Naval Information Warfare Center
PACIFIC

TECHNICAL REPORT 3265
FEBRUARY 2022

2021 Data Challenge: Naval Information Warfare Center Pacific Innovation and Results

Dean Lee
Vincent Siu
Dr. Benjamin Michlin
Jeffrey Bennett
Josh Duclos
Jazlynn Wied

NIWC Pacific

Naval Information Warfare Center Pacific (NIWC Pacific)
San Diego, CA 92152-5001

This page is intentionally blank.

TECHNICAL REPORT 3265
FEBRUARY 2022

2021 Data Challenge: Naval Information Warfare Center Pacific Innovation and Results

Dean Lee
Vincent Siu
Dr. Benjamin Michlin
Jeffrey Bennett
Josh Duclos
Jazlynn Wied

NIWC Pacific

Administrative Notes:
This report was approved through the Release of Scientific and Technical Information (RSTI) process in September 2021 and formally published in the Defense Technical Information Center (DTIC) in February 2022.

NIWC Pacific
San Diego, CA 92152-5001

NIWC Pacific
San Diego, California 92152-5001

A. D. Gainer, CAPT, USN
Commanding Officer

W. R. Bonwit
Executive Director

ADMINISTRATIVE INFORMATION

The work described in this report was performed by the IT Engineering and Support Branch of the Business and Enterprise Information Systems Division, and of the Command & Control and Enterprise Engineering Department, Naval Information Warfare Center Pacific (NIWC Pacific), San Diego, CA. The NIWC Pacific Portfolio Council provided funding for this project.

Released by
Donna Williamson, Division Head
Business and Enterprise Information Systems Division

Under authority of
Scott Crellin, Department Head
Command & Control and Enterprise Engineering Department

Editor: RJP

EXECUTIVE SUMMARY

The 2021 Department of Navy Data Challenge asked teams from across the Department of Navy to find insights from Federal government contracts opportunities data. As part of the process, Naval Information Warfare Center (NIWC) Pacific designed a novel machine learning pipeline that incorporates methods fromNatural Language Processing and Topological Data Analysis for insight extraction. Additionally, NIWC Pacific developed a vulnerability assessment metric using information extracted from contracts award data.The results from the challenge are displayed through a prototype dashboard.

This page is intentionally blank.

CONTENTS

Figures

This page is intentionally blank.

1. INTRODUCTION

The 2021 Department of Navy Data Challenge commenced on March 2021 for approximately three months. The Challenge asked each team to examine the Federal government contracts opportunities data[1] and find non-intuitive insights. As the objective of the Challenge is overly broad, we refined the scope of the challenge to the following objectives:

1. identify the work being done by the various Naval Research and Development Establishment (NR&DE) Systems Commands (SYSCOMs), Systems Centers (SCs), and Warfare Centers (WCs);
2. identify the overlap as well as the unique work being done at each SYSCOM, SC, and WC;
3. identify the trends in focus areas from the work being performed each SYSCOM, SC, and WC;
4. identify information from the data sets that can be leveraged by potential adversaries.

For Objectives 1-3, we focus only on the contracts opportunities data that pertain to the NR&DE and its constituent organizations. The scope of Objective 4 covers all organizations within the Federal government.

This page is intentionally blank.

2. APPROACH

For Objectives 1-3, we developed a novel machine learning pipeline to extract insights. The pipeline starts with the preprocessing of the contracts opportunities data, followed by the application of Natural Language Processing (NLP) to identify and extract salient features. These features are then used as input to Topological Data Analysis (TDA) to transform the features into graphical representations to uncover insights. Finally, the insights are displayed through an interactive dashboard to enable insight exploration by potential stakeholders.

For Objective 4, we looked in the title, description, award amount, and awardee fields across the entire data set to determine if there exist information that can be readily leveraged by potential adversaries.

2.1 DATA CONDITIONING

The Federal government contract opportunities data was initially filtered to NR&DE commands by manually mapping the values in the Office field. Some basic preprocessing was performed, such as using string matching to remove special characters, canceled contracts, and duplicated contracts. Furthermore, as some contracts are awarded in various regional currencies, the contract award values are normalized to US dollars. Due to inconsistent documentation of contract extension and ceiling information, we made best effort to extract contract ceiling information where possible so that the comparison of contract values is fair.

