WRT-1010: Meshing Capability and Threat-based Science & Technology Resource Allocation

Sponsor: US Army

Dr. Carlo Lipizzi

12th Annual SERC Sponsor Research Review
HOSTED VIRTUALLY ON: November 18, 2020
www.sercuarc.org

WRT-1010

- **Title**: Meshing Capability and Threat-based Science & Technology Resource Allocation (Contract [HQ0034-19-D-0003, TO#0150])
- This research was focused on providing a computational model to support the planning cycle injecting relevant threat-based intelligence and operational scenarios into the more traditional capabilities-based planning
- This approach better inform the technical communities charged with future systems developments and has been piloted in late 2016 at the U.S. Army Combat Capabilities Development Command Armaments Center (CCDCAC)

WRT-1010: Key features

- **Replicate the process developed at CCDC AC 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. 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**
- Project phases:
  - Phase I (FY 18): awarded
  - Phase II (FY 19): awarded

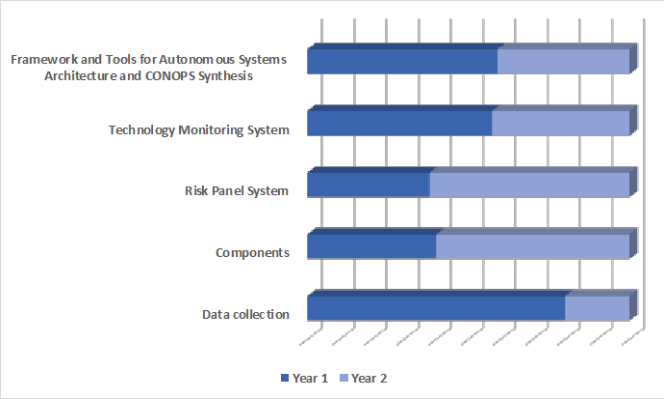
The context: scenario

- “Inter-state strategic competition, not terrorism, is now the primary concern in U.S. national security” (National Defense Strategy, 2018)
- Technology/Innovation is a key factor in this competition: “We cannot expect success fighting tomorrow’s conflicts with yesterday’s weapons or equipment” (National Defense Strategy, 2018)
- "Know the enemy and know yourself; in a hundred battles you will never be in peril. When you are ignorant of the enemy but know yourself, your chances of winning or losing are equal. If ignorant both of your enemy and of yourself, you are certain in every battle to be in peril" (Sun Tzu, The Art of War)
- Among the key capabilities: “Command, control, communications, computers and intelligence, surveillance, and reconnaissance” (National Defense Strategy, 2018)

Overall view

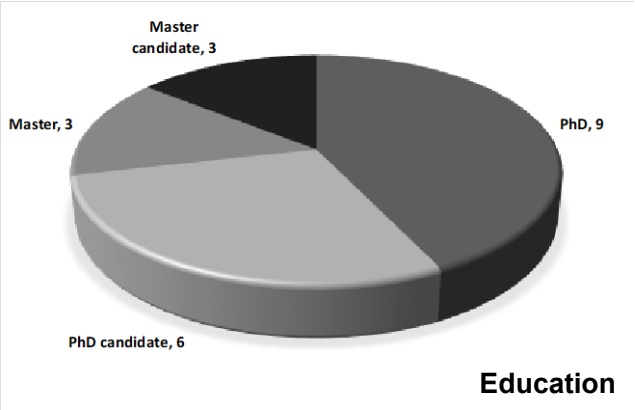
- The project is logically distributed in 2 years, where the 1st year is focused on acquiring the logic used in the planning process, create a corpus to be used for the data/text mining, develop the required components and create proof of concepts for the systems
- The 2nd year is focused adding functionalities to the proof of concepts and making it evolve into a working prototype. The majority of the activities in year 1 will be revised/expanded in year 2

