures Studies

In 2020 it has been estimated that there are already 44 zettabytes of data that come from the increased use of digital experiences such as Facebook, Twitter, Google, and YouTube and by 2025 it is estimated that over 460 exabytes will be created every day from online activity (“How Much Data Is Created Every Day? [27 Powerful Stats]” 2020). Undoubtedly, this data can provide some insights about what people work on and care about and how opinions, values, and preferences shift. In other words, this data can provide insights for understanding the future and thus support the practice of scenario planning.

Scenario planning falls within the domain of Futures Studies, a field that has been using large, structured data sets such as patent data for decades to objectively understand the present with the intent of better understanding the future. Future Studies is increasingly shifting focus to less structured data sources such as websites, news posts, and academic journals to look for signals that could point to future technology or emerging trends (Berg, Wustmans, and Bröring 2019; Jeong, Park, and Yoon 2019; Kose and Sakata 2019; Robinson, Lagnau, and Boon 2019). In doing so, Futures Studies use a variety of document types and text-based processing methods to identify, among others, current trends in a domain or convergence of topics. Table 10 shows an overview of recent studies that primarily relied on a text-based data source, and lists their purpose, data source, and text analysis methods.
Table 10 - Futures Studies using text analysis and big data

<table>
<thead>
<tr>
<th>Authors</th>
<th>Data Type</th>
<th>Purpose</th>
<th>Text Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Kose and Sakata 2019)</td>
<td>Web of Science</td>
<td>Technology Convergence</td>
<td>Clustering</td>
</tr>
<tr>
<td>(Berg, Wustmans, and Bröring 2019)</td>
<td>Patents</td>
<td>Emerging tech dominance</td>
<td></td>
</tr>
<tr>
<td>(Jeong, Park, and Yoon 2019)</td>
<td>Patents</td>
<td>Identify emerging R&amp;BD areas</td>
<td>LDA, topic analysis</td>
</tr>
<tr>
<td>(Li et al. 2019)</td>
<td>Patents, Twitter</td>
<td>Identify emerging technology</td>
<td>(1) semantic hierarchical clustering (SHOC), (2) vector space model (VSM), (3) singular value decomposition (SVD)</td>
</tr>
<tr>
<td>(Y. Zhang et al. 2019)</td>
<td>Web of Science</td>
<td>Discover and forecast interactions</td>
<td>Clustering, bibliometric</td>
</tr>
<tr>
<td>(Wang et al. 2019)</td>
<td>Patents</td>
<td>Identify emerging tech convergence</td>
<td></td>
</tr>
<tr>
<td>(Moehrle and Caferoglu 2019)</td>
<td>Patents</td>
<td>Identify emerging technology speciation</td>
<td>Bi-grams</td>
</tr>
</tbody>
</table>

From these studies, we can see different types of data are used for different applications. For example, Twitter is used to gauge customer expectations because it is a data source that comes directly from customers (Li et al. 2019), where patents and Web of Science (WoS) data are used to identify emerging trends in the field. After all, they are looking at new inventions and knowledge (Moehrle and Caferoglu 2019; Wang et al. 2019; Jeong, Park, and Yoon 2019; Berg, Wustmans, and Bröring 2019). Though multiple
data sources are used in various studies and each study appears to use data that makes the most sense for their context, Futures Studies lacks prescriptive recommendations on what type of dataset to use in what context. It seems to be up to the author of a study to determine what the best type of data is for their research.

5.2.4 Modeling Scenarios with Fuzzy Cognitive Maps

A method that has been demonstrated to create models that represent scenarios is Fuzzy Cognitive Mapping (Amer, Jetter, and Daim 2011; A. Jetter and Schweinfort 2011). Fuzzy Cognitive Maps (FCM) are a popular decision support tool that has been used in a variety of applications (C. W. Davis and Jetter 2021) because they are intuitive and relatively simple to use with significant predictive power (Elpiniki I. Papageorgiou and Salmeron 2013; A. J. Jetter and Kok 2014; Glykas 2010; Felix Benjamín et al. 2019). They are also used as models in scenario planning thanks to their user-friendly nature ability to run what-if simulations and their ability to run quantitative simulations from models built with qualitative data (Amer, Daim, and Jetter 2013). FCM can be created through workshops (A. Jetter 2006; A. J. Jetter and Kok 2014) and there are methods where FCM creation is either automated or learning is assisted through ML (Felix Benjamín et al. 2019).

FCMs are transparent and have interactive qualities because the concepts or nodes that they are composed of representing constructs that have causal relationships such as “studying” and “grades” and the edges that connect them are described with linguistic variables such as “high” or “low”. These concepts typically have fuzzy values in
people’s minds, for example, “studying a lot to get good grades” will have a slightly different meaning depending on who you talk to. This fuzziness makes FCM attractive as a method for scenarios because it allows people to frame concepts and causality in systems as they think of them naturally.

Flexibility and transparency set FCM apart from other ML methods that create black boxes which are impossible for people to understand (Hagras 2018). Transparency in modeling is important for adoption because people are more likely to utilize models they can understand and modify (Dietvorst, Simmons, and Massey 2018; Cai et al. 2019) and less likely to use models if they do not understand how they work even if they are more accurate than other methods, a phenomenon known as algorithm aversion (Dietvorst, Simmons, and Massey 2015). FCM provides system descriptions, consisting of interconnected elements (i.e. nodes), which affect each other element (i.e. directed edges), and the strength of these effects are described using linguistic variables such as ‘strong’, ‘very strong’, ‘weak’, ‘medium’, etc. as demonstrated in Figure 24.

**Figure 24 - An Example FCM**
When the model is run as a simulation the value of each concept $A_i$ is updated according to its relations to other concepts. The FCM formula has evolved over time, but the one more commonly used originated from Kosko (1986) (Felix Benjamín et al. 2019) is demonstrated in Formula 4.

$$A_i^{(k+1)} = f \left( \sum_{j=1, j \neq i}^{n} w_{ij} A_j^{(k)} \right)$$

_Formula 4 – FCM calculation formula._

Where $k$ is the iteration, $n$ is the number of concepts, and $f$ is the squashing function which keeps the resulting value in the range of $[-1,1]$. There are different options for the squashing function, this conceptual system proposes the use of the hyperbolic tangent function demonstrated in Formula 5.

$$\frac{e^x - e^{-x}}{e^x + e^{-x}}$$

_Formula 5 – Hyperbolic Tangent (TANH)_

FCM is a powerful modeling method used today to create transparent models for scenarios that can be created with automation. SAAM will inherit these qualities by using FCM as a quantitative method for modeling scenarios with data scraped from diverse online sources.
5.2.5 Building Fuzzy Cognitive Maps with Big Data and NLP

History is known to be a key informer of scenario planning (Bradfield, Derbyshire, and Wright 2016), and there are plenty of readily available perspectives in the text available in the form of books, journals, and other data sources. These data sources are pre-aggregated and peer-reviewed and are common sources for the desk research used in scenario planning. By processing this data through NLP we can automate much of the desk research that is currently done in scenario planning by domain experts today, as demonstrated in several studies that were reviewed in Chapter 3 (Kwon, Kim, and Park 2017; J. Kim et al. 2016; Kayser and Shala 2020; Gokhberg et al. 2020). NLP identifies patterns in data and hint at future trends in ways that humans (without the help of computers) can only achieve at a large expense of time and effort or not at all. Text Mining (TM) and NLP can be used in concert to process large amounts of data in a way that can be used to create quantitative models used for scenarios:

TM is used to help find the key aspects that help us understand generally what large volumes of text are about. For example, TM can find topics, keywords, and patterns in text to give us an idea of what a series of documents is about, and to what degree of membership each topic has in each document.

NLP is a form of Artificial Intelligence (AI) that specializes in extracting more granular meaning from data. For example, NLP allows semantic analysis of large volumes of text to understand what topics are mentioned, key phrases in documents,
and draw connections across multiple texts known as corpora. This type of semantic analysis has been used to overcome the limitations of bibliometric analysis which will only link documents together based on how they reference each other (Li et al. 2019; Ranaei and Suominen 2017).

5.2.6 Recent Advances in NLP with BERT and Question and Answering Algorithms

The field of NLP is advancing rapidly thanks to pre-trained, so-called transformer models (a form of neural network) that can take different sequences of words into account, find patterns, and then make predictions on new text. When trained on large volumes of text these models can generalize patterns, then “fine-tune” them on much smaller amounts of text to find nuances in specific corpora and create better AI models. For example, a model might initially be trained on all of Wikipedia where it learns sentence structures and general knowledge, then could be fine-tuned on a much smaller document to find specific knowledge using what it has learned before. We can find parallels in this process to how people learn; the more we read the easier it is to comprehend new things thanks to what we have seen in previous texts. If we must read a specific document and test our comprehension of it, we rely on our ability to understand patterns in text we have learned before. The base transformer is like learning how to read, the fine-tuning is like reading and comprehending a specific topic.

Today the state-of-the-art transformer models are based on Bidirectional Encoder Representations from Transformers (BERT) which can perform better than other forms of AI on several NLP-based tasks including question-answering tasks (Q&A)
and natural language inference (NLI) (Devlin et al. 2019). BERT models are trained by first creating a base model on a large unstructured dataset that can make predictions such as what word might appear next in a sentence. Secondly, the previous learnings are transferred, and models are fine-tuned on specific datasets that allow such functionality as answering questions based on the text in the dataset. To achieve this, BERT uses multiple layers of encoding so it can predict context and “understand” the difference between semantically similar terms such as “apple pie” or “apple tree” by encoding (1) the words, (2) the sentences, and (3) the positions of the words in the text. This combination of tokens is then fed into a neural network that creates the base model which can be fine-tuned on specific text for NLP tasks. An example of BERT encoding is shown visually in Figure 25, where the three distinct layers are shown over two different sentences.

<table>
<thead>
<tr>
<th>Input</th>
<th>[cls]</th>
<th>I</th>
<th>love</th>
<th>pie</th>
<th>[sep]</th>
<th>Apple</th>
<th>is</th>
<th>the</th>
<th>best</th>
<th>[sep]</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Token Embeddings</th>
<th>E[cls]</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>E7</th>
<th>E8</th>
<th>E9</th>
<th>E10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment Embeddings</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
<td>EA</td>
</tr>
<tr>
<td>Position Embeddings</td>
<td>E0</td>
<td>E1</td>
<td>E2</td>
<td>E3</td>
<td>E4</td>
<td>E5</td>
<td>E6</td>
<td>E7</td>
<td>E8</td>
<td>E9</td>
<td></td>
</tr>
</tbody>
</table>

Figure 25 - Example BERT base encoding where Tokens represent the words, Segments represent the sentences, and Position represents the position of a word within a chunk of text

BERT can be used for several different tasks in NLP including predicting the next word or phrase in a sentence, summarizing text, translation, and Q&A. Because SAAM is a socio-technical system BERT is highly relevant as a method that can help people
interact with technology in a way that feels natural, such as through Q&A. Another advantage of BERT is the availability of off-the-shelf models built and available for use such as those published by Hugging Face\(^1\), which gives a modeler the ability to quickly swap out models or fine-tune their own if needed.

Q&A also gives us a natural way to determine connections and causality between concepts in the model. By asking the system, “what causes pollution to increase?” a human user identifies a concept of interest through the question – in this case, “pollution”. Q&A systems provide the answer by treating a pre-selected text corpus as the context. For example, when researching the causes of pollution in the fashion industry we would first collect data about the fashion industry and apply the Q&A engine to this material. If we then ask the question “what causes pollution?”, we receive results in the form of keywords that are relevant to the industry and that can be associated with other relevant concepts. We would not identify keywords that also relate to pollution but have no connection to the fashion industry and are therefore not included in the texts – pollutions ‘words’, such radioactive radiation would likely not be identified.

To demonstrate Q&A output I scraped data from several online books about the fashion supply chain and asked the question, “what causes pollution?” using a Q&A BERT-based model from the Hugging Face project. A sample of the output is shown in Table 11 where the answer could represent the corresponding concept of an FCM,

\(^1\) https://huggingface.co/models
confidence is the degree of certainty that the algorithm identified a matching answer, and context contains an excerpt from the most relevant document containing the answer. Note that in this case “fast fashion brands” is returned with high confidence because it is directly referenced in the text as a cause of pollution, where very low confidence was returned for the other concepts because they are mentioned together but do not answer the question based on the text provided. The more text that associates fast fashion brands with pollution the higher the confidence value would be. The ability to view the context also gives a modeler a direct reference back to the text where the answer is found. Through this context, a modeler can potentially identify more relevant concepts that can be used for further questions. For example, if the modeler is interested in understanding pollution in the fashion industry, noting that “sustainable development” is mentioned in the context of an answer (the second answer in Table 11) gives them another potential line of questioning, such as “what types of sustainable development are happening in the fashion industry?” or “what are the benefits of sustainable development?”.

Table 11 - Example output of NLP model given the question, "what causes pollution to increase?"

<table>
<thead>
<tr>
<th>Answer</th>
<th>Confidence</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast fashion brands</td>
<td>0.489</td>
<td>on the other hand, fast fashion brands such as h &amp; m, Zara, Topshop, have been blamed for creating poor labor welfare, severe environmental pollution as well as a massive amount of clothing disposal at the end of the product life cycle.</td>
</tr>
</tbody>
</table>
Global climate change introduction due to the aggravation of environmental pollution and global climate change, sustainable development has attracted more and more attention.

Overconsumption of energy by doing so, these companies alleviate conflicts of interest among participants and reduce pollution and overconsumption of energy.

As the discussion above demonstrates developing Q&A-based engines using BERT for finding connections in text holds promise as a transparent method for building quantitative models. For this reason, this thesis uses Q&A-based NLP as an alternative and updated method to create quantitative models through the natural interaction of asking questions of a collected corpus.

5.2.7 Horizon Scanning Background

Where scenarios help people envision the potential future, Horizon Scanning attempts to help people to respond to the potential future by understanding what is happening today (Sutherland and Woodroof 2009). The big idea behind Horizon Scanning is to help people make sense of what might happen by scanning through data to find “signals” and map those signals over time to understand how they are trending and thus determine how to set policy or make decisions (Konnola et al. 2012). One study categorizes the different types of signals that can be found and applied to scenario development based on how fast they grow over time and how frequently they occur (Gokhberg et al. 2020). This categorization is shown in Figure 26.
It has been shown that combining horizon scanning with scenarios could create an integrated framework to monitor the environment and trigger adjustments to operations as needed with the help of expert integration (Schoemaker, Day, and Snyder 2013). Current data sources in horizon scanning can come from a variety of places including previously aggregated data from large websites such as Google or other sources, raw text data from social websites such as Twitter, academic journals, and patent databases (Kose and Sakata 2019; Robinson, Lagnau, and Boon 2019).

