2020

Natural Language Processing for Modeling Domains in Higher Education

Rebecca Leis
University of Central Florida

NATURAL LANGUAGE PROCESSING FOR MODELING DOMAINS IN HIGHER EDUCATION PLANNING

by

REBECCA ASHLEY LEIS
B.S. University of Central Florida, 2011
M.S. University of Central Florida, 2012

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the School of Modeling, Simulation, and Training
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2020

Major Professor: Bruce Caulkins

ABSTRACT

The present dissertation investigates elements of domain formalization, resource allocation, and student success in higher education to conceptually design a university-wide system to assist in strategic planning efforts. The proposed system is a program-level tool with a modular design to allow scalability and generalizability across the entire university. Higher education strategic planning decisions are under investigation by stakeholders and transparency is needed. University resources allocation models are often outdated lack to adequately support program-level decisions. Further, with the dynamic nature of technology, domain knowledge components are evolving rapidly. This complicates the situation as updating curriculum takes additional time and resources.

Using the University of Central Florida's (UCF) School of Modeling, Simulation, and Training (SMST) as a case study to build and validate the system, I investigate Modeling and Simulation (M&S) domain knowledge, skills, and abilities (KSAs) using a series of natural language, text mining, and machine learning techniques to model topics within domain-specific texts including publication abstracts, job postings, and graduate course descriptions. From there, I use this information to identify and enumerate terms used to develop M&S ontology and expert models for the future university-wide system. This investigation benefits both the M&S field of study, clarifying ill-defined domain components and it helps inform the design of university-wide strategic planning systems.

ACKNOWLEDGMENTS

First, I would like to acknowledge and thank my committee members, Dr. R. Paul Wiegand, Dr. Sabrina Gordon, and Dr. C. Richard Hartshorne for agreeing to guide me as I fumbled my way through this process. One of the challenges I faced was the scope of my project. I appreciate that each of you expressed your concerns honestly but still allowed me to explore this rabbit-hole of a project. The enormity of my idea allowed me to investigate M&S in a way that I had not experienced in my coursework. Next, I would like to thank my Committee Chair, Dr. Bruce Caulkins. Bruce, you have been a wonderful, kind, and understanding human being to me during this entire process no matter how many times I missed a deadline. Further, you built me up when I was frustrated with my progress. I am forever thankful for your patience.

I would like to thank John Lord and Dr. Patricia Bockleman in helping me formulate some of the earlier efforts for this project. Some general thanks also go out to the staff and faculty at the Institute for Simulation and Training that influenced my career path thus far (especially Karla Badillo-Urquiola and Eileen Smith).

I want to thank my parents Thomas Leis and Katherine-Epstein-Leis for providing me opportunities in life for me to get to this point. You are amazing parents. I know it wasn't easy and kids aren't always the most appreciative human beings. Thank you for taking an interest in my education. I know my love of life-long learning and creativity came from you two. I love you both.

Next, I would like to acknowledge and thank my husband, Kyle Lucas. Kyle is my rock. When I feel crazy, he makes me sane. When I'm upset, he always makes me laugh. When I don't believe in myself, he lifts me up higher. Without him there ridiculously saying "I'm good

enough, I'm smart enough, and people like me," I would not have gotten through this process. You are my soul mate and I love you with all my heart.

Special thanks are needed for Dr. Sabrina Gordon and Dr. Avonie Parchment for being my accountability partners during this voyage. They were always there to tell me "done is better than perfect," to talk through mental blocks, and to lend an ear when I needed to vent. You are both (DA) unicorns! I also want to express gratitude for my closest friends and amazing support system: Christina Valdez, Steven Valdez, Tabie Wong, Garrett Effers, Leigh Faircloth, and the best brother ever - Christopher Leis! These people keep me grounded. They never allow me to take myself too seriously.

Further, I would also like to that my co-workers over at Full Sail University for having the patience to deal with my crazy schedule during this process. All of you have been very supporting, checking in frequently and making sure I'm doing dissertation after work hours and not more work. Thank you for supporting my journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS (OR) ABBREVIATIONS

ABD – All but Dissertation

AI – Artificial Intelligence

ARL – Army Research Laboratory

BoK – Body/Book of Knowledge

BoN – Bag-of-n-grams

BoW – Bag-of-Words

CMSP – Certified Modeling and Simulation Professional

DoD – Department of Defense

DTM – Document Term Matrix (Matrices)

FT – Full Time

GPA – Grade Point Average

GRE – Graduate Record Examination

GUI – Graphical User Interface

KSAs – Knowledge, Skills, and Abilities

LDA – Latent Dirichlet Allocation

LSA/I – Latent Semantic Analysis/Indexing

M&S – Modeling and Simulation

M&SPCC – Modeling & Simulation Professional Certification Commission

NACUBO – National Association of College and University Business Office

NLP – Natural Language Processing

NLTK – Natural Language Tool Kit

O*NET – Occupational Information Network

OWL – Web Ontology Language

pLSA – Probabilistic Latent Semantic Analysis

PD – Pandas

PMBOK – Project Management Body of Knowledge

PT – Part Time

PoS – Parts of Speech

Re – Regular Expressions

RDF – Resource Description Framework

SKLearn - SciKitLearn

SMST – School of Modeling, Simulation, and Training

TD-IDF – Term Frequency-Inverse Document Frequency

TTD – Time to Degree

UCF – University of Central Florida

UML – Unified Modeling Language

URL – Uniform Resource Locator

XML – Extensible Markup Language

CHAPTER ONE: INTRODUCTION

In this study, I present the basis for, and conception of a university-wide scalable simulation system created for higher education organizations to automate topic modeling, guide student advising, and maximize student success. To appropriately scope the present dissertation, this chapter will first discuss the basis for the study, conceptualization of the overall system, and then focus on the design and validation of the main component within the system, the domain model. Further development and evaluation of the overall system will also be addressed as design considerations for the entire system.

Background of the Problem

Overview of Strategic Planning in Higher Education

A university's sustainability relies on its ability to effectively plan and allocate resources that meet the strategic goals of the institution. However, stakeholders (e.g., students, government agencies, accreditation boards, etc.) expect universities to keep costs minimal, increase student populations, and improve the quality of higher education simultaneously (UCF Board of Trustees, 2016). These competing objectives make it difficult for university management to divide resources *fairly* among many different programs, each providing varying (and sometimes subjective) levels of value to the university (Kershaw & Mood, 1970). Various stakeholders also want transparency and articulating subjective value can be difficult (Anti-Corruption Risk Assessment Taskforce, 2013; Council of Chief State School Officers, 2017). I believe that

formalizing these systems at multiple levels can help visualize relationships and articulate value within the university.

Overview of Graduate Student Success

Only 40-60% of doctoral students in the United States persist to degree completion, a rate that has remained relatively unchanged for half of a century (C. H. Bair & Haworth, 2004). This is because graduate programs are less structured than undergraduate programs. This system relies on the knowledge, skills, and abilities (KSAs) of the faculty advisors. The relationship between the student and advisor has been shown to be the strongest factor affecting attrition and persistence in doctoral programs (C. H. Bair & Haworth, 2004). One of the goals for the overall system designed in this study is to help determine a student's optimal path for *skill* and *knowledge* acquisition. Therefore, it is important to find out the types of skills and knowledge needed by experts to model our student after. *Abilities* are somewhat fixed and should be factored in when determining appropriate variables for both student and expert profiles/models. It is assumed that students with similar abilities (or internal factors) to experts (e.g., faculty) will succeed in similar careers if given an individualized plan for skill and knowledge acquisition. Another possible future direction for this system may be to help match students with potential advisors.

Overview of Domain Organization

Hiring faculty that can teach domain topics in rapidly changing technology-based fields can also be difficult if standard KSAs are ill-defined. Domains are fields of study and an ontology represents the structure of the domain (e.g., classes, relations, functions; Gruber 1993). A visual representation of the domain can be useful for articulating information and designing

curriculum standards. It can also be used to help software designers, developers, and evaluators model knowledge. A *programmed* ontology is necessary for expert systems or intelligent and adaptive tutors. While building an intelligent or adaptive tutor is outside of the scope of this project, the work presented here is intended to support a system like an adaptive tutor in the future.

Problem Statement

Transparency and accountability are increasingly important to higher education stakeholders; thus, as highly complex systems they must showcase their value to sustain. Metrics of success are necessary to articulate this value but are not always quantitative and explicit. This ambiguity is compounded by the fact that technology-related programs evolve quickly, which makes it difficult for faculty and administrators to determine (and quickly update) appropriate curricula to prepare students for the job market.

Dissertation Purpose

The purpose of this dissertation is to investigate natural language in M&S domain-specific job postings (i.e., what employers request), course descriptions (i.e., what is being taught), and academic literature (i.e., what is applied in practice) to enumerate important terms and determine common relationships between topics, holistically and by job type. Then, using natural language processing techniques to study this qualitative data, topic models were developed to provide data-driven recommendations for graduate program strategic planning.

These recommendations can address needs related to student advising, course planning, faculty hiring, and relevant research directions.

To do this, I will next introduce the subject for this dissertation study. Then, the conceptual model for a potential software tool will be presented to show how it is designed to support program-level decisions and university-level strategic planning goals.

Problem Context

The School of Modeling, Simulation, and Training (SMST) at the University of Central Florida (UCF) is used as a basis for conceptualizing the domain model needed for the system. SMST at UCF was selected because of its location and structure. Orlando, Florida is currently a Simulation Center of Excellence, allowing access to many M&S stakeholder groups both within and outside of the academia. Access to stakeholders will help with verification and validation in future iterations. Further, I have focused on graduate education first, using a backwards design approach. Additionally, SMST is not associated with a matching undergraduate program, consequently allowing for a less convoluted investigation of resources for a *graduate* degree program. Finally, M&S is uniquely positioned to use its own techniques to improve the overall system. SMST is intended to be a *starting point* for determining appropriate student-centered modeling techniques for managing resources and strategic plans within graduate schools-programs.

Simulation Framework/General Approach

There are four levels of detail identified for the project's problem context: 1) strategic planning in higher education at the university level, 2) graduate education and curriculum at the college level, 3) Modeling and Simulation (M&S) domain standards at the program level and 4) the student at the center. Each of these topics is related to each other through various levels of scope/granularity. Specifically, strategic planning in higher education is the highest, most abstract, macro-level of the problem space, and the lowest, least abstract, micro-level is the student level. This relationship is illustrated in Figure 1 below.

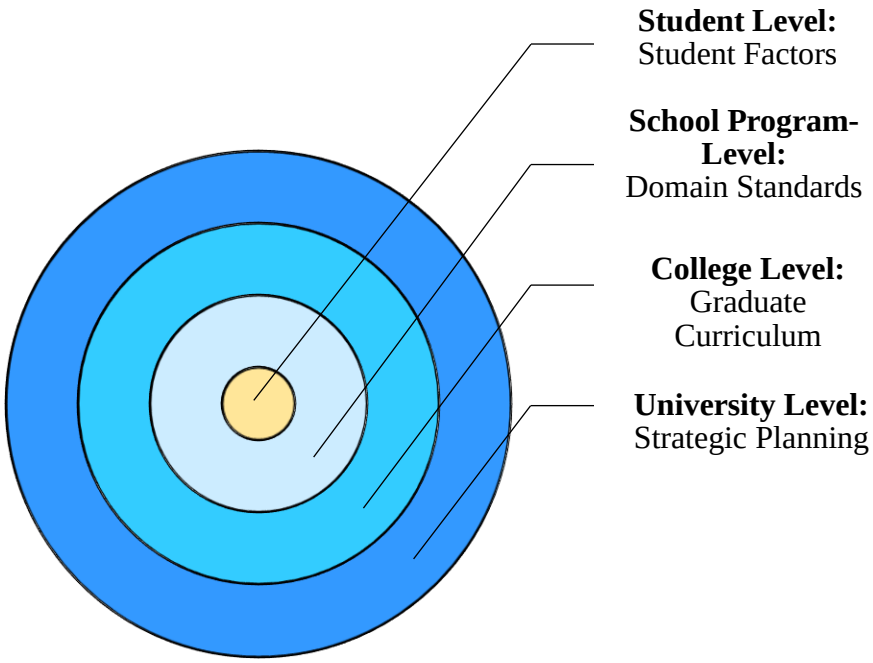

Figure 1: Problem Context Level of Detail

Universities house several colleges. Each college houses related degrees based on similar fields of study. Colleges are degree-granting institutions. SMST is categorized under UCF's

College of Graduate Studies. Each college houses schools, which include departments and programs. In this instance, SMST only houses one program (Modeling and Simulation), so it doesn't have any departments. As such, in this document "school-program level" will be used to include schools, departments, and programs.

The level of detail is important to the context of the present dissertation. While, university-level resource allocation drives changes for the lower levels, less emphasis will be given to the "higher-level" context in the present document. Many universities approach strategic planning efforts using a top-down approach, focusing heavily on the university level decisions, moving downward. However, UCF also calls for program-level strategic planning models (UCF Board of Trustees, 2016). This is the inspiration for the overall system.

There are many interconnected parts following along with the four levels of detail identified in Figure 1, all of which should be considered for the overall system. The framework I present here (Figure 2) is Law's (2003), seven-step approach, which is used to guide the design, development, and assessment of the overall strategic planning simulation system.

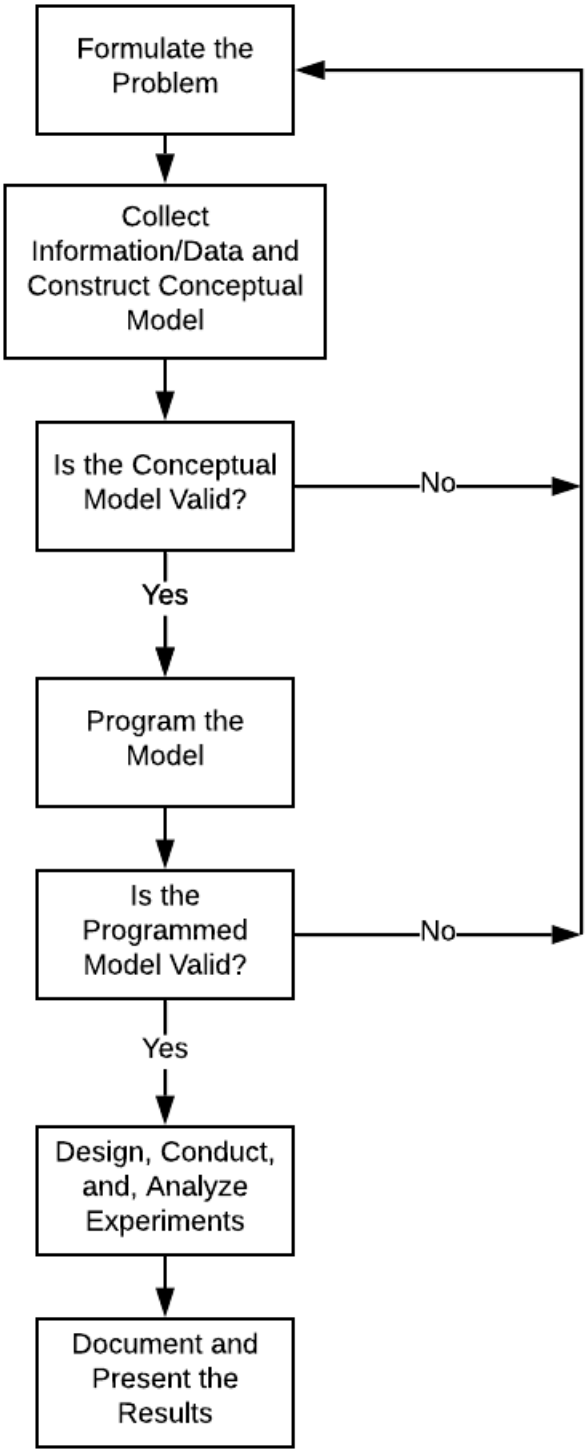

Figure 2: Law's (2003) Seven Step Simulation Study Framework

The overall system will require a few iterations. As such, I will conduct this dissertation study to focus on the first two phases of Law's (2003) High-Level Research Agenda Seven Step Framework (see Table 1) applied to Sottilare's (2015), recommendations for adaptive systems, specifically concentrating on the *domain* model.

Table 1: High-Level Research Agenda based on Law's (2003) Seven Step Framework

Phase 1: System Formation and Conceptual Modeling
- Problem Formation
  - Determine overall goals and objectives of the project
  - Establish *system* and *model* scope
  - Identify appropriate stakeholders
- Collect information
  - Collect information from existing system (if applicable)
  - Identify system configurations
  - Determine system assumptions
- Construct overall conceptual model

Phase 2: Validation of Conceptual *Domain* Model
- Determine specific research questions
- Select appropriate measures for the research questions
- Collect data
- Clean data
- Analyze data
- Visualize data
- Create domain model
- Validate domain model

Phase 3: Validation of Conceptual Student Model
Phase 4: Validation of Conceptual Instructional Design Model
Phase 5: Validation of Conceptual Resource Allocation Model
Phase 6: Validation of User Interface Design
Phase 8: Program and Validate Domain Model
Phase 9: Program and Validate Student Model
Phase 10: Program and Validate Instructional Design Model
Phase 11: Program and Validate Resource Allocation Model
Phase 12: Program User Interface and Validate User Experience
Phase 13: Integrate Overall System Model
Phase 14Validate Simulation
Phase 15: Design and Conduct Strategic Planning Experiments
Phase 16: Report Simulation Results

(Adapted from Law, 2003 & Sottilare, 2015)

Natural Language Processing

To specifically address developing and evaluating the domain model I plan on using natural language processing (NLP), which uses machine learning and statistical techniques to analyze, model, and comprehend human language (Vajjala et al., 2020). Complex systems (e.g. a university wide strategic planning systems) are difficult to build because computers use binary logic, meaning we often have to simplify characteristics of the problem to model them in a way a computer can understand (*Computer Logic vs. Human Logic*, 2018). One common NLP task includes *topic modeling*, which can be defined as "uncovering the topical structure of a large collection of documents," (Vajjala et al., 2020). NLP vectorizes qualitative data in a way that allows computers to perform statistical analyses. I chose to use NLP techniques because they can easily provide data-driven context and meaning to text data (Vajjala et al., 2020).

Definition of Terms

- **Ability**: "a basic capacity for performing a wide range of different tasks, acquiring a knowledge, or developing a skill," (Aamodt, 2010, p. 53).
- **Domain:** a field of study (Gruber, 1993).
- **Knowledge:** "a body of information needed to perform a task," (Aamodt, 2010, p. 53).
- **Model** "a representation of something else," (Sokolowski & Banks, 2009, p.122)
- **Modeling and Simulation**: "a unique discipline that is concerned with understanding and exploring complex problem situations (either real or imaginary) as a basis for training, entertainment, and/or experimentation," (adapted from Gupta & Grover, 2013; Ören, 2014; Padilla, Diallo, & Tolk, 2011).

- **Natural Language Processing**: "an area of computer science that deals with methods to analyze, model, and understand human language," (Vajjala et al., 2020)
- **Ontology:** "[a] specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects," (Gruber 1993).
- **Simulation:** "a method for implementing a dynamic model over time," (adapted from Ören, 2005 and Ören, 2011; Sarjoughian & Zeigler, 2000)
- **Skill:** "the proficiency to perform a learned task," (Aamodt, 2010, p. 53).
- **Topic Modeling:** "This is the task of uncovering the topical structure of a large collection of documents. Topic modeling is a common text-mining tool and is used in a wide range of domains, from literature to bioinformatics," (Vajjala et al., 2020).

Research Objectives

The research objectives are identified to help scope the *university-wide system* and are discussed to varying degrees of specificity within the document.

- *Objective 1:* Investigate current resource allocation and strategic planning models in higher education, automated curriculum management, and graduate student success factors to determine an appropriate plan for designing and developing a holistic university-wide software system.
- *Objective 2:* Conceptualize a university-wide modular decision-making software solution to inform curricula, optimize student paths toward degree completion, and optimize resources to meet program and university objectives.

- *Objective 3:* Develop an M&S ontology using topic modeling techniques from all source types (job listings, course descriptions, and academic publications) and compare each source type to determine if there is a disconnect between requested, taught, and applied KSAs.
- *Objective 4:* Develop M&S expert models using topic modeling techniques.

Conclusion

In summary, educational value is increasingly important to university stakeholders, but current metrics of success are not always quantitative and explicit. Further compounding the issue, technology-related programs evolve quickly, which makes it difficult for faculty and administrators to determine (and quickly update) appropriate curricula to prepare students for the job market. The purpose of this dissertation is to investigate natural language within various domain documents (e.g. job postings, course descriptions, and academic literature) using natural language processing to determine common relationships and model topics to make recommendations for relevant research directions during graduate program strategic planning. The M&S domain is used as a case-study as it is a nascent and ill-defined domain needing a formal ontology. The following chapter presents the background literature for the problem context and the proposed conceptual model theorized based on this literature.

CHAPTER TWO: LITERATURE REVIEW

Faculty and administrators can have a difficult time determining and prioritizing appropriate topics for curriculum, due to limited metrics of success and the speed at which technology is evolving. Modeling and Simulation (M&S), a nascent and ill-defined domain needing a formal ontology, is used as a case-study for the present dissertation. The purpose of this dissertation is to investigate domain-specific documents using natural language processing to enumerate terms and determine topic prioritization and model common relationships between M&S topics. This data-driven method can be applied to make recommendations for relevant research directions during program strategic planning. Therefore, a review of the relevant literature is presented in this chapter on strategic planning in higher education, success factors that affect graduate faculty and students' independent research and curriculum topic choices, the current state of M&S, and methods for automating complex systems.

Strategic Planning in Higher Education

The following section details various components of strategic planning related to higher education including institutional goals and values, resource allocation, and curriculum mapping.

Strategic Goals

To design a university-wide strategic planning system, one should consider the university's strategic goals. UCF's Collective Impact: Strategic Plan (UCF Board of Trustees, 2016), reports UCF's strategic planning initiatives for the university. It includes information such as the goals and values of the institution and the metrics and strategies for reaching said goals.

This information is used to determine appropriate outcomes for the university-wide system. While the goals presented here are UCF specific, these types of values, metrics, and strategic plans are not unique to UCF. Accrediting organizations "require documented evidence that all activities using institutional resources support the institution's mission," (Hinton, 2012).

However, UCF's strategic plan is limited, as the goals outlined in the document directly apply to the overall university and do not address decision-makers at the program-level. I adapt the information in the strategic plan so that it applies to program-level objectives and outcomes. Based on UCF Former-President Hitt's five goals for the university, I've crafted recommended SMST specific goals shown in Table 2.

Table 2: Proposed SMST Program Level Goals

1. To build international prominence in M&S
2. To increase international focus to M&S curricula
3. To increase international focus to M&S research programs
4. To become more inclusive and diverse
5. To strengthen existing partnerships within the university

(Adapted from UCF Board of Trustees, 2016)

The UCF Board of Trustees (2016), plans to address the university's overall goals using 25 specific categories of metrics and strategies, (see strategic plan for full list). In Table 3, I summarize the metrics and strategies related to the proposed program level goals.

Table 3: Applicable Graduate Program Level Metrics

Graduate Student Prominence
• Double the number of graduate students receiving national or international recognition
• Expand [from 8,029] to 10,000 graduate students [about 20% increase]
Faculty Prominence
• Double the number of faculty members receiving national and international recognition in their fields
• Reach 1,200 full-time tenured and tenure-track faculty members
• At least 65% of all faculty members with assigned instructional duties are tenured or tenure-track
Faculty and Staff Diversity and Inclusiveness
• Achieve 25% in employment of under-represented groups among tenured and tenure-track new hires who are retained five or more years
• Achieve 30% in gender diversity in STEM fields among tenured and tenure-track new hires who are retained five or more years
• Achieve 25% in employment of under-represented groups among full-time administrative and professional new hires who are retained five or more years
Student Diversity and Inclusiveness
• Increase by 10% retention and progression of specific diverse student cohorts across all academic disciplines
• Increase by 10% degree attainment of specific diverse student cohorts across all academic disciplines
Research Engagement
• Achieve level at which at least 25% of graduate degrees awarded are research-focused
• Reach at least 200 post-doctoral appointees [at the time the strategic plan was published it was at 52]
Research and Commercialization Commitment
• Double research awards from $133M to at least $250M
• Win ten proposals per year exceeding $1M, five of which exceed $3M
• Create 16 start-up companies annually and execute 36 licenses and options for UCF intellectual property
• Achieve 200 patents awarded over three years
Research Collaborations
• Generate 30% externally funded research expenditures through collaborations with other institutions
• Generate 60% externally funded research through collaborations within UCF
Cost Management
• Develop metrics for fiscal stewardship within each department and academic unit

(Adapted from UCF Board of Trustees, 2016)

In future iterations of the university-wide system, I plan to use these metrics and strategies to inform analyses that I perceive to be useful to the end-user. Each of these bullet points can serve as a separate user scenario for the system. For example, using the highlighted

bullet point in Table 3, I drafted a user story for the software interface, stating: *as a program administrator, I would like to be able to determine the number of faculty we will need to hire to support a 20% increase in incoming students*. To complete a task like this, the system would need to pull staffing information (e.g., expertise, team knowledge gaps) and course information (e.g., topics currently taught), among other factors to output the simulated data. This is another reason I started with the domain model for the present dissertation.

Defining domain KSAs and organizing them into categories is an enormous endeavor taking significant time, effort, and research, reiterating along the way. Further, many professionals and academics disagree on the composition of core versus specialized KSAs. This proves problematic for educational programs, as this disagreement makes determining where resources should be invested difficult for program directors and administrators. For example, program administrators often have to ask questions like *which courses are most important; what type of expertise do we need*; *how often should courses occur*; and *what types of continuing education opportunities are worth investing in*?

Resource Allocation

One of the main objectives of the university as a system is to allow users to investigate ways to minimize costs at the program level and grow the student population, but not at the expense of quality of education (UCF Board of Trustees, 2016). Determining a program's annual budget is often based on antiquated and simple principals such as *incremental budgeting* and *formula-based allocation*. Incremental budgeting occurs when the university sums the previous year's annual budget and multiplies it by a set percentage to account for inflation costs, but ill-suited for volatile markets (Kershaw & Mood, 1970). Formula-based allocation is more flexible

as funding is based upon program full-time head count or total credit hours (Kershaw & Mood, 1970). While formula-based allocation favors popular programs, it punishes programs that are unpopular, ignoring intangible success measures (e.g., a program's ability to follow ethical practices). Additionally, curriculum mapping has been used to informally track resources based on the students' learning opportunities/activities (Harden, 2001).

Curriculum Mapping

Curriculum mapping is a "blend of educational experiences, assessment, the educational environments and the individual students' learning style, personal timetable, and programme of work," (Harden, 2001). This type of mapping process explicitly organizes curriculum components in a way that allows all stakeholders to easily understand the connections between the content, the assessment, the learning outcomes, the staff responsible, etc., using a student-centered approach.

A promising development in resource allocation and curriculum mapping is activity-based costing in which resources are allocated based on the time it takes to design, develop, and deploy a single learning activity (e.g., exam, lecture, discussion; Massy, 2016). William F. Massy (1996, 2016), emeritus professor, and former Vice President of Business and Finance at Stanford University, has been working on refining the model at the course-level. The latest version of the model is supported by the National Association of College and University Business Office (NACUBO) Economic Models Project (Massy, 2016). While potentially useful for modeling the workload and cost of implementing course activities, learning activities should be determined based on the learning outcomes (Nilson, 2010). Learning outcomes are built around measuring performance in common tasks. They include three components, 1) a statement

of measurable performance, 2) a statement of conditions for the performance, and 3) criteria and standards for assessing the performance (Nilson, 2010). It is the job of the faculty to determine how to apply domain specific principles to these three components to design appropriate course outcomes. Thus, a reasonable starting point for a university-wide system should include some type of domain mode or formalization, but I intend to integrate Massy's work into future iterations of the university-wide system. I present it here to show the importance of domain information in relation to resource allocation and strategic planning in higher education.

Graduate Student Success Factors

In higher education, *student success* is a generic term that can mean many things, however degree completion is the easiest and most common measure of student success. Nevertheless, this is an overly simplified measure of success. Bair and Haworth (2004), identified factors of attrition and persistence in doctoral students across multiple universities, in various programs. Table 4 is a reproduction of the general conclusions drawn by Bair and Haworth (2004), during their meta-synthesis (a combined meta-analysis and meta-ethnography) of over 118 articles.

Table 4: Attrition and Persistence Factors

1. Attrition and Persistence Rates Vary by Field of Study and Program of Study
2. Departmental Culture Affects Doctoral Student Persistence:
   - The degree and quality of the relationship between doctoral student and advisor or faculty has a strong, positive relationship to successful completion of the doctorate
   - Student involvement in various programmatic, departmental, institutional, and professional activities and opportunities contributes favorably to doctoral student retention and completion
   - Students' satisfaction with their academic programs —including the perceived fulfillment of their doctoral expectations — contributes favorably to doctoral degree completion
   - Peer interaction is related to persistence, insofar as degree completers are more likely to be involved with their academic peers than non-persisters
   - The financial support offered to doctoral students is related to attrition and persistence; students who hold research assistantships, teaching assistantships, fellowships, or graduate assistantships are more likely to complete their degrees than students who rely on other types of funding
3. Academic Achievement Indicators are Generally not Effective Predictors of Doctoral Degree Completion, with the Exception of Graduate Records Examination (GRE) Advanced Scores
4. Findings are Mixed with Respect to Employment and Financial Factors
5. Personal and Psychological Variables Represent a Relatively New Direction in the Study of Doctoral Student Attrition and Persistence; A Number of these Variables has been Shown to Relate to Persistence
6. Demographic Variables do not Conclusively Distinguish Persisters from Those Who Drop Out
7. Retention and Attrition Rates Vary Widely Among Institutions
8. All but Dissertation (ABD) is Not the Stage Where the Greatest Proportion of Doctoral Students Necessarily Departs
9. Time-to-Degree (TTD) is Related to Attrition
10. Doctoral Programs that Have Smaller Entering Cohorts Have Consistently Lower TTD and Consistently Higher Completion Rates Than Programs with Larger Entering Cohorts

(Reproduced from Bair & Haworth, 2004)

Bair and Haworth (2004), further break down each of these 10 general conclusions into variables and factors that are perceived to influence doctoral degree completion. Figure 3 shows the variables listed in their meta-synthesis. Variables that are shown to have little to no effect on doctoral degree completion or have mixed results are stricken below.

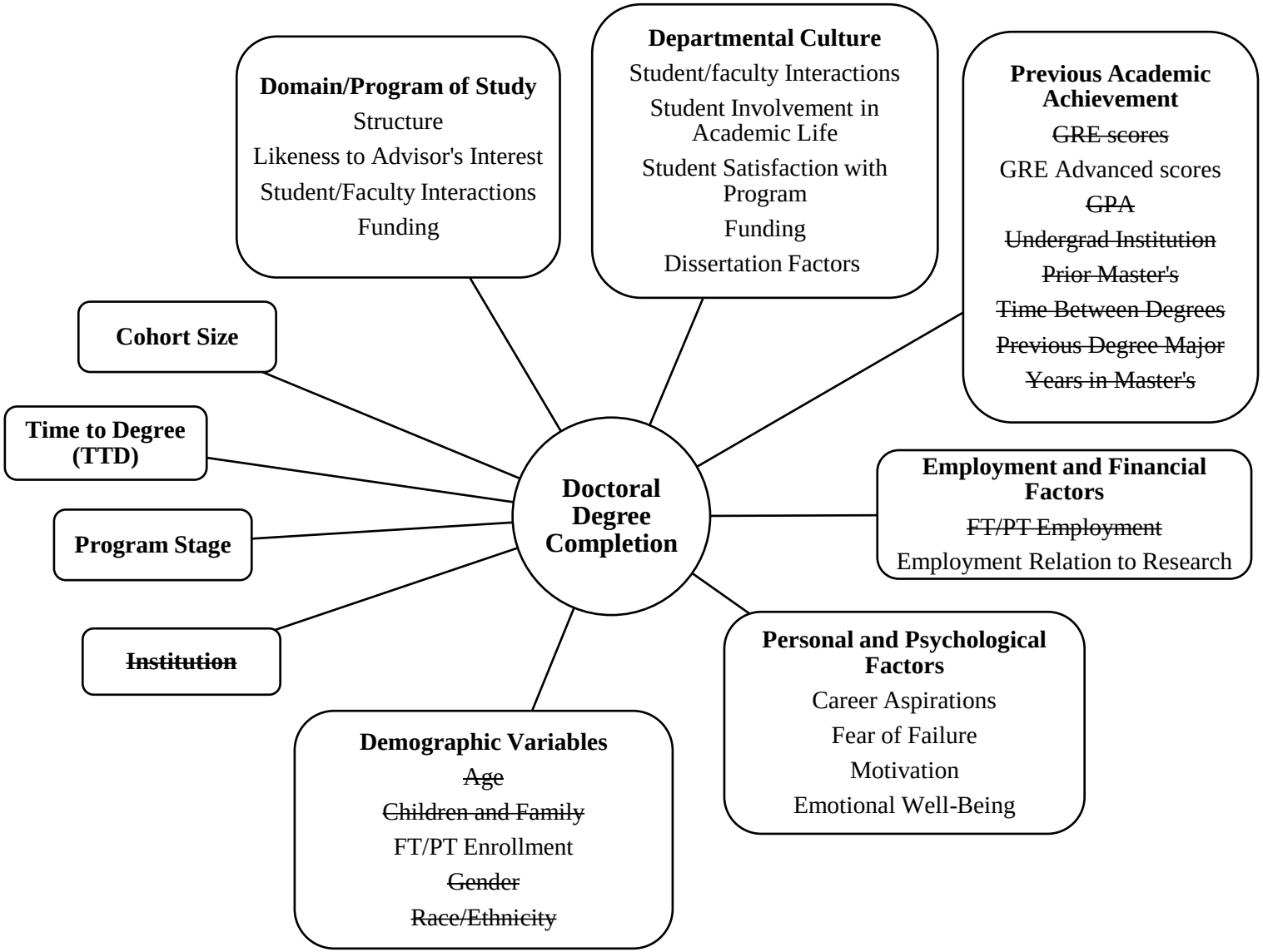

Figure 3: Factors Related to Doctoral Degree Completion (adapted from Bair & Haworth, 2004)

The domain-based student success factors highlighted here include the field of study, likeness to advisor's interest, and the way the program structures the curriculum. Bowen and Rudenstine's (1992), research showed that natural science has the highest rate of degree completion, with lower rates among social studies, and humanities, even when accounting for differences among gender, time, and funding. For instance, Golde (1996), found that the attrition rate for life sciences was 17%, physical sciences was 21%, humanities was 27%, and social sciences was also 27% (Bair & Haworth, 2004). What this means for M&S students is that those focused on the human-centric attributes (Social Sciences, Psychology, Human Performance, etc.) may be more at risk for attrition that others with more technical interests/background.