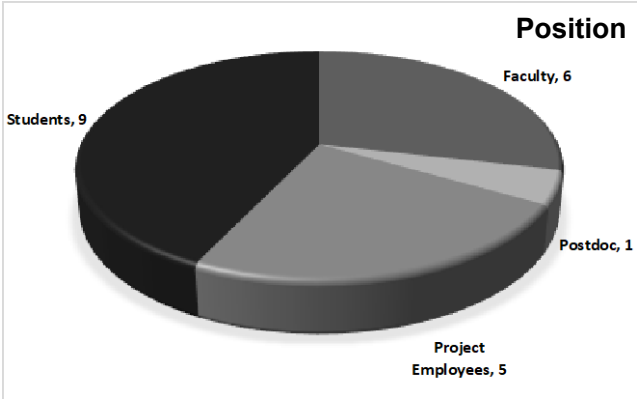
Year	Focus	Key Deliverables
Pre- 2019 (Phase 1)	• Replicate the process developed at ARDEC/ CCDC AC in 2016 to validate the notional computational architecture • Enhance the visualization and analytic capability for a rapid, high fidelity decision-making support tool	• Develop the methodology used in the planning process and create a corpus for data/text mining • Deliver prototypes/proof of concepts with increasing capabilities, using a partially automatic learning approach with an agile approach
2020 (Phase 2)	• Increase the functionalities to the proof of concepts	• Evolve and improve the proof of concepts into a working prototype



Who we are in Phase 2

- **Total number_____21**
  - "Permanent" members____21
    - >= 50% of their time___11
  - "Temporary" members_____0





How we work

Methodology

- Bottom-up, Data/Text-driven approach
- Using a “proxy-domain” to source the data
- Systems are developed as agile growing prototypes with modular components. Most of the components are developed separately for a better reuse

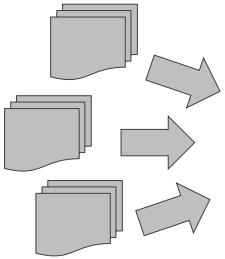
Implementation

- We use a combination of traditional Natural Language Processing (mainly for the preprocessing) and embeddings, that are feature vectors for conversational elements in that text (such as words), calculated via Python using libraries such as Word2Vec
- From the embeddings we extract specific metrics – using our own methodology/algorithms, the “room theory” - for risk evaluation and for visualization

What shaped the system

- The reality we want to monitor first and evaluate then is unstructured and unpredictable
- A top-down, model-based approach wouldn't work
- Majority of the potential sources are textual, while we need measurable semantic insights
- Text/Natural Language may provide different meanings for different people/context
- Most of the insights we could get are from the evolution in time of specific elements
- The evolution in time may contain indications to predict future scenarios

TBDS – Logical view



Text/Data Sources

Risk Decision Support System

Technology Monitoring & Forecasting

Scenarios & Recommendations

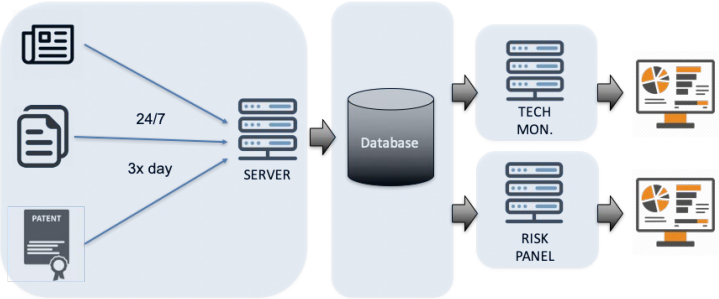
- The system (**Threat Based Decision System** - TBDS) has 2 main components:
  - **Risk Decision Support System** based on the competitive scenario and the “threats” detected from the incoming text
  - **Technology Monitoring System** analyzing streams of domain-specific documents and detecting emerging technologies and forecasting coming/probable future technologies

- TBDS has a user interface to input documents to evaluate, was developed over 2 years, by a team of 20+ people. The minimum viable product has been released at the end of August 2020

TBDS: details

TBDS – the coding

Item	Lines of Code
Server	863
Components	4,920
Database Tools	644
Technology Monitoring System	3,354
Risk Panel System	3,343
TOTAL	**12,124**



The data

Item	# of Documents
Papers Monitoring	3,352
Patents Monitoring	5,123
News Monitoring	2,263,076
TOTAL	**2,271,551**