Horizon scanning has been used to find weak signals using TM to find relevant business opportunities using Degree of Visibility (DoV) and Degree of Diffusion (DoD) (Yoon 2012). DoV is used to indicate the number of times a term is found in a given set of sources over time, and DoD is used to indicate how many different sources mention a term over time. Together DoV and DoD help identify weak signals by looking for terms that are being mentioned more frequently (high DoV) but have not been noticed by
most sources (low DoD). Some studies add a third dimension to HS by giving signs a subjective or interpreted level of importance (Hiltunen 2008), though this is not always found in systems that attempt to automate HS.

The formulas for DoV and DoD are shown below. Keyword \( i \) in the period \( j \) can be defined where \( TF_{ij} \) is the total occurrence frequency of a keyword \( i \) in the period \( j \), \( DF_{ij} \) is the total number of sources keyword \( i \) is found in period \( j \), \( NN_j \) is the total number of news articles in the period \( j \), \( n \) is the number of periods, and \( tw \) is a time-weight. \( tw \) increases as the environment around a given technology c rapidly, while \( tw \) approaches 0 as the environment shows insignificant change (Yoon 2012).

\[
DoV_{ij} = \left( \frac{TF_{ij}}{NN_j} \right) \times \{1 - tw \times (n - j)\},
\]

\[
DoD_{ij} = \left( \frac{DF_{ij}}{NN_j} \right) \times \{1 - tw \times (n - j)\},
\]

*Formula 6 - Degree of Visibility and Degree of Diffusion*

5.3 An Overview of the Conceptual System

Big data, NLP, FCM, and HS combined hold the potential to automate quantitative scenarios in such a way that these scenarios can be used more broadly. SAAM is designed to bring these technologies together incorporating IML to increase accuracy, relevance, and trust in the system by using automation to make sense of big data while bringing people into the loop to increase understanding and transparency. This novel combination of technology brings helpful models to managers, leaders, or
teams that what to explore the future cone of uncertainty in any domain and incorporate the benefits of scenario planning.

At a high level there are three overarching phases to the system; in phase one people define the questions to ask and the data to use to build the model. In Phase 2 NLP is used to build an FCM-based model that represents the current understanding of a domain with the assistance of people in the loop. The final phase incorporates HS and current event data to guide people conducting scenario planning by showing them how the concepts of the model are changing in the present. These phases are demonstrated at a high level in Figure 27.

![Figure 27 - High-level view of three phases of SAAM; setup, model building, and model use](image)

The rest of this chapter uses an illustrative example using public data from the major American fashion retailer, Levi Strauss & Co. (commonly known as Levi’s), to demonstrate how SAAM can quantitatively model a domain based on narrative text and run simulations based on data collected through HS. Public companies in the US are required to file a report commonly referred to as the “10-K” detailing their financial
performance every year with the US Securities and Exchange Commission (SEC). These reports also include company history, risks and opportunities they perceive in the market, and their strategic outlook in a narrative form. For this demonstration, I used the filings from Levi’s from 2021, which are publicly available on the SEC website. I choose this data source to illustrate SAAM because it shows how a company thinks of itself in relation to its environment and how external factors are potentially shaping its future. This makes the data particularly suitable for creating scenarios, which aim to help teams to think more strategically by positioning their work and their organization in relation to plausible future developments in their business environment.

5.3.1 IML Flow - Integrating People and AI

5.3.2 Phase 1 – Setting Up; Collecting Data and Defining Questions

The first two things we need to create scenarios using SAAM are the data source we want to build the model from, and the questions to ask of that data. The data source should represent the domain the modeler wants to create scenarios for, in the case of this example we will attempt to get a better understanding of the external developments that impact competitiveness for Levi’s, a publicly traded American retailer, so we can use public data from the latest financial report to the SEC for that company. In other cases, the modeler may choose data from journals, newspaper articles, or websites that have detailed information for the target domain. As with other methods that use automation to speed up the scenario creation process the decision is up to the modeler to find the data source that best represents their domain.
Secondly, the modeler must determine the questions to ask to find the most important driving factors of potential scenarios. How the modeler goes about determining the questions to ask is up to them and should not require dependency on their computational modeling knowledge but should correspond to the text they are using and what domain they want to build scenarios for. For example, if the text they are using discusses obesity as a personal and public health problem and the modeler wants to understand causes and impacts better, they may ask “what causes obesity?” and “what does obesity cause”? If the modeler wants to better understand the future of self-driving vehicles, then they may ask questions about “self-driving”, “vehicles”, or “self-driving cars”. In other words, it is important to identify concepts (or “nodes” of an FCM) belonging to the domain, that is of interest to the study, and that investigate the causal relationships between them. To achieve the former, I use NLP to find the keywords of the document which characterize the domain before attempting to determine causality in a second step.

To identify the first set of nodes of the FCM I used NLP to identify keywords in the document by first removing common connecting words known as stop-words such as “and”, “the”, “a” etc. then calculating the frequency of the remaining words and how often they are used together. There are existing code libraries that can do this step for us, including RAKE\(^2\) and Gensim\(^3\) that can be used to return a list of keywords ranked by importance in relation to the text. After running such tools the most popular keyword

\(^2\) https://github.com/fabianvf/python-rake

\(^3\) https://radimrehurek.com/gensim/
by far was the word “stock”, which may not be surprising given we are looking at a financial report of a publicly-traded company. Therefore, I started simply with the questions “what causes stocked to increase?” and “what causes the stock to decrease?”

With my first set of questions, I used an NLP-based Question and Answer (Q&A) model to find connections between concepts. Specifically, in this example, I used Hugging Face Q&A pipelines\(^4\) but there are other similar open-source solutions that could be used such as Sentence Transformers\(^5\) or custom BERT-based solutions. Defining questions is a step that requires the modeler to be part of the process. In this example, we want to know what causes the identified terms to increase or decrease, but in other cases, a different line of questioning may be appropriate. For example, if you are building a model that indicates different drivers for a specific technology you may want to define questions that look at Political, Economic, Social, Technical, Environmental, and Legal (PESTEL) aspects of the technology. If you are trying to find cause and effect for a specific phenomenon you may start with two questions around it such as “what causes [phenomenon] to increase?” and “what causes [phenomenon] to decrease?”.

Allowing the modeler to ask questions as they see fit gives them flexibility and allows them to interact with the data naturally. For the sake of this conceptual system, I will not provide hard rules on how someone should define questions but leave it up to the modeler based on the corpus they are working with and the model they are building. This is an example of where IML can provide better models by allowing people to use

\(^4\) https://huggingface.co/docs/transformers/main_classes/pipelines
\(^5\) https://www.sbert.net
their creativity, knowledge, and intuition to direct the system as they see fit in defining questions.

The last option a modeler has is the Question Depth or the number of iterations of follow-up questions from the model. For example, after finding that $A$ causes $B$ to increase the modeler may also want to know what affects $B$ so could run another round of questions, this time asking what causes $B$ to increase and decrease. This would constitute a Question Depth of 1. Setting a higher depth could help build larger models or find more complex connections between concepts. Question depth is illustrated in Figure 28.

![Figure 28 – Example of question depth used by SAAM](image)

A modeler may choose a higher question depth if they only have a single question to start with, or if the corpus used is very large. After a certain number of iterations answers typically start to decrease in confidence because they reach the knowledge limits of the corpus. For example, a text about manufacturing may mention
product logistics, but it will likely not have a comprehensive list of causal factors for

good or bad product logistics.

In subsequent levels of questions depth in this example, I used the same
question type as the first questions, substituting “stock” with terms that were identified
in the first set of answers. In this example, I used the set of questions demonstrated in
Table 12 for each identified keyword to build up a list of cause-effect pairs where

[TERM] is the keyword. Based on the question I create a positive or negative weight that
is later used in creating relations in the FCM. In this case, when I ask what causes

[TERM] to increase, I assume there is a positive relationship between the term and the
answer, and when I ask what causes [TERM] to decrease there is a negative relation
between the term and the answer.

*Table 12 - Questions used in the illustrative use case for SAAM*

<table>
<thead>
<tr>
<th>#</th>
<th>Question</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What causes [TERM] to increase?</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>What causes [TERM] to decrease?</td>
<td>-1</td>
</tr>
</tbody>
</table>

5.3.3 Phase 2 – Building the Model

The values returned contain the answer, a confidence level between 0 and 1
indicating the probability that the model got the correct answer, and token markers
indicating where in the document the answer was found. For example, if a document
contains the sentence “Pollution is a direct cause of a lower standard of living.” and the
Q&A algorithm asks the question, “what causes lower standards of living?”, the model will return “pollution” as the answer, a high probability such as .89, and the beginning position in the document to where the answer was found. From these values, the answer and confidence score are directly relevant to assisting the modeler, and the token marker can be used to find the sentence and the document the answer was found in to give people using SAAM the full context of the answer. In this example, it is as if the model is saying “I am pretty sure that pollution is the answer because of this excerpt from the text you showed me”. Example output from the Q&A step is shown in Figure 29.

Figure 29 - Example output from SAAM’s Q&A that is used for creating an FCM

These results provide quick insight into the text without the modeler having to read it all. There are some obvious answers such as a “failure to compete effectively”
will hurt “sales”, but some not so obvious such as “wage rates” hurting “sales” or “changing market conditions” improving “company image”. If any of these seem counterintuitive the modeler can check the text the answer came from which ultimately gives them a deeper understanding of the model.

Before building the FCM the modeler has the option to define three separate filtering criteria to determine how many of the returned connections to use in their model. The first filter is based on the returned confidence from the model. A modeler may only keep connections that were returned with a high degree of confidence, filtering out results that do not have confidence over a certain threshold. The range of confidence in answers will depend on the Q&A model used and the corpus and thus should be determined by the modeler after reviewing the results. In this case, I chose a confidence threshold of .7 because most answers were clustered above that threshold.

The next filter criterion includes setting a value for the semantic similarity between concepts to combine any like terms when the model is built. There could be answers returned that are so semantically similar that the modeler only wants to keep one of them, for example, “customer demand” and “demand” are different answers but in context may mean the same thing. The semantic distance can be tested using Levenshtein or cosine distance to determine like terms. In this example, I tested Levenshtein distance using the fuzzywuzzy library in python where a threshold of 100 is

---

6 https://github.com/seatgeek/fuzzywuzzy
an exact match, and the closer to 0 the larger the distance between words. In this case, I used a filter value of 90 and no additional words were filtered.

The third filter criterion includes anything that is not a relevant part of speech (POS) that comes back as an answer. For example, sometimes an answer may be an adjective that by itself does not make sense. Using the Spacy\textsuperscript{7} library we can filter out POS that may not make sense to a model user, such as adjectives or interjections alone. The modeler can update the POS tags they wish to filter, in this case, I set the filter to remove adjectives, punctuation, particles, symbols, and interjections.

The three filter criteria are outlined in Table 13.

\textit{Table 13 - Answer filters modelers can set used by SAAM}

<table>
<thead>
<tr>
<th>Filter Criteria</th>
<th>Range</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence threshold</td>
<td>0 - 1</td>
<td>Filter results based on the confidence returned by the Q&amp;A algorithm. The threshold range will vary based on the context thus the cutoff should be up to the modeler.</td>
</tr>
<tr>
<td>Semantic threshold</td>
<td>0-100</td>
<td>Filter concepts based on how semantically similar they are where 100 is the same word. This can use Levenshtein or cosine distance. A lower value will group like concepts, a higher value could result in a model having different concepts with similar meanings.</td>
</tr>
<tr>
<td>POS filter</td>
<td>An array of POS tags</td>
<td>In some cases based on the wording of the context parts of speech may be returned as answers but would not make intuitive sense as concepts. In addition, a modeler may only want certain parts of speech as concepts, for example using only proper nouns. The modeler can filter</td>
</tr>
</tbody>
</table>

\textsuperscript{7}http://spacey.io
The answer, confidence level, and reference to the sentence and document the answer came from returned from Q&A NLP models give the modeler context needed to define their filters. We can also use visual tools such as a categorical diagram to view the range of confidence in answers and network diagrams to visualize the FCM to assist the modeler. One issue in the use of AI-based systems is the inability to easily interpret answers from the system or explain where answers came from (Gunning 2017). Using this method we create an explainable model by showing the modeler all of the connections and concepts in the model visually, allowing them to update the filters used in creating the model, and giving a trail from the concepts, back to the answers, to the sentence they came from, and ultimately the source with specific reference to where the answer was in the text. With SAAM, if someone questions why a connection exists in a model, they can trace the reason back to the source. Examples of these visualizations and answer traceability at Question Depth 0 are shown in Figure 30.

\[^{8}\text{https://universaldependencies.org/u/pos/}\]
Using the categorical diagram, we can see that answers with above 70% confidence are mostly negative, meaning that the corresponding node to the answer will hurt company stock. By viewing our network diagram, we can see that we have a simple model with a single node in the middle and several different nodes connecting to it. This is because we only asked SAAM questions about one thing – stock. To fill out our model a bit more we can use the answers we received from our first question depth to ask what causes them to increase or decrease. For example, since “changing market conditions” was identified as a factor we can ask “what causes changing market conditions to increase?” and “what causes changing market conditions to decrease?”. SAAM can do this for all the identified terms that come back giving us a richer model as
shown in Figure 31. With a question depth of 1 more complex relationships begin to develop; we can see that stock still has several connections, but other nodes such as fluctuations in product cost and change in market conditions also emerge as important concepts with several connections.
Figure 31 - Category diagram after 2 rounds of questions (top), corresponding network diagram (bottom) with a confidence filter of 70%.
Repeating the question cycle can be done as many times as the modeler sees fit. At a question depth of 2, our model starts to become even more complex, as shown in Figure 32.

![Figure 32 - FCM visualized after 3 rounds of questions](image)

Given the size of the text eventually, the number of answers returned within the confidence threshold will diminish. For example, using this text at a 70% confidence threshold we keep 47% of the answers from round 1, 38% of answers from round 2, and 28% from round 3. Because this single document has a specific scope at some point we will start making connections that are not relevant to our model. For example, at question depth 3 we have nodes such as 'transition from LIBOR', 'greenhouse gas emissions, and 'economic disruption' which are called out as risks but are not explained in detail. If the modeler wants to find additional data sources that relate to these concepts they could continue to ask questions of other data sources but in the case of
this example, we have accomplished our goal, which was to understand how this company sees itself in the context of the larger business environment. We now have an FCM that we can use to run scenario simulations.

