

Meshing Capability and Threat-based Science & Technology Resource Allocation

Sponsor: OUSD(R&E) | CCDC

By

Dr. Carlo Lipizzi

11th Annual SERC Sponsor Research Review

November 19, 2019

FHI 360 CONFERENCE CENTER

1825 Connecticut Avenue NW, 8th Floor

Washington, DC 20009

www.sercuarc.org

The Task

- **Title:** Meshing Capability and Threat-based Science & Technology Resource Allocation
- This research is 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 will 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)

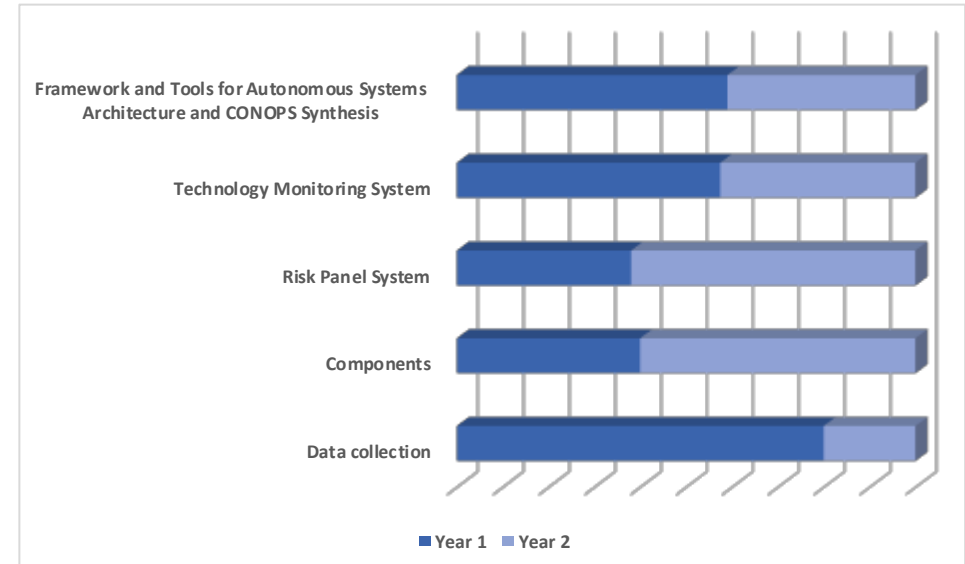
Key features

- **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

Overall view

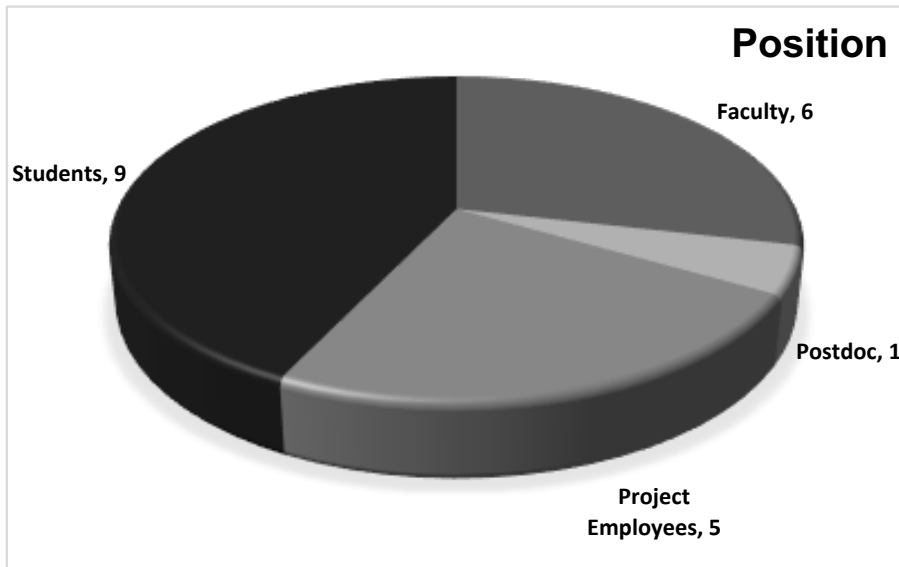
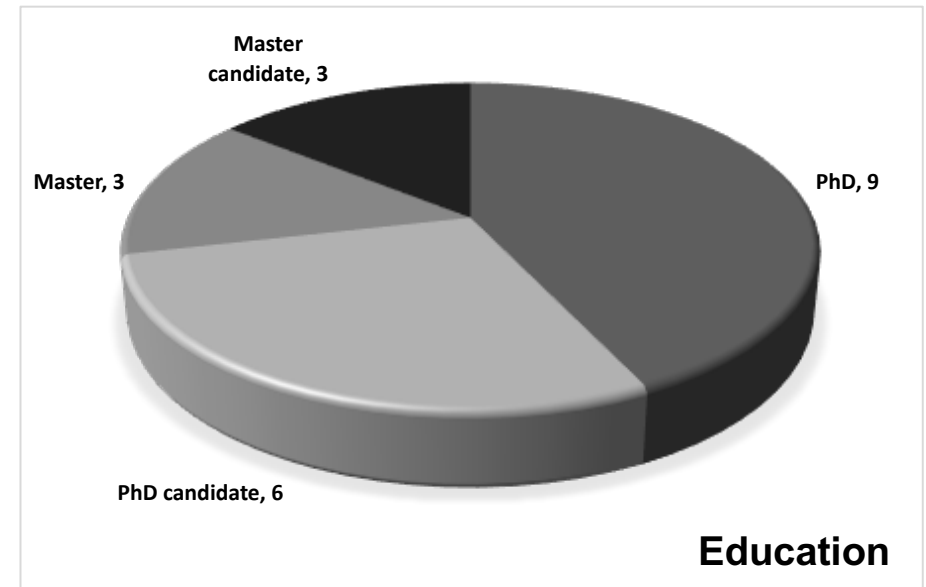
- The project is logically distributed in 2 years, where the 1st year is focused on acquiring the logic used in the planning process, create a corpus to be used for the data/text mining, develop the required components and create proof of concepts for the systems
- The 2nd year is focused adding functionalities 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)	<ul style="list-style-type: none"> • 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 	<ul style="list-style-type: none"> • 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)	<ul style="list-style-type: none"> • Increase the functionalities to the proof of concepts 	<ul style="list-style-type: none"> • Evolve and improve the proof of concepts into a working prototype