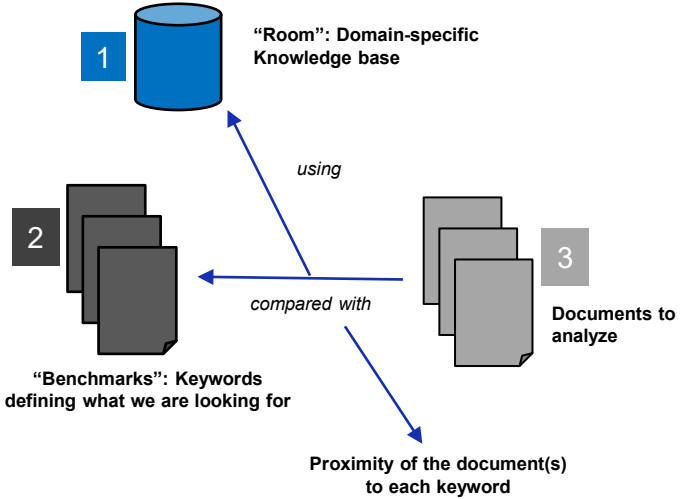
Key Challenge: Extracting Semantic Metrics from Text

- To reduce the risk of wrong/subjective interpretations when making decision based on text, we need to extract metrics out of it
- How do we get numbers from text? Statistical methods provides a limited view, because of their lack of semantic analyses
- If we use semantics, how can we deal with subjectivity/contextuality of the interpretation in text that can or cannot be in given semantic structures (such as ontologies)? Plain use of a generalized reference corpus does not provide any subjectivity
- The task of analyzing a text has a bias, that is related to who is reading/analyzing it. For example, if we want to detect emotions in a text, “ecstasy” may have different meanings for a narcotics officer, a Vatican scholar or a psychologist

TBDS key component: the “room theory”

- The ”room theory” is a framework to address the relativity of the point of view by providing a computational representation of the context we want to use to evaluate the text
- The non computational theory was first released as “schema theory” by Sir Frederic Bartlett (1886–1969) and revised for AI applications as “framework theory” by Marvin Minsky (mid ‘70)
- For instance, when we enter a physical room, we instantly know if it is a bedroom, a bathroom, or a living room
- Rooms/schemata/frameworks …
    - Are mental frameworks that an individual possesses
    - A mental framework is what humans use to organize remembered information
    - Represent an individuals view of reality and are representative of prior knowledge and experiences
- We create computational ”rooms” by processing large corpora from the specific domain/community generating numerical dataset (“embeddings table”). We consider a table as a knowledge base for the context/point of view
- The “room” method makes the whole approach easy to be moved to different domains

