Meshing Capability and Threat-based Science and Technology (S&T) Resource Allocation

Technical Report SERC-2019-TR-010

June 30, 2019

Principal Investigator: Dr. Carlo Lipizzi, Stevens Institute of Technology

Co-Principal Investigators: Dr. Dinesh Verma, Stevens Institute of Technology and Dr. George Korfiatis, Stevens Institute of Technology

Research Team:

Stevens Institute of Technology: Mr. Dario Borrelli, Ms. Fernanda Capela, Ms. Megan Clifford, Mr. Prasad Desai, Mr. Ralph Giffin, Mr. Steven Hespelt, Dr. Steven Hoffenson, Ms. Sravanthi Kanchi, Mr. Mohammed Khan, Dr. Tom McDermott, Ms. Kara Pepe, Dr. Jose Ramirez-Marquez, Mr. Pedro Sá, Dr. Razieh Saremi, Dr. Hoong Yan See Tao, Mr. Rohit Shankar, and Dr. Zhongyuan Yu

The Systems Engineering Research Center (SERC) is a federally funded University Affiliated Research Center managed by Stevens Institute of Technology.

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Contract [HQ0034-19-D-0003, TO#0150].

Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense nor ASD(R&E).

No Warranty.
This Stevens Institute of Technology and Systems Engineering Research Center Material is furnished on an “as-is” basis. Stevens Institute of Technology makes no warranties of any kind, either expressed or implied, as to any matter including, but not limited to, warranty of fitness for purpose or merchantability, exclusivity, or results obtained from use of the material. Stevens Institute of Technology does not make any warranty of any kind with respect to freedom from patent, trademark, or copyright infringement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Our team would like to thank all the individuals that participated to the project, offering their resources, time, and effort.

All the U.S. Combat Capabilities Development Command Armaments Center (CCDCAC) members who worked with us, giving us credit for developing something that was so new we couldn't fully formalize into its details at the beginning. Providing us with valuable feedback, guiding us in applying a completely new text-driven approach to matters relevant to the project and to their core activities.

In the developments of this project, I had few moments when I had to break new ground to address specific roadblocks I was facing. That was a challenge I was happy to take and win, thanks to a great team I had the privilege to manage.

Our work on Natural Language Processing is literally state-of-the-art in this sector. We worked on a technology edge and both the team and the Sponsor demonstrated a high degree of intellectual flexibility to see how what we were developing could have really worked for the project.

I have been honored to work on a project potentially providing value to the core mission of our Sponsor, mission that is critical for our Nation.

Working in projects with this level of complexity and innovation is one of the great reasons to appreciate working for an Academic Research Center: special thanks to the Systems Engineering Research Center (SERC) and the Executive Director of SERC, Dr. Dinesh Verma, in particular for giving me the opportunity to lead this project.

I would finally like to personally thank the current team for the great contribution each one of them provided to the project. To make projects this size and complexity work, you need good people, not "just" good technical experts. And I have been fortunate to have that.

EXECUTIVE SUMMARY

This final technical report summarizes the work performed through the Systems Engineering Research Center (SERC) Research Task (RT 203- Meshing Capability and Threat-based Science and Technology (S&T) Resource Allocation) from April 2018 to June 2019. This research project focused on providing a computational model to support the planning cycle that will inject relevant threat-based intelligence and operational scenarios into the more traditional capabilities-based model. This approach will better inform the technical communities charged with developing future weapons systems and have been piloted in late 2016 at the U.S. CCDCAC in the armament-systems domain.

This research utilizes a data or text-driven approach initially focused on a proxy-domain to source the data. The proxy domain selected for this project is “Artificial Intelligence (AI)/ Machine Learning (ML) in a connected environment”. Threats will be replaced with applications in the proxy universe. The point of view is the one of a provider of AI/ML solutions. The team is working on monitoring applications and the related technologies that makes these applications feasible or prevent unwanted uses of these applications by adversaries. This proxy-domain was a vital choice to have both U.S. and foreign nationals working on the project. At the end of the project, the Principal Investigator (PI) - as a cleared U.S. person – will adapt the system to work on the actual target domain for the Sponsor.

The systems were developed as agile, iterative prototypes with modular components. Most of the components were developed separately for better reusability. The first integrated proof of concept was ready in November 2018 and the team continued to refine and improve the systems. The final results of the prototype are reported in Section 6.

1 Introduction

The purpose of this research is to provide a computational model to support the planning cycle that will inject relevant threat-based intelligence and operational scenarios into the more traditional capabilities-based model. This approach will better inform the technical communities charged with developing future weapons systems and has been piloted in late 2016 at the U.S. CCDCAC in the armament-systems domain.

Using a data or text-driven approach, this research focused on a proxy-domain “Artificial Intelligence (AI)/ Machine Learning (ML) in a connected environment”. In specific, the private security industry marketplace was used as an example for this project. In the U.S., the private security industry is chosen because it is a technology-driven marketplace that has close semantic proximity to the needs at the U.S. CCDCAC. According to the Security Industry Association, cybersecurity impact on physical security, internet of things and the big data effect, cloud computing, workforce development, and AI are the top 5 forecasted security megatrends in 2019.

In this research, two core systems, Technology Monitoring and Risk Panel systems, were designed and developed as agile, iterative prototypes with modular components (refer to Section 5). The modular components are vital building blocks that were designed to be used as components for the overall system and the data collection process for the proxy domain (refer to Section 4). Most of the components are developed separately for better reusability.

1.1 Objectives

The objectives of the computational model are as follows:

- Replicate the aforementioned process developed at the U.S. CCDCAC in 2016 to validate this notional computational architecture
- Enhance the visualization and analytic capability to allow rapid, high fidelity decision making
- Introduce additional parameters and variables to refine the decision-making framework further. Real-world scenarios will be modeled to project evolving threats, doctrine, partner force interoperability, and other operational environmental conditions (political, military, socio-economic, information, infrastructure, physical environment)
- Deliver the results with an agile approach, developing prototypes/proofs of concepts with increasing capabilities, using a partially automatic learning approach.

1.2 Scope

The research project was developed through the following tasks:

***<u>Task 1.1 Systems Engineering Tool</u>*:**

Framework and Tools for Autonomous Systems Architecture and concept of operations (CONOPS) Synthesis

Sub-Task 1.1-1 - A framework to Enable System Composition: The subtask focuses on the design and implementation of abstractions and algorithms that define:

- **An intuitive scenario specification language.** An intuitive specification language allows the definition of scenarios. It contains a catalog of fundamental logical structures and relationships at and across multiple protocol layers.
- **A capability warehouse.** The capability warehouse provides the building blocks for system composition. The capabilities are defined in terms of specifications of sub-systems, devices and protocols. The warehouse may also host several emulation engines to help evaluate system dynamics in a state-aware context.
- **An automated configuration synthesis and repair engine.** This step identifies if a configuration is possible and that it can meet the requirements as defined by the scenario. Furthermore, configuration variables that are non-compliant with requirements are separated, and minimum-cost configuration changes will be identified. If these are unsolvable, a root-cause of insolvability is identified in the form of a typically small part of the requirement set that is itself unsolvable. Furthermore, it finds a maximum weighted solvable subset of requirements, where possible. The key to achieving this is the definition of the system logic and automatic search for solutions using solvers based on Satisfiability (SAT) and Satisfiability Modulo Theory (SMT).
- **Autonomous Verification and generation of CONOPS.** The correctness of the configurations and the systems' ability to generate the defined scenario is verified in this step. Once verified, CONOPS are generated automatically describing how the proposed system can execute the defined scenario.

Sub-Task 1.1-2 - Representation Methods: As the ability to represent scenarios define the success and automatically deduce system requirements that can be compared with capabilities in the warehouse, a rich descriptive specification language is a critical factor. Scenario and specification description languages like (SDL, Q, STSIM/DRIVE) will be evaluated for an easy and preferably graphical description of a scenario. Any language specialization needed to describe scenarios that are applicable for armament system use - such as close combat, area denial and deployment of precision-guided munitions – will be considered.

Subtask 1.1-3 - Formulation of Problem Constraint Set and Solution: In the search sub-task, the requirements as defined by the scenario are matched as constraints on the specifications in the capability warehouse. The result will be the formation of the problem constraint set. In order to translate the problem constraint, one must encode finite domain variables onto propositional variables and define the constraints among the variables onto an effective representation for the SAT. The formula evaluating the constraint is represented in Conjunctive Normal form (CNF), as finite conjunction of clauses, $C_1 \cap C_1 \dots \cap C_m$ defined on a finite set of Boolean variables, $\{x_1, x_2, \dots, x_n\}$ with true (1) or false (0) assignments. A clause C is a finite disjunction of literals, l_i, which are either the Boolean variables, x_i or their negation, $\neg x_i$. Therefore, the clause is satisfied by a true assignment to one of its literals and formula is satisfiable if there is a true assignment to all the clauses. This translation is typically done by one of the two encoding methods - the sparse and the order encodings. The term "sparse" or the direct encoding is the most straightforward way to transform a constraint problem into an SAT problem. A variable V with domain {1, ..., n} is translated into the sparse encoding of n propositional variables, d_i^v $_i$, 1 ≤ i ≤ n, and the assignment V = i is modeled by assigning $d^v{}_i$ to true and all the other propositional variables to false. The sparse encoding requires that exactly one d_i^v variable is assigned to true. Such constraint is achieved by means of a single at-least-one (ALO) clause, $\{d_1^v \cup d_2^v \ldots \cup d_n^v\}$, and a set of the at-most-one (AMO) clauses. Using the AMO clauses retains the equivalence between SAT and Constraint problem solutions. The order encoding represents a constraint variable, V with domain {1, ..., n} by a vector of n − 1 Boolean variables,]. In order to specify V = I, the first i − 1 variables are assigned to true (1) and the remaining variables to false. The encoding is specified by a set of Boolean clauses as follows: $\cap_{i=1}^{n-2} \neg(\neg O_i^v \cap O_{i=1}^v) \equiv \cap_{i=1}^{n-2} (O_i^v \cap \neg O_{i=1}^v)$. The advantage of this encoding is in the representation of interval domains and the propagation of their bounds. Furthermore, constraints can be represented as *conflict clauses* signifying disallowed variable assignments and *support clauses* that specify allowed assignments. The modern Satisfiability Modulo Theories solvers can solve a million dependencies in a million variables in seconds. Thus, SMT solvers provide an expressive and efficient logic for specifying and solving networks of constraints and are much superior to alternatives such as Binary Decision Diagrams, (Constraint) Logic Programming and full first-order logic.

Subtask 1.1-4 - Software tool design, implementation and testing: A software tool is being prototyped that incorporates the algorithms and solution methodologies. We are using rapid tool prototyping processes with low-code options to evaluate the methodology and optimize it quickly. The software tool is designed for use by various stakeholders. Methods to provide varied and secure access to data views, analysis results to users separated by clear access control policies will be implemented.

A novel and innovative feature of the representation schemes and the solution methodology is that access to data and solutions itself can be explicitly written with access control policies defined as the constraints in the solution process. For example, in order to prevent solution composition and presentation to users without necessary privileges, an explicit constraint requiring user privileges can be added to the problem's constraint set. Furthermore, privileges can be added on the use of a capability during the system composition to allow solution computation only for users with appropriate rights.

<u>Task 1.2 System Development:</u>

Acquiring the components and implementing the system.

Our approach is based on the CRISP-DM (Cross Industry Standard for Data Mining), modified for the specific case and expanded to accommodate the decision-support components.

Research Question: Given the available data, what are the most appropriate combinations of metrics, models and visualizations to create a valuable data-driven decision support system to be actively used by the stockholders?

Sub-Task 1.2-1 - Determining a decision framework to be implemented into the system

Considering the proposed system will be driven by both the data streams it is receiving and patterns of behavior, it is essential to get as much information as possible about the processes currently in place. The sub-task focuses on baselining the current decision process and will consist of:

- Extracting and representing the standard armament-systems evolution process. Using high level business process representations such as IDEF-0/SADT, we interview stakeholders to extract the patterns of behavior in the current supply chain for armaments. This will include operational scenarios, and tactics, techniques, and procedures (TTP).
- Extracting and representing the impact of threats in the armament-systems evolution process. We interview stakeholders to evaluate the impact of threats on the armament supply chain. This includes both the level of threats and descriptions of the potential armaments to be used to counterbalance it.
- Extracting and representing risk factors and other external constraints. Besides threats, other constraints to be considered include supply chain delays factors (e.g.: delivery/production capacity) and budget. We interview stakeholders to evaluate the impact of those elements on the armament supply chain.

Sub-Task 1.2-2 - Data collection and understanding.

While 1.2-1 focused on processes, this sub-task is centered on data and consists of:

- Extracting and evaluating the data related to existing and planned armaments. Data related to current inventory, short/medium planned supply and long term planned supply. Once extracted, data is evaluated in terms of their ability to be used as elements to create the decision support system.
- Extracting and evaluating the data related to threats. Data related to how threats are currently collected. This includes threat level and the correlation with armament systems. Once extracted, the data is evaluated in terms of the ability to be used as elements to create the decision support system.
- Extracting and evaluating the data related to past scenarios. In order to provide the system with predictive capabilities, information about past behaviors are required. Once extracted, the data will be evaluated in terms of the ability to be used as elements to create the decision support system

Sub-Task 1.2-3 - Data preparation

- Syntactic data preparation. This phase consists of performing all the steps required to clean and link various data sets. It will focus on standard data cleaning tasks, such as outliers' evaluation, missing values, normalization, etc.
- "Semantic" data preparation. This phase ensures data is semantically consistent. This is particularly important when – like in this case – data is derived from different sources.

Sub-Task 1.2-4 - Modeling

- Metrics definition. Metrics are derived from a combination of synthetic representation of the key performance parameters determined from the interviews in 1.2.1 and new risk indicators.
- Behavior and predictive models. Behavioral models are based on agent-based simulation; predictive models use supervised learning. Both models are being developed based on past scenarios, collected in Sub-Task 1.2-2.

Sub-Task 1.2-5 - Visualization

- Visualization metaphor selection. The most appropriate combination of visualizations is selected and implemented. This requires some interaction with the stakeholders to ensure selected visualizations are sufficiently representative.
- Visualization integration. Various visualizations are integrated in a unified dashboard, which contains wide post-processing capabilities to allow for user interfacing. Voice interaction ("chatbots") may be added.

Sub-Task 1.2-6 - Evaluation

- Predictive portion. Using part of the data from past scenarios, the system will be tested to determine the quality of the predictions. This may require iterative modeling – Sub-Task 1.2-4 - for refinement.
- Overall system. After conducting in-lab tests, systems will be released to a reduced number of stakeholders for use and evaluation. The system will be adjusted based on user feedback.

Sub-Task 1.2-7 - Deployment

- General release. The system will be released to all stakeholders.
- Future steps definition. Based on the feedback received at this stage, future steps will be defined. Those may include expanding the coverage in terms of data, adding new functionalities or porting the current system to different but similar environments.

Each of the task items will be further decomposed into lower-level work items, which will be prioritized and placed on a work queue. These will be allocated to the available researchers and progress will be monitored on a weekly basis. Based on these results, items may be added, removed or reprioritized. The Sponsor was invited to monthly general meetings to understand the progress that is being made and to participate in the decision-making process. Each of the new features and capabilities will be made available for demonstration and evaluation purposes.

1.3 About This Document

This final technical report is organized with the major sections as follows:

- Section 1: Introduction – This section provides an overview of the research project, context, objectives, scope, and organization of this report.
- Section 2: Background – This section summarizes the project, approach, and deliverables.

- Section 3: Research Methodology – This section provides a detailed description of the research methodology used in this project.
- Section 4: Components – This section elaborates on the components developed using a data or text driven approach for the core systems.
- Section 5: Core Systems – This section discusses the Technology Monitoring and Risk Panel Systems developed as agile prototypes in this research.
- Section 6: Results – This section provides the results from the research conducted in this project.
- Section 7: Conclusions – This section summarizes the research outcomes.
- Section 8: Future Research Directions – This section highlights the future research path and continuation of this project.

2 BACKGROUND

This research project is focused on providing a computational model to support the planning cycle that will inject relevant threat-based intelligence and operational scenarios into the more traditional capabilities-based model. This approach will better inform the technical communities charged with developing future weapons systems and has been piloted in late 2016 at the U.S. Combat Capabilities Development Command Armaments Center (CCDCAC) in the armament-systems domain.

This research utilizes a data and text-driven approach initially focused on a proxy-domain to source the data. The proxy domain selected for this project is “AI / ML in a connected environment.” Threats are to be replaced with applications in the proxy universe. The point of view is one of a private security company that wants to use those technologies/applications to provide better services and gain market shares. The team is working on monitoring applications and the related technologies that make these applications feasible or prevent unwanted uses of these applications by adversaries. This proxy-domain was a vital choice to have both U.S. and foreign nationals working on the project. At the end of the project, the PI - as a cleared U.S. person – will adapt the system to work on the actual target domain for the Sponsor.