Who we are in Phase 2

- **Total number** _____ **21**
 - "Permanent" members _____ **21**
 - $\geq 50\%$ 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

Text -> Metrics: the “room theory”

- The “room theory” is addressing the relativity of the point of view providing a computational representation of the viewer view. 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)
- When we enter into a physical room, we instantly know if it’s a bedroom, a bathroom, a living room
- Rooms/schemata/frameworks ...
 - Are mental framework 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 embeddings tables. 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 a different domain

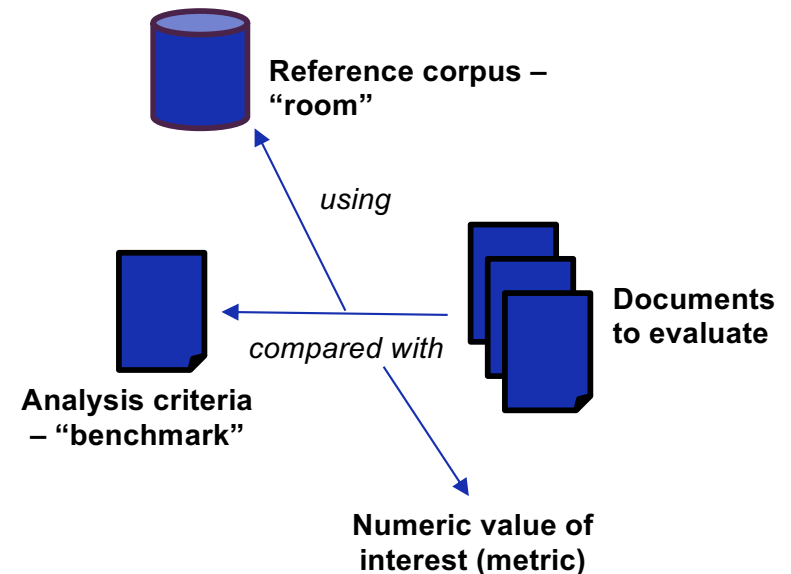
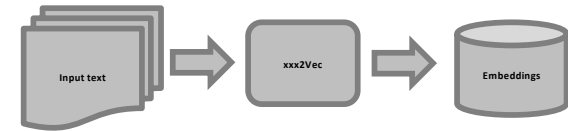
How the “room theory” works

- All the text is transformed into vectors/list of numbers using a text vectorization algorithm (like *Word2Vec*). The result of the transformation is a table (“embeddings”) with all the unique words and a list of numbers per each word.

Vectorization is our enabling technology

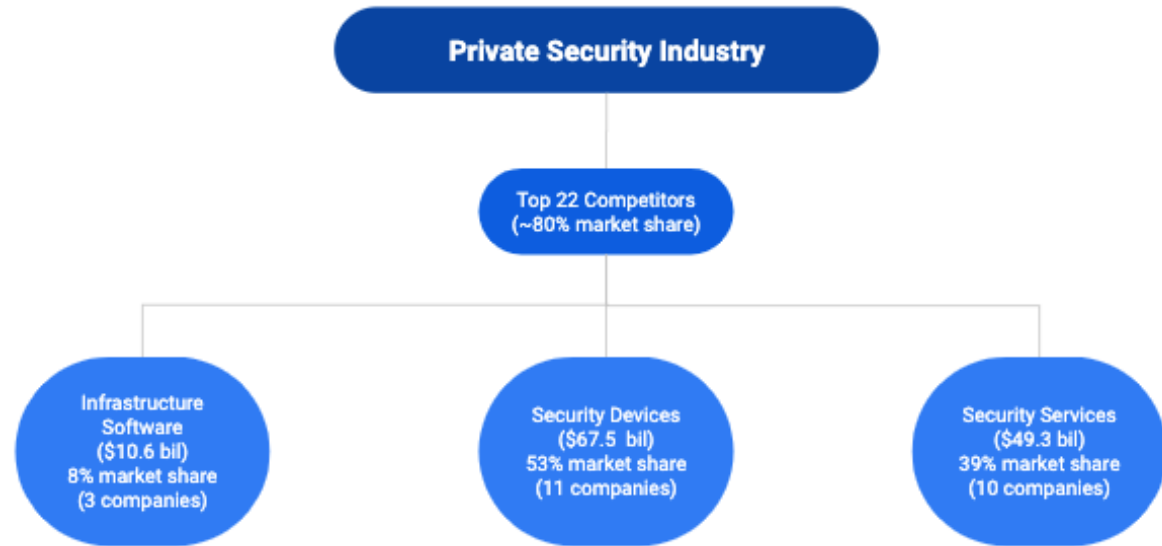
- To take into account the context/subjectivity to analyze the incoming documents, we need:

1. A criteria for the analysis (the “benchmark”). This is represented by a list of words with possible attributes/weights
2. A point of view for the comparison (the “room”). This is represented by an embeddings table extracted from the specific domain