Additionally, Bair and Haworth (2004), mention that the likelihood of degree completion increased if the student's dissertation topic/research interests followed along with the advisor's research. For this reason, administrators should consider faculty specialties and skills when hiring. Another factor related to the field or program of study is the typical amount of interaction each student has with his/her advisor. "The single most frequently occurring finding in this meta-synthesis was that successful degree completion [and lower time to degree (TTD)] is related to *the frequency and quality of contact between a doctoral student and her or his advisor(s)* or other faculty in the student's doctoral program;" not a single study/experiment investigated countered this finding, (Bair & Haworth, 2004).

Bair and Haworth (2004), state that departing students cited inadequate advising, lack of advisor interest, unavailability of faculty, or negative student-faculty relationships as reasons for dropping the program. Bair and Haworth (2004), note a study completed by Muszynski (1988),

in which she used multiple regression analysis and found seven dissertation factors that contributed to degree completion which include (see Table 5):

Table 5: Seven Dissertation Factors Related to Degree Completion

1) good advisor (supportive, interested, competent, secure)
2) good topic choice (quickly manageable, interesting)
3) internal strength (independence, high motivation, ability to endure frustration)
4) self-imposed deadline or goal
5) avoiding or limiting employment
6) delaying internship (until completion of dissertation)
7) externally imposed incentives (such as future employment)

(Pulled from Bair & Haworth, 2004)

Additionally, Bair and Haworth (2004), noted that early selection of a dissertation topic led to a greater chance of successful degree completion. Further, the number of times the topic changed, difficulty scoping the topic, poor topic choice, and inaccessibility of the subject also contributed to the student's ability attain their degree (Bair & Haworth, 2004). In addition to dissertation topic difficulties, the switch from highly structured coursework to the flexibility of "all but dissertation" ABD (also reducing the number of interactions with peers and faculty) hindered degree completion in nearly 50% of ABD students (Bair & Haworth, 2004; Huguley, 1988; Mah, 1986).

Modeling and Simulation Domain Modeling

The benefit of providing background on M&S is two-fold: 1) while higher education is the context, the SMST at UCF is the example used to apply the present domain model; and 2) M&S techniques are utilized as a solution in the present document, thus an explanation of the field these techniques belong to is also beneficial.

Importance of Modeling and Simulation

"[M]odeling and simulation is an important discipline and like mathematics, a vital infrastructure for other disciplines," (Ören, 2011b). M&S is germane to a number of subjects (Ören, 2011b), including engineering, computer science, social sciences, etc. Government agencies believe M&S to be *critical* to our future (Bair & Jackson, 2015), which fuels the need to produce M&S professionals – people that can use interdisciplinary methods and techniques to address complex problems.

There are many different reasons to use M&S techniques and approaches. M&S can be used to determine affordability, increase safety for workers that routinely complete dangerous tasks, increase awareness of how organizations function, train and educate, aid in analysis and decision making (e.g., creating an adaptive system), design and engineer products, perform experiments, and entertain (L. Bair & Jackson, 2013; Loper et al., 2011; Ören, 2005; Tolk, 2009). Although M&S is mainly utilized for military training applications currently, it is spreading quickly to non-military applications (Tolk, 2009).

The Current State of Modeling and Simulation Domain

What efforts have M&S professionals taken to formalize domain knowledge? The field has started to develop its own theories, methods, and standards over the last 40 years. Progress for M&S evolves alongside the growth of technology. However, as new technology emerges, so does the need for a standard M&S discipline-wide foundation that can be frequently updated.

Clearly articulating the impact of M&S will be important to the sustainment of the field. It is also important that M&S experts evaluate the state of the field to solidify the foundational knowledge. M&S is a largely diverse field and many professionals claim that no

one person could attain all the knowledge related to M&S (Bair & Jackson, 2013, 2015). Thus, the domain needs to identify centralized theories and methods to determine an agreed upon set of KSAs necessary to become an M&S professional (or at least specialized M&S professional). This set of KSAs is necessary to inform domain knowledge structuring (e.g., taxonomy, ontology), determine topic prioritization within the graduate curriculum, inform which instructional strategies may be most effective (based on learning outcomes and the strategy's success in similar domains), and help determine measurements of success for the students and program.

M&S is the application of many different skills, methods, and theories borrowed from multiple disciplines. These separate domains are now blending to improve and innovate on creative solutions to complex problems (Tolk, 2009). This shift moves the problem space from one of traditional disciplines – limited, but well-defined theoretical boundaries and well-controlled incremental basic research– into more interdisciplinary type work – messy, unfamiliar, but better suited for an applied space (Tolk, 2009). Thus, fuzzy systems using M&S techniques are an ideal solution for modeling domain knowledge dynamically and simulating program changes.

Although M&S is rooted in borrowed concepts from computer science, engineering, and human factors, many argue that it has evolved into a separate, stand alone, discipline deserving distinction from the others (Mielke, Scerbo, Gaubatz, & Watson, 2009). In their paper entitled "Towards Making Modeling & Simulation into a Discipline," Sarjoughian & Zeigler (2000), proposed an approach borrowed from software engineering to frame elements of M&S into meaningful formalized components (see Figures 4 and 5).

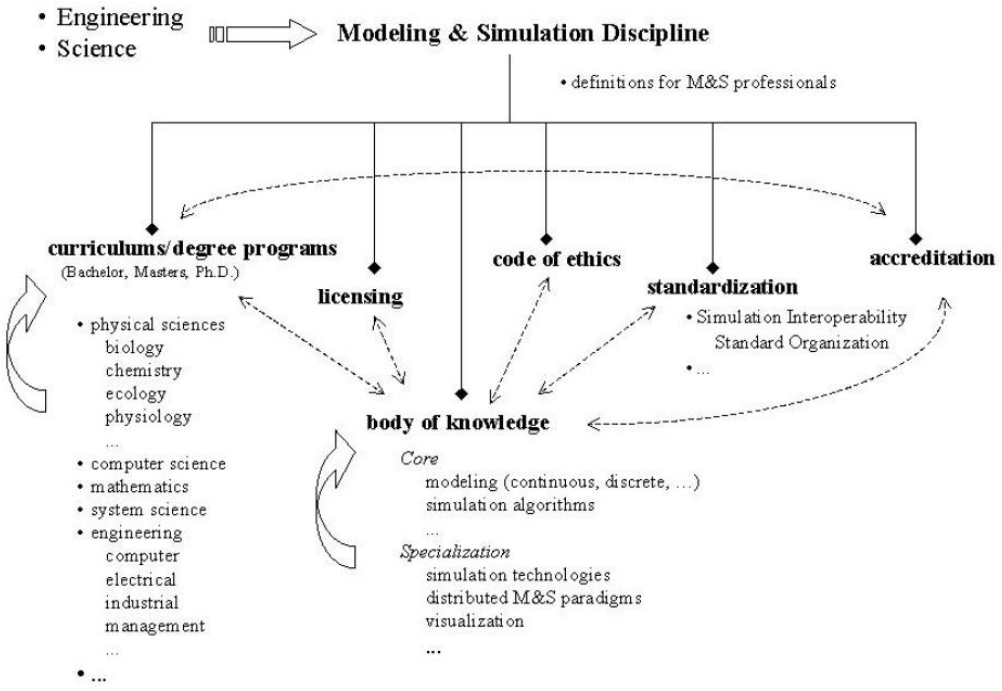

Figure 4: Sarjoughian & Zeigler (2000) elements formalizing the M&S Discipline/Domain

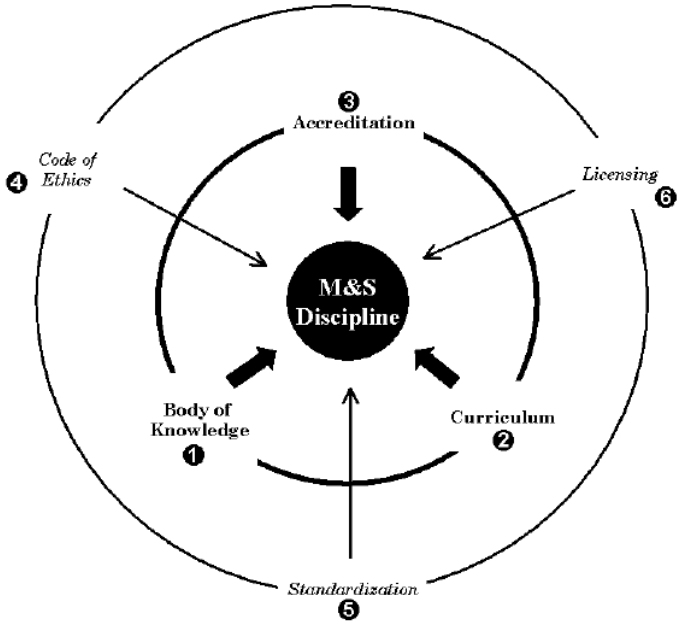

Figure 5: Sarjoughian & Zeigler (2000) a strategic approach to make M&S a discipline

You can see here that these many elements are connected and inform one another. Each component in Figure 5 is numbered. This number designates the order in which the discipline wide standards *should* be established. However, this is not the order in which M&S *has* developed domain formalizations. Most M&S formalization efforts include an attempt at a Body/Book of Knowledge (BoK; Step 1 in Figure 5), standard M&S curricula (Step 2), and licensing (Step 6). Efforts toward formalized accreditation requirements, a code of ethics, standardization (e.g., for interoperability) exist but are sparse and outside of the scope of the present study. However, I plan to address these elements in future research. The efforts toward an M&S BoK and standard curriculum are detailed in the following sections.

Modeling and Simulation Body/Book of Knowledge

While many independent efforts towards establishing M&S as its own unique field of study are reported, "there continues to be a disturbing absence of a coherent and widely accepted statement of the body of knowledge that characterizes the discipline," (Birta, 2003). A BoK "is structured knowledge that is used by members of a discipline to guide their practice or work," (Ören, 2014). A BoK encompasses the accepted ontological foundation of a field of study; thus, its creation must be thoughtful and systematic (Ören, 2014). Here M&S is challenged by its interdisciplinary/specialized nature as core knowledge must be specialized enough to allow for realistic expectations of knowledge for practitioners, while specialized knowledge must be general enough not to back professionals into a too narrow corner (Ören, 2014). "Identification of meaningful specializations that will be generally accepted by the M&S community will likely be a difficult task because of the need to accommodate a large range of

exiting expertise and a large range of marketplace requirements," (Birta, 2003). As such, M&S professionals will need to *prioritize* index topics/KSAs (Ören & Waite, 2010).

The largest and most collective effort to date toward determining a standardized set of core and specialized skills was led by Dr. Tuncer Ören. Ören, together with various M&S organizations, drafted an M&S dictionary (Ören, 2011b), various taxonomies (Ören, 2000), and a BoK index (Ören & Waite, 2010). The BoK is where professionals, academics, and students reference domain agreed upon KSAs to increase their value as professionals. An example of an existing BoK includes the Project Management Body of Knowledge (PMBOK; *PMBOK Guide and Standards*, 2020). While Ören and his team have contributed substantially to the creation of a standard M&S BoK in which many relevant KSAs have been identified, it currently only exists as an index (Ören & Waite, 2010).

It is important to note that there are other efforts towards creating an M&S BoK directed by the Department of Defense (DoD; Department of Defense, 2009). However, these efforts are presented as a *starting* point and are DoD specific. Additionally, Ören states that the DoD's document is focused on fitting M&S KSAs within Bloom's Taxonomy, detracting from its purpose: determining core M&S KSAs, (Ören, 2014). Specifically, Bloom's Taxonomy is not always the most appropriate taxonomy to use. It is hierarchical, which means that the learner *must* master one level to pass onto the next. Other taxonomies and frameworks are better suited for science and engineering courses (e.g., Perry's framework or Baxter-Magolda's framework), while others are also better suited for cumulative and interactive approaches (e.g., Fink's framework; Nilson, 2010). As a result, the present dissertation will focus on the work of Dr. Ören. Below in Table 6 is a listed timeline detailing the efforts to date toward an M&S BoK.

Table 6: Progression of M&S BoK efforts in Chronological Order

- **2003**-Birta documents need for Set KSAs and BoK.
- **2009**-DoD starts on M&S BoK. Since then, little to no published updates.
- **2010**-Ören and Waite attempt to collect perceptions of M&S practitioners and use M&S literature to determine KSAs. Ören and Waite, however, do not detail how these perceptions and the open literature are monitored, what types of key words are used, and how others can help with this effort.
- **2010**-Ören and Waite detail two websites documenting efforts.
- **2011**-Ören details BoK Index efforts.
- **2011**-Ören collects definitions for M&S dictionary to include in BoK.
- **2014**-Ören updates BoK Index efforts.
- **2016**-Ören and team publish M&S Code of Ethics to include in BoK.
- **2017**-Ören updates efforts.

(Birta, 2003; Department of Defense, 2009; Ören, 2011b, 2011a, 2014; Ören & Waite, 2010; *Simulationist Code of Ethics*, 2016)

Although Ören has made significant progress, efforts are slowing and no BoK is static (Birta, 2003); thus, Ören has solicited for assistance in refining the M&S BoK as the field and technology evolve (Ören, 2011a).

There are many different avenues for creating a BoK. Ören (2014), details two approaches that could be used: 1) determining KSAs based on the applied domain or 2) from an M&S perspective highlighting the "(i) purpose of the use of simulation, (ii) problem to be solved, (iii) connectivity of operations of the real system and the simulation, (iv) types of knowledge processing, and (v) philosophy of science." I believe both approaches have merit. Therefore, I explore the M&S domain using both approaches. Taking the first approach, I use NLP to investigate KSAs applied through publications, posing the question:

Research Question 1: *What are the KSAs applied most frequently in M&S academic literature?*

The second approach requires a more in depth investigating of the relationships between M&S topics. As such, I present literature on current licensing and educational efforts to understand how M&S experts have organized information to date.

Modeling and Simulation Licensing Efforts

Design of the Certified Modeling and Simulation Professional (CMSP) exam started in 2000 as a solution for understanding the types of skills an M&S Professional needs to be successful (Lewis & Rowe, 2010). A certification exam could signal to a potential employer that an applicant possesses at the very least a core set of M&S skills and specific skills related to one of the specializations. "[W]ithout [a certification exam], there is no way to determine who is truly qualified to practice that profession," (Lewis & Rowe, 2010). When the Modeling & Simulation Professional Certification Commission (MSPCC) first developed the CMSP, the intention was to start simply with one exam (no specializations) targeted at the Defense Training and Simulation community, (Lewis & Rowe, 2010). As such, the KSAs identified for evaluating professionals were geared towards military and government needs, which is an appropriate starting point as these are the application areas in which M&S was originally designed for, but potentially biased the certification exam.

The creators of the CMSP called for continual improvement and evolution of the certification exam stating: "The initial certification was created, however, with an implicit understanding that the program would evolve through time, and would perhaps have multiple levels, tracks, and/or specialties in the future." (Lewis & Rowe, 2010). As the domain expands beyond defense applications, M&S Professionals should continually evaluate topics and their relationships to make sure they still align with the needs of the field.

The exam now has two *tracks*, the manager and the practitioner (Lewis & Rowe, 2010). The practitioner track is for individuals involved in building, developing, and evaluating M&S using engineering, computer science and other technical skills. The management track on the

other hand is geared towards those that supervise M&S projects and need to know enough about modeling methods, paradigms, and standards to evaluate and manage the work of those on the team (Bair & Jackson, 2013). These people may or may not have a technical background but still have much experience in the field. The index for both the core and the specialized topics is presented below in Figure 6.

M&S-Domain Specific Knowledge	**Areas of Specialized Expertise**	
1. Concepts and context 1.1. Fundamental terms and concepts 1.2. Categories and paradigms 1.3. History of M&S	**2. Applications of M&S** 2.1. Training 2.2. Analysis 2.3. Experimentation 2.4. Acquisition 2.5. Engineering 2.6. Test and evaluation	
6. Supporting tools, techniques, and resources 6.1. Major simulation infrastructures 6.2. M&S resource repositories 6.3. M&S organizations	**4. Modeling methods** 4.1. Stochastic modeling 4.2. Physics-based modeling 4.3. Structural modeling 4.4. Finite element modeling and computational fluid dynamics 4.5. Monte Carlo simulation 4.6. Discrete event simulation 4.7. Continuous simulation 4.8. Human behavior modeling 4.9. Multi-resolution simulation 4.10. Other modeling methods	**5. Domains of use of M&S** 5.1. Combat and military 5.2. Aerospace 5.3. Medicine and healthcare 5.4. Manufacturing and material handling 5.5. Logistics and supply chain 5.6. Transportation 5.7. Computer and communications systems 5.8. Environment and ecology 5.9. Business 5.10. Social science 5.11. Energy 5.12. Other domains of use
7. Business and management of M&S 7.1. Ethics and principles for M&S practitioners 7.2. Management of M&S projects and processes 7.3. M&S workforce development 7.4. M&S business practice and economics 7.5. M&S industrial development		
M&S-Specific Software-Engineering-Related Expertise		
6. Simulation implementation 6.1. Modeling and simulation lifecycle 6.2. Modeling and simulation standards 6.3. Development processes 6.4. Conceptual modeling 6.5. Specialized modeling and simulation languages 6.6. Verification, validation, and accreditation 6.7. Distributed simulation and interoperability 6.8. Virtual environments and virtual reality 6.9. Human-computer interaction and virtual environments 6.10. Semi-automated forces (SAF) / computer generated forces (CGF) 6.11. Stimulation		
Domain Knowledge in Related Fields of Practice		
8. Related communities of practice and disciplines 8.1. Statistics and probability 8.2. Mathematics 8.3. Software engineering and development 8.4. Systems science and engineering		

Figure 6: CMSP Exam Topics (Bair & Jackson, 2015)

The exam developers, however, understand that both the breadth and depth of the information in the exam can be overwhelming. Thus, the exam is designed to be a learning experience as well. It is an online take home exam, which the applicant has 30 days to complete (Lewis & Rowe, 2010). A list of useful references is available on the M&SPCC-CMSP

website (Dwyer, 2020). However, it is still challenging due to the number of sources listed (more than is manageable for a novice in 30 days). The applicant must complete a set of required core M&S questions, but they also have a choice of specialized topics they wish to include and omit from the exam. “All questions are either multiple choice or True/False...Each applicant must complete three categories of questions in…two sections,” (Bair & Jackson, 2015).

In addition to the examination, both experience and education are required at varying levels. Either the applicant must possess a doctoral degree plus three years of experience, a master’s degree plus five years of experience, a bachelor's degree plus 6 years of experience, or an associate degree plus eight years of experience (Lewis & Rowe, 2010). The qualification standards imply that both experience and education are necessary to become a professional.

Bair and Jackson (2013), say “the CMSP program, is doing its best to improve and revise the program to better meet the needs of M&S industry professionals and those who use their services.” However, many questions still need to be addressed: 1) what are the needs of industry professionals; 2) if these needs change, how are these changes reflected in the exam topics; 3) whose needs are considered; and 4) are their needs short-term or long-term needs? These and many more questions (refer to Bair & Jackson, 2013) should be considered when determining an appropriate standard ontology. The developers of the CMSP wish to address these questions by opening a dialog and collecting feedback for improving the exam (Bair & Jackson, 2013). This call for continual feedback provides an opportunity for academics to investigate M&S Professionalism.

Many professionals have tried to narrow down what constitutes an M&S Professional. The top performers in the field have discussed it in informal conversations, at conferences, held

workshops, and still have only concluded: “a common understanding of the M&S Professional only exists as a gestalt—I’ll know it when I see it,” (Bair & Jackson, 2015), but it is clear that this distinction is less than helpful. Thus, in the present dissertation I use the phrase *M&S Professional* here to mean individuals that currently hold a self-identified or domain-identified M&S position in an academic, industry, government, or non-profit setting.

Bair & Jackson (2013), investigated what it means to be an “M&S Professional,” stating "we must scrutinize the M&S profession and its work so that M&S may better define itself as a unique field of study and develop greater unanimity of what it means to be an M&S professional." It is necessary for M&S Professionals to look at the evolution of the domain and explicitly call out the KSAs that are unique to the field and its practitioners. It is not simply enough to say one is an *M&S professional*, because it does not convey the type of work or application areas with which he/she is familiar (Bair & Jackson, 2013). This leads me to believe that M&S should consider specializations the way other disciplines have, such as Engineering (e.g., Electrical, Mechanical, Industrial, Systems, Software). This idea is echoed throughout the field (Bair & Jackson, 2013).

These specializations help the field narrow down skills for employment per kinds of M&S professionals to better understand which specializations are qualified for certain work, but what is not clear is what types of M&S specialists exist or how current M&S curricula map to these specializations. There lies a disconnect between the types of jobs available and the curricula meant to prepare students for these jobs. Therefore, this dissertation is also intended to address these research questions:

Research Question 2*: What are the KSAs most requested in M&S job listings within the United States?*

Research Question 3*: How should M&S job types be categorized?*

Research Question 4*: What are the KSAs most identified per job type?*

Existing Education Programs in Modeling and Simulation

Addressing the interdisciplinarity of M&S often proves problematic utilizing a traditional program structure (Mielke et al., 2009). Within graduate education, students are expected to gather both a wide breath of related topics, as well as, a deep understanding of their specialization, creating a *T* shaped person. This type of educational program requires a different structure and foundation built upon collaboration between various university departments. M&S has historically pulled foundational information from other domains such as systems and industrial engineering, mathematics, computer science, and other technology related fields (Birta, 2003) but it is still finding ways to bring in new information and techniques (e.g., social science and philosophy). This new assortment of information makes it difficult to develop a qualified workforce, particularly when jobs found in industry, academic, and government facilities all require different skills (L. Bair & Jackson, 2013; Kincaid & Westerlund, 2009).

The difficulty of determining which KSAs to address is evident when investigating presently available education programs. Four universities offer well known M&S graduate programs in the United States: 1) Old Dominion University, 2) The University of Alabama Huntsville, 3) The Naval Postgraduate School, 4) The University of Central Florida. Each of the four programs promotes different educational paths and course selections (refer to Table 2 for the

programs' core courses; universities listed in alphabetical order). While some M&S programs are designed based on recommendations given during early M&S workshops, some have forged ahead with *one-off* programs (Sarjoughian & Zeigler, 2000).

Across these four programs there are varying numbers of core courses, which are listed below in Table 7. These topics are relatively similar across the programs. However, the applied fields in which each school focuses on are different. Additionally, each school offers different elective courses. This difference could stem from a disagreement of the KSAs required of an M&S professional. This is pertinent for example in times where only a few faculty members are needed and a gap in potential knowledge forms. Unintentional bias could exacerbate this. Faculty often teach to the techniques, tools, and theories they are most familiar with (as opposed to the most commonly used). Using data-driven techniques to organize and categorize information could potentially alleviate these issues.

Table 7: Core Courses from four M&S Graduate Programs (2017-2018 Academic Year)

University	Core Courses/Competencies
Naval Postgraduate School	List as core competencies • History and Fundamentals of M&S • Applied Mathematics • Computer Systems • Virtual Environments • Training and Human Systems • M&S Systems Lifecycle Management • Modeling (system, combat, real-world physics, VV&A) • JPME level 2 Note: Optional refresher is available in summer
Old Dominion University	Masters (of Engineering or Science) Core Courses • Principles of Visualization • (Advanced) Analysis for Modeling and Simulation • One Advanced Modeling Course (e.g., Machine Learning 1) • One Advanced Simulation Course (e.g., Finite Element Analysis) Doctoral (Doctor of Engineering or Ph.D.) Core Courses • One advanced simulation course (e.g., Cluster Parallel Computing) • Simulation Formalisms • Synthetic Environments • Advanced Analysis for Modeling and Simulation • Two approved technical electives Note: Some students must meet additional pre-requisites
University of Alabama Huntsville	Master of Science Core Courses • Survey of Modeling and Simulation • Intermediate Mathematical Modeling • Statistical Methods for Engineers • Intro to Computer Graphics and/or Artificial Intel. I • Intro to Systems Simulation and/or M&S 1 • Engineering Systems and/or Software Engineering Process Note: "and/or" depends on if the student chose the Thesis (or) or Non-Thesis (and) Thesis option Doctoral (Ph.D.) Core Courses • Survey of Modeling and Simulation • Introduction to Computer Graphics • Introduction to Systems Simulation or M&S 1 • Intermediate Mathematical Modeling • Engineering Systems or Software Engineering Process • Artificial Intelligence I • Statistical Method for Engineers • M&S II

University	Core Courses/Competencies
	• Advanced System Simulation • Computational Models, Adv. Algorithm Design and Analysis, or Artificial Intel. II • Systems Modeling and Analysis, System Modeling, Formal Methods in Software Engineering, or Advanced Software Engineering Topics • Value Decision Theory or Advanced Statistical Applications
University of Central Florida	Master of Science Core Courses • Perspectives of Modeling and Simulation • Quantitative Aspects of Modeling and Simulation • Human Systems Integration for M&S, Human-Computer Integration, or Adv. Human-Computer Interaction • Simulation Techniques • Research Design for Modeling and Simulation • Simulation Research Methods and Practicum Note: Non-thesis students are required to take an additional restricted elective (e.g., Modeling and Simulation for Instructional Design) Doctoral (Ph.D.) Core Courses • Perspectives of Modeling and Simulation • Quantitative Aspects of Modeling and Simulation • Human Systems Integration for M&S, Human-Computer Integration, or Adv. Human-Computer Interaction • Simulation Techniques • Research Design for Modeling and Simulation • Simulation Research Methods and Practicum Note: All doctoral students are required to take an additional restricted elective (e.g., Interdisciplinary Approach to Data Visualization)

(*Academic Programs*, 2017; *Department of Modeling, Simulation and Visualization Engineering*, n.d.; *Education-Arizona Center of Integrative Modeling and Simulation*, 2017; *Modeling and Simulation Doctor of Philosophy*, 2017; *Modeling and Simulation Graduate Courses*, 2017; *Modeling and Simulation Master of Science*, 2017; *Modeling and Simulation MS*, 2017; *Modeling and Simulation Ph.D.*, 2017; *System Engineering*, 2017)

"The result of this observation is that real crossdisciplinary compositions or federations of M&S applications of different domains are nearly without examples. It remains state of the art that M&S applications of one domain gets extended to include desired phenomena instead of

integrating models from human and social sciences," (Tolk, 2009). As part of that shift, M&S professionals must possess mutual awareness (Tolk, 2009), and standards. This dissertation is intended to provide another tool for M&S professionals who wish to continue transforming this domain. Therefore, I ask:

Research Question 5*: What are the KSAs taught most frequently in M&S graduate level course descriptions for Universities within the United States?*

Overall System Configuration

One of the biggest issues with university strategic planning is that despite our advances in technology and data analysis in higher education, often there is no central university-wide system to which university decision makers have access (Guan et al., 2002). Complex systems (e.g. a university-wide strategic planning systems) are difficult to build because computers use binary logic, meaning we often have to simplify characteristics of the problem to model them in a way a computer can understand, which takes time and effort (*Computer Logic vs. Human Logic*, 2018). However, with recent advancements in NLP, machine learning, and artificial intelligence (AI), computers can also be programmed to deal with ambiguous data using fuzzy systems. **Fuzzy systems** can account for incomplete data (lack of information or understanding), imprecise data (the noise or error), and randomness (potential accidents calculated by probability; *Computer Logic vs. Human Logic*, 2018), using programmed rules, machine learning, and/or artificial intelligence. Fuzzy systems can include expert systems, intelligent systems, and adaptive systems, each more complex that the last. An **expert system** is “a creation of rules for a system that can learn and provide expert-level suggestions,” (Expert Systems and Simulations, 2018).

An example of a rule-based system in higher education may include an expert-system modeling naturally observed decisions made by administrative staff. Machine learning could also be used to build an intelligent-system to categorize students by degree and interests, whereas artificial intelligence could be used to predict optimal courses based on frequent paths. An **intelligent tutor** is "a computer system that aims to provide immediate and customized instruction or feedback to learners, usually without intervention from a human teacher," (Sottilare, 2015, p.1). Further, in **adaptive tutors** "the agents observe and interpret each learner's data (behaviors and physiology) to determine learner states (e.g., engagement, emotions, performance) and identify individually tailored learning needs," (Sottilare, 2015, p.2).

Intelligent and adaptive tutors consist of several modular sections, which typically include at least a domain model, learner/student model, pedagogical/instructional design/tutor model, and some type if user interface (Sottilare, 2015). The addition of a resource allocation model to the systems design could yield a more robust product aimed to bridge the gap between education and administration. Expanding on Figure 1, and using the literature presented in this chapter, Figure 7 shows a detailed conceptual model for the university-wide system.

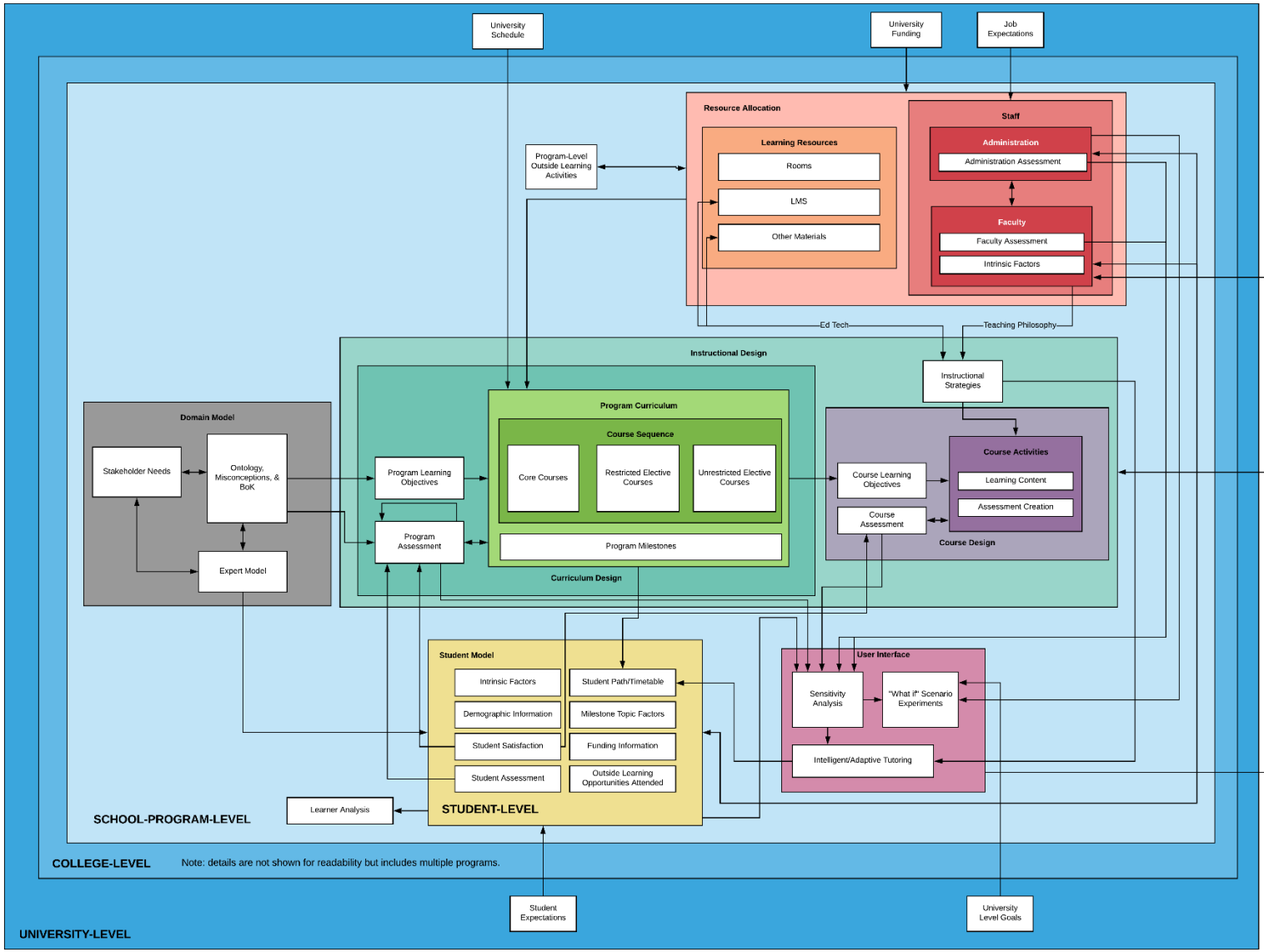

Figure 7: Conceptual Model of Strategic Planning System

Domain Model

The purpose of the domain model is to house information that can be used as an ideal state. The domain model has three main components: stakeholder needs, the ontology, and expert models. The ontology serves as an ideal state of the domain and the expert models are the ideal state of the domain's professionals. Stakeholder needs inform both components.