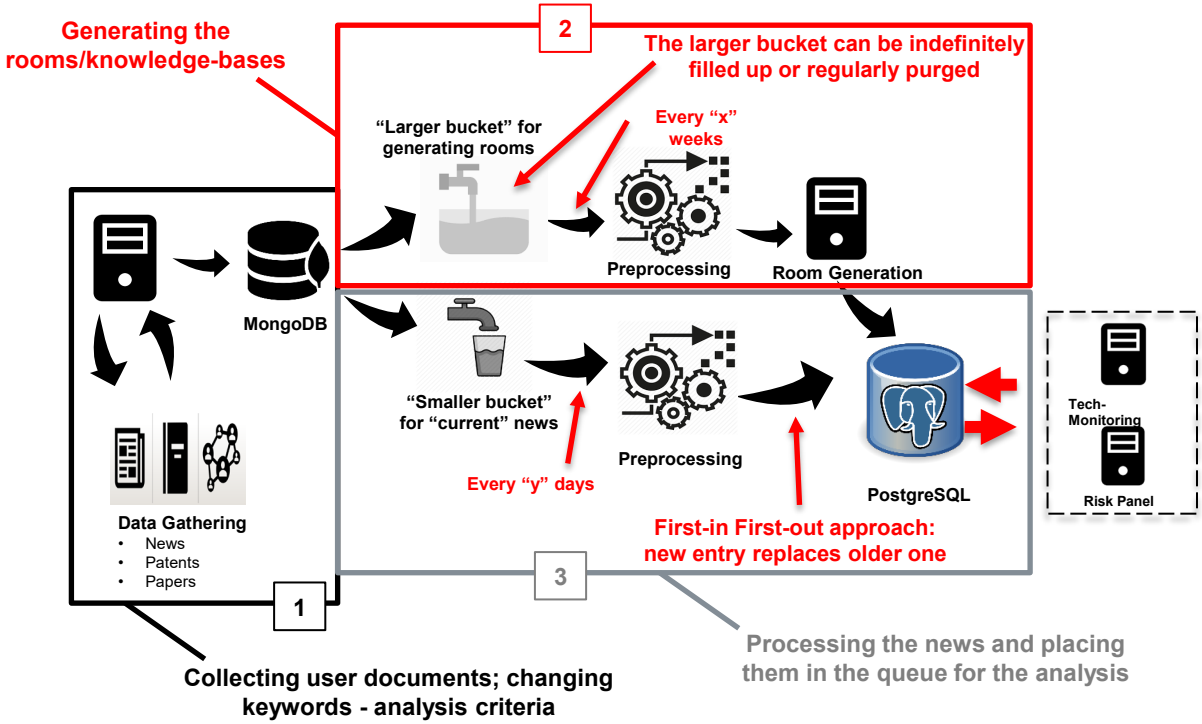
How the "room theory" works



- "**Room theory**" enables the use of context-subjectivity in the analysis of the incoming documents
- Context-subjectivity can be the point of of view of a subject matter expert
- The context-subjectivity in the analysis is represented by a domain specific numerical knowledge base, created from a large domain specific & representative corpus that is then transformed into a numerical dataset ("embeddings table")

- The key components are:
  1. A point of view for the comparison (the "room"). This is represented by the embeddings table extracted from a large/representative corpus from the specific domain
  2. A criteria for the analysis (the "benchmark"). This is a list of keywords defining the "what we are looking for". Different benchmarks would provide different analyses

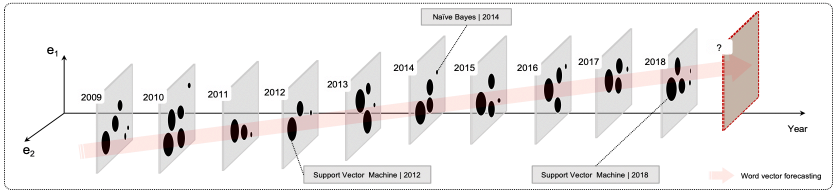
TBDS: Information Flow



TBDS Use case: “Predicting” Technologies

- The future cannot be predicted as such, but in areas such as technology and science, most of the new is based on an evolution of the old
- Leveraging on the “room theory” to provide the point of view, we represent in time the technologies as either points in space or as network of technologies/application

- Using predictive/ML algorithms, we can predict the new interactions and relevance of the technologies/nodes, meaning the new applications for technologies or upcoming technologies



TBDS Use cases: the questions we answer

"What if" analysis on the overall competitive scenario-based arena

- What are the strengths and weaknesses of the players in the competitive arena?
- What are the risks in losing strategic positions from not investing in key technologies?
- What is the return in strategic positioning by investing in a given technology?
- How can I decompose opponents' technologies into their key components (kill chain)?

TBDS Use cases : the questions we answer

Role playing in the competitive scenario-based arena

- What are strengths and weaknesses of my position compared to my competitors?
- What are the technologies that can give me a better return in terms of advantage versus my opponents?
- What are the consequences of disinvesting/reducing the effort in a given technology?
- What are the risks in losing my strategic positions from not investing in key technologies?
- How can I overcome/balance my opponents focusing on technologies in a lower position in the kill chain?

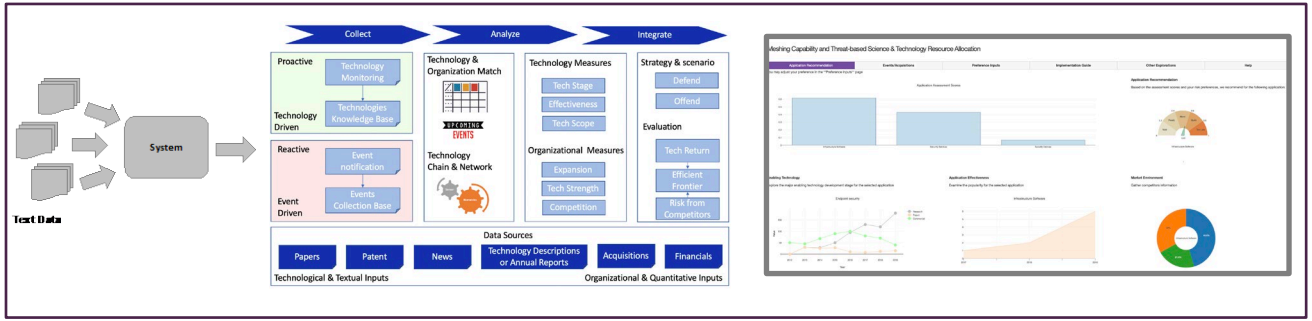
TBDS Use cases : the questions we answer