- Large (USA market)
- Technology-driven
- Semantic proximity
- Fully measurable



Security Industry Association Forecasts 2019 Security Megatrends

1. Cybersecurity Impact on Physical Security
2. Internet of Things (IoT) and the Big Data Effect
3. Cloud Computing
4. Workforce Development
5. AI
6. Emphasis on Data Privacy
7. Move to Service Models
8. Security Integrated in Smart Environments
9. Advanced Digital Identities
10. Impact of Consumer Electronics Companies

OUTSOURCED AND IN-HOUSE
SECURITY INDUSTRY

\$44_{bn}

OUTSOURCED CONTRACT
SECURITY INDUSTRY

\$25.5_{bn}

REVENUES FOR THE 4
INDUSTRY LEADERS

\$14.1_{bn}

REVENUES FOR THE 2
MAJORITY FOREIGN-OWNED
INDUSTRY LEADERS

\$7.1_{bn}



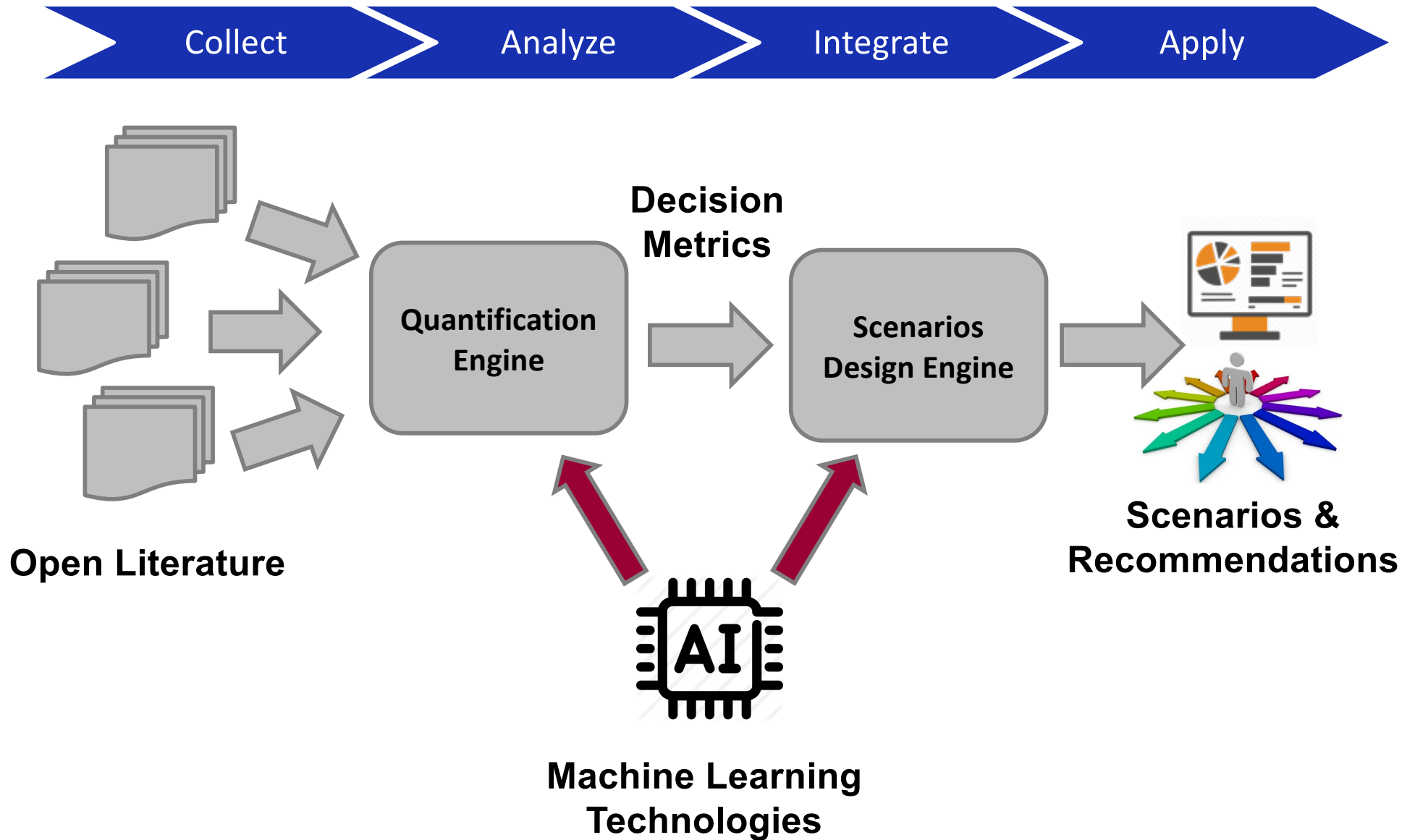
- The proxy domain has:
 - Technology-driven approach
 - Semantic proximity
 - Generalized 3 layers approach: companies, market segments, technologies
 - Metrics reusability. It uses: market shares (for overall market and market segments); technology past/estimated value (for companies and market segments); technology coverage (for companies and market segments); vulnerability/overtaking (for companies within market segments); overall market equilibrium; technology chain
 - Scenarios reusability: status quo vs attack (for overall market and market segments)
 - Dynamic analysis