The two core systems, Technology Monitoring and Risk Panel systems, were developed as agile growing prototypes with modular components with increasing capability (refer to Section 5). The modular components were created to be used as components for the overall system and the crucial data collection process for the proxy domain (refer to Section 4). The quantitative data were gathered using a partially automated learning method and were analyzed using Natural Language Processing (NLP) techniques. Most of the components are developed separately for better reusability. Room theory which was implemented in this research project is an approach that combines theoretical frameworks from cognitive science and AI with new advances in NLP (refer to Section 3).

The key final deliverable for Phase 1 of this project is a working integrated prototype with interactive visualizations and analytical tools for what-if analyses scenarios to support the decision-making process.

3 RESEARCH METHODOLOGY

The research methodology in this project uses customized approach to allow research iterations that continuously provide value to the sponsor. Rapid concepting was performed to assess utility and based on the results obtained at various stages of the project. Using a data or text-driven approach, this research project focused on "AI / ML in a connected environment" as a proxy domain to source the data.

The data collected are texts related to a specific domain. The data were selected to be easily associated – for content and complexity – to the final target domain. A combination of traditional Natural Language Processing (NLP) (mainly for pre-processing) and embeddings were used. Embeddings are feature vectors for conversational elements in the text, calculated via Python libraries based on neural networks, such as Word2Vec. Specific metrics are extracted for risk evaluation and for visualization from the embeddings. This process is shown in Figure 1.

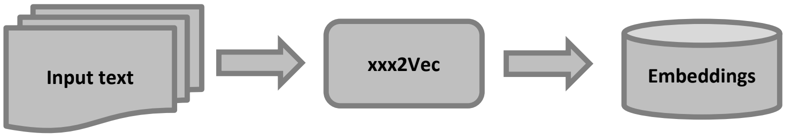


Figure 1: Data or Text-Driven Approach for Processing

An agile development methodology was used throughout the development of the components and systems in this project. Figure 2 shows a high-level overview of the agile development methodology used in this research.

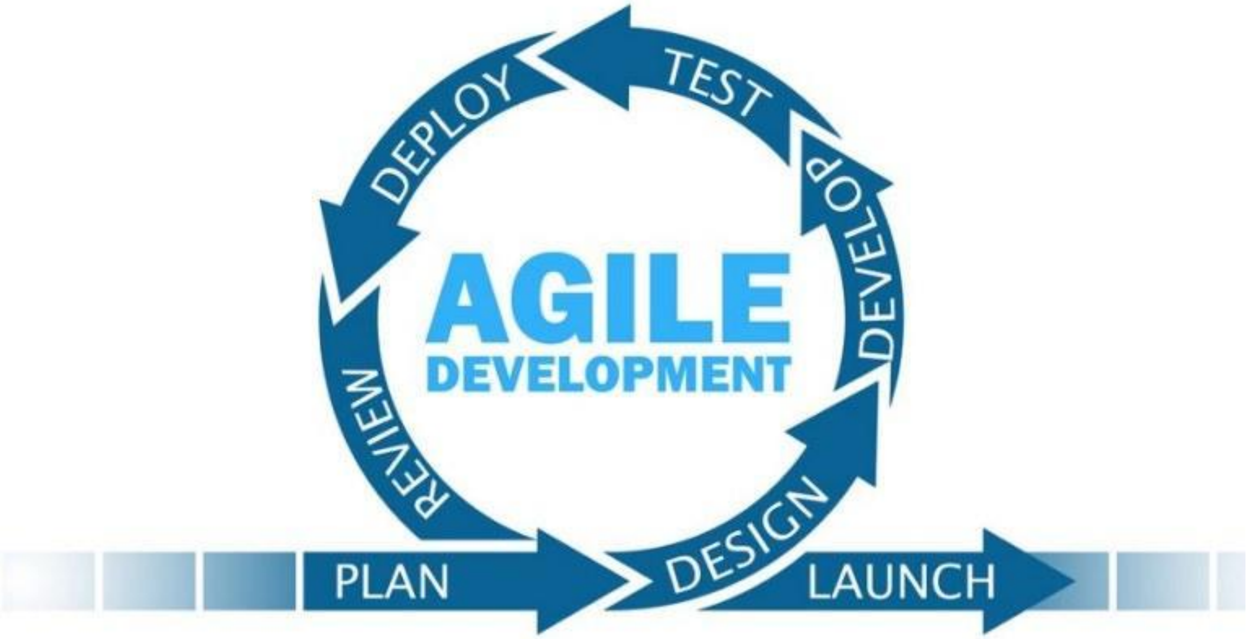


Figure 2: High-Level Overview of the Agile Development Methodology (Hadar, 2019)

This iterative process includes regular engagements with the Sponsor to gather continuous feedback on the development and progress of the project. The first phase includes planning a data and text-driven approach focused on a proxy-domain to source the data. The proxy domain selected for this project is “AI / ML in a connected environment”. Next, the design of this proxy domain takes on the perspective of a private security company that wants to use those technologies/applications in order to provide better services and gain market shares.

The researchers worked on monitoring applications and related technologies that makes these applications feasible or prevent unwanted uses of these applications by adversaries. In the third phase, the systems are being developed as agile growing prototypes with modular components. The code was tested and validated before deployment. Feedback was gathered and reviewed through regular general meetings. Once the product is ready for deployment, it will be launched at the sponsors’ organization site.

Due to the high computational needs for the text processing and the creation of the embeddings in particular, this project acquired two parallel processing units (GPUs) to speed up the processing. The GPUs can be connected to any computer and was selected based on the low cost and preference of insourcing the processing compared to a cloud-based solution. This will mitigate and potentially reduce risks for cybersecurity issues. The dedicated GPUs ran the software developed to create the embeddings from the corpora, reducing the time to obtain the embeddings from days to hours.

3.1 Relationship to Core Competencies and the SERC Research Strategy

The research uses standard methodologies for data mining adapted and detailed for the specific need. In particular our approach is be based on the CRISP-DM (Cross Industry Standard for Data Mining), modified for the specific case and expanded to accommodate the decision-support components.

This research project leverages on the core SERC competencies to deliver its results. In particular, it fits into 2 of the key SERC areas:

1. *Enterprises and System of Systems:*
   - Enterprise Modeling - Create, validate, and transition methods, processes and tools
   - Methods, Processes and Tools (MPTs) to model the socio-technical aspects of complex systems of system and enterprise systems, including developing and populating a framework to integrate models created using diverse methods and tools
   - System of Systems Modeling and Analysis - Create, validate, and transition MPTs for analyzing and evolving systems of systems and provide support for their technical assessment, including through a “workbench” of analytic tools
2. *Systems Engineering and Systems Management Transformation:*
   - Quantitative Risk - Create, validate, and transition methods, processes, and tools to improve risk identification, analysis tracking and management in acquisition and sustainment programs
   - Interactive Model-Centric Systems Engineering (IMCSE) - Create, validate, and transition methods, processes, and tools to rapidly model the critical aspects of systems, especially those that facilitate collaborative system development

3.2 Research Needs

The research needed for this project was developed along the following lines:

• Extract key indicators from the traditional capabilities-based streams of information.

• Create a predictive model from selected sources of threat-based intelligence and operational scenarios to evaluate their impact on the planning cycle. This is based on a combination of text mining, risk evaluation and data science.

• Create an interactive visualization/presentation layer with scenario analysis capabilities, integrating the metrics from the above lines. Visualizations and underlying metrics being developed leverage on the current planning Modus Operandi. This layer is implemented in prototypes with growing capabilities.

• Provide a layer of real-time/near real-time capabilities to the whole process, from modeling/indicators extraction to the dashboard/visualizations.

• Create a system that is able to detect from the way users interact with the visualization/presentation model/prototype the missing or wrongly executed requirements. This provides – once the system will reach the proper maturity level – the input for the following releases of the model/prototype. Using this approach, the model/prototype is released with a growing level of definition and capabilities. This system is based on a machine learning layer working on the log file. In possible future phases of the project, this system will increase its capabilities to influence the following releases of the model/prototype, eventually with some self-implementation of the rules extracted from the analysis of the user- model/prototype interaction.

3.3 Room Theory

We encountered two major issues while addressing the core challenges of this project, both poorly covered by existing work:

- Extracting metrics from text
- Taking into consideration the context/subjective point of view in analyzing the text.

To address the points above, we developed the "room theory", represented by Figure 3.

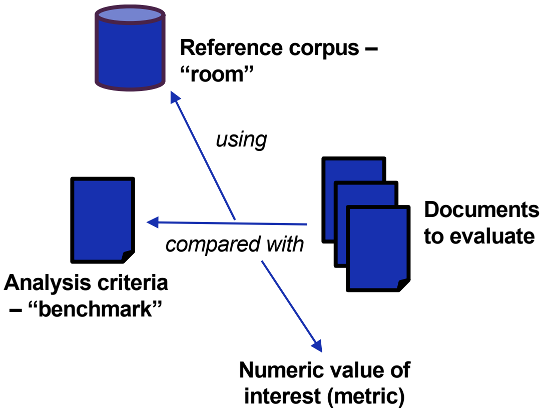

Reference corpus – "room"
using
Documents to evaluate
compared with
Analysis criteria – "benchmark"
Numeric value of interest (metric)

Figure 3: The "room theory"

"Room Theory" is a computational representation of subjective knowledge bases on the non-computational schema theory, was implemented throughout this project. Room theory is an approach that combines theoretical frameworks from cognitive science and AI with new advances in NLP.

We leveraged on a knowledge representation developed by Marvin Minsky (1974) in one of his studies in AI and Cognitive Science - “A Framework for Representing Knowledge” – where he introduced the idea of “frames”. According to his work, “a frame is a data-structure for representing a stereotyped situation like being in a certain kind of living room”.
We used this concept to recreate “rooms” representing the semantic context for a specific social/cultural entity or context.

The first step to have a computational model in this case, it to create a numerical representation of text ("embeddings"). Word2Vec was used to generate a list of vectors from words/text elements. This vectorization is the enabling technology for the room theory.

Then, using this methodology, we

- Create a “room” by generating embeddings from a domain specific corpus, to represent the point of view or the specific context for the analysis
- Define a word set to be used as criteria for the analysis. This is going to be a benchmark for the comparison
- Compare words/chunks in the incoming document (the one to be evaluated) with the words/chunks in the benchmark, using the “room” to calculate their distance
- Adding and normalize the collected similarity values for each word/chunk in the benchmark to have an evaluation of the incoming document based on the elements in the benchmark, according to the point of view represented by the “room”.

The numerical values collected through the similarity values create the input for the models used by the two systems we developed.

4 Components

Part of the whole team has been dedicated to developing components to be used by the systems. The team working on the components ("NLPlab") developed components which are essential building blocks that were designed to be used as components for the overall system composition and the data collection process for the proxy domain. The components were developed to ensure that the technology monitoring and Risk Panel systems are working well on the server and delivering new and regularly updated components for generalized use.

The building blocks were designed to be standalone and usable functions that perform atomic tasks related to NLP. Those blocks were used to build the systems. Using this approach, the NLPlab team increased the code usability, maximize the development time, and optimize the performance of the systems with highly efficient functions/blocks. Besides ensuring the availability of the best tools for the task, the framework enables updates of the functions without having to update the entire system altogether. That is a valuable characteristic for the development and the team assignment that can easily be overlooked.

To further describe the development of the components, this section is divided into the following subsections:

- Development Environment
    - Agile Development Review
    - Tools for Development
    - Building Blocks
- Natural Language Processing (NLP)
    - Natural Language Processing Review
    - NLP Processes
    - Word2vec
- Framework
    - Framework Design
    - Docker for deployment
    - Databases
    - System Deployment
    - System Access

The development environment subsection discusses how the development was set internally for an optimized result. That particular setup enables the whole research project to go from idea-to-code faster since the environment was shared, and always up to date with all the building blocks available at the time. The review of NLP is necessary to discuss a few of the improvements and explorations that were made during the project. Lastly, the framework will be presented, explained, and discussed to lead to a better understanding of the project results and its back-end functionality.

4.1 Development Environment

This subsection will explore the details of how the development environment was setup.

A pipeline was built for NLP so it would be a more straightforward process going from the documents to the analysis process. All the in-between parts will be taken care of with a few configurations of a function. The goal is to have that inside the backend of our systems and provide this additional functionality to the end users.

Figure 4 provides an overview of the components' architecture. The team has implemented the name entity functions and the databases infrastructure. Currently, the data collector is live and collecting and storing data 24/7 on the databases. The team is in the process of using the implemented libraries on the new techniques' developments. The computational Room Theory has been implemented and is currently being tested. The team has also developed the document classifier.

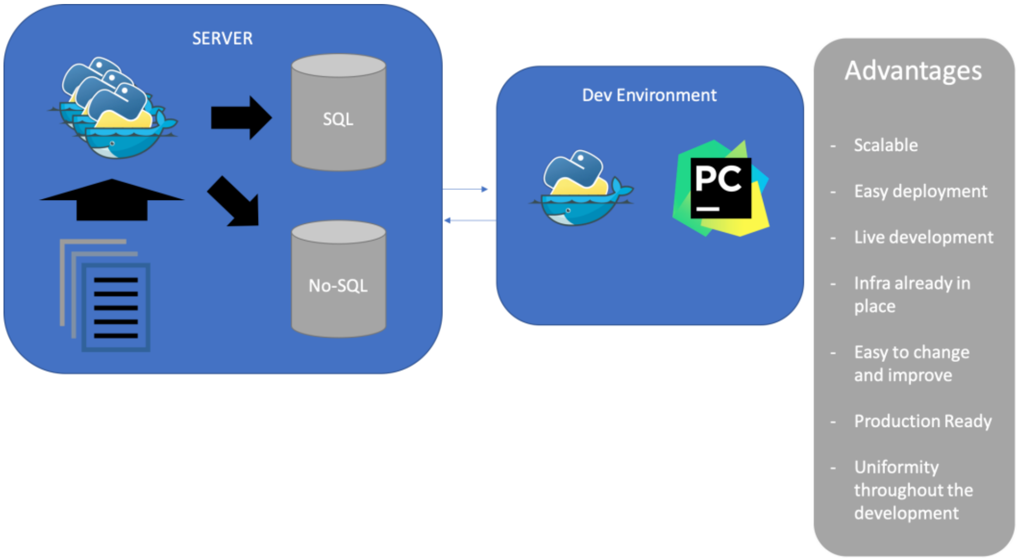


Figure 4: Components Architecture

The team started optimizing the components library by rewriting most of the code with fewer dependencies, which provides a faster runtime and execution. The components cover about 61 tasks, with over 11,900 lines of Python code. Table 1 shows the breakdown of the lines of code for this project are as follows:

Table 1: Breakdown of Lines of Code in this Research Project

Item	**Lines of Code**
Server	1,688
Components	4,254
Database Tools	387
Technology Monitoring System	2,701
Risk Panel System	2,950
TOTAL	**11,980**

4.1.1 Agile Development Review

Project management is a science in itself, and this research project does not seek to develop or deliver any insight on what the best project management methodology and/or line of thought is. However, this report explains why the choice of tools and methodologies to solve the main research question. There are two different approaches for creating software: the first one is the top-down approach, and the second one is the bottom-up approach.

As a norm, for a top-down approach project management, the client states its necessity, draw a solution together with the project manager. That solution is then cascaded to the team, which will receive a task to be completed, with an expected time limit to do so. As a result of this architecture, the team does not have an opening to feedback ideas to the project manager neither to the client. More often than necessary, the result is a solution proposition that might not be feasible in the allocated time frame with the allocated resources.

Often it is necessary to make changes during the project that would affect the underlying tasks, which is another difficulty that arises from a top-down approach. Research projects commonly have changes by nature. It is impossible to be investigating a new knowledge field and already knowing the final and desired result. Moreover, the research will develop itself with incremental steps, and the next step is defined by results that were not yet available.

To address these issues, the PI decided to implement a bottom-up approach. Instead of creating the plan completely, the goal is defined, however, the path to achieve it is not fully pre-defined, and incremental steps create it. The bottom-up approach often starts with a meeting with the client and the project manager, to define the scope of the problem and draw a solution. That solution is then replicated to the team, which is open and welcomed to make suggestions and comments. The first few tasks are determined, and the teams are divided. Another essential characteristic of the bottom-up approach is the data-driven aspect of it. Since the solution is drawn from the goal towards the start of the problem, it is necessary to determine the steps based on actual data. The fact that the bottom-up approach bases itself on actual data and not only a theoretical framework is an important differentiation of the top-down approach.

