# Applied Data Science Project Milestone Portfolio

**School**: Syracuse University
**Program of study**: M.S. in Applied Data Science
**Student**: Patrick Walsh
**Portfolio advisor**: Dr. John Stinnett
**Date submitted**: May 26, 2025

## Table of contents

# Description of Applied Data Science program

The M.S. in Applied Data Science at Syracuse University teaches the theoretical foundations of data science in an interactive learning environment where students learn to apply their knowledge to real-world problems in business, research, and other dynamic sectors. The program combines technical knowledge with an ability to translate data into actionable insights and drive decision-making. The program's unique applied approach emphasizes the use of data science in enterprise operations and processes. It is designed to allow students to gain expertise in the areas of data capture, management, analysis, and communication. It helps students build their resumes through hands-on projects using real-world data sets.

The program is 34 credits and can be completed in as little as one year. The curriculum combines a primary core, a secondary core (your data science track), and electives to give students a strong data science foundation with a focus of their choosing.

# Program learning goals:

1. Collect, store, and access data by identifying and leveraging applicable technologies

2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle

3. Apply visualization and predictive models to help generate actionable insight

4. Use programming languages such as R and Python to support the generation of actionable insight

5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)

6. Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

This portfolio will refer back to these learning goals in the Project highlights section to demonstrate how I applied the learning goals throughout my educational journey.

# Professional background

I am a Data Scientist with over 16 years of experience in data analysis. I have a background in linguistics with an emphasis in Natural Language Processing (NLP). I have 5 years of experience in Computer Vision (CV), software development, and cybersecurity. I am also a military veteran with 8 years of honorable service. I graduated Magna Cum Laude from University of Maryland Global Campus (UMGC) with a B.S. in Software Development & Security and, received an A.A. in Arabic Language Arts from the Defense Language Institute (DLI). My most recent academic achievement is the completion of the M.S. in Applied Data Science at Syracuse University.

# Resume/CV

**Lead Data Scientist**

***ECS Federal***

**February 2024 - Present**

As a Lead Data Scientist working as a contractor with the Department of Homeland Security (DHS)'s Data Services Branch (DSB), I specialize in document classification and text analytics using advanced machine learning techniques. My work focuses on creating and training NLP models, including Naive Bayes, Random Forest, and Support Vector Machines (SVM), implemented through SciKit-Learn. These models are pivotal for classifying multi-page PDF documents, supporting critical DHS operations through accurate multilabel classification.

I also extract key entities from text using Named Entity Recognition (NER) and Regex, enhancing the precision and relevance of data analysis. Beyond model development, I deploy these solutions on AWS platforms like SageMaker and EC2, ensuring scalability and performance. Additionally, I perform UI development with Streamlit, enabling seamless document upload and automated analysis through a user-friendly interface, making actionable insights readily available across DHS.

**Senior Machine Learning Engineer**

***The Cigna Group***

**October 2023 - January 2024**

As a Senior Machine Learning Engineer at Cigna, I specialized in text-to-SQL generation using Generative AI and OpenAI APIs. My responsibilities included formatting Teradata SQL queries,

making API calls to Large-Language Models (LLM), fine-tuning LLM responses, and modifying SQL responses through procedural calls. Additionally, I served as an R&D AI engineer with the Research and Development team, contributing to innovative projects and exploring advancements in AI technologies. This role allowed me to develop a comprehensive skill set in text generation, API integration, and SQL manipulation, further enhancing my expertise in the field.

**Data Scientist**

***ECS Federal***

**January - October 2023**

As a contractor with the Multi-Channel Technologies division at the Department of Veterans Affairs (VA), I played a pivotal role in creating Generative AI applications. Notably, I developed a zero-shot model for automatically tagging case notes based on textual data. My involvement extended to both front-end and back-end development of chatbots, utilizing Large-Language Models (LLM), semantic search, and vector databases. I employed FAISS, Chroma DB, and Pinecone for efficient vector database storage, utilizing them to process, chunk, and store documents as vectorized indexes, thereby enhancing the overall functionality and efficiency of the chatbots. I designed and implemented NLP models for analyzing and routing veteran queries, employing various modern textual content classification techniques. The incorporation of Retrieval-Augmented Generation (RAG) significantly improved the performance of LLMs by leveraging indexed documents stored in a vector database, resulting in heightened user satisfaction and engagement. In addition to serving customers and business stakeholders, I utilized data to provide timely solutions to business problems. I extensively utilized Python for extracting, analyzing, and processing textual data, contributing to the development of deep learning multi-label classification models.