How we develop

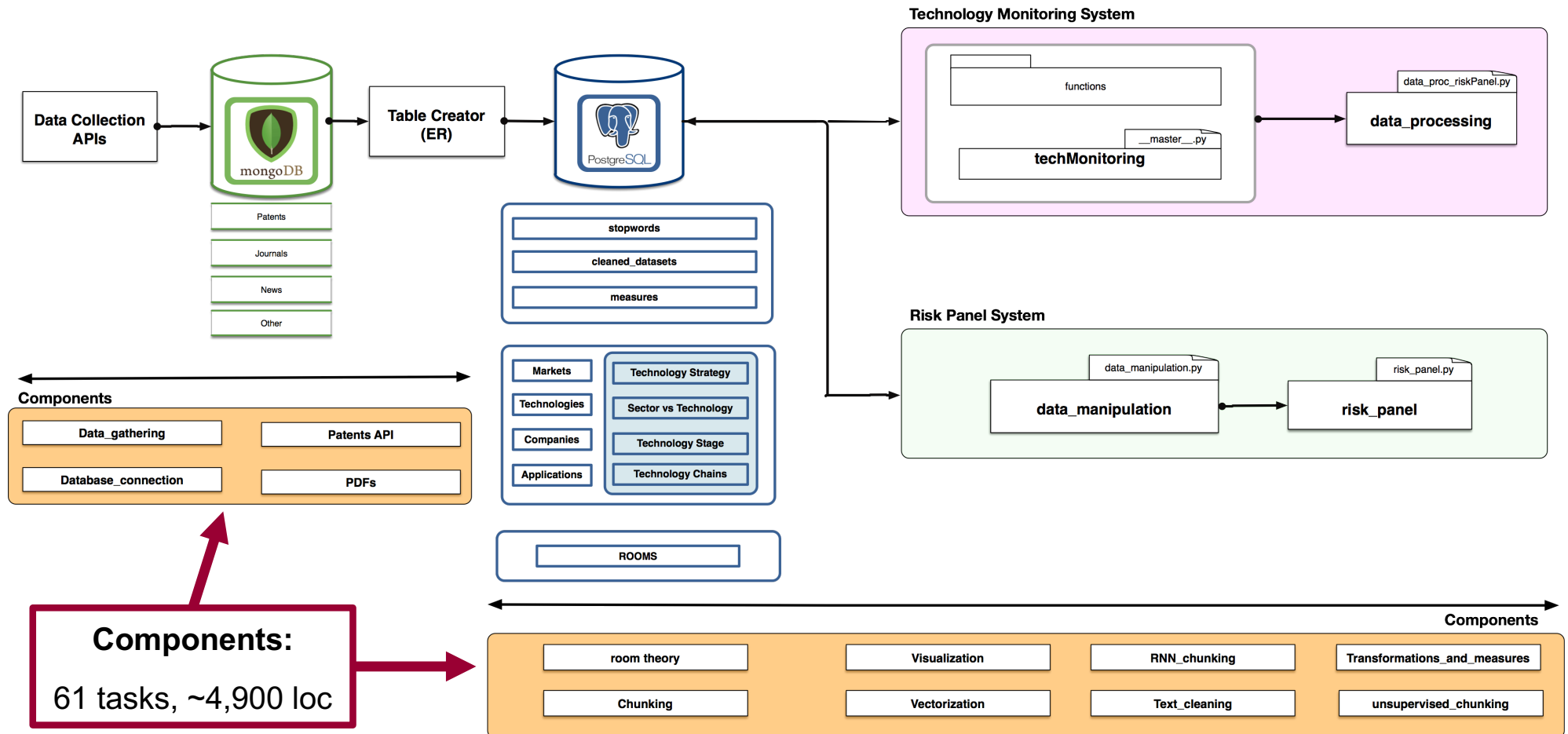


- The “Lego” approach
- Systems are developed as **growing prototypes** based on **components** developed offline
- Systems project leaders design them and build an integration layer to integrate components
- Components cover the tasks the systems have in common plus some system-specific tasks with higher level of complexity
- Components team (*NLP lab*) has dedicated team members; 61 tasks, ~4,900 loc
- Use of components for other CCDCAC activities