A focus was made to use the free text fields in the data, such as the title and description, to extract insights; unfortunately, most of the text had boilerplate contract language which obfuscates the projects' missions. Upon further examination, it was found that about 85% of the data had some form of contract language. We sought to use various text summarization methods to condition the text so that it only contained relevant information, and a list of common contract phrases was created to evaluate the effectiveness of our methods. A heuristic was used to choose the best summarization results between LexRank [1], LSA [2], and TextRank [3] for each record that contained the least contract language. The application of text summarization greatly reduced the amount of irrelevant texts in the data.

The following NLP methods were Figure to further condition the text data:

Bigram Construction: construct two-word phrases from the corpus;

Stopword Removal: manual removal of certain words, such as location, command, contract numbers, acronyms, etc.;

Lemmatization of Words: create a baseline of words by extracting inflected form of words;

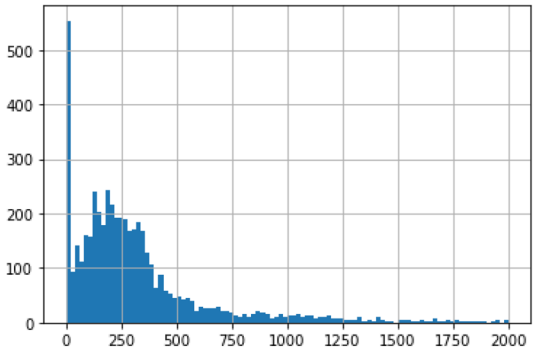
Parts of Sentence (POS) Conditioning: keep only nouns, adjectives, verbs, and adverbs in the text data;

Named Entity Recognition (NER) Conditioning: remove entities, such as person, location, language, etc., from the text data;

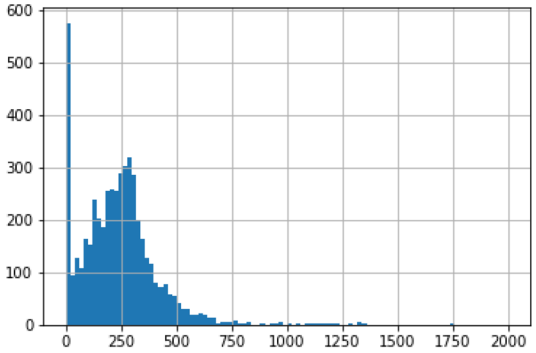
Dictionary Filtering: filter words that are insignificant and provide no value.

word

The distributions of documents by word-count before and after the NLP conditioning are shown in Figure 1.



(a) Distribution of documents by word-count before NLP conditioning.



(b) Distribution of documents by word-count after NLP conditioning.

Figure 1. The distributions of documents by word-count before and after NLP conditioning.

2.2 NLP FEATURE EXTRACTION

Topic modelling is an unsupervised NLP method that automates the organization, understanding, and summarization of large collections of unstructured documents. The Latent Dirichlet Allocation (LDA) [4] method, a generative statistical model for topic modelling, was used in this effort. LDA assumes that each document is a mixture of topics, and that each topic is a mixture of words. The output of a LDA model includes topic distributions for a given document and relevant topics for a given word.

Topic coherence was used to select the optimal number of topics and determine the quality of the topics from LDA. In particular, the coherence score measures how often the topic words appear together in the corpus; by plotting the coherence score against the a specific number of topics from the model, the optimal number of topics in the corpus can be determined heuristically by locating the peak value prior to flattening out. See Figure 2, and note that the optimal number of topics in the figure is around 25.

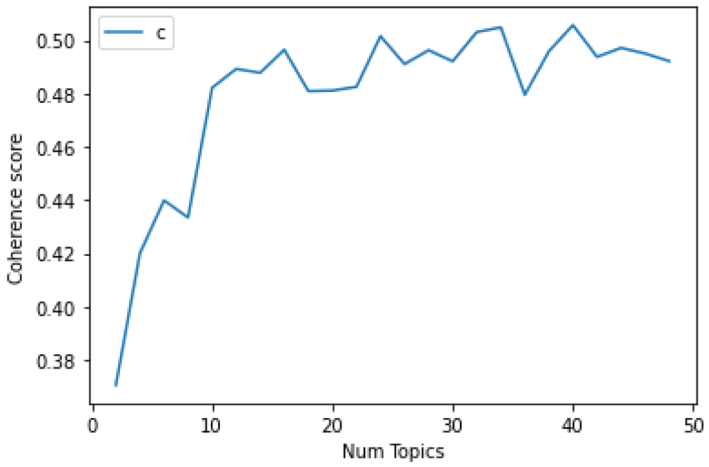


Figure 2. The number of topics and the coherence scores. It can be seen heuristically that the optimal number of topics should be between 20 and 30.