Periodically, the project manager meets with the client to review what has been done to the period, and what is the guidelines for the next steps. That not only enables the client to make changes in the direction in which the project is going, it is also welcomed by the team members to gather feedback from their deliverables up to that point in the project timeline. With periodic reviews of the strategic and tactic directions on the project, the success rate of delivering the final product increases drastically.

This research project is an applied research project, with deliverables that are not only reports but - more relevantly - a useful, agile software prototype. That increases the complexity and introduces a high standard to be met in terms of code infrastructure and architecture. The increase in complexity along with the research nature of the project led this project to implement an agile development methodology.

The agile development methodology is similar in some ways to the bottom-up approach. However, it adds methods to control the requirements and the project production better. Every week there is a short meeting with the project members to address the tasks that were completed, what is backlogged, and what are the next steps in the pipeline. That provides the team members with an opportunity to share experiences and new learning elements, which flattens the learning curve for the less experienced members of the team.

Another characteristic incorporated by the agile development methodology is the short cycles of development. The short-term goals were set every month, and the team would work to deliver these short-term goals. A potential problem that may arise with this methodology is that team members could lose sight of the long-term goals and client requirements. However, meetings were held periodically to ensure that all the team members were on the same page.
The use of this methodology defined the team division on the project. The project was divided into three major sections:

- Components (known as the NLPlab)
- Risk Panel
- Technology Monitoring

Every team was instructed to develop its tasks in a standalone format but sharing the same development environment. The NLPlab team developed the software with a building block design, which means that every function is reusable and atomic enough to be easily replaced. Since the system was not being built as a whole, it was possible to create incremental improvement steps throughout the life cycle of the project. For example, a component would be developed and made available to all the team members, even though it was not the best possible solution for the specific task, keeping in mind that the component could later be improved and in a new release, the system automatically incorporates the new and improved changes.

For the communication plan, the teams had bi-weekly meetings to update the task board. In the meetings, each team member shares her/his progress during that period, and then the team would decide what to do next on the backlogged or incoming new tasks. Highly educated professionals composed the team; therefore, the tasks were higher level, more focused on the problem to be solved rather than the task itself. The team members then would be responsible for researching already implemented solutions, research our components base to ensure that they would not have double developments, and then decide what the best ways to solve given problems are. This is aligned with the agile development methodology and the bottom-up approach, that guaranteed that the project had room for improvement and changes along the way to achieve the goal, even though the path to it was fully defined by the start of the project.

4.1.2 TOOLS FOR DEVELOPMENT

In the project management space, it is crucial to have a communication plan, project management tools and documentation.

Slack was the communication platform of choice. Slack is a platform for collaboration that has various features to share and discuss the software development between teams, which drastically reduced the number of emails sent, and enabled team members to be available for questions and continuous interactions.

This research project was completely developed in Python 3.6, the programming language of choice. Python is a widely used programming language due to its easiness-to-use and readability, which enables developers to go from complex ideas to testing faster than other programming languages. It has a vast variety of packages for data science tools as well largely documented Q&A on internet platforms such as Stackoverflow.

This is a complex research project that required the implementation of complex concepts and the use of a large number of packages. With this complexity, it was also necessary to create a shared development environment that could handle different operational systems, guarantee the uniformity of package versions across all team members, and finally guarantee that code base was protected in a need-to-know-basis for security purposes. Not all team members had open access to the code base; therefore the development environment needed to be able to make the components available without exposing its skeletons.

To address this issue, the project manager decided to use Docker as the main tool for development, which is a self-standing virtual environment for all the components and systems. This provides better segregation of the systems and a more efficient overall infrastructure. Docker performs operating-system-level virtualization, also known as "containerization". This will also add a layer of security to the code, allowing team members to only see the code they are developing while using the components that was developed. Also, because Docker is running on the SERC server, the developments will be more secure than running on each team member's computers.

Docker is an operational-system level virtualization, also known as hyper virtualization. Docker operates with images and containers. Image is a snapshot of a computer and contains all the configurations and files. However, it cannot process any calculations. An image can be transformed into a container, that can perform calculations. Since it is a hyper virtual machine, Docker can run a Linux operational system on top of any operational systems, for example, a container of Ubuntu can be created on a Windows operational system without having to install it directly. That addresses the compatibility issues across different operational systems, which was an actual problem for this project since team members had MacOS, Linux, and Windows computers.

The NLPlab team then created an image that is kept on a protected, online repository that contains all the packages necessary to run the codes, and the latest components already installed on it. With that, a team member needs to run a simple update command that will make available all the latest developments to support the system developments. That minimizes the team members workload, because they were able to reuse the atomic functions on the components as building blocks to their systems.

This development environment strategy also addresses the security issue, since not all team members needed access to the source code, however, they had access to the use of the function itself. Another important advantage of the implemented development environment is that the time-to-deployment has reduced dramatically. Since all the development was made in the same shared environment, it was possible to replicate the development environment on the production server that the system is up and running.

For the source code base, it was used the GitHub repositories. The GitHub is the largest and most well-known source code repository that leverages of the git system (Github, 2018). The git system is a versioning tool that enables developers to make incremental changes in a collaborative environment, guaranteeing the uniformity and reliability of the source code base. It keeps track of all changes made on the project and makes the source code available to all involved developers. To ensure the security compliance of the whole team, each's access was managed to provide access on a need-to-know basis for the team members over the following five different repositories:

- Components
- Server
- Risk Panel
- Technology Monitoring
- Database Tools

The Components repository was designed to hold the source code of the building blocks that support the construction of the system. It is composed of atomic functions, which means each function performs small tasks, a few more complex building blocks are also provided, and are built using the atomic functions. That creates a flexible base that can be assembled as needed and perform complex tasks without losing the sight of simplicity with the atomic functions.

The Server repository was designed to store the source code of the pipelines and tasks involved in running the server. It leverages from atomic functions from the components as well and provides efficient code to run in the server side periodically. It handles the periodical data collection tasks, manages the databases, and ensures that the complete system is healthy, running, and available to the rest of the team. It also lays a foundation for the creation of the natural language pipeline, which enables the systems to combine the building blocks with the database seamlessly.

The Risk Panel repository holds the source code for the Risk Panel (planning support system). The Risk Panel can be separated into two parts, which are the back-end and the visualization. These are completely different approaches requiring different skill sets. The back-end portion of the Risk Panel leverages the components created, and the data collection is realized by the server side. It also implements novelties to address some challenges that arose during one of the improvement steps. On the visualization portion, the latest methodologies were used to display complex information and enable the users to make data-driven decisions faster and more efficient. The system runs on the server and serves the visualization through the web, which makes the software easy to use and provides the users with desirable flexibility to have the solution accessible, yet secured, to support its decision-making.

The Technology Monitoring repository stores the source code performing the tasks related to the Technology Monitoring system. As the Risk Panel, the technology monitoring system can be separated into two different portions, the back-end and the visualization. The back-end runs in the server and leverages from the components functions and interacts with the database to explore the data collection realized by the server side. The technology monitoring visualization portion exploits the latest developments on the visualization science to transform complex data into actionable insights.

The Database Tools repository contains all the necessary source code base that the systems use to interact with the database. It was designed to be installed and self-contained into the Docker image/container. The database tools provide the team members with an easy-to-use interface that does not require previous knowledge of SQL language to perform queries and searches in the database. The code is reusable and can be used with any flavor of SQL Databases.

For the documentation, all available functions, as well as complex parts of the systems are explained and exemplified in a web-based documentation that is easy to update and can be accessed anywhere, as long as the user has the right of access the information.

4.1.3 Building Blocks

The building blocks approach, also known as modular programming, is a software design methodology that emphasizes separating the functionality of a program into independent, interchangeable modules, such that each contains everything necessary to execute only one aspect of the desired functionality.

A module interface expresses the elements that are provided and required by the module. The elements defined in the interface are detectable by other modules. The implementation contains the working code that corresponds to the elements declared in the interface. Modular programming is closely related to structured programming and object-oriented programming, all having the same goal of facilitating construction of large software programs and systems by decomposition into smaller pieces.

By choosing the modular / building blocks approach to develop the project, it was possible to start the developments of parts of the system without having a clear definition of the problem. Given the use of bottom-up approach, the use of textual data was defined since the beginning, even though the complete path to the solution was not yet in place. Then it was possible to start the development of the components to operate with textual data in the initial moments of the project, bringing a workload that usually would be allocated later in the project, to the beginning.

The building blocks approach, together with the agile methodology, enables the component team to create usable functions and the systems team to build prototypes since the early stages of the project. Moreover, since the systems have been built with blocks that could be replaced, as long as the inputs and outputs were cohesive, the systems could evolve in the project and always be in the prototype phase. Since the early stages, it was possible for the client to see a functional prototype, with limited results, due to the constituent's parts that were still under development before reaching the state-of-the-art level.

In the process of guaranteeing the availability of the best functions for the systems to use, the NLPlab team extensively researched natural language processing elements. The components team developed functions that leveraged implemented packages and customized solutions. Most of the implemented packages are previously trained models or models that require supervision, and one of the most complex challenges that this project address is that language it is hard to be supervised, because it can be varied by context, either from the person realizing the supervision and the writer.

4.2 Natural Language Processing

Natural Language Processing (NLP) is a research area specialized in analyzing natural language data. In this digital age, the amount of data produced is growing at an unprecedented rate. Most of this data is produced with written natural language: scientific articles, news articles, Google searches, text messages, emails, social media posts are just a few examples. The goal of NLP is to make machines understand natural language in order to extract valuable information in an automated way.

NLP is a challenging field of research for many reasons. Each language has its own vocabulary, grammar and syntax, and different individuals may use different forms of the same language when writing a text. A word may have different meanings depending on the context. It is crucial to understand and represent the context in which a word is being used to capture its semantic meaning correctly. Further, words and their meanings have a strong nonlinear relationship; this makes the effort of capturing information from text a complex task, that is also why machine learning techniques are key enablers for NLP.

4.2.1 Natural Language Processing (NLP) Review

NLP systems can perform many tasks serving a wide range of applications. Here, we report the most important NLP tasks studied in literature. The first important task is Tokenization (sometimes generalized as text chunking or n-gramming), which refers to the action of separating the text into simpler units. This task may present issues with languages in which words are not delimited by spaces or punctuation marks (e.g. Japanese language).

The way in which text is split strongly affect the results of NLP systems, which is the reason why the chunking step is crucial for a correct analysis. Another aspect to consider is that most current methods available to chunk a text are supervised or semi-supervised methods. This means that they are based on a training corpus that is manually annotated, hence, they can be subjective.
Opinion Analysis is branch that studies the subjectivity of a corpus. Not all textual contents are objective. Many corpora express a subjective point of view, which can be positive-negative, neutral or biased towards a particular topic.

Part-of-speech tagging is the task of labeling each word/token in the corpus with the respective part-of-speech. Part-of-speech are noun, verbs, adjective, adverbs, prepositions. Named Entity Recognition aims at the extraction of entities from a given text. This is particularly useful when there is a need for automatic recognition of specific entities in a text like for example geo-political entities, or name of companies. This can be achieved with a named entity recognition module.
Topic modeling is a field that studies how to extract the argument of discussion in a document. The output of these model is a list of words related to the topic of discussion. The main issue is related to assigning a unique objective label to this list of words.

Another branch of NLP is focused on how to generate text. Language generation is generally performed by training a Recurrent Neural Network (specifically a Long Short-Term Memory in most of cases) to create a model able to predict new text. A related task is Question Answering, which studies how to make a machine interact with a user using textual content, just as in the case of chatbots.

Figure 5 shows a classification of some NLP tasks in different categories.

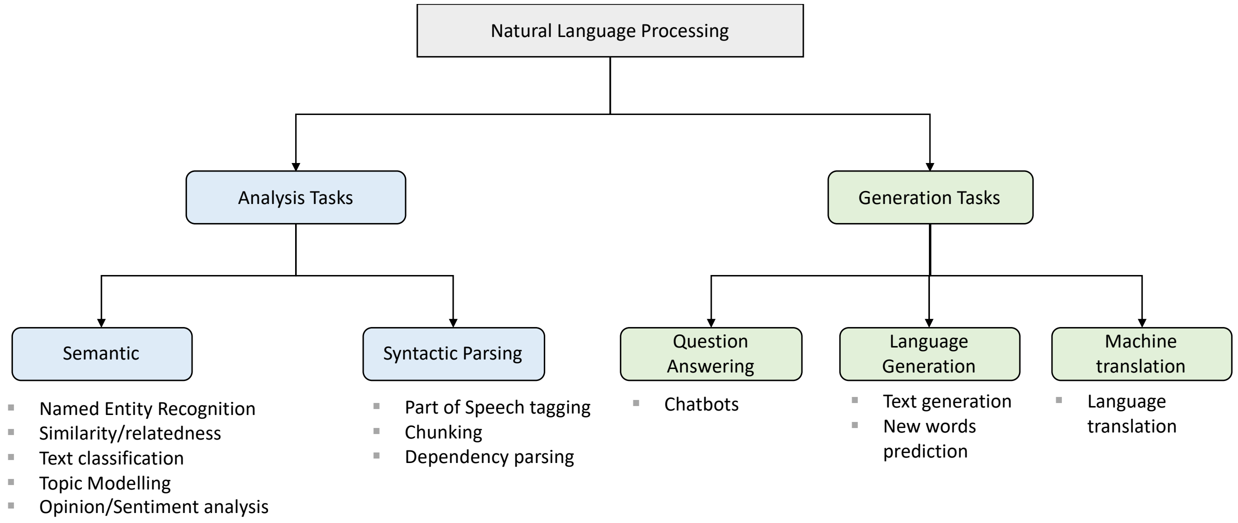


Figure 5: NLP Tasks Categorized in Two Broader Categories: Analysis Tasks (light blue) and Generation Tasks (light green)

4.2.2 Natural Language Processing (NLP) Extended Systems

Customized solutions had to be developed to tackle the unsupervised NLP problem. These solutions were later incorporated by the systems or used as levers to enable better solutions by combining different concepts.

There was a lot of research and development work put into chunking solutions. An extensive literature review was realized and will be presented ahead. However, a few customized solutions were tested. The first customized solution to be developed and tested was called 'brute force chunking'. This methodology creates 1-gram to n-gram combinations of chunks of a given text, counts the frequency of each one of the n-grams. Often the max n-gram to be used was 7, given that each addition of 1 increases the model complexity greatly. With the frequency of each chunk, the algorithm then proceeds to calculate the median frequency for each chunk and ignore the chunks that are below that threshold. That process was performed on all n-grams chunks. Note that this method does not use stopwords neither before nor after the process, only the punctuation was removed.

Another chunking method developed was to use a Recurrent Neural Network, more specifically a LSTM, trained on the given corpus, and for every word in the document, it predicts the next word. If the actual word matches the predicted word, that is added to the current chunk, if not, the current chunk it is ended, and the algorithm proceeds to the next word.

During the development of the project, it was necessary to develop support systems that would provide the metrics to be combined with the larger systems in a framework to provide insight. However, those smaller systems provide enough inside in itself, and were further investigated.

Technology prediction was a very interesting small system that was developed, which leverages room theory, that is further explained in this report. The system can provide a prediction to which other technologies a given technology it is moving in the future time. For example, if a technology of interest has been tracked for the past year and presents a pattern of location in the n-dimensional space, it is possible to fit a mathematical function to it, and then extrapolate its value to the future, as well as its neighbors. Using those positions, which are actually vectors, in the n-dimensional space, to reverse search for the closest known positions in the current context of the room, the system would be capable of identifying technologies that are moving close together, or apart. That would be extremely useful in case of identifying trends, technologies enablers, and technologies chain.

Another interesting system that was created to support the room theory is the identification of incoming documents. By leveraging room theory and Word2Vec algorithms, the system can identify how close an incoming document is from a given technology context. The system uses word2vec to incorporate the context of the text into vectors, that are representative of words and/or chunks of text, and then combines those vectors with an anomaly detection algorithm, such as One-Class Support Vector Machine, to identify the probability of a text belonging to the trained vectors. That enabled the systems to better handle incoming documents, and document classification without any human supervision.

During the process of working with textual data, text position tagging is an extremely complicated task, and to perform it in an unsupervised manner makes it even harder. The focus on unsupervised approach comes from the underlying assumptions that language evolves, and it is hard to compare documents in time with static methods, assuming that language does not change. For example, a paper written 20 years ago do not use the same language compared to a paper published recently. That becomes even more apparent with more informal content, like news, blogs, and social media.

For this particular task, a methodology was developed that leverages deep learning techniques and room theory to identify words positioning from its context. It was based on techniques developed in a new field called Natural Language Understanding, which is an evolution of the NLP field.

The NLPlab team also utilized packages to implement easy-to-use functions that performs functionalities such as Name-Entity-Recognition, which is also an extremely important task to be performed that improve the system’s final results. For this task, the team leveraged existing packages such as Flair and Spacy, which according to literature, are the best available packages to use.