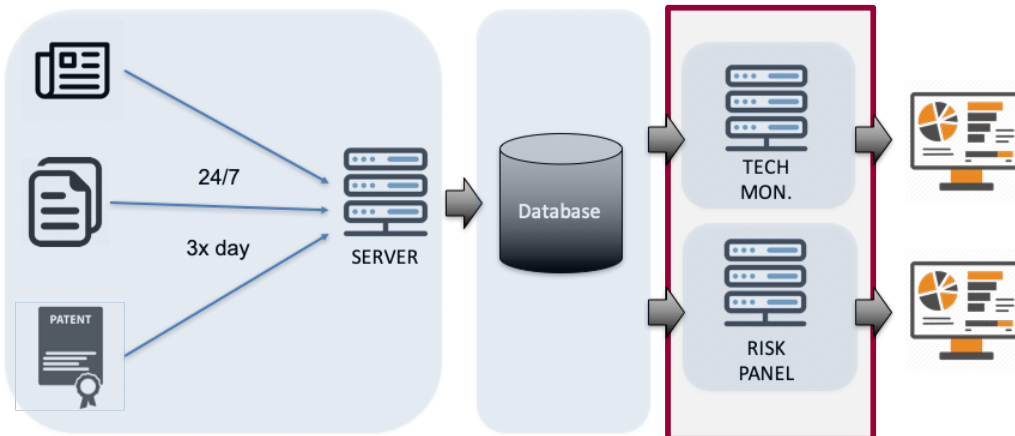
Logical Architecture



The components



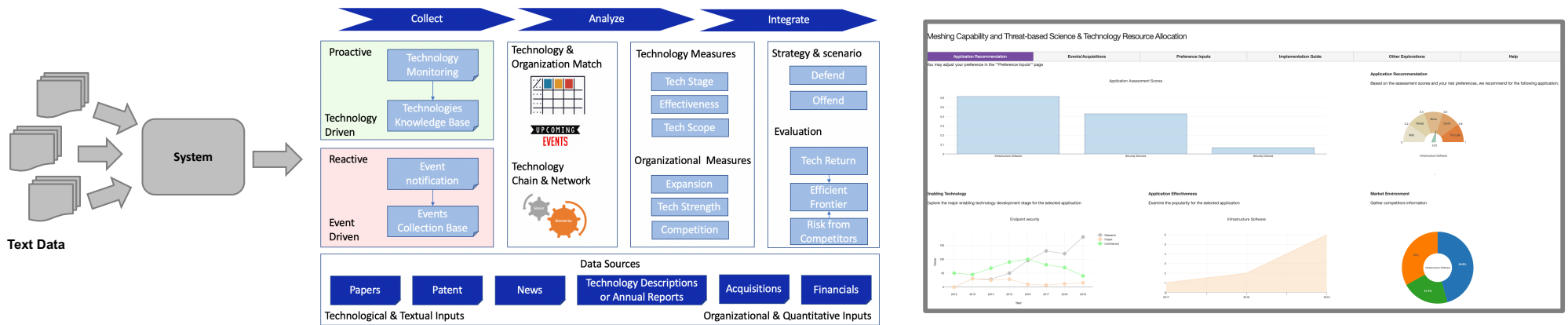
The systems



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

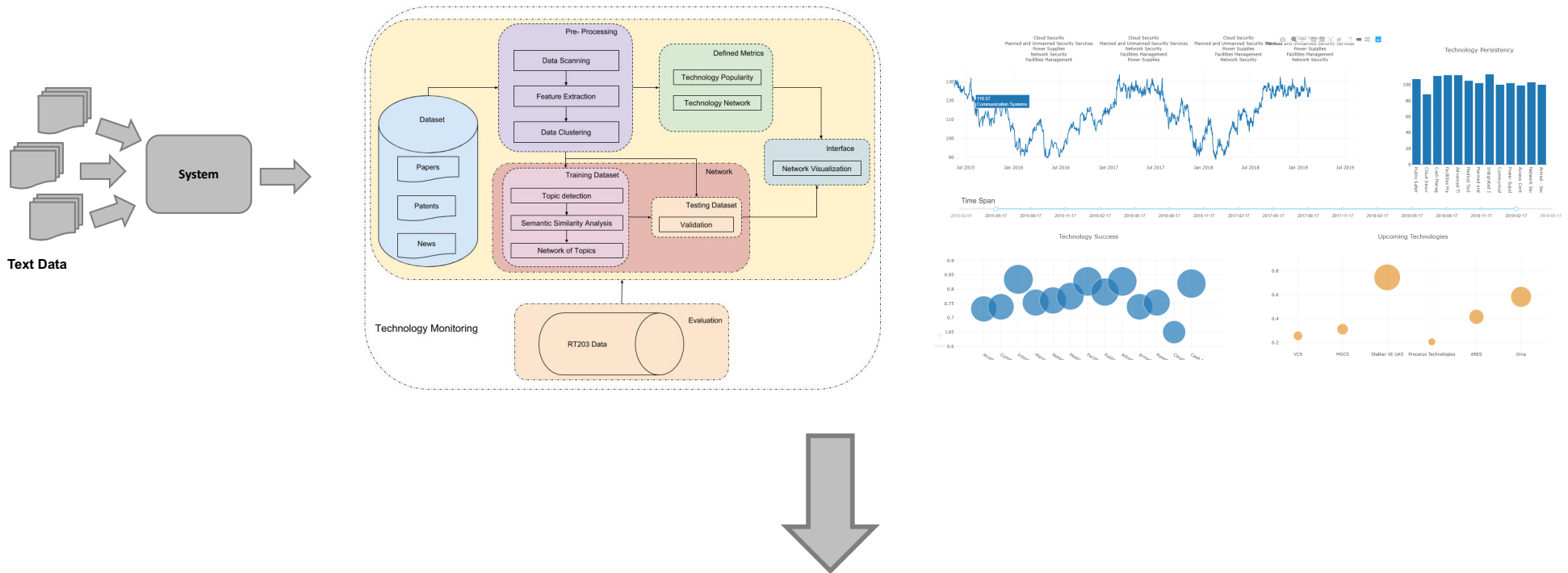
- Risk panel – Planning Decision Support System.** Primary outcome from this system is an interactive panel that can be used for all the what-if analyses, with recommendation layer based on Machine Learning trained based on user-defined “optimal” scenarios
- Technology Monitoring System.** This system looks for emerging technology with potential future impact on the overall scenario. It will provide the Sponsor with a way to be prepared for (or plan for) future technologies that may have an impact on their activities

