DAEN 690
Capstone Project Catalog

Fall 2024

College of Engineering and Computing
DATA ANALYTICS ENGINEERING
George Mason University®

About the Cover

James Baldo is an associate professor and serves as the Director of the George Mason University (GMU), College of Engineering and Computing (CEC), Volgenau School of Engineering (VSE), MS Data Analytics Engineering (DAEN) program. Dr. Baldo has served in this position since Fall 2018 and recently returned this past May after a 2-year sabbatical applying state-of-the-art data analytics engineering concepts and technologies to solve real world problems.

Prior to becoming director of the DAEN program he served 19 years as a CEC adjunct professor while working as a practicing engineer. His career has over 45 years of industry and government experience with roles as a data analytics engineer and software engineer.

His interest in large scale data, data management, analytics, and tools has provided him with opportunities to engage in assessing and applying new technologies across a diverse range of problems. The adoption of new technologies is exciting; however, adoption of new technologies requires careful planning and in addition to the technology, taking special care for planning how to successfully address for both organizational and cultural factors of the enterprise. This requires not only engineering knowledge and skills, but the ability to work on teams that interact across different corporate stakeholders.

Dr. Baldo continues to consult part-time with industry and leverages this knowledge and experience as feedback to the DAEN program. With technology moving at a lightening pace and technology adoption rates increasing, Dr. Baldo closely monitors the skillset needs of industry and how the DAEN program can provide graduates to fulfill these needs.

Dr. Baldo is currently researching applications of data mesh and data fabric for data management and exploring how analytics will be integrated into agentic systems.

CONTENTS

GMU MS DATA ANALYTICS ENGINEERING PROGRAMS

George Mason University's MS Data Analytics Engineering programs in the College of Engineering and Computing prepare students for their future careers in a growing discipline.

All of our programs are taught by our industry-leading faculty members across schools and colleges in Mason, giving our students the ability to see the numerous possibilities a data-driven degree can offer.

- **Master of Science** – The MS in Data Analytics Engineering is a multidisciplinary degree program in the College of Engineering and Computing. It provides students with an understanding of the technologies and methodologies necessary for data-driven decision-making.
- **Graduate Certificate** – The Graduate Certificate in Data Analytics Engineering gives students a foundation of basic data analytics and data science principles.
- **Master of Science Online** – The online MS program gives students the flexibility to earn an advanced degree and expand their knowledge in data analytics in an asynchronous format.
- **Graduate Certificate Online** – The online graduate certificate in data analytics engineering gives students a foundation of basic data analytics and data science principles with the flexibility of an online asynchronous format.

Data analytics engineering is an expanding field. Therefore, all our programs instruct students on current and innovative tools and prepare them to be adaptable to the future of the field.

DAEN 690: DATA ANALYTICS PROJECT COURSE

The DAEN 690 Data Analytics Project course is based on the program's objective to address three data analytics engineering roles: a) data engineer; b) data architect; and c) data analyst. The data engineering area of the program is focused on data conditioning required to fit data into specific data architectures and transform data to be exploitable. The data architecture area is focused creating frameworks that make data driven intelligence possible. The data analysis area is focused on creating repeatable means to draw key insight and signal from data.

The course blends analytics and engineering across a broad group of project domains. This provides students with the opportunity to develop and acquire the following knowledge and skills across the roles described previously:

- Ability to transform data into usable and computationally accessible forms [data engineer]
- Ability to condition data by extraction, cleansing, transformation, and loading [data engineer]
- Ability to plan, design, and implement data systems which separate the data from the application and scale as required [data engineer]
- Ability to plan, design, and implement systems (e.g., procedures, governance, and architectures) to store, manage, process, and preserve or dispose of data [data architect]
- Ability to manage data as an asset with operations such as data usage, cost, and risk [data architect]
- Ability to select, configure, and use algorithms across problem spaces to extract insights from data [data analyst]
- Ability to apply a range of mathematical, computation, modeling, and visualization techniques that enable the key insights and decision making from datasets [data analyst]

In addition to the above the course projects will address various dimensions of the following categories:

- Data Analytics Concepts – for example machine learning, deep learning, data mining, neural networks, regression, linear models, loss functions, optimization, etc.
- Data Analytics Tools – for example cloud based provides such as AWS, Azure, Google, IBM, etc.; Jupyter notebooks, RStudio, Tableau, Weka, RapidMiner, etc.
- Data Analytics Team Problem Solving – Agile or other team-based methodologies.

The course is based on a team project of 4-6 students using the Agile SCRUM methodology and working with an Industry sponsor and problem. Many data analytics engineering problems are solved using this approach in practice today. Therefore, the course provides the students with a real-world problem and experience that is similar to how data analytics engineers practice today.

CAPSTONE PROJECT DOMAINS

DAEN 690 partner projects encompass a spectrum of data science domains and individual projects may encompass one, or more, of the following domain areas.

Systems Engineering

Systems engineering refers to the design, integration, and management of complex systems over their life cycles. At its core, systems engineering utilizes systems thinking principles to organize this body of knowledge. Project teams may be tasked at times to deliver prototype or completed systems at the behest of their partners.

Data Engineering

Data engineering refers to transforming data into a useful format for analysis – also known as *"data wrangling."* This often involves managing the source, structure, quality, storage, and accessibility of the data so that it can be queried and analyzed by other analysts.

Data Mining

Data mining and data analytics are often used interchangeably, but there is a big difference between the two. Data Mining is the process of extracting valuable information from a large dataset. The data can be structured (e.g., data found in a relational database or data warehouse with a fixed or rigid schema), semi-structured (e.g., data that does not conform to a data model but has some structure and lacks a fixed or rigid schema), or unstructured (e.g., audio, video, social media postings, or other natural language text that is not easy for conventional tools to search).

Data Analytics

Data analytics is the process of interpreting data to find trends and patterns. It refers to the application of statistics in the form of exploratory data analysis – as well as descriptive, predictive, and prescriptive analytics – to reveal patterns and trends in data from existing data sources. Project teams will look at a business problem and translate it to a data question, create predictive models to answer the question and story tell about the findings.

Data Modeling and Simulation

Modeling and simulation refer to the up-front work which is required to research and define the problem statement before any data analysis work can be performed. In an academic setting the problem statements are already clearly defined for the student. In a "real world" setting basic research into a subject area is the critical first step to be performed before it can be determined what data is required to perform a data analysis.

Data Visualization

Data visualization refers to being able to present data in a visually appealing way for consumption by an end-user – often in the form of *data dashboards*.

Computer Vision (CV)

Computer vision (CV) is a branch of artificial intelligence that enables computers performing a set of computational techniques to analyze and extract data to derive meaningful information from digital images, videos, and other visual inputs.

Natural Language Processing (NLP)

Natural language processing (NLP) is a branch of artificial intelligence that enables computers to comprehend, generate, and manipulate human language. Natural language recognition and natural language generation are types of NLP. Chatbots, digital assistants, search engine optimization (SEO), analyzing and organizing large document collections, social media analytics, sentiment analysis, and content moderation are applications of natural language processing.

Artificial Intelligence/Machine Learning (Ai/ML)

Artificial Intelligence refers to the broad container term describing the various tools and algorithms that enable machines to replicate human behavior and intelligence. Machine learning refers to a more complex version of data mining and statistical analysis and is a branch of artificial intelligence focused on building applications that learn from data and improve their accuracy over time without being programmed to do so. Deep learning, a subset of machine learning, goes a step further through the use of artificial *neural networks* to discover patterns in the data.

DevOps

The term DevOps is a combination of *development* and *operations*. It is a culture of fostering a collaborative approach among all roles and tasks performed by application development and IT operations teams to formalize and migrate a development project into a production environment. DevOps is a practice that allows a single team to manage the entire application development lifecycle: requirements, design, development, testing, deployment, monitoring, and maintenance.

MLOps

The term MLOps is a combination of *machine leaning operations*. MLOps is an approach to managing machine learning projects and is a core function of machine learning engineering which is focused on streamlining the process of taking machine learning models to production and then maintaining and monitoring them. MLOps is a collaborative function that bridges the gaps between data scientists, devops engineers, and operations teams.

ACCURE, INC.

Project POC: Mr. Shamshad Ansari, CEO & President, sansar3@gmu.edu

https://accure.ai/

Accure is an AI automation company which has developed a software platform, Momentum, that automates phases of AI development across the major cloud platforms – AWS, GCP, and Azure.

Project Title	GRAPHICAL INSIGHTS: AUTOMATED DATA-DRIVEN REPORTING USING GENERATIVE AI
Organization	*Please provide the name of the partnering organization for this project:* Accure, Inc.
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Shamshad Ansari, CEO, sansari@accure.ai
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☐ Data Engineering ☐ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☐ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☒ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Generative AI foundational models (LLM), Retrieval-Augmented Generation (RAG), Prompt Engineering (PE), Amazon Web Services (AWS)
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Spring 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Many organizations struggle to extract meaningful insights from their data, due to the complexity and volume of the information. As a result, decision-makers often rely on manual analysis, which can be time-consuming, error-prone, and limited in scope.

KEY CHALLENGES:

- **Data overload:** The sheer volume and complexity of data make it difficult to identify key trends, patterns, and correlations.
- **Lack of expertise:** Non-technical stakeholders often lack the skills and knowledge to effectively analyze and interpret data.
- **Inefficient analysis:** Manual analysis is often slow, labor-intensive, and prone to errors, leading to delayed or inaccurate insights.
- **Limited scalability:** As data volumes grow, manual analysis becomes increasingly impractical, making it difficult to scale insights across the organization.

DESIRED OUTCOME:

A solution that can efficiently and accurately extract insights from structured data, providing clear and concise summaries, visualizations, and recommendations that support informed decision-making.

SecureGPT by Accure, Inc. is a generative AI solution designed for enterprise use, focusing on security and privacy. Here are some key features:

- **Enterprise-Grade Security:** SecureGPT can be hosted on-premises or in a private cloud, ensuring sensitive data is protected during transmission and storage.
- **Customization:** It can be tailored to meet the unique needs of various industries, such as legal, banking, healthcare, and more.
- **Data Privacy:** The model can be trained using an organization's private data, enhancing accuracy while maintaining strict data privacy measures.

- **Versatility:** SecureGPT can improve customer support, streamline operations, and enhance communication across different domains.

Project Goals

The project goals are to leverage Accure's Generative AI technology, SecureGPT, to automate the creation of industry-agnostic reports, including summaries, graphs, and charts. The solution will ensure input flexibility, offer user-friendly customization, and deliver high-quality visualizations to enhance decision-making across various domains. Rigorous testing and validation will ensure the solution's accuracy, reliability, and ease of use.

Data Sources and Datasets

We will utilize publicly available structured datasets from Kaggle. For the development and testing, we will use dataset from 5 different industries to assess the system's ability to generate industry agnostic reports.

Partner Intellectual Property

Accure Momentum and SecureGPT.

References

1. Arunkumar, L., Darji, V., Dwivedi, A., Kuttamath, A., Mahesh, A., Nguyen, T., "Graphical Insights: Leveraging Language Models for Structured Data Queries and Visualization," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.
2. Anantapalli, R., Annapoorna, S., Kudikala, M., Shrivastava, T., Thota, Y., Vemula, N., "Graphical Insights: Leveraging Language Models for Structured Data Queries and Visualization," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.
3. Accure Inc. SecureGPT (https://accure.ai/securegpt/).

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☒ Other (please specify below)
⇒ Engineered prompts and templates.

ALLWYN CORPORATION

Project POCs: Madhu Garlanka, CEO, madhu@allwyncorp.com
Swathi Young, Lead, Emerging Technologies, syoung@allwyncorp.com

https://allwyncorp.com/

Allwyn Corporation, headquartered in Washington DC, was founded in 2003 with a mission to help companies solve complex technology problems by bringing tools, technologies, experienced resources, processes, methodologies, and project delivery expertise to the table. Our goal is to enable our customers to stay competitive in the global marketplace by helping them implement high quality, cost-efficient software products, and IT applications. Our key differentiators are our approach, attitude, and the top-notch intellectual capital we deploy that will work as an extended arm of your organization to get the job done.

Project Title	GENERATIVE AI RETRIEVAL-AUGMENTED GENERATION (RAG) CHATBOT PROTOTYPE FOR CITY GOVERNMENTS
Organization	*Please provide the name of the partnering organization for this project:* Allwyn Corporation
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Madhu Garlanka, CEO, Allwyn Corporation, madhu.garlanka@allwyncorp.com Swathi Young, CTO, Allwyn Corporation, swathi.young@allwyncorp.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☒ Systems Engineering ☒ Data Engineering ☐ Data Mining ☐ Data Analytics ☐ Data Modeling/Simulation ☐ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☒ Generative AI ☐ DevSecOps ☒ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Generative AI foundational models (LLM), Retrieval-Augmented Generation (RAG), Prompt Engineering (PE), Amazon Web Services (AWS)
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Generative AI Retrieval-Augmented Generation (RAG) chatbots represent a significant advancement in the field of artificial intelligence, combining the strengths of generative models and retrieval-based systems to deliver more accurate and contextually relevant responses.

KEY COMPONENTS OF RAG CHATBOTS

Generative Models: These models, often based on large language models (LLMs) such as OpenAI GPT-4, Anthropic Claude, and Meta AI Llama generate human-like text based on the input they receive. They excel in creating coherent and contextually appropriate responses but can sometimes produce information that is not factually accurate.

Retrieval Systems: These systems enhance the generative models by retrieving relevant information from a predefined dataset or knowledge base. This ensures that the responses are not only coherent but also factually accurate and up-to-date.

HOW RAG CHATBOTS WORK

Query Processing: When a user inputs a query, the chatbot first processes this query to understand the context and intent.

Information Retrieval: The system then searches its knowledge base to retrieve relevant documents or data that can provide accurate information.

Response Generation: The generative model uses the retrieved information to construct a response that is both contextually appropriate and factually accurate.

BENEFITS OF RAG CHATBOTS

Enhanced Accuracy: By combining generative models with retrieval systems, RAG chatbots can provide more accurate and reliable information.

Contextual Relevance: These chatbots can maintain the context of a conversation over multiple turns, making interactions more natural and engaging.

Scalability: RAG chatbots can handle a wide range of queries and are scalable to meet the needs of large enterprises.

CHALLENGES AND CONSIDERATIONS

Complexity: Building and maintaining RAG chatbots is complex, requiring sophisticated engineering and fine-tuning of both the generative and retrieval components.

Data Privacy: Ensuring the privacy and security of the data used by these chatbots is crucial, especially in enterprise settings.

Latency: The process of retrieving information and generating responses can introduce latency, which needs to be managed to ensure a smooth user experience.

APPLICATIONS

Customer Support: RAG chatbots can provide accurate and timely responses to customer inquiries, improving customer satisfaction and reducing the workload on human agents.

Enterprise Solutions: These chatbots can be used to answer questions about company policies, IT support, and other internal processes, enhancing employee productivity.

Healthcare: In the healthcare sector, RAG chatbots can assist in providing medical information and support, ensuring that patients receive accurate and reliable information.

Enhancing Citizen Services: Generative AI RAG chatbots can improve accessibility and efficiency of services for residents.

For example, a city government deploying a generative AI RAG chatbot to improve accessibility and efficiency of city services for residents might do so for the following reasons.

24/7 Availability: The chatbot can provide round-the-clock assistance, answering common queries about city services, events, and regulations.

Multilingual Support: It can communicate in multiple languages, catering to the diverse population of the city.

Real-Time Information: The chatbot can access and provide up-to-date information on public transportation schedules, road closures, and emergency alerts.

Personalized Assistance: Using generative AI, the chatbot can offer personalized responses based on the user's previous interactions and preferences.

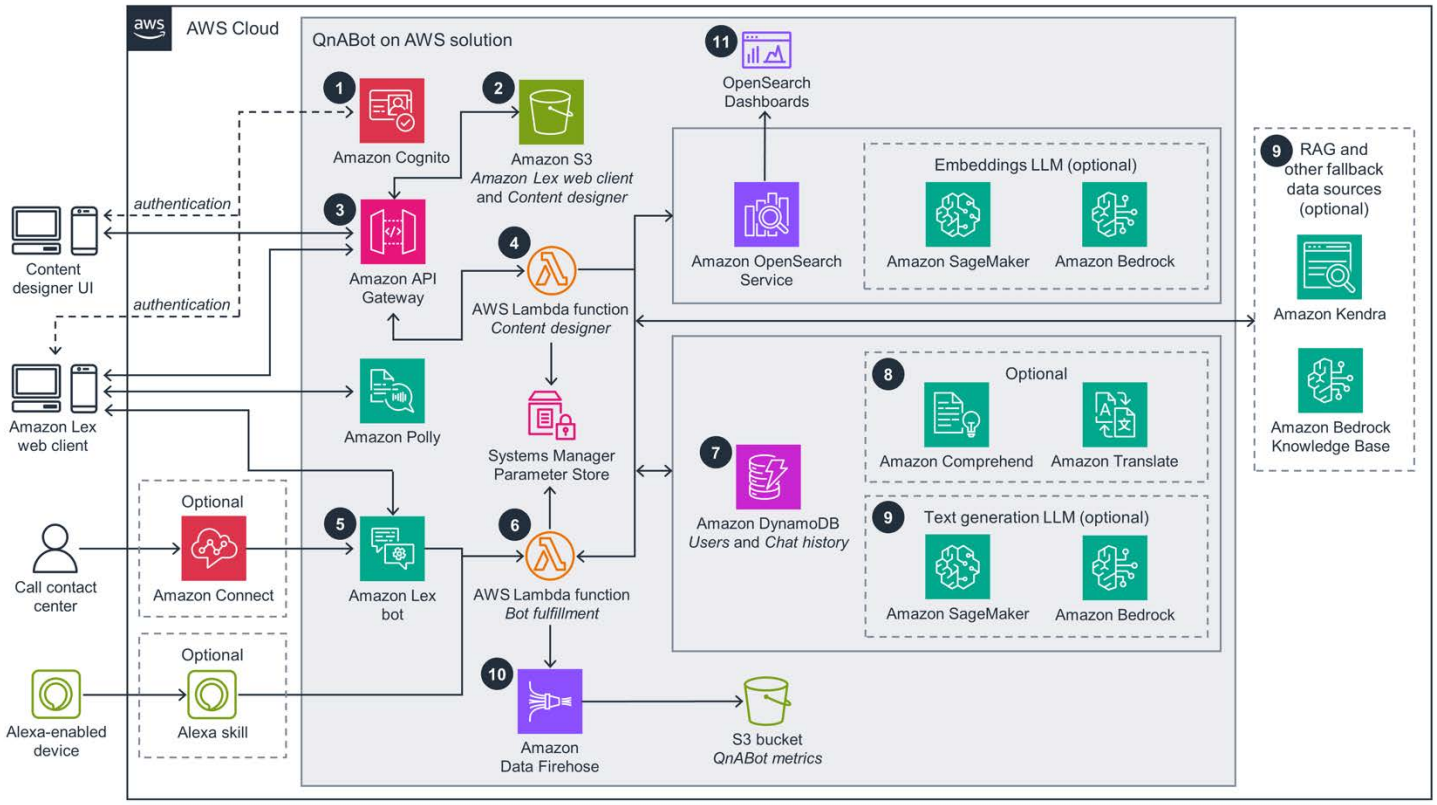
Document Retrieval: Residents can request and receive important documents, such as permits and licenses, through the chatbot.

Feedback Collection: The chatbot can gather feedback from residents on city services, helping the government to identify areas for improvement.

This capstone project will develop a conversational chatbot prototype for a city government to improve citizen engagement and information accessibility using data found on a publicly available city government website. This chatbot will leverage generative AI (LLMs), advanced search capabilities, retrieval-augmented generation (RAG), prompt engineering, and commercial AWS AI services to provide quick, accurate responses to user queries and enhance the overall user experience.

Project Goals

Building on the results from the Spring 2024 capstone project titled *"Generative AI Chatbot with AWS"* the project team will create the Generative AI RAG chatbot prototype using the AWS QnABot solution reference architecture (see figure below). User input to the chatbot will be restricted to a web browser only (i.e., no mobile client development, no call contact center, or Alexa-enabled device).



