Syracuse University, School of Information Studies

M.S. Applied Data Science

Portfolio Milestone

Bryan J. Crigger

SUID: 255676562

Bryan.crigger7@gmail.com

1. Program Overview
   * What term did I take each class, and what were the Course Overview and Learning Objectives
2. Choose what projects I liked the most (Top 4-5 projects)
3. Move these Top projects (reference link) to Portfolio class folder
4. Have a write up for each of these projects
5. Include focus time slots on calendar to allow time to work on Portfolio (reoccurring each week)
6. Should list out the Program Learning Objectives for each project and write out for each how the project did (or didn't) help achieve that objective, and also by how much (i.e. Project X helped me enhance my knowledge for objective Y significantly by doing A, B, C, etc.)
7. Also list out for each project what:
   * Problem(s) were solved or were trying to be solved
   * Tools & techniques were used
   * Insights/results were learned as a result of the project

# Program Overview

## Summer 2021

#### Business Analytics

Course Overview and Learning Objectives:

* This course is intended for the graduate student who is interested in developing a portfolio of skills in business analytics.
* The course learning objectives include:

1. Data collection: using tools to collect and organize data (e.g., Google Analytics)
2. Data analysis: identify patterns in the data via visualization, statistical analysis, and data mining
3. Strategy and decisions: develop alternative strategies based on the data
4. Implementation: develop a plan of action to implement the business decisions

* Final: Exam

## Fall 2021

#### Introduction to Data Science

Course Overview and Learning Objectives:

* The course introduces students to applied examples of data collection, processing, transformation, management, and analysis to provide students with a hands-on introduction to the data science experience. Students will explore key concepts related to data science, including applied statistics, information visualization, text mining, and machine learning. “R,” the open-source statistical analysis and visualization system, will be used throughout the course. R is reckoned by many to be the most popular choice among data analysts worldwide; having knowledge and skill with using it is considered a valuable and marketable job skill for most data scientists.

Learning Objectives:

* After taking this course, the students will be able to understand:
  + Essential concepts and characteristics of data
  + Scripting/code development for data management using R and R-Studio
  + Principles and practices in data screening, cleaning, and linking
  + Communication of results to decision makers
* At the end of the course, students are expected to be able to:
  + Identify a problem and the data needed for addressing the problem
  + Perform basic computational scripting using R and other optional tools
  + Transform data through processing, linking, aggregation, summarization, and searching
  + Organize and manage data at various stages of a project life cycle
  + Determine appropriate techniques for analyzing data
* Final: Project: [Diving into Natural Disasters](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2021\IST687%20-%20Introduction%20to%20Data%20Science\Project\Final%20Project%20Details)
  + Explored U.S. Natural Disaster data using R, to uncover various patterns in the data such as what types of natural disasters occur in what parts of the country, how often, and to what severity.
  + [Need to go through the presentation/code/recording to see if this might be good to include in portfolio, mainly to show a baseline since this was my first class project of the program]

## Spring 2022

#### Applied Machine Learning

Course Overview and Learning Objectives:

* Introduction to data mining techniques, familiarity with particular real-world applications, challenges involved in these applications, and future directions of the field. Hands-on experience with open-source software packages.
* This course will introduce popular data mining methods for extracting knowledge from data. The principles and theories of data mining methods will be discussed and will be related to the issues in applying data mining to problems. Students will also acquire hands-on experience using state-of-the-art software to develop data mining solutions to scientific and business problems. The focus of this course is in understanding of data and how to formulate data mining tasks in order to solve problems using the data.
* The topics of the course will include the key tasks of data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation and analysis. Through the exploration of the concepts and techniques of data mining and practical exercises, students will develop skills that can be applied to business, science or other organizational problems.
* The format of the class meetings will be a combined lecture and lab format, with lectures and class discussions to cover material and lab time to investigate small examples for the topic of the week. There will be weekly readings based on the textbook and on other materials, which will be posted on-line.