Furthermore, I automated business processes using Power Apps, Power Automate, and Power BI, fostering increased efficiency and proactively identifying and preventing data discrepancies.

**Data Scientist**

***SYSCOM, Inc.***

**July 2017 - December 2022**

As a Data Scientist at SYSCOM, I wrote Python code to conduct data wrangling and configure Deep Learning models, fine-tuning hyperparameters, and executing training jobs within both AWS and Azure environments. One significant achievement was the development and refinement of a computer vision (CV) model adept at accurately identifying and classifying

nutrient deficiencies in plants. This breakthrough contributed to improved crop yields and reduced costs for farmers. I took charge of deploying and hosting this model on AWS, ensuring scalability and reliability to cater to clients worldwide.

Additionally, I played a pivotal role in the design and development of an innovative computer vision model that successfully classifies biofilm pathogens in microscope images, leading to a patented technology. My contributions extended to automating the data preprocessing pipeline for CV model development, significantly enhancing efficiency from a 2-hour process to just 30 seconds.

In the realm of Natural Language Processing (NLP), I created models for Topic Modeling, Keyword Extraction, Sentiment Analysis, and Named Entity Recognition (NER). My expertise also encompassed the development of CV models for image classification and object detection.

Engaging directly with customers, I interfaced to understand their business problems, build relationships, and provide effective solutions. In the realm of data management, I queried, designed, and updated relational databases using SQL and Python. Additionally, I developed graphing and visualization programs utilizing Python, R, Tableau, and Power BI to analyze medium to large datasets. These efforts underscore my multifaceted contributions in the dynamic field of Data Science at SYSCOM.

**Data Analyst & Arabic Linguist**

***United States Army***

**February 2009 - March 2017**

In my role as a Data Analyst and Arabic Linguist in the United States Army, I demonstrated exceptional leadership by managing a team of 20 data analysts, contributing single-handedly to 25% of the product output—marking the highest volume for the entire division at that time. My responsibilities extended to translating Arabic text and audio into English for further analysis and delivering high-level classified verbal and written reports to customers. Additionally, I played a crucial role in decrypting secure digital communications to exploit organizational weaknesses.

In pursuit of operational efficiency, I developed and implemented streamlined data analysis methodologies, resulting in a notable 15% reduction in analysis time, all while maintaining the highest standards of data accuracy and quality. My leadership skills were further showcased as I led cross-functional teams in the successful completion of complex projects, ensuring on-time delivery of critical milestones, and demonstrating adept project management skills. Recognizing the importance of data accessibility, I spearheaded the development of customized data visualization dashboards. This initiative significantly improved data accessibility and empowered the organization with data-driven decision-making capabilities.

Furthermore, I actively coordinated with government agencies, providing high-quality, time-sensitive technical support. My multifaceted contributions, from leadership and project management to linguistic and analytical skills, underscore my dedication to excellence in the dynamic field of Data Analysis and Arabic Linguistics within the United States Army.

# Project highlights

Throughout my academic journey in the Applied Data Science program, I participated in a number of projects across multiple classes. These projects allowed me to apply the learning goals (outlined in the section Program learning goals) to my studies and round out my skillset as a data professional.

What follows is a brief description of some of these projects and a demonstration of how I applied the program learning goals within my academic coursework.

## IST 736 - Text Mining

This course introduces concepts and methods for knowledge discovery from a large amount of text data and the application of text mining techniques for business intelligence, digital humanities, and social behavior analysis.

In this project, I applied key learning goals through a comprehensive text mining analysis of ‘The Office’ dataset.

*Goal 1: Collect, store, and access data by identifying and leveraging applicable technologies*

I began by gathering and organizing a large collection of dialogue transcripts from The Office. I identified appropriate file structures and formats and leveraged tools in R and Python to efficiently load and manage the data, ensuring it was clean and accessible for analysis.

*Goal 3: Apply visualization and predictive models to help generate actionable insight*

I used visualization techniques to explore patterns in character dialogue, revealing distinct speaking styles and their relationships to character development. I also experimented with predictive modeling to classify which character said each line. While this task proved challenging using dialogue alone, the process offered valuable insight into the limitations of such models and suggested the potential for more complex, multimodal approaches.

*Goal 4: Use programming languages such as R and Python to support the generation of actionable insight*

Throughout the project, I used both R and Python to process the data, build models, and visualize results. These tools were essential in transforming raw textual data into structured insights and in testing different analytical approaches.

Overall, the project demonstrated how data science techniques can be applied to real-world, unstructured data, offering new ways to understand and appreciate the content we consume.