The AWS services incorporated into the solution will include:

- **Amazon S3** for storage
- **Amazon Lex** for user interactions with QnABot where Amazon Lex processes the input to understand the user's intent and extract relevant information
- **Amazon Kendra** for intelligent search
- **Amazon Bedrock** for high-performing foundation Large Language Models (LLMs)
- **AWS Lambda** for serverless computing
- **Amazon Cognito** for QnABot developer authentications

In addition to developing the prototype solution, the project team will also address the following questions.

1. Research and evaluate which Amazon Bedrock foundation model will provide the *"best"* responses to citizen queries.
2. Research various approaches to evaluating chatbot responses, then apply what the team has learned for evaluating and improving the correctness and efficiency of the prototype chatbot responses.
3. Research the monthly AWS costs expected when deploying the prototype chatbot. Document and model your various operating assumptions.

Data Sources and Datasets

The City of Virginia Beach government website (https://virginiabeach.gov/) will be the source data for the Retrieval-Augmented Generation (RAG) portion of the project. The team will utilize **Amazon Kendra** to crawl and index the website pages for the RAG documents.

Partner Intellectual Property

None.

References

1. **AWS Solutions Library / AWS Solution / QnABot on AWS** (https://aws.amazon.com/solutions/implementations/qnabot-on-aws/)
2. **What is RAG (Retrieval Augmented Generation)?** (https://aws.amazon.com/what-is/retrieval-augmented-generation/)
3. **Amazon Kendra** (https://aws.amazon.com/kendra/)
4. **Amazon Lex** (https://aws.amazon.com/lex/)
5. **Amazon Bedrock** (https://aws.amazon.com/bedrock/)
6. **AWS Lambda** (https://aws.amazon.com/pm/lambda/)
7. Chilakabathini, S., Payaga, K, Staton, B., Vadakattu, A., Veeremalla, N., Vunnam, B., "Generative AI Chatbot with AWS," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☒ Other (please specify below)
⇒ Documentation and procedures to export the project data and environment to the project partner.

ERASMUS.AI

Project POCs: Daniel Erasmus, CEO Erasmus.AI, daniel@dtn.net

https://erasmus.ai/

Erasmus.AI is a pioneering company in the field of artificial intelligence, known for its innovative solutions that address complex global challenges. One of their notable projects is the development of ClimateGPT, an open-source AI platform dedicated to tackling the impacts of climate change. This platform leverages a combination of large language models to assist in decision-making processes for researchers, policymakers, and business leaders. ClimateGPT is designed to be highly efficient in climate-specific tasks, offering insights across various scientific disciplines in over 20 languages.

The company also focuses on creating AI-powered tools for financial institutions and other sectors. For instance, they have developed a Human-Centred Extreme Weather Dashboard in collaboration with the Club of Rome, which helps in delivering innovative climate services and targeting aid more effectively. Erasmus.AI's technology is built on processing vast amounts of data, including billions of web pages and academic articles, to provide comprehensive and actionable insights.

Erasmus.AI's commitment to responsible AI development is evident in their approach to creating equitable and audited models. They emphasize the importance of interdisciplinary research and the inclusion of diverse expert perspectives to ensure their AI solutions are robust and reliable. This dedication to ethical AI practices positions Erasmus.AI as a leader in the field, driving forward the use of AI for social good.

Project Title	IMPROVING THE PERFORMANCE OF CLIMATEGPT
Organization	*Please provide the name of the partnering organization for this project:* Erasmus.AI (Amsterdam, The Netherlands)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Daniel Erasmus, CEO Erasmus.AI, daniel@dtn.net
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☒ Systems Engineering ☐ Data Engineering ☐ Data Mining ☐ Data Analytics ☐ Data Modeling/Simulation ☐ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☐ AI/ML ☒ Generative AI ☐ DevSecOps ☒ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Generative AI foundational models (LLM), Retrieval-Augmented Generation (RAG), Prompt Engineering (PE), Amazon Web Services (AWS)
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☐ 1 ☐ 2 ☐ 3 ☒ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

ClimateGPT is the world's first foundational model on climate change. Developed in partnership with the Club of Rome Climate, the ClimateGPT family of models (https://arxiv.org/abs/2401.09646) enables fundamentally different decision-making for the transformation ahead. Early and forward-looking action on climate change (https://doi.org/10.1038/s41598-020-66275-4) is crucial for several reasons.

1. **Mitigation of Severe Impacts:** Acting early helps to reduce the severity of climate change impacts. By cutting greenhouse gas emissions now, we can limit global temperature rise and avoid the most catastrophic consequences, such as extreme weather events, sea-level rise, and loss of biodiversity.
2. **Cost-Effectiveness:** Early action is generally more cost-effective than delayed action. The longer we wait, the more expensive it becomes to implement necessary measures. Early investments in renewable energy, energy efficiency, and other sustainable practices can save money in the long run.
3. **Adaptation and Resilience:** Forward-looking action allows societies to better prepare for and adapt to the changes that are already underway. This includes building resilient infrastructure, developing early warning systems, and implementing policies that protect vulnerable communities.
4. **Technological Innovation:** Early action drives innovation in clean technologies and sustainable practices. This can lead to new economic opportunities, job creation, and a competitive advantage for countries and companies that lead in the transition to a low-carbon economy.
5. **Intergenerational Equity:** Taking action now ensures that future generations inherit a planet that is livable and thriving. It is a matter of ethical responsibility to address climate change proactively to protect the well-being of future generations.
6. **Global Leadership and Cooperation:** Early and decisive action can position countries as leaders in the global effort to combat climate change. This can foster international cooperation and encourage other nations to follow suit, amplifying the overall impact.

ClimateGPT is designed to tackle the complex and rapidly evolving challenges posed by climate change. Here are some key features and capabilities of ClimateGPT:

1. **Open Source Model:** ClimateGPT is available as an open-source model, allowing researchers, policymakers, and business leaders to access and utilize its capabilities for climate-related decision-making.
2. **Comprehensive Data Analysis:** The platform can analyze vast amounts of data, including over 200 million academic articles and 10 billion webpages, to provide insights on climate change impacts.
3. **Multilingual Support:** ClimateGPT supports translations in over 20 languages, making it accessible to a global audience.
4. **Scenario Testing:** It helps users test various scenarios for resource allocation, mitigation, and adaptation strategies, providing a forward-looking AI decision model.
5. **Specialized Topics:** The model is optimized to explore leading scientific research on topics such as regenerative agriculture, geoengineering, climate-related risks, and sustainable economic systems.
6. **Community and Expert Collaboration:** ClimateGPT is built with input from experts and local stakeholders, ensuring that it provides relevant and actionable insights.
7. **Holistic Approach:** The platform connects the dots between different climate-related events and their broader impacts, helping decision-makers understand the interconnections and interdependencies of climate risks and opportunities.

ClimateGPT is also the first AI system to publish sustainability scores and is built with hydropower with inference run on green power. This innovative feature allows users to evaluate the sustainability performance of various entities, such as companies, projects, or policies, based on comprehensive data analysis. Here are some key aspects of this capability:

1. **Data-Driven Insights:** ClimateGPT leverages a vast amount of data, including academic research, web pages, and climate-specific datasets, to generate accurate and reliable sustainability scores.
2. **Holistic Evaluation:** The AI system considers multiple dimensions of sustainability, including environmental impact, social responsibility, and governance practices, to provide a well-rounded assessment.
3. **Transparency and Accessibility:** By making these scores publicly available, ClimateGPT promotes transparency and encourages organizations to improve their sustainability practices.
4. **Support for Decision-Making:** These scores can be used by investors, policymakers, and business leaders to make informed decisions that align with sustainability goals.
5. **Continuous Improvement:** The AI model is continuously updated and fine-tuned with new data and feedback from experts and stakeholders, ensuring that the sustainability scores remain relevant and accurate.

Climate change poses a significant threat to global stability, with extreme weather events becoming more frequent and intense. To mitigate this challenge, AI can play a crucial role in addressing the underlying factors driving climate change. By leveraging the power of generative AI, we can create a more accurate and detailed understanding of climate change, ultimately enabling more effective and targeted solutions to this global challenge.

To ground model responses in existing numerical climate databases, (e.g., Cost of a metric ton of Carbon, NOAA Climate Data Online (CDO), Climate Hotspots Database, Climate-ADAPT, NASA’s Climate Data Online, IPCC Data Distribution Centre (DDC), European Union's Climate-ADAPT database, World Bank's Climate Change Knowledge Portal, etc.) the team(s) should (1) develop an AI classifier that can parse user queries into core model responses or/and ancillary databases, (2) develop a framework for structured queries (typically SQL) into the databases with a focus on long term utility of the interfaces (e.g., API), (3) format the multiple responses into ClimateGPT appropriate prompts to (4) allow for rich user responses (possibly multi-modal, with references) The system will form part of ClimateGPT 3 roadmap release and will be used in a public/policy facing space underpinning real life climate resilience decisions from Maui to Madagascar.

Project Goals

The project goals can be broken out into a series of steps to improve the performance of ClimateGPT.

Step 1: Develop an **AI classifier** that can parse user queries into **core model responses** and/or **ancillary database responses**. Core model user queries can be passed directly on to the ClimateGPT foundational model while ancillary database user queries will need to retrieve additional data before being passed on to the ClimateGPT foundational model.

Step 2: Develop a **framework for structured queries** (typically SQL) into the ancillary databases with a focus on long-term utility of the interfaces (i.e., API). The framework will need to access several public databases (e.g., NOAA Climate Data Online (CDO), Climate Hotspots Database, Climate-ADAPT, NASA's Climate Data Online, IPCC Data Distribution Centre (DDC), European Union's Climate-ADAPT database, NetZero Cloud, World Bank's Climate Change Knowledge Portal, etc.). The exact set of databases will be determined according to the difficulty executing, the value of responses, etc.

Step 3: Format the multiple responses from Step 2 into ClimateGPT appropriate prompts. An architectural consideration the project team will need to address is whether the databases should be run and synced to ensure near real time user responses.

Step 4: Allow for **rich user responses** – possibly multi-modal with references – as output generated by the Climate GPT foundational model.

Data Sources and Datasets

For Step 2 of the project, the team will utilize publicly available climate databases to include, but not be limited to, the following list of databases.

1. **NOAA Climate Data Online (CDO)** (https://www.ncei.noaa.gov/cdo-web/) is a comprehensive resource provided by the National Centers for Environmental Information (NCEI). It offers free access to a vast archive of historical weather and climate data.
2. **Climate Hotspots Database** is a tool designed to identify and track areas that are particularly vulnerable to the impacts of climate change. There are several databases available including the **United National Environment Program Strata** platform (https://unepstrata.org/).
3. **Climate-ADAPT** (https://climate-adapt.eea.europa.eu/en) is the European Climate Adaptation Platform, a partnership between the European Commission and the European Environment Agency (EEA). It serves as a comprehensive resource for sharing knowledge on climate adaptation to support a climate-resilient Europe.
4. **NASA** provides extensive climate data through various platforms, including the **NASA Center for Climate Simulation (NCCS)** (https://www.nccs.nasa.gov/services/climate-data-services). This center offers a centralized location for accessing large, complex climate model data, which is beneficial for the climate science community and the broader public. Additionally, NASA's **Climate Change: Vital Signs of the Planet** website (https://science.nasa.gov/climate-change/) offers detailed climate data and research, including current news, data streams, and interactive tools like the Climate Time Machine.
5. The **IPCC Data Distribution Centre (DDC)** (https://www.ipcc-data.org/) is a key resource established by the Intergovernmental Panel on Climate Change (IPCC). Its primary purpose is to provide a transparent, traceable, stable, and accessible archive for climate, socio-economic, and environmental data and scenarios used in IPCC reports.
6. **Net Zero Cloud**, developed by Salesforce, is a comprehensive platform designed to help organizations manage their environmental, social, and governance (ESG) data. Net Zero Cloud (https://www.salesforce.com/net-zero/cloud/) is a commercial product requiring paid subscriptions.
7. The World Bank's **Climate Change Knowledge Portal (CCKP)** (https://climateknowledgeportal.worldbank.org/) is an online platform designed to provide comprehensive data and information on climate change. It serves as a one-stop shop for global, regional, and country-level climate data, offering insights into historical and future climate conditions, vulnerabilities, and impacts.

The project team will assess and evaluate other publicly available climate databases as necessary for the project.

Partner Intellectual Property

The ClimateGPT codebase (https://huggingface.co/eci-io/climategpt-7b) is published under the **Apache License 2.0** permissive open-source software license to allow the ClimateGPT team to continue to work on the framework in an unencumbered fashion.

References

1. **ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change** (https://arxiv.org/abs/2401.09646)

2. **Assessing the Costs of Historical Inaction on Climate Change** (https://doi.org/10.1038/s41598-020-66275-4)

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

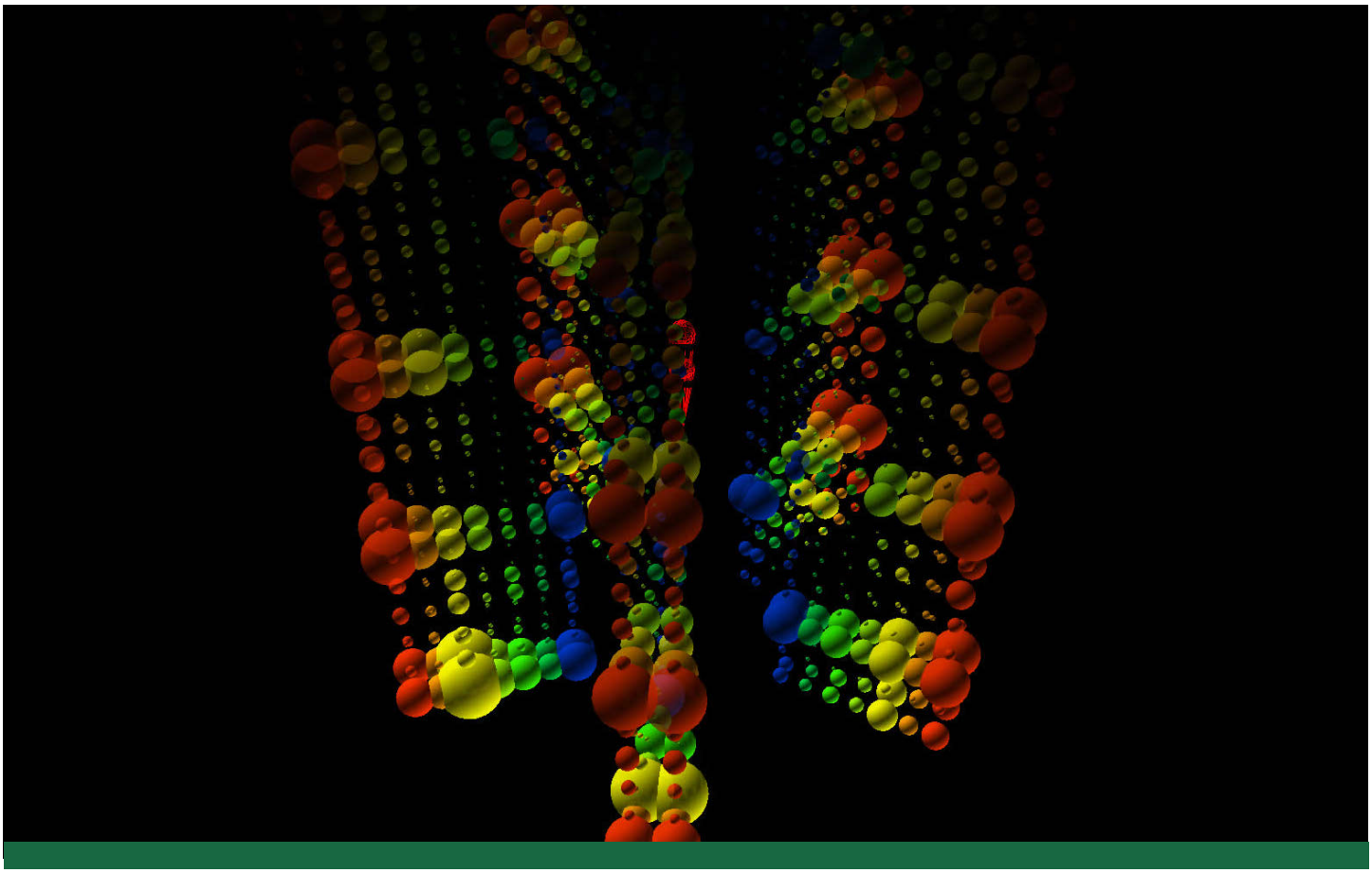
At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☐ Working Prototype
☒ Other (please specify below)
⇒ Contribution to the ClimateGPT 3 academic paper.



GAIAVIZ, LLC

Project POCs: Shane Saxon, CEO & co-founder, shane@gaiaviz.com
Claire B. Saxon, COO & co-founder, claire@gaiaviz.com
Lucas D. Erickson, Software Engineer, lucas@gaiaviz.com

https://gaiaviz.com/

GaiaViz is a cutting-edge company specializing in 3D data fusion solutions. They provide custom software and equipment designed to offer live, immersive 3D data visualizations. This technology facilitates the exploration, analysis, and presentation of complex data sets. GaiaViz integrates real-time graphics, multiple input sources such as IoT devices, cameras, and databases, along with physical controllers and powerful processing capabilities.

The company's solutions are tailored to various industries, including business intelligence, healthcare, government, and education. GaiaViz's platform is known for its advanced 3D visualization engine, which allows users to interact with data in a highly intuitive and engaging manner. This interactive approach helps users gain deeper insights and make more informed decisions in real-time.

Founded by Shane Saxon and Claire B. Saxon, GaiaViz has over 14 years of development experience, supported by governmental and academic grants. The company prides itself on its ability to deliver high-performance, cognitive-friendly 3D data visualization solutions. GaiaViz also offers a range of services, from turnkey systems and content creation to live installations and website embedding, ensuring that clients receive comprehensive support for their data visualization needs.

Project Title	GAIAVIZ HIGH-DIMENSIONAL WEATHER EVENT EXPLORER
Organization	*Please provide the name of the partnering organization for this project:* GaiaViz LLC (Paris, France)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Shane Saxon, CEO, shane@gaiaviz.com Claire B. Saxon, COO, claire@gaiaviz.com Lucas D. Erickson, Software Engineer, lucas@gaiaviz.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☐ Data Mining ☒ Data Analytics ☒ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Python Data Wrangling (required), 3D Modeling (Spatial Reasoning is a plus)
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Climate change increases the frequency and intensity of extreme weather events that impact civil society, businesses and the environment alike. The effects of which can be felt by communities long after the initial occurrence. Such is the case when a wildfire structurally alters terrain leading to landslides several months later, or when water infrastructure is overwhelmed to the point it causes sanitation and epidemic issues. Improved planning and agile disaster response is key to better outcomes.

There are a growing number of relevant data sources, but they are time consuming to analyze and usually difficult to grasp the complex multi-dimensional relationships. We have historic data, real-time data, and predictive models, but not the tools to comprehend the complete picture, especially in real-time. New tools are needed to make better decisions to protect infrastructure, supply chains, property and life.

Project Goals

Develop a real-time weather data-fusion explorer using the GaiaViz app (3D data engine). A proof-of-concept that incorporates real-time IoT sensors (water monitoring, traffic, etc.) and weather based news events (data provided by Erasmus.AI).

We aim to enhance the interactive exploration of extreme weather events to aid decision-making at a higher-level and facilitate community planning at the grass-roots level. By enabling stakeholders to intuitively visualize complex 3D weather data in real-time, we intend to improve resilience and disaster management.

Data sources will include a year of extreme weather news events, and other public data sources, such as telemetric data streams from MQTT brokers and REST (JSON). Deliverables include a 3D interactive visualization that fuses these multiple data types. Visualizing human based news events and layering them with real-time telemetric field sensors may improve early detection and improve disaster response effectiveness.