Stakeholders

Table 8 presents an example list of stakeholders identified for the present system. I have identified the *key* stakeholders as belonging to the group *UCF M&S Agents* (except for alumni) as these are the stakeholders most impacted by any program decisions. Additional stakeholders are expected to benefit from the outcomes of the system, should the end user decided to implement recommended changes. For this project, I group and define the *key* stakeholder sets here as:

- **UCF M&S Students:** In the present document, students are identified as individuals that are currently enrolled in either the SMST master's, and/or doctoral programs. Students are the stakeholders most affected by the model outcomes/recommended changes.
- **UCF M&S Administrators:** Administrators are individuals whose main responsibility to the SMST is to assist with the program structure or university organization. This would include the program coordinator, assistant director (if applicable), and the director. UCF administrators are going to be the most likely end-user (or target) population.
- **UCF M&S Faculty:** Individuals that teach M&S courses, assist with curriculum, admission, and administrative decisions, and advise students. This is a very general statement and the "grouping" of SMST faculty is much more complicated than this

definition may present. In interdisciplinary programs, faculty in various departments can meet this definition without being on the SMST payroll. In the 2016-2017 academic year, only two tenure-track faculty members were compensated by the M&S program. However, Institute for Simulation and Training Research Faculty have previously served as faculty (in some cases free of charge to the program). Starting in 2016-2017, these faculty members will be compensated for their time to encourage realistic job expectations and to reduce faculty workload.

Table 8: List of Stakeholders

- UCF SMST Set
  - UCF SMST Students
    - Doctoral
    - Master's Thesis
    - Master's Non-Thesis
    - Certificate
    - Non-Degree Seeking
  - UCF SMST Administrative Personnel
    - Director
    - Assistant Director
    - Coordinator
    - Other Admin Staff
  - UCF SMST Faculty
    - Full Professor
    - Associate Professor
    - Adjunct Professor
  - UCF SMST Alumni
- Employer/General M&S Set
  - M&S Government Professionals
  - M&S Industry Professionals
  - M&S Professional Organization Members
- Non-UCF Academic Set
  - Other M&S Students
  - Other M&S Administrators
  - Other M&S Faculty
  - Other M&S Alumni
- "Other" Interdisciplinary Set
  - Interdisciplinary Students
  - Interdisciplinary Administrators
  - Interdisciplinary Faculty Agents

Interdisciplinary Alumni

Ontology

An **ontology** "defines a common vocabulary for researchers who need to share information in a domain," (Noy & McGuinness, 2004). Ontologies consist of overt explanations of domain concepts and concept properties in a way that makes is easy for machines to read the content, (Noy & McGuinness, 2004). Ontologies can be general (e.g., common core public education) or specific (e.g. marksmanship), ontologies. This distinction is related to the idea of general/strong versus narrow/weak AI. General AIs (e.g., what Amazon's Echo strives to be) typically have more variables making the combination of information and functions almost impossible to parse out. Narrow/weak AI is focused on one subject/domain (e.g., expert system; *General and Narrow AI*, 2019). In this instance I am creating an M&S specific ontology to inform a *future* weak AI system (i.e., M&S adaptive tutor). Ontologies are used for several reasons some of which include (see Table 4):

Table 9: Ontology Uses

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

(Noy & McGuinness, 2004)

There are also several levels of ontology formalization. Table 5 refers to Uschold and Gruninger's (1996), classification of ontology formality.

Table 10: Ontology Formalization Categories

- **Highly Informal**: expressed loosely in natural language
- **Semi-formal**: expressed in an artificial formally defined language [e.g., programmed version]
- **Rigorously formal**: meticulously defined terms with formal semantics, theorems, and proofs of such properties as soundness and completeness

(Uschold & Gruninger, 1996)

An example of a highly informal (also called lightweight) ontology analysis is a set of terms, a data dictionary, or structure glossaries (Wong et al., 2012). Semi-formal ontologies can include Extensible Markup Language (XML) schemas, formal taxonomies, or data models (Wong et al., 2012). Rigorously formal ontology instances include description and general logistics (Wong et al., 2012). It is my intention to create and validate a semi-formal ontology for the present dissertation. Additional efforts towards increased formality are detailed in the Future Research section in Chapter Five. To create an ontology, I used Noy and McGuinness (2004), methodology for Building an Ontology (see Table 11).

Table 11: Noy & McGuinness's Steps for Creating an Ontology

- Step 1: Establish domain and scope of the ontology
  - Determine competency questions
- Step 2: Consider reusing existing ontologies
- Step 3: Enumerate important terms in the ontology
- Step 4: Define the classes and the class hierarchy
- Step 5: Define the properties of classes-slots
  - Determine the intrinsic and extrinsic properties of the class
- Step 6: Define the facets of the slots
  - Slot Cardinality
  - Slot Value-Type
  - Domain and Range of Slot
- Step 7: Create Instances

(Noy & McGuinness, 2004)

Establish Ontology Scope

Noy and McGuinness (2004), explain that the first step consists of scoping the ontology and determining general competency questions. The authors provide a few questions to help a designer scope the ontology needed. The questions include "What is the domain that the ontology will cover?", "For what [are we] going to use the ontology?", and "Who will use and maintain the ontology?", (Noy & McGuinness, 2004). Noy and McGuinness (2004), further detail that the intended use of the ontology can help the designer determined the types of data

that would be most useful. For example, if the intention of the domain is to manage inventory items within a warehouse (e.g. wine) for a business that distributed their product online (e.g., monthly wine club). An adaptive system can use the domain ontology to suggest not only wines that follow along with the customer's tastes but also those that need to be promoted because it is overstocked (Noy & McGuinness, 2004). In that case, the designer should engage in data collection of variables like wholesale pricing, retail pricing, stock/availability, wine type (e.g., red), wine sub-types (e.g., red blend), blend percentage (e.g., 85% Cabernet Sauvignon and 15% Merlot), or brand (e.g., Chateau Picard; example adapted from Noy & McGuinness, 2004; startrek.com staff, 2019). However, if the ontology was designed to assist natural language processing of wine related literature, the designer may be more concerned with components like wine-types, common wine characteristics (e.g., smoky, spicy, bright, fresh), synonyms (e.g., hazy, peppery, light, crisp), parts-of-speech, etc. (example adapted from Noy & McGuinness, 2004; startrek.com staff, 2019).

While scoping the ontology, Noy and McGuinness (2004), also instruct designers to list competency questions related to the domain. A competency question is defined as "questions that a knowledge base based on the ontology should be able to answer," (Noy & McGuinness, 2004). Competency questions serve as questions later for verification purposes. Examples of competency questions are: "Which wine characteristics should I consider when choosing a wine?", "Is Bordeaux a red or white wine?", and "Does Cabernet Sauvignon go well with seafood?" (Noy & McGuinness, 2004). These questions can be used to help designers narrow down the types of data they plan to include in the system. The CMSP exam questions could be

used as competency questions. Thus, I don't focus on developing competency questions for the present dissertation

Reuse Existing Ontologies

Ontologies are used in AI to provide the machine a structure used to inform outputs from the system; thus, many programmed ontologies currently exist. "It is almost always worth considering what someone else has done and checking if we can refine and extend existing sources for our particular domain and task," (Noy & McGuinness, 2004). It benefits us to build upon an existing ontology rather than start from scratch. There are several general and domain specific ontologies. However, M&S is a new domain with little formalization. To the best of my knowledge, there exists no M&S ontology. As a result, I looked at other related topics (e.g., Computer Science, Mathematics, Social Sciences) as well M&S specific formalizations (e.g., Birta's taxonomy; Ören's BoK efforts detailed previously) to inform the creation of an M&S ontology.

Enumerate Important Items

Next, Noy & McGuinness (2004), instruct the designer to identify/brainstorm terms related to the domain. Specifically, they mention "[i]nitially it is important to get a comprehensive list of terms without worrying about overlap between concepts they represent, relations among the terms, or any properties that the concepts may have, or whether the concepts are classes or slots," (Noy & McGuinness, 2004). The index designed by Ören (2014), can serve as a starting point. However, I plan to use natural language processing techniques to collect and weight common M&S terms within domain-related documents. A relative frequency analysis can

inform the weights for each enumerated word. This is where the majority of my efforts will be focused for the present dissertation.

Define the Classes and Class Hierarchy

From the list of words identified and prioritized, Noy and McGuinness (2004), instruct the designer to choose *anchor* words to serve as the classes for the overall domain hierarchy. Anchor words are identified by determining the independence of each word. "From the list created, we select the terms that describe objects having independent existence rather than terms that describe these objects," (Noy & McGuinness, 2004). Noy & McGuinness (2004), mention that these are the "most important steps in the ontology-design process." There are several possible approaches for this process: top-down, bottom-up, or combination classification (Noy & McGuinness, 2004; Uschold & Gruninger, 1996). A combination approach is most common.

Define Properties of Classes (Slots)

These objects will become class *anchors* and we can fill in information around them. "Most remaining terms are likely to be properties of these classes," (Noy & McGuinness, 2004). The remaining terms should be paired to a class and organized into intrinsic (e.g. wine flavor) and extrinsic (e.g., company name) properties (Noy & McGuinness, 2004). Like many software solutions the sub-classes will inherit the properties of the class (Noy & McGuinness, 2004). As such, the designer should add the property to the most general class (Noy & McGuinness, 2004).

Define the Facets of the Slots

The facets of the slots describe details like the value type (e.g., student name – value type: string), allowed values (e.g., "Melody Pond" – birth name or "River Song" – other names), the number of values (called cardinality; e.g., multiple cardinality – two name options), and other

features of slot values (Noy & McGuinness, 2004). "**Slot cardinality** defines how many values a slot can have." (Noy & McGuinness, 2004). The slot can have a single cardinality or multiple cardinality and can also have a minimum or maximum cardinality (Noy & McGuinness, 2004). For example, expanding on our student name above, the minimum cardinality should be one. The student needs at least one name in the system and the maximum is *n* since a student can legally change their name as many times as they wish (or you can add common nicknames to the system). "A **value-type** facet describes what type of values can fill in the slot," (Noy & McGuinness, 2004). Common value types include string (as seen in example above), number, Boolean, enumerated, and instance. Defining the domain and the range of the slot is also useful during this process (Noy & McGuinness, 2004).

Create Instances

"The last step is creating individual instances of classes in the hierarchy," (Noy & McGuinness, 2004). To do this the designer must first chose a class, then create an instance of the class and then define and fill slot values (Noy & McGuinness, 2004).

Expert Models

Profiles of expert performers will be used to guide SMST students through the university-wide system. The student is compared to the expert profile throughout the system to determine the differences between the two. The system will determine the difference and come up with a plan for the student. However, expert models need to be developed first. Considering the definition of an M&S Professional and the types of specialties outlined for one are still nebulous, I use a natural language processing categorization technique to determine appropriate expert categories.

Natural Language Processing

A promising research area that addresses some of tasks includes **Ontology Learning**, which is a method of using NLP and machine learning methods to investigate texts to create a semi-automated or full-automated domain knowledge systems (Wong et al., 2012). As mentioned in chapter one, NLP is an area of computer science that deals with methods to analyze, model, and understand human language," (Vajjala et al., 2020). NLP is also involved in many commonly used technology products (e.g, Google, Amazon Alexa) to help users intuitively interact with technology. To build these types of products experts complete NLP tasks like language modeling, text classification, information extraction, information retrieval, creating conversational agents, text summarization, building question and answer systems, develop machine translators, and topic modeling (Vajjala et al., 2020). A full discussion of these text-mining and machine learning tasks is outside of the scope of the present dissertation, however, one can be seen in Vajjala et al. (2020). I plan on using a collection of text mining and topic modeling for the present dissertation.

Text Mining and Topic Modeling

Computers use binary logic, meaning we have to numerically represent words in a way a computer can understand (*Expert Systems and Simulations*, 2018; Vajjala et al., 2020). To numerically represent the data, the data needs to be vectorized into a matrix. **Bag-of-Words** (BoW) "is a way of extracting features from text [converting it to numeric form] for use in modeling, such as with machine learning algorithms," (Brownlee, 2017a). Essentially, the algorithm counts the number of times these words occur and places it in a matrix (Brownlee, 2017a). If you were to break up a document by word and place them all in a bag you would have

a certain number of each word but no sentence structure to determine meaning. More complicated bag-of-words approaches can also measure occurrences of common phrases with n number of words, called ***n*-grams** (Vajjala et al., 2020). The BoW method for *n*-grams is called **Bag-of-*n*-grams** [(BoN); Vajjala et al., 2020]. For example, a unigram is a one-word phrase (e.g., pilots), a bigram is a two-word phrase (e.g. pilots Serenity) and a trigram is a three-word phrase (e.g., Wash pilots Serenity). I could explore this data to help identify KSAs. However, to prioritize KSAs using BoW methods requires an assumption that if a word occurs frequently then it is a highly valued KSA and if it occurs seldomly, it is not as highly valued by M&S employers. However, there is a more robust method, called **Term Frequency-Inverse Document Frequency** (TD-IDF), can be used instead to mathematically rank and weight each word. Using TD-IDF as a method for analyzing data not only allows me to determine important concepts for the M&S ontology but it can also inform natural language processing for future M&S domain formalization efforts (e.g., automated collection, categorization, and generation of M&S domain information).

I considered using either text classification or topic modeling. "Text classification is sometimes also referred to as *topic classification*, *text categorization*, or *document categorization*. … topic classification is different from *topic detection*, which refers to the problem of uncovering or extracting "topics" from texts," (Vajjala et al., 2020). Text classification algorithms are also considered supervisory machine learning algorithms, which means it uses pre-labeled (known) data (Vajjala et al., 2020). For the present study, the data collected is not labeled. However, appropriate labels can be detected using topic modeling techniques (Vajjala et al., 2020).

"**Topic modeling** generally refers to a collection of unsupervised statistical learning methods to discover latent topics in a large collection of text documents," (Vajjala et al., 2020). Topic models generally assume that 1) every document is a mix of topics and 2) each topic is a mix of words (Xu, 2018; Zhao, 2018). Topic modeling algorithms include latent semantic analysis/indexing (LSA/I), probabilistic latent semantic analysis (pLSA), and latent Dirichlet allocation [(LDA); Vajjala et al., 2020; Zhao, 2018] Each one of these algorithms assumes a different distribution of words within a topic and distribution of the topics within the document.

LSA uses singular value decomposition for dimension reduction (feature extraction), which means it assumes that the distribution of words is *not* probabilistic (Xu, 2018). It simply looks at the existing data and produces a model based on only this data. This type of model does not do well with new documents (Xu, 2018). Further while this method is easy, quick, and cheap to compute, it is hard to interpret, it needs a *large* set of documents and vocabulary to accurately get results (Xu, 2018).

Instead of using singular value decomposition pLSA uses a probabilistic distribution, meaning it assumes "topics are nothing but a mixture of keywords with a *probability* distribution, and documents are made up of a mixture of topics, again with a *probability* distribution," (Vajjala et al., 2020). The addition of the *probability* distribution allows the model to be *somewhat* generative or predictive of how new data will perform when introduced (Magesh, 2019; Xu, 2018). However, the model isn't truly generative because it is not Bayesian and is prone to overfitting (Magesh, 2019; Xu, 2018). Bayesian probability distribution allows data scientists to *update* their belief the distribution (StataCorp LLC., 2016).

Like pLSA, LDA is also a probabilistic approach (Blei et al., 2002). LDA is different from pLSA in that it is fully Bayesian (Xu, 2018). Using LDA for topic modeling allows data scientists to make inferences with this model and apply it as an algorithm to future documents, making LDA a truly generative topic modelling approach (Blei et al., 2002; Magesh, 2019; Xu, 2018). An additional benefit of LDA over other models is that LDA can account for when documents contain multiple topics (e.g, Captain's Log entry consists of 30% Topic A-prime directive and 70% Topic B-cultural customs; Blei et al., 2002). LDA is the most commonly used in practice and the most popular (Vajjala et al., 2020; Xu, 2018). Thus, I chose to use it to model topics within each source type. As such, I hypothesize:

Hypothesis 1*: Scraping multiple types of M&S documents (job postings, course descriptions, and academic publications) will produce a difference in the KSAs (topics) most frequently mentioned in each source type.*

Occupational Information Network (O*NET)

The output of the analyses discussed will produce several key terms. However, it is up to the researcher to assign meaning to these terms (Vajjala et al., 2020). As such, I use the Occupational Information Network (O*NET®) coding schema used by the U.S. Department of Labor to organize terms into knowledge components, various skill types, and abilities. O*NET is database a "rich set of variables that describe work and worker characteristics" (U.S. Department of Labor et al., 2020). Tables 12, 13, and 14 list the knowledge components, skills, and abilities defined by O*NET, respectively.

Table 12: O*NET Knowledge Components

Knowledge Component(s)	Definition of Knowledge Component
Administration and Management	"Knowledge of business and management principles involved in strategic planning, resource allocation, human resources modeling, leadership technique, production methods, and coordination of people and resources."
Biology	"Knowledge of plant and animal organisms, their tissues, cells, functions, interdependencies, and interactions with each other and the environment"
Building and Construction	"Knowledge of materials, methods, and the tools involved in the construction or repair of houses, buildings, or other structures such as highways and roads."
Chemistry	"Knowledge of the chemical composition, structure, and properties of substances and of the chemical processes and transformations that they undergo. This includes uses of chemicals and their interactions, danger signs, production techniques, and disposal methods."
Clerical	"Knowledge of administrative and clerical procedures and systems such as word processing, managing files and records, stenography and transcription, designing forms, and other office procedures and terminology."
Communication and Media	"Knowledge of media production, communication, and dissemination techniques and methods. This includes alternative ways to inform and entertain via written, oral, and visual media."
Computers and Electronics	"Knowledge of circuit boards, processors, chips, electronic equipment, and computer hardware and software, including applications and programming."
Customer and Personal Service	"Knowledge of principles and processes for providing customer and personal services. This includes customer needs assessment, meeting quality standards for services, and evaluation of customer satisfaction."
Design	"Knowledge of design techniques, tools, and principles involved in production of precision technical plans, blueprints, drawings, and models."
Economics and Accounting	"Knowledge of economic and accounting principles and practices, the financial markets, banking and the analysis and reporting of financial data."

Knowledge Component(s)	Definition of Knowledge Component
Education and Training	"Knowledge of principles and methods for curriculum and training design, teaching and instruction for individuals and groups, and the measurement of training effects."
Engineering and Technology	"Knowledge of the practical application of engineering science and technology. This includes applying principles, techniques, procedures, and equipment to the design and production of various goods and services."
English Language	"Knowledge of the structure and content of the English language including the meaning and spelling of words, rules of composition, and grammar."
Fine Arts	"Knowledge of the theory and techniques required to compose, produce, and perform works of music, dance, visual arts, drama, and sculpture."
Food Production	"Knowledge of techniques and equipment for planting, growing, and harvesting food products (both plant and animal) for consumption, including storage/handling techniques."
Foreign Language	"Knowledge of the structure and content of a foreign (non-English) language including the meaning and spelling of words, rules of composition and grammar, and pronunciation."
Geography	"Knowledge of principles and methods for describing the features of land, sea, and air masses, including their physical characteristics, locations, interrelationships, and distribution of plant, animal, and human life."
History and Archeology	"Knowledge of historical events and their causes, indicators, and effects on civilizations and cultures."
Law and Government	"Knowledge of laws, legal codes, court procedures, precedents, government regulations, executive orders, agency rules, and the democratic political process."
Mathematics	"Knowledge of arithmetic, algebra, geometry, calculus, statistics, and their applications."
Mechanical	"Knowledge of machines and tools, including their designs, uses, repair, and maintenance."
Medicine and Dentistry	"Knowledge of the information and techniques needed to diagnose and treat human injuries, diseases, and deformities. This includes symptoms, treatment alternatives, drug properties and interactions, and preventive health-care measures."

Knowledge Component(s)	Definition of Knowledge Component
Personnel and Human Resources	"Knowledge of principles and procedures for personnel recruitment, selection, training, compensation and benefits, labor relations and negotiation, and personnel information systems."
Philosophy and Theology	"Knowledge of different philosophical systems and religions. This includes their basic principles, values, ethics, ways of thinking, customs, practices, and their impact on human culture.
Physics	"Knowledge and prediction of physical principles, laws, their interrelationships, and applications to understanding fluid, material, and atmospheric dynamics, and mechanical, electrical, atomic and sub- atomic structures and processes."
Production and Processing	"Knowledge of raw materials, production processes, quality control, costs, and other techniques for maximizing the effective manufacture and distribution of goods."
Psychology	"Knowledge of human behavior and performance; individual differences in ability, personality, and interests; learning and motivation; psychological research methods; and the assessment and treatment of behavioral and affective disorders."
Public Safety and Security	"Knowledge of relevant equipment, policies, procedures, and strategies to promote effective local, state, or national security operations for the protection of people, data, property, and institutions"
Sales and Marketing	"Knowledge of principles and methods for showing, promoting, and selling products or services. This includes marketing strategy and tactics, product demonstration, sales techniques, and sales control systems."
Sociology and Anthropology	"Knowledge of group behavior and dynamics, societal trends and influences, human migrations, ethnicity, cultures and their history and origins."
Telecommunications	"Knowledge of transmission, broadcasting, switching, control, and operation of telecommunications systems."
Therapy and Counseling	"Knowledge of principles, methods, and procedures for diagnosis, treatment, and rehabilitation of physical and mental dysfunctions, and for career counseling and guidance."

Knowledge Component(s)	Definition of Knowledge Component
Transportation	"Knowledge of principles and methods for moving people or goods by air, rail, sea, or road, including the relative costs and benefits."

(National Center for O*NET Development, 2020b)

Table 13: O*NET's Skill Components

Skill Component(s)		Definition of Skill Component
Basic Skills	Active Learning	"Understanding the implications of new information for both current and future problem-solving and decision-making."
	Active Listening	"Giving full attention to what other people are saying, taking time to understand the points being made, asking questions as appropriate, and not interrupting at inappropriate times."
	Critical Thinking	"Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems."
	Learning Strategies	"Selecting and using training/instructional methods and procedures appropriate for the situation when learning or teaching new things."
	Mathematics	"Using mathematics to solve problems."
	Monitoring	"Monitoring/Assessing performance of yourself, other individuals, or organizations to make improvements or take corrective action."
	Reading Comprehension	"Understanding written sentences and paragraphs in work related documents."
	Science	"Using scientific rules and methods to solve problems."
	Speaking	"Talking to others to convey information effectively."
	Writing	"Communicating effectively in writing as appropriate for the needs of the audience."
Complex Problem Solving	Complex Problem Solving	"Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions."
Resource Management	Management of Financial Resources	"Determining how money will be spent to get the work done, and accounting for these expenditures."
	Management of Material Resources	"Obtaining and seeing to the appropriate use of equipment, facilities, and materials needed to do certain work."

Skill Component(s)		Definition of Skill Component
	Management of Personnel Resources	"Motivating, developing, and directing people as they work, identifying the best people for the job."
	Time Management	"Managing one's own time and the time of others."
Social Skills	Coordination	"Adjusting actions in relation to others' actions."
	Instruction	"Teaching others how to do something."
	Negotiation	"Bringing others together and trying to reconcile differences."
	Persuasion	"Persuading others to change their minds or behavior."
	Service Oriented	"Actively looking for ways to help people."
	Social Perceptiveness	"Being aware of others' reactions and understanding why they react as they do."
Systems Skills	Judgement and Decision Making	"Considering the relative costs and benefits of potential actions to choose the most appropriate one."
	Systems Analysis	"Determining how a system should work and how changes in conditions, operations, and the environment will affect outcomes."
	Systems Evaluation	"Identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system."
Technical Skills	Equipment Maintenance	"Performing routine maintenance on equipment and determining when and what kind of maintenance is needed."
	Equipment Selection	"Determining the kind of tools and equipment needed to do a job."
	Installation	"Installing equipment, machines, wiring, or programs to meet specifications."
	Operation and Control	"Controlling operations of equipment or systems."
	Operation Monitoring	"Watching gauges, dials, or other indicators to make sure a machine is working properly."

Skill Component(s)		Definition of Skill Component
	Operations Analysis	"Analyzing needs and product requirements to create a design."
	Programming	"Writing computer programs for various purposes."
	Quality Control Analysis	"Conducting tests and inspections of products, services, or processes to evaluate quality or performance."
	Repairing	"Repairing machines or systems using the needed tools."
	Technology Design	"Generating or adapting equipment and technology to serve user needs."
	Troubleshooting	"Determining causes of operating errors and deciding what to do about it."

(National Center for O*NET Development, 2020c)

Table 14: O*NET's Ability Components

Ability Component(s)		Definition of Ability Component
Cognitive Abilities	Cognitive Flexibility	"The ability to generate or use different sets of rules for combining or grouping things in different ways."
	Deductive Reasoning	"The ability to apply general rules to specific problems to produce answers that make sense."
	Flexibility of Closure	"The ability to identify or detect a known pattern (a figure, object, word, or sound) that is hidden in other distracting material."
	Fluency of Ideas	"The ability to come up with a number of ideas about a topic (the number of ideas is important, not their quality, correctness, or creativity)."
	Inductive Reasoning	"The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events)."
	Information Ordering	"The ability to arrange things or actions in a certain order or pattern according to a specific rule or set of rules (e.g., patterns of numbers, letters, words, pictures, mathematical operations)."
	Mathematical Reasoning	"The ability to choose the right mathematical methods or formulas to solve a problem."
	Memorization	"The ability to remember information such as words, numbers, pictures, and procedures."
	Number Facility	"The ability to add, subtract, multiply, or divide quickly and correctly."
	Oral Comprehension	"The ability to listen to and understand information and ideas presented through spoken words and sentences."
	Oral Expressing	"The ability to communicate information and ideas in speaking so others will understand."
	Originality	"The ability to come up with unusual or clever ideas about a given topic or situation, or to develop creative ways to solve a problem."
	Perceptual Speed	"The ability to quickly and accurately compare similarities and differences among sets of letters, numbers, objects, pictures, or patterns. The things to be compared may be presented at the same time or one

Ability Component(s)		Definition of Ability Component
		after the other. This ability also includes comparing a presented object with a remembered object."
	Problem Sensitivity	"The ability to tell when something is wrong or is likely to go wrong. It does not involve solving the problem, only recognizing there is a problem."
	Selective Attention	"The ability to concentrate on a task over a period of time without being distracted."
	Spatial Orientation	"The ability to know your location in relation to the environment or to know where other objects are in relation to you."
	Speed of Closure	"The ability to quickly make sense of, combine, and organize information into meaningful patterns."
	Time Sharing	"The ability to shift back and forth between two or more activities or sources of information (such as speech, sounds, touch, or other sources)."
	Visualization	"The ability to imagine how something will look after it is moved around or when its parts are moved or rearranged."
	Written Comprehension	"The ability to read and understand information and ideas presented in writing."
	Written Expression	"The ability to communicate information and ideas in writing so others will understand."
Physical Abilities	Dynamic Flexibility	"The ability to quickly and repeatedly bend, stretch, twist, or reach out with your body, arms, and/or legs."
	Dynamic Strength	"The ability to exert muscle force repeatedly or continuously over time. This involves muscular endurance and resistance to muscle fatigue."
	Explosive Strength	"The ability to use short bursts of muscle force to propel oneself (as in jumping or sprinting), or to throw an object."
	Extent Flexibility	"The ability to bend, stretch, twist, or reach with your body, arms, and/or legs."

Ability Component(s)		Definition of Ability Component
	Gross Body Coordination	"The ability to coordinate the movement of your arms, legs, and torso together when the whole body is in motion."
	Gross Body Equilibrium	"The ability to keep or regain your body balance or stay upright when in an unstable position."
	Stamina	"The ability to exert yourself physically over long periods of time without getting winded or out of breath."
	Static Strength	"The ability to exert maximum muscle force to lift, push, pull, or carry objects."
	Trunk Strength	"The ability to use your abdominal and lower back muscles to support part of the body repeatedly or continuously over time without 'giving out' or fatiguing."
Psychomotor Abilities	Arm-Hand Steadiness	"The ability to keep your hand and arm steady while moving your arm or while holding your arm and hand in one position."
	Control Precision	"The ability to quickly and repeatedly adjust the controls of a machine or a vehicle to exact positions."
	Finger Dexterity	"The ability to make precisely coordinated movements of the fingers of one or both hands to grasp, manipulate, or assemble very small objects."
	Manual Dexterity	"The ability to quickly move your hand, your hand together with your arm, or your two hands to grasp, manipulate, or assemble objects."
	Multilimb Coordination	"The ability to coordinate two or more limbs (for example, two arms, two legs, or one leg and one arm) while sitting, standing, or lying down. It does not involve performing the activities while the whole body is in motion."
	Rate Control	"The ability to time your movements or the movement of a piece of equipment in anticipation of changes in the speed and/or direction of a moving object or scene."
	Reaction Time	"The ability to quickly respond (with the hand, finger, or foot) to a signal (sound, light, picture) when it appears."

Ability Component(s)		Definition of Ability Component
	Response Orientation	"The ability to choose quickly between two or more movements in response to two or more different signals (lights, sounds, pictures). It includes the speed with which the correct response is started with the hand, foot, or other body part."
	Speed of Limb Movement	"The ability to quickly move the arms and legs."
	Wrist-Finger Speed	"The ability to make fast, simple, repeated movements of the fingers, hands, and wrists."
Sensory Abilities	Auditory Attention	"The ability to focus on a single source of sound in the presence of other distracting sounds."
	Depth Perception	"The ability to judge which of several objects is closer or farther away from you, or to judge the distance between you and an object."
	Far Vision	"The ability to see details at a distance."
	Glare Sensitivity	"The ability to see objects in the presence of glare or bright lighting."
	Hearing Sensitivity	"The ability to detect or tell the differences between sounds that vary in pitch and loudness."
	Near Vision	"The ability to see details at close range (within a few feet of the observer)."
	Night Vision	"The ability to see under low light conditions."
	Peripheral Vision	"The ability to see objects or movement of objects to one's side when the eyes are looking ahead."
	Sound Location	"The ability to tell the direction from which a sound originated."
	Speech Clarity	"The ability to speak clearly so others can understand you."
	Speech Recognition	"The ability to identify and understand the speech of another person."
	Visual Color Discrimination	"The ability to match or detect differences between colors, including shades of color and brightness."

(National Center for O*NET Development, 2020a)

Conclusion

M&S consists of a wide breadth of knowledge. It is expected that graduate level students possess a deep understanding of concepts related to their research area; however, it is also pertinent for these students to communicate knowledge with other team members and domain experts. To achieve this, students must understand general concepts from a range of related disciplines including industrial engineering, mathematics, computer science, digital media, philosophy, human performance, human-computer interaction, education, etc. To date, the identification of such KSAs necessary for the standard M&S professional has proven difficult as professionals and other stakeholders often disagree on which KSAs are *most* important for the M&S discipline (Birta, 2003). Further, Ören & Waite (2010), imply it is possible that more than one type of M&S professional is necessary to meet the requirements of M&S users. The dynamic nature of the current domain presents some difficulties to determining and updating appropriate learning requirements for graduate level M&S education programs. Some elements are the same across M&S graduate programs but there are also many differences.

To clarify topics and specializations in M&S I use an NLP technique called LDA to determine topics from domain documents. I will then use the information from the NLP used to help identify KSAs for the ontology and expert models. Table 15 summarizes the problem space and restates the research objectives, questions, and hypotheses listed throughout the chapter. The following chapter will build upon the literature I have reviewed by addressing, in detail, the methods utilized for this dissertation study.

Table 15: Restated Problem Statement and Research Objectives, Questions, and Hypotheses

Problem Statement	
Problem	Transparency and accountability are increasingly important to higher education stakeholders; thus, as highly complex systems they must showcase their value to sustain. Metrics of success are necessary to articulate this value but are not always quantitative and explicit. This ambiguity is compounded by the fact that technology-related programs evolve quickly, which makes it difficult for faculty and administrators to determine (and quickly update) appropriate curricula to prepare students for the job market.
Research Objectives	
Objective 1	Investigate current resource allocation and strategic planning models in higher education, automated curriculum management, and graduate student success factors to determine an appropriate plan for designing and developing a holistic university-wide software system.
Objective 2	Conceptualize a university-wide modular decision-making software solution to inform curricula, optimize student paths toward degree completion, and optimize resources to meet program and university objectives.
Objective 3	Develop an M&S ontology using topic modeling techniques from all source types (job listings, course descriptions, and academic publications) and compare each source type to determine if there is a disconnect between requested, taught, and applied KSAs.
Objective 4	Develop M&S expert models using topic modeling techniques.
Research Questions	
Research Question 1	What are the KSAs applied most frequently in M&S academic literature?
Research Question 2	What are the KSAs most requested in M&S job listings within the United States?
Research Question 3	How should M&S job types be categorized?
Research Question 4	What are the KSAs most identified per job type?
Research Question 5	What are the KSAs taught most frequently in M&S graduate level course descriptions for Universities within the United States?
Research Hypotheses	
Hypothesis 1	Scraping multiple types of M&S documents (job postings, course descriptions, and academic publications) will produce a difference in the KSAs (topics) most frequently mentioned in each source type.