Risk DSS



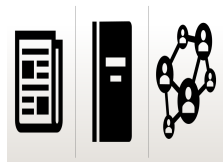
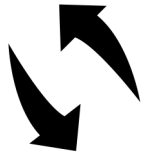
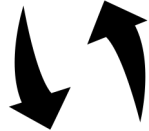
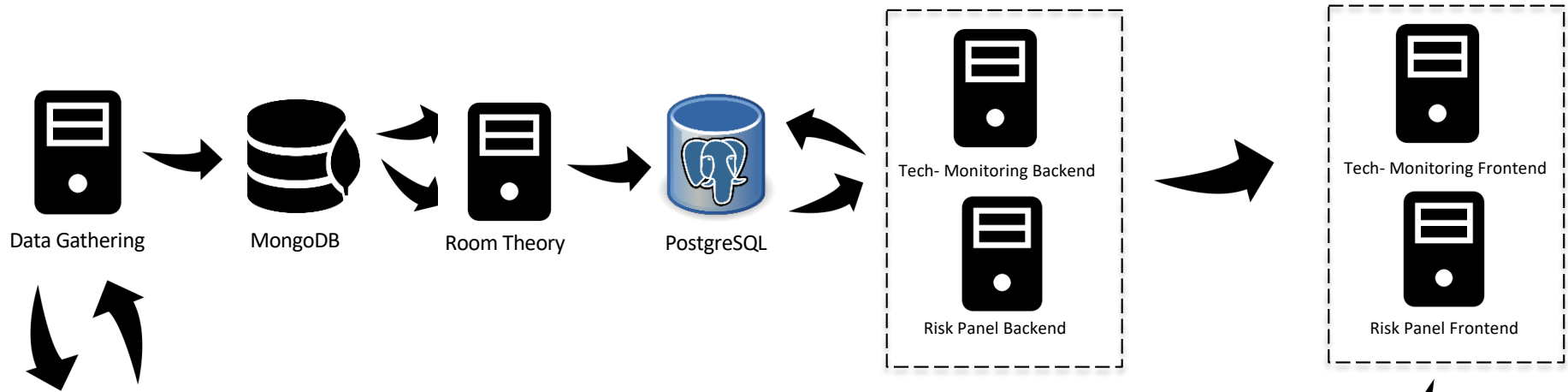
Text-based risk decision support elements for scenarios creation

Technology monitoring




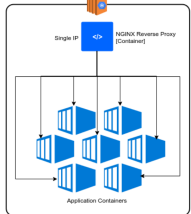
A radar screen for coming and “future” technologies, along with a technology taxonomy generator

Infrastructure



Papers, news & social media

- **Docker** allows developers to package up applications with all the parts needed and ship it all out as one package
- **NGINX** Adds a layer of security and anonymity


“By-products”

- We worked on papers focused on specific aspects of our research
 - *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”. Expert Systems Journal (in-progress).*

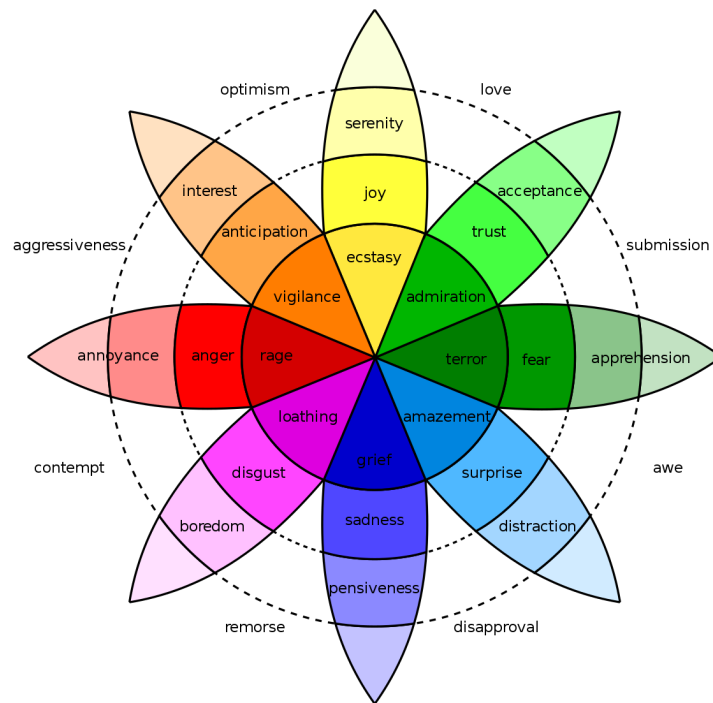
Coming Features

Emotional Reaction to products

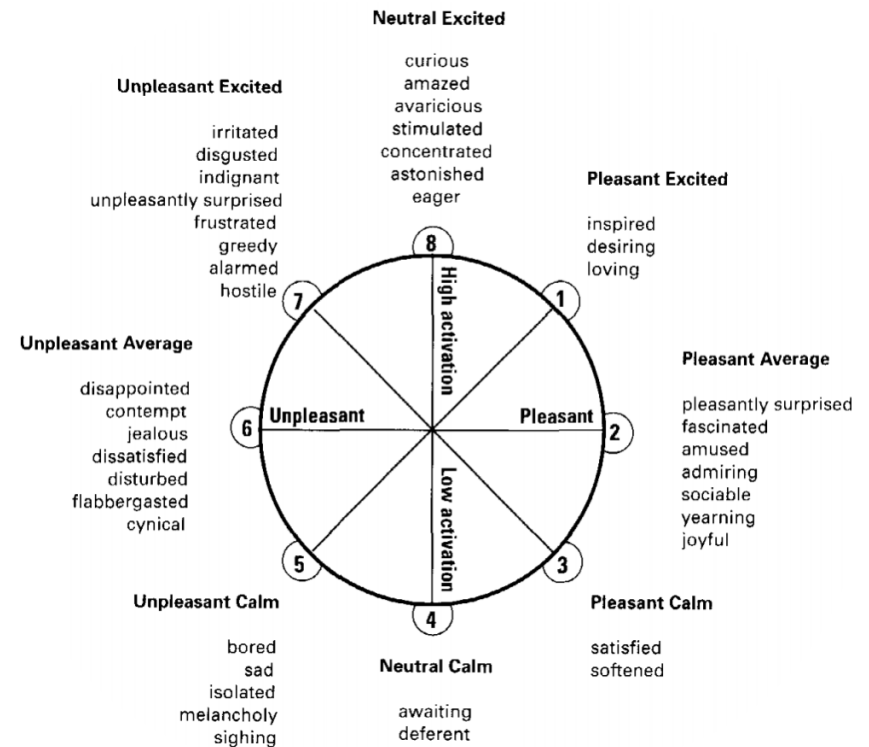
- This is going to be an additional metric for the Risk Panel
- The problem we are trying to address here is how to consider a subjective/user evaluation of a technology. Some technologies or products may be great from the pure technical standpoint, but not successfully applicable in the target domain. VHS vs Betamax is a typical example, some can be from military domain
- It is what we believe can be a logically equivalent to cumulative on-the-field experts opinion reports