We then inspected the top words in each topic generated by the LDA model, and found that the model grouped together contractual and financial words. This resulted in several iterations of the conditioning methods described in Section 2.1 until the coherence and the top words produced satisfactory results.

Finally, the extraction of topic distribution was added as another step in the data pipeline, as the incorporation of the distribution information seem to improve the quality of the insights even more.

2.3 TOPOLOGICAL DATA ANALYSIS

Topological Data Analysis (TDA) is a set of techniques based in algebra and topology. Mapper [5] is a TDA algorithm that condenses high dimensional data into an undirected graph, while preserving the topological structure of the original data. The resulting graphs can be used to identify non-obvious insights from the data.

Mapper is made up of three stages: 1) filter, 2) cover, and 3) cluster. In each stage, a function is chosen, and the stages are applied successively. Determining the appropriate choice of functions for each stage is an iterative process until a suitable graph is found. For the data challenge, the functions were chosen so that the similarities between the work performed at the various SYSCOMs and Warfare Centers would be made obvious through Mapper.

More specifically, the geodesic distance of the Mapper graph should describe the similarity in work performed between the organizations represented by the graph nodes. For example, if two organizations perform highly similar work, their respective nodes should be direct neighbors of each other; and the more dissimilar the work, the more edges between the two nodes. Through experimentation, it was determined that the eccentricity function [6], together with the cubical cover [7] and DBSCAN [8] as the filter, cover, and cluster functions, respectively, help guide Mapper to produce visually intuitive graphs.

2.4 VULNERABILITY ASSESSMENT

2.4.1 Project Classification Labels

One of the most surprising discovery in the contracts opportunities data is the existence of project classification labels. Out of the projects which have classification labels, we further determined if these contracts have awardee name and address, and technical point of contact (TPOC) information. These information, when combined with the contract award amount, may indicate location of classified labs, as well as the type of the classified work being performed. We designed the following vulnerability metric to assess the vulnerability of a project to espionage:

$$\mathsf{VulnerabilityScore} = \mathsf{ClassificationScore} + \mathsf{AddressScore} + \mathsf{TPOCScore}, \tag{1}$$

where

$$\mathsf{ClassificationScore} = \begin{cases} 0 & \text{if a project has no classification labels,} \\ 20 & \text{if a project is labeled as Confidential,} \\ 40 & \text{if a project is labeled as Secret,} \\ 60 & \text{if a project is labeled as Top Secret;} \end{cases}$$

$$\mathsf{AddressScore} = \begin{cases} 20 & \text{if a project has a physical address in the data,} \\ 0 & \text{otherwise;} \end{cases}$$

$$\mathsf{TPOCScore} = \begin{cases} 20 & \text{if the data contains project TPOC information,} \\ 0 & \text{otherwise.} \end{cases}$$

2.4.2 SolarWinds Breach

The SolarWinds hack recently made headlines due to the severity of the breach. While the extent of the exploit was never publicized, we investigated the FY2019 and FY2020 contract awards with SolarWinds to identify organizations that may still be vulnerable.

3. RESULTS

3.1 OBJECTIVES 1-3

To answer Objectives 1 and 2, the Mapper algorithm is applied to the embeddings generated from the NLP feature extraction steps. See Figure 3 for the Mapper output. It is immediately clear that there are work that is unique to each SYSCOM, as well as similar work that is performed across multiple SYSCOMs. For example, NAVWAR is the only SYSCOM performing work in topic ID 1, which are made up of keywords that indicate information warfare and command and control. On the other hand, Figure 3 shows that NAVAIR and NAVSEA overlap in much of their work.

