IST 782 Applied Data Science Portfolio

Dan Burke

June 2024

Syracuse University

Dr. Stinnett

Complied Course Work

Please use the following URL for Access to All Course Work:
https://github.com/dudemandando/SYR_Complied_CourseWork

Introduction

Pursuing a Master's degree in Applied Data Science at Syracuse University has been akin to undertaking a journey through a dense, vibrant ecosystem of knowledge, technology, and innovation. My journey within this program spanned approximately two years, mirroring the lifespan of the ruby-throated hummingbird. These hummingbirds live brief but dynamic lives of about two years, characterized by unparalleled speed and energy—a fitting metaphor for my transformative experience throughout this program.

Just as the ruby-throated hummingbird must swiftly adapt and navigate its environment, I evolved through my coursework's practical applications, developing skills in applying data-driven technologies as well as new thought processes for addressing both data-centric and non-data-centric problem sets. The curriculum encompassed an extensive range of topics, from the foundational principles of data administration and database management in IST 659 to the intricate techniques of Natural Language Processing (NLP) in IST 664. Each course subsequently

influenced my development in building a comprehensive understanding of data science and the application of machine learning and data science technologies.

Immersed in the complexities of Big Data Analytics (IST 718) and the practicalities of Scripting for Data Analysis (IST 652), much like the hummingbird rapidly learns to fly, forage, and survive, I too learned to forage through data, fly through scripting, and survive the application of machine learning algorithms and data analysis. The process was demanding, requiring dedication and adaptability. Yet, it was through this intense and focused effort that I acquired the skills which I was able to apply directly to my current employment, developing data-driven sophisticated software systems utilizing Computer Vision (OpenCV Python Distribution) for geospatial image analysis and real-time fiducial marker tracking (ARUCO), leveraged within broader Augmented and Virtual Reality Applications.

My motivations for completing this program were twofold: to expand my technical proficiency beyond software engineering and to gain the skills necessary to manage Data Science and Machine Learning/AI projects, programs, and employees with this skill set. This journey has equipped me with the expertise needed to oversee complex projects and lead teams in data science initiatives.

The Master's of Applied Data Science program aims to produce graduates who can:

- Collect, store, and access data by identifying and leveraging applicable technologies.
- Create actionable insights across a range of contexts (e.g., societal, business, political) using data and the full data science life cycle.
- Apply visualization and predictive models to help generate actionable insights.
- Use programming languages such as R and Python to support the generation of actionable insights.
- Communicate insights gained via visualization and analytics to a broad range of audiences, including project sponsors and technical team leads.
- Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).

As I delve deeper into the specifics of my journey, the comparison to the ruby-throated hummingbird serves to highlight the intensity, purpose, and rapid development that defined my pursuit of a Master's degree in Applied Data Science. Each moment was utilized to its fullest potential, resulting in a profound metamorphosis that has equipped me to tackle the challenges of the modern data-driven world.

Course Work

- **IST 659** - Data Administration Concepts and Database Management (3 credits): Focused on the principles of data administration and the management of database systems.
- **IST 687** - Introduction to Data Science (3 credits): Provided a foundational understanding of data science, covering essential concepts and techniques.
- **IST 707** - Applied Machine Learning (3 credits): Explored practical applications of machine learning algorithms and models.
- **IST 718** - Big Data Analytics (3 credits): Delved into the analysis of large datasets using big data technologies and tools.
- **IST 772** - Quantitative Reasoning for Data Science (3 credits): Developed quantitative skills necessary for data science, including statistical analysis and reasoning.
- **SCM 651** - Business Analytics (3 credits): Applied data analytics techniques to solve business problems and generate insights.
- **IST 664** - Natural Language Processing (3 credits): Covered techniques and applications of processing and analyzing natural language data.
- **IST 736** - Text Mining (3 credits): Focused on extracting useful information and patterns from large text datasets.
- **IST 652** - Scripting for Data Analysis (3 credits): Taught scripting skills for data analysis using programming languages like Python.
- **IST 722** - Data Warehouse (3 credits): Examined the design, implementation, and management of data warehouses.
- **FIN 654** - Financial Analytics (3 credits): Applied data science techniques to financial data for analysis and decision-making.