## IST 664 - Natural Language Processing

This course explores all the levels of linguistic analysis, going from tokenization, word-level semantics, part-of-speech tagging, syntax, and semantics up to the discourse level. The course also uses NLP techniques on unstructured data using Python, including information retrieval, question-answering, sentiment analysis, summarization, and dialogue systems.

In this project, I used Natural Language Processing (NLP) techniques to build a predictive model that classifies gender based on the last letter of a first name. This hands-on experience allowed me to apply the following learning goals:

*Goal 2: Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle*

This project followed the complete data science life cycle—from feature engineering and data preparation to model training, evaluation, and interpretation. By exploring the relationship between the final letter of a name and gender classification, I generated insights into how seemingly small linguistic features can hold predictive value. Although the model's accuracy (around 75%) suggests room for improvement, the process highlighted how data science can be used to uncover patterns that may have applications in fields such as marketing, user personalization, or form design, where gender inference is sometimes used.

*Goal 4: Use programming languages such as R and Python to support the generation of actionable insight*

I used Python and the NLTK library to preprocess the data, engineer features, train a Naive Bayes classifier, and evaluate the model’s performance. This project deepened my ability to use Python for machine learning and NLP tasks while reinforcing the importance of clean, well-structured code for reproducible results.

Overall, this project helped me better understand both the potential and the limitations of using simple textual features for classification tasks while strengthening my practical programming and modeling skills.

## IST 615 - Cloud Management

In this course, I learned about cloud services creation and management. The course focuses on practical experience in using, creating, and managing digital services across data centers and hybrid clouds. It also covers strategic choices for cloud digital service solutions across open data centers and software-defined networks.

*Goal 2: Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle*

The final project centered on a fictional business scenario—migrating GlobeTrekker's legacy on-premise data infrastructure to AWS cloud services. We began by assessing current limitations in scalability, performance, and cost. From there, we conducted a full financial and technical analysis using the data science life cycle: problem definition, data collection, modeling of cost savings and efficiency gains, and interpretation of outcomes. The result was a strategic recommendation grounded in data: AWS migration could reduce operational costs by up to 40% and save the company an estimated $3.6 million over ten years, with a payback period of under three years. These insights directly inform business decisions in IT investment, marketing analytics, and long-term strategic planning, making this project an excellent example of deriving actionable insight in a business context.

*Goal 5: Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)*

To effectively convey the findings and recommendations, we structured our communication to appeal to both technical and executive-level stakeholders. For technical teams, we detailed how specific AWS tools (like Redshift, Glue, and S3) would support performance improvements and advanced analytics. For business leaders and sponsors, we translated technical outcomes into meaningful business value, highlighting cost savings, customer engagement improvements, and growth potential. The use of visual aids such as cost-benefit charts, architecture diagrams, and predictive ROI models helped bridge the gap between technical complexity and business strategy. This dual-level communication ensured that all stakeholders understood the value of cloud migration from their respective perspectives.

## SCM 651 - Business Analytics

This course focused on business analytics, including advanced spreadsheets, relational databases, and SQL queries; statistical analysis in R, including multi-linear regression, interactions, tests for regression assumptions, logit, and probit; neural networks; and dashboards.

One of the projects in this course involved analyzing a dataset of financial data from Universal Bank. This project required our team to analyze real banking data in order to understand what factors influence a customer's likelihood of taking out a personal loan.

*Goal 2: Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle*

We applied the full data science life cycle—from exploratory data analysis and model building (logit, probit, and neural networks) to sensitivity testing—to identify which customer attributes had the greatest impact. For instance, we found that having a certificate of deposit account, higher income, and higher education levels significantly increased loan acquisition likelihood. These findings offer actionable insight that could directly inform the bank's marketing strategies or customer outreach programs.

*Goal 5: Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)*

Throughout the project, we used visual aids and screen captures of R output to clearly communicate the results of our models. We presented both statistical findings (e.g., p-values, coefficient directions) and conceptual interpretations (e.g., why the interaction between education and family size might matter) in a concise, structured report. This approach made the analysis understandable to both technical team members and non-technical stakeholders, like project sponsors or decision-makers at the bank.

*Goal 6: Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)*

As we built our models, we remained aware of ethical considerations, particularly regarding fairness and potential bias in our predictions. For example, while education level and income were strong predictors, we were cautious about how these variables could reflect or reinforce societal inequalities if used irresponsibly in decision-making. We also ensured that our model interpretation avoided discriminatory conclusions and that we presented the limitations of our analysis transparently, particularly in relation to how certain variables (like family size or ZIP code) could inadvertently encode sensitive or demographic information.