5.3.4 Phase 3 – Using the Model

FCM simulations are run by setting input concepts to corresponding linguistic variables such as “high” or “low”. For example, if you study (“study” as an input set to high) then you will get good grades (“grades” as an output result as high) and if you do not study (“study” as an input set to low) you will not get good grades (“grades” as output results in low). Using the model we built we can create different potential scenarios by setting specific concept values in our model to high and see how it affects all of the other concept’s values where high would the value 1 and low would be the value -1. The full list of concepts identified after filters were applied in the example model include the following: 'Adverse publicity', 'Brexit', 'Failing to make timely deliveries', 'Fluctuations in product costs', 'Increases in market interest rates', 'Increases in raw material costs', 'Social media', 'changing market conditions', 'complexity', 'compliance costs', 'costs', 'decreased demand', 'demand', 'economic disruption', 'equity securities', 'future issuances of debt', 'global climate change', 'government', 'greenhouse gas emissions', 'increasing costs', 'interest rates', 'inventory shortages', 'lawsuits', 'loyalty', 'product development cycle lead times', 'product pricing actions', 'protection risks', 'quality problems', 'reduction in workforce', 'repurchase stock', 'seasonal', 'shifts in U.S. immigration policy', 'short-term or other non-controlling investors', 'shortages of
An example interface for setting input levels for an FCM in a user-friendly way is demonstrated in (C. W. H. Davis, Jetter, and Giabbanelli 2020), but for the sake of this example, I will assume the user of the system can change the inputs of any identified concept to high or low based on their judgment. This gives the user of the system the ability to run many different simulations and it is up to them to pick which combinations of inputs together make sense. For example, a scenario where there is high economic growth, high unemployment, and low consumer demand does not seem very plausible. Typically, in scenario planning, only a handful of scenarios are chosen to work with that are plausible, relatable, and do not contradict themselves (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007). To illustrate how this is done I have created 4 different scenarios that a management team might want to understand while creating a strategy. Because quantitative scenarios output quantitative values I have created narratives of the results in each summary based on the output values from the model. Full results from the simulations are shown below in Table 14.

In the first scenario, I set the concepts 'government', 'shifts in U.S. immigration policy', and 'state-imposed restrictions' to high to represent the government deciding to create strict regulations and policies, increasing the minimum wage, tightening immigration policy, and imposing restrictions on trade. I called the first scenario “Big
Brother’s Heavy Hand” to represent a possible future where the company has to navigate the repercussions of a heavy-handed government. This scenario results in a stable market with minimal economic disruption but increases costs and supply chain complexity with the ultimate result of decreasing the company stock price. In a scenario where there is a major climate disaster, I set ‘global climate change’, 'Increases in raw material costs' too high which resulted in decreased demand, higher product development lead time, an overstock of goods, and a lower stock value. In the third scenario called “Material Shortages”, I set the concepts 'Failing to make timely deliveries', 'Fluctuations in product costs', 'Increases in raw material costs', and 'inventory shortages' to high to reflect a potential future where creating products becomes more difficult. This resulted in a major increase in demand, a glowing public image (adverse publicity became low and social media ended at high), and a higher stock price. The final scenario called “Fashion Champion” represents a potential future where denim becomes extremely popular, the concepts 'Adverse publicity', 'decreased demand' is set to low and the concepts 'Social media', 'demand', 'short-term or other non-controlling investors' are set to high. In this case, even though all the factors seem to be in Levi’s favor the stock price decreases because of major fluctuations in the market.

In our Fashion Champion scenario, we saw a surprising result, where stock price decreased when nearly all indicators pointed to the contrary. This underscores the importance of a transparent model that can be traced back to the source. In this case, if
we look at the final network diagram in Figure 33 combined with the final results we can see there are complex relationships around changing market conditions that are difficult to predict and ultimately bring the stock price down. This should make us reflect on our view of the world given the model and how we would respond in such a situation or what other variables we might be able to control to get a more positive result on the stock price. This type of critical reflection is exactly what scenarios are intended to do, they help us challenge our assumptions, think through why things could happen, and help us determine the best strategy based on what we know today.
<table>
<thead>
<tr>
<th>Big Brother’s Heavy Hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Result</td>
</tr>
<tr>
<td>Climate Disaster</td>
</tr>
<tr>
<td>Final Result</td>
</tr>
<tr>
<td>Fashion Champion</td>
</tr>
<tr>
<td>Final Result</td>
</tr>
<tr>
<td>Material Shortages</td>
</tr>
<tr>
<td>Final Result</td>
</tr>
<tr>
<td>Adverse publicity</td>
</tr>
<tr>
<td>Brexit</td>
</tr>
<tr>
<td>Failing to make timely deliveries</td>
</tr>
<tr>
<td>Fluctuations in product costs</td>
</tr>
<tr>
<td>Increases in market interest rates</td>
</tr>
<tr>
<td>Increases in raw material costs</td>
</tr>
<tr>
<td>Social media</td>
</tr>
<tr>
<td>changing market conditions</td>
</tr>
<tr>
<td>complexity</td>
</tr>
<tr>
<td>compliance costs</td>
</tr>
<tr>
<td>costs</td>
</tr>
<tr>
<td>decreased demand</td>
</tr>
<tr>
<td>demand</td>
</tr>
<tr>
<td>economic disruption</td>
</tr>
<tr>
<td>equity securities</td>
</tr>
<tr>
<td>future issuances of debt</td>
</tr>
<tr>
<td>global climate change</td>
</tr>
<tr>
<td>government</td>
</tr>
<tr>
<td>greenhouse gas emissions</td>
</tr>
<tr>
<td>increasing costs</td>
</tr>
<tr>
<td>interest rates</td>
</tr>
<tr>
<td>inventory shortages</td>
</tr>
<tr>
<td>lawsuits</td>
</tr>
<tr>
<td>loyalty</td>
</tr>
<tr>
<td>product development cycle lead time</td>
</tr>
<tr>
<td>product pricing actions</td>
</tr>
<tr>
<td>protection risks</td>
</tr>
<tr>
<td>quality problems</td>
</tr>
<tr>
<td>reduction in workforce</td>
</tr>
<tr>
<td>repurchase stock</td>
</tr>
<tr>
<td>seasonal</td>
</tr>
<tr>
<td>shifts in U.S. immigration policy</td>
</tr>
<tr>
<td>short-term or other non-controlli</td>
</tr>
<tr>
<td>shortages of inventory</td>
</tr>
<tr>
<td>social distancing</td>
</tr>
<tr>
<td>state imposed restrictions</td>
</tr>
<tr>
<td>stock</td>
</tr>
<tr>
<td>supply chain cost efficiencies</td>
</tr>
<tr>
<td>transition from LIBOR</td>
</tr>
<tr>
<td>wage rates</td>
</tr>
</tbody>
</table>
5.3.5 Phase 3 – Using HS to Inform Model Inputs

After a long description of technology components, the four example scenarios show us what is in it for managers: a quantitative model that can be run with many different inputs to fully explore the cone of uncertainty. This example demonstrates how the model can be used by managers or any team of people who want to better understand how driving forces might affect them in the future. With a quantitative model, we could run as many scenarios as we wanted with any combination of inputs the users of the model consider plausible, but these models do not evolve and they do not give us any indication on which potential combinations we should try together. HS is the practice of monitoring weak signals that carry “information on potential change of a system towards an unknown direction” (J. Kim et al. 2016) with an eye on the future (van Rij 2010b; 2010a). HS could be used to monitor terms over time that may be the same terms used as concepts in quantitative scenarios, thus HS could be used with quantitative models to keep them up to date and identify the most plausible combination of inputs given changes in the current environment.

Some of the terms identified in the model above relate to the COVID-19 pandemic and how it could potentially further impact Levi’s business because at its peak COVID-19 was a major disruptor to supply chains, impacted the climate by reducing travel, and in the case of denim decreased demand as people opted to wear jogging pants while stuck at home. Using HS we could have made changes to strategic direction as all these factors changed drastically and quickly. As an example output of HS, I have
used data from news stories in the US between late 2019 and early 2021 to show the cycle of the term “virus” through the rise of COVID in the US. At first, we can see the term “virus” start as a weak signal in US news in late 2019. In March it becomes such a main topic that it starts to drown out all other news until the summer of 2019 where it stays in the stable area quadrant, only to settle into niche news in early 2021. This is demonstrated in the charts shown in Figure 33.

![Figure 33 – The term “virus” moving through HS quadrants as COVID-19 becomes prevalent in the U.S.](image)

In this case, we could have created a new set of plausible inputs as early as January 2020 when COVID was identified as a weak signal, setting demand to low (as people are stuck at home), reduction in workforce to high (when people stay home sick), complexity to high (as supply chains are disrupted), and global climate change to low (as people stop traveling) to help inform potential strategies in the case of a pandemic.

One could use HS to look further into the future by monitoring trends such as how the workforce is changing with the integration of more AI and robotics, or new
automation that improves supply chain efficiency. These could directly impact the potential plausible combinations of inputs of our model and help us create new scenarios to help prepare for the potential future. Using HS we can monitor terms that relate to the concepts identified in our model to understand how they are trending today to get a head start on planning for potential repercussions in the future. For example, if a signal starts to emerge as a weak signal we could use it in a new group of plausible input combinations to better understand how it might affect us in the future. If we see a term transition from a weak signal to a growth area we may create a new set of plausible inputs that could happen in the shorter term. HS would serve as a guide for people to rethink their strategy based on what is happening in the world today, helping teams stay current with a data-based reference as to how they can run simulations to get more plausible results. The Integration flow of HS in SAAM is shown visually in Figure 34.

Figure 34 - Using HS to assist people in running FCMs with plausible inputs

These examples demonstrate that once a quantitative model is built the output of HS helps bring the modelers along, after all, setting strategy is a social process where people benefit the most from thinking through problems and with the aid of data.
5.4 Concluding Summary

This research describes the conceptual system SAAM, which uses NLP to identify scenario drivers and their interdependencies using Q&A methods on any text corpus, utilizing off-the-shelf models by integrating people and AI into the model building process with IML to check results and provide filters. These scenario drivers can then be turned into a quantitative model in the form of an FCM to be used for modeling potential futures that cover the range of uncertainty by simulating different scenarios. Finally, SAAM uses HS to identify signals in real-time text data to help modelers find alternate futures that are plausible and decision-relevant.

SAAM can automate the bottlenecks in the scenario process while integrating people at key points for more transparent and plausible scenarios. This process has the potential to bring scenarios that help break through cognitive bias to a broad range of people, anyone who wants to have a better understanding of the potential future to make better decisions. SAAM could be used in a wide range of applications from financial advising to product development to career development to name a few. In this illustrative example, we used data from a financial report as an example to understand how well a company may fare in its current environment, but many other data sources could be explored including blogs, news sources, books, or academic journals. Future research for SAAM includes moving the system from conceptual to prototypical and measuring the various parts of the system from a technical perspective. There are also
several applications that SAAM can be tested on to see how people can incorporate scenarios in their work.
6 Evaluation

The previous chapter provided an overview of SAAM by demonstrating each component, describing how they were implemented and how people interact with them. In this chapter, I describe how I have evaluated selected SAAM components through 3 experiments. The evaluation provides insights suitable for answering the research questions of my work, which were identified in Chapter 4. They are:

RQ1: How can NLP be used to identify scenario drivers and their interdependencies from diverse knowledge sources using a Q&A method?

RQ2: How can scenario drivers identified in RQ1 be modeled in an FCM and used to generate alternative futures that are plausible, decision-relevant, and cover the range of uncertainty?

RQ3: How can information about possible changes to scenario drivers – so-called weak signals - be collected through text-based automation and used to update scenarios?

The first two questions pertain to the scenario functions of SAAM. The first asks how SAAM can identify scenario drivers (or “concepts” in an FCM) and their relationships (or “causal links” in FCM). The second asks how “good” (i.e., plausible, meaningful, relevant for a decision, etc.) the resulting scenarios are. The third question concerns the horizon scanning function of SAAM and how the weak signals it identifies can be used to find different plausible configurations of concepts based on current events and how they can be used to update models over time. None of these
questions map to single modules of SAAM but require the involvement of most or all its three main elements (Figure 27): Setup, Model Building, and Use. Evaluating such a complex system requires planning and prioritization and I use the framework of Design Science Research (DSR) to achieve this: in the subsequent sections, I first describe the overall evaluation approach and, building on it, present data and results of three separate experiments.

6.1 Overview of the Evaluation Approach

According to Baskerville et al., there are three aspects of evaluation in DSR that need to be considered: (1) evaluation of the artifact, (2) evaluation of the knowledge outcomes, and (3) evaluation of the research methodology (Baskerville, Kaul, and Storey 2018). There are several ways to evaluate an artifact in DSR, which include observational methods through case studies and usage monitoring, analytics methods that study the structure of the artifact itself, experimental methods where the artifact is used within the scope of controlled experiments or simulations, testing methods where functionality and design are tested, and descriptive methods where an artifact is used to inform an argument to demonstrate utility (A. R. Hevner et al. 2004). To address my research questions, I utilize analytical methods to test the technical ability of SAAM, which in the language of DSR is “the artifact”. Analytical methods are used to test the fit of the architecture of the artifact against the specified research questions. Other validation methods, not used here, can be used in future research, some of which are outlined in Chapter 7 with plans on how to potentially execute them. Table 15 outlines the full list
of possible evaluation methods as defined by Hevner et al. (2004), designating which categories are used in this research.