Data Sources and Datasets

1. Erasmus AI *"Human Centered Extreme Weather Dashboard"* dataset will be provided in JSON format.
2. Publicly available REST (JSON) and MQTT data streams (traffic, weather sensors, etc.).

Partner Intellectual Property

The GaiaViz app is provided under the **Apache License 2.0** permissive open-source software license.

References

1. **MQTT** (https://mqtt.org) is an OASIS standard messaging protocol for the Internet of Things (IoT). It is designed as an extremely lightweight publish/subscribe messaging transport that is ideal for connecting remote devices with a small code footprint and minimal network bandwidth. MQTT today is used in a wide variety of industries, such as automotive, manufacturing, telecommunications, oil and gas, etc.
2. **MIDIMonster** (https://midimonster.net) is a universal control and translation tool for most show control protocols in the entertainment industry. It bridges MQTT to OSC (Open Sound Control) supported by GaiaViz app.
3. **Erasmus.AI *"Human Centered Extreme Weather Dashboard"*** can be viewed at https://erasmus.ai/extreme_weather/map.html.
4. **GaiaViz LLC** can be found at https://www.gaiaviz.com/.

Project Development Environment

GaiaViz is a Windows 11 application. All project application development will need to be performed on a PC or laptop system with the following minimum specifications.

- Windows 11 Pro (or Windows 11 Education)
- Intel Core I7 or I9 processor (or AMD equivalent)
- 16 GB RAM
- 1 TB storage
- Dedicated Nvidia GPU

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☒ Other (please specify below)
⇒ Screenshot images and video capture of the 3D datascape in the GaiaViz app.
⇒ A diagram of the data flow process.

Project Title	UNITED NATIONS SUSTAINABLE DEVELOPMENT GOAL (SDG) EXPLORER
Organization	*Please provide the name of the partnering organization for this project:* GaiaViz LLC (Paris, France)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Shane Saxon, CEO, shane@gaiaviz.com Claire B. Saxon, COO, claire@gaiaviz.com Lucas D. Erickson, Software Engineer, lucas@gaiaviz.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☐ Data Mining ☒ Data Analytics ☒ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☐ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Python Data Wrangling (required), 3D Modeling (Spatial Reasoning is a plus).
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Spring 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Since 2015, the United Nations rallies governments, private sector and civil society around 17 Sustainable Development Goals (SDGs) for a better future, in a roadmap to alleviate health, social, political and economic problems.

In a call for partnerships between stake holders of all levels, Goal 17, *"Partnerships For The Goals"* encourages to "*Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development"* (Goal 17 | Department of Economic and Social Affairs, n.d.), through higher inclusion of the least developed countries in the trade system as well as resources, expertise and technology sharing to build their self-sufficiency.

The Sustainable Development Goals (SDGs) progress is monitored by series of indicators. It is a complex and multidimensional dataset that has reached 2.7 millions records in 2023, provided in CSV and/or JSON formats. There is a need for a more effective visualization approach (than traditional 2D graphs), as to enable a holistic vantage point that allows project participants to better understand the complex relations surrounding the goals.

Project Goals

Building on the work of two capstone project teams from the Spring 2024 semester, we aim to enhance the presentation of the Sustainable Development Goals (SDGs) to aid decision-making at the higher-level and facilitate actionability at the grass-root level through improved communication of key insights within these inherently complex systems. Enabling stakeholders to see how their piece of the puzzle relates to other key players, so that they may better know who to communicate with and how to communicate actionable information.

Deliverables include a 3D interactive visualization of the partnerships surrounding all 17 of the Sustainable Development Goals using the GaiaViz app. This is a follow-on project that will enhance the existing datascape by adding EU open data that shows the funding (links) between partnerships, which is particularly important for Goal 17.

Data Sources and Datasets

The **UN SDG (Sustainable Development Goals)** database (https://unstats.un.org/sdgs/dataportal) is a comprehensive platform provided by the United Nations. It offers access to data on over 210 indicators related to the SDGs, which are part of the 2030 Agenda for Sustainable Development.

This database allows users to:

- Analyze data availability and trends for each indicator globally and regionally.
- Compare trends for different countries and indicators.
- Access detailed SDG profiles for 132 countries.

It's a valuable resource for tracking progress, informing policy, and ensuring accountability in achieving the SDGs.

To find open data on **SDG aid funding in Europe**, particularly focusing on partnerships for Goal 17, you can explore the following resources:

1. **Eurostat:** The European Union's statistical office provides detailed data on SDG 17, including financial resources and partnerships. You can access their reports and datasets at https://ec.europa.eu/eurostat/statistics-explained/index.php/SDG_17_-_Partnerships_for_the_goals.
2. **International Aid Transparency Initiative (IATI):** This platform offers comprehensive data on international development aid, including funding links and partnerships. Their data can be accessed at https://stats.oecd.org/Index.aspx?DataSetCode=Table1.
3. **World Bank SDG Atlas:** This resource provides visualizations and data on SDG progress, including partnerships and funding. You can explore it at https://datatopics.worldbank.org/sdgatlas/goal-17-partnerships-for-the-goals/?lang=en.

Partner Intellectual Property

The GaiaViz app is provided under the **Apache License 2.0** permissive open-source software license.

References

1. Calve, R., Cholleti, S., Donapati, T., Kanumuri, N., Patel, J., "United Nations Sustainable Development Goals (SDGs) High-Dimensional Explorer," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.
2. Clark, J., Coombs, S., Jain, S., Myslewski, J. Wadsworth, R., "United Nations Sustainable Development Goals (SDGs) High-Dimensional Explorer," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.
3. GaiaViz LLC can be found at https://www.gaiaviz.com/.

Project Development Environment

GaiaViz is a Windows 11 application. All project application development will need to be performed on a PC or laptop system with the following minimum specifications.

- Windows 11 Pro (or Windows 11 Education)
- Intel Core I7 or I9 processor (or AMD equivalent)
- 16 GB RAM
- 1 TB storage
- Dedicated Nvidia GPU

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☒ Other (please specify below)
- ⇒ Screenshot images and video capture of the 3D datascape in the GaiaViz app.
- ⇒ A diagram of the data flow process.

GMU CENTER FOR AIR TRANSPORTATION SYSTEMS RESEARCH (CATSR)

Project POC: Dr. Lance Sherry, Director CATSR, lsherry@gmu.edu

https://catsr.vse.gmu.edu/

The GMU Center for Air Transportation Systems Research (CATSR) mission is to foster excellence in education and research in Air Transportation Systems Engineering.

Project Title	AIRLINE FINANCIAL ECONOMICS DASHBOARD
Organization	*Please provide the name of the partnering organization for this project:* GMU Center for Air Transportation Systems Research (CATSR)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Dr. Lance Sherry, Director, Center for Air Transportation Systems Research (CATSR), lsherry@gmu.edu
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☐ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☒ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☐ 1 ☒ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Form 41 is a comprehensive reporting system used by the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS). It requires most American passenger and cargo airlines to report detailed financial and operational information. This includes balance sheets, income statements, and various operating statistics such as revenue passenger-miles and available seat-miles.

The data collected through Form 41 is crucial for understanding the financial health and operational performance of the airline industry. It helps in regulatory oversight, policymaking, and providing transparency to the public.

Form 41 collects a wide range of detailed information from airlines. Here are some key types of data reported:

1. **Financial Data:**
    a. **Balance Sheets:** Assets, liabilities, and equity.
    b. **Income Statements:** Revenues, expenses, and net income.
    c. Cash Flow Statements: Cash inflows and outflows from operations, investing, and financing activities.
2. **Operational Data:**
    a. **Traffic Statistics:** Revenue passenger-miles (RPMs), available seat-miles (ASMs), and load factors.
    b. **Fleet Data:** Number and types of aircraft, aircraft utilization, and age of the fleet.
    c. Employee Data: Number of employees, hours worked, and labor costs.
3. **Performance Metrics:**
    a. **On-Time Performance:** Arrival and departure times, delays, and cancellations.
    b. **Fuel Consumption:** Amount of fuel used, fuel costs, and efficiency metrics.
4. **Market Data:**
    a. **Passenger and Cargo Data:** Number of passengers, amount of cargo transported, and revenue generated from these services.
    b. **Route Information:** Details about routes operated, including distances and frequencies.

Airlines are required to submit Form 41 data to the U.S. Department of Transportation on a **monthly**, **quarterly**, or **semi-annual basis** depending on the specific type of information being reported. This ensures that the Bureau of Transportation Statistics has up-to-date and comprehensive data for analysis and regulatory purposes.

The MIT Airline Data Project (ADP) (https://web.mit.edu/airlinedata/www/Revenue&Related.html) is an initiative by the Massachusetts Institute of Technology's Global Airline Industry Program. It aims to provide a comprehensive and user-friendly repository of data and analysis on the U.S. commercial airline industry. The project used to provide a tabular version of the data but stopped updating the website in 2020.

KEY FEATURES OF THE ADP:

- **Data Source:** The project primarily uses data from the U.S. Department of Transportation's Form 41.
- **Scope:** It covers various aspects of the airline industry's performance, including financial health, operational efficiency, and market trends.
- **Purpose:** The ADP is designed to support academia, the financial community, and the media in monitoring and understanding the evolution of the airline industry.
- **Updates:** The data is updated annually, typically in June, following the release of Form 41 data.

The ADP provides detailed analysis by individual airlines, helping to identify trends, opportunities, and challenges within the industry.

LIMITATIONS OF THE ADP:

- Website is manual (not updated automatically)
- Data is not up to date
- Data is tabular (not visualized)
- Stock price data is not included

Project Goals

The project team will develop a public-facing website portal with the following features:

1. Updated (semi) automatically with most recent data
2. Provide tables and charts of the data
3. Include stock price
4. Use analytical methods (e.g. Machine Learning) to establish trends and correlations between stock price and cost/revenue/productivity data

The project team ***must interview*** industry experts to assess what data visualizations and analysis are required. Industry experts the team are to interview, at a minimum, include:

Dr. Darryl Jenkins – Dr. Darryl Jenkins (airjenkins@aol.com) is a well-known aviation expert and former director of The Aviation Institute at George Washington University. He has extensive experience in the airline industry and has authored numerous studies and reports on aviation economics and airline management. Dr. Jenkins has also been involved in various consulting roles, providing insights and analysis on airline operations and market trends.

Dr. Antonio A. Trani – Dr. Antonio A. Trani (vuela@vt.edu) is a professor at Virginia Tech in the Department of Civil and Environmental Engineering. His areas of expertise include air transportation, simulation and modeling, airport engineering, and systems engineering. Dr. Trani has been involved in numerous research projects related to aviation demand modeling, airport noise studies, and runway capacity improvements.

Dr. Frederick Weiland – Dr. Frederick Wieland (fwieland@mosaicatm.com) is the Chief Research Scientist at Mosaic ATM. He has over thirty years of experience with NASA, the FAA, and the Department of Defense in fields such as analysis, modeling, and simulation. Dr. Wieland is also involved in business development at Mosaic ATM and has a background in teaching systems engineering and related subjects at George Mason University.

Dr. Seth Young – Dr. Seth Young (young.1460@osu.edu) is a prominent figure in aviation studies at The Ohio State University. He holds the position of Associate Professor in the Department of Civil, Environmental, and Geodetic Engineering and serves as the McConnell Chair of Aviation. Dr. Young is also the founding Director of The OSU Center for Aviation Studies. His academic background includes a Ph.D. in Civil and Environmental Engineering/Transportation and an M.S. in Industrial Engineering/Operations Research from the University of California, Berkeley, as well as a B.A. in Applied Mathematics from the State University of New York at Buffalo. Dr. Young is an Accredited Airport Executive and a certified flight instructor.

Ben Baldanza – Ben Baldanza (bbaldanz@gmu.edu) is an adjunct professor of economics at George Mason University. He is well-known for his tenure as the CEO of Spirit Airlines, where he transformed the company into an ultra-low-cost carrier. Baldanza also serves on the boards of JetBlue Airways and Six Flags Entertainment. Additionally, he co-hosts a weekly podcast called "Airlines Confidential".

Kent Duffy – Kent Duffy (kent.duffy@faa.gov) is a National Resource Expert for Airport/Airspace Capacity at the Federal Aviation Administration (FAA) located in Arlington, Virginia. He is involved in various aspects of airport capacity and delay modeling, NextGen integration with airport development, and runway length evaluations. Additionally, he works on critical aircraft determinations and the release of NAS-Data (Radar) for airports.

Robert Samis – Robert Samis (robert.samis@faa.gov) is an economist at the U.S. Department of Transportation. His work involves analyzing economic data and trends related to transportation, which can help inform policy decisions and regulatory actions.

Megan James – Megan James (megan.james@faa.gov) is an Operations Research Analyst with the Federal Aviation Administration (FAA) located in Mays Landing, New Jersey. As an Operations Research Analyst, Megan utilizes fast-time simulation modeling and analysis to study airport capacity. These studies determine present airport capacities and delays. The results of which can be used to suggest ways to increase airport capacities, reduce flight delays, increase airport efficiency, provide estimates of cost benefits of proposed improvements, provide hourly capacity rates for investment decisions, assist with planning for upcoming construction projects, and develop airport improvement action plans.

The team will also consider other industry experts as necessary.

In terms of infrastructure, the project team will implement the solution using the College of Engineering and Computing IT Services-provided AWS environment. The project team should consider various ways in which to capture, store, and analyze the data using AWS services and implement the most cost-effective and easily maintainable solution. The project team needs to also consider whether the dashboard visualizations should be implanted via code (e.g., Python or R) or via a 3rd-party commercial visualization tool (e.g., Tableau Public). The project team can consider this project a prototype systems solution in a development environment but should take into consideration what would be necessary for deploying the solution to a production AWS environment (i.e., DevSecOps).

Data Sources and Datasets

1. **United States Department of Transportation | Bureau of Transportation Statistics | Form 41 Data** (https://www.transtats.bts.gov/databases.asp?Z1qr_VQ=E&Z1qr_Qr5p=N8vn6v10&f7owrp6_VQF=D).

Partner Intellectual Property

None.

References

1. **MIT Global Airline Industry Program | Airline Data Project** (https://web.mit.edu/airlinedata/www/Revenue&Related.html).

Project Development Environment

The project team will be required to use a cloud-based project development environment such as the College of Engineering Computing (CEC) AWS team-based environment.

If the team selects the CEC AWS team-based environment, at the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their ***minimum*** specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☒ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
⇒

GMU CENTER FOR RESILIENT AND SUSTAINABLE COMMUNITIES (C-RASC)

Project POCs: Dr. Lance Sherry, Professor, College of Engineering and Computing, Systems Engineering & Operations Research department, lsherry@gmu.edu
Dr. Paul Houser, Professor, College of Science, Geography & Geoinformation Science department, phouser@gmu.edu

https://c-rasc.gmu.edu/

The **Center for Resilient and Sustainable Communities (C-RASC)** at George Mason University (GMU) is a transdisciplinary research center established in 2020. C-RASC is dedicated to fostering an inclusive environment that promotes active research collaborations and innovation in the fields of community resilience and sustainability.

C-RASC's mission is to engage stakeholders in developing community-based solutions to resilience challenges. The center emphasizes a bottom-up, community-led approach, addressing resilience in comprehensive and measurable ways. This approach involves integrating the impacts and policy implications of converging, accelerating technological changes. The center collaborates with various partners, including local, regional, and state governments, nonprofits, and other organizations, to support initiatives that turn adversity into opportunity.

The center's work is supported by six colleges and schools within GMU: the College of Health and Human Services, the College of Science, the Schar School of Policy and Government, the School of Business, the Carter School for Peace and Conflict Resolution, and the College of Engineering and Computing.

Project Title	GMU Campus Carbon Absorption Capacity From Computer Vision Analysis of Drone Treetop Imagery
Organization	*Please provide the name of the partnering organization for this project:* GMU Center for Resilient and Sustainable Communities (C-RASC)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Dr. Lance Sherry, Professor, College of Engineering and Computing, Systems Engineering & Operations Research department, lsherry@gmu.edu Dr. Paul Houser, Professor, College of Science, Geography & Geoinformation Science department, phouser@gmu.edu
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☒ Systems Engineering ☒ Data Engineering ☐ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☒ Computer Vision ☐ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☐ 1 ☒ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Advances in drone and camera technology have made it feasible to collect large sets of hyperspectral images using drones. Hyperspectral imaging involves capturing and processing information from across the electromagnetic spectrum. When applied to treetops using drones, this technology can provide detailed insights into various aspects of forest health and management.

Applications of Hyperspectral Imaging of Treetops via Drones

1. **Tree Species Classification:**
   - Hyperspectral images can help differentiate between tree species by capturing unique spectral signatures of each species. This is crucial for forest inventory and biodiversity studies.
2. **Health Monitoring:**
   - By analyzing the spectral data, researchers can detect signs of disease, pest infestations, or nutrient deficiencies in trees. This allows for early intervention and better forest management.
3. **Biomass Estimation:**
   - Hyperspectral data can be used to estimate the biomass of trees (i.e., leaf density), which is important for carbon stock assessments and understanding the role of forests in carbon sequestration.
4. **Environmental Monitoring:**
   - This technology can monitor changes in forest composition and structure over time, providing valuable data for studying the impacts of climate change and human activities on forests.
5. **Precision Forestry:**
   - Hyperspectral imaging supports precision forestry practices by providing detailed information that can guide selective logging, reforestation efforts, and sustainable forest management.

ADVANTAGES OF USING DRONES

1. **High Spatial Resolution:**
    - Drones can capture high-resolution images, allowing for detailed analysis of individual trees and small forest patches.
2. **Flexibility and Accessibility:**
    - Drones can access remote or difficult-to-reach areas, making it easier to gather data from diverse forest environments.
3. **Cost-Effectiveness:**
    - Compared to traditional methods, using drones for hyperspectral imaging can be more cost-effective and time-efficient.

This combination of hyperspectral imaging and drone technology is revolutionizing how we study and manage forests, providing critical data for conservation and sustainable use.

Project Goals

The project team will focus on delivering the following to the partners.

1. **Data Engineering:** Develop a website that accepts the image files and processes the data
2. **Data Analysis/Machine Learning:** Develop algorithms to process and identify tree types and leaf density
3. **Data Visualization:** Develop a dashboard to visualize the results of the analysis

In terms of infrastructure, the project team will implement the solution using the College of Engineering and Computing IT Services-provided AWS environment. The project team should consider various ways in which to process the image data including Jupyter notebooks with Python, AWS SageMaker Studio, Accure Momentum AI workbench, etc. and implement the most cost-effective and easily maintainable solution. The project team needs to also consider whether the dashboard visualizations should be implanted via code (e.g., Python or R) or via a 3rd-party commercial visualization tool (e.g., Tableau Public). The project team can consider this project a prototype systems solution in a development environment but should take into consideration what would be necessary for deploying the solution to a production AWS environment (i.e., DevSecOps).

Data Sources and Datasets

Image dataset to be provided by Dr. Paul Houser.

Partner Intellectual Property

None.

References

None.

Project Development Environment

The project team will be required to use a cloud-based project development environment such as the College of Engineering Computing (CEC) AWS team-based environment or the Office of Research Computing High-Performance Argo or Hopper Clusters.

If the team selects the CEC AWS team-based environment, at the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☒ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
⇒

GMU MEASURABLE SECURITY LAB (MSL) — GMU CENTER FOR ASSURANCE RESEARCH AND ENGINEERING (CARE)

Project POCs: Dr. Eric Osterweil., Director Measurable Security Lab (MSL), eoster@gmu.edu
Dr. Jean-Pierre Auffret, Director Center for Assurance Research and Engineering (CARE), jauffret@gmu.edu

https://msl.cs.gmu.edu/
https://care.gmu.edu/

The **Center for Assurance Research and Engineering (CARE)** and the **Measurable Security Lab (MSL)** at George Mason University (GMU) are both integral parts of the university's efforts to advance cybersecurity research and education.

CARE focuses on developing innovative cybersecurity technologies and policy solutions. It emphasizes a multidisciplinary approach, integrating technology, business, and governance to address cybersecurity challenges. CARE's research is oriented towards practical applications, aiming to transform theoretical research into real-world security solutions1.