4.2.3 Word2Vec

How is it possible to switch from natural language to vectors? The idea is based on what is called the distributional hypothesis, which was introduced in the 1957 by Dr. J.R. Firth ("You shall know a word by the company it keeps"). This has generated an interest of approaching semantics of words with a new perspective.

In recent years, with the increase in computational power, trends in the usage of Artificial Neural Network (ANN), and the huge amount of textual data available, the distributional hypothesis found a relevant match in a research conducted by the Google Brain team in 2013.

The Word2Vec methodology, introduced by Mikolov (2013), consists in training a neural network with a corpus as an input and returns for each word a n-dimensional vector in output. These vectors, called word embeddings, are distributed representations of words, meaning that similar vectors correspond to words that appear close in the corpus.

Word2Vec is an effective method for a variety of NLP tasks. Its independence from the language makes it a powerful tool for a variety of applications. Having n-dimensional vectors representations (Figure 6) of words enables the possibility to perform mathematical operations with words that can be useful for discovering interesting semantic relationships.

For instance, it is possible to calculate the semantic similarity between two words, or between a given word and a set of other words. Word embeddings are an intelligent way to represent text, which offers new possibilities to understand the semantics in an automated way.

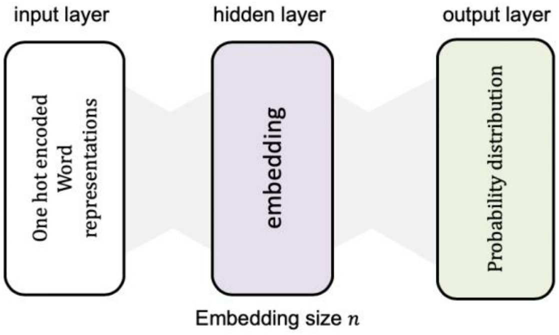


Figure 6: General Schema of a Word Embedding Generation Architecture

4.3 Framework

This section will elaborate on how the system's framework is structured and provide a better understanding on how the systems are communicating with the databases, and an overview of the previously discussed pipeline.

One of this project's requirements is that it should be able to deliver its insights if dealing with streaming data or batch data. That is a challenge on the server development standpoint. We overcame that challenge by creating processes that run on Dockers, and therefore can be deployed without a connection with the internet. Even though the system is able to run without an internet connection, or in a local network, that means that the document (data) collection will be compromised, therefore the system will reflect a snapshot of that point in time.

4.3.1 Framework Design

As discussed above, the system framework was designed to be able to be replicated, scalable, and operate either online or offline (see Figure 7). With those requirements in mind, the system is based on the use of different Docker images, and GitHub repositories to perform its tasks. The first step in the pipeline is the data collection process, which is performed using the source code in the Server repository in the GitHub. The data collection runs every 8 hours daily (3 times per day) to collect papers, patents, news, and blogs, from a set of defined sources and keywords. The keywords were derived from the proxy domain that was already defined. From the proxy domain, the sectors, with its respective technologies were derived, and further keywords to easily search the specific technologies.

Currently the server runs on a Stevens Institute of Technology only private network, and the server it is responsible to run the hyper virtual machines (Docker containers).

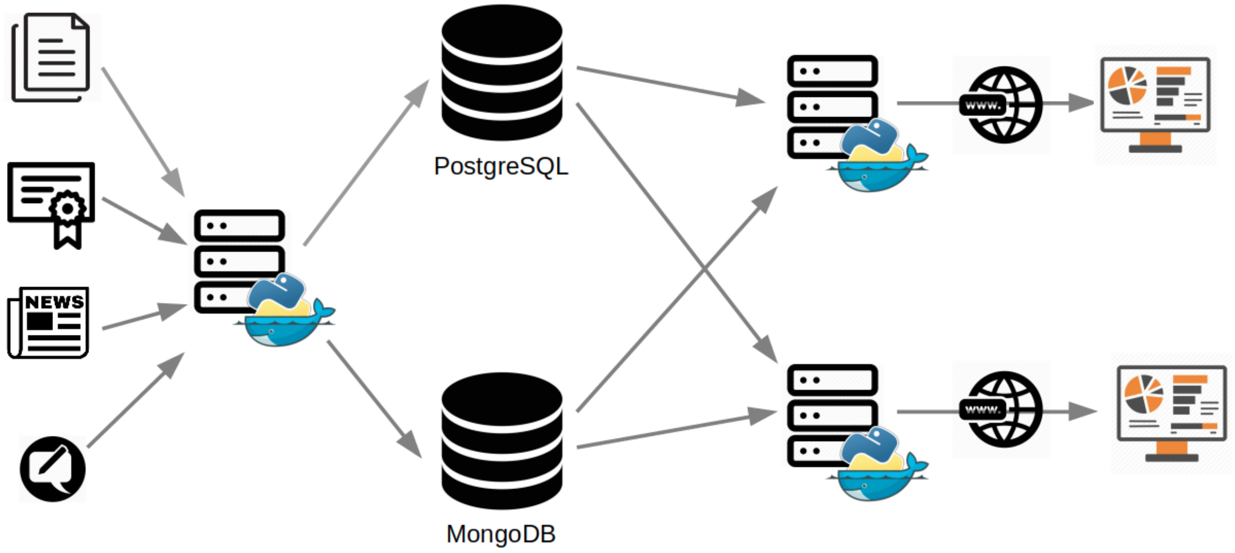


Figure 7: System Framework

The first is the Docker container which runs on the default image, and it is responsible to gather the data, pre-process it with the basic pre-processing techniques, such as punctuation, and lowering all the characters. It uses multiprocessing to speed up the process and first requests the news updates from the previous days and retrieves its textual content into a pandas DataFrame. After collecting the textual data, it proceeds to insert and update (if necessary) any news in the database. The process of gathering, basic cleaning, and inserting/updating the database continues with a search based on keywords for papers and patents. For blog searches, it is based on the sources, yielding broader (sometimes noisy) results.

For the databases, there are two different systems of databases: the structured data, for that we decided to use PostgreSQL, and the unstructured data, using MongoDB. The main difference is that SQL databases expects data to come in a structured and expected format, whereas the unstructured database do not require type definition, nor columns definition, each item is indexed by a unique ID, and the search happens inside the object.

Each of those databases runs inside of a Docker image that has been prepared to run and communicate with the server Docker container. That makes the services reliable, since Docker containers are persistent, when it comes to data, and scalable, since the container can be deployed in any hardware with the correct configurations. The databases will be further explored in the next item.

Each system runs on another Docker container that is based on the default image and has all the necessary installations on it. Given that the production environment derives from the development environment, there are fewer problems when deploying updates in the systems. Each system requests data from the databases, that have been populated with PDF (MongoDB), or structured textual data (PostgreSQL). Systems that make use of room theory code and pre-configured rooms makes those requests from the unstructured database. The systems then proceed to perform its respective pre-processing pipeline, such as chunking, data cleaning, stopwords, or any other possible combination of those.

After pre-processing the textual data, the system then proceeds to perform its respective calculations of metrics, which are explored in the systems sections of this report. After processing the metrics, the system loads the visualizations and servers in a web-based platform. Each system has a different address in the network and can be accessible through the documentation web page.

This framework design follows the characteristics of the bottom-up approach and the agile development, considering that the project's goal was to create a prototype and improve the prototype in iterations and cycles. There is enough flexibility in this design that each part of it can be easily updated without affecting the subsequent or previous parts of the pipeline.

4.3.3 Databases

The two different databases run in two different instances of Docker containers, that means those can be updated, copied and paste, or even deleted, without affecting the functionality of the framework. The structured database uses the engine of a PostgreSQL database, and has a database called 'data'. This database, contains a schema, called 'public'. This schema contains 4 different tables. The tables are:

- news_monitoring
- papers_monitoring
- patents_monitoring
- rss_feed

Each table stores the collected data from the respective source mentioned in the name. The system that performs the data collection is on the Server repository and runs on the Docker image Server.

The unstructured database uses the engine of a MongoDB database and has a database called RT203. This database contains several distinctive information alongside with a file system. The file system is mostly used to save raw downloads from the sources, in case there is missing information, and also to save the binaries from the room theory rooms that can be loaded and used in the framework.

The MongoDB also handles configurations on the server, and communication between the servers when that is not required to be streamed but step wise. The use of two different databases provides the system with flexibility to deal with a wide variety of incoming textual data, and the use of Docker to deploy it brings the scalability to the system as well.

The following table list the number of documents currently collected for each category.

Table 2: Data Collection with the Total Number of Documents

Item	# of Documents
Papers Monitoring	53,570
Patents Monitoring	3,990
RSS feed (blogs)	1,123
News monitoring	2,050
TOTAL	60,733

4.3.4 System Deployment

Given the requirements from the Sponsor, the system can be deployed in an online base or an offline base. The deployment consists in the following steps:

1. Installing Docker in the hardware where the systems will run

2. Setup the Docker environment using the following steps:

    a. Download or load from a hard drive the Docker images with all the necessary libraries and configurations files necessary to run the systems.

    b. Start the containers, based on the previously loaded images, that will be necessary to run the systems. If the system is to run offline, the Server Docker container it is not necessary. If the system is running online, the Server container needs to be started.

    c. The databases Docker containers should be started and loaded with previously downloaded data, the mechanism to do that it is called volumes. Therefore, the loaded volumes should be shipped as well with the hard drive and/or made available online, through a secured link.

3. Free the ports on the network device to ensure that the Dockers can communicate between themselves, and if online, to ensure that the Docker containers can communicate with the internet.

More instructions are provided as part of the final deliverables to the Sponsor.

4.3.5 System Access

The systems were built to deliver insightful information of high dimension complex data. To achieve that goal, the systems were built in a web-based platform that can easily be extended. With the framework setup, all systems can accessible by a URL on a web page. It is up to the users to decide whether the URL will be open to the public, or only be available in an internal network. The system can also be configured with users and passwords to increase the security level.

The components are available in a source code format, which allows the user to have full access to the developed code, its comments and it is free to make any necessary changes. The components will also be available in a Docker image, that can be transformed into a container at any time, for further use. Another delivery method it is through a Jupyter Notebook, which is a data science development tool that is easy-to-use and enable fast development serving as an IDE. The Jupyter Notebook can be accessible online or locally only, and similarly with the systems, it is enabled with users and passwords.

5 Core Systems

The core systems are divided into two subsystems: Technology Monitoring and the Risk Panel (Planning Support System). The Technology Monitoring System has been designed to be used as a stand-alone system, providing insights to the Sponsor on new emerging technologies. It will also provide inputs to the Risk Panel System. Each subsystem is described in detail in the following subsections.

5.1 Technology Monitoring System

With the growing popularity of internet platforms as an important sharing media, it has contributed to the decision-making process in various domains. Accordingly, significant technologies have been developed to process and analyze online data using techniques from different fields such as text mining, machine learning, natural language processing, statistics, and semantic web. Such amalgamation of multiple techniques within a common framework have provided feature-rich analytical tools (Purohit and Sheth, 2013; Davis et. al, 2016) leading to valid, reliable, and robust solutions. A smart service system is one of such domains. A smart service system is a service system capable of learning, dynamic adaptation, and decision making (Medina-Borja, 2015) that requires an intelligent object (Wunderlich et. al, 2015) and involves intensive data and information interactions among people and organizations (Lim et. al, 2018a). One of the most efficient ways of developing a smart service system is using text mining methods to analyzed wide range of online data and documents such as patents, papers, news, reviews, and even customer opinions.

Online platforms provide a multi-modal data structure containing text, images, and videos, along with contextual and social metadata such as temporal and spatial information, and information about user connectivity and interactions. Data such as news, patents, and research papers can be used for predictive analysis in many application areas to understand the technologies that are available in order to predict the emergent technologies in the industry.

To achieve this goal, researchers have focused on a document-centric method to understand the similarity metrics between documents. For example, Phuvipadawat et. al (2010) tried to solve the problem of detecting breaking new topics in Twitter. In this research, tweets were converted to a bag of words and then assigned to clusters based on textual similarity between incoming tweets and existing clusters. Sankaranarayanan et. al (2009) applied both text and temporal distribution methods to output trending topics, however, the presented model suffers from noise sensitivity and fragmentation of clusters.

Generally, feature-centric methods are based on statistical models to extract a set of key words that represent a topic in specific set of documents. As of today, most approaches were based on LDA (Latent Dirichlet Allocation) (Blei et. al, 2003) and some extensions of LDA (Blei et. al, 2006). LDA based approaches identify a set of burst words and uses these words to define clusters defining topics. A newer approach in predictive text analysis is the graph-based approach which detects important keywords based on their pair-wise similarity score. Graph-based approach creates a term co-occurrence graph, where each node represents a token and an edge depicts occurrence of 2 words/tokens in the same text and uses community detection to create topical clusters (Sayyadi et. al, 2009).

To overcome the challenges in the industry, this research project proposes a technology monitoring system (TMS) to scan all the possible source of detecting emerging technologies in the given domain. To do so, this system will target available news, technical papers, and patents as the possible proxy dataset. The TMS comprises of 10 different components in 6 phases of this project. Figure 8 illustrates the overall view of the TMS.

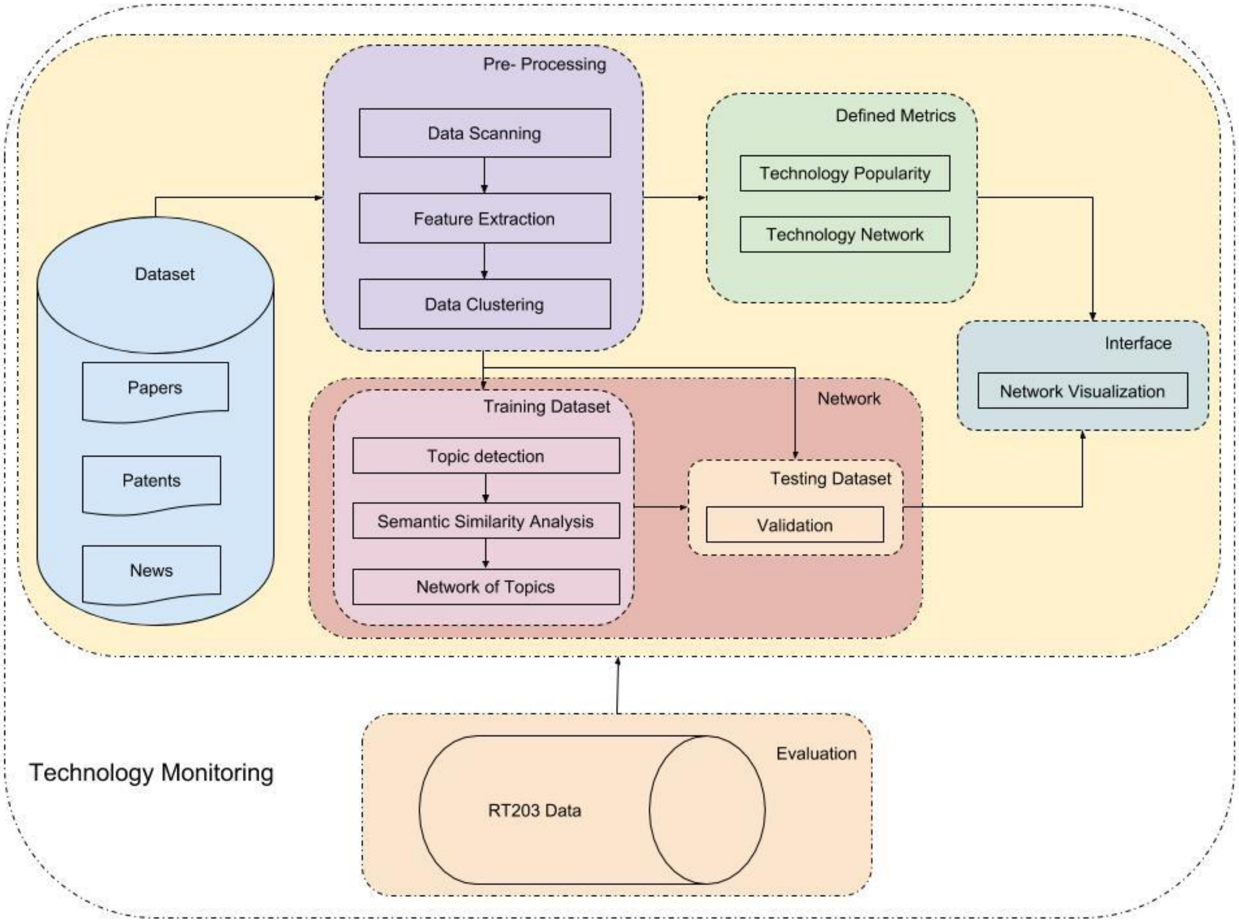
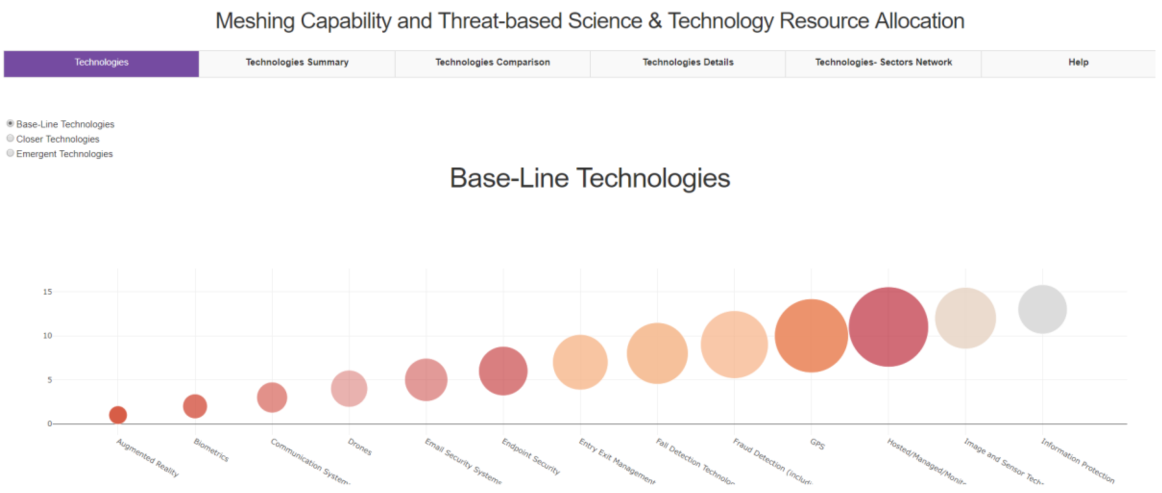