Table 15 - DSR evaluation methods, adapted from (A. R. Hevner et al. 2004)

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Used here</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observational</td>
<td>Case Study: Study artifact in-depth in a business environment</td>
<td>Future work, See Ch. 7</td>
</tr>
<tr>
<td></td>
<td>Field Study: Monitor use of artifact on multiple projects</td>
<td>Future works, See Ch. 7</td>
</tr>
<tr>
<td>Analytical</td>
<td>Static analysis: Examine the structure of the artifact for static qualities (e.g., complexity)</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Architectural analysis: Study fit of architecture into technical IS architecture</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Optimization: Demonstrate inherent optimal properties of an artifact or provide optimality bounds on artifact behavior</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Dynamic analysis: Study artifact in use for dynamic qualities (e.g., performance)</td>
<td>No</td>
</tr>
<tr>
<td>Experimental</td>
<td>Controlled experiment: Study artifact in a controlled environment for qualities (e.g., usability)</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Simulation: Execute artifact with artificial data</td>
<td>Yes</td>
</tr>
<tr>
<td>Testing</td>
<td>Functional (Black Box) testing: Execute artifact interfaces to discover failures and identify defects</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Structural (White Box) testing: Perform coverage testing of some metric (e.g., execution paths) in the artifact implementation</td>
<td>No</td>
</tr>
<tr>
<td>Descriptive</td>
<td>Informed argument: Use information from the knowledge base (e.g., relevant research) to build a convincing argument for the artifact’s utility</td>
<td>No</td>
</tr>
</tbody>
</table>
This research tests SAAM from an experimental perspective; tests the architecture and uses historical data for verification. Note that the table above shows possible methods for any artifact, not all of which align with the scope of my research. Descriptive methods are designed to test hypothetical artifacts and I am building a tangible artifact so I will not be using descriptive methods at all in my evaluation. Observational methods are outside the scope of this phase of research but potential further experiments are outlined in Chapter 7.

DSR can be used to evaluate an artifact from several different perspectives, often while using an artifact in context, understanding how it performs, and getting feedback from people who use it. A logical first step before any full-scale implementation of a novel IT artifact and its test with users is to assess its modules independently and to review how their respective outputs fit with the input needs of other modules. This course of action is particularly advisable when modules are very novel, as is the case in this work (A. R. Hevner et al. 2004). In this chapter, I conduct three different experiments to test the effectiveness of SAAM. The first experiment tests SAAM’s ability to identify scenario drivers and their interdependencies (RQ1) by comparing models built by SAAM with models created by people in a previous study. This will test SAAM’s ability to identify drivers and dependencies against the ability of people who are used in the manual, time-consuming scenario modeling SAAM is
designed to accelerate. My hypothesis for this experiment is that SAAM can create models that are technically indistinguishable from those created by people, thus showing SAAM’s models are as good or better than those created by people using the same text.

The second experiment also informs RQ1 by identifying scenario drivers and their interdependencies in a diverse set of documents. In addition, it tests how models created by SAAM can generate alternative futures that are plausible, decision-relevant, and cover a broad range of uncertainty (RQ2) by comparing generated scenarios to those of another study that uses an alternate method for accelerating scenarios, specifically LSA and FARM. I hypothesize that SAAM can create scenarios that are comparable to those created by other methods that utilize NLP.

In chapter 5 I demonstrated how possible changes of scenario drivers can be collected through text-based automation and used to update scenarios (RQ3). The third experiment of this chapter tests how HS output could have been used on the model generated from the second experiment to see what insight, if any, could have been found in using HS with the generated scenarios over time.

Figure 35 revisits the dataflow diagram from Figure 19 to show which steps in the system are tested by each experiment.
Figure 35 - Figure 19 revisited, showing which experiments validate the components of SAAM
6.2 Experiment 1 – SAAM vs Humans

Technology has competed with people since the earliest industrial revolutions. The ultimate test of any form of AI for decades has been the Turing Test, where one cannot distinguish the difference between a person and a machine (Saygin, Cicekli, and Akman 2003). SAAM is designed to accelerate the scenario process by speeding up parts of the process traditionally done by people; thus the first experiment compares FCM based mental models created by people with those created by SAAM to determine if they have similar content, structure, and dynamic system behavior. In this comparison, I will use data from a previously published study titled *Are We Modeling the Evidence or Our Own Biases? A Comparison of Conceptual Models Created from Reports* (Freund and Giabbanelli 2021) where several people read through a paper on the causes and effects of obesity and coded it into an FCM. The original study demonstrates how people bring an inherent bias to their model and catalogs the differences between them by comparing the number of concepts, connections, and semantic similarities between them. I ran SAAM on the same text and compared the output to the results from the human subjects, firstly measuring static concepts, such as number of connections and semantic similarity as compared against people’s models, and secondly dynamic consistency across each model by running simulations with the generated models. This experiment validates SAAM in two ways: (1) by showing that important concepts are captured from the text by comparing semantic similarity against a sample of concepts captured by people, and (2) by showing that generated models are at least as consistent as models generated by people.
6.2.1 Method Overview

This experiment starts with building two different models through two different lines of questioning with SAAM, a naïve approach, and an informed-question approach, using the corpus from the original study. Because SAAM uses a Q&A-based AI model to find connections, it could be possible to create very different models on the same corpus with different questions. Using two different lines of questions allows us to test if models using different types of question strategies create different models about those created by people.

Because FCM are directed graphs, comprised of vertices $V$ (nodes/concepts) and edges $E$ (causal connections) the original study used network science to analyze, compare and contrast different FCM as demonstrated in (Freund and Giabbanelli 2021) so I adopted the same approach. We can compare the graphs by the number of nodes in the graph and the number of edges between the nodes. Additional metrics that help us understand the nature of the relations in the graph are density, or the number of nodes compared to how many ties between nodes are possible, and number of cycles which is the number of unique, closed paths in the network that start and end at the same node. I will use these metrics in the first step of static analysis step to compare similarities and differences between all FCM models.

As a secondary means of comparison, semantic similarity of the concepts calculated with Levenshtein distance and cosine similarity across each model will show us if the different models contain the same or similar concepts, or if different concepts
were identified between human subjects and SAAM. Levenshtein distance (Levenshtein 1966), or the minimum number of single-character edits required to change one word to another tests how similar each word is from a spelling perspective where cosine similarity (Rahutomo, Kitasuka, and Aritsugi 2012) converts words into numerical representations or vectors and measures the distance between them in context. Levenshtein distance shows us the similarity between exact words, so “eating” and “eaten” would have a high similarity, and “eating” and “food intake” would have a very low similarity even though they appear similar in meaning. This loss of meaning can be mitigated by using cosine similarity, which compares representations of words in relation to one another in the text. Thus, if two different sentences use different words in the same context those words would have a higher cosine similarity than if they were used in completely different contexts. In the comparison study (Freund and Giabbanelli 2021) the modelers were asked to use specific words from the corpus as concepts in their model which warrants the use of both Levenshtein and cosine distance, these two metrics combined give us a more complete picture of how similar concepts are between each model.

Finally, to test how well the SAAM-generated and the person-generated FCM represent the dynamic behavior described in the underlying text, I ran a simulation with each model and compared the outputs to passages in the text. These steps are shown in Figure 36 and detailed in the following sections.
6.2.2 Model Building with SAAM

SAAM’s components, the technology used, and data flow were demonstrated in the previous chapter. This chapter will focus on specific aspects of validation and rather than detailing exactly how the models were built, we will describe the inputs, parameters, and outputs of the model assuming we are using the previously described method. The data used in this experiment was provided by the authors of the original study (Freund and Giabbanelli 2021).

6.2.2.1 Model 1 – Asking Naïve Questions

In the original study, the group of people was asked to model the cause and effect of obesity, so for the Q&A strategy in SAAM, I started with a naïve approach, asking (1) “what causes obesity?” with the subject being “Obesity” with a positive causal connection from the answer to the subject and (2) “what does obesity cause? ” where again “Obesity” is the subject but with a positive causal connection from the subject to the question. This assumes that whatever causes obesity will cause it to increase and
what obesity causes will have a positive effect on obesity. The questions used in this approach are shown in Table 16.

Table 16 - First set of questions used in validation experiment 1

<table>
<thead>
<tr>
<th>Question</th>
<th>Subject</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>What causes obesity?</td>
<td>Obesity</td>
<td>1</td>
</tr>
<tr>
<td>What does obesity cause?</td>
<td>Obesity</td>
<td>1</td>
</tr>
</tbody>
</table>

An excerpt of these results is shown in Table 17.

Table 17 - Example results from experiment 1, first set of questions

<table>
<thead>
<tr>
<th>Start node</th>
<th>End node</th>
<th>weight</th>
<th>answer</th>
<th>score</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>environmental influences</td>
<td>obesity</td>
<td>1</td>
<td>trends in environmental influences on physical activity and food intake</td>
<td>0.293415</td>
<td>{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}</td>
</tr>
<tr>
<td>national guidance</td>
<td>obesity</td>
<td>-1</td>
<td>national guidance, policy, research directions, and partnership initiatives</td>
<td>0.465117</td>
<td>{'neg': 0.0, 'neu': 0.87, 'pos': 0.13, 'compound': 0.4404}</td>
</tr>
</tbody>
</table>

The next step in the process involves the modeler deciding how to filter results. I used a confidence threshold to .29, a semantic threshold of .9, and removed and adjectives, punctuation, particles, symbols, and interjections returned as answers. The filtered results are shown in Table 18.
I used a Question Depth of 1, for each returned concept C asking, “what causes C to increase?” and “what causes C to decrease?” where C is the subject. This resulted in a model with 19 concepts and 26 connections. At this point, 75% of the answers fell below the set threshold and I did not run a third round of follow-up questions.

6.2.2.2 Model 2 – Asking Informed Question

The second line of questioning involved asking informed questions that more specifically ask about the details of the text. To do this without having to parse the entire document I based questions on the headers in the document. In this case, the headers are easily identifiable, as they are a larger size and different color font as demonstrated in Figure 37.
I manually formulated questions by turning these into questions as demonstrated in Table 19.

Table 19 – How headlines were rephrased to questions from the corpus

<table>
<thead>
<tr>
<th>Headline</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human and Societal consequences of the obesity epidemic</td>
<td>What are the consequences of the obesity epidemic?</td>
</tr>
<tr>
<td>Obesity Prevalence and Trends</td>
<td>What trends does obesity affect?</td>
</tr>
<tr>
<td>Contributory Trends</td>
<td>What contributes to obesity?</td>
</tr>
<tr>
<td>Advances during the past decade and barriers to further progress</td>
<td>What advances have been made? What barriers are there?</td>
</tr>
<tr>
<td>Tracking progress on outcomes</td>
<td>What progress has been made?</td>
</tr>
</tbody>
</table>

This led me to the questions shown in Table 20. Note in the table below an “- -“ in front of the subject indicates that the answers returned should be the subject, and the term here should be the cause.

Table 20 – Follow up questions used in experiment 1

<table>
<thead>
<tr>
<th>Question</th>
<th>Subject</th>
<th>Correlation</th>
</tr>
</thead>
</table>

Figure 37 - Screenshot of headers from original corpus used in validation experiment 1
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>What does obesity cause?</td>
<td>Obesity</td>
<td>-1</td>
</tr>
<tr>
<td>What are multiple social and economic problems?</td>
<td>Obesity</td>
<td>-1</td>
</tr>
<tr>
<td>What are trends in environmental influences?</td>
<td>Obesity</td>
<td>-1</td>
</tr>
<tr>
<td>How can we accelerate preventative efforts?</td>
<td>preventative efforts</td>
<td>-1</td>
</tr>
<tr>
<td>What can serve as a roadblock?</td>
<td>roadblock</td>
<td>-1</td>
</tr>
<tr>
<td>What causes poor quality of life?</td>
<td>quality of life</td>
<td>-1</td>
</tr>
<tr>
<td>How does obesity affect self-esteem?</td>
<td>self-esteem</td>
<td>1</td>
</tr>
</tbody>
</table>

This line of questioning resulted in 112 total results. I applied the same filtering criteria as before, dropping any answers with a confidence score lower than .29 or were semantically similar to existing terms. This resulted in a first model with 31 concepts and 35 connections.

6.2.3 Static Analysis

For static analysis, I rebuilt the models from the original study and calculated the number of nodes, number of edges, density, and cycles for each subject and SAAM, then plotted each metric to show how they relate to each other. For each metric, both models generated by SAAM fell within the range of the models created by people, which shows that based on a static analysis the automated results are indistinguishable from the manually created results. These metrics are shown in Table 21 and plotted as violin charts in Figure 38 with the models created by SAAM noted in the charts.
Table 21 – Network metrics across all models in validation experiment 1

<table>
<thead>
<tr>
<th>Subject</th>
<th>Nodes</th>
<th>Edges</th>
<th>Density</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>29</td>
<td>69</td>
<td>.073</td>
<td>23</td>
</tr>
<tr>
<td>B</td>
<td>36</td>
<td>34</td>
<td>.026</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>25</td>
<td>36</td>
<td>.06</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>15</td>
<td>34</td>
<td>.161</td>
<td>65</td>
</tr>
<tr>
<td>E</td>
<td>12</td>
<td>24</td>
<td>.181</td>
<td>10</td>
</tr>
<tr>
<td>SAAM1</td>
<td>28</td>
<td>28</td>
<td>.037</td>
<td>1</td>
</tr>
<tr>
<td>SAAM2</td>
<td>16</td>
<td>30</td>
<td>.125</td>
<td>47</td>
</tr>
</tbody>
</table>

Figure 38 - Static analysis from validation experiment 1 visualized in violin plots where the y-axis represents the values for each metric and the x-axis represents the average number of participants at each value. Thin sections represent fewer participants at that value. SAAM is explicitly shown to demonstrate where they fall in the distribution.

These results also show that even when given the same text different people will create different mental models with a different number of concepts and different connections between concepts. In comparison to A’s model with 29 concepts and 60 connections, E has a much smaller model with only 12 concepts and 24 connections. B has identified 36 concepts, but only 12 of them have outward causality which indicates of the 36 only 12 are drivers of potential simulation, where A has 25 drivers out of 29
concepts, C has 21 drivers out of 25 concepts, D has 12 out of 15 and E has 11 of 12.
Such a wide range of cognitive maps makes it difficult to compare SAAM directly to the
output of others, though SAAM’s results are not an anomaly as compared to the rest,
they fit within the range of possible outputs as defined by the original set of subjects.

Secondly, both Levenshtein and cosine distance was calculated by comparing
each term against each other terms, then averaging the similarity across all terms as
shown in Formula 7.

\[ \mathcal{C} = \left( \forall a \in A \left( L(a, \forall b \in B) \right) \right) \]

*Formula 7 – Calculating Levenshtein distance across nodes*

Where \( \mathcal{C} \) is a set of semantic similarity measures, \( a \) is a concept in set \( A \)
identified by Subject A, \( b \) is a concept in set \( B \) identified by Subject B, and \( L \) is the
Levenstein distance between 2 concepts. Pearson correlation is used across each \( \mathcal{C} \) to
determine a correlation matrix.