The MSL, on the other hand, specializes in creating metrics and methodologies to measure and improve the security of systems. This lab focuses on quantifying security aspects to provide measurable and verifiable security assurances. The MSL's work is crucial for developing standards and benchmarks that can be used to evaluate the effectiveness of security measures2.

By working together, these two entities contribute to a comprehensive approach to cybersecurity at GMU, combining theoretical research, practical application, and measurable outcomes to advance the field.

Project Title	DATA-DRIVEN IMPACT ANALYSIS OF DNSSEC OUTAGES
Organization	*Please provide the name of the partnering organization for this project:* GMU Center for Assurance Research and Engineering (CARE) and the GMU Measurable Security Lab (MSL)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Dr. Eric Osterweil, Associate Director, GMU Center for Assurance Research and Engineering (CARE) and Director, GMU Measurable Security Lab, eoster@gmu.edu Dr. Jean-Pierre Auffret, Director, GMU Center for Assurance Research and Engineering (CARE), jauffret@gmu.edu
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☒ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☒ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Computer/Communications networking, DNS, DNSSEC, network operations, programming.
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☐ 1 ☐ 2 ☒ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Spring 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

The Domain Name System's (DNS') Security Extensions (DNSSEC) is a ~20 year old security protocol, deployed in over 13 million separate administrative domains (zones), worldwide. Recent concerns have arisen that when there are problems in these zones, errors may cause "outages" to portions of the Internet. However, with the extreme redundancy and complex nature of the protocols, failures of components may not result in qualitative outages of services.

In this project, the team will use the ~56 billion measurement longitudinal corpus of active measurements curated by CARE's and MSL's Internet Namespace Security Observatory (INSO) (https://inso.gmu.edu/) to quantify and evaluate the natures of observed failures in conjunction with natural disasters including hurricanes and earthquakes and develop a model of DNS resiliency.

Project Goals

Starting from a reference set of candidate outages which will be provided to the project team:

a) Create classifications of different "types" of outages (from, but not limited to, the reference set),
b) Using existing longitudinal measurement corpus (access to be provided) develop data-driven analyses to confirm/deny/augment outages in the reference set,
c) Correlate outages with natural disasters including hurricanes and earthquakes, and
d) Develop a model of DNS resiliency.

Data Sources and Datasets

CARE/MSL curate an ongoing data set of ~20 years of DNSSEC active measurements (the INSO), which is roughly 54 billion measurements. This data set is globally unique, as it spans back to the beginning of DNSSEC's rollout and does not exist elsewhere. Project teams will be given access to this data set during this project.

Partner Intellectual Property

The data sets are proprietary to the GMU Measurable Security Lab (MSL).

References

1. Osterweil, Eric, Pouyan Fotouhi Tehrani, Thomas C. Schmidt, and Matthias Wählisch. "From the beginning: Key transitions in the first 15 years of DNSSEC." *IEEE Transactions on Network and Service Management* 19, no. 4 (2022): 5265-5283.
2. Osterweil, Eric, Dan Massey, and Lixia Zhang. "Deploying and monitoring dns security (dnssec)." In *2009 Annual Computer Security Applications Conference*, pp. 429-438. IEEE, 2009.
3. Jain, B., Daing, C., Dunaboyina, D., Idris, H., Franzinetti, L., Pati, T., "Data-Driven Impact Analysis of DNSSEC Outages," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.

Project Development Environment

The project team will be required to use a cloud-based project development environment such as the College of Engineering Computing (CEC) AWS team-based environment.

If the team selects the CEC AWS team-based environment, at the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their minimum specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☐ Working Prototype
☐ Other (please specify below)
⇒

GMU MASON STUDENT SERVICES CENTER (MSSC) — GMU INFORMATION TECHNOLOGY SERVICES (ITS)

Project POCs: Andrew Bunting, Executive Director, Enrollment Services, Office of the Provost • GMU, abunting@gmu.edu
Kimberly E. Shumadine, Director, Mason Student Services Center, Office of the Provost • GMU, kshumadi@gmu.edu

Charlie Span, Assistant VP, Enterprise Service Delivery, and Deputy CIO, GMU Information Technology Services, cspann2@gmu.edu

https://mssc.gmu.edu/
https://its.gmu.edu/

The **Mason Student Services Center (MSSC)** at George Mason University is a comprehensive resource hub designed to assist students with a variety of administrative needs. It serves as a "one-stop shop" for services related to registration, enrollment, financial aid, billing, and academic records. The MSSC aims to streamline student support by providing solutions and information in one convenient location, reducing the need for students to visit multiple offices.

George Mason University's **Information Technology Services (ITS)** provides comprehensive technology support and solutions to advance the university's educational and business goals. ITS offers a wide range of services, including technical support, network infrastructure, cybersecurity, and software solutions, to ensure the success of students, faculty, and staff.

Project Title	GENERATIVE AI CHATBOT 6-MONTH PILOT SETUP
Organization	*Please provide the name of the partnering organization for this project:* GMU Mason Student Services Center GMU Information Technology Services
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Andrew Bunting, Executive Director, Enrollment Services, GMU Office of the Provost, abunting@gmu.edu Kimberly E. Shumadine, Director, Mason Student Services Center, GMU Office of the Provost, kshumadi@gmu.edu Greg Grieff, Sr. Solutions Architect – Enterprise Higher Education, AWS, ggrieff@amazon.com Charlie Span, Assistant VP, Enterprise Service Delivery, and Deputy CIO, GMU Information Technology Services, cspann2@gmu.edu
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☐ Data Mining ☐ Data Analytics ☐ Data Modeling/Simulation ☐ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☒ Generative AI ☐ DevSecOps ☒ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Generative AI foundational models (LLM), Retrieval-Augmented Generation (RAG), Prompt Engineering (PE), Amazon Web Services (AWS)
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Spring 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

The GMU Mason Student Services Center (MSSC) (https://mssc.gmu.edu/) is the first stop and the central resource for information and solutions related to registration, enrollment, financial aid, billing, academic records and other student support services. A team of cross-trained Mason Student Services Center Representatives provide assistance to new and continuing students at all points of their academic career, in one convenient location, thus eliminating the need to visit multiple offices on campus. MSSC services are available on both a walk-up and virtual basis.

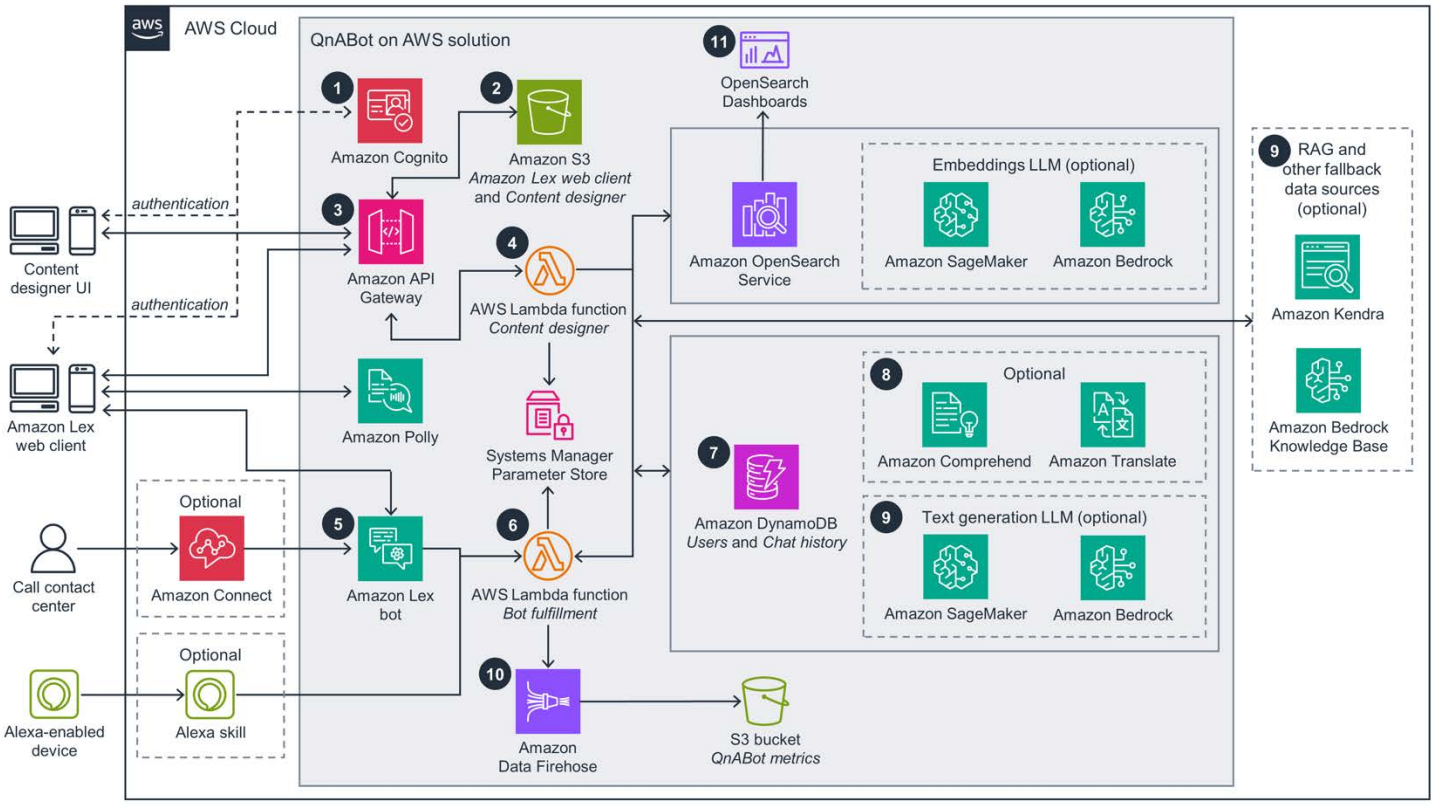
The MSSC maintains a knowledge base (https://mason.my.site.com/SelfServiceHC/s/) of articles containing answers to the most commonly asked questions by students along with a topic catalog (https://mason.my.site.com/SelfServiceHC/s/topiccatalog).

In an effort to make the MSSC knowledge base more user friendly to students, the Spring 2024 Capstone project built a proof-of-concept generative AI chatbot with Retrieval-Augmented Generation (RAG) and Prompt Engineering (PE) techniques used to enhance the performance of Large Language Models (LLMs). The platform utilized an Amazon Web Services (AWS) architecture based on AWS Kendra, AWS QnABot, and AWS Bedrock employing the Anthropic Claude LLM.

Though the Spring 2024 capstone project successfully demonstrated an AWS-based proof-of-concept generative AI chatbot solution, there are additional steps which need to be taken before the solution can be deployed and supported as a production Information Technology Services (ITS) enterprise system.

Project Goals

Building on the results from the Spring 2024 capstone project titled *"Generative AI Chatbot with AWS"* the project team will continue with the Generative AI RAG chatbot prototype using the AWS QnABot solution reference architecture (see figure below). User input to the chatbot will be restricted to a web browser only (i.e., no mobile client development, no call contact center, or Alexa-enabled device).



The AWS services incorporated into the solution will include:

- **Amazon S3** for storage
- **Amazon Lex** for user interactions with QnABot where Amazon Lex processes the input to understand the user's intent and extract relevant information
- **Amazon Kendra** for intelligent search
- **Amazon Bedrock** for high-performing foundation Large Language Models (LLMs)
- **AWS Lambda** for serverless computing
- **Amazon Cognito** for QnABot developer authentications

This Fall 2024 capstone project is a direct follow-up to the Spring 2024 capstone project and focuses on:

a) Updating the previous Generative AI RAG Chatbot with the latest Anthropic Claude LLM version,
b) Updating the search index with new MSSC knowledge base articles,
c) Addressing the Virginia IT Agency (VITA) Policy and Governance standards on Artificial Intelligence use by state agencies,
d) Addressing the GMU ITS Guidance on Using Artificial Intelligence (AI) Tools for Administrative Purposes at Mason (https://its.gmu.edu/knowledge-base/its-guidance-on-using-ai/),
e) Testing and evaluating the quality and effectiveness of the Generative AI chatbot responses, and
f) Producing the operational documentation required for the 6-month limited access pilot.

Data Sources and Datasets

The **Mason Student Services Center (MSSC) knowledge base articles** accessed via screen scraping the knowledge base topic catalog (https://mason.my.site.com/SelfServiceHC/s/topiccatalog).

Partner Intellectual Property

None.

References

1. **Virginia IT Agency (VITA) Commonwealth of Virginia Enterprise Architecture Standard (EA-225)** (https://www.vita.virginia.gov/artificial-intelligence/).
2. **Governor Youngkin's Executive Order 30** (EO 30 PDF | news release).
3. **GMU ITS Guidance on Using Artificial Intelligence (AI) Tools for Administrative Purposes at Mason** (https://its.gmu.edu/knowledge-base/its-guidance-on-using-ai/).
4. **AWS Solutions Library / AWS Solution / QnABot on AWS** (https://aws.amazon.com/solutions/implementations/qnabot-on-aws/).
5. **What is RAG (Retrieval Augmented Generation)?** (https://aws.amazon.com/what-is/retrieval-augmented-generation/)
6. **Amazon Kendra** (https://aws.amazon.com/kendra/).
7. **Amazon Lex** (https://aws.amazon.com/lex/).
8. **Amazon Bedrock** (https://aws.amazon.com/bedrock/).
9. **AWS Lambda** (https://aws.amazon.com/pm/lambda/).
10. Chilakabathini, S., Payaga, K, Staton, B., Vadakattu, A., Veeremalla, N., Vunnam, B., "Generative AI Chatbot with AWS," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

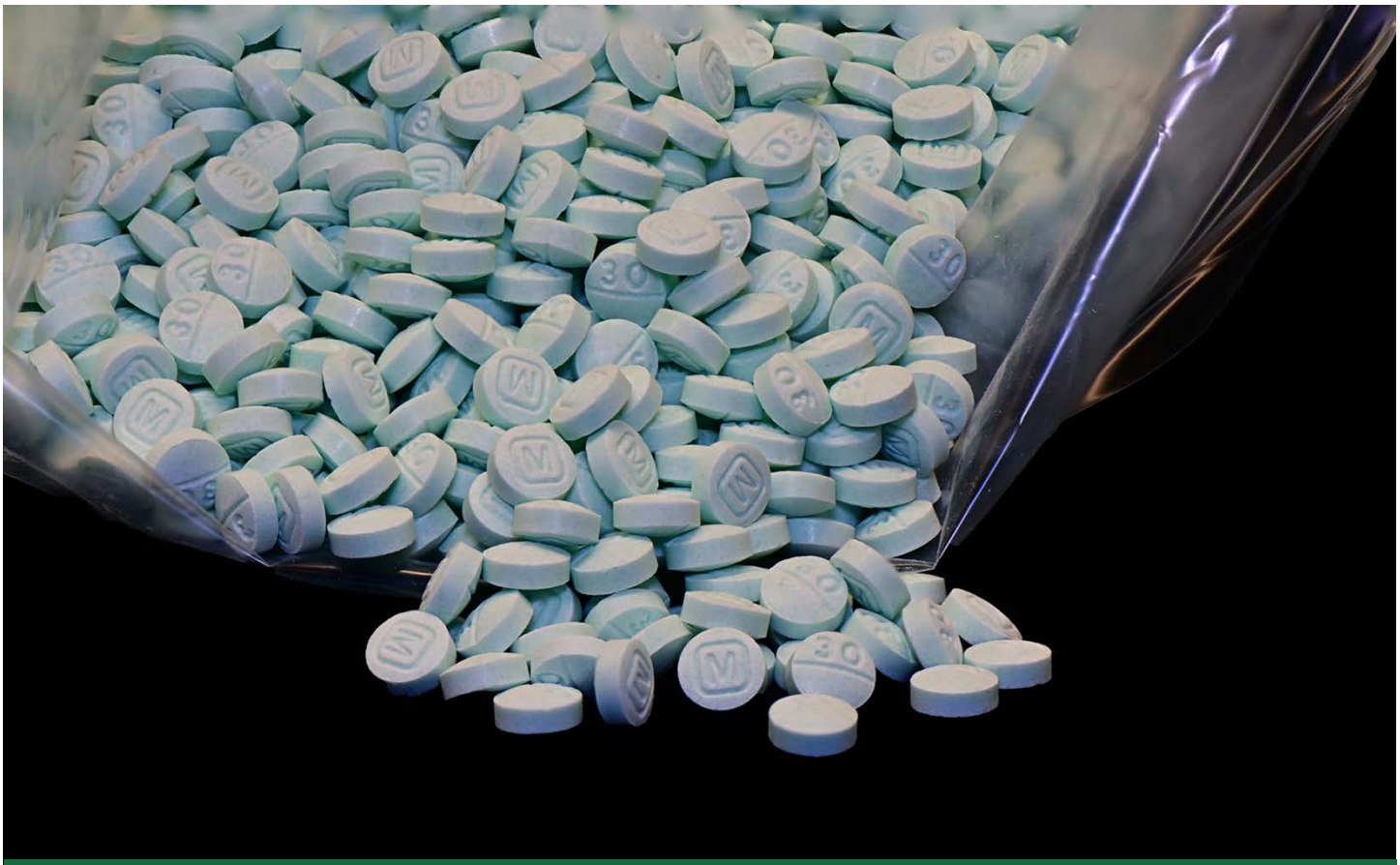
Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☒ Other (please specify below)

⇒ All documentation necessary to address the Virginia IT Agency (VITA) Policy and Governance standards on Artificial Intelligence use by state agencies.

⇒ All documentation necessary to address the GMU ITS Guidance on Using Artificial Intelligence (AI) Tools for Administrative Purposes at Mason.

⇒ All documentation and procedures necessary to operate and maintain a 6-month limited pilot of the MSSC Generative AI RAG Chatbot.



GMU TERRORISM, TRANSNATIONAL CRIME AND CORRUPTION CENTER (TRACCC) — THE MITRE CORPORATION

Project #1 POCs: Dr. Louise Shelley, TraCCC Director, lshelley@gmu.edu
Dr. Randy Howard, Data Director for Investigation Research for TraCCC, choward@gmu.edu and MITRE Enterprise Data Architect / Engineer / Scientist, choward@mitre.org

Project #2 POC: Dr. Mahmut Cengiz, Associate Research Professor, TraCCC, mcengiz@gmu.edu

https://traccc.gmu.edu/
https://www.mitre.org/

The **Terrorism, Transnational Crime and Corruption Center (TraCCC)** at George Mason University is a pioneering research institution dedicated to understanding the complex interplay between terrorism, transnational crime, and corruption. Established within the Schar School of Policy and Government, TraCCC is the first center in the United States to focus exclusively on these interconnected issues. The center's mission is to inform policy, conduct cutting-edge research, and provide training to address these global challenges effectively.

The MITRE Corporation is a not-for-profit organization that operates federally funded research and development centers (FFRDCs) on behalf of U.S. government agencies. Established to advance national security and serve the public interest, MITRE applies systems thinking to solve complex challenges across various domains, including cybersecurity, healthcare, and aviation safety. With dual headquarters in Bedford, Massachusetts, and McLean, Virginia, MITRE collaborates with government, industry, and academia to pioneer innovative solutions for a safer world.

Project Title	INTERDICTING FENTANYL SUPPLY CHAINS
Organization	*Please provide the name of the partnering organization for this project:* GMU Terrorism, Transnational Crime and Corruption Center (TraCCC) and The MITRE Corporation
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Dr. Louise Shelley, Director, GMU TraCCC, lshelley@gmu.edu, Dr. Randy Howard, Data Director for Investigation Research for TraCCC, choward@gmu.edu and MITRE Enterprise Data Architect / Engineer / Scientist, choward@mitre.org
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☒ Systems Engineering ☒ Data Engineering ☐ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☒ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☐ 1 ☐ 2 ☒ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Spring 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

The illicit fentanyl epidemic has emerged as a devastating public health crisis in the United States, claiming over 112,000 lives in 2023 alone[1]. Fentanyl, a synthetic opioid, is significantly more potent than heroin and morphine, making it a preferred choice for illicit drug manufacturers and dealers. Its high potency means that even a small amount can be lethal, leading to a surge in overdose deaths. The widespread availability of fentanyl, often mixed with other drugs without users' knowledge, has exacerbated the crisis, making it difficult for individuals to gauge the risk of overdose. This epidemic has not only overwhelmed healthcare systems but also strained law enforcement and social services, highlighting the urgent need for comprehensive strategies to address the issue.