Motivations for Data Warehouse and Scripting For Data Analysis Electives

I have chosen to take electives in Data Engineering and Data Pipelines. I believe this field forms the foundation for larger client/customer-facing technology products and applications. Data Engineering is crucial for increasing runtime efficiency, decentralizing systems, and integrating various data sources to create a rich and seamless experience for clients and customers. These skills are vital for developing scalable and reliable data infrastructures that support advanced analytics and machine learning applications.

For a software engineer, studying Data Engineering and Data Pipelines offers additional benefits. It enhances the ability to design and implement robust data architectures, ensuring data quality and integrity. Knowledge in this field enables the automation of data workflows, reducing manual effort and minimizing errors. Furthermore, understanding data engineering principles is essential for real-time data processing, which is increasingly important in applications such as IoT and streaming analytics. This expertise also supports cross-functional collaboration with data scientists, analysts, and business stakeholders, leading to more integrated and effective technology solutions. Overall, these skills are indispensable for building the next generation of intelligent, data-driven applications.

Motivations for NLP Specialization

I chose to specialize in Natural Language Processing (NLP) for several compelling reasons. While the allure of popular machine learning applications such as image generation and prediction attracts many, I recognized the often overlooked power and potential of NLP.

As a working software engineer specializing in virtual and augmented reality, NLP was a foreign skillset. I saw this new skill as an opportunity for learning and a challenge which would expand my technical repertoire and perspective. NLP stands out due to its ability to process and analyze vast amounts of textual data, extracting valuable features, themes, and characteristics that are critical for understanding and leveraging human communication.

This capability is especially important in applications like sentiment and intent analysis, where understanding the nuances of language can provide deep insights into emotions, intentions, and behaviors. These insights are essential for improving user experience, enhancing customer service, and informing decision-making processes across various contexts all exceptionally vital within my field of work; technology consulting within the federal space (government contracts).

By specializing in NLP, I aimed to harness and further leverage this powerful toolset within my current work, to address complex problems and develop my technical skills and obtain a new lens of which to view technical problem sets. This

focus aligns with my broader goals of expanding my technical proficiency beyond software engineering and gaining the skills necessary to manage data science and machine learning projects effectively. My motivation for specializing in NLP is rooted in its profound ability to transform unstructured text into actionable insights, making it a critical component of the data science landscape and a valuable addition to my expertise in virtual and augmented reality.

Data Engineering and Data Pipelines Electives

I choose electives in Data Engineering and Data Pipelines—specifically, IST 652 (Scripting for Data Analysis) and IST 722 (Data Warehouse). I believe that data pipelines and data engineering are crucial for creating large, well-functioning software systems, even though they often don't attract much attention.

These fields provide the foundation for advanced analytics, ensuring data quality and integrity, and support machine learning and AI by making data available and usable. My primary motivation, however, is their critical role in supporting data-driven applications. Data engineering and pipelines provide the necessary infrastructure for these applications, enabling real-time data processing, integration of various data sources, and efficient data workflows. This ensures that data-driven applications, such as recommendation systems and business intelligence tools, function seamlessly and effectively, driving better decision-making and innovation.

Additionally, data engineering and pipelines are crucial for modeling and simulation systems, which are often used in military and video game simulations. These systems rely heavily on accurate, real-time data to create realistic and responsive environments. By ensuring robust data infrastructure, data engineering supports the development and maintenance of these complex simulation systems, enhancing their performance and reliability.

Application of Skills Acquired

Throughout my tenure as a student in the Master of Science in Applied Data Science program, I've been fortunate to pursue this program on a part-time basis while working as a software engineer, allowing me to directly apply new skills to my existing work. Within the past year, there have been multiple projects in which I applied my new machine learning competency.

First, "Project Goldeneye," an augmented reality "system of systems," that allows military planners to create "what if" logistical scenarios on a digital augmented reality "game board." Second, a virtual reality training application required the recognition of user-performed hand and arm signals. Third, creating an "Elevation Server", a server which is capable of delivering (via API) topographical elevation data of a user-provided area of interest (latitudinal and longitudinal bounds).