The results show a strong correlation between the two SAAM-generated models,
but a weak correlation between all other models. I charted the results of both
Levenshtein distance and cosine similarity in correlation matrixes to visually represent
how different each model was from the others. These are shown in Figure 39, where
higher numbers correspond to high positive correlation and lower numbers correspond
to high negative correlation. For Levenshtein distance on average SAAM’s results were
middle of the pack, with Subject A and Subject B having a stronger similarity on average
but only slightly and with Subject C right behind. Subject D and Subject E were on average more dissimilar to the rest, but they also identified fewer concepts that can account for the semantic differences. Using cosine distance SAAM2 was on the lower end of correlation, but was not very different from Subject E.

![Figure 39 - Correlation matrixes showing cosine and Levenshtein distance across all models in validation experiment 1](image)

For secondary visualization I compared each subject against the rest and plotted them on violin charts, marking the mean and median in each chart. For SAAM1 and 2 I removed the comparison to each other because the correlation was much stronger than the others and was more concerned with the correlation of the output of the algorithm than comparisons of different versions of the algorithm against people. Again, there is not a strong distinguishing factor, by looking at these charts for cosine and Levenshtein distance there is no obvious indicator of what was created by a person and what was
created by SAAM from either questioning method. Results are shown in Figures 40 and 41.

**Semantic Similarity – Cosine distance**

Figure 40 - Semantic similarity - cosine distance plotted for each model in validation experiment 1 where the y-axis represents the range of semantic similarity for each participant and the x-axis represents the average terms at each level that match other participants. Thicker sections on each chart represent values that are the most like other participants.
From this static analysis we can conclude that people will create mental models of varying sizes using different terms even after reading the same corpus. This makes distinguishing SAAM’s result from this perspective difficult, but we can show that SAAM can create models that are within the range of people who created models from the same text and are not distinguishable by these metrics. We can also conclude that from a static analysis perspective SAAM can create models comparable to those created by people using the same text corpus even with different question strategies. The different question strategies did create different models but were more alike to each other than...
to those created by people, and more like each other than any person’s model was to any other person.

6.2.4 Dynamic Analysis

Static analysis shows that SAAM can create maps that are statistically within the bounds of those created by people, but the more important question is if they are good representations of the existing knowledge about the system they describe? In scenario planning, consistency, or the lack of contradiction within the model, is a key measure of a good model (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007). To test consistency, I ran simulations with all models to determine if the outputs of these models would generate results that were consistent with the defined correlations. In this case, we are working with a corpus that defines complex relationships between obesity, society, the economy, natural security, and public health. FCM are tools that can be used to model such complexity, and in theory and FCM that represents the arguments made in this document should reflect its conclusions. To measure consistency from this perspective I manually coded the resulting values of each concept after simulations with a 1 if the resulting value reflected the text and a 0 if the resulting value contradicted the text. I then calculated the percentage of values that resulted in a 1 to determine consistency.

The only concept that was included in every model was, (not surprisingly, given the topic of the text) “obesity”, which was relevant as something that causes other concepts to change (e.g. health) and as something that is impacted by other concepts
(e.g. nutrition): To compare models against each other I clamped the term “obesity” to 1, ran the model for each of the subjects and SAAM, and created consistency scores for each subject by comparing the final result of each model with the effect of each subject noted in their definition. For example, if obesity is set to 1 and a subject noted that had a positive effect on depression, then depression should also be 1 or close to 1. For each concept if the final value from the simulation reflected the text I scored it as a “1”, if not I scored it a “0”. I then created a score by calculating the percentage of concepts that scored a “1” for each model. The lowest human subject had a consistency score of 40%, with the highest score of 54.4% and all people together averaged 46.2%. SAAM 2 outperformed all people with a consistency rating of 61.5%, and SAAM2 significantly outperformed the entire group at 88.9%. Results are shown in Table 22 below, full results are shown in Table 23.

Table 22 - Consistency results from validation experiment 1

<table>
<thead>
<tr>
<th>Subject</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>45%</td>
</tr>
<tr>
<td>B</td>
<td>41.6%</td>
</tr>
<tr>
<td>C</td>
<td>40%</td>
</tr>
<tr>
<td>D</td>
<td>54.4%</td>
</tr>
<tr>
<td>E</td>
<td>50%</td>
</tr>
<tr>
<td>SAAM1</td>
<td>88.9%</td>
</tr>
<tr>
<td>SAAM2</td>
<td>61.5%</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Activity level</td>
<td>12</td>
</tr>
<tr>
<td>Feeding</td>
<td>0.95118</td>
</tr>
<tr>
<td>Breastfeeding</td>
<td>0.250399</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>0.957836</td>
</tr>
<tr>
<td>Chronic illness</td>
<td>0.957836</td>
</tr>
<tr>
<td>Depression</td>
<td>0.86414</td>
</tr>
<tr>
<td>Disability</td>
<td>0.96118</td>
</tr>
<tr>
<td>Employment</td>
<td>0.957806</td>
</tr>
<tr>
<td>High blood pressure</td>
<td>0.96118</td>
</tr>
<tr>
<td>Legislative action</td>
<td>0.250399</td>
</tr>
<tr>
<td>Military recruitment</td>
<td>0.46748</td>
</tr>
<tr>
<td>National security</td>
<td>0.87165</td>
</tr>
<tr>
<td>Number of ethnic minorities</td>
<td>0.250399</td>
</tr>
<tr>
<td>Obesity</td>
<td>0.957836</td>
</tr>
<tr>
<td>Physical education</td>
<td>0.78622</td>
</tr>
<tr>
<td>Public awareness</td>
<td>0.250399</td>
</tr>
<tr>
<td>Quality of life</td>
<td>0.957836</td>
</tr>
<tr>
<td>Recession</td>
<td>0.250399</td>
</tr>
<tr>
<td>School performance</td>
<td>0.947124</td>
</tr>
<tr>
<td>School wellness program</td>
<td>0.250399</td>
</tr>
<tr>
<td>Use of medication</td>
<td>0.957836</td>
</tr>
<tr>
<td>Use of medication / do</td>
<td>0.957836</td>
</tr>
</tbody>
</table>

Table 23 - Full dynamic comparison results from SAAM
6.2.5 Discussion and Limitations

The data collected in this first experiment measures differences in FCM models created by SAAM and those created by people statically through network metrics and dynamically by running a simulation. Even though we were able to show that SAAM created models that fall within the range of the original participants the depth of comparison possible is limited due to the limited number of people that built very different models in the original study. Additionally, the original study did not run simulations with the models, but it did conclude that because people bring an inherent bias to modeling that they would build less consistent models. FCM models are intentionally fuzzy and can uncover hidden complexities that seem contradictory at first but make more sense as a network of connections is built. In this case, we used a corpus that has empirically grounded information on what happens to other factors when obesity becomes prevalent and the various causes of obesity. This allowed me to create a consistency score grounded on the empirically presented data. With these results, we can conclude that SAAM can generate models with more consistent dynamic behavior than people using the same corpus and in theory, could help future modelers overcome inherent bias when building models by recommending causal pairs. However, it could also be possible that using different, more diverse data sources that describe a domain in different ways measuring consistency from SAAM would be much more difficult.

The results from this experiment seem to add another dimension to the conclusion from the original study, showing that the final output of the models created
by people was of a much lower consistency than those created with SAAM – a system
based on AI that is intended to reflect the text and, in this case, more specific
questioning led to the most consistent model. We can also conclude that models
created by SAAM fall within the statistical boundaries of a set of people asked to
perform the same task of creating a mental model based on a specific test. This could
mean that SAAM is well suited for creating FCM that are comparable to those created
by people that could be used for dynamic analysis.

6.3 Experiment 2 – SAAM vs LSA/FARM

Experiment 2 evaluates SAAM’s ability to extract concepts and causal links from
a text, combine them into an FCM model, and use the model to run simulations to
generate alternative future scenarios that are plausible, decision-relevant, and cover the
range of uncertainty. Evaluation occurs by comparing SAAM to a different approach
with similar objectives that use LSA and FARM. The approach is documented in a study
titled *Futuristic data-driven scenario building: Incorporating text mining and fuzzy
association rule mining into fuzzy cognitive map* (J. Kim et al. 2016) and was already
introduced in Chapter 4, Figure 18. This study is part of a series of similar studies by the
same research group that uses the NLP-based method to speed up creating scenarios
(Kwon, Kim, and Park 2017; J. Kim et al. 2016; Kwon and Park 2018; Son, Kim, and Kim
2019) and is, to my knowledge, still the only known combination of NLP techniques with
FCM scenario building. I investigate similarities and differences between this study and
SAAM are illustrated in Figure 42. I hypothesize that SAAM, using the same data, can generate alternative future scenarios that perform at least as well as the prior approach.

![Comparison of steps between SAAM and LSA/FARM model building](image)

**Figure 42 - Comparison of steps between SAAM and LSA/FARM model building**

### 6.3.1 Method Overview

This experiment begins with collecting data, then building a model with SAAM, and finally running simulations to compare scenarios with those from the original study. The authors from the original study were looking at political, economic, social, technological, environmental, and legal (PESTEL) factors in their models as they are used commonly in scenario planning (Oliver and Parrett 2018; Norouzi, Fani, and Ziarani 2020) so I designed questions for SAAM to ask questions based on PESTEL. I then compared the SAAM-generated scenario against the scenarios from the study. This requires answering the question “what makes a good scenario?” The most common measure of scenario quality is plausibility, where there is a conceivable path to a
potential scenario from the present state (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007). Several studies use internal consistency to measure scenario quality to ensure scenarios do not contradict themselves (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007) or coherency to ensure that the scenarios make logical sense (Durance and Godet 2010; Bradfield, Derbyshire, and Wright 2016). The relevance of the scenarios to the team is another important factor in scenario quality (Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Durance and Godet 2010). There are several other measures which include differentiation between a set of selected scenarios (Wilson 1998), the creativity of the scenarios (Alcamo and Henrichs 2008), whether they reflect uncertainty and help people create new perspectives (Heijden 2011), and how transparent they are (Durance and Godet 2010). In this experiment, because we have a previous model using the same data, I run simulations using SAAM with similar inputs as defined in the comparison study. I assume that the previous, peer-reviewed and published study contributed scenarios that are plausible, decision-relevant, and cover a wide range of uncertainty, therefore by comparing these scenarios to those created with SAAM with similar inputs we can determine if SAAM can generate scenarios with the same qualities.

These steps are outlined in Figure 43 and demonstrated throughout the rest of this section.
6.3.2 Data Collection

The authors of the comparison study did not publish the data they used, however, based on their description I was able to largely reconstruct what they used: data was gathered through web scraping from five websites: Siemens\(^9\), MIT technology review\(^{10}\), Kurzweil Accelerating Intelligence\(^{11}\), World Future Society\(^{12}\), and FutureTimeLine\(^{13}\). These sites were selected by the original authors because they all published articles that were “futuristic” and thus already contained an analysis of trends and expert insight on the potential future. They did not make their data source accessible and did not describe how they filtered data, but I was able to access them from the web archive Wayback Machine\(^{14}\) with the following assumptions: (1) since the paper was written in 2016 I scraped data from each site that was available on March of 2016 which should reflect most of the articles that were scraped and used in the original

---

\(^{10}\) http://www.technologyreview.com/topics/  
\(^{11}\) http://www.kurzweilai.net  
\(^{12}\) http://www.wfs.org  
\(^{13}\) http://futuretimeline.net/index.htm  
\(^{14}\) http://web.archive.org/
study and (2) I only scraped articles that discussed electric vehicles or alternative energy as the original authors noted they had done.

6.3.3 Formulating Questions

Scenarios often consider PESTEL factors to understand the potential future, as did the authors of the original study. I structured my questions by asking about electric vehicles from the lens of each aspect of PESTEL with a few open general questions about electric vehicles. For example, for “environmental” factors I ask, “what are benefits to the environment” and “what hurts the environment”. Note the structure of questions is the same as the first experiment where the subject of each question is predetermined along with the correlation between the subject and answer. The full list of questions used is shown in Table 24.

Table 24 – Questions used in Validation Experiment 2 based on PESTEL for Electric Vehicles

<table>
<thead>
<tr>
<th>Question</th>
<th>Subject</th>
<th>Correlation</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>What technology is needed for electric vehicles?</td>
<td>EV adoption</td>
<td>1</td>
<td>Technology</td>
</tr>
<tr>
<td>Why use an electrified vehicle?</td>
<td>EV adoption</td>
<td>1</td>
<td>Open</td>
</tr>
<tr>
<td>What are impediments?</td>
<td>EV adoption</td>
<td>-1</td>
<td>Open</td>
</tr>
<tr>
<td>What are political factors?</td>
<td>EV adoption</td>
<td>1</td>
<td>Political</td>
</tr>
<tr>
<td>What are the benefits to the environment?</td>
<td>--EV adoption</td>
<td>1</td>
<td>Environmental</td>
</tr>
<tr>
<td>What hurts the environment?</td>
<td>Environment</td>
<td>-1</td>
<td>Environmental</td>
</tr>
<tr>
<td>What are social benefits?</td>
<td>--EV adoption</td>
<td>1</td>
<td>Social</td>
</tr>
<tr>
<td>What are social problems?</td>
<td>-- EV adoption</td>
<td>-1</td>
<td>Social</td>
</tr>
<tr>
<td>What social aspects affect electric vehicles?</td>
<td>EV adoption</td>
<td>1</td>
<td>Social</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-------------</td>
<td>---</td>
<td>-------</td>
</tr>
<tr>
<td>What are the economic benefits?</td>
<td>economy</td>
<td>1</td>
<td>Economic</td>
</tr>
<tr>
<td>What are economic problems?</td>
<td>economy</td>
<td>-1</td>
<td>Economic</td>
</tr>
<tr>
<td>What are economic drivers?</td>
<td>economy</td>
<td>-1</td>
<td>Economic</td>
</tr>
<tr>
<td>What are legal problems?</td>
<td>EV adoption</td>
<td>-1</td>
<td>Legal</td>
</tr>
<tr>
<td>What are legal drivers?</td>
<td>EV adoption</td>
<td>1</td>
<td>Legal</td>
</tr>
<tr>
<td>What are legal benefits?</td>
<td>-- EV adoption</td>
<td>1</td>
<td>Legal</td>
</tr>
</tbody>
</table>

After filtering the model consisted of resulted in 52 unique concepts with 110 connections, as compared to the 15 concepts and 44 connections from the original study. The terms identified include the following: 'EV', 'a completely carbon neutral transportation option', 'a comprehensive charge station network', 'aboriginal training', 'artificial intelligence', 'batteries', 'biomimicry', 'business development', 'carbon pricing', 'cities conservation', 'clean renewable energy sources', 'confidence', 'current unit sales', 'durability', 'economic activity', 'economic and safety benefits', 'electric motor', 'employment', 'energy efficiency', 'energy pollution', 'environmentally conscious citizens', 'evs cost', 'fear', 'gaps', 'generic super charging stations', 'governments', 'greenhouse gas emissions', 'harmony', 'incentives', 'information technology', 'infrastructure', 'lack of hydrogen infrastructure', 'liability', 'no exhaust emissions', 'oil and gas volatility', 'polarisation systems', 'potential roadblocks', 'power and mileage limits', 'public investment', 'rare earth metals', 'recharge speed', 'regulation', 'remote communities', 'save lives', 'scholarships', 'self - driving vehicles', 'significant technology
improvements', 'sustainability', 'the air', 'the falling price of batteries', 'the power and mileage limits', 'thinking globally and acting locally', 'traffic congestion', 'transform mobility', 'wealth', 'your gas guzzler'.