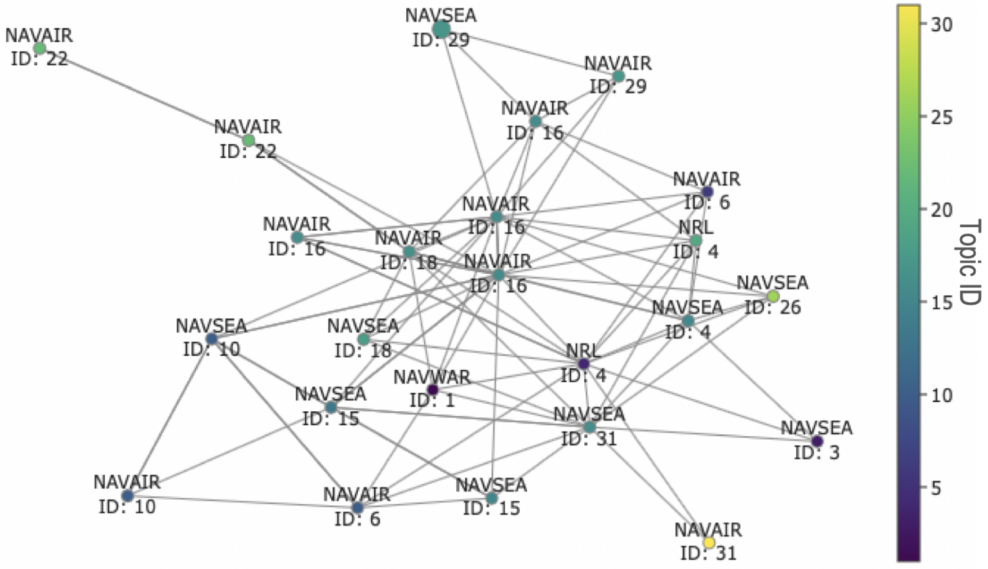


Figure 3. An undirected graph that illustrates the unique work being done at each SYSCOM, as well as similar work that is being done across the multiple SYSCOMs. Each node in the graph is labeled by the representative SYSCOM, as well as the dominant topic ID for that node.

To answer Objective 3, a manual mapping is created between the 2018 Modernization Priorities (MP) [9] and the topics created from NLP topic models. This step provides contextual information to the extracted topics. Mapper is applied to the NLP embeddings in 5 year increments, starting in FY2001 and ending in FY2020. Figure A-1 in Appendix A illustrates not only a shift in focus from Laser and Autonomy to AI, Autonomy, and Space, but also serves to identify the particular SYSCOMs doing work in these areas over the years.

To get a better sense of the investment in each of the MP focus areas, we investigate the contract awards and the operating costs at each of the Systems Centers from FY2001 to FY2020. FigureA-2 in Appendix A shows the total award amount, as well as the award amount relative to the operating costs each Systems Center. The shift in investment over the years is clear for all the Systems Centers. It is also striking to see that the contract awards in the MP focus areas make up only a small portion of the operating costs at each Systems Center.

3.2 OBJECTIVE 4

3.2.1 Project Classification Labels

We restricted the search for classification labels to contracts awarded by the NR&DE commands, and determined that roughly 13% of the projects have classification labels, while 12% have awardee name and address, and 2% have TPOC contact info, such as name, email, or phone number. These results seen in Figure B-1 in Appendix B.

Based on these results, we perform a vulnerability assessment using Equation (1) on the NR&DE commands. Figure 4 shows vulnerability scores at the NR&DE SYSCOMs, and more importantly, demonstrates how the vulnerability scores quantify the amount of exploitable information in the contracts award data.