Project Goldeneye

The key stakeholders' request for Project Goldeneye was to increase efficiency in military logistical planning and operations, allowing for as many "reps and sets" (iterations) of "what if" (wargaming) simulated scenarios as possible. The primary requirement was to build a system that minimizes the need for labor or user effort in consulting logistical models (typically an Excel spreadsheet with estimated fuel burn, provision usage, etc.) while wargaming.

This problem set requires tracking real world physical game pieces within a real world space and then translating them to a digital space. Initially, the solution was to repurpose virtual reality body tracking "pucks" to act as game pieces. However, this approach had a scalability issue, as only 12 VR pucks can be successfully tracked at a time with commercial off-the-shelf solutions (HTC Vive).

Approaching this problem set from a new data science perspective allowed me to consider using computer vision to track the game pieces via an overhead camera. The first iteration utilized OpenCV (Python distribution) with QR codes. However, it became apparent that QR codes are wildly inefficient for this use case. Each QR code must be recognized, its physical pose estimated, and then decoded. The system only needed to match a QR code to a digital entity stored in an adjacent system rather than decode high-resolution data within the QR code itself. This led me to consider a lower resolution tracking solution: fiducial markers (specifically

ARUCO markers), which only encode an integer, decreasing the overall computational complexity.

ARUCO markers are often used in the autonomous robotics field to allow a vehicle or robotic system to determine the direction and distance of its camera to an ARUCO marker within its field of view. This is slightly different from the needs of Project Goldeneye, which required estimating the position of a marker within a physical space rather than relative to the observing camera. This was remedied by placing ARUCO markers at the corners of the game board, acting as reference points from which the relative position of the game pieces could be determined.

The final challenge was twofold: translating marker positions and pose from camera space to real-world game board space, as well as correcting for imperfect camera placement. In an ideal setting, the camera would be placed exactly perpendicular to the plane of the game board and pieces at an exact predetermined distance. However, budget restrictions and the possibility of human error prevented equipment that could allow for precision camera placement. This led to the exploration and implementation of homography within OpenCV, or in more simple terms, projective transformation between two planar projections of an image. Homography allowed the observing camera to be placed in multiple static locations above the game board, morphing the perspective of the resultant image using the pose of game board boundary ARUCO markers as if the camera were perpendicular to the physical plane of the game board. This adjusted result, combined with known

physical measurements of the game board and known ARUCO marker dimensions, allowed for millimeter-accurate distance measurements of game pieces to game board boundary ARUCO markers.

Gesture Recognition

The second application of the skills learned throughout this program was to create a code base within a virtual reality application that can recognize user-performed hand and arm signals (of low complexity) by tracking controller positions. This task presented multiple challenges: there was no ability to leverage computer vision, it had to be done in real-time at such a pace that it was perceived as instantaneous, and it must not be computationally intensive.

Within game development, the use of "colliders" (digital game objects that provide collision/interception events within the digital game space) are widely used. One solution to gesture recognition of hand and arm signals is to leverage colliders to receive controller positional detection events; however, they are wildly inefficient at runtime. Without the use of colliders and with the only available data being VR controller positional data, I approached this task from a data perspective.

VR controllers are attached to the user's hands, which are attached to the user's arms. This means that when a user performs a hand and arm signal, the controller will move similarly to a brush stroke in three-dimensional space. If one were to look at the user from the back, they would observe the "brush stroke" along

the X and Y axes. If one were to observe the user's hand and arm signal from the side, they would see the "brush stroke" from the Y and Z axes.

Given these two perspectives, one can take the controller positional points and plot them on two sets of axes (perspectives). Then, using regressions created from reference datasets for each hand and arm signal, the residual sum of squares would be computed from the user-generated controller positions for each model. If the computed residual sum of squares fell within a development-defined threshold, it would be marked as a valid gesture.

This approach, utilizing regressions and the residual sum of squares, unfortunately, has not been as successful as intended. However, it served as the foundation for which to iterate, engage this problem set, focusing on user-generated data rather than an inefficient, analogue-like collider approach.