Figure 8: Overall View of the Technology Monitoring System

In Phase 1 of the TMS, the datasets were collected using the component module to crawl online websites in real-time to search for all the available news, technical research papers, and the patents. The results from the searches were added to the available dataset. The following components were used in Phase 2 which is the pre-processing stage:

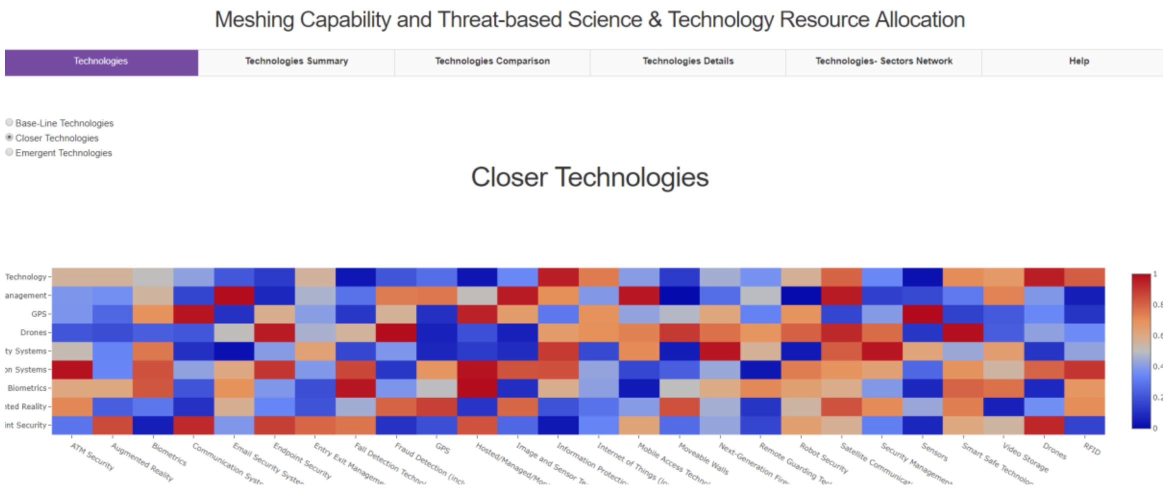
1. Data scanning: this component is scanning the identified documents to clean them
2. Feature Extraction: Cleaned documents will be analyzed to find out the required/related features
3. Data Clustering: Clustering the extracted features based on the chosen baseline technologies updated by the end user to the system. Then system will provide access to list of identified

clusters of closer technologies, technologies which are performing similar job to base line technologies and were in the system for some time, and emergent technologies, group of identified technologies which are new to the market and may or may not perform similar analytical job to the baseline technologies, as shown in Figure 9.



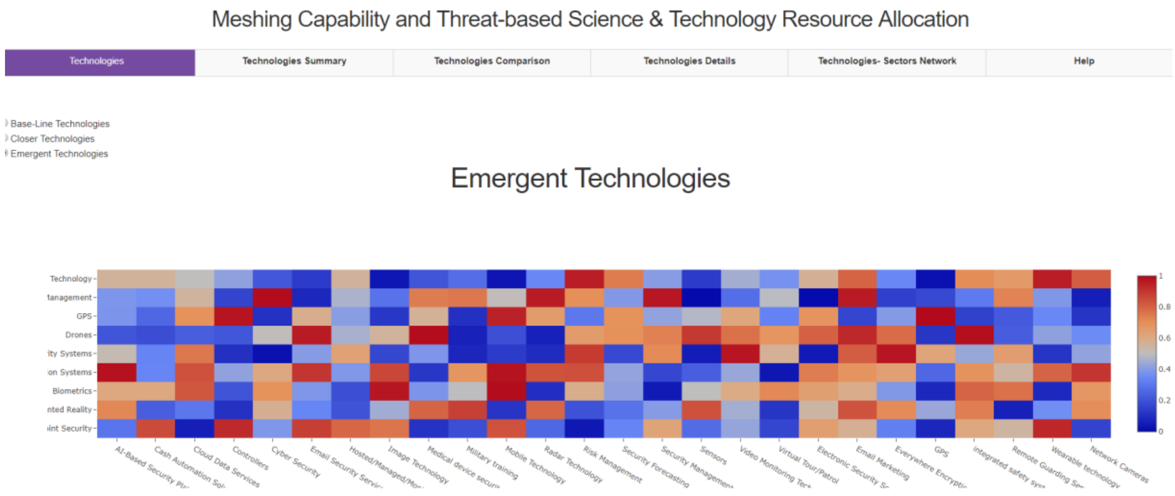
Meshing Capability and Threat-based Science & Technology Resource Allocation

Base-Line Technologies



Meshing Capability and Threat-based Science & Technology Resource Allocation

Closer Technologies



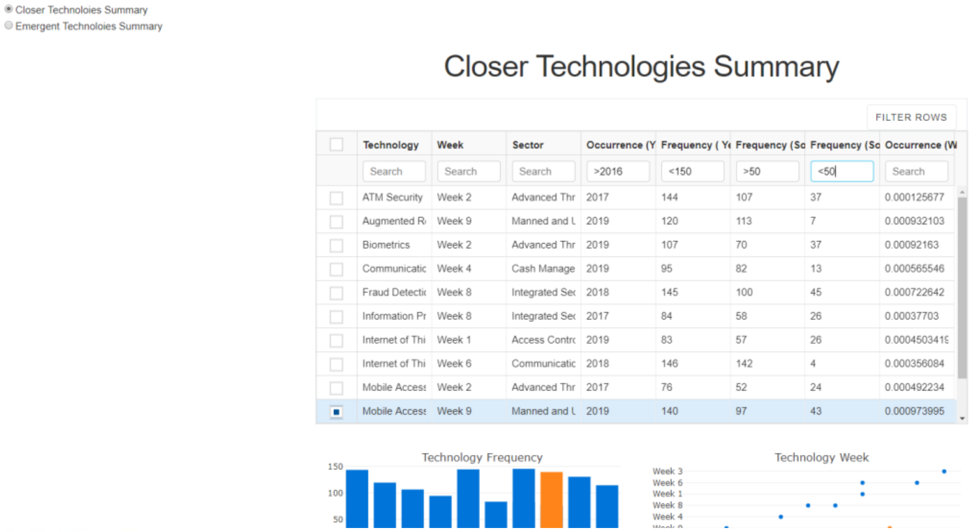
Meshing Capability and Threat-based Science & Technology Resource Allocation

Emergent Technologies

Figure 9: Overview of Technologies Preparation from the Technology Monitoring System

Once the documents corpuses are obtained, different strategies can be used to establish the links between the information. The TMS obtained sub-systems that consists of the topic layer, network layer, and interface layer. The topic layer aggregates data from the dataset and shares the results with the network layer to easily manage the available technology. The highest-level topic will be extracted and divided into sub-networks. It is reported that maximum topics of patent networks tend to be spare compare to paper networks (Shibata et. al, 2010).

In Phase 3, the metrics were defined based on identified technologies. The required metrics to address the goal of the system which is technology popularity and the technology network will be extracted in this phase. During this phase, the summary of identified technologies is presented. The summary table contains the frequency of each technology and source during specific periods, the sector used the technology during that time, and the ratio of technology usage among the list of identified technologies in created dataset. The summary table will be used for comparing different technologies used in different sectors. Figures 10 and 11 display the systems view in Phase 3.

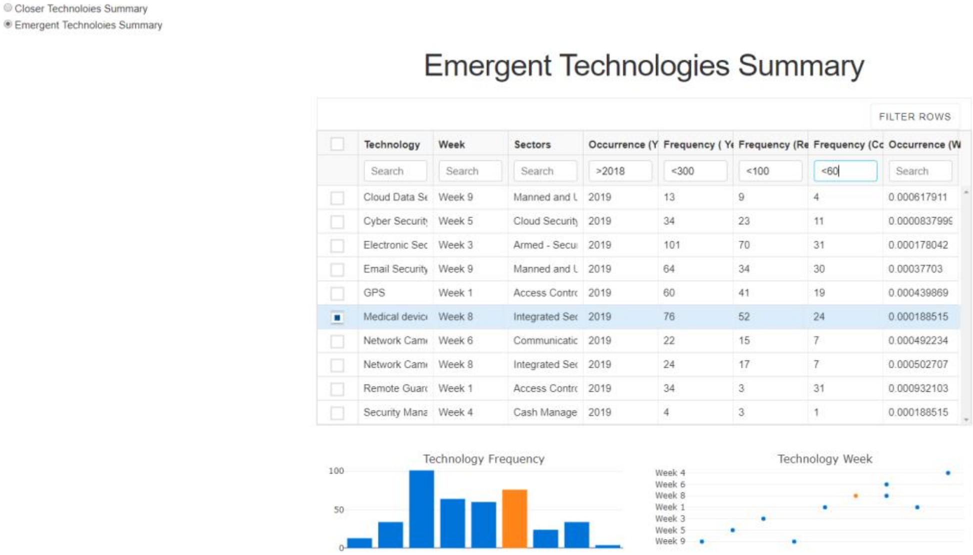


● Closer Technoloies Summary
○ Emergent Technoloies Summary

Closer Technologies Summary

FILTER ROWS

□	Technology	Week	Sector	Occurrence (Y	Frequency (Ye	Frequency (Sc	Frequency (Sc	Occurrence (W
	Search	Search	Search	>2016	<150	>50	<50	Search
□	ATM Security	Week 2	Advanced Thr	2017	144	107	37	0.000125677
□	Augmented R	Week 9	Manned and U	2019	120	113	7	0.000932103
□	Biometrics	Week 2	Advanced Thr	2019	107	70	37	0.00092163
□	Communicatic	Week 4	Cash Manage	2019	95	82	13	0.000565546
□	Fraud Detecti	Week 8	Integrated Sec	2018	145	100	45	0.000722642
□	Information Pr	Week 8	Integrated Sec	2017	84	58	26	0.00037703
□	Internet of Thi	Week 1	Access Contro	2019	83	57	26	0.0004503419
□	Internet of Thi	Week 6	Communicatic	2018	146	142	4	0.000356084
□	Mobile Access	Week 2	Advanced Thr	2017	76	52	24	0.000492234
☒	Mobile Access	Week 9	Manned and U	2019	140	97	43	0.000973995



○ Closer Technoloies Summary
● Emergent Technoloies Summary

Emergent Technologies Summary

FILTER ROWS

□	Technology	Week	Sectors	Occurrence (Y	Frequency (Ye	Frequency (Re	Frequency (Cc	Occurrence (W
	Search	Search	Search	>2018	<300	<100	<60	Search
□	Cloud Data Se	Week 9	Manned and U	2019	13	9	4	0.000617911
□	Cyber Security	Week 5	Cloud Security	2019	34	23	11	0.0000837999
□	Electronic Sec	Week 3	Armed - Secu	2019	101	70	31	0.000178042
□	Email Security	Week 9	Manned and U	2019	64	34	30	0.00037703
□	GPS	Week 1	Access Contro	2019	60	41	19	0.000439869
☒	Medical devic	Week 8	Integrated Sec	2019	76	52	24	0.000188515
□	Network Cam	Week 6	Communicatic	2019	22	15	7	0.000492234
□	Network Cam	Week 8	Integrated Sec	2019	24	17	7	0.000502707
□	Remote Guar	Week 1	Access Contro	2019	34	3	31	0.000932103
□	Security Mana	Week 4	Cash Manage	2019	4	3	1	0.000188515

Figure 10: Overview of the Technologies Summaries from the Technology Monitoring System

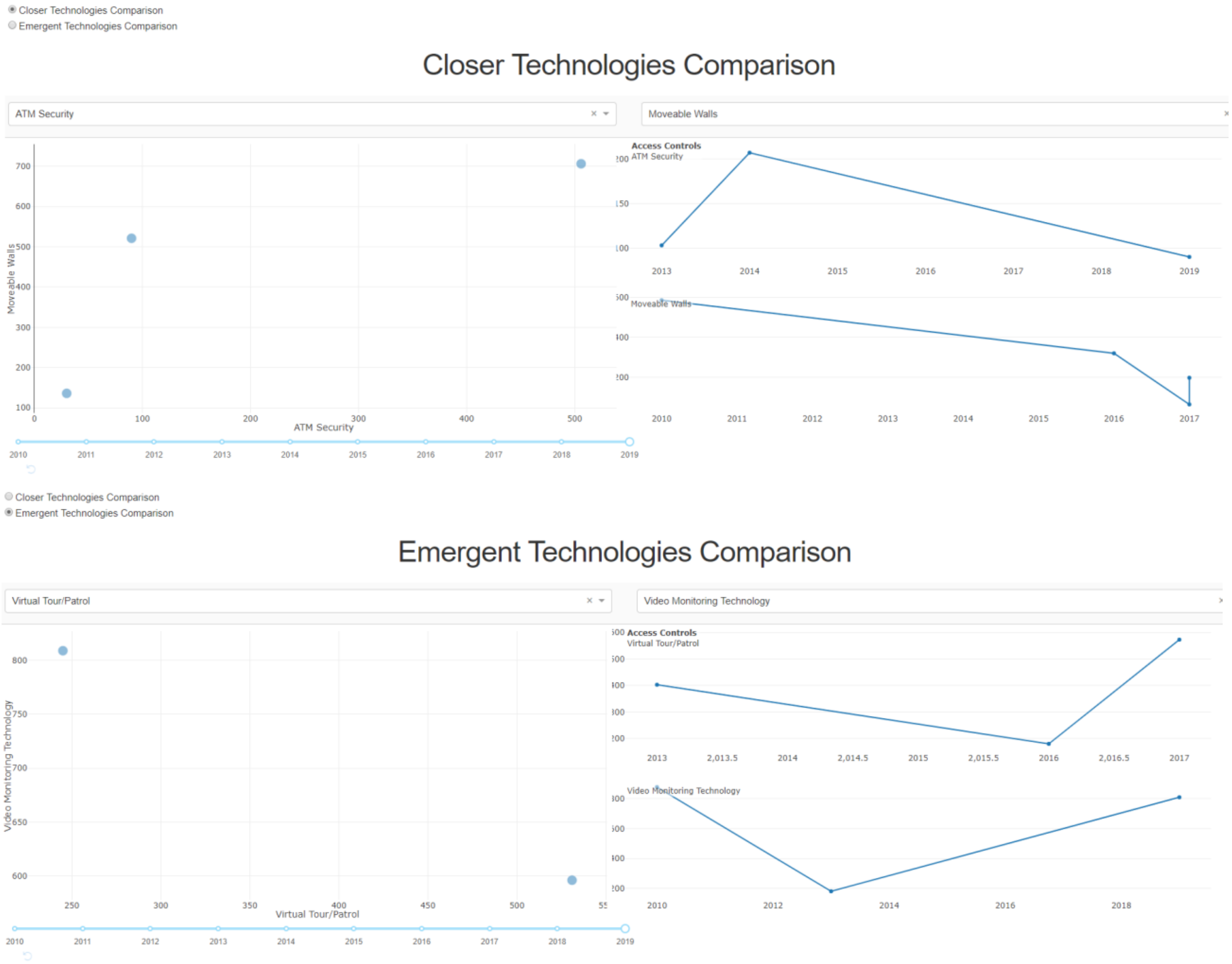


Figure 11: Overview of Technologies Comparison from the Technology Monitoring System

In Phase 4 which contains the technologies details, the TMS will provide the lifecycle of each identified technology, the top five sectors which used the chosen technology during the past year, and the list of usage persistency by sector based on analyzed data in the technology summary table. Moreover, the system provides the successful predictions of usage by sector based on a feasibility analysis and the chance of usage in upcoming technologies based on vectorizing analysis. This phase will provide the opportunity to select the most relevant technologies in the ongoing project as well as upgrading projects with the newest relevant upcoming technologies, if necessary. Figure 12 shows the overview of the technology details from the system.