Learning Objectives:

* After taking this course, the students will be able to:
  + Document, analyze, and translate data mining needs into technical designs and solutions.
  + Apply data mining concepts, algorithms, and evaluation methods to real-world problems.
  + Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.
* Final: Project: [Predicting U.S. Opioid Prescribers](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2022\IST707%20-%20Applied%20Machine%20Learning\Project)

#### Database Administration Concepts and Management

Course Overview and Learning Objectives:

* IST 659 is an introductory course to database management systems. This course examines data structures, file organizations, concepts, and principles of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementation. More specifically, it introduces hierarchical, network, and relational data models; entity-relationship modeling; basics of Structured Query Language (SQL); data normalization; and database design. Using Microsoft’s Access and SQL Server DBMSs as implementation vehicles, this course provides hands-on experience in database design and implementation through assignments, lab exercises, and course projects. This course also introduces advanced database concepts such as transaction management and concurrency control, distributed databases, multitier client/server architectures, web-based database applications, data warehousing, and NoSQL.
* After taking this course, the students will be able to:
  + Describe fundamental data and database concepts
  + Explain and use the database development lifecycle
  + Create databases and database objects using popular database management system products
  + Solve problems by constructing database queries using Structured Query Language (SQL)
  + Design databases using data modeling and data normalization techniques
  + Develop insights into future data management tool and technique trends
  + Recommend and justify strategies for managing data security, privacy, audit/control, fraud detection, backup and recovery
  + Critique the effectiveness of DBMS in computer information systems
* Final: Project: [Household Item Database](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2022\IST659%20-%20Dasebase%20Administration%20Concepts%20and%20Management\Project)
  + Created a business use case, designed the ERD (Entity-Relationship Diagram), and built the tables with sample data with SSMS and Access.
  + Summary from Project write-up:

“This data model shows a summary of the items we have in our house, along with other important information I wanted to track including store purchased from, date of purchase, item type, and others. This is an initial framework to build and add additional items too in the future. There were 5 questions that I wanted to be able to answer from building this database that revolved around how much have we spent on items around the house. Luckily with SSMS and Access I was able to do just that. For now, with myself and my boyfriend being the Stakeholders and myself being the sole user, Access seemed like a great, easy option to have as the UI. My next steps for this database will be to add additional items to have even more representative data, and to better customize reports for nicer data summary views.”

## Summer 2022

#### Advanced Database Management

Course Overview and Learning Objectives:

* An analysis of relational and nonrelational databases and their corresponding database management system architectures. Learn to build complex database objects to support a variety of needs from both the big data and traditional perspectives. Data systems performance, scalability, and security.

Learning Objectives:

* Understand advanced issues with the relational database model, such as transactions, performance, and security, as to understand the need for other database models.
* Explain the CAP theorem and describe how any given database system’s architecture fits within the CAP context.
* Compare different database models such as document, key-value, column-family, streaming, and relational.
* Identify the most suitable database systems for a specific application’s data storage requirements.
* Evaluate relational, Hadoop, and noSQL database tooling as to understand their underlying similarities and necessary differences.

## Fall 2022

#### Enterprise Risk Management

Course Overview and Learning Objectives:

Learning Objectives:

* The design of this course focuses on the following knowledge and performance objectives:
* **ERM process:** You will become familiar with the Enterprise Risk Management (ERM) process in a variety of different organizational contexts. This includes evaluating challenges and benefits of transitioning to an enterprise-wide risk approach.
* **Definitions/vocabulary:** You will learn and apply the common risk management vocabulary so that you may more effectively participate in discussions about enterprise-wide risk management.
* **Risk culture and risk tolerance:** You will learn and apply different theoretical frameworks that underlay individual and organizational perception of risk, risk tolerance, and risk culture. This includes understanding how risk culture and risk tolerance influence perceptions of risk priorities, as well as shape ERM activities. This will enable you to better participate in and develop risk management programs, strategies, and practices
* **Value of ERM:** You will evaluate ERM in context of both upside (positive) and downside (negative) risk, including influences on how concepts of “risk, opportunity, loss, and benefit” are defined or measured in different industry contexts.
* **Risk categories:** You will analyze, synthesize, and evaluate cases involving a wide range of enterprise risks, including environmental disasters, hazards, legislative, social, political, financial, and operational risks. This will enable you to discern more effectively the broad range of enterprise-wide risks facing different types of firms.
* **Risk measurement:** You will learn and apply different qualitative and quantitative risk measurement approaches so you can evaluate and recommend appropriate methods for measuring and monitoring different types of risk.
* **Risk management practices:** Managing risk includes mitigation of risk, transfer of risk, sharing of risk, or acceptance/avoidance of risk. We will explore examples of these techniques and their rationale.
* **Hands-on practice:** You will research, synthesize, and evaluate risk management practices in one or more real organizations in order to add to your practical experience and understanding of enterprise-wide risk management.

## Winter 2023

#### Managing Data Science Projects

Course Overview and Learning Objectives:

* Improve data science project outcomes by improving the process a team uses to execute their project. Explore how to integrate data science life cycles / workflows (e.g., CRISP-DM, TDSP) with agile collaboration frameworks (e.g., Kanban, Scrum).