CHAPTER THREE: METHODOLOGY

The methods presented in this chapter are intended to address the purpose of this dissertation, which is to investigate natural language within the Modeling and Simulation (M&S) field using domain-related documents to determine the priority of and relationships between M&S knowledge, skills, and abilities (KSAs). To do this, I have organized this chapter to address important aspects of the dissertation study methodologies to include study design, procedures materials, and planned analyses and outputs.

Study Design

The methodology for the present dissertation includes a qualitative investigation of natural language within various M&S professional documents. The independent variable for this dissertation includes the source type (job postings, course descriptions, and academic articles). The KSAs identified are the dependent variable. The output will include qualitative categorical data, in the form of salient key terms. I then categorize key terms identified them based the O*NET KSA schema using qualitative thematic analysis.

Study Procedures

The process of gathering and creating the model includes several steps. In this instance, I use Vajjala and colleagues' (2020) NLP Pipeline. I considered this and another procedure from Ameisen, (2020), however Ameisen takes a general machine learning approach versus Vajjala and colleagues' (2020), who outline procedures specifically for Natural Language Processing

(NLP) with topic modelling examples. Other applied examples use a similar approach to the one outlined by Vajjala and colleagues (Galli, 2018, 2020; Zhao, 2018). The steps of their process are included in Table16. For the present dissertation, I will not be completing steps seven and eight in the pipeline below because that includes deploying the NLP model into the overall-university wide system, which is yet to be developed.

Table 16: Vajjala and Colleagues'(2020) NLP Pipeline

- Step 1: Data Acquisition
- Step 2: Text Cleaning
- Step 3: Pre-Processing
- Step 4: Feature Engineering
- Step 5: Modeling
- Step 6: Evaluation
    - Can iterate and improve the model and repeat from pre-processing step
- Step 7: Deployment
- Step 8: Monitoring and Model Updating
    - Can iterate back to beginning and refine model

(Vajjala et al., 2020)

Data Acquisition

The present dissertation looks at various M&S documents to determine and prioritize appropriate domain wide KSAs. I wanted to investigate the requested KSAs, the KSAs taught, and the KSAs applied in practice. Thus, data selection included a job-based dataset, a course-based dataset, and a publication-based dataset, which was suggested by the dissertation committee. Machine Learning and NLP best practices encourage finding an existing dataset before creating one (Ameisen, 2020; Vajjala et al., 2020). As such, I referenced existing, public datasets which included general repositories such as Google's Dataset Search, Internet Archive's Academic Torrents, and the University of California Irvine's Machine Learning Repository (*Academic Torrents*, 2014; Dua & Graff, 2019; Google, 2020). Search terms used included

"Modeling and Simulation," "Education," "Higher Education," "Technology Jobs," and "Technology Job Skills".

Existing Publication Data

While many article databases exist, to the best of my knowledge there are no existing M&S specific publication-based, natural language-based datasets.

Existing Job-Based Datasets

Data selection for jobs included the investigation of the Occupational Information Network (O*NET) 25.0 database a "rich set of variables that describe work and worker characteristics, including skill requirements"(U.S. Department of Labor et al., 2020). The O*NET Database collects KSA data on worker characteristics, worker requirements, experience requirements, occupational requirements, workforce characteristics, and occupation-specific information (U.S. Department of Labor et al., 2020). However, the usefulness of this source for this study was limited in that there is no occupation or category of occupations directly associated with M&S. For example, when searching *modeling*, the results yielded 20 occupations, two of which are related to fashion modeling, which is not applicable to the present dissertation. The other occupations listed are jobs aligned with similar domains, such as computer science, software development, and mathematics. This contributes to the confusion between M&S and related domains like Computer Science. Identifying and prioritizing M&S

KSAs is especially pertinent if M&S intends to forge forward as a stand-alone domain rather than a sub-field of computer science.

Existing Course Data

Another source of M&S information comes from practitioners focused on teaching their skills through university courses. Collection of course data for this study included consideration of syllabi, required texts, graduate program listings, and alternative datasets. The National Center for Education Statistics has published course datasets but were determined to be too general for use in the present dissertation (i.e., didn't discuss M&S specifically) or related to a different level of education (e.g., primary and secondary education rather than post-secondary; U.S. Department of Education, n.d.).

Another source considered was big data project the Open Syllabus Explorer, which includes a mixture of information about textbooks as well as syllabi (Karaganis et al., n.d.). Information about textbooks collected on this site included the book's identifying information (e.g. title, author), ranking, appearances, score (unique to the site), field of study most often associated with the text, location of the institution offering the course using the text, and texts frequently paired with current text (Karaganis et al., n.d.). While the developers of the Open Syllabus Explorer have organized some topics (e.g., Computer Science and Mathematics), there is no M&S specific section. However, this data is still helpful because it can be used to verify other domains in the future, but it is outside of the scope of the present dissertation. Since my hypothesis suggests is that there is a disparage between the current M&S job expectations and curriculum, I wanted to investigate M&S specific program information rather than a more generalized population.

Web Scraping New Datasets

If an existing dataset doesn't exist, the alternative is to instead create one (Vajjala et al., 2020). As such, I used a web parsing method to *scrape*, *crawl*, or *collect*, data from various web-based sources. Web scraping (data extraction) is "a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format," (SysNucleus, n.d.). "Even when you have to develop some sort of generic search engine (say, a blog search engine), you're unlikely to encounter a scenario where you should design your own crawler. Production-ready crawlers, such as Apache Nutch and Scrapy , can be customized and used for your project in such scenarios, (Vajjala et al., 2020).

Using one such program, called Octoparse©, I built a few web scrapers to extract data from each website grouping (*Octoparse*, 2020). Octoparse is an application that allows the user to operate a drag-and-drop graphical user interface (GUI) to build a web scraper. Each job listing website required two scrapers – one to pull the uniform resource locator (URL) information and another to pull the job details listed at each URL pulled. Data exports into a comma separated (.csv) file.

One of the limitations of using web parsing is that it uses XML to locate the information on the web page. However, some of the websites, for example Indeed.com, do not have uniform fields on their web page beyond the job title and company (see Figure 8). Even the ratings and job location are often switched and/or missing. Further, the rest of the job description is posted as one large chunk of text. This prompted the use of data cleaning methods on text collected.

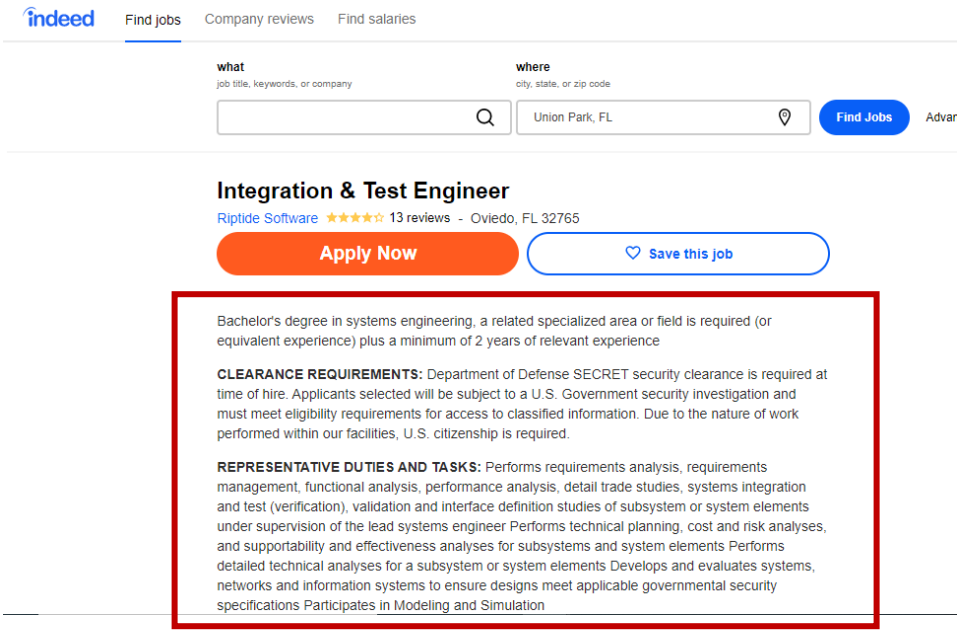

Figure 8: Indeed.com Example Job Posting (Integration & Test Engineer, 2019)

Scraping Job Data

I sought to find data through "scraping job postings websites such as Indeed.com and the *Chronicle of Higher Education*'s job center, *Chronicle Vitae*. While other career and job-related websites (such as LinkedIn) were publicly available, they were eliminated from the scope of this dissertation, 1) due to the nature of how they were structured (not easily scrapable) and 2) to focus on data sources that both produce a high volume of job postings and higher caliber academic positions (jobs for which a graduate degree should prepare students).

Scraping Course Data

Course syllabi were considered first for the web scraped data as one of the committee members suggested this as a means of looking at university and course related data. However, it is not common practice to publicly post syllabi. The syllabus for a course often has information related to *how* the course is taught as well. Therefore, some university administrators discourage

posting syllabi for public use because they feel like *how* a course is taught is what makes the university, program, and course unique. To automate the system, I wanted to utilize publicly available data. While it is not common practice to post syllabi, it is common practice to post general course descriptions and program requirements (see Figure 9). As such, I chose to scrape M&S course descriptions for the present dissertation.

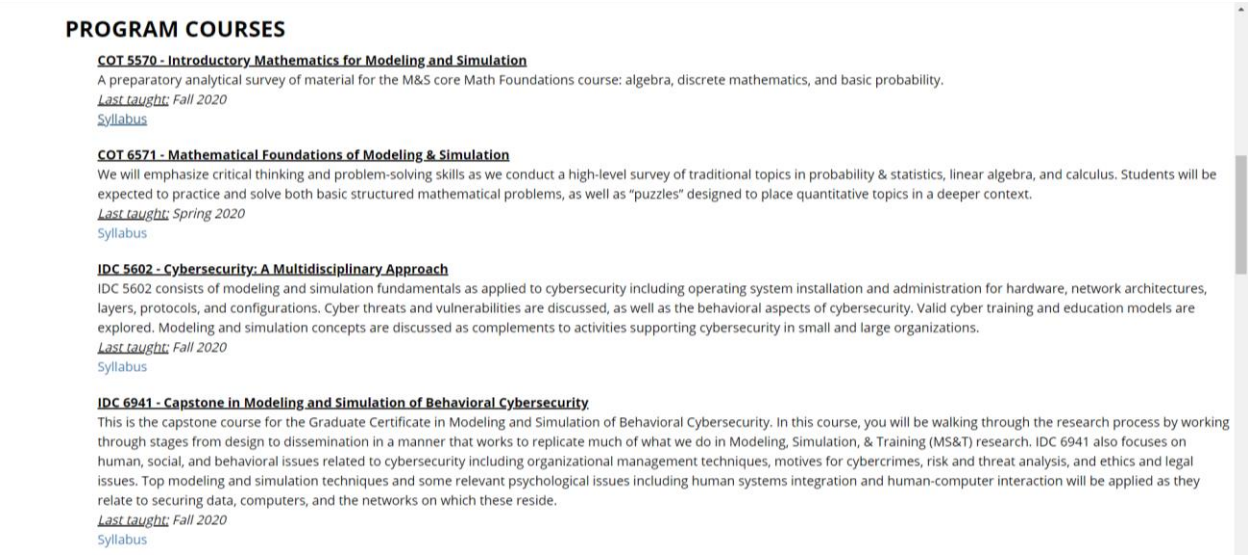

Figure 9: Program Course Description Example

Scraping Publication Data

Open source journals were selected for this study as a sample of convenience. Originally, I intended to scrape Google Scholar, however I kept running into issues with the browser noticing the scraper was a robot and would discontinue the scrape. As such I started looking for open-source journal databases as an alternative to Google Scholar. Scientific Research (2020), an open-source publishing company for academic peer-reviewed articles, had some M&S specific related journals such as, the Open Journal for Modeling and Simulation. So, I searched for

"Modeling and Simulation," and then scraped all the abstracts and keywords from the resulting articles.

Sample Size

"In an ideal setting, we'll have the required datasets with thousands—maybe even millions—of data points," (Vajjala et al., 2020). The data is *largely* affected by the number of *available* documents and the *quality* of the data (*The Size and Quality of a Data Set*, 2019). A few general rules exist. Some of these rules include collecting 1,000 documents per factor or 10 times the degrees of freedom (e.g., 3 factors - 30 data points; *The Size and Quality of a Data Set*, 2019). There seems to be little consensus on what an appropriate sample size of documents per set of text is appropriate. However, I referenced the original source article for LDA and looked at the sample size used in their analysis of the model. In their experiments Blei et al. (2002) analyzed two different sized data sets. One consisted of. 2,500 new articles with a vocabulary size of $|V| = 37{,}871$ words, and the other included 1400 technical abstracts with a vocabulary size of $|V| = 7{,}747$ words (Blei et al., 2002). As such, I collected as many documents as I could while scraping each source. Octoparse automatically cleans up repeated URLs in the output files and deletes repeated job posts and articles from the corpus based on the information within the posting. This is to reduce repeat data/postings (convoluting the data) with the same job opening. The resulting sample sizes of unique documents found are listed in Table 17.

Table 17: Sample Population of Documents

Data Source	Unique Documents Found
Academic Publications	*n* = 2001
Job Postings	*n* = 640
Course Descriptions	*n* = 130

Text Cleaning

Next, I created a **corpus** (plural **corpora**), which is a collection of texts/documents (Zhao, 2018). Typically, the data from these sources has additional characters (e.g., HTML code; Zhao, 2018). Therefore, I used **regular expressions,** "a special text string for describing a search pattern," to clean the text within the corpus to make it easier to analyze (Goyvaerts, 2020). This data cleaning process involved changing all letters to lower case, removing punctuation, removing numbers within words, and removing other non-sensical text such as line breaks. I also used regular expressions to remove text within curly and square brackets, removing large chunks of white space, and correct common misspellings.

Pre-Processing Data

I then **tokenized** the data, which breaks down the corpus text into smaller parts – either sentences or words (Vajjala et al., 2020). At the same time, I also removed stop words. **Stop words** include smaller connecting words that do not really add value but can help connect thoughts (Brownlee, 2017a). For example, "a," "the," "which," "that," "those," etc. are common stop words. For this study, I used the Natural Language Tool Kit's (NLTK's) English stop word library. However, prior to removing stop words, I used regular expressions to replace any instance of "modeling and simulation" with M&S. The reason for this was to keep any instance of M&S mentioned as a field of study versus any individual instance of the words "modeling," and "simulation." This reduces the chance of misinterpreting the context for the phrase with the

removal of the word “and”. “Modeling,” and “simulation,” are different from M&S. From there I **lemmatized** the data, which truncates the word into a **lemma** or root word. For example, the lemmatizer changes “engineers” to “engineer” and thus counts the term “engineer” twice rather than as two separate terms (Vajjala et al., 2020). For the present dissertation I used the WordNet Lemmitizer because it is a commonly used lemetizer in NLP tasks (Zhao, 2018).

Feature Engineering

The output of the BoW and BoN analyses are a collection of **Document Term Matrices** [(DTM); Zhao, 2018]. These will be used to determine the most frequently occurring words (or features) in each corpus. **Term Frequency-Inverse Document Frequency** (TD-IDF) results are similar to a DTM, however they show the likelihood that a term will occur in a document, but also can account for the number of times that topic is discussed in all of the documents within the corpus.

Modeling

The DTM is used as the input for the **Latent Dirchlet Allocation** (LDA), the topic modeling technique for the present dissertation. When broken down, LDA encompasses a few different components. *Latent* in this case means hidden because the labels or features are not yet known (Vajjala et al., 2020; Zhao, 2018). *Dirichlet* is a type of probability distribution, (Zhao, 2018). Essentially, Dirchlet distribution is a probability distribution of probability distributions (Vajjala et al., 2020). “In this sense, the model generates an *allocation* of the words in a document to topics. When computing the probability of a new document, this unknown allocation induces a mixture distribution across the words in the vocabulary. There is a many-to-many relationship between topics and words as well as a many-to-many relationship between

documents and topics," meaning Dirichlet is a component of *likelihood* (Blei et al., 2002). To run the LDA analysis, the function will need the DTM, a number of topics (starting with two is common practice), and the number of iterations the model will go through (Vajjala et al., 2020; Zhao, 2018).

Evaluation

Two common LDA specific evaluation methods of inference also include perplexity and coherence (Magesh, 2019). "**Perplexity** is the measure of uncertainty, meaning lower the perplexity better the model," (Magesh, 2019). While perplexity is a common score in NLP tasks, optimizing the model results based on perplexity alone may yield non-sensical results to the human; coherence scores can be used in this case to determine how interpretable a model is (Magesh, 2019). "**Coherence** is the measure of semantic similarity between top words in our topic. Higher the coherence better the model performance," (Magesh, 2019).

Additionally, LDA model parameters can be used to fine tune the model. Looking at perplexity and coherence during model tuning will help determine an appropriate topic model. These parameters include the number of topics, the number of iterations or the words within the DTM (Zhao, 2018). Determining the appropriate number the of topics in corpus is a trial and error process, but it is common best practices to start with two features and work your way up (Zhao, 2018). I plotted a coherence model by number of topics to help determine the largest coherence score as a way of estimating the optimal number of topics based on the data (Blei et al., 2002; Magesh, 2019).

The number of times the model iterates can increase the likeliness that the model will make more sense to the human later because of this reason. A common number of iterations to

start with is 10, however, others have used 50 or 100 iterations. In the present dissertation, I chose to implement 10 iterations. I can also look at the words within the DTM to determine reliability.

Reliability refers to the consistency or trustworthiness of a measure (*The Size and Quality of a Data Set*, 2019). In NLP unreliable data can appear as omitted values, duplicate examples, bad labels, and bad feature values (*The Size and Quality of a Data Set*, 2019). In the present dissertation, I took several steps to increase data reliability. Omitted values were dropped using a regular expression looking for non-value data. In this case the document is omitted, but the "document" is one chunk of text, meaning no incomplete data was used. As mentioned, Octoparse automatically detects duplicate examples of the URLs scrapped and discards them, thereby dealing with duplicate documents. Further, I did not use labeled data because the intention of the present dissertation is to investigate unlabeled data using unsupervised machine learning algorithms to find appropriate categories in which to label data. Therefore, the potential of having bad labels due to human error was a non-issue. Bad feature value factors can be introduced by error during the data acquisition process; examples include experimenter accidently entering an extra digit or equipment malfunctioning (*The Size and Quality of a Data Set*, 2019). In the present dissertation, bad feature values were identified during the data cleaning steps.

I only identified one issue with the feature values. There were common instances where the web crawler incorrectly scraped data and combined words. This was particularly true of the job posting data, it occurred a few times in the publication abstract data but did not occur in the course descriptions data. In this instance, I went through the terms and manually cleaned the data

using regular expressions. Many of these functions included searching for a commonly combined word and adding a space after it. Various forms of the word were taken into account while using regular expressions.

Study Materials

I used several technological applications for my study analysis. Jupyter Notebook 6.0.3 an open-source web-based computational integrated development environment, using the Python 3 programming language to clean, pre-process, and build NLP Models (Project Jupyter, 2020). I used the Pandas, Regular Expressions, Natural Language Tool Kit 3.5, SciKit Learn, NumPy 1.19.0, GenSim 4.0, Matplotlib 3.3.2, and pyLDAvis programming libraries to help me clean, fit, transform, model, visualize, and evaluate the data based on various tutorials and documentation (Galli, 2018, 2020; Kinsley, 2015; Matplotlib Development Team, 2020; NTLK Development Team, 2020; Numpy Development Team, 2020; Pandas Development Team, 2020; Rehurek, 2020; Sievert & Shirley, 2015; Zhao, 2018). In Table 18, below, a high-level description of each library and its purpose is provided to summarize these study materials.

Table 18: Study Materials

Library	Description	NLP Step	Purpose
Pandas (PD)	Python library that allows users to import data into a data frame for manipulation.	Text Cleaning	Creates corpora
Regular Expressions (Re)	Python library for processing text via regular expressions	Text Cleaning	Cleans corpora
Natural Language Tool Kit (NLTK) 3.5	Python library for processing text via classification, tokenizing, stemming, etc.	Pre-Processing Data	PoS Tagging, Remove Stop Words, Tokenize, and Lemmatize Data
SciKit Learn (SKLearn)	Python library for machine learning, allows transformation and normalization	Feature Engineering	Manipulates corpora into DTM
GenSim 4.0	Python library for topic modeling and data visualization	Modeling	Models LDA &
NumPy 1.19.0	Python library used for scientific computing.	Modeling, Evaluation	Performs various calculations on text
Matplotlib 3.3.2	Python library for creating visualizations (static, animated, and interactive)	Modeling, Evaluation	Visualizes TF-IDF and Coherence results
pyLDAvis	R library (adapted for python) for interactive data visualizations	Modeling, Evaluation	Visualizes LDA results

(Matplotlib Development Team, 2020; NTLK Development Team, 2020; Numpy Development Team, 2020; Pandas Development Team, 2020; Rehurek, 2020; Sievert & Shirley, 2015; Vajjala et al., 2020)

Planned Outputs

Various steps have different inputs and outputs. Theses inputs and outputs are summarized below in Table 19. The final visualizations will be able to help determine inference. These visualizations include the most salient terms in each source type (job postings, course descriptions, and open-source academic publications) and intertopic distance maps, showing the relationships between topics within the job corpus. Additionally, perplexity and coherence scores are plotted as well to determine the most appropriate job posting model parameters.

Table 19: Summary of Methodology

Research Question	Data Sources /Input(s)	Analysis	Output(s)
1. What are the KSAs applied most frequently in M&S academic literature?	Open source peer-reviewed journal articles on M&S topics	TF-IDF Vectorization	M&S Applied KSAs - Top Most Salient Terms
2. What are the KSAs most requested in M&S job listings within the United States?	Chronicle and Indeed job posting data	TF-IDF Vectorization	M&S Requested KSAs - Top Most Salient Terms
3. How should M&S job types be categorized?	M&S Requested KSAs DTM	LDA Modeling	M&S Requested KSAs – Intertopic Distance Map
4. What are the KSAs most identified per job type?	M&S Requested KSAs DTM and Topic Model Categories	TF-IDF Vectorization of Models	M&S Requested KSAs - Top Most Salient Terms by Topic
5. What are the KSAs taught most frequently in M&S graduate level course descriptions for Universities within the United States?	Publicly available M&S program course descriptions	TF-IDF Vectorization	M&S Requested KSAs - Top Most Salient Terms

Operationalizing Unigrams and Bigrams

"A topic model only gives a collection of keywords per topic. What exactly the topic represents and what it should be named is typically left to human interpretation," (Vajjala et al., 2020). This statement means, I will have to assign labels to the data based on the results. To complete this step, I used a **thematic analysis**, which is "a search for themes that emerge as being important to the description of the phenomenon," (Boyatzis, 1998; Fereday & Muir-Cochrane, 2006). The goal of thematic analysis to determine patterns by visually iterating over the data, performing a "careful reading and re-reading of the data" (Boyatzis, 1998; Fereday & Muir-Cochrane, 2006). First step in the thematic analysis is to recognize important terms prior to interpreting them (Boyatzis, 1998; Fereday & Muir-Cochrane, 2006). Recognizing the terms as domain specific will help identify appropriate themes to categorize data (Boyatzis, 1998; Fereday & Muir-Cochrane, 2006) For example, the term *agent-based* could signal the use of the discrete, continuous, and agent-based simulation paradigm. In the present dissertation a **theme** is defined as "a pattern in the information that at minimum describes and [organizes] the possible observations and at maximum interprets aspects of the phenomenon" (Boyatzis, 1998 p. 161). Once salient and notable terms are identified, I map these terms common themes and to the scheme outlined by O*NET (U.S. Department of Labor et al., 2020). The purpose of this step is to de able to determine which KSAs are most important to M&S professionals but also to start determining the relationships within and between the terms and themes/classes, which will be used to inform the later ontology.

Conclusion

The present dissertation compromises of a qualitative investigation of natural language within various M&S professional documents including job postings, course descriptions, and publication data. I used steps one through six of Vajjala and colleagues' (2020) NLP pipeline as the procedure for the present dissertation, which includes data acquisition, text cleaning, pre-processing data, feature engineering, modeling, and evaluation. To complete these steps, I used a number of Python libraries including Pandas, Regular Expressions, Natural Language Tool Kit 3.5, SciKit Learn, NumPy 1.19.0, GenSim 4.0, Matplotlib 3.3.2, and pyLDAvis. Topic modeling planned outputs include most salient terms in each source type (job postings, course descriptions, and open-source publication abstracts), intertopic distance maps, and perplexity and coherence scores, showing the relationship between topics within the job posting corpus.

CHAPTER FOUR: RESULTS

In the present dissertation I investigated natural language within the M&S domain to include M&S job postings (i.e., what employers request), M&S course descriptions (i.e., what is being taught), and M&S academic literature (i.e., what is applied in practice) to determine common relationships between M&S topics. To do this, I used a combination of Term Frequency-Inverse Document Frequencies (TF-IDF) and Latent Dirichlet Allocation (LDA) models. The output of these models included the top most salient terms for each source type (job postings, course descriptions, and open-source publication abstracts/keywords) and for the LDA Model I also produced intertopic distance maps, and coherence scores, showing the relationship between topics within the job posting corpus. The following sections present the analyzed data for the stated research questions and hypothesis. The results for each of these are presented in detail using the publication abstracts/keywords (applied KSAs), job postings (requested KSAs), and course descriptions (taught KSAs).

Publication Abstracts TF-IDFs

Research Question 1*:* W*hat are the KSAs applied most frequently in M&S academic literature*? To answer my first research question, I investigated natural language in publication abstracts and created two TF-IDF Document Term Matrices (DTM; i.e., unigrams, bigrams) to help enumerate important terms in the overall ontology. For the DTMs, I chose to limit the unigram DTM to only nouns, adjectives, and verbs to attempt to find greater meaning within the salient terms. Bigrams inherently have more meaning than unigrams. As such I chose not to tag

the parts of speech (PoS). The top 30 most salient unigrams within the publication abstracts corpus were generated. Results were highest for the terms *model* (*M*=.044) and *simulation* (*M*=.029), which is expected in a corpus based on a search term like, "Modeling and Simulation." The next most salient terms included *system* (*M*=.025), *result* (*M*=.023), and *method* (*M*=.021). The lowest results were generated for the terms *structure* (*M*=.012), *show* (*M*=.011), and *performance* (*M*=.011). Figure 10 shows the top 30 most salient unigrams within the publication abstracts corpus.

While some of the top most salient words are terms expected in all types of publication abstracts, notable terms based on the thematic analysis were identified. When mapped these terms fall into O*NET knowledge component categories of Communication and Media (e.g., *paper*), Design (e.g., *design, structure*), Engineering and Technology (e.g., *simulation, data, process*), Mathematics (e.g., *analysis, distribution, equation*), Physics (e.g., *energy, power*) and Production and Processing (e.g., *process*) categories. These were further mapped to the related O*NET skills of critical thinking, mathematics, monitoring, reading comprehension, science, writing, complex problem solving, management of material resources, management of personnel resources, time management, persuasion, judgement and decision making, systems analysis, systems evaluation, equipment selection, operation and control, operation analysis, quality control analysis and technology design. O*NET abilities of cognitive flexibility, deductive reasoning, flexibility of closure, inductive reasoning, number facility, perceptual speed, speed of closure, time sharing, visualization, written comprehension, written expression, and control precision were also identified. Further thematic analysis and O*NET coding are discussed in the following sections.

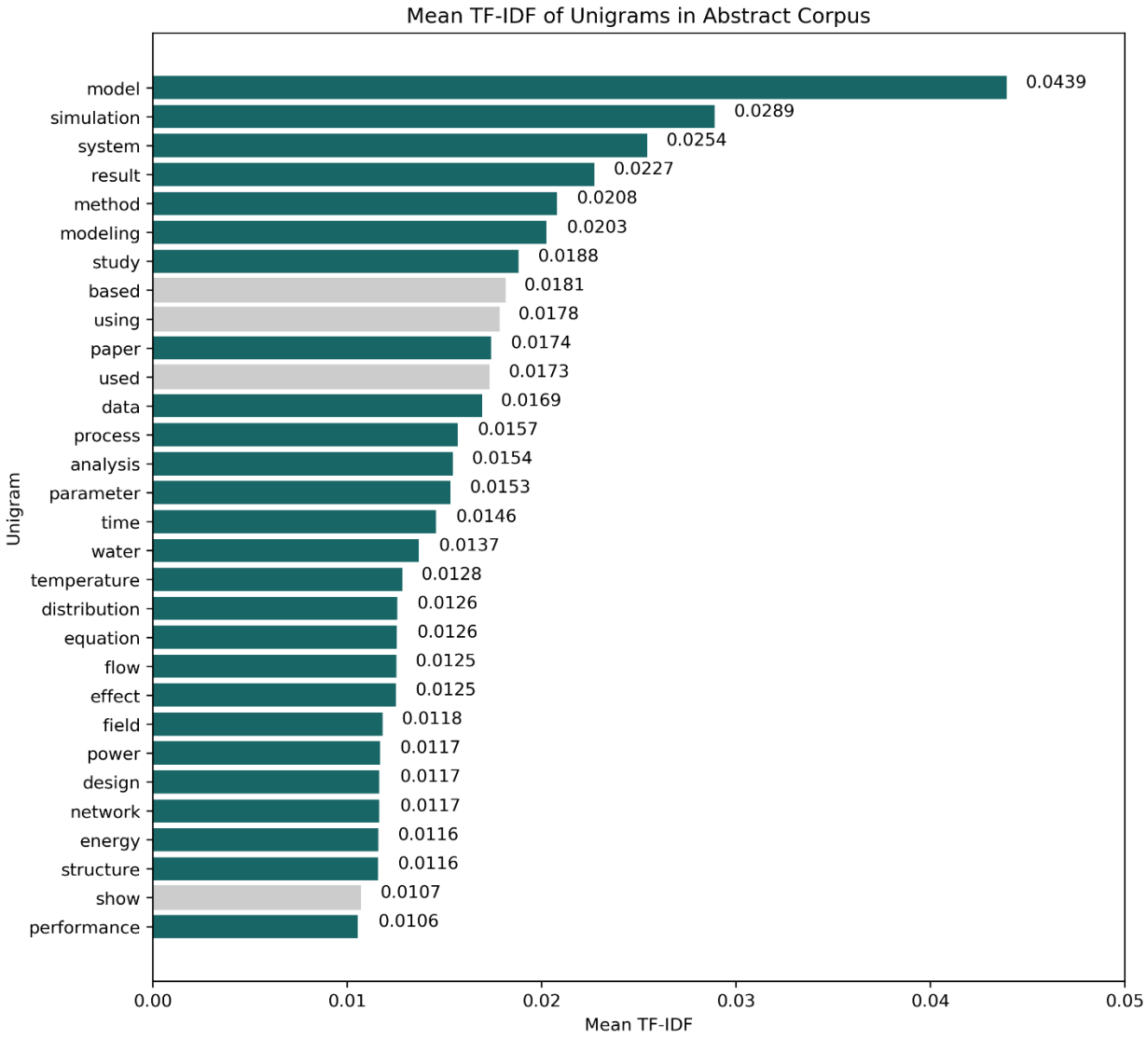

Figure 10: Top 30 Most Salient Unigrams within Publication Abstracts Corpus

A bigram model was unable to be computed due to corpus size and computing power. I had a similar issue with my job posting corpus data. As a result, I attempted to reduce the corpus size and only used publication keywords rather than the entire abstract. The data collected with Octoparse labeled the keywords separately from the abstract text. As such, I also looked at both the unigrams and bigrams for the publication keywords.

The top 30 most salient unigrams within the publication keywords were also generated. Results were highest for the terms *simulation* (*M*=.045), *modeling* (*M*=.044), and *model* (*M*=.024), all of which are expected. The lowest results were generated for the terms *field* (*M*=.006), *structure* (*M*=.006), and *design* (*M*=.006). Figure 11 shows the top 30 most salient unigrams within the publication keyword corpus. When mapped these terms fall into O*NET knowledge component categories, terms fall within the Computers & Electronics (e.g., *computer*), Design (e.g., *design*), Mathematics (e.g., *mathematical, optimization, distribution, equation*), Production and Processing (e.g., *process, structure, design*) categories These were further mapped to the related O*NET skills of critical thinking, mathematics, monitoring, reading comprehension, science, writing, complex problem solving, management of material resources, management of personnel resources, time management, persuasion, judgement and decision making, systems analysis, systems evaluation, equipment selection, operation and control, operation analysis, quality control analysis and technology design. O*NET abilities of cognitive flexibility, deductive reasoning, flexibility of closure, inductive reasoning, number facility, perceptual speed, speed of closure, time sharing, visualization, written comprehension, written expression, and control precision. Further thematic analysis and O*NET coding are discussed in the following sections.