## IST 652 - Scripting for Data Analysis

The goal of this class is to teach students the tools and skills of scripting needed to solve problems of accessing and preparing data in a variety of formats and situations, sometimes known as data wrangling. The scripting will provide the skills needed to form data science pipelines, from acquiring and cleaning data to accessing data and transforming data for analysis or visualization.

I applied several of the learning goals during this class:

*Goal 1: Collect, store, and access data by identifying and leveraging applicable technologies*

For this project, I identified the appropriate data sources and technologies required for analysis. The dataset was obtained from Kaggle, and I leveraged web scraping techniques to gather the data from 'officequotes.net.' By selecting the dataset that contained the dialogue for all seasons of The Office, the team and I were able to ensure the data was both comprehensive and suitable for your analysis. Storing the data in a CSV file allowed for easy access and manipulation, providing a strong foundation for the analysis.

*Goal 3: Apply visualization and predictive models to help generate actionable insight*

Throughout the analysis, we utilized various methods of visualization and predictive modeling. We applied descriptive statistics to gain insights into word counts per character and per season, which helped to quantify patterns in dialogue. The use of a Naive Bayes model to classify character dialogue was an attempt to apply a predictive model to a qualitative dataset despite the model's lower accuracy. Additionally, we used N-gram analysis to extract key phrases, which helped provide more context to Michael’s dialogue and characterize the most common phrases in the script.

*Goal 4: Use programming languages such as R and Python to support the generation of actionable insight*

Python was the main tool I used to conduct the analysis. I made use of libraries such as Pandas for data manipulation, SciKit-Learn for Naive Bayes classification, and Flair for sentiment analysis. Python allowed me to automate the processing of the dataset, apply machine learning models, and generate various outputs such as confusion matrices, sentiment scores, and N-grams. This enabled me to derive actionable insights, such as the character with the most dialogue or the overall tone of the show.

*Goal 5: Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)*

In the conclusion, I effectively communicated the findings to a broader audience by summarizing the key insights in a clear and concise manner. I highlighted significant findings, such as

Michael being the most dominant character in terms of dialogue, the overall positive tone of the show, and the bias in the Naive Bayes model. Additionally, the use of visual aids, such as the confusion matrix and the ranking of N-grams, helped make the insights more accessible. By doing so, I communicated not only the results but also the implications of those results for understanding the character dynamics within The Office.

## Conclusion

Through my academic journey in the Applied Data Science program, I have had the opportunity to apply key learning goals across a variety of projects, gaining valuable skills and experience in data science, machine learning, and analytics. These projects have not only broadened my technical expertise but also deepened my understanding of how to use data to generate actionable insights across different domains.

In the IST 736 - Text Mining project, I applied fundamental data collection and management techniques, as well as advanced text mining and predictive modeling methods. This allowed me to explore the unique ways in which data science techniques can be applied to unstructured text, such as character dialogue in a popular TV show. While challenges such as model accuracy arose, the project reinforced the importance of using multiple analytical approaches to generate deeper insights.

The IST 664 - Natural Language Processing project provided an opportunity to dive into more advanced linguistic analysis, applying Python-based NLP techniques to classify names based on gender. This project underscored the real-world applicability of data science methods in everyday scenarios while also highlighting the limitations of simplistic models and the need for continuous improvement.

The IST 615 - Cloud Management project offered hands-on experience with cloud technologies, particularly in analyzing the cost and efficiency gains of migrating an organization's infrastructure to AWS. This experience reinforced the critical importance of data-driven decision-making in IT and business strategy, demonstrating how data science can lead to significant operational improvements and cost savings.

In the SCM 651 - Business Analytics course, I worked on a project analyzing financial data to uncover patterns that influence personal loan decisions. By utilizing statistical models and machine learning techniques, I was able to extract actionable business insights that could help a bank refine its marketing and customer outreach efforts. This project also emphasized the importance of considering ethical considerations in data analysis, particularly in terms of fairness and bias.

Finally, in the IST 652 - Scripting for Data Analysis project, I combined various programming languages and tools to clean, process, and analyze large datasets. This project reinforced the

importance of scripting skills in building data science pipelines and effectively communicating insights to both technical and non-technical audiences.

Overall, these projects demonstrate how data science techniques, from text mining and machine learning to cloud management and business analytics, can be applied to real-world challenges. They also emphasize the importance of using appropriate technologies, effective programming, and clear communication to derive actionable insights that inform decision-making in a variety of domains.