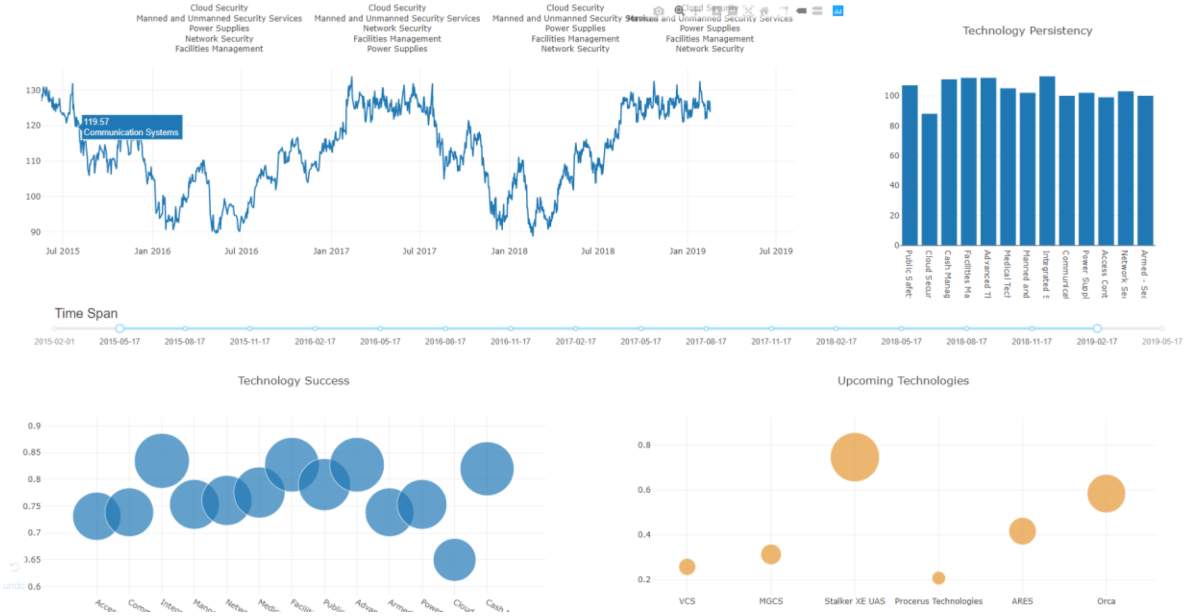
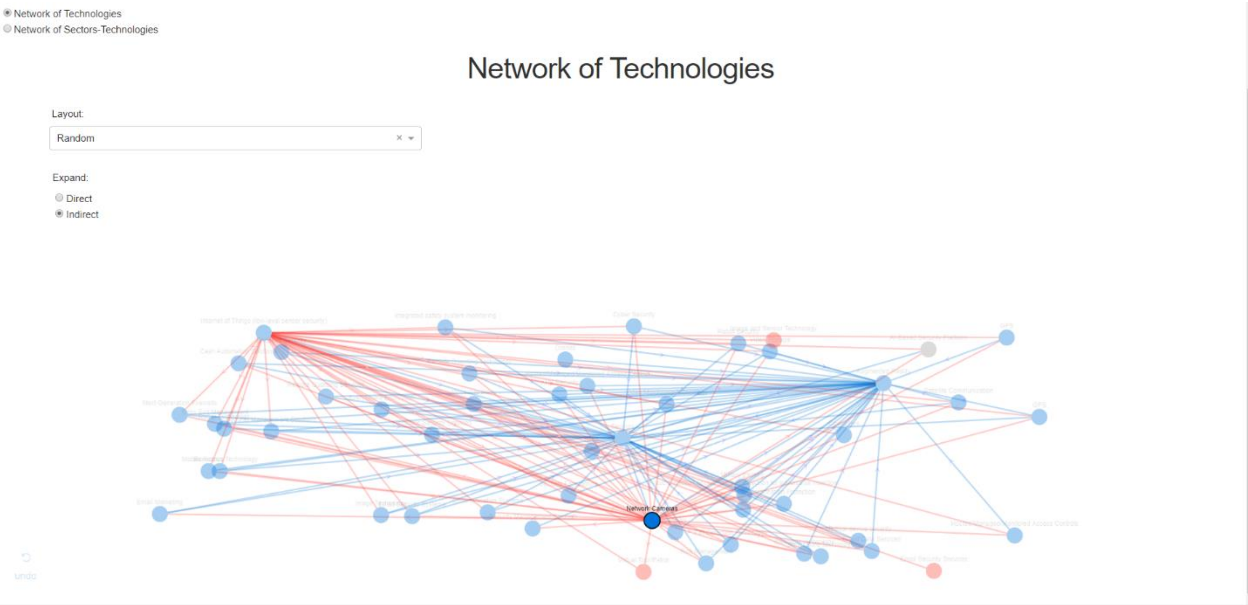


Figure 12: Overview of Technology Details from the Technology Monitoring System

In Phase 5 which is the technologies sectors network, the system calculates the relation among Technologies and Sectors, based on the following:

1. Technology-Technology network, which calculates based on the frequency of each two technologies were used by the same Sector, and
2. Technologies-Sectors network, which calculates based on the frequency of any individual technologies used by any 2 different sectors at the same time.

Figure 13 illustrates the overview of Technologies-Network in the system.

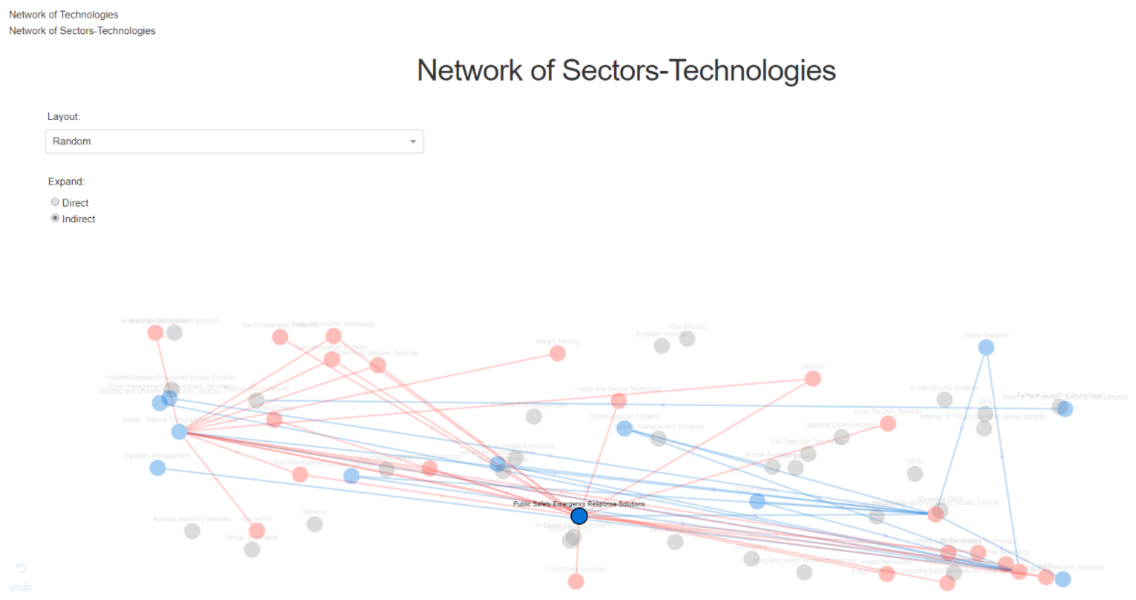


● Network of Technologies
○ Network of Sectors-Technologies

Network of Technologies

Layout:
Random

Expand:
○ Direct
● Indirect



Network of Technologies
Network of Sectors-Technologies

Network of Sectors-Technologies

Layout: Random

Expand:
○ Direct
● Indirect

Figure 13: Overview of Technologies-Sectors Network in the System

5.2 Risk Panel (Planning Support System)

Risk Panel, or the Planning Support System is an interactive panel that is used for what-if analyses. It is a data-intensive decision support system that collects rich information in real-time, analyze technological changes and organizations activities periodically, and recommend technological applications according to users’ preferences and strategic scenarios. The proposed Risk Panel framework in Figure 14 illustrates how the data is used to shape strategic decisions: from problem framing, to data collection and preparation, to exploratory analysis, to modeling and integration, and dashboard representation.

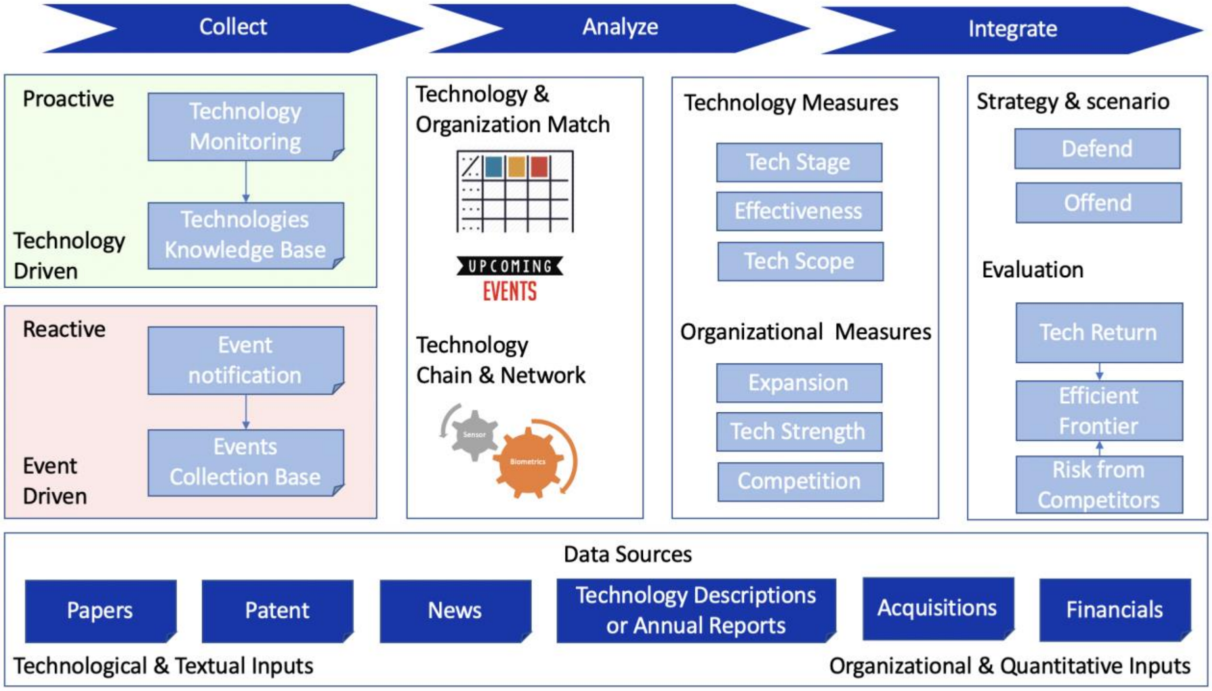


Figure 14: Risk Panel Framework

The Risk Panel identifies and collects information. The enormous amount and variety of data collection form the foundation of Risk Panel development. Data ranges from qualitative papers, patents, and news articles about technologies, to the quantitative numbers regarding organizations financials, as well as merger and acquisition deals. Both qualitative and quantitative information were aggregated as inputs to the Risk Panel.

There are two compatible approaches regarding sources of information: proactively tracking technology development, or reactively collecting ongoing organization activities. The raw data, including papers, patents, and news articles were downloaded in real-time, saved in a database, and used to generate intermediate outputs for risk analysis periodically. Examples of intermediate outputs are related news counts regarding each technology, dependency among technologies and technology chains, technologies belongings to each application, and technologies ownership by various organizations.

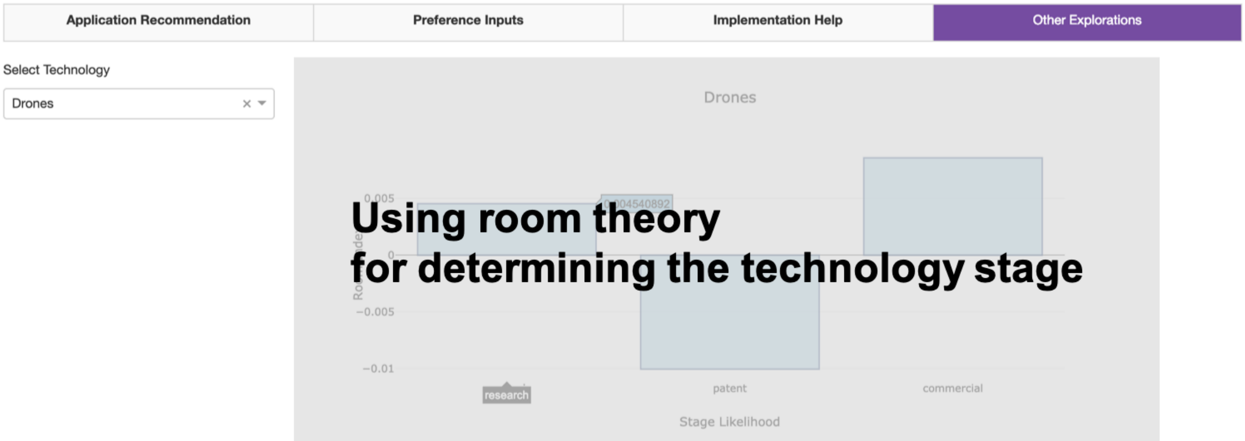
Evaluating technology-driven applications have similar characteristics as optimizing an investment portfolio. It is a decision and tradeoff regarding expected return and risk. The return in the Risk Panel is defined as the maturity, effectiveness, and scalability of a technology-driven application. The risk is defined as the market concentration, organizational expansion rate, and overall technological strength of all organizations within a specific sector. An aggregated score for both returns and risk is derived and illustrated in a Pareto efficient frontier view.

Users are allowed to define their preferences and choose their strategies. For example, users can choose from a general risk preference (risk-seeking, risk-aversion, and neutral). Users may have a particular interest in matching all technologies where the majority have been widely used or competing with a specific company to invest in promising technologies in the top of the technology chain which the competitors have not yet adopted. The final recommendation will be updated under these preferences and scenarios. Table 3 shows the different resources for the data input and application of the data as well as the purpose of the data used in the Risk Panel.

Table 3: Data Input and Application of the Data in the Risk Panel

Resources	Technology	Organization	Purpose
Research Articles (from google scholar)	√		To construct a deep understanding about technology, for room theory use
Patent (from US patent database)	√	√	Technology: to predict technology growth and define the scope of technology; Organization: to evaluate organizations' technology strength
News (currently CNN, upon expansion)	√	√	Technology: to extract technology popularity; Organization: company news are for complementary use
Wikipedia (through API)	√		To gain the general understanding about certain technology, for room theory use
Annual reports		√	To acquire general strategic plans about specific company, for room theory use
Company financials (Hanlon lab Bloomberg financial terminal)		√	To predict company and industry growth
Company mergers and acquisitions (MA)	√	√	To extract MA activities and related sectors

This research project applied an innovative room theory – a word2vec-based text analytics algorithm to determine development stage for technologies. First, a list for queries was created for research, patent, and commercial stages respectively as benchmark. Then, the distance from a list of technology to the benchmark queries was measured to predict the stage for each technology using the embedding rooms that were created. For example, drones are more likely to be either in research or commercial stage, as shown in Figure 15.



Select Technology: Drones

Using room theory
for determining the technology stage

Figure 15: Application of Room Theory to the Forecasting of Technology Stages

The objective of the technology/application recommendation system is to filter and recommend users with the most appropriate items according to their preferences. Instead of providing a one-fit-all solution of technology recommendation, a dynamic weighting system prototype was developed for aggregating multiple measurements. The current system applies an optimization-based approach, with future scalability of machine learning approaches based on user generated historical data. After the aggregated scores have been derived for each technology, the efficient frontier for technologies and a tradeoff regarding expected return and risk were developed, as illustrated in Figure 16.