Efforts to combat the fentanyl epidemic must be multifaceted, involving public health initiatives, law enforcement, and community support. Public health campaigns aimed at raising awareness about the dangers of fentanyl, along with increased access to addiction treatment and harm reduction services, are crucial. Law enforcement agencies need to focus on disrupting the supply chains of illicit fentanyl, while also ensuring that individuals struggling with addiction are treated with compassion and provided with the necessary resources to recover. Community support systems, including mental health services and support groups, play a vital role in helping individuals and families affected by the epidemic. Addressing the fentanyl crisis requires a coordinated and sustained effort from all sectors of society to reduce the number of overdose deaths and support those impacted by addiction.

PROBLEMS WITH INTERDICTION INCLUDE:

- There are insufficient resources appropriated to prevent the illicit flow of Fentanyl into the US.
- Fentanyl precursors (e.g., common analogues, synthetic cannabinoids, etc.) are key ingredients to produce fentanyl.
- PRC corporations, or active producers, are the only suppliers able to produce these pre-cursors at the scale necessary to illicitly produce Fentanyl.
- The PRC government is complicit as they have key associations with these active producers.

THE TECHNICAL PROBLEM SPACE ENTAILS:

- Evidence of this is readily available in Publicly Available Information (PAI) data.
- Data sources and structures far too disparate to manually identify the corporations and their inter-relations at scale.

The Fall 2024 capstone project solution approach continues the Spring 2024 capstone project solution approach to:

- Provide ground truth evidence that influences US policy to interdict the supply of fentanyl pre-cursors; thereby, curbing the influx of fentanyl across the world.
- Employ big data methodologies and technologies must be applied to produce this necessary evidence at scale.
- Structure data into discrete values with discrete labels

Project Goals

To extend the work from the Spring 2024 semester capstone projects the Fall 2024 semester capstone projects goals, along with background information, are detailed below:

- Define Data Models Using LLM
    - The Spring 2024 data is contained in Excel tables. Due to extremely timelines, defining formal data models and data definitions was prohibitive
    - The project explores how LLMs can expedite defining data models for the following structures to support their associated purposes:
        - Relational: organizing and managing the data
        - Key-Value: easy data capture for ever-changing data that is both human and machine readable
        - Graphical: network analysis
        - Geospatial: location analysis
- Extract additional evidence by applying Natural Language Processing (NLP) to web scrapes and Optical Character Recognition (OCR) processing
    - Unstructured data in web scrapes has additional evidence that needs to be gathered via NLP
    - Images also have additional evidence that needs to be gathered via OCR (and using NLP)
- Improving Network Analysis
    - Incorporate WHOIS data into the identifiers (e.g., phone#, emails) network analysis
    - A more reliable manner to process phone numbers needs to be developed
    - A more reliable fuzzy matching of company names is also needed

Data Sources and Datasets

TraCCC Illicit Fentanyl Supply Chain Data.

Partner Intellectual Property

Sensitive to source of data.

References

1. Mann, B., Pattani, A., Bebinger, M., “In 2023 the overdose death toll surpassed 112,000, driven largely by fentanyl,” NPR, Dec. 28, 2023. [Online]. Available: https://www.npr.org/2023/12/28/1220881380/overdose-fentanyl-drugs-addiction#:~:text=In%202023%20the%20overdose%20death,for%20Disease%20Control%20and%20Prevention . [Accessed: Aug. 20, 2024].
2. Amudalapalli, S., Bhau, D., Gududuru, P., Musaligari, A., Pandit, S., “Interdicting Fentanyl: Government Complicit Associations,” presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.

3. Dindigala, B., Gundeti, S., Komirishetty, T., Middhuddi, S., Nalla, D., Penmathsa, G., “Interdicting Fentanyl: Active Producers,” presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.
4. Gangarapollu, S., Jaganath, S., Kurra, A., Malyala, T., Patel, F., Singh, S., “Interdicting Fentanyl: TraCCC Current Operators,” presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.
5. Ampomah, A., Bethi, S., Daddala, Y., Satambakkam, K., Sunkara, T., Vallamkonda, G., “Interdicting Fentanyl: TraCCC Corporate Aliases,” presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☒ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
⇒

Project Title	TERRORISM RESEARCH AND DASHBOARD
Organization	*Please provide the name of the partnering organization for this project:* GMU Terrorism, Transnational Crime and Corruption Center (TraCCC)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Dr. Mahmut Cengiz, Associate Research Professor, GMU Terrorism, Transnational Crime and Corruption Center TraCCC, mcengiz@gmu.edu
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Visualization Dashboard best-practices, User Interface/User Experience (UI/UX) best-practices.
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Since the September 11 attacks, terrorism has been a major threat to global security. While terrorist organizations have not carried out another significant attack on the US since then, their activities have increased in Africa, the Middle East, and Asia. The Global Terrorism Trends and Analysis Center (GTTAC) (https://gttac.com/) and the Terrorism, Transnational Crime and Corruption Center (TraCCC) (https://traccc.gmu.edu/), working with the Department of State, tracks terrorist incidents through a project mandated by Congress to produce an annual report. Since 2018, GTTAC has recorded between 7,000 and 10,000 incidents each year, which have resulted in approximately 25,000 deaths annually. The data, gathered from *reliable media sources* and *meeting specific inclusion criteria*, includes detailed information on over 53,000 incidents, covering aspects such as perpetrators, tactics, targets, weapons, victims, facilities, and logistics. This data is publicly available through the GTTAC Data Portal (https://gttac.com/data/).

This project will utilize GTTAC data and terrorism-specific domain knowledge to address the following research questions:

1. Which regions or countries recorded the most terrorist incidents from 2018 to 2023? Which months saw the highest number of people killed and wounded during the same period?
2. Which ten groups had the highest number of casualties (people killed and wounded) from 2018 to 2023?
3. How many types of tactics (e.g., suicide attacks, kidnapping, car ramming, assassinations, stabbing, shooting, bombing, ambush, and planting mines/IEDs) were used from 2018 to 2023? Which ten groups employed these tactics the most, and how has their use of these tactics changed from 2018 to 2023?
4. What are the most commonly used weapon types (e.g., firearms, explosives, UAVs, incendiary devices, and melee weapons)? Which five groups use firearms, IEDs, rockets, mortars, grenades, missiles, and UAVs the most?
5. Which victim categories are most frequently targeted (e.g., general population/civilians, government officials, professionals, political figures)? Which five groups primarily target civilians, military personnel, politicians, government officials, and professionals?
6. What are the similarities and differences between Iran-backed and jihadist terrorist groups?
7. What tactics have lone actors used from 2018 to 2023? Which ideologies (jihadist, far right, or nationalist) most influence lone actors?

8. Is there a positive correlation between the types of terrorist perpetrators and the types of tactics used?
9. Do weapon types and target types influence the tactics used by terrorist organizations?

Project Goals

The project specific goals fall into two areas.

1. Performing data analysis on the GTTAC data to answer the nine research questions.
2. Researching and creating a terrorism dashboard based on the GTTAC data which displays the following information with an intuitive easy-to-use user interface/user experience (UI/UX). The dashboard should present the following information.
    a. **Countries:** People Killed, Wounded, Kidnapped, the number of perpetrators
    b. **Tactics:** Suicide attacks, kidnapping, car ramming, assassinations, stabbing, shooting, bombing, ambush, and planting mines/IEDs
    c. **Weapon types:** Firearms, IEDs, Rockets, Mortars, Grenades, Missiles, and UAVs
    d. **Perpetrators:** categories (jihadist, right-wing extremist, anarchist. etc.) number of incidents, killed, wounded, kidnapped, suicide attacks, shooting, bombing, ambush, and
    e. **Victim types:** civilians, military, politicians, governments, and professions by countries and perpetrators

In the case of the research questions, the project team will need to consider the various statistical analysis they have learned (or should have learned) in their coursework and propose an analytical method and visualization which addresses each research question. When completing the project report, the project team will be required to document their answers not just with a visualization, but a well-written analysis explaining their basis for their response to include specific domain knowledge of terrorism gleaned from their interactions with the project partner which backs up their analysis. Any visualizations produced will be expected to be of the highest quality and have a uniform look-and-feel across research questions where applicable – meaning that simply using the default visualization settings (either programmatically or using a commercial tool) for creating a visualization will not be acceptable for this project.

In the case of the terrorism dashboard, the project team will research programmatic and commercial dashboard tools. The team will also be expected to research best-practices for dashboard user experiences and visualizations.

Project success will be defined as:

1. The thoroughness of their research question analysis and narrative responses as well as the quality and uniformity of their research question visualizations.
2. The intuitiveness and ease-of-use for accessing the dashboard information as well as the layout, quality, and uniformity of their visualizations and user interface.

Data Sources and Datasets

The primary source of data will be the GMU Terrorism, Transnational Crime and Corruption Center (TraCCC) terrorism database maintained by the Global Terrorism Trends and Analysis Center (GTTAC). The GTTAC Data Portal can be accessed from https://gttac.com/data/. The database is updated weekly on Monday mornings, so the project team will be expected to devise a data engineering solution to access and update data on a weekly basis for their data analysis and their dashboard.

Partner Intellectual Property

None.

References

1. **GTTAC Codebook** (https://gttac.com/wp-content/uploads/2023/11/2023_GRID_Codebook_V2.pdf)
2. **GTTAC Methodology** (https://gttac.com/methodology/)

3. Alapati, N., Daftary, N., Devarapalli, P., Kolanu, S., & Tummala, A., “Terrorism Dashboard,” presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
- ⇒ The processes and procedures documenting the data engineering required to access the GTTAC database weekly.
- ⇒ The processes and procedures documenting the data engineering required for updating and maintaining the Terrorism Dashboard.

GMU VIRGINIA CLIMATE CENTER (VCC)

Project POCs: Sophia Whitaker, Communications Manager, GMU Virginia Climate Center (VCC) swhitak9@gmu.edu
Dr. Luis E. Ortiz, VCC Principal Investigator and Assistant Professor, AOES, lortizur@gmu.edu

https://www.vaclimate.gmu.edu/

The Virginia Climate Center's (VCC) mission is to engage with Virginia's municipal officials, businesses, and other community leaders as well as co-develop information and tools that will inform municipal decisions, enhance Virginia's resiliency, save tax dollars, and improve the productivity and profitability of Virginia's businesses.

Operating as a climate extension service to Virginia municipalities and businesses, the Virginia Climate Center partners with stakeholders to co-produce locally relevant data, products, and services. They aim to help municipalities and businesses adopt preventive and mitigation strategies to protect lives and property, enhance the standard of living, and promote wise resource management and sustainable entrepreneurship. Their large interdisciplinary team of Mason experts conduct research on Virginia's vulnerability and risks to the impacts of climate change in order to provide local decision makers with actionable climate information.

The VCC is a two year, congressionally directed community project funded through the National Oceanic and Atmospheric Administration (NOAA). Initially partnered with Northern Virginia municipalities, VCC is building a network of local governments and businesses throughout the Commonwealth of Virginia to collaborate on best practices and solutions.

Project Title	VIRGINIA CLIMATE INDICATORS
Organization	*Please provide the name of the partnering organization for this project:* GMU Virginia Climate Center (VCC).
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Sophia Whitaker, VCC Communications Manager, swhitak9@gmu.edu Dr. Luis E. Ortiz, VCC Principal Investigator and Assistant Professor, AOES, lortizur@gmu.edu
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☐ Data Engineering ☐ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Summer 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

There is both too much information and hard to reach information for municipal planners looking to take on mitigation and adaptation work for their municipalities. Information can be hard to access. Information can be at too high a scale (state level verse county, or national verse state). Information can also be in too great a size or in an unusable format. This information is needed to enable municipal planners to (more easily) take on planning and preparing their municipalities for the impacts of climate change. This includes the ability to easily produce key information points that are easily visualized to support community education and support of these types of planning efforts. End user would be planners/practitioners in Virginia.

Project Goals

The goal is to create a series of usable climate data products (actionable insights) for practitioners in the State of Virginia for the purpose of climate adaptation and resilience planning activities. The proposed data product is a dashboard, broken down for each county in Virginia, looking at historical trends for temperature (extremes, humidity) and precipitation (extremes and totals). It could also include possible projections (multimodel ensemble) looking at future temperatures/precipitation depending on students' comfort with the data. Additionally, depending on capacity and availability of data, this project could factor in vector borne disease (disease resulting from mosquito and tick bites).

Data Sources and Datasets

1. Weather station records from NOAA (~MB-GB).
2. Climate reanalysis data (GB - TB).
3. Global model projections (GB-TB).
4. Data is encoded in csv files or packaged in self-describing binary formats (e.g. NetCDF, HDF5, TIFF).

Partner Intellectual Property

None.

References

1. Bekele, K., Comer, C., Gutierrez, J., Khan, S., Macharia, A., Nguyen, P., "Virginia Climate Center Data Portal Dashboard," presented at the Summer 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, August 2, 2024.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
⇒

HUMAN-CYBER PERFORMANCE TECH, LLC

Project POC: Dr. Curt Rasmussen, Lead Researcher, rasmussenc@outlook.com

Human-Cyber Performance Tech, LLC focuses on the intersection of human performance and cyber technology, aiming to enhance human capabilities through advanced technological solutions. This involves integrating cyber systems with human performance metrics to optimize efficiency, productivity, and overall performance in various industries. Their work is particularly relevant in fields that require high levels of precision and reliability, such as defense, healthcare, and industrial operations.

Human-Cyber Performance Tech, LLC leverages a multidisciplinary approach, combining expertise from engineering, computer science, psychology, and social sciences to develop innovative solutions that augment human abilities. Their goal is to create systems that not only support but also enhance human performance, making tasks easier, safer, and more efficient.

Project Title	ARTIFICIAL INTELLIGENCE ALGORITHM TAXONOMY FOR HUMAN-SYSTEM INTEGRATION
Organization	*Please provide the name of the partnering organization for this project:* Human-Cyber Performance Tech, LLC
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Dr. Curt Rasmussen, Lead Researcher, rasmussenc@outlook.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☐ 1 ☒ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

A taxonomy (or other structure) of artificial intelligence (AI) algorithms for choosing algorithms that best support different human-systems integration (HSI) projects currently does not exist. At present, algorithm selection often rests on the developers' knowledge of the different types of algorithms and their usage. Not only is the reliance upon developers' knowledge of algorithms and algorithm application but could affect the ability to explain how the AI develops its output. Developing trust in algorithms is a common goal of explainable AI (XAI). However, the difficulty of developing XAI increases with inconsistent algorithm usage. By developing a taxonomy (or other structure) of AI algorithms for HIS could help with development of effective XAI and the integration of AI in the workplace.

Project Goals

The project team will focus on achieving the following project goals:

1. Develop a taxonomy (or other structure) of AI algorithms for supporting HSI projects.
2. Identify if determining where the human is in the loop is necessary for XAI.
3. Identify whether differing levels of explanation are necessary for trust to be built based on:
    a. Occupation
    b. Industry

Data Sources and Datasets

1. O*NET OnLine Career Data (https://onetonline.org)
2. NAICS, North American Industry Classification System (NAICS) U.S. Census Bureau (https://www.census.gov/naics/)
3. Overview of businesses by NAICS, NAICS & SIC Identification Tools | NAICS Association (https://www.naics.com/search/)
4. OpenML AI algorithms (https://www.openml.org/)

Partner Intellectual Property

None.

References

1. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., . . . Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99. doi:10.1016/j.inffus.2023.101805.
2. Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence Program. AI Magazine, 40(2), 44-58.
3. Gunning, D., Vorm, E., Wang, Y., & Turek, M. (2021). DARPA's Explainable Artificial Intelligence Program: A retrospective. FAccT 2021 Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 598-609. doi:10.1145/3442188.344.5921.
4. Hind, M. (2019). Explaining explainable AI. XRDS Spring, 16-19. doi:10.1145/3313096.
5. Hu, B., Tunison, P., Vasu, B., Menon, N., Collins, R., & Hoogs, A. (2021). XAITK: The Explainable AI Toolkit. Special Issue: DARPA's Explainable Artificial Intelligence (XAI) Program, 2(4). doi:10.1002/ail.40.
6. Ribes, D., Henchoz, N., Portier, H., Defayes, L., Phan, T., Gatica-Perez, D., & Sonderegger, A. (2021). Trust indicators and explainable AI: A study on user perceptions. 18th IFIP Conference on Human-Computer Interaction (INTERACT), (pp. 662-671). Bari, Italy. doi:10.1007/978-3-030-85616-8_39.
7. Tiwari, R. (2023). Explainable AI (XAI) and its application in building trust and understanding in AI decision making. International Journal of Scientific Research in Engineering and Management, 7(1). doi:10.55041/IJSREM17592.
8. Tsiakas, K., & Murray-Rust, D. (2022). Using human-in-the-look and explainable AI to envisage new future work practices. PETRA, 588-594. doi:10.1145/3529190.3534779.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☐ Working Prototype
☐ Other (please specify below)
⇒

Project Title	EXPLAINABLE AI: BUILDING TRUST BY INDUSTRY AND OCCUPATION
Organization	*Please provide the name of the partnering organization for this project:* Human-Cyber Performance Tech, LLC
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Dr. Curt Rasmussen, Lead Researcher, rasmussenc@outlook.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☐ 1 ☒ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Explainability is one of the current challenges of integrating artificial intelligence (AI) into the workplace. According to Tiwari (2023), a growing need for Explainable AI (XAI) exists to help build trust and understanding of human employees. However, several challenges exist including defining what XAI is, whether the humans' position in the decision loop affects the trust needed, and how detailed the explanation for the AI needs to be. Additionally, a knowledge gap exists whether challenges differ between occupations (e.g., executive assistant, production line worker) and industries (e.g., manufacturing, hospitality).

Project Goals

The project team will focus on achieving the following project goals:

1. Develop a definition of Explainable AI (XAI) in the context of the project.
2. Identify the point Explainable AI (XAI) becomes necessary and relevant to the human in the loop.
3. Identify whether differing levels of explanation are necessary for trust to be built based on:
    a. Occupation
    b. Industry
4. Could any of the following Explainable AI (XAI) tools help explainability or build trust within a specific occupation or industry? If yes, then explain the reasoning. If no, then explain the reasoning.
    a. LIME or Local Interpretable Model-Agnostic Explanations is a technique developed by researchers from the University of Washington. It helps attain a higher level of transparency within an algorithm. LIME - Local Interpretable Model-Agnostic Explanations – Marco Tulio Ribeiro – (washington.edu)
    b. DeepLIFT is a comparative technique for activation of each neuron to its "reference activation". It also allocates resources to contribute scores according to their comparisons. GitHub - kundajelab/deeplift: Public facing deeplift repo
    c. Shapley Value SHAP or SHapley Additive exPlanations refers to the aggregate marginal contributions within a feature value for various coalitions. An introduction to explainable AI with Shapley values — SHAP latest documentation

d. AIX360 or AI Explainability 360 offers an open-source library. IBM develops this framework to enable the interpretability and explainability of various datasets in a machine learning model. It functions as a Python package that includes comprehensive algorithms. Moreover, these algorithms monitor various dimensions of explanations and their proxy explainability metrics. GitHub - Trusted-AI/AIX360: Interpretability and explainability of data and machine learning models
e. Activation Atlases is one of the most robust Explainable AI Frameworks. Google collaborates with OpenAI to develop this novel technique to visualize the interaction between neural networks. It also monitors the way neural networks expand their horizon with information and various layers. Introducing Activation Atlases | OpenAI
f. Rulex is a company that develops predictive models for first-order conditional logic rules. Moreover, it
g. helps to provide immediate comprehensive results for everyone to use. Explainable AI: why it is important for business - Rulex

Data Sources and Datasets

1. O*NET OnLine Career Data (https://onetonline.org)
2. NAICS, North American Industry Classification System (NAICS) U.S. Census Bureau (https://www.census.gov/naics/)
3. Overview of businesses by NAICS, NAICS & SIC Identification Tools | NAICS Association (https://www.naics.com/search/)
4. United States Census Bureau | Business Counts (https://www.census.gov/topics/business-economy/counts.html)

Partner Intellectual Property

None.