Coming Features

Emotional Reaction to products is not the same as Emotion Reaction to “life events”



Reactions to “life events”



Reactions to products

Emotional Reaction to products

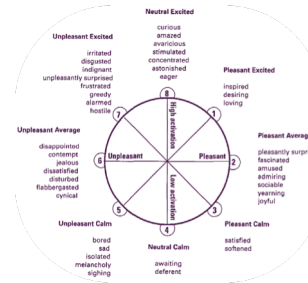


Opinionated text
data

NEWSPAPERS
SPECIALIZED BLOGS



Text Analytics
+
Room Theory



Detect
Product-related
Emotions



Extract METRICS
from results

- The metric extracted from the emotion evaluation is based on the same stream of document we use for the 2 systems and will provide an additional element for the analyses in the systems

Technology “forecast”

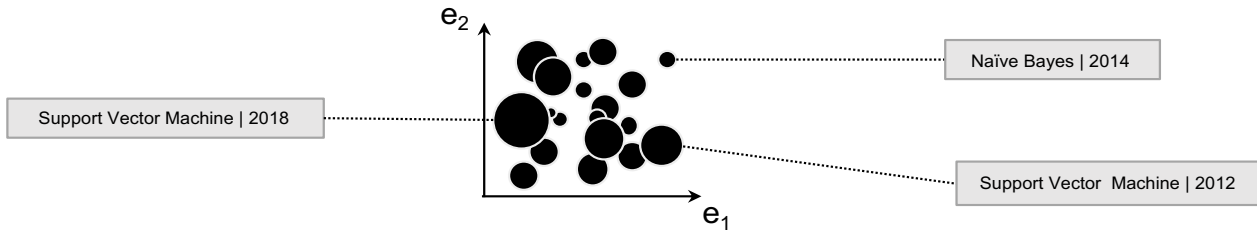
- Future cannot be predicted as such, but in some areas, such as technology and science, most of the new is based on an evolution of the old
- Our overall approach for this project is based on transforming text into numbers, creating matrixes (“embeddings”) representing a numerical knowledge base of the specific domain
- Collecting over time news, patents, papers, blogs on the topic (technology, in this case) we have a text dataset, that we transform into the numerical knowledge base/matrix, where each technology/year has a row with n (we use 300) numbers, defining it
- The matrix represents the n -dimensional space for the specific domain (technology in this case), where each point is a technology/year
- Some of the technologies evolve in time in a “coherent” way. This “coherence” may be extrapolated in time, determining an area/set of points in the future
- We then determine what are the current technologies more related/closer to that “future” set of points. This will provide a “description” of the future technologies either in a narrative form or as a semantic network

Technology “forecast”

Machine Learning papers dataset (10 years)



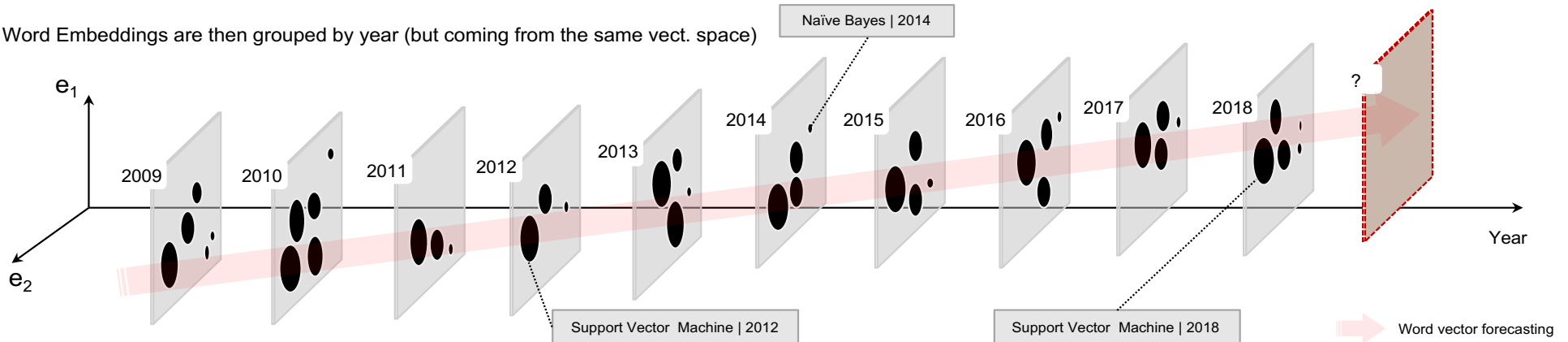
Word Embeddings
(with tag of the year - single embedding model)



↕ Vector space ● Word/chunk vectors

Note: 2-dimensional space for easier representation, case study is 300-dimensional

Word Embeddings are then grouped by year (but coming from the same vect. space)