Learning Objectives:

* After taking this course, the students will be able to:
  + Leverage agile concepts within a data science project context
  + Describe the key differences between data science projects and software development projects
  + Explain the responsibilities of the different roles within a data science project
  + Identify and minimize the potential ethical implications of a data science project
  + Explain the general data science life cycle, and be able to use several standard workflows frameworks, such as CRISP-DM and TDSP.
  + Articulate the key aspects of coordination frameworks, such as Kanban and Scrum.
  + Select and use the most appropriate team process framework for a specific project
* Managing Data Science Projects: [Designed Data Science Workflow Process for BigNorth Airlines](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST644%20-%20Managing%20Data%20Science%20Projects)

## Spring 2023

#### Natural Language Processing

Course Overview and Learning Objectives:

* This course is designed to develop an understanding of how natural language processing (NLP) can process written text and produce a linguistic analysis that can be used in other applications. This goal will be achieved by:
  + Readings, lectures, and class discussions of the multiple levels of linguistic analysis required for a computer to accept natural language input, interpret it, and carry out a particular application.
  + Lab exercises and assignments in using some of the computational techniques required to perform these levels of natural language processing of text.
  + Studies of real-world applications that incorporate substantive NLP modules.
* The course primarily covers the techniques of NLP in the levels of linguistic analysis, going through tokenization, word level semantics, part-of-speech tagging, syntax, semantics, and on up to the discourse level. It also includes the use of the NLP techniques, such as information retrieval, question answering, sentiment analysis, summarization, and dialogue systems, in applications.

Learning Objectives:

* At the end of the course the student will be able to:
  + Demonstrate the levels of linguistic analysis, the computational techniques used to understand text at each level, and what the challenges are for those techniques.
  + Process text through the language levels using the resources of the Natural Language Toolkit (NLTK) and some rudimentary use of the programming language Python.
  + Describe how NLP is used in many types of realworld applications.
* Natural Language Processing: [Classifying Spam vs. Non-Spam Emails with NLP](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST664%20-%20Natural%20Language%20Processing\Final%20Project)

## Summer 2023

#### Big Data Analytics

Course Overview and Learning Objectives:

* A broad introduction to analytical processing tools and techniques for information professionals. Students will develop a portfolio of resources, demonstrations, recipes, and examples of various analytical techniques.
* You will find if much easier to succeed if you have completed IST687, IST777 or both. Familiarity with command-line interfaces, basic quantitative skills, including statistics, as well as programming skills with languages such as R or Python. Most of the course work will be using Python, Spark, and Tensorflow.

Learning Objectives:

* During the course, we will emphasize:
  + Experiential learning through reading and practical exercises.
  + Collaborative learning through online discussions between instructors and peers.
  + Self-learning with appropriate instructional support and timely feedback using analytical case studies.
* In order to be successful in this course, the student will:
  + Pro-actively research solution options vs. relying solely on textbook content.
  + Actively code while completing the reading assignments.
  + Present results in a professional manner. Comments – Clarity – Correctness – Credit.
  + Submit their assignments on time.
* After taking this course, the students will be able to:
  + Obtain data and explain data structures and data elements.
  + Scrub data by applying scripting methods, to include debugging, for data manipulation in Python, R or other languages.
  + Explore data by analyzing using qualitative techniques including descriptive statistics, summarization, and visualizations.
  + Model relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users.
  + INterpret the data, model, analysis, and findings. Communicate the results in a meaningful way.
  + Select an applicable analytical methodology for real problems in areas such as business, science, and engineering.
* Final: Project: [Quantifying the Impact that features within a home have on Overall Home Price](file:///C:\Users\bryan\_Syracuse%20Jupyter%20Notebook\Big%20Data%20Analytics\IST718_Big_Data_Analytics\Project)

## Fall 2023

#### Deep Learning in Practice

Course Overview and Learning Objectives:

* This course will introduce deep learning methods for classifying data and predicting outcomes. Both the theories and the practice of deep learning techniques will be discussed and applied to real-world problems. Students will acquire hands-on experience developing deep learning–based solutions using state-of-the-art deep learning–based software.
* The format of the class meetings will be a combined lecture and lab format, with lectures and class discussions to cover material and lab time to investigate small examples for the topic of the week. There will be weekly readings based on the textbook and on other materials, which will be posted online.