References

1. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., . . . Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99. doi:10.1016/j.inffus.2023.101805.
2. Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence Program. AI Magazine, 40(2), 44-58.
3. Gunning, D., Vorm, E., Wang, Y., & Turek, M. (2021). DARPA's Explainable Artificial Intelligence Program: A retrospective. FAccT 2021 Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 598-609. doi:10.1145/3442188.344.5921.
4. Hind, M. (2019). Explaining explainable AI. XRDS Spring, 16-19. doi:10.1145/3313096.
5. Hu, B., Tunison, P., Vasu, B., Menon, N., Collins, R., & Hoogs, A. (2021). XAITK: The Explainable AI Toolkit. Special Issue: DARPA's Explainable Artificial Intelligence (XAI) Program, 2(4). doi:10.1002/ail.40.
6. Ribes, D., Henchoz, N., Portier, H., Defayes, L., Phan, T., Gatica-Perez, D., & Sonderegger, A. (2021). Trust indicators and explainable AI: A study on user perceptions. 18th IFIP Conference on Human-Computer Interaction (INTERACT), (pp. 662-671). Bari, Italy. doi:10.1007/978-3-030-85616-8_39.
7. Tiwari, R. (2023). Explainable AI (XAI) and its application in building trust and understanding in AI decision making. International Journal of Scientific Research in Engineering and Management, 7(1). doi:10.55041/IJSREM17592.
8. Tsiakas, K., & Murray-Rust, D. (2022). Using human-in-the-look and explainable AI to envisage new future work practices. PETRA, 588-594. doi:10.1145/3529190.3534779.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☐ Working Prototype
☐ Other (please specify below)
⇒

NIRA, INC.

Project POCs: Wen Zhu, Chief Architect, wzhu@nira-inc.com
John Oh, .Net Developer, joh@nira-inc.com

https://www.nira-inc.com/

Nira, Inc. is a dynamic and innovative company specializing in enterprise program management and IT solutions. As a participant in the Small Business Administration's (SBA) 8(a) Program and a Disadvantaged Women-Owned Small Business, Nira is committed to delivering high-quality services tailored to meet the unique needs of its clients. The company excels in financial management, system engineering, and the development of state-of-the-art software solutions based on open standards and agile methodologies.

Nira's expertise extends to supply chain information discovery, case management and workflow, and inter-organizational information sharing. The company has been recognized for its innovative solutions, such as its supply chain transaction platform, which enhances food safety and traceability. Nira's commitment to excellence is further demonstrated by its recent award of the GSA OASIS+ contract in both the Management and Advisory Domain and the Technical and Engineering Domain.

Project Title	PROACTIVE IDENTIFICATION OF PRODUCT SAFETY ISSUES
Organization	*Please provide the name of the partnering organization for this project:* NIRA, Inc.
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Wen Zhu, Chief Architect, wzhu@nira-inc.com John Oh, .Net Developer, joh@nira-inc.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☐ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Natural Language Processing (NLP), Data Virtualization/Business Intelligence (BI), Generative AI (LLM)
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

To derive actionable insights from customer review data to enable early detection of consumer safety product issues and automatically identify product safety issues using NLP (Natural Language Processing) on Amazon's customer review data and matching it with SaferProducts.gov complaints and recall descriptions from the Consumer Product Safety Commission (CPSC). Optionally, to identify potential approaches to mitigate the risks identified leveraging Generative AI.

Project Goals

The project team will focus on achieving the following project goals:

1. Leveraging Machine Learning and NLP to predict product recalls based on online reviews and consumer feedback, for proactive identification of product risks to consumers and supplier risks for important consumer products. This project will explore various methods to predict product risks, like sentiment analysis, anomaly detection, clustering, similarity analysis, or even a combination of these techniques.
2. Exploring Generative AI's ability to propose mitigation strategies for product safety risks identified. This project will explore several approaches to Generative AI, including Retrieval-Augmented Generation (RAG), integrating generative AI with knowledge graphs to connect risks, products, and mitigation strategies, or using prompt engineering to create effective strategies (i.e., design specific prompts that guide the model to generate actionable and practical mitigation strategies based on identified risks).

Data Sources and Datasets

1. Amazon Reviews Dataset, https://amazon-reviews-2023.github.io/
2. CPSC Recalls and Unsafe Product Reports: https://www.saferproducts.gov/SPDB.zip

Partner Intellectual Property

None.

References

1. Haque, Tanjim Ul, Nudrat Nawal Saber, and Faisal Muhammad Shah. "Sentiment analysis on large scale Amazon product reviews." *2018 IEEE international conference on innovative research and development (ICIRD)*. IEEE, 2018. https://www.researchgate.net/publication/325756171_Sentiment_analysis_on_large_scale_Amazon_product_reviews
2. Bleaney, Graham, et al. "Auto-detection of safety issues in baby products." *Recent Trends and Future Technology in Applied Intelligence: 31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2018, Montreal, QC, Canada, June 25-28, 2018, Proceedings 31*. Springer International Publishing, 2018. https://arxiv.org/pdf/1805.09772.pdf

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☐ Working Prototype
☐ Other (please specify below)
⇒

PEARMUND CELLARS WINERY

Project POCs: Chris Pearmund, Managing Partner, Chris@pearmundcellars.com
John Memoli, General Manager, John@pearmundcellars.com
Mark Ward, Wine Maker, Mark@pearmundcellars.com

https://www.pearmundcellars.com/
https://www.effinghammanor.com/
https://www.vinthillcraftwinery.com/

Pearmund Cellars is located in the beautiful foothills of eastern Fauquier County, VA, conveniently close to Northern Virginia and Washington, DC. Their 7500-square-foot geothermal winery and 25-acre vineyard produces Chardonnay, Viognier, Riesling, Late Harvest Vidal, Merlot, Cabernet Franc, Cabernet Sauvignon, Petit Verdot, Ameritage, and other award-winning Virginia wines.

Established in 1976, the vineyard originally hosted nine different grape varieties. Today they stick to chardonnay on the property, as they are the most successful grapes to grow with respect to this terroir. They have 15 acres and 11,000 vines of Chardonnay, cultivars or clones of Chardonnay to help increase the complexity of wine. They also source from the premier vineyards of Virginia that specialize in one particular grape variety.

Chris Pearmund has been instrumental in the founding of more than a dozen new wineries and countless vineyards, and is currently Managing Partner at Pearmund Cellars, Vint Hill Craft Winery, and Effingham Manor and Winery along with General Manager John Memoli and Winemaker Mark Ward.

Project Title	PEARMUND CELLARS WINERY DATA WAREHOUSE
Organization	*Please provide the name of the partnering organization for this project:* Pearmund Cellars Winery
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* John Memoli, Operations Manager, Pearmund Cellars Winery, john@pearmundcellars.com Chris Pearmund, Managing Partner, Pearmund Cellars Winery, chris@pearmundcellars.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☐ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☐ Generative AI ☒ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Data warehouse design and implementation best practices for a small craft winery business.
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Spring 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Pearmund Cellars Winery, located in Broad Run, Virginia, has been one of Virginias top wineries for over 20 years and has gathered over 200 gold medals in international competitions and national acclaim for its production of 100% Virginia grown wine. The winery's vineyard, Meriwether, is the oldest Chardonnay vineyard in Virginia producing wine from its 30 acres for over 40 years. Pearmund Cellars produces a variety of wines from Virginia grown fruit such as Viognier, Petit Manseng, Cabernet Franc, Petit Verdot and more. Their signature Ameritage, a blend of 5 different Bordeaux varietals, has won Best in Class at the Tasters Guild International.

Chris Pearmund, managing partner and founder has consulted on over 20 winery openings in the state; most recently Effingham Manor located in Nokesville, Virginia. A sister winery to Pearmund Cellars, it marries the history of Virginia and the history of Virginia winemaking. Using varietals found in Virginia such as Traminette, Viognier, Petit Verdot and Tannat, Effingham Manor tells the story of Virginia's wine history. Together Effingham and Pearmund provides guests with education, history, quality and service unrivaled in the Virginia Wine business, and carry a reputation as the leaders in the Virginia Wine Industry.

There are currently a total of three wineries associated with Chris Pearmund – Pearmund Cellars Winery (https://www.pearmundcellars.com/), Effingham Manor Winery (https://www.effinghammanor.com/), and Vint Hill Craft Winery (https://www.vinthillcraftwinery.com/).

Pearmund Cellars Winery uses several distribution channels for the sale of their wines – wholesale, Internet, and direct-to-consumer – with various 3rd-party information technology systems used to track those distribution channels. Unlike national wineries based out of California where the bulk of wine sales are through the wholesale and Internet distribution channels, Virginia craft wineries experience the bulk of their wine sales through the direct-to-consumer distribution channel via wine clubs, tasting rooms, and on-site restaurants.

During the Fall 2023 DAEN Capstone project John Memoli, the wineries' operations manager, recognized that the direct-to-consumer (D2C) point-of-sale (POS) system at that time – ShopKeep (now known as ShopKeep by Lightspeed) – did not have the

integrated software features nor the data capture and reporting capability necessary to provide the operational insights into their winery direct-to-consumer operations. This is primarily because ShopKeep is tailored for:

- Retail Shops
- Quick-Service Restaurants
- Full-Service Restaurants and Bars

After a brief evaluation period John selected a new D2C system, OrderPort (https://orderport.net/), which is a specialized point-of-sale system designed primarily for wineries. OrderPort is tailored specifically for:

- Wineries
- Tasting Rooms
- Wine Clubs

Additional features which make OrderPort attractive to Virginia craft wineries include:

- **State Compliance Validation:** Ensures compliance with state regulations for D2C (Direct-to-Consumer) wine sales.
- **Tasting Tracking and Sample Management:** Helps manage tasting sessions and sample bottles.
- **Club Member Management:** Provides alerts for expired club member credit cards, orders waiting for pick-up, and more.

OrderPort is a cloud-based system that is intended as a single winery solution and is not designed to support multiple wineries under the same owner. During the Spring 2024 DAEN Capstone this limitation became a hinderance to the project team attempting to perform data analytics and visualizations on sales across both Pearmund Cellars and Effingham Manor wineries since many of their customers visit both Pearmund Cellars and Effingham Manor and are even wine club or barrel club members at both wineries. Even though a small percentage of customers also visit Vint Hill Craft winery or are wine club members, that winery was not part of the data analysis.

In order to simplify the data analysis and visualizations during the Spring 2024 DAEN Capstone, the project team attempted to create a rudimentary data warehouse using Microsoft SQL Server. No formal data warehouse analysis was performed and what was produced by the project team was inadequate for use as a production data warehouse system.

Project Goals

For the Fall 2024 DAEN Capstone, the focus will be on the data engineering associated with creating a formal production data warehouse which can be maintained by Pearmund Cellars Winery after the DAEN Capstone project has concluded. The data warehouse will involve the Pearmund Cellars Winery and Effingham Manor Winery only and will exclude the Vint Hill Craft Winery.

Four primary data sources will feed into the data warehouse:

1. Pearmund Cellars Winery Tock reservation system
2. Pearmund Cellars Winery OrderPort system
3. Effingham Manor Winery Tock reservation system
4. Effingham Manor Winery OrderPort system

The project team will research and implement best practices into the design and implementation of a production data warehouse for a small craft winery. Creating a data warehouse for a small craft winery business involves several key steps.

1. **Define Business Requirements**: Identify your business goals and the specific objectives for the data warehouse. Determine the types of data you need and the key users and stakeholders.
2. **Choose a Platform**: Select the optimal platform and technology stack for your data warehouse. This could be an on-premises solution or a cloud-based service.

3. **Design the Data Model**: Create a data model that defines how data will be structured and organized in the warehouse. This includes designing tables, relationships, and schemas.
4. **Build the ETL Pipeline**: Develop the Extract, Transform, Load (ETL) processes to move data from various sources into the data warehouse. This involves extracting data, transforming it into a suitable format, and loading it into the warehouse.
5. **Develop Reporting and Analytics**: Implement tools for querying and reporting to extract actionable insights from the data. This could include dashboards, visualizations, and ad-hoc query capabilities.
6. **Implement Ongoing Maintenance and Optimization**: Regularly maintain and optimize the data warehouse to ensure it continues to meet business needs. This includes monitoring performance, updating data models, and ensuring data quality.

These steps provide a foundational approach to building a data warehouse, but the specifics can vary based on the unique needs and goals of the winery. The project team is expected to conduct research and provide a detailed set of recommendations throughout the duration of the capstone project. The following is a non-exhaustive list of tasks the team can expect to complete the during each of the Sprints. It is expected that the project team will create detailed tasking during the first two days of each Sprint.

- **Sprint 1** – The project team will focus on conducting multiple interviews with Pearmund Cellars Operations Manager, John Memoli, to define detailed business requirements necessary for the selection of a platform and overall system design. This will include the documenting and detailed understanding of how the wine club programs operate at each winery. The team will be expected to produce, at a minimum, a *professional* systems architecture diagram and data flow diagram of the data warehouse solution; as well as any additional design documents associated with the requirements analysis and design of the system.
- **Sprint 2** – The project team will focus on the design of the data model and the building of the ETL pipeline with an emphasis on easily maintained ETL operational processes and procedures.
- **Sprint 3** – The project team will focus on implementing the data warehouse and developing simple reporting and analytics to test the correctness and operational effectiveness of their data warehouse design.
- **Sprint 4** – The project team will focus on implementing ongoing maintenance and optimization of the data warehouse.
- **Sprint 5** – The project team will focus on completing the final report and preparing for the public showcase presentation at the end of the semester.

Capstone project success will be defined by whether the project team creates a data warehouse system that Pearmund Cellars Winery will be able to take over and operationally maintain after the conclusion of the capstone project.

Data Sources and Datasets

The source data for the data warehouse will come from the data feeds of the following systems:

1. Pearmund Cellars Winery Tock reservation system
2. Pearmund Cellars Winery OrderPort system
3. Effingham Manor Winery Tock reservation system
4. Effingham Manor Winery OrderPort system

The project team *will pay particular attention* to any **Personally Identifiable Information (PII)** that is part of any operating business (e.g., customer and staff information) system and will take steps to follow and maintain appropriate PII protocols when discussing the project both inside and outside the classroom. This does not apply when discussing the project directly with Pearmund Cellars staff or the course instructor.

Partner Intellectual Property

All data residing in Pearmund Cellars Winery and Effingham Manor Winery business systems.

References

There are multiple resources freely available to students on the design and implementation of data warehouses. The project team will be expected to conduct and report on research *throughout the semester* (not just during Sprint 1) regarding best practices for implementing a data warehouse for a small craft winery business.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

- ☒ Capstone Showcase Presentation & PowerPoint Slides
- ☒ Final Project Report
- ☒ Repository for Data and Code Artifacts
- ☐ Machine Learning Model
- ☒ Working Prototype
- ☐ Other (please specify below)
  - ⇒ All requirements analysis, design, and operational documentation expected in the production implementation of a small craft winery data warehouse.

PRECISE SOFTWARE SOLUTIONS, INC.

Project #1 POCs: Ben Duan, Chief Technology Officer, ben.duan@precise-soft.com
Eckart Bindewald, Lead Data Scientist, eckart.bindewald@precise-soft.com

Project #2 POC: Xu Yang, Vice President of Engineering, xu.yang@precise-soft.com

https://preceise-soft.com/

Precise Software Solutions, Inc. (Precise), an SBA 8(a) program participant, is an innovative small business with a proven record of success delivering quality services and solutions to government organizations. A CMMI Level 3 company, Precise serves as a trusted advisor to senior technology executives and helps government agencies enhance and expand their information technology capabilities. Precise helps their customers capitalize on the efficiencies offered by technological advancements and ensures the integrity of their IT systems and programs so they can perform their public mission more effectively. The company is known for delivering agile and innovative solutions and specializes in strategic consulting, system modernization and integration, digital transformation and experience, infrastructure and cloud implementation, and data management and analytics.

Project Title	LARGE LANGUAGE MODELS FOR KNOWLEDGE GRAPH EXTRACTION AND REASONING BASED ON COMPLEX DATA
Organization	*Please provide the name of the partnering organization for this project:* Precise Software Solutions
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Ben Duan, Chief Technology Officer, ben.duan@precise-soft.com Eckart Bindewald, Lead Data Scientist, eckart.bindewald@precise-soft.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☒ Data Modeling/Simulation ☐ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☒ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Combining AI in form of Large Language Models (LLM) with existing information like biomedical data and unstructured text and converting that to a knowledge graph is an important goal as it facilitates sophisticated reasoning on complex real-world data that is not limited in terms of dataset size and can lead to data with increased interpretability. It involves building synthetic reference data sets, determine performance metrics of AI/LLM-based knowledge extraction systems, implementation of different knowledge extraction systems and implementation of proof-of-concept system.

Project Goals

The project team will focus on achieving the following project goals:

1. Implement performance metric for question answering based on AI-based knowledge graph extraction.
2. Implement LLM-based question answering based on a database containing a knowledge graph.
3. Develop approach extracting knowledge graphs from unstructured texts.
4. Develop approach to harmonize, merge and query knowledge graphs.
5. Develop approach to convert a knowledge graph back to unstructured texts.

Data Sources and Datasets

1. Reference data sources that have downloadable data like bioarxiv.
2. Reference data in form of structured data and external databases.

Partner Intellectual Property

None.

References

None.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☒ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
⇒

Project Title	PRODUCT LABEL RECOGNITION FOR FDA
Organization	*Please provide the name of the partnering organization for this project:* Precise Software Solutions
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Xu Yang, Vice President of Engineering, xu.yang@precise-soft.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☐ Data Visualization ☒ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* LLMs, Computer Vision, Database design, API, RAG and Knowledge Graphs(LLMs)
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Spring 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

An FDA product code describes a specific product and contains a combination of five to seven numbers and letters. This code is composed of five components: Industry Code, Class, Subclass, Process Indicator Code (PIC), and Product (Group). The product code submitted with each FDA line item should accurately match the actual product name and/or invoice description. Products with multiple names (e.g., a fish known by different regional names) may have several synonymous definitions associated with them.

Currently, FDA investigators use a product code builder tool to manually create product codes for all examined or collected products during field activities. This manual process can be time-consuming and prone to errors. Automating this process using Machine Learning algorithms could significantly enhance work efficiency by leveraging information such as product labels and packaging.

During the Spring 2024 capstone project, a prototype was developed to classify 4 out of the 5 elements of the FDA product code using an OpenAI model. While the initial results were promising, the model's accuracy and consistency need improvement. Additionally, one element of the product code string still requires development.

Project Goals

The project team will focus on achieving the following project goals:

1. **Data Engineering and Mining:** Collecting product information and images from the Internet to build a comprehensive dataset.
2. **Algorithm Development:** Develop and improve machine learning algorithms to classify FDA product codes based on product images and labels using ai.
3. **Web-based Application:** Develop and enhance a web-based application to demonstrate the capability of the automated system.
4. **Complete FDA Product Code Detection:** Continue work to detect all five components of the FDA product code.
5. **Model Validation:** Validate model performance using reliable statistical methods to ensure accuracy and consistency.

6. **Workflow and Performance Improvement:** Improve the workflow and model performance through prompt design, better database management, data mining, and building LLM applications with retrieval-augmented generation (rag) or graph knowledge frameworks.
7. **Mobile Application Prototype (optional):** Develop a mobile application prototype to extend the system's accessibility and usability.

Data Sources and Datasets

1. Photos of product labels and packages.

Partner Intellectual Property

None.

References

1. **FDA Product Codes and Product Code Builder** (https://www.fda.gov/industry/import-program-tools/product-codes-and-product-code-builder#structure).
2. Bothra, V., Burugu, S., Diddi, S., Moturu, S., Sinha, S., Sourivong, M., "Product Label Recognition," presented at the Spring 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, April 30, May 3, & May 6, 2024.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☒ Machine Learning Model
☒ Working Prototype
☒ Other (please specify below)
⇒ Web-based Application.