Elevation Server

Our final example of applying the skills learned throughout this program is the creation of an “elevation server.” This project involved creating a lightweight server that could ingest Shuttle Radar Topography Mission (SRTM) data, index it, and return SRTM imagery for a user-defined area of interest. This server was utilized both at my current employment and for my final project in IST 652 - Scripting for

Data Analysis, further expanding it to leverage OpenCV for conducting edge detection on output satellite imagery.

SRTM (Shuttle Radar Topography Mission) data is a collection of high-resolution elevation data obtained by NASA during an 11-day mission in February 2000 using radar interferometry. The mission produced detailed maps of the Earth's surface with a resolution of about 30 meters for most regions and 90 meters for areas outside the United States, covering approximately 80% of the Earth's land surface between latitudes 60° North and 56° South.

This project began with searching GitHub for open-source code that could be leveraged. I initially found a Python-based elevation server that utilized SRTM data to output the elevation of a specific geographic point (latitude and longitude). I further adapted this open-source code to leverage its SRTM data indexing ability and combined it with my own written code to fetch, concatenate, and truncate SRTM imagery tiles. Running within a local Linux environment, I was able to fetch data and ingest it into a Python notebook, which utilized OpenCV to conduct edge detection.

Conclusion

Completing the Master’s degree in Applied Data Science at Syracuse University has been a transformative experience, much like the dynamic and adaptive life of a ruby-throated hummingbird. The program’s diverse curriculum has

equipped me with a comprehensive understanding of data science and machine learning technologies and techniques. From foundational courses in database administration and management, machine learning and financial analytics to specialized topics in Natural Language Processing and Data Warehousing, each class has significantly contributed to my professional and technical growth.

Choosing electives in Data Engineering and Data Pipelines has proven invaluable, providing the skills necessary to build robust, scalable data infrastructures that support advanced analytics and machine learning applications. My focus on NLP has opened new avenues for leveraging textual data, enhancing my ability to develop innovative solutions and gain deeper insights into human communication, which are critical in my role as a software engineer.

Applying these skills to real-world projects such as Project Goldeneye, gesture recognition in virtual reality, and the creation of an elevation server has demonstrated the practical value and impact of the knowledge acquired. These projects have not only advanced my technical capabilities but also reinforced the importance of data-driven solutions in various applications, from military logistics to enhancing user experience..

Overall, this journey has not only expanded my technical proficiency but also prepared me to lead and manage complex data science and machine learning projects. With the skills and experiences gained from this program, I am well-

equipped to tackle the challenges of the modern data-driven world and continue driving innovation in my field and with those whom I work with.

Looking Forward

Moving forward from my studies in Applied Data Science, I plan to use these new skills and competencies to expand my professional work into autonomous systems, vehicles, and robotics, specifically focusing on unmanned underwater vehicles (UUVs) and the control of those vehicles. I aim to leverage these vehicles to create data-centric systems for capturing bathymetric data from underwater surveys.

A key component of this endeavor will be developing autonomous "command and control systems" to manage and operate the UUVs. These systems will enable precise and efficient control of the vehicles, allowing them to navigate complex underwater environments autonomously. By integrating advanced data science techniques, machine learning algorithms, and predictive models, I can enhance the accuracy and efficiency of these surveys.

In addition to UUVs and data systems, I plan to leverage my new skills within the field of distributed sensors (sensor fusion). I aim to build systems that collect and monitor real-time data streams from optical electrical cameras, infrared

cameras, acoustical sensors, and radio frequency sensors. This approach will enable the creation of comprehensive, multi-modal datasets that provide a richer and more detailed understanding of the environments being studied.

By combining autonomous vehicle control, data-centric systems, and distributed sensor networks, I will be able to develop innovative solutions that push the boundaries of current technology. This integration will provide valuable insights into underwater terrains and other environments, contributing to advancements in marine exploration, research, and beyond. The combination of data science with autonomous command and control systems and sensor fusion represents the next frontier in my career, enabling me to apply my expertise to innovative and impactful projects in various fields.