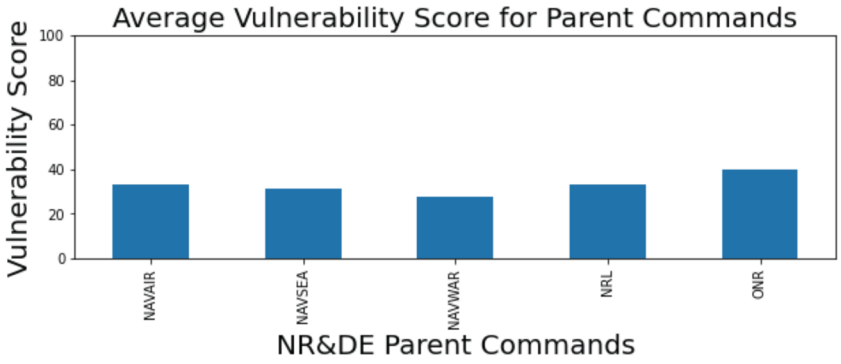
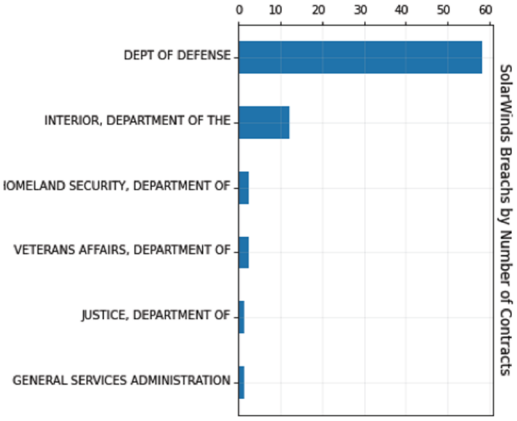


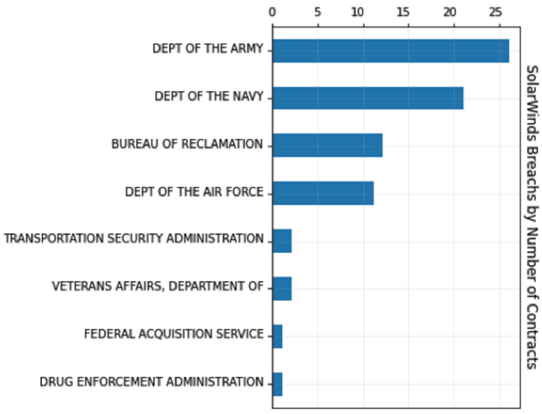
Figure 4. Vulnerability assessment of NR&DE commands.

3.2.2 SolarWinds Breach

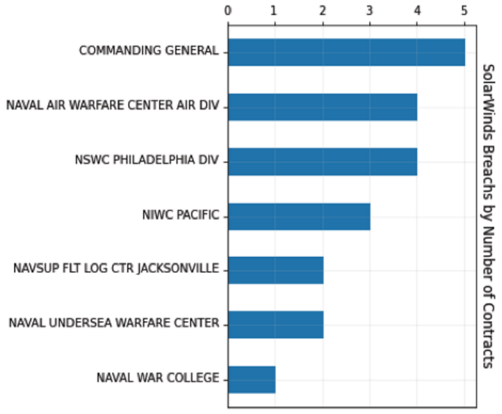
Finally, we look at the FY2019 and FY2020 contract awards with SolarWinds to determine (albeit circumstantially) the scope of the SolarWinds breach in the Federal government. As can be seen in Figure 5, the Department of Defense holds the most number of contracts with SolarWinds; and within the DoD, Army and Navy have the majority of these DoD contracts. It can be surmised that Army and Navy were the most vulnerable to the SolarWinds hack.



(a) SolarWinds contracts by Federal government departments.



(b) SolarWinds contracts by DoD departments.



(c) SolarWinds contracts by Navy commands.

Figure 5. SolarWinds contracts search results.

This page is intentionally blank.

4. PROTOTYPE DASHBOARD

While the duration of the challenge was short, we believe that the machine learning pipeline and the results from the challenge would be of interest to stakeholders. Thus a prototype dashboard Figure C-1 in Appendix C) was built to demonstrate how the results can be used by stakeholders. The dashboard is designed with an intuitive user-interface for insight exploration. Furthermore, the dashboard is connected to the machine learning pipeline, so the dashboard could be used for real-time analyses if there were real-time data sources made available.

This page is intentionally blank.

5. CONCLUSION

In the 2021 Department of Navy Data Challenge, participants were asked to uncover insights from the contracts opportunities data. To meet this objective, we designed a novel machine learning pipeline that integrated state of the art techniques from NLP and TDA to extract insights from the contracts data. We also aggregated data from the 2018 Modernization Priority and the annual NR&DE operating costs to provide additional interpretation to the extracted insights. Specifically, the Modernization Priority added a framework to find similar work being performed across the NR&DE; the annual operating cost information provides a measure of investment in different technical focus areas year over year.