PUERTO RICO SCIENCE, TECHNOLOGY & RESEARCH TRUST — CARIBBEAN CENTER FOR RISING SEAS

Project POCs: Carlos I. Gómez Borges, Engineering Specialist, cgomez@prsciencetrust.org
Gilberto Guevara Velázquez, Senior Manager, gguevara@prsciencetrust.org

https://prsciencetrust.org/
https://prsciencetrust.org/ccrs/

The **Puerto Rico Science, Technology & Research Trust (PRSTRT)** is a nonprofit organization established in 2004 under Public Law 214. Its primary mission is to foster innovation, promote the commercialization of technology, and create high-tech jobs in Puerto Rico. The Trust aims to advance the island's economy and improve the well-being of its citizens by investing in and facilitating the growth of science and technology sectors.

The **Caribbean Center for Rising Seas (CCRS)** is a pivotal initiative by the Puerto Rico Science, Technology & Research Trust (PRSTRT), launched in 2021. The center's mission is to prepare Puerto Rico and the broader Caribbean region to adapt and thrive amidst the increasing risks of flooding due to storms, tides, and sea level rise. This initiative aims to position Puerto Rico as a leader in resilience and innovation, addressing the urgent challenges posed by climate change. The CCRS focuses on resiliency and adaptation by promoting design guidelines and best practices for the built environment. This includes creating Safe Advanced Flood Estimates (SAFE™) to guide coastal design, planning, and engineering with a margin of safety over the latest projections for flooding. The center collaborates with professionals in architecture, engineering, finance, and legal fields to ensure that buildings and infrastructure are designed to withstand future flooding events.

Project Title	PUERTO RICO SEA LEVEL RISE CENTER - DATA PORTAL DASHBOARD
Organization	*Please provide the name of the partnering organization for this project:* Puerto Rico Science, Technology & Research Trust – Caribbean Center for Rising Seas (San Juan, Puerto Rico)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Carlos I. Gómez Borges, Engineering Specialist, cgomez@prsciencetrust.org Gilberto Guevara Velázquez, Senior Manager, gguevara@prsciencetrust.org
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☒ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:* Data analytics, Python, LLM, Dashboard creation, Power BI
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

The **Puerto Rico Science, Technology & Research Trust (PRSTRT)** is a nonprofit organization established in 2004. Its mission is to promote innovation, technology transfer, and the creation of high-tech jobs in Puerto Rico. The Trust focuses on several key areas, including entrepreneurship, public health, and research and development. By investing in these areas, PRSTRT aims to advance Puerto Rico's economy and improve the well-being of its citizens.

The **Caribbean Center for Rising Seas (CCRS)** is a program under the PRSTRT, launched to address the increasing flood risks from storms, tides, and sea level rise in Puerto Rico and the Caribbean. The CCRS aims to prepare the region to adapt and thrive in this new era of climate challenges. It focuses on promoting resilient design and construction practices, creating educational courses, and developing a global resource hub for adaptation solutions. The Center also works on policy advocacy to ensure durable infrastructure and communities.

Both organizations play a crucial role in fostering innovation and resilience in Puerto Rico and the broader Caribbean region.

Puerto Rico faces significant climate challenges that impact its environment, economy, and communities. Here are some of the key issues:

1. **Rising Temperatures:** Since 1950, temperatures in Puerto Rico have increased by approximately 2°F. This rise in temperature contributes to more frequent and intense heatwaves, which can affect public health, agriculture, and energy demand.
2. **Stronger and More Frequent Hurricanes:** Puerto Rico's location in the Caribbean makes it particularly vulnerable to hurricanes. Climate change has led to more powerful storms, such as Hurricane Maria in 2017 and Hurricane Fiona in 2022, which caused widespread devastation, including loss of life, infrastructure damage, and long-term power outages.
3. **Rising Sea Levels:** Sea level rise poses a significant threat to Puerto Rico's coastal areas. This can lead to increased flooding, erosion, and damage to coastal infrastructure. Communities living near the coast are particularly at risk.

4. **Increased Flooding:** Along with rising sea levels, more intense rainfall events contribute to flooding. This can disrupt daily life, damage property, and strain emergency response systems.
5. **Environmental Injustice:** Despite contributing minimally to global emissions, Puerto Rico experiences disproportionate impacts from climate change. This highlights issues of environmental injustice, where vulnerable populations bear the brunt of climate-related challenges.
6. **Economic and Social Impacts:** The economic and social fabric of Puerto Rico is also affected. Climate-related disasters can lead to job losses, displacement, and increased poverty, exacerbating existing social inequalities.

Addressing these challenges requires comprehensive climate policies, resilient infrastructure, and community engagement to build a sustainable and equitable future for Puerto Rico.

Of the climate challenges facing Puerto Rick, sea level rise information needs to be available to coastal communities at risk. Based on Intergovernmental Panel on Climate Change (IPCC) and National Oceanographic and Atmospheric Administration (NOAA) projections residents need to know if they are at risk of sea level rise at 1-, 3-, 7-, and 10-feet. Consolidating databases and data from various federal agencies and local sources is necessary. AI technology can help project through algorithms to assess risks, giving risk managers the ability to monitor and predict volatility in losses to structure and content. The algorithm can be applied to residential, commercial, and government structures to assign responsibilities more effectively, plan for risk reduction, value ecosystem services, and cover losses.

Project Goals

GENERAL APPROACH:

Step 1: Conduct a thorough assessment of data sources from United States Government agencies (e.g., United States Census Bureau Community Resilience Estimates) to complement Puerto Rico's locally available information. This includes evaluating the quality and relevance of each data source to ensure it aligns with the project's objectives. The end goal is to produce an integrated hazard assessment visualization tool using Power BI that can provide accurate and comprehensive insights.

Step 2: Create a data library that compiles all relevant data sets related to Puerto Rico's risks and hazards. This will serve as the foundation for further analysis. The library will be designed to be publicly accessible, fostering transparency and enabling widespread use.

Step 3: Use artificial intelligence (AI), large language models (LLM), bots, and any other appropriate tools to centralize existing data from different platforms that can provide accurate results from user queries on hazards, at-risk structures and urban services, and social determinants of climate change vulnerability, into a single tool that can make data accessible and easy to understand for end users.

Step 4: Design and implement algorithms that leverage AI to perform real-time asset risk analysis. These algorithms will enhance the tool's ability to anticipate and project physical risks, structural volatility, and economic impacts due to sea level rise, floods, and other hazards. The algorithms will also support cost-benefit analyses that incorporate the value of ecosystem services and nature-based solutions.

FUTURE WORK:

Step 5: Integrate the algorithm to refresh the data as it becomes available to update models.

Data Sources and Datasets

The following is a non-comprehensive list of datasets applicable to the project.

1. **Esri (Environmental Systems Research Institute)** is a leading company in the field of geographic information system (GIS) software, location intelligence, and spatial analytics with their main product being **ArcGIS** (https://www.esri.com/en-us/arcgis/products/index). GMU maintains several computer labs in which ArcGIS is installed. Students can also access

ArcGIS Online. ArcGIS Online is a cloud-based mapping and analysis solution. Use it to make maps, analyze data, and to share and collaborate. ArcGIS Online contains a subset of the tools available within the desktop clients.

2. **OpenStreetMap (OSM)** (https://www.openstreetmap.org/) is a collaborative project to create a free, editable map of the world. Launched in 2004, OSM is built and maintained by a community of volunteers who contribute and update map data. The project was started in the UK due to the lack of freely available map data and has since grown into a global initiative.
3. **Arkly** (https://www.arkly.com/pr) is a platform designed to help homeowners understand their flood risk and explore potential solutions. It provides tools to look up flood zones, assess flood risks, and get flood insurance quotes. Arkly also offers consultations with flood mitigation experts to help users protect their properties.
4. The **NASA Sea Level Change Portal** (https://sealevel.nasa.gov/) is a comprehensive resource for understanding and tracking sea level changes globally.
5. The **U.S. Census Community Resilience Estimates (CRE)** (https://www.census.gov/programs-surveys/community-resilience-estimates.html) provide a metric for assessing how socially vulnerable every neighborhood in the United States is to the impacts of disasters, such as wildfires, flooding, hurricanes, and pandemics like COVID-19.
6. The **Whole Community Resilience Planning (WCRP) Program** (https://recuperacion.pr.gov/wcrp/tools-portal.html) is an initiative designed to support communities in Puerto Rico in developing comprehensive resilience plans.
7. The **NOAA Sea Level Rise Viewer** (https://coast.noaa.gov/slr/) is a powerful tool designed to help visualize and understand the impacts of sea level rise and coastal flooding.
8. The **Caribbean Center for Rising Seas (CCRS)** internal data sets and contributing institutions data set.

Partner Intellectual Property

None.

References

None.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Creative Commons Zero (CC0)** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☐ Working Prototype
☐ Other (please specify below)
⇒

SEMBRANDO SENTIDO

Project POCs: Issel Masses, Executive Director, Sembrando Sentido, imasses@sembrandosentido.org
Dr. Eva Villalon, Research Director, Sembrando Sentido, evillalon@sembrandosentido.org

https://www.sembrandosentido.org/en/home

Sembrando Sentido is a non-profit organization dedicated to promoting transparency, accountability, and efficiency within the central government of Puerto Rico. Their mission is to create a more inclusive and responsible government that serves all Puerto Ricans effectively. By focusing on transparency in public contracting and policy analysis, Sembrando Sentido aims to foster a robust and open public procurement system.

One of their key initiatives is the **Contracts in Law program**, which seeks to enhance transparency in government contracts. This program has led to significant achievements, such as the development of a registry of convicted corruption offenders by the Department of Justice, which had not been enforced until Sembrando Sentido's advocacy. Additionally, they have facilitated over 30 workshops on public contracting, increasing the centralization and transparency of government contracting data by 91.7%.

Sembrando Sentido also emphasizes the importance of citizen participation and education. They support the implementation of transparency mechanisms in infrastructure projects and promote an open, corruption-free government through research, advocacy, and education. Their vision is to cultivate an informed and engaged civil society that collaborates with the government to maximize public resources for the long-term well-being of Puerto Rico.

Project Title	DATA FOR EQUITABLE RESULTS: TRACKING FEDERAL FUNDS AND IMPACT IN THE UNITED STATES
Organization	*Please provide the name of the partnering organization for this project:* Sembrando Sentido (San Juan, Puerto Rico)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Issel Masses, Executive Director, Sembrando Sentido, imasses@sembrandosentido.org Dr. Eva Villalon, Research Director, Sembrando Sentido, evillalon@sembrandosentido.org
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☒ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Amidst the climate crisis and the pandemic, federal spending has surged, totaling around $4.6 trillion for COVID-19 recovery and approximately $1.5 trillion for clean energy, climate action, and environmentally responsive infrastructure projects. Despite efforts to enhance transparency and monitoring, fragmented disclosure of information hinders access to detailed spending data, risking fund mismanagement and hampering effective oversight and participation. Ensuring these funds meet community needs requires empowering governments and communities to track funding flows and assess impact.

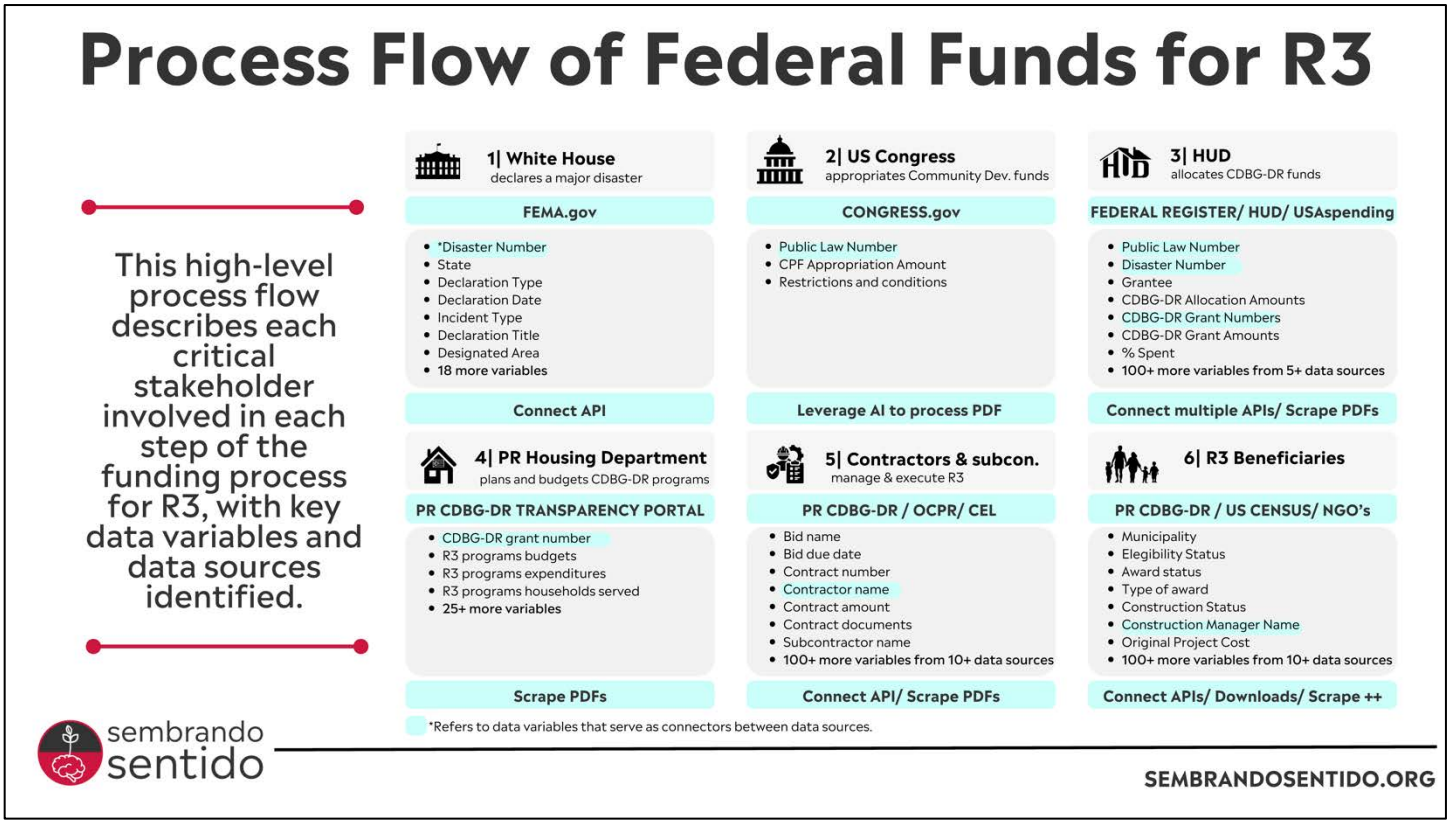
To address these challenges, Sembrando Sentido and partners are launching a federal-wide Initiative aimed at addressing data limitations at the federal and local levels to create a digital transparency tool that improves access to, and understanding of federal spending, from its initial appropriation to its impact on communities. Our focus starts in Puerto Rico, where over 40% of overall funds come from federal allocations, providing critical resources to help address the high climate vulnerabilities it faces. The Initiative, however, aims to address data limitations and develop solutions that serve and are replicable in other jurisdictions in the U.S.

The process flow of federal funds, including those for R3 (Resilient, Reliable, and Robust), typically follows a structured path within the broader federal budget process. Here's a simplified overview:

1. **President's Budget Request:** Early each year, the President submits a budget request to Congress, outlining spending priorities for federal agencies and programs, including R3 initiatives.
2. **Budget Resolution:** Congress develops a budget resolution, which serves as a blueprint for federal spending and revenue. This resolution guides the appropriations process but does not have the force of law.
3. **Appropriations Bills:** Congress passes appropriations bills to allocate specific funding to federal agencies and programs. These bills must be approved by both the House and Senate and signed by the President.
4. **Apportionment:** Once appropriations are made, the Office of Management and Budget (OMB) apportions the funds to various federal agencies. This step ensures that funds are distributed effectively and economically.

5. **Obligation and Expenditure:** Federal agencies then obligate the funds for specific projects and programs, including R3 initiatives. These obligations are followed by actual expenditures as the funds are spent on approved activities.
6. **Monitoring and Reporting:** Throughout the fiscal year, agencies monitor the use of funds and report back to OMB and Congress to ensure compliance with budgetary guidelines and objectives.

This process ensures that federal funds are allocated, distributed, and used in a controlled and transparent manner. The figure below describes each critical stakeholder involved in each step of the funding process for R3, with key data variables and data sources identified.



Project Goals

The first phase of this Initiative entailed (1) researching the needs and interests of local communities regarding information on federal spending and impact, and (2) compiling, evaluating, and identifying opportunities to connect data sources in order to develop a data model and architecture for the digital transparency tool (leading to the data source catalog). We now have identified a pilot project, aimed at integrating data sources that can help better track public investments in housing, particularly the "Repair, Reconstruction and Reallocation" programs funded by the U.S. Department of Housing, and critical in the post-disaster reconstruction efforts across many different jurisdictions, including Puerto Rico, Florida, Texas, California, Louisiana, and others.

We have mapped the data integration and tracking efforts for Puerto Rico and are exploring additional locations for this work. Students at George Mason University (GMU) could support the team in the development of solutions for data extraction, data integration and visualization tools to show "housing needs", investments and change over time.

1. **Understanding Project Scope and Objectives:**

    a. **Review Documentation:** Students will thoroughly review all project documentation to understand the objectives, requirements, and expected outcomes. This includes getting familiar with all relevant data sources, including federal,

state, and local databases, especially those related to the "Repair, Reconstruction, and Reallocation" programs relevant to Puerto Rico and the comparative case of selection (TBD, likely Louisiana).

b. **Meet with Project Leads:** Students will attend initial meetings with project leads to gain clarity on specific goals and any nuances of the project.

2. **Data Extraction:**

   a. **Develop Extraction Scripts:** Students will choose/be assigned to develop several code scripts that automate the extraction of data from various data sources. Depending on the data source different scraping approaches may be necessary, headless crawling, API usage and PDF scanning are all possibilities.

   b. **Data Cleaning:** Students will also perform initial data cleaning and transformations to remove any inconsistencies or errors in the raw data and prepare it for the next steps.

3. **Data Modeling:**

   a. **Design Data Models:** Students will help design data models that will support the integration of various datasets. This involves defining tables, relationships, and indexes.

   b. **Implement Data Models:** They may also implement these models in a database management system (e.g., MySQL).

4. **Data Integration and Normalization:**

   a. **Schema Mapping:** Using previous analysis of the datasets, students could be assigned to help create a schema in a relational database, support normalization of key data to ensure it will seamlessly integrate into the database, and support its upload into the db.

5. **Develop Visualization Tools:**

   a. **Identify Visualization Requirements:** Using pre-determined metrics and insights that need to be visualized (e.g., housing needs, fund allocation, project progress), students will assess and help Sembrando's team select appropriate data visualization tools (e.g., Tableau, D3.js, web app).

   b. **Create Dashboards:** Students will be encouraged to develop an interactive dashboard and visualization that presents the data in an accessible and understandable manner.

Data Sources and Datasets

Specific datasets within the main sources listed below have been identified.