6.3.4 Comparing Outputs

For the sake of comparison, I used the labels from the original study to group SAAM’s concepts into categories. SAAM identified some of the same terms that were identified in the original study, but SAAM identified other concepts that were not detected in the original study. For example, SAAM identified aspects such as consumer confidence, infrastructure investments needed, and natural resources required to build required batteries. These could be valuable details that give us deeper insight into the data creating more robust models. Asking specific questions about social impacts led to answers such as ‘thinking globally and acting locally’ which was completely lost in the LSA method. On the other hand, there were some topics identified that do not make sense without some context, such as “your gas guzzler” and “aboriginal training”. Upon closer inspection “your gas guzzler” is referring to cars that run on gas, in particular cars that are still on the road today. “Aboriginal training” was from an Australian article about retraining people in poor communities to work in the high-tech sector driven by electric vehicles.

The concepts from SAAM are shown in Table 25 with corresponding categories from the original study for comparison.
<table>
<thead>
<tr>
<th>Category</th>
<th>SAAM Concepts</th>
<th>LSA Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air pollution</td>
<td>greenhouse gas emissions, no exhaust emissions, the air, your gas guzzler, energy pollution</td>
<td>Temperature, environment, pollution, atmosphere, carbon dioxide emission, greenhouse gas, CO2, eco</td>
</tr>
<tr>
<td>Alternative energy technology</td>
<td>clean renewable energy sources, polarization systems</td>
<td>Renewable energy, diesel, biofuel, biomass, geothermal, petroleum, gasoline, hybrid, photovoltaic, solar energy</td>
</tr>
<tr>
<td>Battery technology</td>
<td>batteries, power and mileage limits, recharge speed</td>
<td>Lithium battery, ion battery, acid battery, storage, battery life, lightweight, BMS, lithium ion battery</td>
</tr>
<tr>
<td>Charging technology</td>
<td>a comprehensive charge station network, generic supercharging stations</td>
<td>Wireless power, charger, recharge, power transmission, charger</td>
</tr>
<tr>
<td>Costs reduction</td>
<td>EVs cost, the falling price of batteries</td>
<td>Cost reduction, incentive, support, maintenance cost</td>
</tr>
<tr>
<td>Economic revenue</td>
<td>business development, current unit sales, wealth, economic activity</td>
<td>Economy, growth, sales, investment, revenue, GDP, trade, import, export</td>
</tr>
<tr>
<td>Energy efficiency</td>
<td>energy efficiency</td>
<td>Energy efficiency, energy consumption, efficiency improvement, energy density, mileage</td>
</tr>
<tr>
<td>Government regulation</td>
<td>carbon pricing, cities conservation, governments, incentives, public investment, regulation</td>
<td>Regulation, incentive, policy, government, limitation, standard, tax reduction, policy</td>
</tr>
<tr>
<td>Industry-university collaboration</td>
<td>scholarships, aboriginal training</td>
<td>Company, startup, university, laboratory, investment, partnership, entrepreneur, grid</td>
</tr>
<tr>
<td>Job creation</td>
<td>employment</td>
<td>Job, worker, manufacturing, services, employment</td>
</tr>
<tr>
<td>Motor technology</td>
<td>electric motor</td>
<td>Engine, inverter, magnet, DC, AC, torque, capacity, motor</td>
</tr>
<tr>
<td>Usability</td>
<td>information technology</td>
<td>Automation, sensor, network connection, software, comfort, assistant, internet</td>
</tr>
<tr>
<td>Public transportation</td>
<td>self-driving vehicles</td>
<td>Transportation, electric bus, driver, passenger</td>
</tr>
<tr>
<td>Safety</td>
<td>economic and safety benefits</td>
<td>Safety, driverless, collision, vibration, pressure, security, stability, obstacle warning, monitoring</td>
</tr>
<tr>
<td>Other</td>
<td>thinking globally and acting locally, a completely carbon neutral transportation option, biomimicry, confidence, durability, environmentally conscious citizens</td>
<td></td>
</tr>
<tr>
<td>Application to tourism</td>
<td>None</td>
<td>Consumer, customer, tourism, growth, economy</td>
</tr>
</tbody>
</table>

Using the category mapping in Table 25 we can run the model created by SAAM by setting corresponding inputs to the ones used in the original study and comparing the outputs of both. There were 4 scenarios in the comparison study which included (1)
the application of EV to tourism, (2) failure to develop battery technology, (3) failure of EV adoption in general, and (4) relaxation of government regulation. One inconsistency between models is the absence of an “Application to tourism” from the original study that was not identified by SAAM. This could be because there was data used in the original study that I did not capture for SAAM, but it is also because the original author’s method grouped the terms “economy”, “consumer”, “customer”, “growth”, and “tourism” into the tourism category assuming that tourism is driven by consumers and is directly related to the economy. Because SAAM did not identify tourism I substituted application to tourism with economic benefits in my comparison scenarios. In addition, though the third scenario claims to be about the failure of battery adoption it represents the inverse of the first scenario or a widespread battery adoption. As a result, I ran simulations where (1) the economy is good or the economy is bad to show how economic factors affect EV adoption, (2) what happens if battery technology does not develop, and (3) what happens if the government decides to not help the EV industry at all by removing any incentives for EV and stopping any regulation efforts to increase adoption.

Scenario 1 – No increase in applying EV to tourism/economic indicators:

The original study showed applying EV to tourism resulted in increased employment, a better economy, lower pollution levels, and improved energy efficiency. In the data, SAAM used there were no articles about tourism, so none of SAAM’s scenarios reflected results in that category. However, in the original paper tourism is
grouped with economic benefits, so I was able to run scenarios to test conditions in a
good or poor economy by setting the terms 'employment', 'business development',
'current unit sales', 'economic activity', 'economic and safety benefits', and 'wealth' to
high for one simulation and low for another.

SAAM’s model output a different result than the original study, noting that in a
good economy ‘no exhaust emissions’ are adopted, but ‘greenhouse gas emissions’
increase which has a negative effect on ‘the air’. In addition, we get richer results with
SAAM, as some concepts that were not identified with LSA show some interesting
results such as ‘think globally act locally’ decreasing in a good economy, ‘public
investment’ increasing, and ‘lack of infrastructure’ decreasing implying that
infrastructure will start to improve. In a good economy ‘EV adoption’ decreases and
‘your gas guzzler’ which represents existing gas-powered vehicles increases where the
inverse is true in a bad economy. This may seem counterintuitive, but on closer
inspection, we can see that while technology, consumer confidence, and battery
technology are high, a focus on sustainable decreases and volatility in gas prices
decreases which hurts the adoption of EVs. This scenario implies that in a good economy
even though EV infrastructure improves, battery technology improves, and energy
efficiency improves, there is no strong driver for consumers to adopt EV technology.

**Scenario 2 – Battery technology fails to develop:** The original study showed that
if battery technology is not improved, then there will be less job creation, less tourism, a
poor economy, and an increase in pollution. For this scenario I set the following terms to
low: 'batteries', 'lithium-air batteries', 'lithium-ion', 'lithium-ion batteries', 'recharge speed', 'power and mileage limits' and 'energy efficiency'.

SAAM also found that the term 'employment' decreased in this scenario, terms associated with the economy ('economic activity', 'business development', 'current unit sales', 'wealth') all ended on low values, but as with the first scenario, SAAM found an inverse relationship on the economy and the environment, so 'greenhouse gas emissions' decrease and 'the air' increases. In this scenario EV adoption starts to improve even though the cost of EVs ('evs cost') is driven up. In this case the desire for sustainable solutions ('sustainability') and 'public investment' end on high values implying that though battery technology fails to improve a drive for sustainability and investment from the government help offset the high cost of EV.

**Scenario 3 - Deregulation:** Finally by simulating the relaxation of government regulation by setting the concepts “regulation, incentive, policy, government, limitation, standard, tax reduction, and policy” to low the original study found a potential increase in safety, reduction in costs, and an increase in energy efficiency. Rather than finding a single stable result SAAM when I set all the concepts related to government regulation to high (regulation, public investment, incentives, scholarships) the simulation reached a limit cycle rather than a stable state, indicating that if the government does nothing then consumers would oscillate between EV adoption and rejection as the environment shifted from one preference to another because of competing factors. Because scenarios are meant to help stimulate thought and understanding I argue that this is a
richer result from the previous study as it implies that regulation is a key concept in the adoption of EV, and it should be considered carefully when considering future strategies for this domain. Though it is difficult to see the individual concepts in the resulting diagram mapping results, Figure 44 shows us the resulting simulation results from the deregulation scenario where a clear pattern can be distinguished. In this pattern, we see that nearly all the concepts in the model play off each other causing most of the factors in the model to waver from one extreme to another in the absence of regulation.

![Figure 44 - Validation experiment 2 – the pattern that emerges in the deregulation scenario](image)

The full results of different scenarios are shown in Table 26. Note that in the Deregulation scenario because an end state was never reached, the results only represent a part of the cycle and do not represent final values.
6.3.5 Discussion and Limitations

The results from this experiment show that SAAM can create quantitative scenarios based on data as has previously been demonstrated with LSA, but with some key differences. Firstly, LSA is used to find topics in text and inherently groups terms together. That grouping is lost with SAAM, which finds specific terms because of asking direct questions. This can lead to more seemingly granular results, but also can require...
some interpretation such as understanding that “your gas guzzler” or “the air” have representative meaning based on the context. Secondly, SAAM and the LSA method both use people in the scenario creation process but at different times. While SAAM uses people upfront to define questions then again to create filters, the LSA method uses people to set topic cluster size and give names to the final topics. These different methods require very different skill sets; while using LSA requires a background in data science or programming, SAAM’s Q&A interface and threshold setting could be used without such expertise.

One limitation of this experiment was the inability to use the same data as the original study. This caused us to miss the grouping that was labeled as “application to tourism” in the original study. Another potential limitation was the method used for questioning. Where in a previous experiment I tested SAAM with different question methods only one was used here because that was enough to fulfill the objective of the experiment, though future research could explore the use of different lines of questioning across different data sets.

Scenarios are supposed to help us step back and see the bigger picture, think outside the box, and consider alternatives that might not be obvious. In this experiment I have shown that SAAM, using the same data as the existing study using LSA, was able to generate alternative future scenarios that met this objective and therefore we can conclude that SAAM performed at least as well at scenario generation as the comparison method.
6.4 Experiment 3: integrating Horizon Scanning

This final experiment evaluates how information about possible changes to scenario drivers can be collected through text-based automation and used to update scenarios by using HS with historical data. Using the scenario models created in the previous experiment we use HS to monitor the development of specific concepts over time and show how HS can be used to inform people in real-time how to run models and update strategies with an eye to the future informed by the present. This experiment tests how DoD and DoV can be used with quantitative models to keep them up to date and identify combinations of inputs to help people identify the most plausible futures. I hypothesize that HS outputs naturally inform the potential inputs for quantitative models.

6.4.1 Horizon Scanning Data

In Section 5.3, Figure 27 shows a very high level of data flow in SAAM. Figure 45 revises that figure slightly to show us where this experiment fits in the overall process.

*Figure 45 - Figure 27 revisited, “YOU ARE HERE” – the part of the system validated by experiment 3*
I used the method described by (Yoon 2012) where DoV and DoD were calculated for specific terms over time. To simulate monitoring data over time I used historical news data from The GDELT Project\textsuperscript{15}, a repository of thousands of news sources scraped every 15 minutes and stored for open research. I collected data between 8/6/2019 – 11/3/2021 and parsed a representative sample; I used 56 2-day samples of data from sources in the United States throughout the period where I parsed every 100\textsuperscript{th} article. This resulted in 641,123 unique URLs with news stories and 4,827,706 terms identified.

6.4.2 Results using Historical Data

In this experiment, I refer to the 4 quadrants of HS as weak signals, growth areas, niche areas, and stable areas. To connect HS to quantitative scenarios I used the model created in Experiment 2, where data from future-facing websites was used to create scenarios about the potential adoption of electric vehicles. I picked 4 terms from the model which correspond to the 4 different scenarios: “electric vehicle”, “pollution”, “self-driving”, and “bad economy”. As control measures, I also monitored the terms “police” and “virus”. Because the source is comprised of news many of the stories talk about the police in some way so “police” should consistently be a stable area, and the selected period encompassed the time when COVID-19 began spreading in the United States and took over much of the news so we should see the term “virus” move from a

\textsuperscript{15} https://www.gdeltproject.org/
weak signal – when COVID was still growing overseas – to a stable area – when COVID became prevalent in the US.

As expected, the term “police” is consistently in the stable area quadrant or the entire timespan, but the rest of the terms except self-driving are mostly stable in the niche areas category. We can see that “self-driving” is emerging as a weak signal while EV, pollution, and a bad economy are consistently in the niche areas quadrant. Results mapped onto the HS quadrant are shown for March 2021, June 2021, and September 2021 as an example.
In early 2022 the Consumer Electronics Show (CES)\textsuperscript{16}, where companies announce new tech-based products every major auto manufacturer debuted an electric vehicle of some kind, but autonomous vehicles have yet to hit the market. We can conclude that today EV is no longer a weak signal but has earned its place in the “niche” quadrant by finding a stable cycle since the comparison study was conducted in 2016. Self-driving vehicles, however, are still emerging and are classified correctly in the weak signal quadrant, yet to diffuse throughout all media.