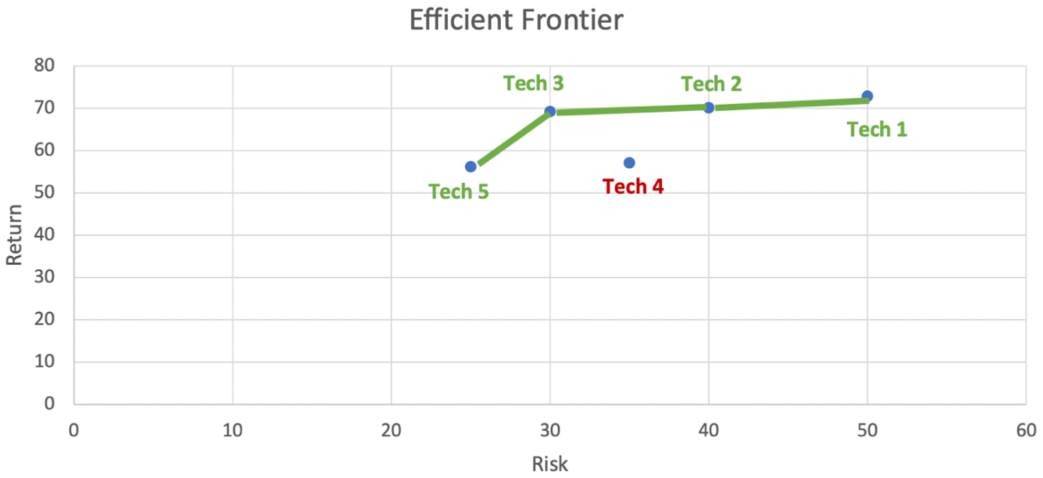


Figure 16: Example Output from the Dynamic Weighting Recommendation and Efficient Frontier Calculation System

6 Results

The overarching purpose of this research project is to provide a computational model to support the planning cycle that will inject relevant threat-based intelligence and operational scenarios into the more traditional capabilities-based model. A proxy domain, "Artificial Intelligence/ Machine Learning in a connected environment" was selected for this project. Specifically, the marketplace of private security companies was used as an example in this project. Data was gathered from the top 22 competitors in this market, which owns 80% of the total market share. Figure 17 shows the breakdown of these findings. The competitors selected were international publicly owned companies with accessible public information. A few private companies within the industry were omitted due to the difficulty in gathering information on these private companies.

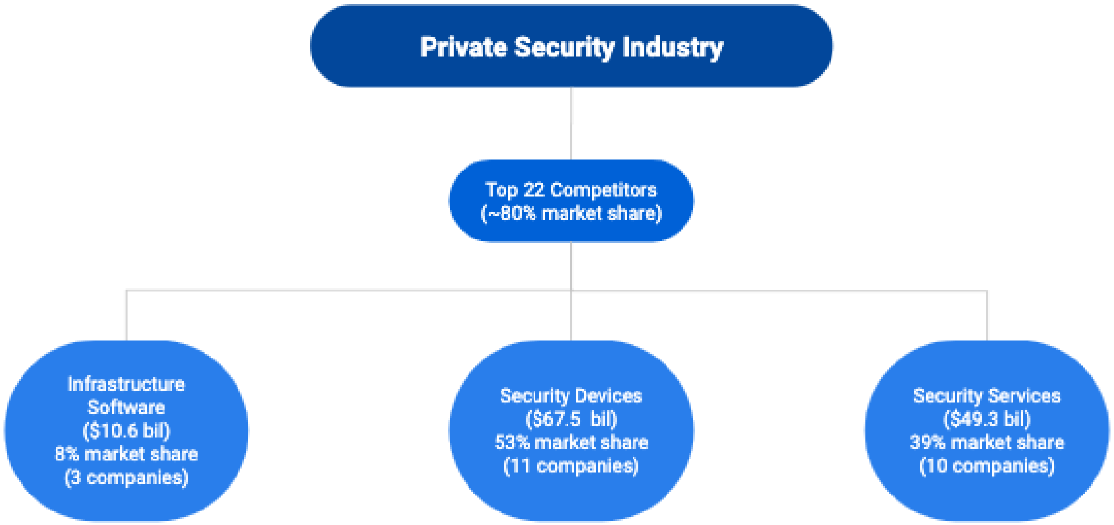

Private Security Industry
Top 22 Competitors (~80% market share)
Infrastructure Software ($10.6 bil) 8% market share (3 companies)
Security Devices ($67.5 bil) 53% market share (11 companies)
Security Services ($49.3 bil) 39% market share (10 companies)

Figure 17: Breakdown of the Private Security Industry Market Share

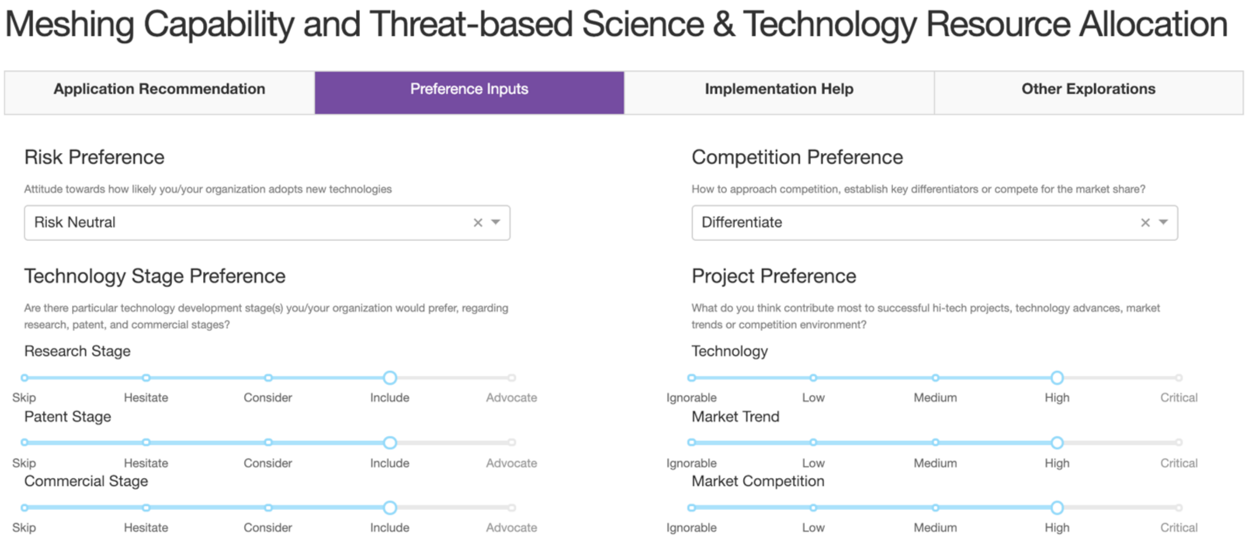
In addition, data collection included general and open source information such as news, patents, and research papers that was used for predictive analysis in the Technology Monitoring System. The predictive analysis enables users to better understand the relevant technologies that are currently available in order to predict the emerging technologies in the industry, specifically, technologies from competitors to maintain a competitor's advantage. In the Risk Panel (planning support system), the following four major risk decision sub panels were developed as a result of this research project:

1. Application recommendation tab — filtering and ranking applications based on embedding technology, market trends, competitors, and others
2. Preference inputs tab — allowing users further adjust inputs and preferences to accommodate their beliefs

3. Application implementing help tab — implementing selected applications by taking competitors’ actions into consideration
4. Acquisition insights tab — monitoring the organizations’ acquisition activities and trends

The preference inputs tab allows users accommodate their beliefs, as displayed in Figure 18. Users can adjust the following inputs:

- Risk preference (risk-seeking, risk-aversion, and neutral);
- Technology stage preference regarding research, patent or commercial stage;
- Evaluation preference which users may weight technical or social issues differently;
- Competition preference whether users may prefer defending or offending; and
- Technology preference in general which users may choose from a set of ideal technologies with intuitions.



Meshing Capability and Threat-based Science & Technology Resource Allocation

Risk Preference

Attitude towards how likely you/your organization adopts new technologies

Risk Neutral

Technology Stage Preference

Are there particular technology development stage(s) you/your organization would prefer, regarding research, patent, and commercial stages?

Research Stage: Include

Patent Stage: Include

Commercial Stage: Include

Competition Preference

How to approach competition, establish key differentiators or compete for the market share?

Differentiate

Project Preference

What do you think contribute most to successful hi-tech projects, technology advances, market trends or competition environment?

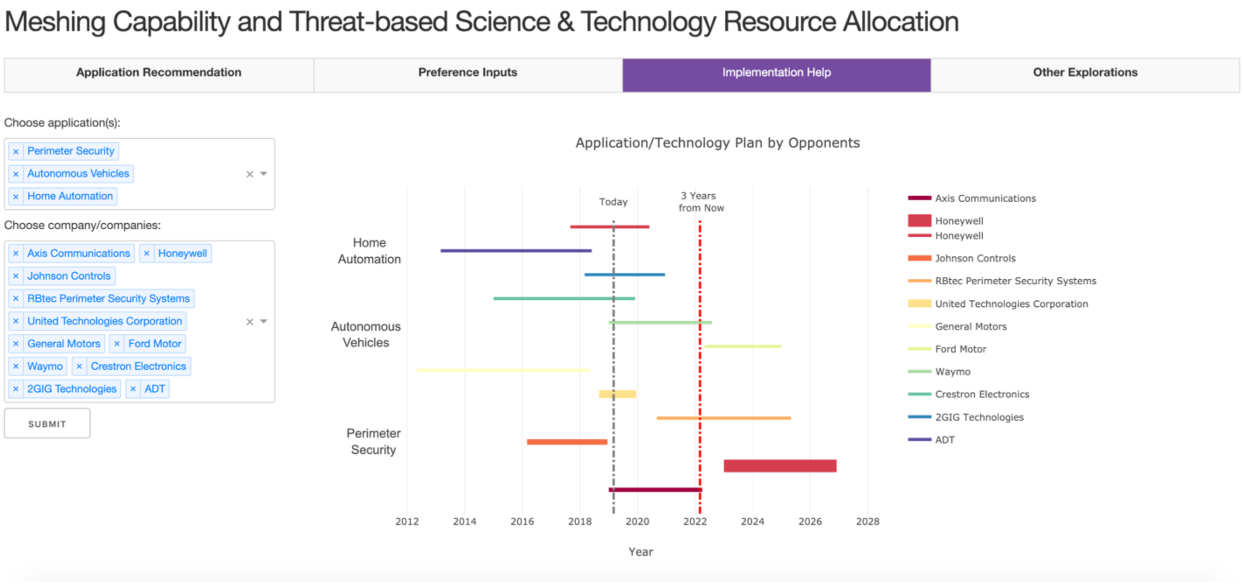
Technology: High

Market Trend: High

Market Competition: High

Figure 18: The Preference Inputs Tab in the Risk Panel

After identifying a set of application and technology candidates, users can view competitors’ estimated product launch schedule and develop a detailed implementation plan using the Risk Panel. Users are expected to change their own start or end dates of selected applications to accommodate the competitors’ actions, as displayed in Figure 19.



Meshing Capability and Threat-based Science & Technology Resource Allocation

Application Recommendation	Preference Inputs	Implementation Help	Other Explorations

Choose application(s):

× Perimeter Security
× Autonomous Vehicles
× Home Automation

Choose company/companies:

× Axis Communications × Honeywell
× Johnson Controls
× RBtec Perimeter Security Systems
× United Technologies Corporation
× General Motors × Ford Motor
× Waymo × Crestron Electronics
× 2GIG Technologies × ADT

SUBMIT

Figure 19: Implementation Tab where Users can View the Competitors' Estimated Product Launch Schedule and Adjust their own Schedule to Accommodate the Competitors' Actions

As mentioned earlier, there are two compatible approaches regarding source of information when users make decisions about technology adoption: proactively tracking technology development, and reactively collecting ongoing organization activities. The reactive approach includes collection of organization activities such as tracking organization news release as well as mergers and acquisitions (MA) events. From the MA activities, the security domain was collected and there was a growing interest for security companies making acquisitions involving technologies such as cloud computing and network securities (see Figure 20), while companies in the traditional commercial and electronics are unlikely to be acquisition targets (Figure 21).

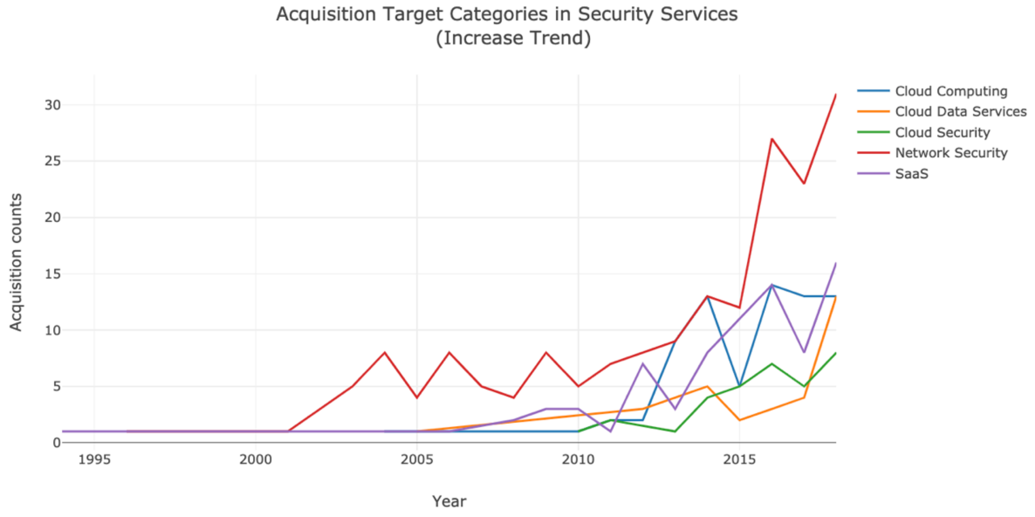


Figure 20: Acquisitions Trends (increasing) in the Security Domain

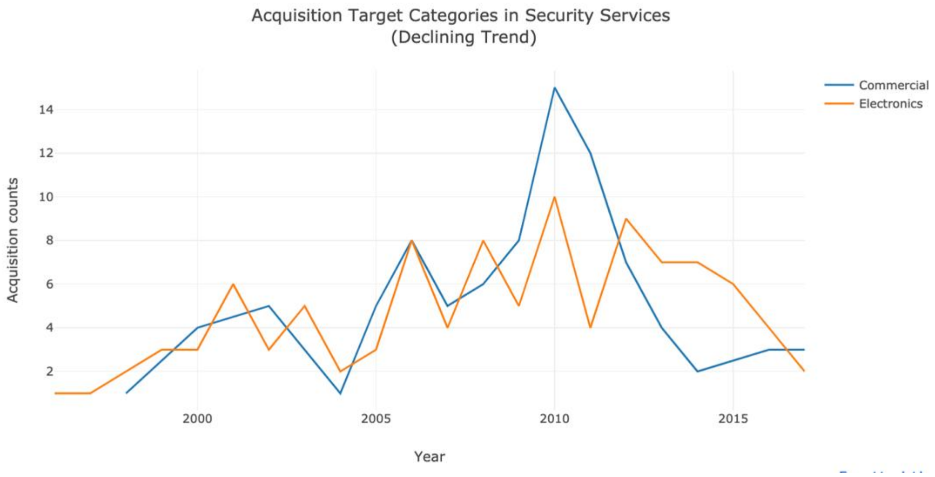


Figure 21: Acquisitions Trends (decreasing) in the Security Domain

Besides the industry landscape, users are allowed to explore acquisitions made by a specific company (see Figure 22) to strategically adjust their own technology planning decisions.