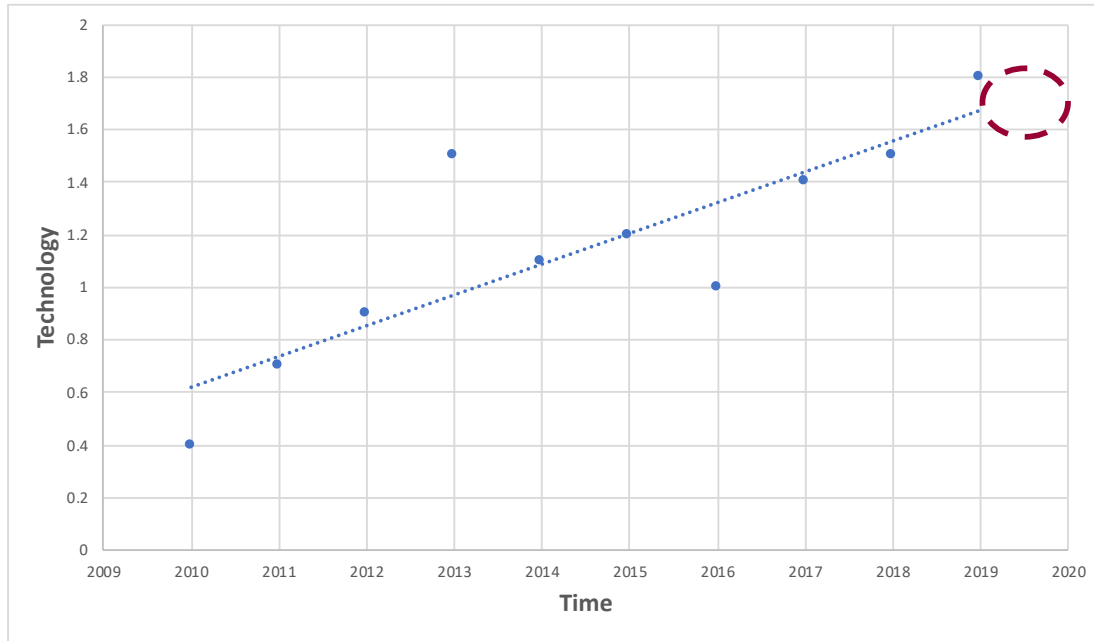


Coming Features

Technology “alternate evolutions”

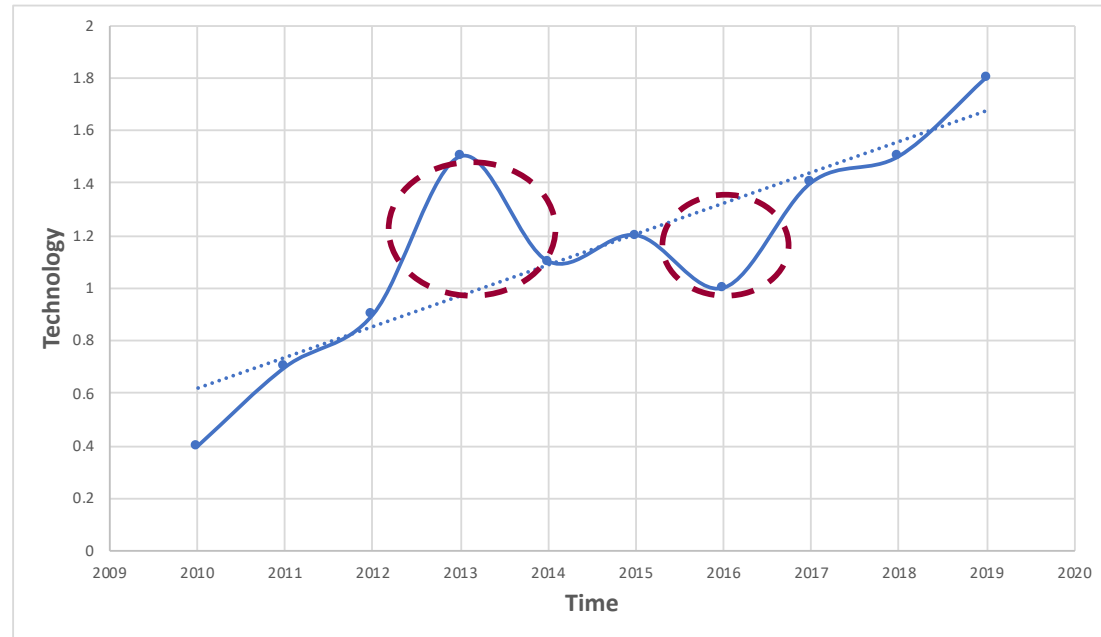
- The points in the the n -dimensional space described before are collected from public sources, meaning publicly available information on specific technologies
- Some companies (or equivalent entities) may not publicly disclose their developments, being of strategic value for them
- Nevertheless, those developments would most likely be based on public domain technologies/existing points in the space
- This divergency may reflect into the distance between the actual points and a linear trend. The gap would be – for a given time interval – between the interpolating curve and the linear regression
- When that distance is above a certain average (measured for example in number of standard deviations), we can start focus on the gap area and determine what are the current technologies that are more related/closer to the points in that area

Coming Features



Technology “forecast”

Technology “alternate evolutions”



“Quantum” algorithms

- Our overall approach is based on transforming text into sequences of numbers. The text will become an n -dimensional Euclidian space. This is providing a large but fixed representation of the text
- Expanding the Euclidian space by having functions instead of points, we may have a better accuracy in analyzing the text
- Those functional space/Hilbert space are a generalization of Euclidian spaces and are commonly used for quantum algorithms



SYSTEMS
ENGINEERING
RESEARCH CENTER



STEVENS
INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

Thank you!

Dr. Carlo Lipizzi
clipizzi@stevens.edu



NLP
L A B