1. FEMA.gov: Disaster Declaration Summaries: https://www.fema.gov/openfema-data-page/disaster-declarations-summaries-v2
2. US Congress appropriates Community Development Funds: https://www.congress.gov/advanced-search/legislation
3. HUD allocation of CDBG-DR Funds: https://www.federalregister.gov/documents/search#advanced
4. HUD.gov: CDBG-DR Reports https://www.hud.gov/program_offices/comm_planning/cdbg-dr/reports
5. USAspending.gov: Award Search (Transaction Level): https://www.usaspending.gov/search
6. Puerto Rico Housing Department plans and budgets CDBG-DR programs:
   a. Quarterly reports: https://recuperacion.pr.gov/en/transparency-portal/finance/reports/
   b. Contractors & subcontractors manage and execute R3 programs: https://recuperacion.pr.gov/en/procurement-and-nofa/procurement/
   c. R3 Program Managers: https://recuperacion.pr.gov/recursos/contratos/contratistas/gerentes-de-programas/

   d. R3 Construction Managers: https://recuperacion.pr.gov/recursos/contratos/contratistas/gerentes-de-construccion/
   e. CDBG Contracts: https://recuperacion.pr.gov/en/resources/contracts/contracts-cdbg/
7. Impact: R3 Beneficiaries: https://recuperacion.pr.gov/en/transparency-portal/transparency-reports/housing-reports/r3-dashboard/
8. US Census Profiles: https://data.census.gov/profile?g=040XX00US72
9. US Census Bureau: Community Resilience Estimates: https://www.census.gov/programs-surveys/community-resilience-estimates/data/cre-pr.html

Partner Intellectual Property

After consulting with our Lead Developer and considering that our focus will likely be on extracting (and processing) key data from PDF documents, we would like to request that if we share any scrapers or data schemas previously developed by **Sembrando Sentido** for its **Contratos En Ley** (https://contratosenley.org/en) platform as examples for students to use or draw from in developing their own tools for this project, the following conditions be respected:

1. The intellectual property (IP) of these shared resources should be acknowledged as belonging to Sembrando Sentido.
2. Any key/sensitive information regarding these resources would also be noted in advance, to be kept and shared internally (i.e., within the team, advisor, and leadership).
3. An example of this may include .pdf scrapers that have been developed by Sembrando Sentido's team to extract data from pdf documents on corporations.

References

1. **Sembrando Puerto Rico Federal Funding Data Source Catalog** (https://docs.google.com/spreadsheets/d/1MlF7223oNq_d-n-lnMC0KbkrwdTtbiKwzxnp0pwxSGg/edit)
2. **Schema.org** (https://schema.org) is a collaborative initiative founded by Google, Microsoft, Yahoo, and Yandex. Its primary goal is to create, maintain, and promote schemas for structured data on the internet. These schemas help webmasters and developers embed structured data on their web pages, making it easier for search engines and other applications to understand and use the information.
3. Guide to the Federal Budget Process | Bloomberg Government
4. Apportionment of Appropriated Funds | www.dau.edu
5. How DoD Gets Money: A Primer on the US Federal Budget Process
6. Introduction to the Federal Budget Process - CRS Reports

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Creative Commons Zero (CC0)** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
⇒

UNITED STATES POSTAL SERVICE — CORPORATE INFORMATION SECURITY OFFICE (CISO)

Project POC: Michael Billingsley, Director, Cybersecurity Engineering, michael.a.billingsley@usps.gov

https://postalpro.usps.com/ciso/

The **Corporate Information Security Office (CISO)** of the United States Postal Service (USPS) plays a crucial role in safeguarding the organization's digital infrastructure. Established to address the increasing cyber threats that impact USPS's customers, partners, and employees, the CISO is responsible for protecting the USPS network, monitoring threats, and responding to incidents. This office ensures that the USPS can continue to operate securely and efficiently, maintaining the trust and safety of its stakeholders.

The CISO's responsibilities include a range of activities aimed at enhancing cybersecurity. These activities involve 24/7 monitoring through the CyberSecurity Operations Center, which proactively identifies and responds to incidents and threats. The office also leads incident response efforts, ensuring that any cyber events are managed swiftly and effectively. Additionally, the CISO emphasizes the importance of cybersecurity awareness and training, having educated hundreds of thousands of USPS employees on cybersecurity essentials.

Overall, the CISO's efforts are integral to the USPS's mission of protecting its critical infrastructure and information systems against cyber threats. By maintaining a robust cybersecurity posture, the CISO helps the USPS achieve its business goals while safeguarding the data and privacy of its employees and customers.

Project Title	USPS Counterfeit Label Detection Challenge
Organization	*Please provide the name of the partnering organization for this project:* United States Postal Service — Corporate Information Security Office (CISO)
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Michael Billingsley, Director, Cybersecurity Engineering, michael.a.billingsley@usps.gov
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☐ Systems Engineering ☒ Data Engineering ☒ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☒ Natural Language Processing (NLP) ☒ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☐ New project ☒ Follow-on project from a previous semester (Semester Year): Summer 2024
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☒ Yes - U.S. citizenship required ☐ No - U.S. citizenship not required

Problem Description

The USPS provides commercial shippers with unique Mailer Identification Numbers (MID) to manifest and pay for their postage. Fraudulent shippers use a combination of hijacked MIDs, where a fraudulent shipper steals a MID assigned to a USPS customer, and unregistered MIDs, a MID that has not been setup to enter packages into the USPS package stream. USPS package fraud detection is currently a retroactive process to identify fraudulent packages. The time lag creates an opportunity for fraudulent MIDs to go undetected while also providing time for fraudulent shippers to create new fraudulent MIDs before USPS can address.

USPS would like to present the participants with a challenge to develop analytics and AI/ML approaches, utilizing a sample of USPS shipment data, that allows USPS to shorten the detection time-lag and deter fraudulent package shipments. Work from a prior semester will be provided as a starting point for this cohort. The team will evaluate the prior work, consider the recommended next steps for employing supervised methods, and determine a preferred path forward to achieve the project goals.

Project Goals

The project will involve analyzing shipping data associated with a sample of USPS labels created over a short period of time (e.g., one week). The project will focus on understanding mailer patterns between fraudulent and non-fraudulent labels based on characteristics of the mailer and shipping label. The team will identify several mailers from the data that appear to have hijacked volume (some legitimate, paid labels and some unpaid labels generated by fraudsters) and will build machine learning models on a per mailer basis that use label characteristics to classify counterfeit labels. Emphasis should be placed on ease of use for executive personnel (e.g., developing a natural language processing-based interface to query model data).

The team will evaluate the effectiveness of these models using model precision and recall and may use additional metrics if applicable. Teams will provide insights into model features that are influential in predicting fraudulent behavior. The project will also involve developing recommendations on how to improve data collection and model performance. The final project report should encapsulate the following:

1. **Methodology and Data Sources Used:** A detailed description of the approach used to identify counterfeit labels, including the validation methodology and data sources used.

2. **Effectiveness:** Evaluate the effectiveness of the model with appropriate ML performance metrics.
3. **Features:** Provide insights into model features that are influential in predicting fraudulent behavior.
4. **Conclusions and Recommendations:** A summary of conclusions and recommendations for next steps to improve data collection and model performance.

Data Sources and Datasets

Access to relevant data sources will be provided by USPS and sanitized of any Personally Identifiable Information (PII) or sensitive data. Only a sample of data will be provided due to large data volumes. The data will include information on the mailers, label types / details, manifest details, shipping partner details, and induction scan events.

Partner Intellectual Property

The USPS-provided sample data for packages will be part of the project solution. Identifying this partner intellectual property at the beginning of the project will ensure it is not part of any graduate student intellectual property that is part of the project solution.

References

1. Participants are encouraged to familiarize themselves with the USPS operations and business-to-business customers utilizing USPS's eVS (Electronic Verification System) and PC Postage services.
    a. **Publication 205 Electronic Verification System (eVS®) Business and Technical Guide** (https://postalpro.usps.com/pub205)
    b. **Publication 199: Intelligent Mail Package Barcode (IMpb) Implementation Guide for Confirmation Services and Electronic Payment Systems** (https://postalpro.usps.com/pub199)
    c. **Postal Terms** (https://about.usps.com/publications/pub32/pub32_terms.htm)
2. Choy, R., Chughtai, K., Koziol, M., Vega, B., Walden, H., "USPS Counterfeit Label Detection Challenge," presented at the Summer 2024 DAEN 690 Capstone Presentation Showcase, Fairfax, Virginia, August 2, 2024.

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☒ Machine Learning Model
☒ Working Prototype
☐ Other (please specify below)
⇒

WEB3NITY FOUNDATION — MYTIKI.COM

Project POCs: Sean Koh, WEB3NITY Foundation, skoh@koherentinc.com
Mike Audi, CEO, MyTiki.com, mike@mytiki.com

https://www.proofofreception.org/
https://www.mytiki.com/

The **WEB3NITI (pronounced "web trinity") Foundation** is a non-profit humanitarian organization with the mission to free, empower, and protect the data, time, and skills of 8 billion humans around the world in order to end the online trafficking of our humanity once and for all. WEB3NITI's core principles include: 3N "Necessary Natural Nourishment," 3I "Intrinsic Innovative Intelligence," 3T "Transformational Truthful Transparency," 3I "International Inclusive Interoperability." WEB3NITI educates and empowers individuals with "Proof of Reception" (PoR) a proprietary edge validation protocol that enables a better internet based on our Human Intelligence (HI) with unprecedented security, efficiency, and sustainability for a variety of use cases globally.

MyTiki provides infrastructure for the exchange of data to unlock humanity's most valuable assets. With a mission to improve the accessibility, safety, and compensation for shared data; a gray market directly and indirectly impacting all of us. They automate the technical, legal, and sales complexities of turning data into revenue generating assets for businesses and their users.

Project Title	PROOF OF RECEPTION (POR) UNIVERSAL PROFILE IDENTIFICATION SYSTEM
Organization	*Please provide the name of the partnering organization for this project:* WEB3NITY Foundation and MyTiki
Project POC(s)	*Please provide the name, title, email, and phone contact information of all organization individuals supporting this project:* Sean Koh, WEB3NITY Foundation, skoh@koherentinc.com Mike Audi, CEO, MyTiki.com, mike@mytiki.com
Knowledge Domain(s)	*Please select all knowledge domains which apply to this project:* ☒ Systems Engineering ☒ Data Engineering ☐ Data Mining ☒ Data Analytics ☐ Data Modeling/Simulation ☒ Data Visualization ☐ Computer Vision ☐ Natural Language Processing (NLP) ☐ AI/ML ☐ Generative AI ☐ DevSecOps ☐ MLOps
Specialized Skills	*Please indicate any specialized skills required to work on this project:*
Max Number of Project Teams	*Please indicate the maximum number of project teams which can work on this project during the semester:* ☒ 1 ☐ 2 ☐ 3 ☐ 4
New/Follow-on Project	*Please indicate whether this is a new project or a follow-on project from a previous semester:* ☒ New project ☐ Follow-on project from a previous semester (Semester Year): Fall/Spring/Summer 202x
U.S. Citizenship Requirement	*Please indicate whether U.S. citizenship is a requirement to work on this project:* ☐ Yes - U.S. citizenship required ☒ No - U.S. citizenship not required

Problem Description

Currently, fragmented data of our personal humanity are extracted, exploited, and exchanged generating billions of dollars for others without our knowledge; furthermore, the value of online data becomes more enriched when it is consolidated around unique identifiers and updated in real-time which can only be done by the actual Human Intelligence (HI) of the individuals of whom fragmented data describes. In order to introduce a more humanitarian and sustainable data paradigm while freeing our personal humanity online from third party trafficking, Proof of Reception's (PoR) unique edge validation enables an unprecedented solution to benefit everyone.

Unifying fragmented data across large data sets under individual human identities and enabling the owners of the respective HI to claim, control, and consent themselves as personal data is a fundamental human right.

Sean Koh, also known by his stage name "ESKOH" (which stands for "Every Situation Kan Offer Hope"), is a multifaceted entrepreneur and expert in cross-border transactions across finance, technology, infrastructure, media, and entertainment. He is the founding partner of Koherent Incorporated, a family office involved in various industries, including private equity, project financing, and blockchain technologies.

In addition to his business ventures, Sean Koh has made significant contributions to the music industry as a producer, writer, and performer. He is also the inventor of the Proof of Reception (PoR) technology, which ensures secure and verifiable data transactions. His diverse expertise and innovative approaches have positioned him as a prominent figure in both the business and technology sectors. Proof of Reception (PoR) is a blockchain-based solution that ensures the secure and verifiable receipt of data. This technology creates an immutable record of data transactions, enhancing trust and accountability. By verifying and logging the receipt of data, PoR helps prevent fraud and ensures compliance with data privacy regulations. This innovative approach is designed to build a more reliable and transparent ecosystem for data exchanges.

MyTiki.com is a pioneering platform that empowers users to take control of their data and monetize it effectively. The company offers a suite of tools designed to aggregate, clean, and deliver data in a compliant manner. By creating a branded, self-service

storefront, MyTiki.com allows businesses to source, test, and integrate unique datasets seamlessly. This approach not only simplifies the process of data management but also ensures that users can turn their data into valuable revenue streams. The platform's Receipt OCR feature, for instance, extracts SKU-level data from physical and email receipts, enabling users to monetize aggregate, de-identified data while covering costs.

MyTiki.com utilizes Proof of Reception (PoR) technology to ensure that data transactions are securely and transparently recorded. PoR is a mechanism that verifies and logs the receipt of data by the intended recipient, providing an immutable record of the transaction. This technology is crucial for maintaining trust and accountability in data exchanges, as it ensures that all parties involved can verify that the data has been received as intended.

By implementing PoR, MyTiki.com enhances the security and reliability of its data monetization platform. Users can confidently share their data, knowing that there is a verifiable record of its receipt. This not only helps in preventing data fraud but also ensures compliance with data privacy regulations. Overall, PoR technology plays a key role in building a trustworthy ecosystem for data transactions on MyTiki.com.

ABOUT WEB3NITI FOUNDATION

The WEB3NITI (pronounced "web trinity") Foundation (https://www.proofofreception.org/) is a non-profit humanitarian organization with the mission to free, empower, and protect the data, time, and skills of 8 billion humans around the world in order to end the online trafficking of our humanity once and for all. WEB3NITI's core principles include: 3N "Necessary Natural Nourishment," 3I "Intrinsic Innovative Intelligence," 3T "Transformational Truthful Transparency," 3I "International Inclusive Interoperability." WEB3NITI educates and empowers individuals with "Proof of Reception" (PoR) a proprietary edge validation protocol that enables a better internet based on our Human Intelligence (HI) with unprecedented security, efficiency, and sustainability for a variety of use cases globally.

ABOUT MYTIKI

MyTiki (https://mytiki.com/) provides infrastructure for the exchange of data to unlock humanity's most valuable assets. With a mission to improve the accessibility, safety, and compensation for shared data; a gray market directly and indirectly impacting all of us. We automate the technical, legal, and sales complexities of turning data into revenue generating assets for businesses and their users.

Project Goals

The overall project goal will be to build a Proof of Reception (PoR) universal profile identification system.

A universal identification system is a framework designed to provide a unique identifier to individuals or entities that can be used across various platforms and services. This system ensures that each individual or entity has a single, consistent identity that can be recognized and verified universally.

Key Features:

1. **Unique Identifier:** Each individual or entity is assigned a unique identifier that is consistent across different systems and services.
2. **Interoperability:** The system is designed to work across various platforms, ensuring that the identifier can be used in multiple contexts without confusion.
3. **Security and Privacy:** Robust measures are in place to protect the security and privacy of the individuals or entities using the system.

Uses:

1. **Government Services:** Universal identification systems are often used by governments to streamline access to public services. For example, national ID systems can be used for voting, taxation, social security, and healthcare services.

2. **Digital Marketing:** In the digital marketing ecosystem, a universal ID can help track user interactions across different platforms, providing a more cohesive understanding of user behavior.
3. **E-commerce:** Universal IDs can simplify the user experience by allowing seamless login and transaction processes across different e-commerce platforms.
4. **Healthcare:** In healthcare, a universal identification system can ensure that patient records are accurately matched and accessible across different healthcare providers, improving the quality of care.

Overall, universal identification systems aim to simplify and secure the process of identifying individuals or entities across various domains, enhancing efficiency and interoperability.

During development, the team will need to consider the use of an AI-based entity matcher which is a sophisticated tool designed to identify and link entities (such as names, organizations, locations, etc.) across different datasets or within a single dataset. This process involves recognizing and matching entities that may be represented in various forms or formats, ensuring that they are correctly identified as the same entity. For example, an AI-based entity matcher can recognize that "IBM," "International Business Machines," and "I.B.M." all refer to the same company.

The primary use of an AI-based entity matcher is to enhance data integration and consistency. In many applications, data comes from multiple sources, each with its own way of representing entities. By matching these entities accurately, organizations can consolidate data, eliminate duplicates, and ensure that their datasets are comprehensive and accurate. This is particularly useful in fields like customer relationship management (CRM), where understanding the full scope of interactions with a single customer across various touchpoints is crucial.

Additionally, AI-based entity matchers are employed in natural language processing (NLP) tasks, such as information retrieval, text mining, and knowledge graph construction. They help in extracting meaningful information from unstructured data by linking mentions of entities to their corresponding records in a database. This capability is essential for applications like search engines, recommendation systems, and automated customer support, where understanding the context and relationships between entities can significantly improve the quality of service and user experience.

The project team will focus on achieving the following project goals:

1. Evaluate and select a scalable AI-based entity matcher such as **AWS Entity Resolution** (https://aws.amazon.com/entity-resolution/) or **Zingg** (https://github.com/zinggai/zingg).
2. Understand the Proof of Reception (PoR) protocol and tools from the published application document (https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/20230169510).
3. Design and build the Proof of Reception (PoR) matching system connected to MyTiki's data system containing over 500M hashed emails and 5M payment profiles to bootstrap the initial universal identification system.
4. Using Proof of Reception (PoR) collect and deliver tagged user profile data to the matching system updating the universal profiles.
5. Perform semi-supervised learning to fine-tune the matching system.

Data Sources and Datasets

MyTiki's data system containing over 500M hashed emails and 5M payment profiles.

Partner Intellectual Property

1. Proof of Reception (PoR) Patent Filing – United State Patent and Trademark Office (US PTO) (https://patentcenter.uspto.gov/applications/18071163).
2. Proof of Reception (PoR) Patent Filing –World Intellectual Property Organization (WIPO) (https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2023097101&_cid=P10-LZH9T6-33453-1).
3. MyTiki's data system containing over 500M hashed emails and 5M payment profiles.

References

1. Sean “ESKOH” Koh (https://linkin.bio/eskoh/)
2. Proof of Reception (PoR) patent application (https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/20230169510)
3. The **2024.06.25 People Centered Internet (PCI) Community Call - Sean Koh** YouTube video (https://youtu.be/e8MCwXKqkEk?si=A_hjDUnDm8hI_AUe) is a mandatory watch by the project team as Sean goes into great detail explaining Proof of Reception (PoR).
4. Proof of Reception (PoR) DEVPOST (https://devpost.com/software/proof-of-reception-por).

Project Development Environment

The project team will be required to use the College of Engineering Computing (CEC) AWS team-based environment.

At the end of the first week of the course the project team will provide to the course instructor the list of AWS services and their *minimum* specifications (e.g., EC2 instance selected, S3 bucket size, etc.) necessary to successfully complete the project. The course instructor will review the project team request to ensure it meets CEC ITS guidelines for AWS environment provisioning. The course instructor will then be responsible for providing that information to CEC ITS staff so that they can, in turn, provision the AWS environment for the project team.

Project Open Source Licensing

All code and project deliverables generated by the project team will be published under the **Apache License 2.0** permissive open-source software license.

Project Deliverables

☒ Capstone Showcase Presentation & PowerPoint Slides
☒ Final Project Report
☒ Repository for Data and Code Artifacts
☐ Machine Learning Model
☐ Working Prototype
☐ Other (please specify below)
⇒

For more information about becoming an MS Data Analytics Engineering Program capstone project partner

PLEASE CONTACT

Bernard Schmidt
Instructor and Assistant Director, MS Data Analytics Engineering Program

George Mason University | College of Engineering and Computing | Volgenau School of Engineering
Research Hall
10401 York River Road
Suite 359, MS 6B1
Fairfax, VA 22030
DAEN Program URL: https://analyticsengineering.gmu.edu/
Faculty Bio: https://cec.gmu.edu/node/1421
Email: bschmid5@gmu.edu
Office Location: Fairfax Campus | Research Hall | Room 368
Office Phone: (703) 993-6548

College of Engineering and Computing
DATA ANALYTICS ENGINEERING
George Mason University