6.4.3 Applying HS to Quantitative Scenarios

The results from HS help us add another dimension to the terms used in our quantitative model. We can experimentally set inputs of our model to see what happens

\textsuperscript{16} https://www.ces.tech
when the economy does not do well or how EV adoption may affect self-driving vehicle technology, but HS gives us real data to help inform modelers how the terms they care about are evolving in real-time to help them determine what inputs they might want to explore to investigate potential scenarios that reflect emerging and pertinent trends.

Terms that are classified in the “growth leaders” quadrant are evolving rapidly in the present, so setting such terms to “high” in our model and running simulations should help us understand what is plausible in the near future and understand which scenarios currently have the most momentum. Terms in the “weak signal” quadrant are emerging and are usually still on the horizon, so terms from this quadrant can be set to “high” to see what potential disruptors are coming in the medium and long term. Terms found in the “Stable areas” quadrant could be set to low or high where low shows the repercussions of disruption and high keeps the status quo. Finally, terms in the “Niche areas” represent terms that have diffused but are not frequently present in the data set used for HS. To get a more granular understanding of these you may need a more specific data source, such as a group of journals or news sources that specialize in the topic.

Results from HS can also help us keep our model up to date. For example, our model showed a strong correlation between the economy and EV, however, by tracking them both with HS we can see that the two terms could have a much weaker correlation than previously found. We could use this knowledge to update our model with a weaker correlation to get more plausible results moving forward. For example, when we
trended away from battery failure, we could have rebuilt our model to be more future-looking. When we validated that the economy did not have a strong effect on EV adoption, we could have updated our weights. Since deregulation never happened, we can assume that things will continue to trend in the same direction. Self-driving vehicles seem to be a current weak signal, so we might want to rebuild our model with that at the center rather than EV adoption.

6.4.4 Discussion and Limitations

HS gives us algorithms to make sense of mountains of data in real-time, helping us bubble things to the top that are difficult to find in all the noise of everyday events. To demonstrate HS on historical data I used samples of data over time with the assumption that the samples were representative of the full set of data. In a real-time system it is more likely that this data would be parsed as it becomes available. In a previous experiment, I showed how to find semantic similarity between terms, and in a real-time HS system, one would likely want to monitor for the concepts in their model as well as semantically similar terms. In this experiment determining what to search for in hindsight is much easier than figuring out what we may want to look for moving forward.

6.5 Discussion on Overall Validation

This study used a few diverse examples to test the effectiveness of SAAM and though the results are promising there are some overarching limitations to this research. Firstly, as with any system that incorporates AI the data used to train the
models is a crucial factor. In testing SAAM against people I only used a single dataset and only a handful of people to compare results against. In the comparison with an alternate method, I did not find the exact data used in the original study but instead used a best guess to find data that was described in the study. Because of limitations on compute power, I only used samples of data for HS rather than the millions of news articles that were possible to scrape. These studies only used relatively small datasets to fine-tune the Q&A models used, there could be more exploration on how different size datasets improve or hurt the performance of SAAM. In addition to data, the models used to ask questions of the data are part of the quickly evolving field of NLP. SAAM relies on off-the-shelf generic models for Q&A but could be updated to use either domain-specific models or custom models that may yield different results.

Secondly, I only tested a few question strategies. Because SAAM relies on the answers to build models determining the best types of questions to ask and how to formulate them could generate better results for future models. Question strategy might also be tied back to the underlying Q&A models used for SAAM, as different Q&A models use different coded text that might map better to certain types of documents in different domains. The number of iterations of questions could also be explored further to understand how far we can probe to find answers and how that affects the resulting models.

Finally, one aspect of FCM that was not explored in this research is the variability of weights between concepts. The FCM generated uses simple weights, either 1 or -1.
Much FCM research has more granular sets of weights, where the value can be any number (other than 0) between 1 and -1. It could be possible to generate better models if determining more granular weights between concepts in the FCM, for example, it could be possible to use the confidence returned from the Q&A algorithm as the weight, or some other value. Though I did not explore changing the weights in this research it could be a topic for future research.

6.6 Concluding Remarks about Evaluation

In this chapter, I have conducted three different experiments to evaluate (1) if NLP can be used to identify scenario drivers and their interdependencies from diverse knowledge sources using a Q&A method, (2) How can scenario drivers identified in RQ1 be modeled in an FCM and used to generate alternative futures that are plausible, decision-relevant, and cover the range of uncertainty, and (3) How can information about possible changes to scenario drivers – so-called weak signals - be collected through text-based automation and used to update scenarios? I conducted these experiments using DSR through the evaluation of an artifact, SAAM, which aggregates technologies well known in computer science to create scenarios that can help managers, leaders, and teams make better decisions by gaining a better understanding of plausible futures.

In the first experiment, SAAM used a text corpus that was used by people in a previous study to build cognitive maps as FCMs. Using static and semantic analysis I showed that SAAM created models that were within the range of those created by
people and thus had similar content and structure. Running the models created by people and by SAAM showed that while the dynamic system behavior was similar SAAM had a much higher degree of consistency than people. Because the original experiment concluded that the human participants brought biases to their model and were subsequently outperformed in this research, we could imply that the use of an artifact such as SAAM could help reduce bias in modeling, thus becoming a valuable tool for people creating quantitative models in future studies. Because people have such diverse interpretations of the same text from a modeling perspective it is difficult to conclude that SAAM is better or worse than people at creating mental models, but we can conclude that if modelers are using a tool such as SAAM they can create models faster than by relying on the analysis of the system with traceability back to the text, rather than reading an entire corpus and creating their own models.

In the second experiment, I compared SAAM to a previous study where a different method of automation, namely LSA and FARM, was used to speed up the scenario process to test if SAAM could create models that are at least as good as models created with other methods documented in the literature. SAAM was able to create a model on the corpus described in the comparison study, however upon running simulations SAAM output some similar and some dissimilar results. The original method grouped and generalized terms that were returned from the automated analysis, thus losing much of the detail from the text. SAAM was able to capture these details and as a result, brought out potential interpretations and richer descriptions to explore the cone
of uncertainty. I would conclude that given scenarios intend to help broaden perspectives and encourage people to think through what could potentially happen that these more granular results are an improvement on previously existing methods for automating scenarios. More complex results encourage people to think harder about relationships between driving forces and make strategic decisions that may rely on information that is not intuitive.

Finally, this research showed how the output of HS can be used to inform the users of the model on how to run them to create the most plausible futures based on what we know today by incorporating HS. Using HS also gives the modeler information on how to keep the model up to date over time by monitoring terms that align to concepts with the model and understanding how the output of the model aligns to current trends.

In conclusion, the validation used in this study shows that SAAM is a viable method to create quantitative models used for scenarios and assist users of the models in how to keep them up to date and find the most plausible outputs. SAAM can identify scenario drivers and their interdependencies from knowledge sources using a Q&A method within the same statistical bounds as people can create plausible and decision-relevant alternative futures that cover the cone of uncertainty and can use information about possible changes to scenario drivers to help users of the system update their models and identify new plausible futures as they emerge. As the code for SAAM has
been open sourced there is a variety of future research that can be conducted using these methods, some of which are detailed in the next chapter.
Discussion of Limitations and Future Research

In this dissertation, I showed how NLP can be used to identify scenario drivers and their interdependencies from diverse knowledge sources using a Q&A method, how such drivers can be modeled in an FCM and used for simulation to generate alternative future scenarios, and how information about possible changes to scenario drivers can be collected through text-based automation and used to update scenarios. I accomplished this using the DSR framework by creating a technical artifact, SAAM, and testing it from an experimental perspective against external data sources. Such experiments are a necessary first step towards implementing SAAM in the context of an agile product development team but cannot provide rich data for understanding several aspects of interaction with the system.

In future research we can study how often teams feel the need to question and update scenarios. In previous chapters I discussed that scenarios are typically used before projects start to set strategic direction (Gausemeier, Fink, and Schlake 1998; Derbyshire and Giovannetti 2017; Nissen, Pretorius, and Klerk 2017; Fasterholdt et al. 2017; Schuh et al. 2014; Betten et al. 2019) but are not typically updated as the project evolves, and are not used continuously by the development team. This presents a gap in the literature we could further explore to better understand if scenarios were used continuously, how this would work and how teams interpret data over time. Additionally we can study how incorporating scenarios with SAAM impacts their team processes, particularly the creation of a shared product vision (Zasa, Patrucco, and
Pellizzoni 2021), strategic focus, and perceptions of uncertainty (Hannay 2021). It could also be important to understand the perceived trustworthiness of SAAM because trust in AI is a topic of growing importance as more AI is adopted in our daily lives. Where an inherent trust in AI has been shown in decision-making (Nguyen et al. 2018) it is possible that when that trust is shaken people may simply stop using it (Stephanie M. Merritt 2011), thus missing out on the potential benefits. Understanding how trustworthy (or not) results are would further the knowledge of collaborative systems and inform the overall usefulness of SAAM. Future research that follows DSR can address these questions through observational techniques in a series of case studies that monitor the use of SAAM on multiple projects. Such field studies can help us understand the quality of scenarios as perceived by teams that use them or any potential benefits they gain from using scenarios.

Moreover, the work did not investigate the process of sensemaking in real-time, such as how people interpret weak signals or future scenarios when they first emerge and there is no hindsight. Instead, experiments in previous chapters created scenarios from historical data, so the future-facing aspect of them was visible by looking backward. Because scenarios are designed to help better understand the future rather than look backward future experiments can create scenarios with data today and test them on people in different contexts. In Sections 2.2.2 and 2.3 I detailed the negative side-effects of agile processes in product development such as losing sight of the broader environment and suffering from groupthink (Perri 2018; Coyle, Conboy, and
Acton 2013; Przybilla, Wiesche, and Krcmar 2019) and how accelerating the processes to create scenarios could help mitigate them if teams could harness benefits of scenarios already documented in the literature (Collier et al. 2018; Postma, Broekhuizen, and van den Bosch 2012; Derbyshire and Giovannetti 2017; Meissner and Wulf 2013; Relich et al. 2020; Wheatcroft et al. 2019). It would be possible to create scenarios across multiple domains using different data sources to study the use of SAAM on such agile teams in longitudinal studies to test if the same benefits gained in strategic thinking apply to mitigate the side effects of agile teams. We could also test across different teams to understand if scenarios help them create better products. If we did discover an agile, fast-moving team accruing benefits from using scenarios we could test if scenario planning can be incorporated into their development process, and if so how.

Naturally, such investigations would require a massive, multi-year effort and are only possible when SAAM, which is currently a conceptual system, has matured into robust and user-friendly software that can be deployed in different teams and projects. Such a development effort, however, is only warranted if the scenarios generated with SAAM are perceived to be plausible and useful for decisions. In the experiments in chapter 6, I have presented evidence in support, but this was largely based on the concept of face validity – the scenarios seemed reasonable and useful to the researcher and in comparison to other studies on the same topic. What is needed next is to investigate the quality of the scenarios with people. In the following sections, I outline how such research could be undertaken.
7.1 Background - Measuring Scenario Quality

Scenarios are a tool used in the inherently social process of strategic thinking and measures of quality would be held to a subjective as well as an objective standard when being used in context. In Chapter 6 I outlined the different measures of scenario quality, the most common being plausibility followed by internal consistency and relevance. Chapter 6 used technical measures but by testing the use of SAAM on product development teams we can also test relevance and plausibility from a user’s perspective. The list of measures for scenarios found in the literature are detailed below in Table 27.

*Table 27 - Scenario validation criteria found in the literature*

<table>
<thead>
<tr>
<th>Validation Criteria</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plausibility</td>
<td>(Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007; Durance and Godet 2010)</td>
</tr>
<tr>
<td>Internal Consistency</td>
<td>(Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Bradfield, Derbyshire, and Wright 2016; Burt 2007)</td>
</tr>
<tr>
<td>Relevance/Importance</td>
<td>(Wilson 1998; Heijden 2011; Alcamo and Henrichs 2008; Durance and Godet 2010)</td>
</tr>
<tr>
<td>Coherency</td>
<td>(Durance and Godet 2010; Bradfield, Derbyshire, and Wright 2016)</td>
</tr>
<tr>
<td>Differentiation</td>
<td>(Wilson 1998)</td>
</tr>
<tr>
<td>Creativity</td>
<td>(Alcamo and Henrichs 2008)</td>
</tr>
<tr>
<td>Reflect Uncertainty</td>
<td>(Heijden 2011)</td>
</tr>
<tr>
<td>Produce new perspectives</td>
<td>(Heijden 2011)</td>
</tr>
</tbody>
</table>
Future experiments can integrate SAAM into the product development process and focus on the most referenced types of scenario measurement; namely plausibility, internal consistency, and relevance to understand how users of SAAM perceive them in context.

7.2 Potential Research Questions

One of the bottlenecks in the scenario process that SAAM alleviates is the need for domain experts to do heavy research upfront. In one experiment we could further validate SAAM with the question *are the scenarios generated by SAAM in line with what current domain experts think of the future of their field.* This could include creating scenarios with SAAM and validating their quality through semi-structured interviews with domain experts.

Secondly, if SAAM can generate scenarios in a way that they can be created and used by product development teams we could ask *to what extent are the scenarios resulting from SAAM interpretable and useful to agile product development teams?* – or *To what extent does the use of SAAM impact the strategic focus of agile product development teams?* We could potentially test this through a longitudinal study where we track a team that uses SAAM through the development of a new product or through multiple case studies where we monitor the use of SAAM across multiple teams in different domains.
7.3 Proposed future experiments

The research questions above can be explored in multiple ways, but I have detailed two additional experiments in this section. The first is where we choose a domain, generate multiple scenarios without the help of domain experts including a random set of scenarios, then present them to domain experts to rank the scenarios. If scenarios generated by SAAM are plausible, internally consistent, and relevant domain experts should rank SAAM’s scenarios high in all three categories.

In other experiments, we can create scenarios with SAAM and incorporate them into the ongoing planning process of one or more agile product development teams. This will help us measure if SAAM is useful to teams in an agile environment, what kind of impact scenarios makes to these teams, and if scenarios help teams in the same way across multiple domains.

7.3.1 Experiment 1: Expert Validation

SAAM was designed to speed up the scenario generation process by addressing two major bottlenecks: detailed up-front research by domain experts and lengthy workshops. The first experiment proposed for future research includes validating scenarios generated by SAAM in a certain domain with experts from the same domain. These same experts are relied upon in traditional scenario planning so their opinions and perspectives will help validate the plausibility of SAAM’s scenarios.