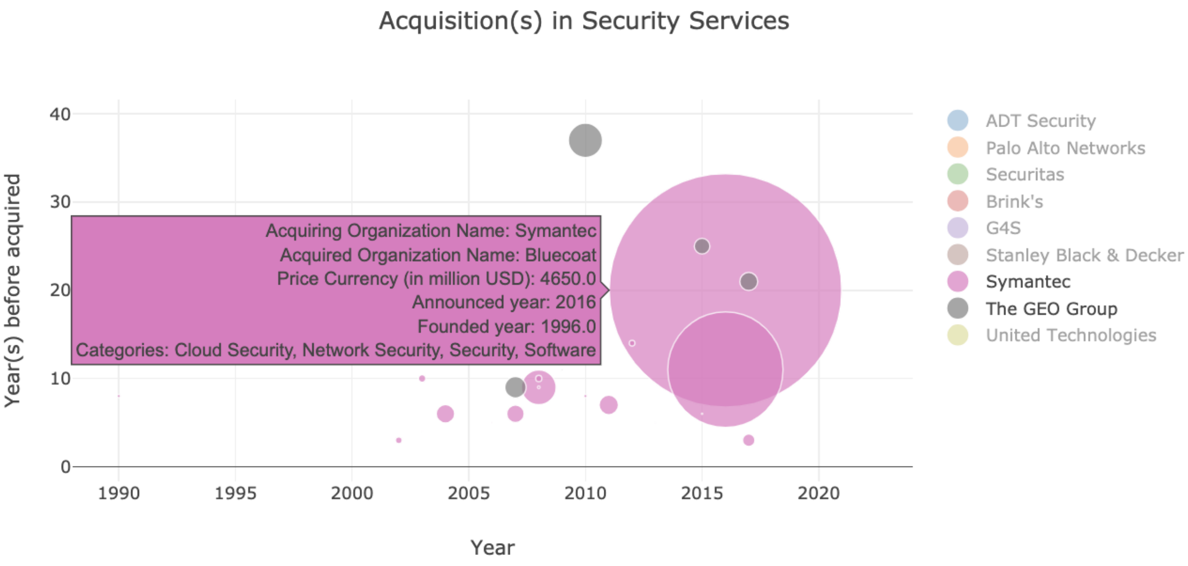


Figure 22: A Series of Acquisitions Made by Individual Companies

Among the top selected private security companies, not all of them are competing in all the technologies. A competitors' network was built using the Mergers and Acquisitions data (see Figure 23). The results were used to identify competitors, adjust competitor-based strategies, and estimate the likelihood of how new technologies affect the existing industry landscape.

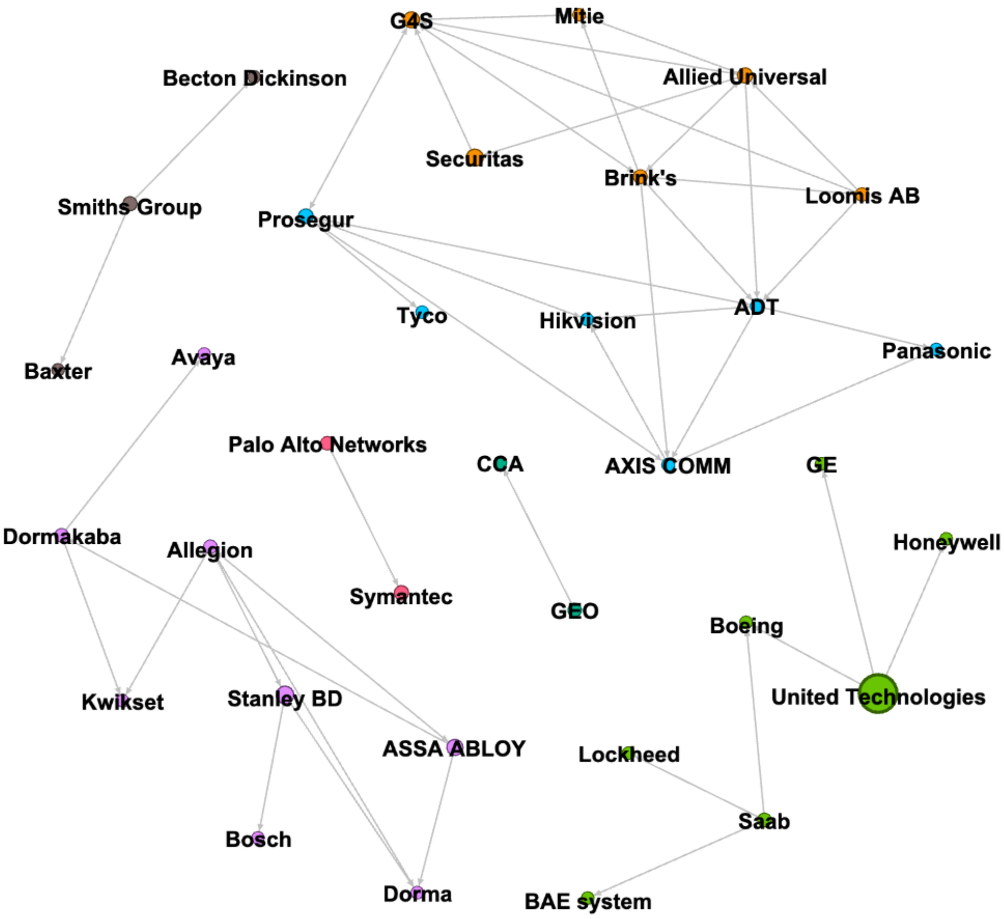


Figure 23: Competitors Network in the Security Domain

7 CONCLUSIONS

A proxy domain was chosen as an example for this project, which was the “AI/ML in a connected environment”. The prototype developed in the first phase of this project comprised of two core systems, the Technology Monitoring and Risk Panel systems. These core systems were developed using modular components based on an agile development methodology. The research project implemented Room Theory, a computational representation of subjective knowledge bases based on the non-computational schema theory.

The Technology Monitoring System has been designed to provide insights on new emerging technologies as a stand-alone system. The Risk Panel system (also known as the planning support system), is an interactive panel used for what-if analyses with a machine learning layer trained by user interactions and provide suggestions for forecasting “optimal” scenarios. The data used as inputs to the Technology Monitoring and Risk Panel systems are based on texts (i.e. news, technical papers, and patents) collected related to the private security industry marketplace, which was relevant to the proxy domain and also easily associated – for content and complexity – to the final domain. NLP was used in pre-processing the data to make machines understand natural language in order to extract valuable information in an automated way.

The research work in this project resulted in the proof of concept that was essential to provide a computational model to support the planning cycle that will inject relevant threat-based intelligence and operational scenarios into the more traditional capabilities-based model for the U.S. CCDCAC.

8 FUTURE RESEARCH DIRECTIONS

Moving forward, this research project will continue with Phase 2 for 2019-2020. The team plans to continue improving the prototype and to develop more components to enhance the visualizations and analytic capability, as well as introduce additional parameters and variables to refine the framework. In doing so, the team will work with the Sponsor to consolidate the list of market segments and technologies, all while continuing to extract data on each. The proof of concept allowed the team and Sponsors to see the generated "rooms" from the data as well. The components will also be validated and made available as standalone. With the inclusion of team members from the Sponsor side, the research team will work to continue to update wikis for the systems and components, providing greater insight, including developing use-case scenarios to help bound the scope while exercising the art-of-the-possible.

APPENDIX A: LIST OF PUBLICATIONS RESULTED

Desai, P., Saremi, R., Hoffenson, S., Lipizzi, C. (2019). “Agile and Affordable: A Survey of Supply Chain Management Methods in Long Lifecycle Products”. 2019 IEEE Systems Conference, Orlando, FL

Lipizzi, C. (2018). “Text Mining in an Evolving Society: Getting Insights from Text in Times of Minimally Structured Conversations”. CESUN, Tokyo, Japan.

Lipizzi, C. (2019). “Extracting Decision-Making Metrics from Text and Placing the Human Feedback in the Quantitative Loop”. INFORMS 2019 Annual Meeting, October 20-23, 2019, Seattle, WA (accepted).

Lipizzi, C., Borrelli, D., Capela, F. (2019). “A Computational Model Implementing Subjectivity with the “Room Theory” – The case of Detecting Emotion from Text”. Computers in Human Behavior Journal (in-progress).

APPENDIX B: CITED AND RELATED REFERENCES

Ajah, A. N., & Herder, P. M. (2005). Addressing Flexibility During Process and Infrastructure Systems Conceptual Design: Real Options Perspective. *IEEE International Conference on Systems, Man and Cybernetics*, *4*, pp. 3711-3716. Delft, Netherlands.

Banerjee, P., & de Weck, O. (2004). Flexibility Strategy--Valuing Flexible Product Options. *INCOSE/ICSE Conference on Synergy Between Systems Engineering and Project Management*, (p. 8). Las Vegas, NV.

Baykasoglu, A. (2009). Quantifying Machine Flexibility. *International Journal of Production Research, 47*(15), 4109-4123.

Blei, D. M., and Lafferty, J. D. (2006). Dynamic topic models, in Proc. ICML: 23rd Int. Conf. Machine Learning, New York, NY, USA, 2006, pp. 113-120, ACM.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, J. Mach. Learn. Res., vol. 3, pp. 993-1022, Mar. 2003.

Boden, B., & Ahlen, A. (2011). *Real Options Analysis: A Study of Implementation Impediments.* Thesis, University of Gothenburg, School of Business, Economics, and Law, Gothenburg, Sweden.

Brown, O., & Eremenko, P. (2008). *Application of Value-Centric Design to Space Architectures: The Case of Fractionated Spacecraft.* American Institute of Aeronautics and Astronautics, Reston, VA.

Brown, O., & Eremenko, P. (2009). Value-Centric Design Methodologies for Fractionated Spacecraft: Progress Summary From Phase 1 of the DARPA System F6 Program. *AIAA Space 2009 Conference and Exposition.* Pasadena, CA.

Brown, O., Long, A., Shah, N., & Eremenko, P. (2007). System Lifecycle Cost Under Uncertainty as a Design Metric Encompassing the Value of Architectural Flexibility. *AIAA Space Conference*, (pp. 216-229).

Christensen, D., Searle, D., & Vickery, C. (1999). The Impact of the Packard Commission's Recommendations on Reducing Cost Overruns on Defense Acquisition Contracts. *Acquisition Review Quarterly*(Summer), 251-262.

Collopy, P., & Hollingsworth, P. (2009). Value-Driven Design. *9th AIAA Aviation Technology, Integration, and Operations Conference.* Hilton Head, SC.

Copeland, T., & Keenan, P. (1998). How Much is Flexibility Worth. *McKinsey Quarterly*(2), 38-49.

Cormier, P., Olewnik, A., & Lewis, K. (2008). An Approach to Quantifying Design Flexibility for Mass Customization in Early Design Stages. *ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *4*, pp. 203-216. Brooklyn, NY.

de Neufville, R., & Scholtes, S. (2011). *Flexibility in Engineering Design.* Cambridge, MA: MIT Press.

Davis, C. A., Ciampaglia, G. L., Aiello, L. M., Chung, K., Conover, M. D., Ferrara, E., Flammini, A., Fox, G. C., Gao, X. Goncalves, B., Grabowicz, P. A., Hong, K., Hui, P.M., McCaulay, S., McKelvey, K., Meiss, M. R., Patil, S., Kankanamalage, C. P., Pentchev, V., Qiu, J., Ratkiewicz, J., Rudnick, A., Serrette, B., Shiralkar, P., Varol, O., Weng, L., Wu, T.-L., Younge, A. J., and Menczer, F. (2016). “OSoMe: the IUNI observatory on social, media,” PeerJ Computer Science

Defense Acquisition University. (2012, February). *Defense Acquisition Guidebook.* Retrieved February 2012, from https://dag.dau.mil/Pages/Default.aspx

Drezner, J., & Krop, R. (1997). *The Use of Baselining in Acquisition Program Management.* RAND, Santa Monica, CA.

Drezner, J., Jarvaise, J., Hess, R., Hough, P., & Norton, D. (1993). *An Analysis of Weapon System Cost Growth.* RAND, Santa Monica, CA.

Ekstrom, M., & Bjornsson, H. (2005). Valuing Flexibility in Architecture, Engineering, and Construction Information Technology Investments. *Journal of Construction Engineering & Management, 131*(4), 431-438.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Fitzgerald, G., Barad, M., Papazafeiropoulou, A., & Alaa, G. (2009). A Framework for Analyzing Flexibility of Generic Objects. *International Journal of Production Economics, 122*(1), 329-339.

GAO. (2009). *DEFENSE ACQUISITIONS: Assessments of Selected Weapons Programs.* Government Accountability Office. Washington, D.C.: GAO.

Github. (2018). "NetworkX: Software for Complex Networks". Retrieved from https://networkx.github.io/

Hadar, Yonatan (2019). The Best Tips for Agile Data Science Research. Retrieved from https://www.kdnuggets.com/2019/03/best-tips-agile-data-science-research.html

Kumar, R. (1999). Understanding DSS Value: An Options Perspective. *Omega, 27*(3), 295-304.

Leach, P. (2006). *Why Can't You Just Give Me The Number? An Executive's Guide to Using Probabilistic Thinking to Manage Risk and to Make Better Decisions.* Sugar Land, TX: Probabilistic Publishing.

Lim, C., and Kim, M. (2018a). Using data to advance service: Managerial issues and theoretical implications from action research. Service Theory Practice 28(1):99–128.

Lipizzi, C., Dessavre, D. G., Iandoli, L. & Ramirez-Marquez, J. E. (2016). Towards computational discourse analysis: A methodology for mining Twitter backchanneling conversations. *Computers in Human Behavior, 64, pp. 782-792.*

Lipizzi, C., Iandoli, L., and Marquez, J. (2015). Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter , International Journal of Information.

Lipizzi, C., Ramirez-Marquez, J. E., Dessavre, D. G. and Iandoli, L. (2016). Social Media Conversation Monitoring: Visualize Information Contents of Twitter Messages Using Conversational Metrics. In M. Connolly (ed.), *ICCS,* pp. 2216-2220: Elsevier.

Mayer, Z., & Kazakidis, V. (2007). Decision Making in Flexible Mine Production System Design Using Real Options. *Journal of Construction Engineering and Management, 133*(2), 169-180.

Medina-Borja A (2015). Smart things as service providers: A call for convergence of disciplines to build a research agenda for the service systems of the future. Service Science.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Minsky, M. (1974). A framework for representing knowledge.

Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in Twitter, in Proc. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM Int. Conf., 2010, vol. 3, pp. 120-123

Purohit, H. and Sheth, A. (2013). “Twitris v3: From Citizen Sensing to Analysis, Coordination and Action,” in ICWSM.

Rajan, P., Van Wie, M., Campbell, M., Wood, K., & Otto, K. (2005). An Empirical Foundation for Product Flexibility. *Design Studies, 26*(4), 405-438.

Ross, A. M. (2006). *Managing Unarticulated Value: Changeability in Multi-Attribute Tradespace Exploration.* Cambridge, MA: Massachusetts Institute of Technology.

Rumelhart, D. E. (1983, June 7). *DTIC.* Retrieved from DTIC: https://apps.dtic.mil/dtic/tr/fulltext/u2/a130662.pdf

Ryan, E., Schubert, C., Jacques, D., & Ritschel, J. (2013). A Macro-Stochastic Model for Improving the Accuracy of DoD Life Cycle Cost Estimates. *Journal of Public Procurement, 13*(1), 103-132.

Saleh, J., Hastings, D., & Newman, D. (2003). Flexibility in System Design and Implications for Aerospace Systems. *Acta Astronautica, 53*(12), 927-944.

Saleh, J., Mark, G., & Jordan, N. (2009). Flexibility: A Multi-disciplinary Literature Review and a Research Agenda for Designing Flexible Engineering Systems. *Journal of Engineering Design, 20*(3), 307-323.

Sankaranarayanan, J., Samet, H., eB. E. Teitler, M. D. Lieberman, and J. Sperling, (2009). Twitter stand: News in tweets, in Proc. GIS: 17th ACM Int. Conf. Advances in Geographic Information Systems, New York, NY, USA, 2009, pp. 42-51.

Sayyadi, H., Hurst, M. and Maykov, A. (2009). Event detection and tracking in social streams, in ICWSM, E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, Eds. Palo Alto, CA, USA: AAAI Press, 2009

Shah, N., Viscito, L., Wilds, J., Ross, A., & Hastings, D. (2008). Quantifying Flexibility for Architecting Changeable Systems. *6th Conference on Systems Engineering Research.* Los Angeles, CA.

Shibata N., Kajikava Y., and Sakata I. (2010). Extracting the commercialization gap between science and technology – Case study of a solar cell. Technol Forecast. Soc. Change 77

Sivanthi, T., & Killat, U. (2008). Valuing the Design Flexibility of a Distributed Real-time Embedded System. *International Symposium on Industrial Embedded Systems*, (pp. 163-168). La Grande Motte, France.

Suh, E., de Weck, O., & Chang, D. (2007). Flexible Product Platforms: Framework and Case Study. *18*(2), 67-89.

United States Navy. (2009, February 18). *America's Navy*. Retrieved September 2, 2013

Wünderlich N. V., Heinonen K., Ostrom A. L., Patricio L., Sousa R., Voss C., Lemmink J. G. (2015). Futurizing smart service: Implications for service researchers and managers. Services Marketing 29(6/7):442–447.

Younossi, O., Arena, M., Leaonard, R., Roll, C., Jain, A., & Sollinger, J. (2007). *Is Weapon System Cost Growth Increasing?* RAND Corporation, Santa Monica, CA.