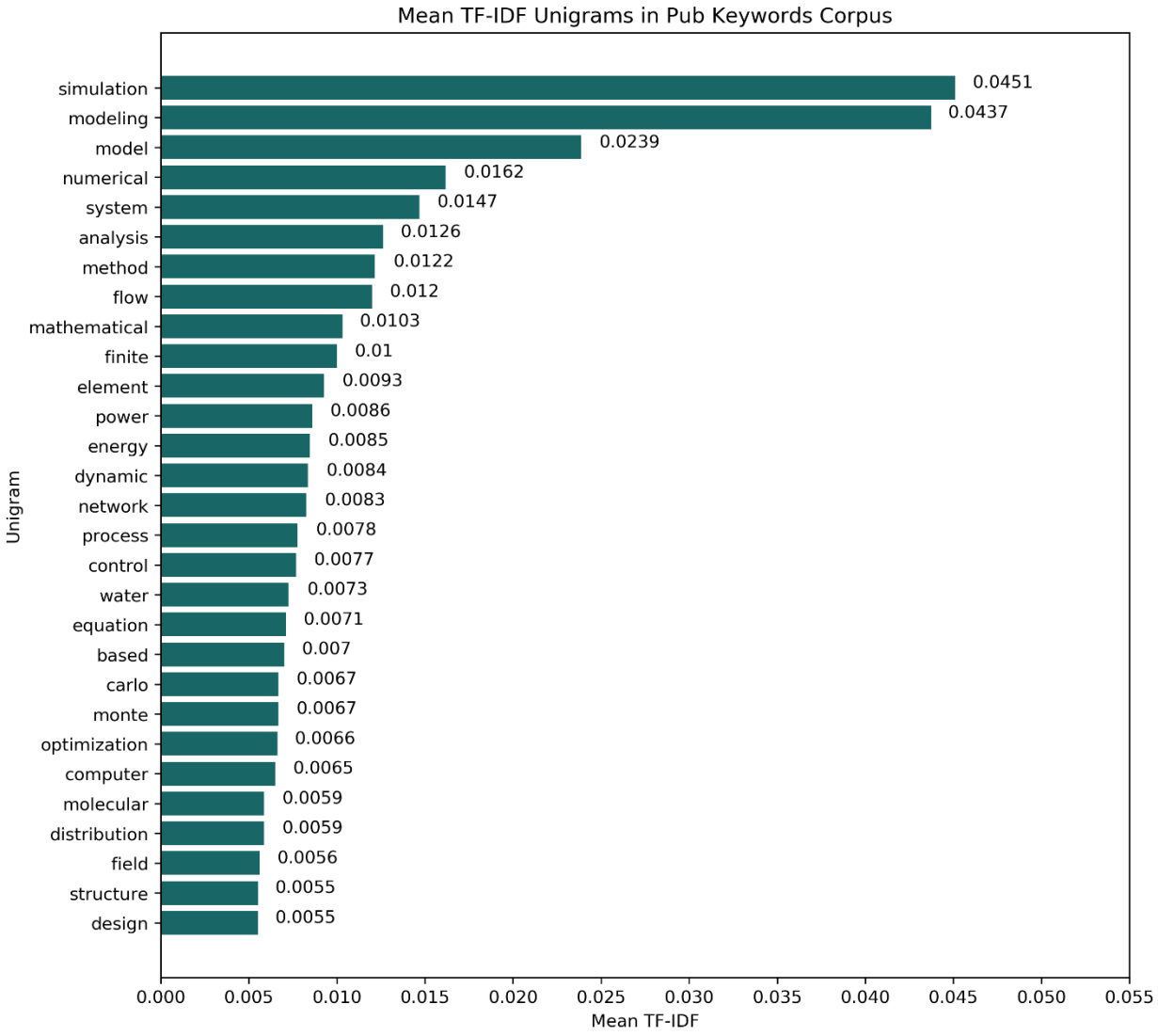

Figure 11: Top 30 Most Salient Unigrams within Publication Keyword Corpus

The top 30 most salient bigrams within the publication keywords were also generated. Results were highest for the terms *numerical simulation* (*M*=.05), *finite element* (*M*=.04), and *mathematical modeling* (*M*=.03). The lowest results were generated for the terms *based modeling* (*M*=.007), *business process* (*M*=.007), and *molecular dynamic* (*M*=.006). Figure 12 shows the

top 30 most salient bigrams within the publication keywords corpus. All of the terms identified as salient were meaningful. When mapped these terms fall into O*NET knowledge component categories, terms fall within the Administration and Management (e.g., *business process*), Biology (*neural networks artificial neural, molecular dynamic*), Chemistry (*element method, element modeling*, *neural networks artificial neural*, *molecular dynamic*). Computers & Electronics (e.g., *computer simulation*), Engineering and Technology (e.g., *power system*), Mathematics (e.g., *monte carlo, numerical simulation, mathematical modeling*), and Production and Processing (e.g., *simulation optimization, business process*) categories. These were further mapped to the related O*NET skills of critical thinking, mathematics, monitoring, reading comprehension, science, writing, complex problem solving, management of material resources, management of personnel resources, time management, persuasion, judgement and decision making, systems analysis, systems evaluation, equipment selection, operation and control, operation analysis, quality control analysis and technology design. O*NET abilities of cognitive flexibility, deductive reasoning, flexibility of closure, inductive reasoning, number facility, perceptual speed, speed of closure, time sharing, visualization, written comprehension, written expression, and control precision. Further thematic analysis and O*NET coding are discussed in the following sections.

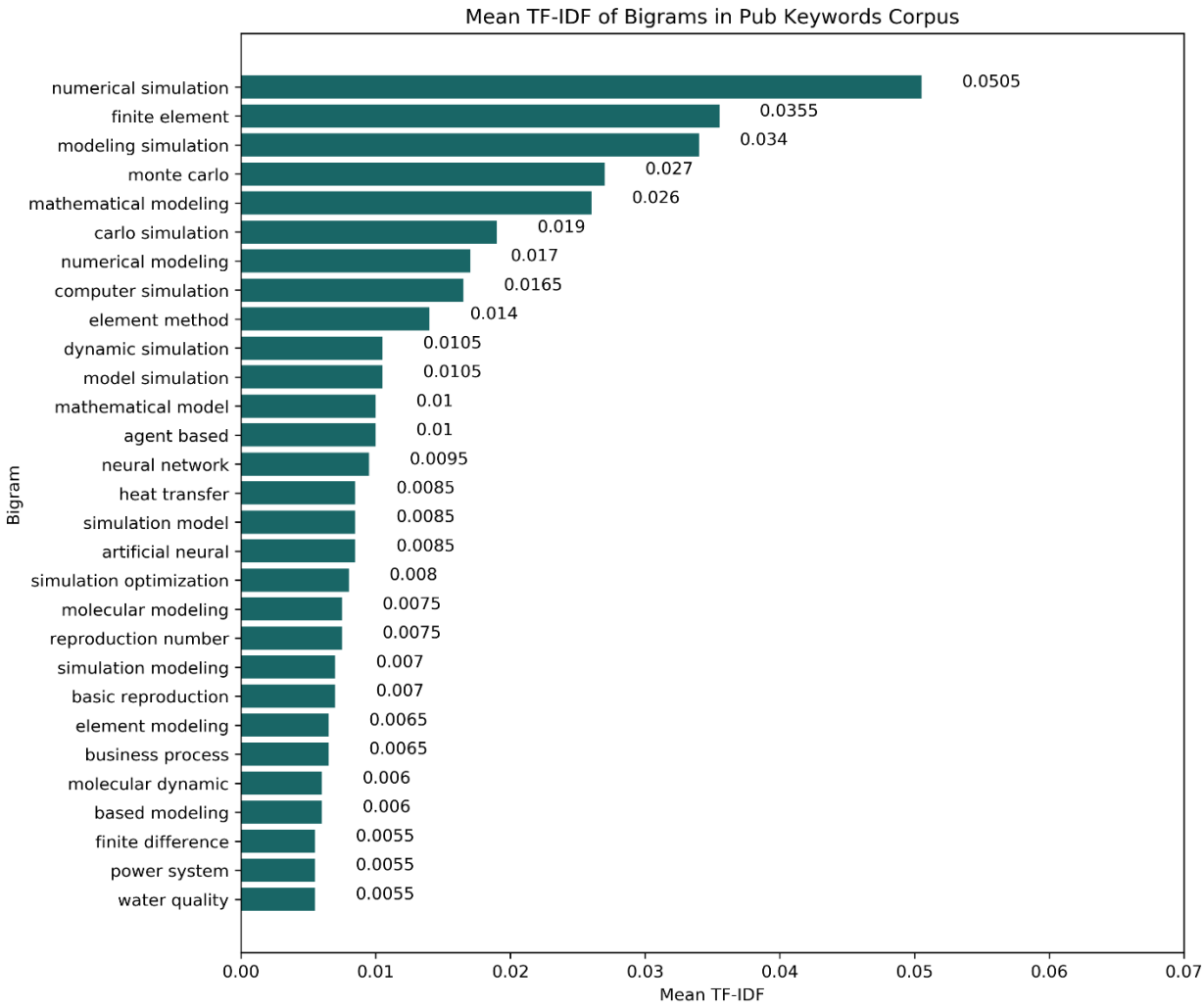

Figure 12: Top 30 Most Salient Bigrams within Publication Keyword Corpus

Using the three different salient lists, I map the terms to the knowledge components, skills, and abilities and identified by O*NET in Tables 20, 21, and 22 respectively

Table 20: O*NET's Knowledge Components Compared to TF-IDF Salient Terms in Abstract Corpora

Knowledge Component(s)	Definition of Knowledge Component	Notable Related Salient Terms
Administration and Management	"Knowledge of business and management principles involved in strategic planning, resource allocation, human resources modeling, leadership technique, production methods, and coordination of people and resources."	*process, performance, business process, agent based*
Biology	"Knowledge of plant and animal organisms, their tissues, cells, functions, interdependencies, and interactions with each other and the environment"	*water, neural networks artificial neural, molecular dynamic*
Chemistry	"Knowledge of the chemical composition, structure, and properties of substances and of the chemical processes and transformations that they undergo. This includes uses of chemicals and their interactions, danger signs, production techniques, and disposal methods."	*element method, element modeling, neural networks, artificial neural, molecular dynamic*
Communication and Media	"Knowledge of media production, communication, and dissemination techniques and methods. This includes alternative ways to inform and entertain via written, oral, and visual media."	*paper*
Computers and Electronics	"Knowledge of circuit boards, processors, chips, electronic equipment, and computer hardware and software, including applications and programming."	*computer, network*
Design	"Knowledge of design techniques, tools, and principles involved in production of precision technical plans, blueprints, drawings, and models."	*model, process, design, structure, simulation model, mathematical model, mathematical modeling, element modeling*

Knowledge Component(s)	Definition of Knowledge Component	Notable Related Salient Terms
Engineering and Technology	"Knowledge of the practical application of engineering science and technology. This includes applying principles, techniques, procedures, and equipment to the design and production of various goods and services."	*simulation, system, design, structure*
Mathematics	"Knowledge of arithmetic, algebra, geometry, calculus, statistics, and their applications."	*model, simulation, modeling, data, analysis, parameter, distribution, equation, simulation results, numerical simulation, finite element, simulation model, mathematical model, monte carlo, mathematical modeling, neural network*
Physics	"Knowledge and prediction of physical principles, laws, their interrelationships, and applications to understanding fluid, material, and atmospheric dynamics, and mechanical, electrical, atomic and sub- atomic structures and processes."	*water, temperature, energy, power, field, flow, heat transfer*
Production and Processing	"Knowledge of raw materials, production processes, quality control, costs, and other techniques for maximizing the effective manufacture and distribution of goods."	*system, method, process, analysis, effect, design, show, performance*

(National Center for O*NET Development, 2020b)

Table 21: O*NET's Skills Compared to TF-IDF Salient Terms in Abstract Corpora

Skill Component(s)		Definition of Skill Component	Notable Related Salient Terms
Basic Skills	Critical Thinking	"Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems."	*analysis, modeling*
	Mathematics	"Using mathematics to solve problems."	*numerical, mathematical, parameter, equation, analysis*
	Monitoring	"Monitoring/Assessing performance of yourself, other individuals, or organizations to make improvements or take corrective action."	*performance*
	Reading Comprehension	"Understanding written sentences and paragraphs in work related documents."	*paper*
	Science	"Using scientific rules and methods to solve problems."	*method*
	Writing	"Communicating effectively in writing as appropriate for the needs of the audience."	*paper*
Complex Problem Solving	Complex Problem Solving	"Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions."	*performance, analysis, effect*
Resource Management	Management of Material Resources	"Obtaining and seeing to the appropriate use of equipment, facilities, and materials needed to do certain work."	*water, energy, power*
	Management of Personnel Resources	"Motivating, developing, and directing people as they work, identifying the best people for the job."	*agent based*

Skill Component(s)		Definition of Skill Component	Notable Related Salient Terms
	Time Management	"Managing one's own time and the time of others."	*time*
Social Skills	Persuasion	"Persuading others to change their minds or behavior."	*show*
Systems Skills	Judgement and Decision Making	"Considering the relative costs and benefits of potential actions to choose the most appropriate one."	*model, method, based*
	Systems Analysis	"Determining how a system should work and how changes in conditions, operations, and the environment will affect outcomes."	*system*
	Systems Evaluation	"Identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system."	*system, performance*
Technical Skills	Equipment Selection	"Determining the kind of tools and equipment needed to do a job."	*method, optimization, simulation*
	Operation and Control	"Controlling operations of equipment or systems."	*control*
	Operations Analysis	"Analyzing needs and product requirements to create a design."	*design, structure, system*
	Quality Control Analysis	"Conducting tests and inspections of products, services, or processes to evaluate quality or performance."	*performance, effect*
	Technology Design	"Generating or adapting equipment and technology to serve user needs."	*design, simulation, computer simulation.*

(National Center for O*NET Development, 2020c)

Table 22: O*NET's Abilities Compared to TF-IDF Salient Terms in Abstract Corpora

Ability Component(s)		Definition of Ability Component	Notable Related Salient Terms
Cognitive Abilities	Cognitive Flexibility	"The ability to generate or use different sets of rules for combining or grouping things in different ways."	*modeling, optimization, analysis, flow*
	Deductive Reasoning	"The ability to apply general rules to specific problems to produce answers that make sense."	*analysis, based*
	Flexibility of Closure	"The ability to identify or detect a known pattern (a figure, object, word, or sound) that is hidden in other distracting material."	*structure, process,*
	Inductive Reasoning	"The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events)."	*analysis*
	Information Ordering	"The ability to arrange things or actions in a certain order or pattern according to a specific rule or set of rules (e.g., patterns of numbers, letters, words, pictures, mathematical operations)."	*mathematical*
	Mathematical Reasoning	"The ability to choose the right mathematical methods or formulas to solve a problem."	*mathematical, numerical, equation, analysis, parameter, distribution*
	Number Facility	"The ability to add, subtract, multiply, or divide quickly and correctly."	*mathematical, numerical, equation, analysis*

Ability Component(s)		Definition of Ability Component	Notable Related Salient Terms
	Perceptual Speed	"The ability to quickly and accurately compare similarities and differences among sets of letters, numbers, objects, pictures, or patterns. The things to be compared may be presented at the same time or one after the other. This ability also includes comparing a presented object with a remembered object."	*analysis, result, effect, time*
	Speed of Closure	"The ability to quickly make sense of, combine, and organize information into meaningful patterns."	*analysis, time*
	Time Sharing	"The ability to shift back and forth between two or more activities or sources of information (such as speech, sounds, touch, or other sources)."	*time, flow*
	Visualization	"The ability to imagine how something will look after it is moved around or when its parts are moved or rearranged."	*model, modeling, data, optimization*
	Written Comprehension	"The ability to read and understand information and ideas presented in writing."	*paper*
	Written Expression	"The ability to communicate information and ideas in writing so others will understand."	*paper*
Psychomotor Abilities	Control Precision	"The ability to quickly and repeatedly adjust the controls of a machine or a vehicle to exact positions."	*control*

(National Center for O*NET Development, 2020a)

Thus, the KSAs applied most frequently in M&S academic literature are listed in Table 23.

Table 23: O*NET KSAs Derived from M&S Publication Abstracts and Keywords

Knowledge Component	Skills	Abilities
Administration and Management	Critical Thinking	Cognitive Flexibility
Biology	Mathematics	Deductive Reasoning
Chemistry	Monitoring	Flexibility of Closure
Communication and Media	Reading Comprehension	Inductive Reasoning
Design	Science	Information Ordering
Engineering and Technology	Writing	Mathematical Reasoning
Mathematics	Complex Problem Solving	Number Facility
Physics	Mgmt. of Material Resources	Perceptual Speed
Production and Process	Mgmt. of Personnel Resources	Speed of Closure
	Time Management	Time Sharing
	Persuasion	Visualization
	Judgement and Decision Making	Written Comprehension
	Systems Analysis	Written Expression
	Systems Evaluation	Control Precision
	Equipment Selection	
	Operation and Control	
	Operation Analysis	
	Quality Control Analysis	
	Technology Design	

Job Posting TF-IDFs

Research Question 2: *What are the KSAs most requested in M&S job listings within the United States*? I investigated natural language in job posting data to answer my second research question and created a TF-IDF DTM to help enumerate important terms in the overall ontology. Figure 13 shows the top 30 most salient unigrams within the job posting corpus. A bigram model was unable to be computed due to corpus size and computing power. I generated the top 30 most salient unigrams within the job postings. Results were highest for the terms *system* (*M*=.073), *experience* (*M*=.063), and *job* (*M*=.063). The lowest results were generated for the terms *information* (*M*=.024), *customer* (*M*=.024), and *time* (*M*=.022).

While some of the top most salient words are terms expected in all types of job postings, notable terms based on the thematic analysis were identified. When mapped these terms fall into O*NET knowledge component categories, terms fall within the Communication and Media (e.g., *paper*), Design (e.g., *design, structure*), Engineering and Technology (e.g., *simulation*, *data, process*), Mathematics (e.g., *analysis, distribution, equation*), Physics (e.g., *energy, power*) and Production and Processing (e.g., *process*) categories. These were further mapped to the related O*NET skills of active learning, critical thinking, learning strategies, mathematics, science, writing, complex problem solving, management of material resources, time management, coordination, instruction, service oriented, judgement and decision making, systems analysis, systems evaluation, operation and control, operation analysis, programming, quality control analysis, and technology design. O*NET abilities of cognitive flexibility, deductive reasoning, flexibility closure, fluency of ideas, inductive reasoning, information ordering, mathematical reasoning, number facility, originality, perceptual speed, speed of closure, time sharing, written comprehension, and written expression. were also identified. Further thematic analysis and O*NET coding are discussed in the following sections.

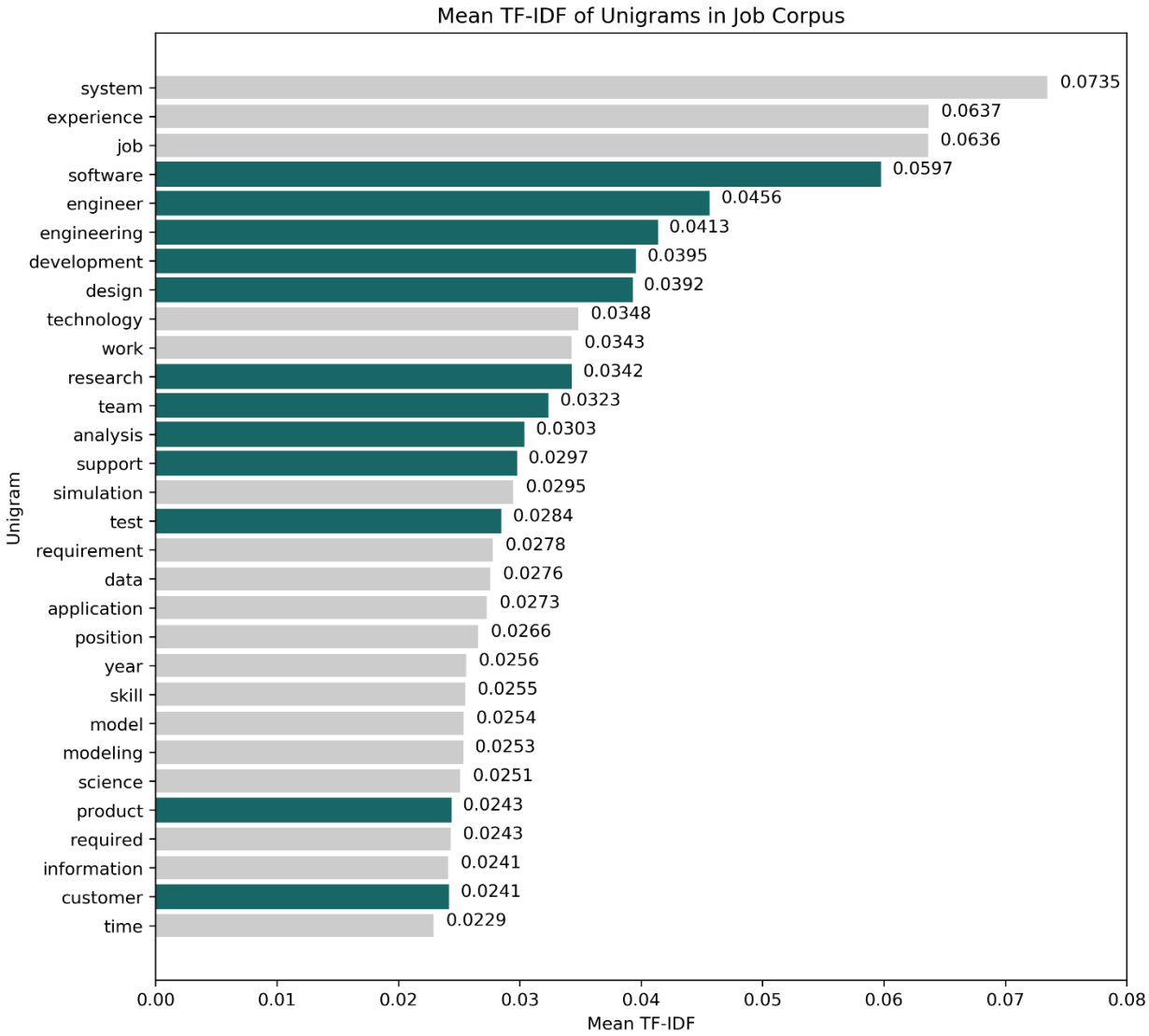

Figure 13: Top 30 Most Salient Unigrams within Publication Job Corpus

Job Posting LDM Model

Research Question 3: *How should M&S job types be categorized?* To answer this question, I used a unigram BoW model including only nouns, adjectives, and verbs, which will be used by the LDA model. LDA is limited in that it assumes a set number of topics, typically unknown to the researcher (Blei et al., 2002). Several LDA modes using various number of topics should be run to determine the model with the best fit to the data. To help inform the number of topics I can look at the coherence score. The largest coherence score is the best fit. From there I can re-run the LDA model with the optimal number of topics to determine appropriate terms for categorization. Table 24 shows the coherence scores found when running various iterations with different topic numbers. Figure 14 plots these scores and easily visualizes the highest coherence score. The highest coherence score was at nine topics with a score of 0.4197.

Table 24: Coherence Scores by Number of Topics

Number of Topics	Coherence Score
2	0.3623
3	0.3246
4	0.3567
5	0.3646
6	0.3653
7	0.3523
8	0.4051
9	0.4197
10	0.3952
11	0.3983
12	0.3832
13	0.4052
14	0.4173
15	0.4048
16	0.3590
17	0.3490
18	0.3908
19	0.3593
20	0.3410
21	0.4023
22	0.4151
23	0.3794
24	0.3962

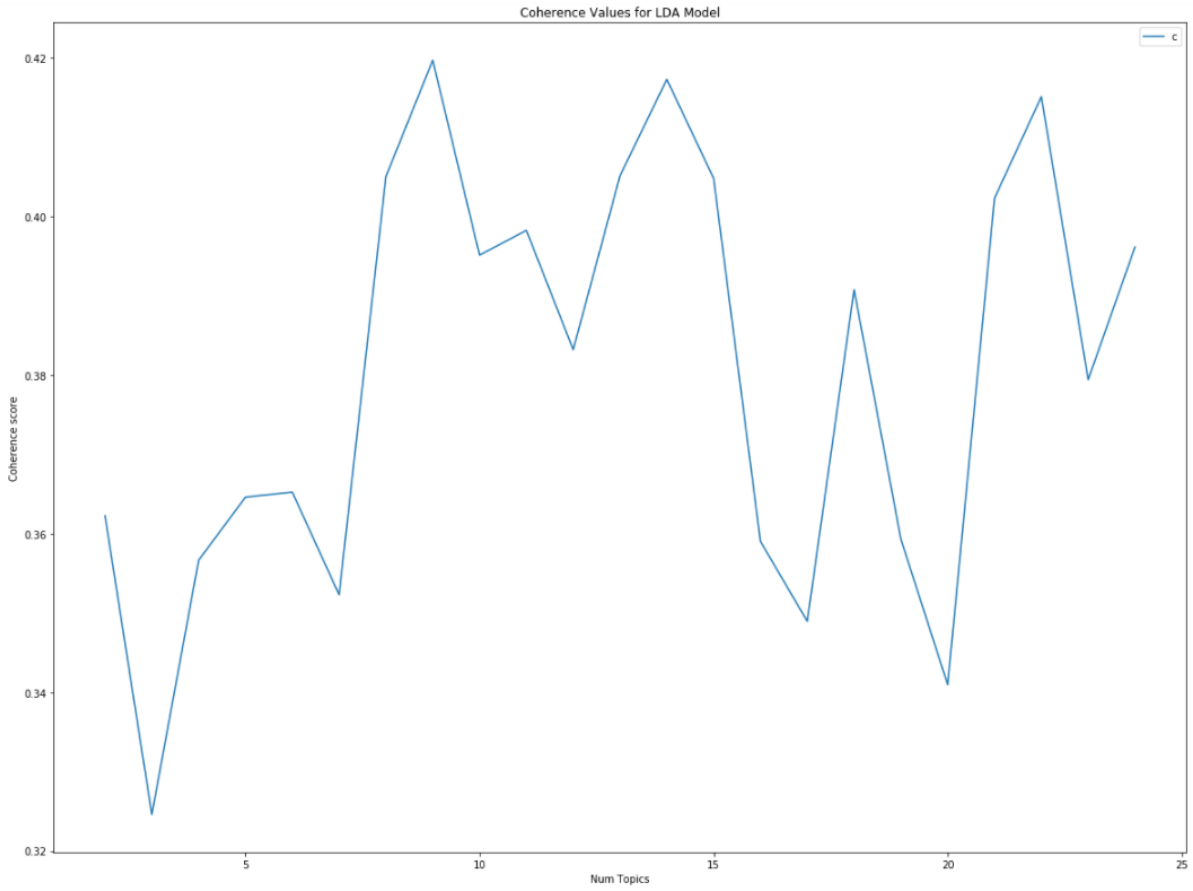

Figure 14: Line Plot of Coherence Scores by Number of Topics

Figure 14 shows the intertopic distance map and the top 30 most salient terms using nine topics. Note the metric used in the LDA figures below for most salient terms in this case are the number of times, rather than the relative frequency of the words and documents.

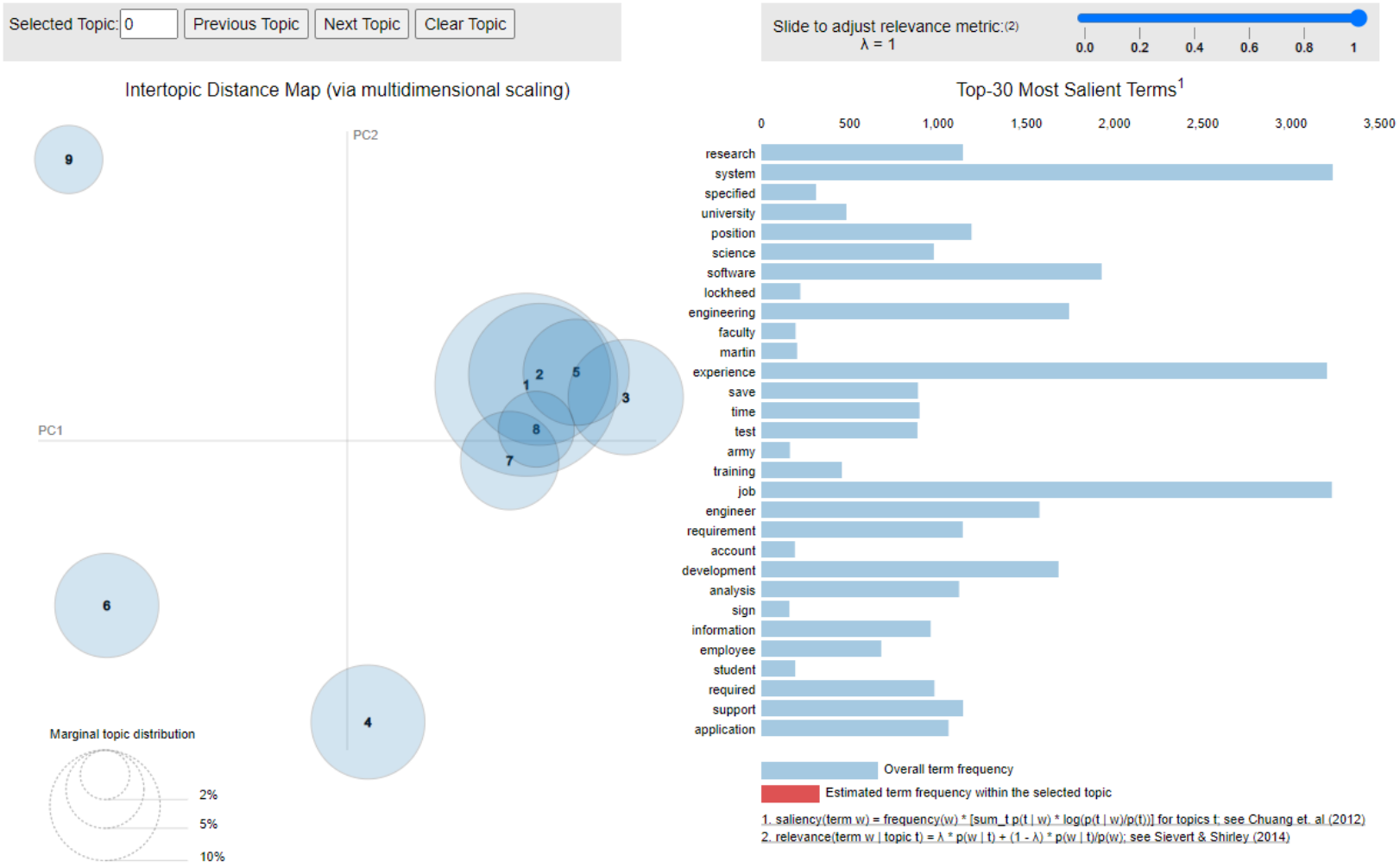

Figure 15: Topic Modeling Visualizations for Job Posting Corpus -Overall

Research Question 4*: What are the KSAs most identified per job type?* It was difficult to distinguish the differences between the groups due to how much they were overlapped, making inference difficult. However, three topics stood out from the rest in terms of intertopic distance: Topic four, topic six, and topic nine, which are shown in Figures 16, 17, and 18 respectively. I noticed that topics six and nine had some notable terms. Topic four included many employment related words like *position*, *requirement*, and *required*. Topic six include notable words like *university*, *faculty*, and *student*, unique to that topic. Topic nine included notable terms such as *army* and *arl*. I suspect that topics six and nine are related to a common characterization grouping for professionals (Academic, Government, and Industry). As such, I will use these categories as titles for topic four (Industry), topic six (Academic), and topic nine (Government).

Topic four (Industry), knowledge components included computers and electronics, education and training, engineering and technology, production and process, and public safety and security. Topic four skills included learning strategies, writing, complex problem solving, management of personnel resources, time management, coordination, instruction, service oriented, judgement and decision making, and technology design. Lastly, the abilities identified in topic four include cognitive flexibility, flexibility of closure, perceptual speed, speed of closure, time sharing, written comprehension, and written expression.

For topic six (Academic), knowledge components included administration and management, computers and electronics, education and training, engineering and technology, and geography. The skills identified in topic six include active learning, critical thinking, learning strategies, science, complex problem solving, management of personnel resources, time management, instruction, and judgement and decision making. Topic six abilities identified

included cognitive flexibility, deductive reasoning, flexibility of closure, inductive reasoning, mathematical reasoning, number facility, perceptual speed, speed of closure, time sharing, written comprehension, and written expression.

Further, topic nine (Government), knowledge components identified included administration and management, education and training, engineering and technology, geography, mathematics, and physics. Skills for topic nine included active learning mathematics, science, writing, complex problem solving, management of material resources, management of personnel resources, time management, coordination, service oriented, judgement and decision making, systems analysis, system evaluation, operation and control, operations analysis, programming, quality control analysis, and technology design. Finally, the abilities identified in topic nine include cognitive flexibility, flexibility of closure, inductive reasoning, mathematical reasoning, perceptual speed, speed of closure, and time sharing. Using the Job posting TF-IDF and LDA analysis output, I map the terms to the knowledge components, skills, and abilities and identified by O*NET in Tables 25, 26, and 27 respectively.