Separately, we analyzed the contracts data to identify information that may be leveraged by potential adversaries. We discovered that classification labels can be found in the contracts data, and developed a vulnerability metric to perform vulnerability assessment based on the presence of classification labels, physical address, TPOC contact info, and other crucial pieces of information. Additionally, we used the SolarWinds breach as an exemplar to show that contracts information can be leveraged to exploit known vulnerabilities.

Finally, all the results are brought together in an intuitive dashboard that may be further developed for use by stakeholders.

This page is intentionally blank.

REFERENCES

1. Erkan, G. and Radev, D. 2011. "LexRank: Graph-based Lexical Centrality As Salience in Text Summarization," *Journal of Artificial Intelligence Research - JAIR*, vol. 22.

2. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. "Indexing by Latent Semantic Analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, vol. 41, pp. 391–407.

3. Mihalcea, R. and Tarau, P. 2004. "TextRank: Bringing Order into Texts," *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

4. Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, URL `http://portal.acm.org/citation.cfm?id=944937`.

5. Chazal, F. and Michel, B. 2021. "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists," .

6. Singh, G., Memoli, F., and Carlsson, G. 2007. "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition," M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, eds., *Eurographics Symposium on Point-Based Graphics*, The Eurographics Association.

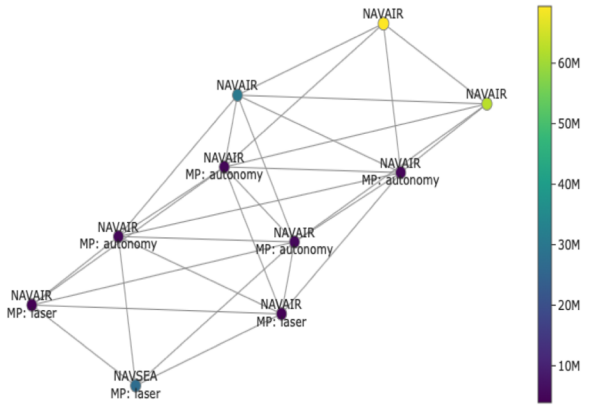
7. Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A., Dassatti, A., and Hess, K. 2020. "giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration," .

8. Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231).

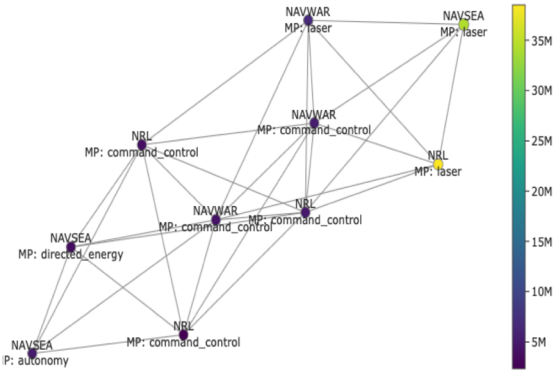
9. U.S. Department of Defense. 2018. "2018 National Defense Strategy of the United States of America," `https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf`.

This page is intentionally blank.

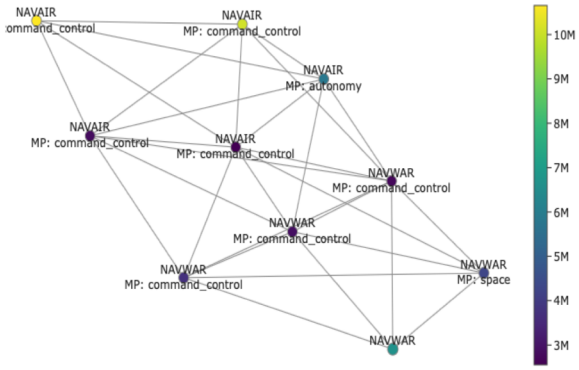
APPENDIX A
OBJECTIVES 1 to 3 RESULTS



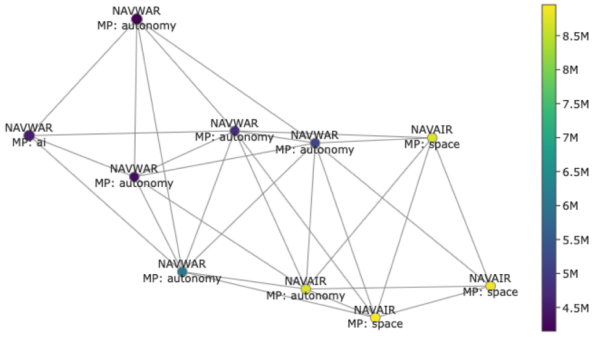
(a) FY2001-FY2005. The dominant MP top-ics are laser and autonomy.