Horizon scanning:
Emerging and predicted technologies radar screen

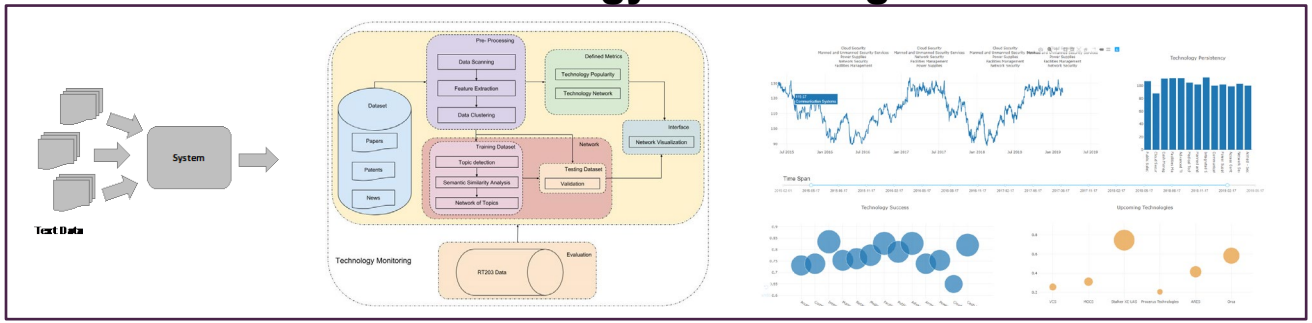
- What are the technologies that are emerging and I may have been missing?
- How the new technologies are related to the older ones?
- How is a given technology in terms of life cycle: is it growing or fading out?
- How a given technology has been applied?
- Considering the past and present technologies, what are the possible technologies in the future?
- Considering the use of the past and present technologies, what are the possible uses of the same technologies in the future?

TBDS - Demo

Risk DSS



Technology Monitoring



TBDS vs commercial Solutions

- TBDS is the result of an applied research and as such is optimized for the specific need
- A System Integration or Consulting companies can do it, but they will focus on maximizing their ROI: reuse what already developed, use lower skilled people and increase the headcount over time
- Off-the-shelf commercial solutions have low degrees of customization, forcing the user to learn their own “language”, that is made with user-locking in mind. Real customizations are special projects
- No available commercial solution is able to analyze text by extracting semantic insights based on a predefined point of view

TBDS – Future Developments

- Supporting the adaptation of TBDS to CCDC-AC/Picatinny mission
  - adjustment to the way Room Theory works on the actual CCDC-AC data by creating a unique CCDC-AC/military point of view
  - development of plug-ins for specific streams of data sources
  - integration of the additional informational content from the new sources
  - redefinition of the competitive scenario
  - adjustment/redefinition of the metrics to better serve the goals of CCDC-AC
  - adjustment and customization of the visualizations
  - development of additional user interfaces for possible uses of subset of TBDS functionalities
  - expand and adapt the kill chain automatic creation

TBDS – Future Developments

- Enhancing the TBDS capabilities
  - Opponent behavior modeling: dynamically position opponents in the range from supportive to non-hostile to hostile
  - Develop a dynamic opponent “technology footprint”, representing their strengths, weaknesses, and the related evolution in time
  - Introduce a time-based view to most of the metrics with additional derived metrics on the trends
  - Work on a “Room Theory 2.0” and developing a triage system to dispatch the incoming document to the most appropriate Room
  - Expand the Technology forecasting module and integrating it with the risk panel
  - Create scenarios for “technology alternative evolution”

What we learned

- Extracting semantic metrics from text is a challenging and widely under addressed task
- Developing a ML/AI system is “developing a system”: system engineering, software engineering and DevOps need to play a central role
- Working on projects with some level of restrictions – both in military or civil environments – in a multinational/open environment requires the creation of a “digital relative”, a proxy domain with available and accessible data semantically close to the target environment
- Developing product-level research in academy environments is challenging, for the different dynamics in place, but it can push the capabilities of the deliverables well beyond what would be feasible outside Academy

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Thank you!

Dr. Carlo Lipizzi
clipizzi@stevens.edu

NLP LAB
https://nlplab.sercuarc.org