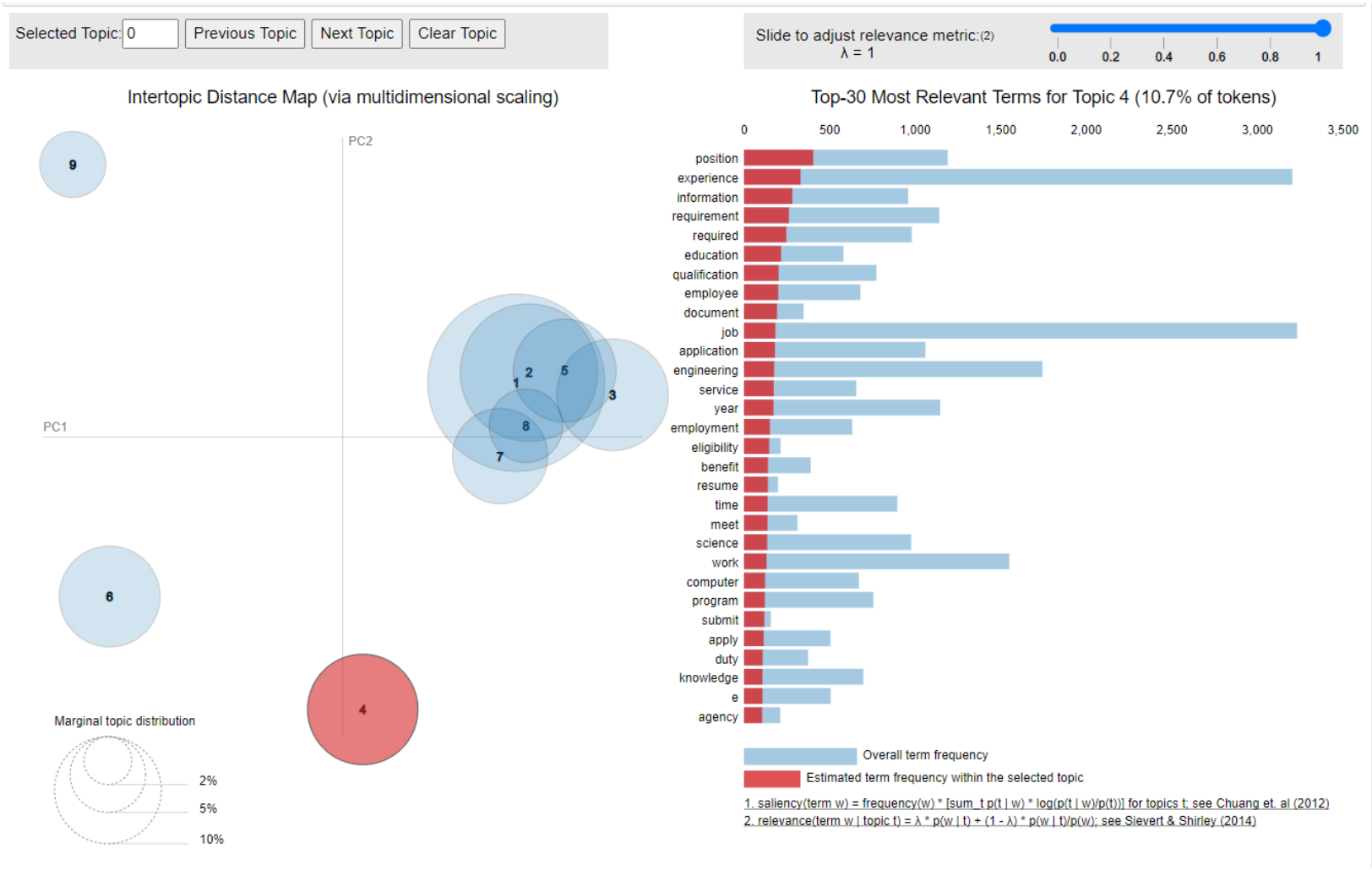

Figure 16:Topic Modeling Visualizations for Job Posting Corpus - Topic #4

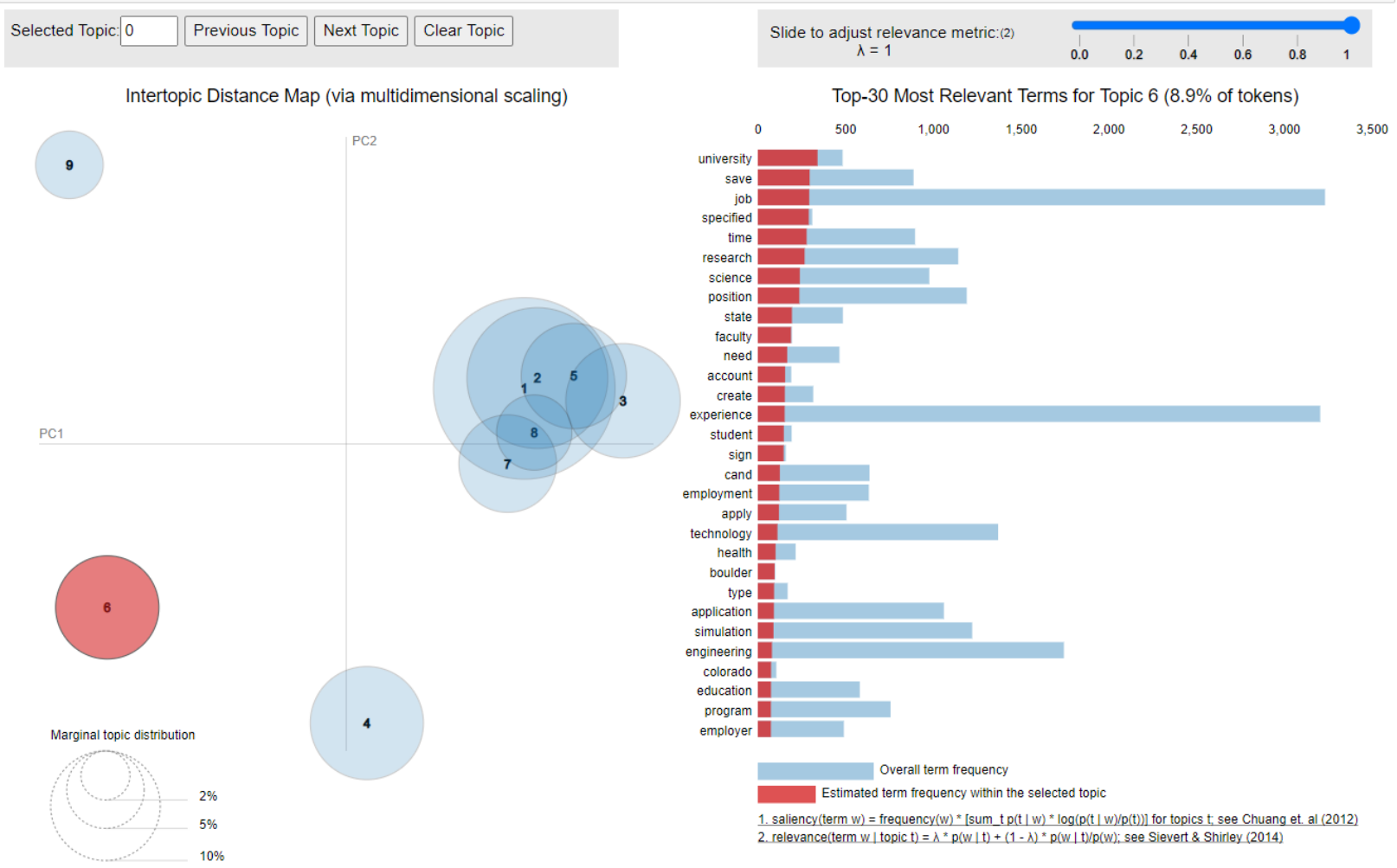

Figure 17: Topic Modeling Visualizations for Job Posting Corpus - Topic #6

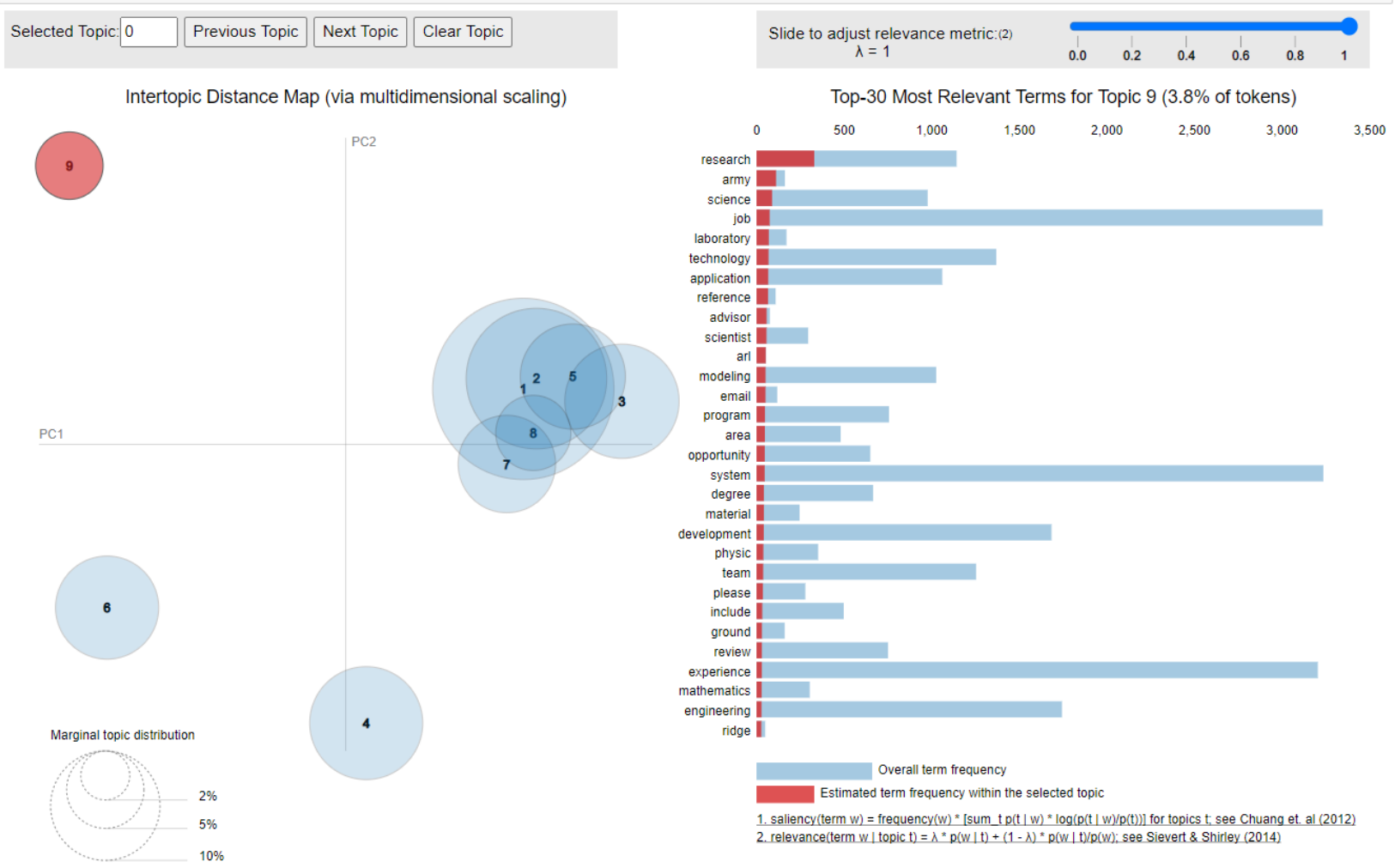

Figure 18: Topic Modeling Visualizations for Job Posting Corpus - Topic #9

Table 25: O*NET's Skills Compared to TF-IDF Salient Terms in Job Corpus

Knowledge Component(s)	Definition of Knowledge Component	Notable Related Salient Terms
Administration and Management	"Knowledge of business and management principles involved in strategic planning, resource allocation, human resources modeling, leadership technique, production methods, and coordination of people and resources."	*research, design, development, position, skills, product* Topic 6: *research, position* Topic 9: *research, advisor*
Computers and Electronics	"Knowledge of circuit boards, processors, chips, electronic equipment, and computer hardware and software, including applications and programming."	*simulation, software, application* Topic 4: *computer, program* Topic 6: *technology, simulation*
Design	"Knowledge of design techniques, tools, and principles involved in production of precision technical plans, blueprints, drawings, and models."	*design, product, model, modeling*
Education and Training	"Knowledge of principles and methods for curriculum and training design, teaching and instruction for individuals and groups, and the measurement of training effects."	*training* Topic 4: *education* Topic 6: *university, faculty, student, education* Topic 9: *advisor*
Engineering and Technology	"Knowledge of the practical application of engineering science and technology. This includes applying principles, techniques, procedures, and equipment to the design and production of various goods and services."	*engineer, engineering, technology, simulation, computer* Topic 4: *engineering* Topic 6: *engineering, technology* Topic 9: *engineering*
Geography	"Knowledge of principles and methods for describing the features of land, sea, and air masses, including their physical characteristics, locations, interrelationships, and distribution of plant, animal, and human life."	*n/a* Topic 6: *boulder, colorado* Topic 9: *ground, area, ridge*

Knowledge Component(s)	Definition of Knowledge Component	Notable Related Salient Terms
Mathematics	"Knowledge of arithmetic, algebra, geometry, calculus, statistics, and their applications."	*analysis* Topic 9: *mathematics*
Physics	"Knowledge and prediction of physical principles, laws, their interrelationships, and applications to understanding fluid, material, and atmospheric dynamics, and mechanical, electrical, atomic and sub- atomic structures and processes."	*time* Topic 9: *material, physic*
Production and Processing	"Knowledge of raw materials, production processes, quality control, costs, and other techniques for maximizing the effective manufacture and distribution of goods."	*product, requirement, required* Topic 4: *requirement, required*
Public Safety and Security	"Knowledge of relevant equipment, policies, procedures, and strategies to promote effective local, state, or national security operations for the protection of people, data, property, and institutions"	*data* Topic 4: *army*
Sale and Marketing	"Knowledge of principles and methods for showing, promoting, and selling products or services. This includes marketing strategy and tactics, product demonstration, sales techniques, and sales control systems."	*customer, product*

(National Center for O*NET Development, 2020b)

Table 26: O*NET's Knowledge Components Compared to TF-IDF Salient Terms in Job Corpus

Skill Component(s)		Definition of Skill Component	Notable Related Salient Terms
Basic Skills	Active Learning	"Understanding the implications of new information for both current and future problem-solving and decision-making."	Topic 6: *research* Topic 9: *research*
	Critical Thinking	"Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems."	*analysis* Topic 6: *research*
	Learning Strategies	"Selecting and using training/instructional methods and procedures appropriate for the situation when learning or teaching new things."	*training* Topic 4: *education, knowledge* Topic 6: *university, faculty, student, education*
	Mathematics	"Using mathematics to solve problems."	Topic 9: *mathematics*
	Science	"Using scientific rules and methods to solve problems."	*science* Topic 4: *science* Topic 6: *science* Topic 9: *science*
	Writing	"Communicating effectively in writing as appropriate for the needs of the audience."	Topic 4: *resume, document* Topic 9: *email*
Complex Problem Solving	Complex Problem Solving	"Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions."	*test, experience, development, analysis* Topic 4: *experience* Topic 6: *experience, research* Topic 9: *experience development*
Resource Management	Management of Material Resources	"Obtaining and seeing to the appropriate use of equipment, facilities, and materials needed to do certain work."	Topic 9: *material*

Skill Component(s)		Definition of Skill Component	Notable Related Salient Terms
	Management of Personnel Resources	"Motivating, developing, and directing people as they work, identifying the best people for the job."	*experience, job, skill, support* Topic 4: *resume, employment, benefit, position, experience, qualifications, requirements* Topic 6: *experience, employment, apply, application* Topic 9: *job opportunity, experience, degree, advisor*
	Time Management	"Managing one's own time and the time of others."	*time* Topic 4: *time* Topic 6: *time* Topic 9*: time*
Social Skills	Coordination	"Adjusting actions in relation to others' actions."	*support* Topic 4: *meet* Topic 9: *team*
	Instruction	"Teaching others how to do something."	*training student* Topic 4: *education* Topic 6: *university, faculty, education, student*
	Service Oriented	"Actively looking for ways to help people."	*support* Topic 4: *service* Topic 9: *advisor team*
Systems Skills	Judgement and Decision Making	"Considering the relative costs and benefits of potential actions to choose the most appropriate one."	*experience* Topic 4: *experience* Topic 6: *experience* Topic 9: *experience*
	Systems Analysis	"Determining how a system should work and how changes in conditions, operations, and the environment will affect outcomes."	*system, analysis* Topic 9: *system*
	Systems Evaluation	"Identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system."	*system, test* Topic 9: *system*

Skill Component(s)	Definition of Skill Component	Notable Related Salient Terms
Operation and Control	"Controlling operations of equipment or systems."	*system,* Topic 9: *system*
Operations Analysis	"Analyzing needs and product requirements to create a design."	*analysis, design, system,* Topic 9: *system*
Programming	"Writing computer programs for various purposes."	Topic 9: *program*
Quality Control Analysis	"Conducting tests and inspections of products, services, or processes to evaluate quality or performance."	*test,* Topic 9: *review*
Technology Design	"Generating or adapting equipment and technology to serve user needs."	*design, technology, requirement* Topic 4: *engineering, computer, program* Topic 9: *technology*

(National Center for O*NET Development, 2020c)

Table 27: O*NET's Abilities Compared to TF-IDF Salient Terms in Abstract Corpora

Ability Component(s)		Definition of Ability Component	Notable Related Salient Terms
Cognitive Abilities	Cognitive Flexibility	"The ability to generate or use different sets of rules for combining or grouping things in different ways."	*research, modeling, analysis* Topic 4: *apply* Topic 6: *model, research, review* Topic 9: *research, apply*
	Deductive Reasoning	"The ability to apply general rules to specific problems to produce answers that make sense."	Topic 6: *mathematics, laboratory, review*
	Flexibility of Closure	"The ability to identify or detect a known pattern (a figure, object, word, or sound) that is hidden in other distracting material."	*research, modeling, analysis* Topic 4: *apply* Topic 6: *model, research, review* Topic 9: *research, apply*
	Fluency of Ideas	"The ability to come up with a number of ideas about a topic (the number of ideas is important, not their quality, correctness, or creativity)."	*design*
	Inductive Reasoning	"The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events)."	Topic 6: *laboratory, review, research* Topic 9*: research*
	Information Ordering	"The ability to arrange things or actions in a certain order or pattern according to a specific rule or set of rules (e.g., patterns of numbers, letters, words, pictures, mathematical operations)."	*data*
	Mathematical Reasoning	"The ability to choose the right mathematical methods or formulas to solve a problem."	*analysis, data* Topic 6: *mathematics, engineering* Topic 9: *engineering*

Ability Component(s)	Definition of Ability Component	Notable Related Salient Terms
Number Facility	"The ability to add, subtract, multiply, or divide quickly and correctly."	*analysis* Topic 6: *mathematics*
Originality	"The ability to come up with unusual or clever ideas about a given topic or situation, or to develop creative ways to solve a problem."	*design, model, develop, modeling*
Perceptual Speed	"The ability to quickly and accurately compare similarities and differences among sets of letters, numbers, objects, pictures, or patterns. The things to be compared may be presented at the same time or one after the other. This ability also includes comparing a presented object with a remembered object."	*time* Topic 4: *time* Topic 6: *time, review, research* Topic 9: *time, research*
Speed of Closure	"The ability to quickly make sense of, combine, and organize information into meaningful patterns."	*time* Topic 4: *time* Topic 6: *time, review* Topic 9: *time*
Time Sharing	"The ability to shift back and forth between two or more activities or sources of information (such as speech, sounds, touch, or other sources)."	*time* Topic 4: *time* Topic 6: *time* Topic 9: *time*
Written Comprehension	"The ability to read and understand information and ideas presented in writing."	Topic 4: *document, resume* Topic 6: *email*
Written Expression	"The ability to communicate information and ideas in writing so others will understand."	Topic 4: *document, resume* Topic 6: *email*

(National Center for O*NET Development, 2020a)

Thus, the KSAs applied most frequently in M&S job postings overall and per topic are listed in Table 28, 29, 30 and 31.

Table 28: O*NET KSAs Derived from M&S Job Postings Overall

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Computers and Electronics	Critical Thinking	Deductive Reasoning
Design	Learning Strategies	Flexibility of Closure
Education and Training	Mathematics	Fluency of Ideas
Engineering and Technology	Science	Inductive Reasoning
Geography	Writing	Information Ordering
Mathematics	Complex Problem Solving	Mathematical Reasoning
Physics	Mgmt. of Material Resources	Number Facility
Production and Process	Mgmt. of Personnel Resources	Originality
Public Safety and Security	Time Management	Perceptual Speed
Sale and Marketing	Coordination	Speed of Closure
	Instruction	Time Sharing
	Service Oriented	Written Comprehension
	Judgement and Decision Making	Written Expression
	Systems Analysis	
	Systems Evaluation	
	Operation and Control	
	Operation Analysis	
	Programming	
	Quality Control Analysis	
	Technology Design	

Table 29: O*NET KSAs Derived from M&S Job Postings Topic 4

Knowledge Component	Skills	Abilities
Computers and Electronics	Learning Strategies	Cognitive Flexibility
Education and Training	Science	Flexibility of Closure
Engineering and Technology	Writing	Perceptual Speed
Production and Process	Complex Problem Solving	Speed of Closure
Public Safety and Security	Mgmt. of Personnel Resources	Time Sharing
	Time Management	Written Comprehension
	Coordination	Written Expression
	Instruction	
	Service Oriented	
	Judgement and Decision Making	
	Technology Design	

Table 30:O*NET KSAs Derived from M&S Job Postings Topic 6

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Computers and Electronics	Critical Thinking	Deductive Reasoning
Education and Training	Learning Strategies	Flexibility of Closure
Engineering and Technology	Science	Inductive Reasoning
Geography	Complex Problem Solving	Mathematical Reasoning
	Mgmt. of Personnel Resources	Number Facility
	Time Management	Perceptual Speed
	Instruction	Speed of Closure
	Judgement and Decision Making	Time Sharing
		Written Comprehension
		Written Expression

Table 31: O*NET KSAs Derived from M&S Job Postings Topic 9

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Education and Training	Mathematics	Flexibility of Closure
Engineering and Technology	Science	Inductive Reasoning
Geography	Writing	Mathematical Reasoning
Mathematics	Complex Problem Solving	Perceptual Speed
Physics	Mgmt. of Material Resources	Speed of Closure
	Mgmt. of Personnel Resources	Time Sharing
	Time Management	
	Coordination	
	Service Oriented	
	Judgement and Decision Making	
	Systems Analysis	
	Systems Evaluation	
	Operation and Control	
	Operation Analysis	
	Programming	
	Quality Control Analysis	
	Technology Design	

Course Description TF-IDFs

Research Question 5: *What are the KSAs taught most frequently in M&S graduate level course descriptions for Universities within the United States?* Further, I used another TF-IDF DTM to investigate course description text data, investigating my fifth research question. Unigram results were highest for the terms *system* (*M*=.05), *simulation* (*M*=.05), *course* (*M*=.04),

topic (*M*=.03), *m&s* (*M*=.03), and *technique* (*M*=.03). The lowest results were generated for the terms *introduction* (*M*=.02) *instructor* (*M*=.02), *visualization* (*M*=.02), and *optimization* (*M*=.02). Figure 19 shows the top 30 most salient unigrams within the job posting corpus.

While some of the top most salient words are terms expected in all types of course descriptions, notable terms based on the thematic analysis were identified. These terms are highlighted below and include *m&s* (*M*=.037), *analysis* (*M*=.029), *network* (*M*=.028), *computer* (*M*=.028), *program* (*M*=.027), *security* (*M*=.027), *research* (*M*=.027), *design* (*M*=.022), *visualization* (*M*=.020), *and optimization* (*M*=.020). I would like to note here that this is the only salient term list that includes M&S as a domain rather than variations of tools. O*NET skills of critical thinking, mathematics, monitoring, reading comprehension, science, writing, complex problem solving, management of material resources, management of personnel resources, time management, persuasion, judgement and decision making, systems analysis, systems evaluation, equipment selection, operation and control, operation analysis, quality control analysis and technology design were also identified Further thematic analysis and O*NET coding are discussed later in the chapter.

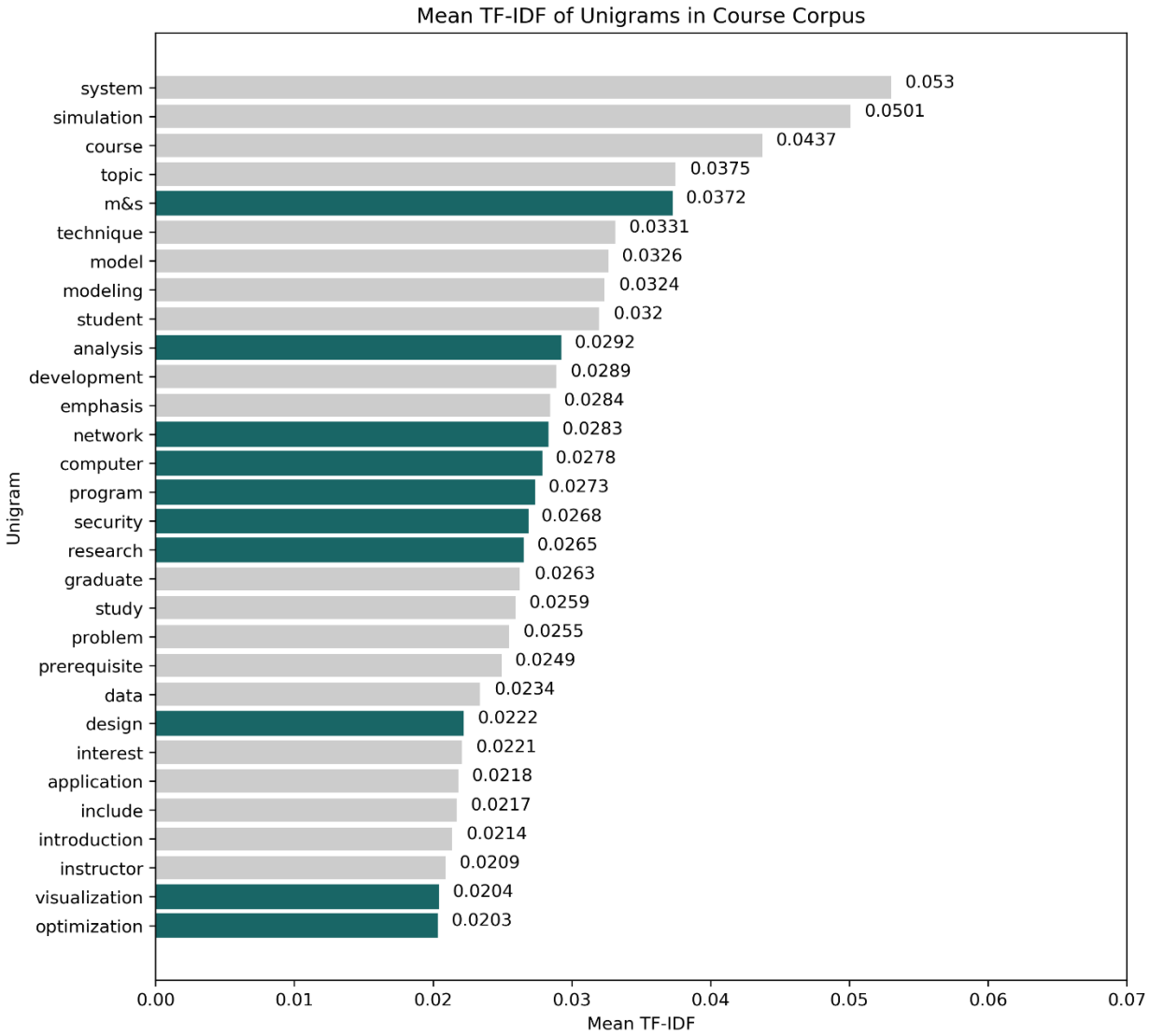

Figure 19: Top 30 Most Salient Unigrams within Course Description Corpus

Figure 19 shows the top 30 most salient bigrams within the course description corpus. Results were highest for the terms *graduate program* (*M*=.004), *program director* (*M*=.004), and *special topic* (*M*=.004), all of which are expected in course descriptions. The lowest results were generated for the terms *listed cee* (*M*=.002), *probability statistics* (*M*=.002), and *distributed simulation* (*M*=.002).

While some of the top most salient words are terms expected in course descriptions, notable terms based on the thematic analysis were identified. These terms are highlighted below and include *development m&s* (*M*=.002), *case study* (*M*=.002), *design analysis* (*M*=.002), *medical image* (*M*=.002), *probability statistic* (*M*=.002), and *distributed simulation* (*M*=.002). O*NET skills of critical thinking, mathematics, monitoring, reading comprehension, science, writing, complex problem solving, management of material resources, management of personnel resources, time management, persuasion, judgement and decision making, systems analysis, systems evaluation, equipment selection, operation and control, operation analysis, quality control analysis and technology design were identified.

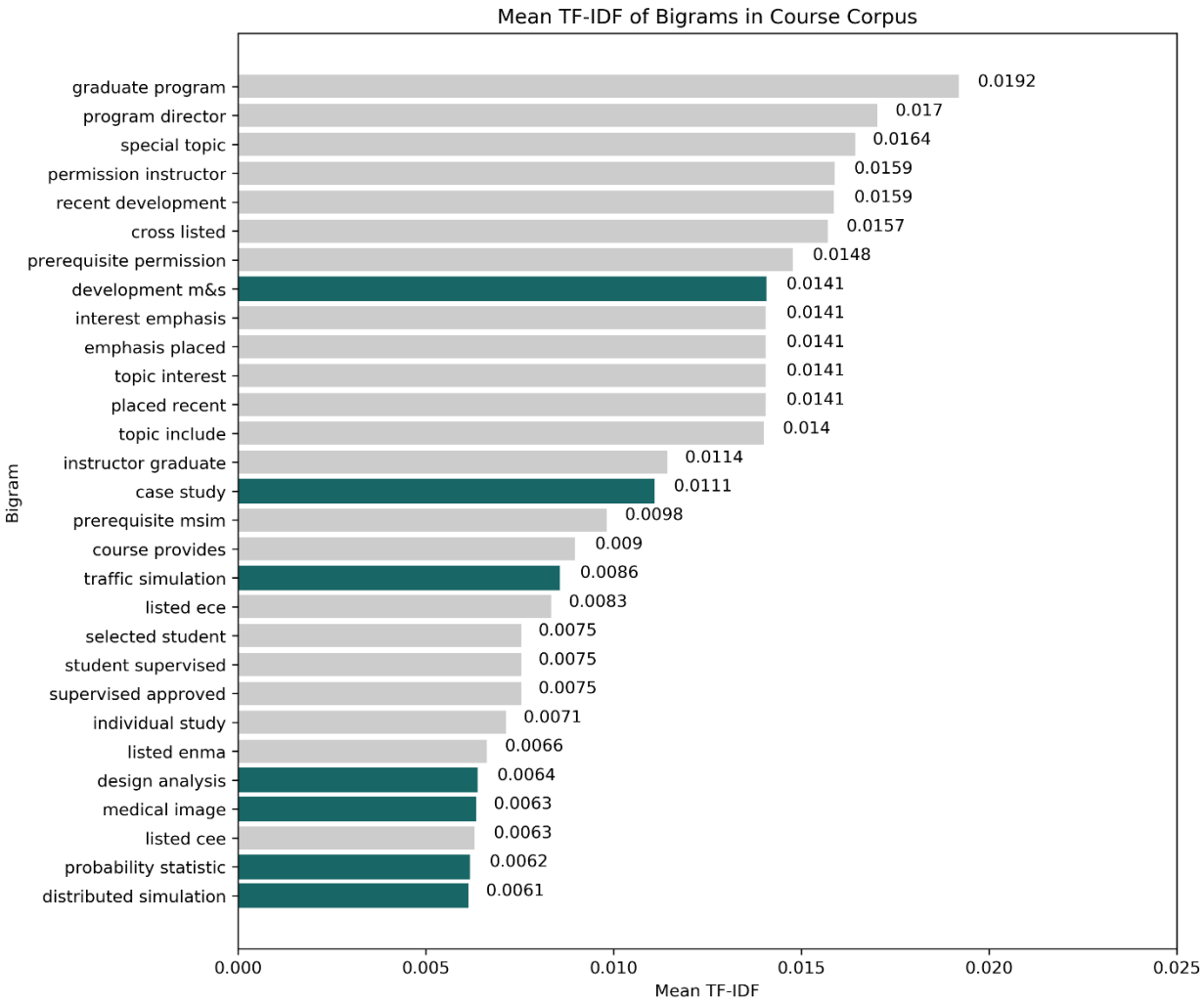

Figure 20: Top 30 Most Salient Bigrams within Course Description Corpus

Table 32: O*NET's Knowledge Components Compared to TF-IDF Salient Terms in Course Description Corpus

Knowledge Component(s)	Definition of Knowledge Component	Notable Related Salient Terms
Administration and Management	"Knowledge of business and management principles involved in strategic planning, resource allocation, human resources modeling, leadership technique, production methods, and coordination of people and resources."	*instructor, student, program director, student supervised*
Computers and Electronics	"Knowledge of circuit boards, processors, chips, electronic equipment, and computer hardware and software, including applications and programming."	*computer, program, optimization, simulation, distributed simulation*
Design	"Knowledge of design techniques, tools, and principles involved in production of precision technical plans, blueprints, drawings, and models."	*design, design analysis*
Education and Training	"Knowledge of principles and methods for curriculum and training design, teaching and instruction for individuals and groups, and the measurement of training effects."	*technique, student, graduate, study, problem, course, prerequisite, instructor, program director, graduate program, special topic, permission instructor, prerequisite permission, topic interest, emphasis, emphasis placed, topic include, instructor graduate, case study, prerequisite msim, course provides, selected student, student supervised, individual study*
Engineering and Technology	"Knowledge of the practical application of engineering science and technology. This includes applying principles, techniques, procedures, and equipment to the design and production of various goods and services."	*simulation, traffic simulation*
Fine Arts	"Knowledge of the theory and techniques required to compose, produce, and perform works of music, dance, visual arts, drama, and sculpture."	*medical image*