(b) FY2006-FY2010. The dominant MP top-ics are command and control, laser, and autonomy.



(c) FY2011-FY2015. The dominant MP top-ics are command and control, autonomy, and space.



(d) FY2016-FY2020. The dominant MP top-ics are AI, autonomy, and space.

Figure A-1. Modernization Priorities from FY2001 to FY2020 in 5 year increments. The color scale indicates the contract award amount in U.S. dollars.

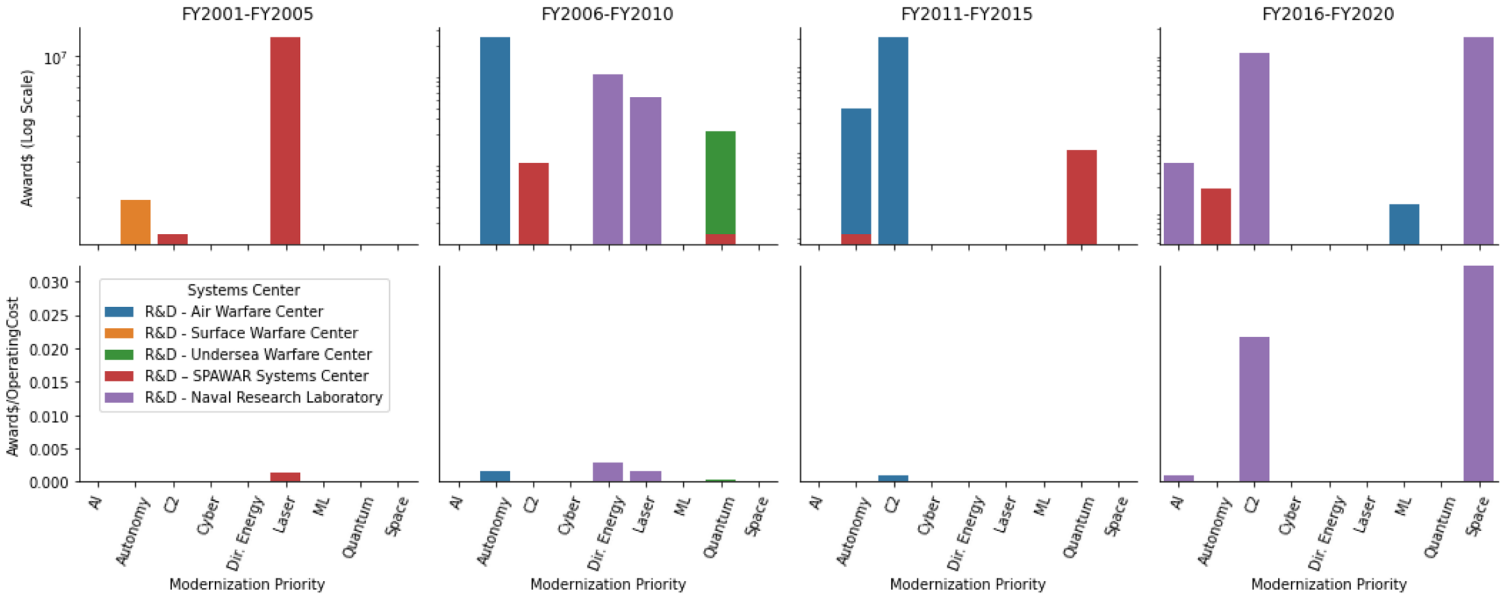
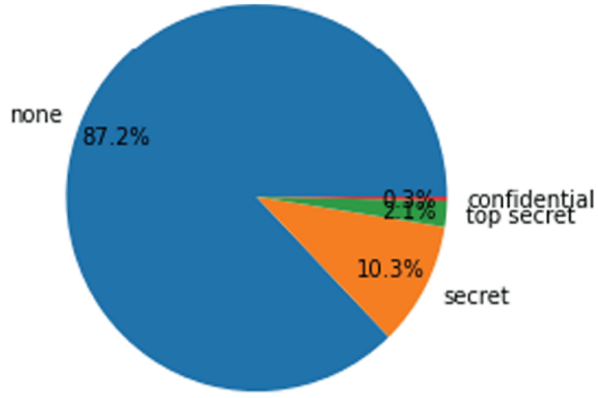
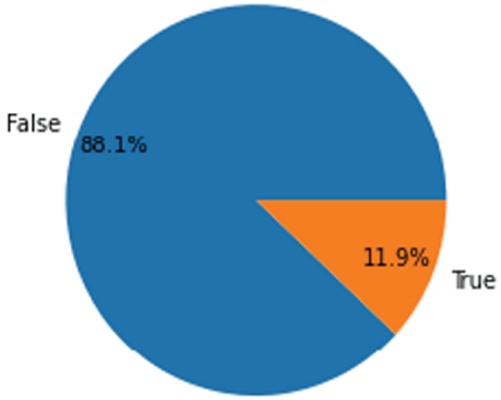