After recruiting these domain experts, future research could present them with three sets of scenarios; one output from SAAM and two that have random changes to
the underlying FCM model. The experts could then be asked a set of questions to rank the scenarios by choosing the ones that they find most plausible and asking them what changes they would make to the underlying model to improve it. I hypothesize that scenarios created by SAAM will be the highest-ranked.

Questions will either be answered on a Likert scale where answers range from 1 to 5 where 1 represents the answer “strongly disagree” and 5 represents the answer “strongly agree”. For example, answers for rating questions are demonstrated in Table 28. This scale will be used across both experiments.

Table 28 - Likert scale ratings

<table>
<thead>
<tr>
<th>Answer</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td>1</td>
</tr>
<tr>
<td>Agree</td>
<td>2</td>
</tr>
<tr>
<td>Neither agree nor disagree</td>
<td>3</td>
</tr>
<tr>
<td>Agree</td>
<td>4</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>5</td>
</tr>
</tbody>
</table>

The first set of questions could be asked to domain experts after they have been shown the generated scenarios along with two randomly varied sets of scenarios. These questions will determine if the generated scenarios match the expert’s view of the domain and give us insight into the quality of the generated scenarios.
Table 29 - Ranking question

<table>
<thead>
<tr>
<th>#</th>
<th>Question</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rank each set of scenarios from most plausible to least plausible, based on your expert understanding of the domain.</td>
<td>Rank</td>
</tr>
</tbody>
</table>

Once the scenarios are ranked future research could probe further, as demonstrated in the next set of questions in Table 30.

Table 30 - Post-observation interview questions for expert team

<table>
<thead>
<tr>
<th>#</th>
<th>Question</th>
<th>Type</th>
<th>Measuring</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Given where we are today, there is a conceivable path to this future.</td>
<td>Likert</td>
<td>Plausibility</td>
</tr>
<tr>
<td>2</td>
<td>This scenario does not contradict itself in any way.</td>
<td>Likert</td>
<td>Internal consistency</td>
</tr>
<tr>
<td>3</td>
<td>If this scenario does contradict itself, please explain.</td>
<td>Open</td>
<td>Internal consistency</td>
</tr>
<tr>
<td>4</td>
<td>These scenarios cover all uncertainties that exist in the given domain.</td>
<td>Likert</td>
<td>Relevance</td>
</tr>
<tr>
<td>5</td>
<td>The scenarios presented are easily interpretable.</td>
<td>Likert</td>
<td>Interpretability</td>
</tr>
</tbody>
</table>

For each question of rating type, additional open questions can be asked to understand more about why a subject answered the way they did. For example, if a question is answered with a “1”, open questions allow me to explore the reasons why they do not agree.
The outcome of this experiment could help align the output of SAAM with an expected output from someone who is an expert in a particular domain. As domain experts should have a front-row seat to the changing landscape they work in, the ability for SAAM to find insights that such experts find useful would be a strong validation of its ability.

7.3.2 Experiment 2: Product Team Feedback

The second future experiment includes observing an agile team as they use the generated scenarios in their planning process to understand how useful scenarios are to them and what impact scenarios have on the strategic thinking of an agile team. Though it has been shown that scenarios help teams understand driving forces and identify emerging trends that help prioritize the highest-value work (van Notten 2006), this has not been demonstrated in an agile environment to my knowledge. Future research could conduct a longitudinal study where generated scenarios as narratives are shown to an agile team working on a new product, then study if this affects how they perceive their environment and prioritize their work moving forward. This could be done by asking them to draw mental models of their understanding of the domain as FCMs and tracking differences to them over time. It is possible that after working on a product for some time a team will naturally start to learn more about its domain, so to mitigate this it would be possible to split the team into two different groups: an experimental group that uses scenarios (Group A) and a control group that does not (Group B). Before showing the experiment group scenarios each member of the team could be asked to
draw their understanding of the domain their product is being developed in as an FCM. After each development iteration, the experiment group can be given generated scenarios related to the product they are developing to read as narratives, then ask them to draw their understanding of the model again. This data will demonstrate how each person in the experiment group models their domain at the beginning of the study and over time as they are exposed to the generated scenarios which can be compared to the maps drawn by the control group to measure the difference. The FCMs will be used as quantitative models that can be compared over time to determine how they change when the subjects are exposed to scenarios and to measure how closely their models map to the FCM that were used to generate the narratives. I hypothesize that the models drawn by the participants in the experiment group will change over time to move closer toward the FCM used to generate the scenarios shared with the team and will show a greater depth in understanding their domain and thus their ability to think more strategically within it.

Finally, future research could interview the agile team to collect their opinion on if they found the scenarios interpretable, useful, and if they had a positive impact on their strategic focus. Each member of the team that is incorporated into SAAM’s process could be asked a set of structured questions with open-ended follow-ups to add depth to their answers. The data collection phases are demonstrated visually below in Figure 47.
In this experiment, interviews include asking the participants to draw a mental model of their understanding of their domain and is intended to measure how participants change their point of view over time while using scenarios created by SAAM. The final set of questions is set up as an exit interview for participants from the agile team to understand their opinions on how scenarios affected their strategic thinking.

First future research could split the team into two groups; Group A and Group B. Each participant of both groups would receive detailed instructions and examples on drawing mental maps as FCMs. Then before showing them any scenarios ask them to draw their understanding of the domain their product is a part of as an FCM to get a starting point for each participant. Before the next development iteration, future research could show the scenarios to the team members in Group A as narratives before they prioritize their work, then ask them to redraw their understanding of the domain. The members of Group B will not see any scenarios but will also redraw their understanding of the domain as product development continues. This process could be repeated throughout several development iterations to understand how their mental models change over time and the difference in any changes between Group A which
uses scenarios and Group B which has no exposure to scenarios. Future research could also track updates to nodes, connections, and weights for each participant over time and the similarity of their models to the original FCM used to generate the scenarios given to Group A. This data will show if scenarios affect people’s mental models over time, the difference between those who use scenarios and those who do not, and how similar models drawn by participants are to the quantitative representation of the narrative scenarios.

After collecting data over time in Phase two future research could conduct an exit interview with the participants from Group A on the agile team to understand their opinion on the usefulness of using scenarios. The questions for the exit interview are shown below in Table 31.

*Table 31 - Agile team exit interview questions*

<table>
<thead>
<tr>
<th>#</th>
<th>Question</th>
<th>Type</th>
<th>Measuring</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The scenarios help me make better decisions in developing this product.</td>
<td>Likert</td>
<td>Relevance</td>
</tr>
<tr>
<td>2</td>
<td>The scenarios presented are easily interpretable.</td>
<td>Likert</td>
<td>Interpretability</td>
</tr>
<tr>
<td>3</td>
<td>The scenarios are useful to me to help create a better product.</td>
<td>Likert</td>
<td>Relevance</td>
</tr>
<tr>
<td>4</td>
<td>The scenarios are useful to help me better understand the environment my product is being developed for.</td>
<td>Likert</td>
<td>Relevance</td>
</tr>
<tr>
<td>5</td>
<td>The scenarios help me maintain a sharper strategic focus.</td>
<td>Likert</td>
<td>Relevance</td>
</tr>
</tbody>
</table>
7.3.3 Analyzing and Reporting Results

For both experiments, survey results will be aggregated and charted. Because the second experiment would capture mental maps of the participants, the same methods used in Chapter 6 could be used to view the network metrics of each subject along with their semantic distance from one another.

The open-ended exploratory questions could be reviewed in detail for insights on why subjects answer questions the way they did. Insights will be reported in detail along with the charts for the corresponding questions to show not only the overall ratings but the reasons behind them.

Plausibility, relevance, and internal consistency could be measured with the first set of interview questions collected from domain experts. Usefulness, interpretability, and the ability to impact strategic thinking will be measured by tracking the changes in Group A’s mental models over time, as compared with the control Group B’s mental models. Finally, the exit interview is intended to get the opinions of those on the team that used the scenarios.

7.4 Conclusion

I have outlined two additional experiments in this chapter that can be used to validate SAAM using DSR, one using domain experts to validate their quality and one using agile teams to study the use of SAAM in context. SAAM is a viable, modern system for assisting with the help of creating scenarios that cover many potential futures in the
cone of uncertainty. SAAM can be used in several different ways for future research, in particular as a tool to assist anyone who is embarking on a scenario-driven research project as demonstrated in this dissertation. Because SAAM dramatically speeds up the scenario process there is also the potential for future research around how product development teams, agile teams, or others can incorporate scenarios into their work and the benefits they gain from using them.
8 Contributions

8.1 Contributions

Based on the outcomes from this research, the following sections discuss the theoretical, methodological, and practical contributions.

8.1.1 Theoretical Contributions

8.1.1.1 Moved the Academic Research on Scenario Planning

Today over half of the publications on scenarios are regarding methodology (Tiberius, Siglow, and Sendra-García 2020) and this work contributes to such literature with a novel method that incorporates AI. Practitioners can use the steps demonstrated in this research to create scenarios faster regardless of whether they use the artifact created for this research. This novel method helps speed the long workshops normally used and reduce the need for experts, which are two known shortcomings of scenarios (Schoemaker, Day, and Snyder 2013; Rafael Ramirez et al. 2017; Hodgkinson et al. 2006; Franco, Meadows, and Armstrong 2013).

This research also furthers the state of the art in creating scenarios faster with the help of NLP. There have been other approaches to creating scenarios with the help of NLP using LSA and LDA (J. Kim et al. 2016; Kwon, Kim, and Park 2017; Kwon and Park 2018; Kayser and Shala 2020; Gokhberg et al. 2020; Feblowitz et al. 2021), and I build on this research by incorporating the latest in ML technology with BERT (Devlin et al. 2019).

Finally using FCM for quantitative scenarios is an emerging practice, (Amer, Jetter, and Daim 2011; A. Jetter and Schweinfort 2011). This research furthers this by
demonstrating the creation of FCM through a socio-technical system specifically for the use of scenario planning.

8.1.1.2 Moved the Academic Research on FCM

While FCM is by no means new (Kosko 1986) there is no standard way to generate FCM from data. While FCM have been built using various methods such as Hebbian learning, hybrid learning, and error-driven learning algorithms (Felix Benjamín et al. 2019) to the author’s knowledge-creating FCM using Q&A-based techniques is a novel approach. This research contributes to the FCM literature by demonstrating how they can be created using Q&A-based AI to identify concepts in data and the connections between them.

FCM are a popular decision support tool (C. W. Davis and Jetter 2021) thanks to their relatively simple ease of use and significant predictive power (Elpiniki I. Papageorgiou and Salmeron 2013; A. J. Jetter and Kok 2014; Glykas 2010; Felix Benjamín et al. 2019). This research also advances the state of the art in FCM by providing a demonstration of how FCM can be used for decision support, specifically in the case of strategic thinking.

8.1.1.3 Defined a Novel Approach to Create Scenarios with IML

The practice of IML (Fails and Olsen Jr 2003; Amershi et al. 2014) has been shown to create better models than ML alone while increasing people’s understanding of the model (Robert et al. 2016). This research furthers the field of IML by showing that SAAM, a system coordinated through IML, was able to create scenarios that are at least
as good as those created with other methods that use automation and demonstrates a practical example of implementing an IML workflow to create scenarios through a socio-technical system.

This approach could be integrated into any team or can be adopted in traditional scenario planning to help accelerate the process in a transparent and traceable way. Using Q&A with AI shows that we can utilize NLP-based automation in a very human way to define and use scenarios.

8.1.2 Methodological Contributions

8.1.2.1 Combined Horizon Scanning with FCM to Find Plausible Scenarios

HS scans large amounts of data to find weak or emerging signals which can be used to aid in decision making (Yoon 2012; Gokhberg et al. 2020; Konnola et al. 2012). To my knowledge combining HS with FCM to find plausible scenarios by associating signals found in real-time data with concepts in an FCM is a novel method that can help users of scenarios regardless of how they are created, helping them find new plausible collections of inputs for quantitative models based on data from current events. This method also contributes to the literature on HS by demonstrating this novel application.

8.1.2.2 Building FCM using Q&A-based AI From Disparate Data Sources

FCM have been created in previous studies from data including using NLP (Son, Kim, and Kim 2019; Kwon and Park 2018; Sandhu et al. 2019) through conversation interfaces (Reddy, Giabbanelli, and Mago 2019; Reddy, Srivastava, and Mago 2020; Anjum et al. 2021), from summaries (Villalon and Calvo 2011), and large numbers of
documents (Hajek, Prochazka, and Pachura 2017; Son, Kim, and Kim 2019; Pillutla and Giabbanelli 2019) but to my knowledge-creating FCM with Q&A based AI from disparate data is a novel way to create quantitative modes from data. By combining people and AI this method allows people to generate mental models with reference data and questions that represent what they want to learn from modeling it.

8.1.3  Practical Contributions

8.1.3.1  Created a practical tool for creating scenarios

One of the gaps this research identified was because of limited expert availability, time, resources, and the expertise needed to interpret and use scenarios, current techniques for scenario planning are inaccessible for most product development teams. SAAM as a tool can be used by bringing your data and set of questions. The use of scenarios in product development is not common practice today, even though it has been identified as a needed component (Kazmi, Naarananoja, and Kytola 2015) and has been proposed as an addition to the product development process (Derbyshire and Giovannetti 2017; Randt 2015). Despite its benefits scenario planning has been shown to take too long to fit into short, iterative cycles and are not granular enough to support tactical decisions (Leau et al. 2012; Sahoo and Pattnaik 2020; Gabriel et al. 2021). This research addresses this gap by publishing SAAM as an open-source system that could allow any person or team with access to data to create their scenarios today.
8.1.3.2 Demonstrated the Use of Quantitative Scenarios on Public Data

FCM has been used to demonstrate scenarios in several domains from autonomous vehicles, natural resource production and allocation, energy development, and new product development to name a few (Elpiniki I. Papageorgiou and Salmeron 2013; Felix Benjamin et al. 2019). However, to my knowledge FCM has not been demonstrated as a tool to analyze financial reports of public companies to create scenarios that could help with strategic direction. The use of 10-K documents is a common way for people to gain an understanding of a publicly traded company quickly and by using this method anyone can quickly build a quantitative model of a company’s view of the world.


Betten, Thomas, Raed Bouslama, Daniel Wehner, and Ville Uusitalo. 2019. Unlocking Sustainability Potentials in Product Development through Extended Knowledge and Predictions about the Product Use Phase.