Knowledge Component(s)	Definition of Knowledge Component	Notable Related Salient Terms
Mathematics	"Knowledge of arithmetic, algebra, geometry, calculus, statistics, and their applications."	*model, model, analysis, problem, data, optimization, visualization, probability statistic*
Medicine and Dentistry	"Knowledge of the information and techniques needed to diagnose and treat human injuries, diseases, and deformities. This includes symptoms, treatment alternatives, drug properties and interactions, and preventive health-care measures."	*medical image*
Production and Processing	"Knowledge of raw materials, production processes, quality control, costs, and other techniques for maximizing the effective manufacture and distribution of goods."	*development, design, research, analysis, optimizing, prerequisite, development m&s, recent development*
Psychology	"Knowledge of human behavior and performance; individual differences in ability, personality, and interests; learning and motivation; psychological research methods; and the assessment and treatment of behavioral and affective disorders."	*interest, emphasis, interest emphasis*
Public Safety and Security	"Knowledge of relevant equipment, policies, procedures, and strategies to promote effective local, state, or national security operations for the protection of people, data, property, and institutions"	*computer, security, network, data*
Sociology and Anthropology	"Knowledge of group behavior and dynamics, societal trends and influences, human migrations, ethnicity, cultures and their history and origins."	*traffic simulation, case study*
Transportation	"Knowledge of principles and methods for moving people or goods by air, rail, sea, or road, including the relative costs and benefits."	*traffic simulation, case study*

(National Center for O*NET Development, 2020b)

Table 33: O*NET's Skills Compared to TF-IDF Salient Terms in Course Description Corpus

Skill Component(s)		Definition of Skill Component	Notable Related Salient Terms
Basic Skills	Active Learning	"Understanding the implications of new information for both current and future problem-solving and decision-making."	*problem, analysis, research*
	Critical Thinking	"Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems."	*analysis, research, study, optimization, case study,*
	Learning Strategies	"Selecting and using training/instructional methods and procedures appropriate for the situation when learning or teaching new things."	*study, course, topic, student, graduate, problem, prerequisite, graduate program, special topics, topic interest, instructor graduate, case study*
	Mathematics	"Using mathematics to solve problems."	*probability, statistics*
	Science	"Using scientific rules and methods to solve problems."	*problem, research*
Complex Problem Solving	Complex Problem Solving	"Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions."	*technique, design, analysis, optimization, visualization, develop m&s*
Resource Management	Management of Material Resources	"Obtaining and seeing to the appropriate use of equipment, facilities, and materials needed to do certain work."	*traffic simulation*
	Management of Personnel Resources	"Motivating, developing, and directing people as they work, identifying the best people for the job."	*program director*
Social Skills	Coordination	"Adjusting actions in relation to others' actions."	*program director, student supervised*

Skill Component(s)		Definition of Skill Component	Notable Related Salient Terms
	Instruction	"Teaching others how to do something."	*study, course, topic, student, graduate, problem, prerequisite, graduate program, special topics, topic interest, instructor graduate, case study*
	Service Oriented	"Actively looking for ways to help people."	*student supervised, supervised approved*
Systems Skills	Judgement and Decision Making	"Considering the relative costs and benefits of potential actions to choose the most appropriate one."	*model, modeling, analysis, technique, problem*
	Systems Analysis	"Determining how a system should work and how changes in conditions, operations, and the environment will affect outcomes."	*system, analysis*
	Systems Evaluation	"Identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system."	*system*
	Operation and Control	"Controlling operations of equipment or systems."	*system*
	Operations Analysis	"Analyzing needs and product requirements to create a design."	*analysis*
	Programming	"Writing computer programs for various purposes."	*program*
	Quality Control Analysis	"Conducting tests and inspections of products, services, or processes to evaluate quality or performance."	*analysis, research system*
	Technology Design	"Generating or adapting equipment and technology to serve user needs."	*design, simulation, computer*

(National Center for O*NET Development, 2020c)

Table 34: O*NET's Abilities Compared to TF-IDF Salient Terms in Course Description Corpus

Ability Component(s)		Definition of Ability Component	Notable Related Salient Terms
Cognitive Abilities	Cognitive Flexibility	"The ability to generate or use different sets of rules for combining or grouping things in different ways."	*include, emphasis, development, system, topic, network, analysis, research, optimization, model, modeling*
	Deductive Reasoning	"The ability to apply general rules to specific problems to produce answers that make sense."	*analysis, research, optimization, problem, technique*
	Flexibility of Closure	"The ability to identify or detect a known pattern (a figure, object, word, or sound) that is hidden in other distracting material."	*research, model, modeling*
	Inductive Reasoning	"The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events)."	*analysis*
	Information Ordering	"The ability to arrange things or actions in a certain order or pattern according to a specific rule or set of rules (e.g., patterns of numbers, letters, words, pictures, mathematical operations)."	*system, emphasis, interest, optimization*
	Mathematical Reasoning	"The ability to choose the right mathematical methods or formulas to solve a problem."	*modeling, model, problem, analysis,*
	Number Facility	"The ability to add, subtract, multiply, or divide quickly and correctly."	*analysis*

Ability Component(s)	Definition of Ability Component	Notable Related Salient Terms
Perceptual Speed	"The ability to quickly and accurately compare similarities and differences among sets of letters, numbers, objects, pictures, or patterns. The things to be compared may be presented at the same time or one after the other. This ability also includes comparing a presented object with a remembered object."	*analysis, research, optimization, visualization, emphasis*
Problem Sensitivity	"The ability to tell when something is wrong or is likely to go wrong. It does not involve solving the problem, only recognizing there is a problem."	*problem*
Selective Attention	"The ability to concentrate on a task over a period of time without being distracted."	*study*
Speed of Closure	"The ability to quickly make sense of, combine, and organize information into meaningful patterns."	*visualization, analysis, design, include*
Visualization	"The ability to imagine how something will look after it is moved around or when its parts are moved or rearranged."	*visualization, optimization, model, modeling*

(National Center for O*NET Development, 2020a)

Table 35: O*NET KSAs Derived from M&S Course Descriptions

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Computers and Electronics	Critical Thinking	Deductive Reasoning
Design	Learning Strategies	Flexibility of Closure
Education and Training	Mathematics	Inductive Reasoning
Fine Arts	Science	Information Ordering
Engineering and Technology	Complex Problem Solving	Mathematical Reasoning
Mathematics	Mgmt. of Material Resources	Number Facility
Medicine and Dentistry	Mgmt. of Personnel Resources	Perceptual Speed
Production and Processing	Coordination	Problem Sensitivity
Psychology	Instruction	Selective Attention
Public Safety and Security	Service Oriented	Speed of Closure
Sociology and Anthropology	Judgement and Decision Making	Visualization
Transportation	Systems Analysis	
	Systems Evaluation	
	Operation and Control	
	Operation Analysis	
	Programming	
	Quality Control Analysis	
	Technology Design	

Comparison of TF-IDF Models

Hypothesis 1: *Scraping multiple types of M&S documents (job postings, course descriptions, and academic publications) will produce a difference in the KSAs (topics) most frequently mentioned in each source type.* To address my hypothesis and visually compare the results, I looked at each of the top 30 most salient unigrams for each source type and identified the terms unique to each of those lists. Unigram analyses for publication abstracts included terms such as result, method, and based; for job postings included terms such as experience, job, and engineer, and for course descriptions included terms such as course, topic, and M&S. Table 36 shows the terms identified by source type.

Table 36: Unique Unigrams in Top 30 Most Salient Lists by Source Type

Publication Abstracts	Job Postings	Course Descriptions
Result	Experience	Course
Method	Job	Topic
Based	Engineer	M&S
Using	Engineering	Techniques
Paper	Technology	Student
Used	Work	Emphasis
Process	Team	Computer
Parameter	Support	Program
Water	Test	Security
Temperature	Requirement	Graduate
Distribution	Position	Problem
Equations	Year	Prerequisite
Flow	Skill	Interest
Effect	Science	Include
Field	Product	Introduction
Power	Required	Instructor
Energy	Information	Visualization
Structure	Customer	Optimization
Show		
Performance		

Further analyses revealed details about overlap. For example, terms that were common across the two source types of job postings and course descriptions included *development*, *research*, and *application*. Terms common between publication abstracts and course descriptions included *study* and *network*. And, the term *time* was found to be in in common between job postings and publication abstracts. These results are listed in more detail below in Table 36.

Table 37: Unique Unigrams to Two of Three Top 30 Most Salient Lists by Source Type

Term	Source Corpora
Development	Job Postings and Course Descriptions
Research	Job Postings and Course Descriptions
Application	Job Postings and Course Descriptions
Study	Publication Abstracts and Course Descriptions
Network	Publication Abstracts and Course Descriptions
Time	Job Postings and Publication Abstracts

There were also terms that were similar across all three source types which included *system*, *simulation*, *model*, *modeling*, *analysis*, *data*, and *design*. Figure 21 shows a Venn-diagram of the salient terms per corpus.

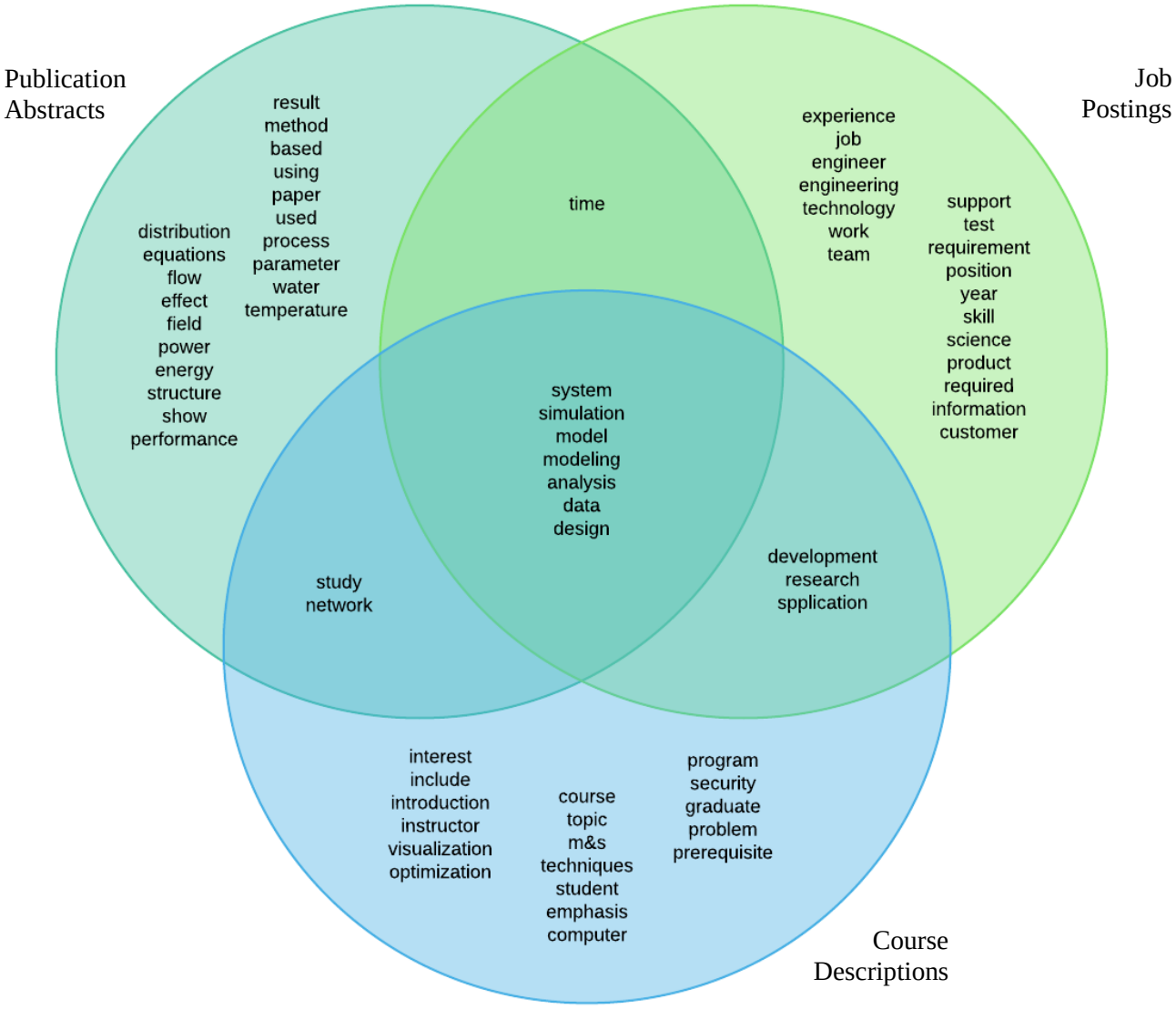

Figure 21: Comparison of Unigrams Between Three Corpora

Unexpected/Interesting Terms Per Source Type

Publication Abstracts and Keywords

Within the publication unigrams a theme of natural resources and physics emerged. These terms include *time* (*M*=.014), *water* (*M*=.014), *temperature* (*M*=.013), *flow* (*M*=.013), *field* (*M*=.012), *power* (*M*=.012), and *energy* (*M*=.012). The Physics knowledge category was mapped from other corpora as well. Publication keyword unigrams and bigrams also seemed to fall under Physics or Product and Processing. Terms fell into mathematical concepts and the steps of the software production process.

Job Postings

The notable terms identified by the job postings include *software* (*M*=.060), *engineer* (*M*=.046), *engineering* (*M*=.041) *development* (*M*=.040), *design* (*M*=.040), *research* (*M*=.034), *team* (*M*=.032), *analysis* (*M*=.030), *support* (*M*=.030), *test* (*M*=.028), *product* (*M*=.024), and *customer* (*M*=.024). These terms all seems to fall under the umbrella of the Product and Processing knowledge component. Specific emphasis is placed on software engineering within the enumerated terms from the model.

Course Descriptions

All terms within the course description corpus were expected. However, note the lack of physics, engineering and technology related terms

Conclusion

This chapter included a reiteration of the study purpose statement and addressed in detail the data collected for each research question. Analyses were presented using BoW, BoN, TD-

IDFs, and LDA Models. Results from my first research question revealed that *modeling simulation* and *system* were the most salient terms within the publication abstract corpus. Within the publication unigrams a theme of natural resources and physics emerged. Publication keyword unigrams and bigrams also seemed to fall under Physics or Product and Processing. Terms fell into mathematical concepts and the steps of the software production process. Results from the second research question revealed that the most salient terms within the job posting corpus included *system*, *experience*, and *job*. These terms all seems to fall under the umbrella of the Product and Processing knowledge component. Specific emphasis is placed on software engineering within the enumerated terms from the model. The results from my third research question show that the LDA model determined nine topics was the optimal number of topics. Of the topics identified, three emerged as different from the rest with terms related to academia, government, and industry/employment. The results from my fourth research question showed that notable terms within each topic included position, requirement, required for topic four, university, faculty, and education for topic six and army and ARL for topic nine. The results from the fifth research question revealed that the most salient terms within the course descriptions corpus included *system*, *simulation*, *course, topic*, *m&s,* and *technique*. All terms within the course description corpus were expected. However, note the lack of physics, engineering and technology related terms. An investigation of the hypothesis for this study revealed that while many of the terms identified are in all three corpora, there are variations in the KSAs requested, taught, and applied. The next chapter provides a discussion of these results and recommendations for future work on the topic of this study.

CHAPTER FIVE: DISCUSSION

The purpose of this dissertation was to investigate natural language in Modeling and Simulation (M&S) domain-specific job postings (i.e., what employers request), course descriptions (i.e., what is being taught), and academic literature (i.e., what is applied in practice) to enumerate important terms and determine common relationships between M&S topics. Based on the results presented in the previous chapter, I have included a brief summary of them in this chapter along with a discussion of the findings, practice implications, and recommendations for future work.

Summary of Results

Research Question 1: *What are the KSAs applied most frequently in M&S academic literature*? To answer this research question, I created a corpus of text scraped from an open-source journal database and used this to run a Term Frequency- Inverse Document Frequency (TF-IDF) analysis the results showed that the top most salient words within this corpus were *model*, *simulation* (both expected), *system*, *result*, and *method*. Within the publication unigrams a theme of natural resources and physics emerged. Terms followed mathematical concepts and the steps of the software production process. Publication keyword unigrams and bigrams also seemed to fall under Physics or Product and Processing knowledge components. This is interesting because Bowen and Rudenstine's (1992), research showed that natural science dissertation topics have the highest rate of degree completion. This may suggest that students who chose a topic investigating natural science using M&S are most likely to graduate and apply

the techniques and principles learned within academic publications than those investigating other topics using M&S techniques.

Research Question 2: *What are the KSAs most requested in M&S job listings within the United States?* To answer this research question, I created a corpus of text scraped from job postings and used this to run a TF-IDF analysis. The results showed that the top most salient words within this corpus were *system, experience,* and *job.* These terms all seems to fall under the umbrella of the Product and Processing knowledge component. Specific emphasis is placed on software engineering within the enumerated terms from the model.

Research Question 3: *How should M&S job types be categorized?* ***Research Question 4:*** *What are the KSAs most identified per job type?* To answer these two questions, I used the job posting corpus to run an LDA analysis. The results from my third research question show that the LDA model determined nine topics was the optimal number of topics. Results showed that based on intertopic distance, three topics stood out. The fourth topic's most salient terms included *position*, *experience,* and *information*. Additionally, the sixth topic's most salient terms included *university*, *save,* and *job*. Further, the ninth topic's most salient terms included *research*, *army,* and *science*. The results from my fourth research question showed that notable terms within each topic included *position*, *requirement*, and *required* for topic four, *university*, *faculty*, and *education* for topic six, and *army* and *arl* for topic nine. Of the topics identified, three emerged as different from the rest with terms related to academia, government, and industry/employment, which is a common job categorization scheme (L. Bair & Jackson, 2013; Kincaid & Westerlund, 2009).

Research Question 5: *What are the KSAs taught most frequently in M&S graduate level course descriptions for Universities within the United States?* To answer this research question, I created a corpus of text scraped from course descriptions from M&S program websites and used this to run a TF-IDF analysis the results showed that the top most salient words within this corpus were *system*, *simulation,* and *course*. All terms within the course description corpus were expected. However, note the lack of physics, engineering and technology related terms

Hypothesis 1*: *Scraping multiple types of M&S documents (job postings, course descriptions, and academic publications) will produce a difference in the KSAs (topics) most frequently mentioned in each source type. To address my hypothesis, I examined the results and looked for terms unique to each salient term list. Unique terms to the top 30 most salient list for the publication abstracts corpus were *result, method, based, using, paper, used, process parameter, water, temperature, distribution, equations, flow, effect, field, power, energy, structure, show,* and *performance.* Unique terms to the top 30 most salient list for the job posting corpus were *experience, job, engineer, engineering, technology, work, team, support, test, requirement, position, year, skill, science, product, required, information,* and *customer*. Unique terms to the top 30 most salient list for the course description corpus were *course, topic, M&S, techniques, student, emphasis, computer, program, security, graduate, problem, prerequisite, interest, include, introduction, instructor, visualization,* and *optimization.*

Unique unigrams to two of three top 30 most salient lists by source type show that the terms *development*, *research*, and *application* occurred between the job posting corpus and the course description corpus, the terms *study* and *network* occurred between the publication abstract corpus and course description corpus, and the term *time* occurred between the Job posting and

publication abstract corpora. There were also terms that were similar across all three source types which included *system*, *simulation*, *model*, *modeling*, *analysis*, *data*, and *design*. An investigation of the hypothesis for this study revealed that while many of the terms identified are in all three corpora, there are variations in the KSAs requested, taught, and applied. Based on this, the following discussion is provided to address implications of this work, limitations of the study, and future directions for research.

Conclusions

Modeling and Simulation Ontology

While the most salient terms are discussed above, NLP and LDA often require human input to make sense of the data presented, as previous identified by Blei and colleagues (2002), Vajjala and colleagues (2020), and Zhao (2018), in the literature review. As such, I went through each of the salient terms lists to identify common M&S terms against the CSMP topics list. Then I map common M&S and notable terms to O*NET's list of KSAs (National Center for O*NET Development, 2020).

Modeling and Simulation Specific Terms Identified

Referring back to the list of CMSP topics (reproduce below in Figure 22 for readability), terms that specifically stood out compared to the silent terms lists presented throughout were *analysis*, *engineering*, *test*, (possibly referring to areas of expertise – category two), *numerical simulation*, *simulation model*, *mathematical model*, *monte carlo*, *mathematical modeling*, *agent based*, and *simulation based* (modeling methods – category four).

M&S-Domain Specific Knowledge	Areas of Specialized Expertise	
1. Concepts and context 1.1. Fundamental terms and concepts 1.2. Categories and paradigms 1.3. History of M&S	**2. Applications of M&S** 2.1. Training 2.3. Experimentation 2.5. Engineering 2.2. Analysis 2.4. Acquisition 2.6. Test and evaluation	
6. Supporting tools, techniques, and resources 6.1. Major simulation infrastructures 6.2. M&S resource repositories 6.3. M&S organizations	**4. Modeling methods** 4.1. Stochastic modeling 4.2. Physics-based modeling 4.3. Structural modeling 4.4. Finite element modeling and computational fluid dynamics 4.5. Monte Carlo simulation 4.6. Discrete event simulation 4.7. Continuous simulation 4.8. Human behavior modeling 4.9. Multi-resolution simulation 4.10. Other modeling methods	**5. Domains of use of M&S** 5.1. Combat and military 5.2. Aerospace 5.3. Medicine and healthcare 5.4. Manufacturing and material handling 5.5. Logistics and supply chain 5.6. Transportation 5.7. Computer and communications systems 5.8. Environment and ecology 5.9. Business 5.10. Social science 5.11. Energy 5.12. Other domains of use
7. Business and management of M&S 7.1. Ethics and principles for M&S practitioners 7.2. Management of M&S projects and processes 7.3. M&S workforce development 7.4. M&S business practice and economics 7.5. M&S industrial development		
M&S-Specific Software-Engineering-Related Expertise		
6. **Simulation implementation** 6.1. Modeling and simulation lifecycle 6.2. Modeling and simulation standards 6.3. Development processes 6.4. Conceptual modeling 6.5. Specialized modeling and simulation languages 6.6. Verification, validation, and accreditation 6.7. Distributed simulation and interoperability 6.8. Virtual environments and virtual reality 6.9. Human-computer interaction and virtual environments 6.10. Semi-automated forces (SAF) / computer generated forces (CGF) 6.11. Stimulation		
Domain Knowledge in Related Fields of Practice		
8. Related communities of practice and disciplines 8.1. Statistics and probability 8.2. Mathematics 8.3. Software engineering and development 8.4. Systems science and engineering		

Figure 22: CMSP Exam Topics (Bair & Jackson, 2015)

Mapping terms to O*NET

KSAs elements are further broken down into subcategories. This categorical information paired with the terms identified as most salient can help determine appropriate M&S KSAs. Tables 38, 39, and 40 discuss the KSAs identified by source type.

Table 38: KSAs Identified in Publication Abstracts and Keywords

Knowledge Component	Skills	Abilities
Administration and Management	Critical Thinking	Cognitive Flexibility
Biology	Mathematics	Deductive Reasoning
Chemistry	Monitoring	Flexibility of Closure
Communication and Media	Reading Comprehension	Inductive Reasoning
Design	Science	Information Ordering
Engineering and Technology	Writing	Mathematical Reasoning
Mathematics	Complex Problem Solving	Number Facility
Physics	Mgmt. of Material Resources	Perceptual Speed
Production and Process	Mgmt. of Personnel Resources	Speed of Closure
	Time Management	Time Sharing
	Persuasion	Visualization
	Judgement and Decision Making	Written Comprehension
	Systems Analysis	Written Expression
	Systems Evaluation	Control Precision
	Equipment Selection	
	Operation and Control	
	Operation Analysis	
	Quality Control Analysis	
	Technology Design	

Table 39:KSAs Identified in Job Postings

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Computers and Electronics	Critical Thinking	Deductive Reasoning
Design	Learning Strategies	Flexibility of Closure
Education and Training	Mathematics	Fluency of Ideas
Engineering and Technology	Science	Inductive Reasoning
Geography	Writing	Information Ordering
Mathematics	Complex Problem Solving	Mathematical Reasoning
Physics	Mgmt. of Material Resources	Number Facility
Production and Process	Mgmt. of Personnel Resources	Originality
Public Safety and Security	Time Management	Perceptual Speed
Sale and Marketing	Coordination	Speed of Closure
	Instruction	Time Sharing
	Service Oriented	Written Comprehension
	Judgement and Decision Making	Written Expression
	Systems Analysis	
	Systems Evaluation	
	Operation and Control	
	Operation Analysis	
	Programming	
	Quality Control Analysis	
	Technology Design	

Table 40: KSAs Identified in Course Descriptions

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Computers and Electronics	Critical Thinking	Deductive Reasoning
Design	Learning Strategies	Flexibility of Closure
Education and Training	Mathematics	Inductive Reasoning
Fine Arts	Science	Information Ordering
Engineering and Technology	Complex Problem Solving	Mathematical Reasoning
Mathematics	Mgmt. of Material Resources	Number Facility
Medicine and Dentistry	Mgmt. of Personnel Resources	Perceptual Speed
Production and Processing	Coordination	Problem Sensitivity
Psychology	Instruction	Selective Attention
Public Safety and Security	Service Oriented	Speed of Closure
Sociology and Anthropology	Judgement and Decision Making	Visualization
Transportation	Systems Analysis	
	Systems Evaluation	
	Operation and Control	
	Operation Analysis	
	Programming	
	Quality Control Analysis	
	Technology Design	

Modeling and Simulation Expert Models

Multiple Bodies/Books of Knowledge (BoKs) may be necessary, as many M&S curriculum and domain experts have started that M&S should be broken out into specializations (Bair & Jackson, 2013, 2015; Birta, 2003; Mielke et al., 2009; Ören, 2011b, 2014; Ören & Waite, 2010; Padilla et al., 2011; Sarjoughian & Zeigler, 2000) What is not necessarily agreed upon is the way in which M&S jobs should be categorized into these specializations. Several early M&S domain articles concerning M&S formalization suggest specialization categories (L. Bair & Jackson, 2015; Birta, 2003; Mielke et al., 2009, 2008; Padilla et al., 2011). Four of these articles break M&S Professionals into two categories, either "user/manager" or "developer/technical," (L. Bair & Jackson, 2015; Mielke et al., 2009, 2008; Padilla et al., 2011). Birta (2003) conversely categorizes M&S specialists into three different categories: model developer, simulation program developer, and end-user support. Further investigation into

the techniques specifically associated with these titles can help categorize M&S jobs, which will make it easier to determine and organize the types of KSAs necessary to M&S students' success. However, I think it is important to note that this categorization scheme does not include the design process. Each corpus listed *design* as one of the top 30 most salient terms, which makes me think that designing is an extremely pertinent component of M&S that the field is largely ignoring. Table 41, 42 and 43 show the KSAs identified per job posting topic identified.

Table 41: KSAs Identified in Job Postings Topic 4 (Proposed Category: Industry)

Knowledge Component	Skills	Abilities
Computers and Electronics	Learning Strategies	Cognitive Flexibility
Education and Training	Science	Flexibility of Closure
Engineering and Technology	Writing	Perceptual Speed
Production and Process	Complex Problem Solving	Speed of Closure
Public Safety and Security	Mgmt. of Personnel Resources	Time Sharing
	Time Management	Written Comprehension
	Coordination	Written Expression
	Instruction	
	Service Oriented	
	Judgement and Decision Making	
	Technology Design	

Table 42: KSAs Identified in Job Postings Topic 6 (Proposed Category: Academia)

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Computers and Electronics	Critical Thinking	Deductive Reasoning
Education and Training	Learning Strategies	Flexibility of Closure
Engineering and Technology	Science	Inductive Reasoning
Geography	Complex Problem Solving	Mathematical Reasoning
	Mgmt. of Personnel Resources	Number Facility
	Time Management	Perceptual Speed
	Instruction	Speed of Closure
	Judgement and Decision Making	Time Sharing
		Written Comprehension
		Written Expression

Table 43: KSAs Identified in Job Postings Topic 9 (Proposed Category: Government)

Knowledge Component	Skills	Abilities
Administration and Management	Active Learning	Cognitive Flexibility
Education and Training	Mathematics	Flexibility of Closure
Engineering and Technology	Science	Inductive Reasoning
Geography	Writing	Mathematical Reasoning
Mathematics	Complex Problem Solving	Perceptual Speed
Physics	Mgmt. of Material Resources	Speed of Closure
	Mgmt. of Personnel Resources	Time Sharing
	Time Management	
	Coordination	
	Service Oriented	
	Judgement and Decision Making	
	Systems Analysis	
	Systems Evaluation	
	Operation and Control	
	Operation Analysis	
	Programming	
	Quality Control Analysis	
	Technology Design	

Recommendations

Since many of the terms identified belong to product/software development, I use a software development lifecycle approach, organizing jobs into research-based, design-based, development-based, evaluation-based, and deployment/maintenance-based jobs and provide examples of common M&S titles in each category (Figure 23).

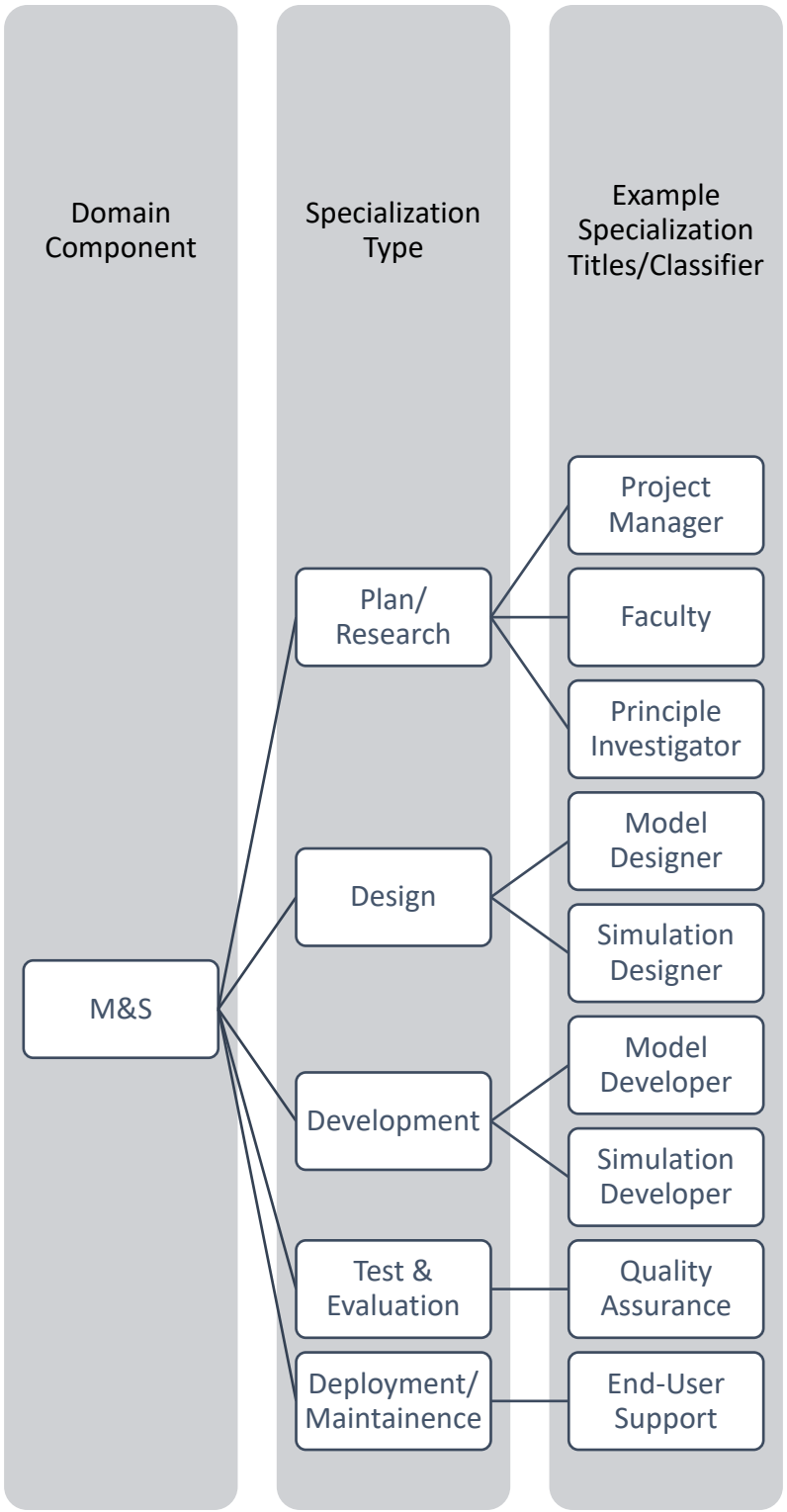

Figure 23: M&S Specialization Categories from Literature Review

Limitations

There are three common limitations to the BoW family of methods (e.g. BoW, BoN, TF-IDF, and LDA), which include vocabulary, sparsity, and meaning (Brownlee, 2017). The size of the potential library can impact the sparsity of the documents used to inform the text analysis (Brownlee, 2017). If the vocabulary is too large there will be a lot of *noise*, whereas, with a smaller vocabulary you may not get enough data to represent the domain adequately.

The corpora in the present dissertation used similar size documents (publication abstracts, job postings, course descriptions). The publication abstract corpus and the job posting corpus were noisy compared to the course description and required quite a bit of data cleaning to produce meaningful results. This may be alleviated in the future by writing a stand-alone scraper versus using Octoparse. Additionally, sparsity can also make it difficult computationally represent this information (Brownlee, 2017). Data sparsity can be addressed by gathering and updating data frequently. This can be accomplished through Octoparse, which allows the user to schedule automated scrapes periodically, however this is a paid feature for the application (*Octoparse*, 2020).