Figure A-2. The contract awards in terms of absolute dollar amount and the dollar amount relative to the total operating cost at each Systems Center from FY2001 to FY2020.

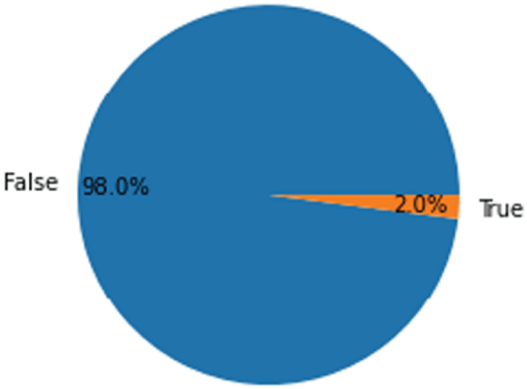
APPENDIX B
OBJECTIVES 4 RESULTS



(a) Percentage of NR&DE contracts with classification labels.



(b) Percentage of NR&DE contracts with awardee address.



(c) Percentage of NR&DE contracts with TPOC information.

Figure B-1. NR&DE project vulnerability assessment.

This page is intentionally blank.

APPENDIX C
PROTOTYPE DASHBOARD

Figure C-1. Prototype dashboard.

This page is intentionally blank.

INITIAL DISTRIBUTION

84310	Technical Library/Archives	(1)
53424	D. Lee	(1)
55123	V. Siu	(1)
53424	Dr. B. Michlin	(1)
53624	J. Bennett	(1)
71740	J. Duclos	(1)
53424	J. Wied	(1)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (1)

This page is intentionally blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
February 2022	Final	

4. TITLE AND SUBTITLE

2021 Data Challenge: Naval Information Warfare Center Pacific Innovation and Results.

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Dean Lee
Vincent Siu
Dr. Benjamin Michlin
Jeffrey Bennett
Josh Duclos
Jazlynn Wied
NIWC Pacific

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NIWC Pacific
53560 Hull Street
San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER

TR 3265

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

The NIWC Pacific Portfolio Council
53560 Hull Street
San Diego, CA 92152

10. SPONSOR/MONITOR'S ACRONYM(S)

NDIA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The 2021 Department of Navy Data Challenge asked teams from across the Department of Navy to find insights from Federal government contracts opportunities data. As part of the process, Naval Information Warfare Center (NIWC) Pacific designed a novel machine learning pipeline that incorporates methods from Natural Language Processing and Topological Data Analysis for insight extraction. Additionally, NIWC Pacific developed a vulnerability assessment metric using information extracted from contracts award data. The results from the challenge are displayed through a prototype dashboard.

15. SUBJECT TERMS

NLP; Data conditioning; NR&DE; bigram construction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dean Lee
U	U	U	SAR	36	19B. TELEPHONE NUMBER *(Include area code)* 619-553-7203

This page is intentionally blank.

This page is intentionally blank.

Naval Information Warfare Center Pacific (NIWC Pacific)
San Diego, CA 92152-5001