Lastly, BoW methods ignore context, meaning, word arrangement, and synonyms (Brownlee, 2017). For example, in an M&S context the word requirements pulled from job postings could mean job requirements related to work tasks or it could mean system requirements as in job seekers should be aware of how to formulate system requirements. "The evaluation and interpretation of topic models is still challenging, and there's no consensus on it yet. Parameter tuning for topic models can also take a lot of time. …As mentioned previously, there's no straightforward procedure to know the number of topics; we explore with multiple

values based on our estimates about the topics in the dataset." (Vajjala et al., 2020). One way to infer meaning from unigrams is to explore the *n*-gram models. Bigrams are used more to infer context. In the present dissertation, of the bigram models that could compute, the terms identified were more meaningful than the unigram models. However, inferring meaning and mapping to O*NET was still difficult and time consuming. Further NLP analysis and feature/topic labeling could reduce the burden of inferring meaning. Proposed labels include job type (academia, industry, and government) and lifecycle type (plan/research, design, development, test & evaluation, and deployment & maintenance).

Future Work

With a goal like a university-wide simulation system, there is plenty of future work needed to design develop and deploy all the moving components. This process will, require iteration, stakeholder verification, and updates to the programming components along the way. As such, the following paragraphs discuss direction for this work concerning the M&S ontology, M&S expert models, and the overall university-wide system.

Future Work for Modeling and Simulation Ontology

The present dissertation focuses on the first iteration of a semi-formal M&S domain ontology. This is by no means a completed product. It is a living formalization, requiring frequent updates. Again, this may be another justification for using software (e.g., Octoparse) that can automatically schedule regular web scrapers to run and gather data for more advanced NLP techniques and time series analysis to view changes to the field over time.

Future work could also focus on automating document collection, text classification, and article dissection using supervised learning algorithms to inform domain changes. Once data sets are labeled we can start using them for supervisory machine learning methods. Additionally, in future work, deep learning techniques could be used to generate additional data points in instances where data is lacking or incomplete (Brownlee, 2017b). NLP text generation algorithms can be used in the future to quickly generate course descriptions, course content, and provide individualized adaptive tutoring in the form of a digital mentor. I also suggest future work on investigating other variables using a similar NLP approach. For example, future investigations could focus on number of views, number of downloads, or index values within the academic publications and/or different levels of education within M&S within job postings. Further, I could also refine the LDA model by exploring additional permutations.

Further, work is also needed in creating a rigorously formal ontology, and thus should start considering appropriate tools to program such an ontology. Using Unified Modeling Language (UML), I can start to outline some example classes, slots, and values. Figure 24 shows and example of this notation in the context of an M&S ontology.

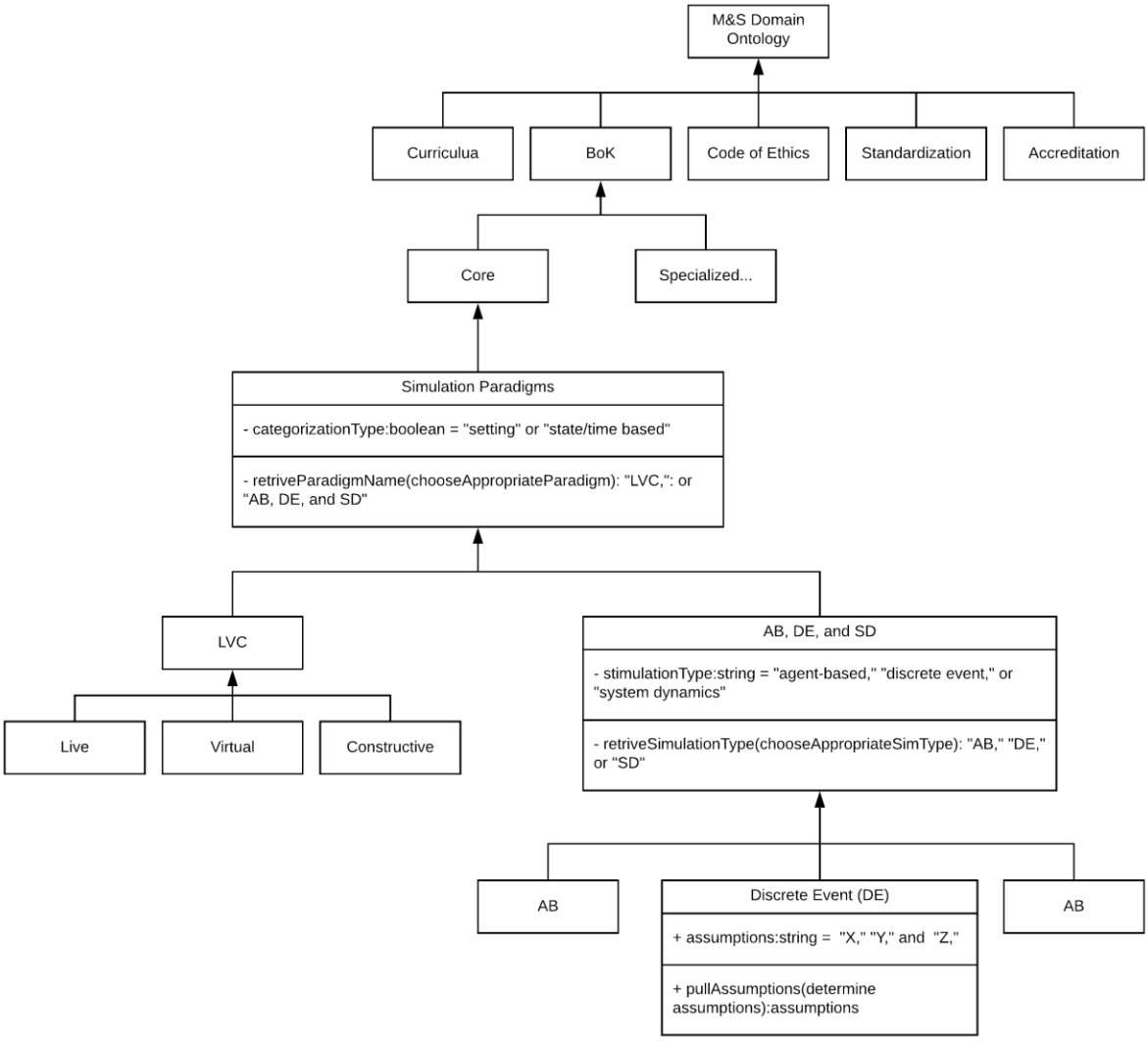

Figure 24: Example M&S Ontology UML Diagram

Ontology learning is a growing field concerned with automating ontology creation. Within ontology learning however, "Object Oriented Programming centers primarily around methods on classes—a programmer makes design decisions based on the operational properties of a class, [a]s a result, a class structure and relations among class in an ontology are different from the structure for similar domain in an object-oriented program." (Noy & McGuinness,

2004). **Web Ontology Language** (OWL) is "is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit," (*Web Ontology Language (OWL)*, 2013). Object Oriented Programming is too slow and less flexible than OWL, (Noy & McGuinness, 2004). Ontology standards exist and are rooted in the **Resource Description Framework** (RDF) which is "a framework for expressing information about resources. Resources can be anything, including documents, people, physical objects, and abstract concepts," (W3C Working Group, 2014b). RDF Schema is used to draw relationships between what the W3C Working group identify as *resources* (note different meaning to resource allocation mentioned in previous chapters). The terms identified in the present dissertation could be programmed in as various classes and class properties (W3C Working Group, 2014a). This is one of the next steps in creating a programmed and rigorously formal ontology. Another step in M&S domain modeling is to clarify theoretical versus applied components.

Theory builds the foundation of a discipline. It collects the axioms and rules that govern the discipline," (Padilla et al., 2011). Delving into the theoretical basis for a discipline leads to greater understanding of the guidelines that help shape the discipline (Padilla et al., 2011), and thus its ontology and curriculum. Unfortunately, there is a lack of M&S specific theory is due to 1) M&S being an applied field and 2) M&S is a *victim of its own success*, (Padilla et al., 2011). M&S has grown quickly over the last few years and because of its *criticality* to the government. The field has had little/no time to solidify M&S specific theoretical foundations

(Padilla et al., 2011). However, I expect with technology advances and the growing need for solutions to complex problems, M&S educational programs will grow rapidly. With the emergence of additional M&S programs, it will be necessary to record theoretical foundations unique to M&S; examples include the theories of *composability* and *interoperability* (Padilla et al., 2011). Composability and interoperability "seek to combine models and simulations for reuse," (Padilla et al., 2011).

Confusion concerning the philosophical foundation of M&S (beyond normal philosophical debate) comes from the fact that the field borrows from many other disciplines. For example, verification and validation, an important aspect of M&S is historically based in empiricism; however, the mathematical and physics-based formulas used to develop accurate models is a more rationalist endeavor (Padilla et al., 2011). So, should M&S curriculum and theory derive from empiricism or rationalism?

M&S is based firstly upon empiricism, the idea that our senses help us derive our knowledge. It could be argued that M&S is routed in observation of everyday world objects and processes, but what about theoretical models? This leads to positivism: "A model is a simplification of reality," (Padilla et al., 2011). However, simplification for simplifications sake can be perceived as *lazy* or *incomplete*. Thus, from a post-positivism point of view: "[A model] is a purposeful simplification of a perception of a situation in order to generate a theory or an explanation," (Padilla et al., 2011). What this means for M&S as a discipline is that, scope is an important factor and should be considered by all stakeholders (e.g., which KSAs are most important?). What this means for the present dissertation, is that many elements of the

system have been omitted. The reader should expect some holes in the overall system. I encourage others to investigate these holes and develop more robust and complex models.

Future Work in Modeling and Simulation Expert Models

One constraint of this research is the current state of the M&S job market. "In 2010, The DOD M&S [Human Capital Strategy] HCS estimated that its M&S workforce has approximately 30,000 military, government civilians, and contractors costing $2.25B in labor alone,"(L. Bair & Jackson, 2015; Department of Defense, 2009). In 2011, Military M&S projects were estimated to be worth $5 Billion (L. Bair & Jackson, 2013; Kimla et al., 2012). The Florida High Tech Corridor Council estimates that M&S has contributed to $4.8 out of the $8 billion in state sales in Florida alone (L. Bair & Jackson, 2013; *Modeling, Simulation, and Training*, 2017). Even during volatile economic changes, M&S shows a growing trend (L. Bair & Jackson, 2013). These details speak to the value of M&S.

While many M&S organizations have collected data on the financial impact of M&S, the way our government classifies M&S businesses and jobs makes it difficult to measure M&S's *true* impact on our country. In the United States, job types are organized using the North American Industry Classification System (NAICS) categorization codes, which allow officials the ability to collect, analyze, and publish economic data, (*North American Industry Classification System*, 2017). These are the types of codes used by Occupational Information Network (O*NET) to determine KSAs per job type. "By the NAICS code itself and its defining characteristics; the U.S. government implicitly defines an industry as a group of business endeavors having or using similar processes to produce goods or services," (L. Bair & Jackson, 2013). In this instance, M&S would be considered an industry. However, M&S entities are often

classified into multiple categories, based upon the application areas/domains rather than M&S processes or products. Thus, M&S's economic impact is not accurately reflected with existing NAICS categorizations (L. Bair & Jackson, 2015).

Many established M&S government and industry organizations[5] cosponsored a proposal to establish one single NAICS code for M&S during the 2012 NAICS revision process; however, the proposal was denied, (L. Bair & Jackson, 2013). This denial of one code by the NAICS, further demonstrates the complexity and interconnectedness of M&S as an industry (L. Bair & Jackson, 2013). "[I]t is incumbent upon a nascent M&S industry not only to differentiate itself from these other NAICS-defined industries... but also to expand its myopic view of M&S beyond those industry support-areas with which its sponsors are most familiar into the greater breadth of M&S's use," (L. Bair & Jackson, 2013). In the meantime, M&S can still move forward and thrive without the NAICS code, but M&S professionals and recent graduates will need to be vigilant in looking for appropriate employment. Very few jobs are listed as *Modeling and Simulation Professional* and those that are listed as such, vary in the types of skills required. Thus, it may behoove M&S practitioners to determine the standardized set of *specializations*. Until then, M&S Professionals will need to determine a way of marketing their unique set of skills on a case-by-case basis, relying heavily on network connections within the field.

Another future work avenue would be to further research the labels provided by O*NET. Labels in other fields could generalize to M&S and labeled data will allow for further investigation of terms using supervisory machine learning methods. Additional validation of mapping of terms should also occur. A potential guiding question for this research could be, *how do M&S experts sort a set of terms*? This study could include a card sort task for M&S experts.

Future Work for University-Wide System

In future iterations, I plan to use this information to create a university wide simulation system. Table 44 shows the remaining steps of Law's (2003) seven-step simulation framework specifically applied to common adaptive tutor components, (Sottilare, 2015).

Table 44: High-Level Research Agenda based on Law's (2003) Seven Step Framework

Phase 1: System Formation and Conceptual Modeling
- Problem Formation
  - Determine overall goals and objectives of the project
  - Establish *system* and *model* scope
  - Identify appropriate stakeholders
- Collect information
  - Collect information from existing system (if applicable)
  - Identify system configurations
  - Determine system assumptions
- Construct overall conceptual model

Phase 2: Validation of Conceptual *Domain* Model
- Determine specific research questions
- Select appropriate measures for the research questions
- Collect data
- Clean data
- Analyze data
- Visualize data
- Create domain model
- Validate domain model

Phase 3: Validation of Conceptual Student Model
Phase 4: Validation of Conceptual Instructional Design Model
Phase 5: Validation of Conceptual Resource Allocation Model
Phase 6: Validation of User Interface Design
Phase 8: Program and Validate Domain Model
Phase 9: Program and Validate Student Model
Phase 10: Program and Validate Instructional Design Model
Phase 11: Program and Validate Resource Allocation Model
Phase 12: Program User Interface and Validate User Experience
Phase 13: Integrate Overall System Model
Phase 14Validate Simulation
Phase 15: Design and Conduct Strategic Planning Experiments
Phase 16: Report Simulation Results

(Adapted from Law, 2003; Sottilare, 2015)

I completed a notable amount of research on the literature for sections of the university-wide system including the student model, instructional design model, resource allocation, and user interface components. I present some of that information here to inspire further direction for future work.

Student Model

An adaptive tutoring system that will help identify the optimal path for student *skill* and *knowledge* acquisition. *Abilities* are somewhat fixed and should be factored in when determining appropriate variables for both student and expert profiles/models. It is assumed that students with similar abilities (or internal factors) to experts will succeed in similar careers if given an individualized plan for skill and knowledge acquisition. While an adaptive tutor is outside of the scope of the present document, thinking about the curriculum map structure in terms of adaptive tutoring components could benefit the sustainability of the program moving forward.

Instructional Design Model

The purpose of this model is to house information the system can use to emulate the program and course structure. The instructional design model has two sub-models: the curriculum design and the course design. The curriculum design components will utilize ontology information to inform program (terminal) learning objectives and course sequencing. The course design component will house information about specific courses (e.g. course – or enabling – learning objectives, topics covered, resources needed). Thus, a rigorously formal ontology is necessary to inform domain knowledge structuring (e.g., taxonomy, ontology), determine topic prioritization within the graduate curriculum, inform which instructional strategies may be most effective (based on learning outcomes and the strategy's success in similar domains), and help determine measurements of success for the students and program.

Resource Allocation Model

The purpose of the resource allocation model is to house information about the number, status (e.g., in use), and condition of the inanimate resources and agents that contribute to labor.

The resource allocation model has two main components: learning resources and staff. Learning resources consist of the materials needed to teach including classrooms (on campus), a learning management system (for on campus and online students), and other materials (e.g., potential textbooks, smart boards). Staff includes both administrators and faculty. Both groups of people will be simulated using agent-based techniques in future iterations of the project to simulate more realistic relationships between staff and students. Massy's (Massy, 2016), activity-based costing model mentioned in chapter two can be incorporated to tie course activities to specific resources.

User Interface Model

In future iterations of the university-wide system, I plan to use these metrics and strategies outlined by the UCF Board of Trustees, (2016) to inform analyses that I perceive to be useful to the end-user. As mentioned in chapter two, each of these bullet points can serve as a separate user scenario for the system and a means of creating what-if scenarios for end-users.

Research Benefit and Implications

The research benefits are detailed below and include information on the project's contribution to the field and broader impact.

Contribution to the Field

This document also has the potential to contribute to economic growth by recommending changes for improving the current and future workforce. For example, in Florida alone, M&S is a $5-billion-dollar industry supporting nearly 30,000-60,000 jobs (*Modeling, Simulation, and Training*, 2017). The workforce will continue to grow as demands for M&S professionals grow,

requiring a larger number of highly qualified graduates. Additionally, building an M&S ontology can potentially contribute to curriculum standards within M&S, NLP labels for categorizing future M&S articles, and adaptive tutoring for teaching M&S in the future.

Broader Impact

M&S techniques are utilized for many applications across multiple disciplines; thus, there is a potential that the present dissertation could improve education and research metrics beyond the immediate domain. First, the university-wide system has the potential to improve strategic decisions for hiring and allocating tasks to faculty in higher education programs. Second, the university-wide system has the potential to contribute to the way in which *student success* and *program success* are modeled, measured, and assessed across multiple program types. The system is modular, allowing the system potentially to generalize to other higher educational programs. This investigation could potentially change the foundation of higher-education by providing a tool for creating sustainable data-driven topics and prioritization of topics. This data-driven information could be used to strengthen proposals submitted to funding agencies or inform policy changes within the university.

.

LIST OF REFERENCES

Academic Programs. (2017). The MOVES Institute-Naval Postgraduate School. https://www.movesinstitute.org/academic-programs/

Academic Torrents. (2014, December 31). Internet Archive. https://archive.org/details/academictorrents?and%5B%5D=Education&sin=

Ameisen, E. (2020). *Building Machine Learning Powered Applications*. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/building-machine-learning/9781492045106/

Anti-Corruption Risk Assessment Taskforce. (2013). *A Guide for Anti-Corruption Risk Assessment*. United Nations Global Compact Office. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2FRiskAssessmentGuide.pdf

Bair, C. H., & Haworth, J. G. (2004). Doctoral Student Attrition and Persistence: A Meta-Synthesis of Research. In *Higher Education: Handbook of Theory and Research: Vol. XIX* (pp. 481–534). Kluwer Academic Publishers.

Bair, L., & Jackson, J. J. (2013). *M&S Professionals Domains, Skills, Knowledge, and Applications*. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL. http://iitsec.ndia.org/about/PublicationsProceedings/Documents/BP_HSE_13313_Paper.pdf

Bair, L., & Jackson, J. J. (2015). *Modeling and Simulation Professionals-Meeting the Demand*. Interservice/industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

Birta, L. G. (2003, February 19). *The Quest for the Modelling and Simulation Body of Knowledge*. The Sixth Conference on Computer Simulation and Industry Application, Instituto Technologico de Tijuana, Mexico.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*, *14*. https://proceedings.neurips.cc/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf

Bowen, H. G., & Rudenstine, N. J. (1992). *In Pursuit of the PhD*. Princeton University Press.

Boyatzis, R. (1998). *Transforming Qualitative Information: Thematic analysis and code development.* Sage.

Brownlee, J. (2017a). *A Gentle Introduction to the Bag-of-Words Model*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Brownlee, J. (2017b, July 24). *How Much Training Data is Required for Machine Learning?* Machine Learning Mastery. https://machinelearningmastery.com/much-training-data-required-machine-learning/

Council of Chief State School Officers. (2017, August). *Transparency in Stakeholder Engagement: A Tool to Help Demonstrate How Stakeholders Informed the State ESSA Plan*. https://ccsso.org/resource-library/transparency-stakeholder-engagement

Department of Defense. (2009). *DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)* (5000.61; Instruction, pp. 1–10). http://www.msco.mil/documents/_25_M&S%20BOK%20-%2020101022%20Dist%20A.pdf

Department of Modeling, Simulation and Visualization Engineering. (n.d.). Old Dominion University. Retrieved December 12, 2017, from http://catalog.odu.edu/graduate/frankbattencollegeofengineeringandtechnology/modelingsimulationvisualizationengineering/#masterofengineeringmodelingsimulationtext

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/citation_policy.html

Dwyer, C. (2020). *CMSP*. NTSA. https://www.trainingsystems.org/cmsp

Education-Arizona Center of Integrative Modeling and Simulation. (2017). ASU Engineering Arizona Center of Integrative Modeling and Simulation. https://acims.asu.edu/education/

Expert Systems and Simulations. (2018). https://www.lynda.com/Software-Development-tutorials/Expert-systems-simulations/667369/726676-4.html

Fereday, J., & Muir-Cochrane. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach for Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods, 5*(1), 80–92.

Galli, K. (2018, October 25). *Complete Python Pandas Data Science Tutorial! (Reading CSV/Excel files, Sorting, Filtering, Groupby)*. https://www.youtube.com/watch?v=vmEHCJofslg&t=2264s

Galli, K. (2020, July 13). *Solving Real World Data Science Tasks with Python Pandas!* https://www.youtube.com/watch?v=eMOA1pPVUc4&t=2172s

General and Narrow AI. (2019). https://www.lynda.com/IT-Infrastructure-tutorials/General-narrow-AI/765326/5022609-4.html

Golde, C. M. (1996). *How departmental contextual factors shape doctoral student attrition* [Doctoral Dissertation]. Stanford University.

Google. (2020). *Dataset Search*. Dataset Search. https://datasetsearch.research.google.com/

Goyvaerts, J. (2020, October 5). *Regular-Expressions.info*. Regular-Expressions.Info. https://www.regular-expressions.info/

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Guan, J., Nunez, W., & Welsh, J. F. (2002). Institutional strategy and information support: The role of data warehousing in higher education. *Campus-Wide Information Systems*, *19*(5), 168–174.

Gupta, N., & Grover, S. (2013). Introduction to Modeling and Simulation. *Int. J. IT, Eng. Appl. Sci. Res*, *2*(4), 45–50.

Harden, R. M. (2001). AMEE Guide No. 21: Curriculum mapping: A tool for transparent and authentic teaching and learning. *Medical Teacher*, *23*(2), 123–137.

Hinton, K. (2012). *A Practical Guide to Strategic PLanning in Higher Education*. Society for College and University Planning. https://www.bu.edu/strategic-plan-taskforce/files/2018/08/A-Practical-Guide-to-Strategic-Planning-in-Higher-Education.pdf

Huguley, S. (1988). *An investigation of obsticals to completion of the disseration and of doctoral student attitudes toward the dissertation experience* [Doctoral Dissertation]. Pepperdine University.

Karaganis, J., McClure, D., Boyer, J., Leonard, J., Ryu, M., & Genosha. (n.d.). *Open Syllabus Explorer*. https://opensyllabus.org/

Kershaw, J. A., & Mood, A. M. (1970). Resource allocation in higher education. *The American Economic Review*, *60*(2), 341–346.

Kimla, D., Pannu, A., Srimoolanathan, B., & Webb, S. (2012). *Global military training and simulation market assessment: Stable growth despite defence budgets cuts* (No. M7C5-16).

Kincaid, J. P., & Westerlund, K. K. (2009). *Simulation in education and training*. 273–280.

Kinsley, H. (2015, May 7). *Lemmatizing—Natural Language Processing With Python and NLTK p.8* [YouTube]. https://www.youtube.com/watch?v=uoHVztKY6S4&list=PLQVvvaa0QuDf2JswnfiGkliBInZnIC4HL&index=8

Law, A. M. (2003). Designing a simulation study: How to conduct a successful simulation study. *Proceedings of the 35th Conference on Winter Simulation: Driving Innovation*, 66–70. http://dl.acm.org/citation.cfm?id=1030829

Lewis, F., & Rowe, P. (2010). The Certified Modeling and Simulation Professional (CMSP) Program: Why it was created, where it stands now, and what you can do to support it. *SCS M&S Magazine*, *January*, 1–5.

Loper, M. L., Henninger, A., Diem, J. W., Petty, M. D., & Tolk, A. (2011). Educating the workforce: M&S professional education. *Proceedings of the Winter Simulation Conference*, 3968–3978.

Magesh. (2019, December 26). *Topic Modeling: Art of Storytelling in NLP*. Technovators. https://medium.com/technovators/topic-modeling-art-of-storytelling-in-nlp-4dc83e96a987

Mah, D. M. (1986). *The process of doctoral candidate attrition: A study of the ABD phenomenon* [Doctoral Dissertation]. University of Washington.

Massy, W. F. (Ed.). (1996). *Resource Allocation in Higher Education*. Michigan Publishing.

Massy, W. F. (2016). *Course-Level Activity-Based Costing as an Academic and Financial Tool* (TIAA Institute for Higher Education: Understanding Academic Productivity). National Association of College and University Business Office. https://www.ucr.edu/sites/g/files/rcwecm986/files/2018-03/tiaa-course.pdf

Matplotlib Development Team. (2020, September 15). *Matplotlib*. https://matplotlib.org/

Mielke, R. R., Scerbo, M. W., Gaubatz, K. T., & Watson, G. S. (2009). A model for multidisciplinary graduate education in modelling and simulation. *International Journal of Simulation and Process Modelling, 5*(1), 3–13.

Mielke, R. R., Scerbo, M. W., Gaubatz, K. T., & Watson, G. S. (2008). A multidisciplinary model for M&S graduate education. *Proceedings of the 2008 Spring Simulation Multiconference*, 763–769. http://dl.acm.org/citation.cfm?id=1400671

Modeling and Simulation Doctor of Philosophy. (2017). The University of Alabama in Huntsville. https://www.uah.edu/cmsa/academics/msdegrees/444-research/cmsa/6301-msdegrees-phd

Modeling and Simulation Graduate Courses. (2017). The University of Alabama in Huntsville. http://www.uah.edu/cmsa/academics/msdegrees/444-research/cmsa/6326-modeling-and-simulation-graduate-courses

Modeling and Simulation Masters of Science. (2017). The University of Alabama in Huntsville. https://www.uah.edu/cmsa/academics/msdegrees/444-research/cmsa/6300-msdegrees-masters

Modeling and Simulation MS. (2017). UCF Graduate Catalog 2017-2018. http://www.graduatecatalog.ucf.edu/programs/program.aspx?id=1326&program=Modeling%20and%20Simulation%20MS

Modeling and Simulation Ph.D. (2017). UCF Graduate Catalog 2017-2018. www.graduatecatalog.ucf.edu/programs/program.aspx?id=1328&program=Modeling%20and%20Simulation%20PhD

Modeling, Simulation, and Training. (2017). The Florida High Tech Corridor. http://www.floridahightech.com/high-tech-industries/modeling-simulation-training/

Muszynski, S. Y. (1988). *The relationship between demographic/situationa factors, cognitive/affective variables, and needs and time to completion of the doctoral program in psychology* [Doctoral Dissertation]. Kent State University.

National Center for O*NET Development. (2020a, August 18). *Browse by O*NET Data: Abilities*. O*NET Online. https://www.onetonline.org/find/descriptor/browse/Abilities/

National Center for O*NET Development. (2020b, August 18). *Browse by O*NET Data: Knowledge*. O*NET Online. https://www.onetonline.org/find/descriptor/browse/Knowledge/

National Center for O*NET Development. (2020c, August 18). *Browse by O*NET Data: Skills*. O*NET Online. https://www.onetonline.org/find/descriptor/browse/Skills/

Nilson, L. B. (2010). *Teaching at its best* (3rd ed.). John Wiley & Sons, Inc.

North American Industry Classification System. (2017). United States Census Bureau. https://www.census.gov/eos/www/naics/

Noy, N. F., & McGuinness, D. L. (2004). *Ontology Development 101: A Guide to Creating Your First Ontology*. https://protege.stanford.edu/publications/

NTLK Development Team. (2020, April 13). *Natural Language Toolkit*. NLTK 3.5 Documentation. https://www.nltk.org/

Numpy Development Team. (2020). *NumPy*. https://numpy.org/

Octoparse. (2020). https://www.octoparse.com/

Ören, T. (2011a). A Basis for a Modeling and Simulation Body of Knowledge Index: Professionalism, Stakeholders, Big Picture, and Other BoKs. *SCS M&S Magazine*, *2*(1), 40–48.

Ören, T. (2011b). A Critical Review of Definitions and About 400 Types of Modeling and Simulation. *SCS M&S Magazine*, *July*, 142–151.

Ören, T. (2014). The Richness of Modeling and Simulation and an Index of Its Body of Knowledge. In M. S. Obaidat, J. Filipe, J. Kacprzyk, & N. Pina (Eds.), *Simulation and Modeling Methodologies, Technologies and Applications* (Vol. 256, pp. 3–24). Springer International Publishing. http://link.springer.com/10.1007/978-3-319-03581-9_1

Ören, T. (2005). Toward the body of knowledge of modeling and simulation. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 1–19.

Ören, T., & Waite, B. (2010). Modeling and Simulation Body of Knowledge Index: An Invitation for the Final Phases of its Preparation. *SCS M&S Magazine*, *October*(4). https://www.researchgate.net/profile/Tuncer_Oeren/publication/228749480_Modeling_a

nd_Simulation_Body_of_Knowledge_Index_An_Invitation_for_the_Final_Phases_of_its_Preparation/links/0c960530f882e26424000000.pdf

Padilla, J. J., Diallo, S. Y., & Tolk, A. (2011). Do We Need M&S Science? *SCS M&S Magazine*, *Oct*(4), 161–166.

Pandas Development Team. (2020, February). *About Pandas*. Pandas. https://pandas.pydata.org/about/index.html

PMBOK Guide and Standards. (2020). PMBOK Guide and Standards. https://www.pmi.org/pmbok-guide-standards

Project Jupyter. (2020, October 21). *About Us*. https://jupyter.org/about

Rehurek, R. (2020, May 3). *Gensim 3.8.3*. Python Package Index (PyPI). https://pypi.org/project/gensim/

Sarjoughian, H. S., & Zeigler, B. P. (2000). *Towards Making Modeling & Simulation into a Discipline*. International Conference on Simulation and Multimedia in Engineering Education, Western Multi-Conference, Phoenix, AZ. http://s3.amazonaws.com/academia.edu.documents/40238041/0303d.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1486053676&Signature=56Ss2wrBsVmdu8zG9znTNfoibR4%3D&response-content-disposition=inline%3B%20filename%3DTowards_Making_Modeling_and_Simulation_i.pdf

Scientific Research. (2020). *Open Journal of Modeling and Simulation* [Open- source, academic article database]. https://www.scirp.org/journal/ojmsi/

Sievert, C., & Shirley, K. (2015). *PyLADvis*. https://pyldavis.readthedocs.io/en/latest/readme.html#:~:text=pyLDAvis%20is%20designed%20to%20help,an%20interactive%20web%2Dbased%20visualization.

Simulationist Code of Ethics. (2016). The Society for Modeling and Simulation International (SCS). http://scs.org/ethics/

Sokolowski, J. A., & Banks, C. M. (2009). *Principles of Modeling and Simulation: A Multidisciplinary Approach*. Wiley.

Sottilare, R. A. (2015). *Fundamentals of Adaptive Intelligent Tutoring Systems for Self-Regulated Learning* (ARL-SR-0318). US Army Research Laboratory. https://www.gifttutoring.org/attachments/download/1416/ADA614161.pdf

startrek.com staff. (2019). *Drink in the News of Chateau Picard Wine: Wines Thank Rock introduces new "Trek"-Inspired Wines, including one from the real Chateau Picard.* Star Trek. https://www.startrek.com/news/picard-wine-bordeaux-sirah

StataCorp LLC. (2016, March 3). *Introduction to Bayesian Statistics, Part 1: The Basic Concapts*. https://www.youtube.com/watch?v=0F0QoMCSKJ4

SysNucleus. (n.d.). *What is Web Scraping*. WebHarvy. https://www.webharvy.com/articles/what-is-web-scraping.html#targetText=Web%20Scraping%20(also%20termed%20Screen,in%20table%20(spreadsheet)%20format.

System Engineering. (2017). Johns Hopkins Whiting School of Engineering. https://ep.jhu.edu/programs-and-courses/programs/systems-engineering

The Size and Quality of a Data Set. (2019, July 11). Data Preparation and Feature Engineering for Machine Learning. https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality

Tolk, D. (2009). Emerging M&S challenges for human, social, cultural, and behavioral modeling. *Proceedings of the 2009 Summer Computer Simulation Conference*, 462–469. http://dl.acm.org/citation.cfm?id=2349570

UCF Board of Trustees. (2016). *UCF Collective Impact: Strategic Plan* (pp. 1–39). http://www.ucf.edu/wp-content/uploads/2012/08/UCF-Strategic-Plan-BOT-FINAL-052616-Web.pdf

U.S. Department of Education. (n.d.). *Integrated Postsecondary Education Data System (IPEDS)*. IES-NCES National Center for Education Statistics. https://nces.ed.gov/ipeds/use-the-data

U.S. Department of Labor, Employment & Training Administration, & National Center for O*NET Development. (2020, October 13). *O*NET Resource Center*. https://www.onetcenter.org/

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, *11*(2).

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical Natural Language Processing*. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/

W3C Working Group. (2014a, February 25). *RDF Schema 1.1*. W3C. https://www.w3.org/TR/rdf-schema/

W3C Working Group. (2014b, June 24). *RDF 1.1 Primer*. W3C. https://www.w3.org/TR/rdf11-primer/

Web Ontology Language (OWL). (2013, December 11). W3C Semantic Web. https://www.w3.org/OWL/

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology Learning from Text: A look back and into the future. *ACM Computing Surveys*, *44*(4). https://doi.org/10.1145/2333112.2333115

Xu, J. (2018, May 25). *Topic Modeling with LSA, PLSA, LDA & lda2Vec*. Nanonets Machine Learning APIs. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

Zhao, A. (2018, July 29). *Natural Language Processing in Python*. PyOhio, Ohio. https://www.youtube.com/watch?v=xvqsFTUsOmc&t=2019s