Learning Objectives:

* Document, analyze, and translate data analytics needs into technical designs and solutions.
* Apply deep learning–based concepts, algorithms, and evaluation methods to address real-world problems.
* Final: Project: [Classifying Pistachio Images by Species](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST691%20-%20Deep%20Learning%20in%20Practice\IST691-Deep-Learning-in-Practice\Project)

## Winter 2024

#### Scripting for Data Analytics

Course Overview and Learning Objectives:

* Scripting for the data science pipeline. Acquiring, accessing, and transforming data in the forms of structured, semistructured, and unstructured data.
* 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.
* The main content focus is on information access and processing tasks on the types of structured, semistructured, and unstructured data in current use in information applications. For these three types of data, the course includes the use of structured numeric and text data such as that from a spreadsheet or database, the use of data obtained through standard data exchange formats such as HTML or XML from web pages or JSON from web-based APIs, and the use of data obtained by pattern matching from text or log files. The scripting language Python was chosen because of its ease of use and available packages to work with data in many information applications. The skills learned in this class are intended to complement the analytical and visualization skills learned in other data science courses. The scripting language Python will be taught, but it will be assumed that students already have a programming background, through either coursework or online study.

Learning Objectives:

* Upon successful completion of this course, the student will be able to:
  + Write scripts to access and amass data from fields in structured data, access fields in semistructured data, and define and find patterns of data in unstructured data
  + Prepare and transform data to produce data summaries, lists, and networks
  + Analyze and solve data access problems for the three types of data and to find and deploy appropriate software packages that can be integrated into the problem solution
  + Frame real-world data questions and show how they can be answered from data

#### Applied Data Science Portfolio

Course Overview and Learning Objectives:

* This portfolio

# Projects that I want to Include in Portfolio (Top 4-5 projects)

* Applied Machine Learning: [Predicting U.S. Opioid Prescribers](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2022\IST707%20-%20Applied%20Machine%20Learning\Project)
* Managing Data Science Projects: [Designed Data Science Workflow Process for BigNorth Airlines](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST644%20-%20Managing%20Data%20Science%20Projects)
* Natural Language Processing: [Classifying Spam vs. Non-Spam Emails with NLP](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST664%20-%20Natural%20Language%20Processing\Final%20Project)
* Big Data Analytics: [Quantifying the Impact that features within a home have on Overall Home Price](file:///C:\Users\bryan\_Syracuse%20Jupyter%20Notebook\Big%20Data%20Analytics\IST718_Big_Data_Analytics\Project)
* Deep Learning in Practice: [Classifying Pistachio Images by Species](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST691%20-%20Deep%20Learning%20in%20Practice\IST691-Deep-Learning-in-Practice\Project)

## More In-depth writeup on how these classes fulfilled the Learning Objectives of the program:

### Program Learning Objectives:

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)

## For the following classes/projects for the Portfolio, note if they fulfilled each of Program Learning Objectives, and if so, to what degree:

*Write out for each Learning Objective above, how the project did (or didn't) help achieve that objective, and by how much (i.e. Project X helped me enhance my knowledge for objective Y significantly by doing A, B, C, etc.).*

### Applied Machine Learning: [Predicting U.S. Opioid Prescribers](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2022\IST707%20-%20Applied%20Machine%20Learning\Project)

* Learning Objectives Fulfilled:
  1. Yes; Collected data from the Centers for Medicare & Medicaid Services website. Used R to do analysis, insights, and prediction models.
  2. Yes; The goal of our project was to classify and evaluate opioid prescribers to 1) highlight possible anomalies, 2) identify high risk areas, and 3) identify high opioid-prescriber types to assist early intervention programs.
  3. Yes
  4. Yes; used R in R Studio
  5. Yes; Presented results and findings from our project to the class. We highlighted the types of physicians that were the highest opioid prescribers, what types of opioids are most often prescribed by which types of physicians, and the analyses that we did to get these insights.
  6. Yes (something about how we may sure the results or anything that we shared was statistically sound / had a high accuracy)
* For the class’s final project, what:
  1. Problem(s) were solved or were trying to be solved
  2. Tools & techniques were used
  3. Insights/results were learned as a result of the project

### Managing Data Science Projects: [Designed Data Science Workflow Process for BigNorth Airlines](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST644%20-%20Managing%20Data%20Science%20Projects)

* Learning Objectives Fulfilled:
* For the class’s final project, what:
  1. Problem(s) were solved or were trying to be solved
  2. Tools & techniques were used
  3. Insights/results were learned as a result of the project

### Natural Language Processing: [Classifying Spam vs. Non-Spam Emails with NLP](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST664%20-%20Natural%20Language%20Processing\Final%20Project)

* Learning Objectives Fulfilled:
  1. Yes (something about how we may sure the results or anything that we shared was statistically sound / had a high accuracy)
* For the class’s final project, what:
  1. Problem(s) were solved or were trying to be solved
  2. Tools & techniques were used
  3. Insights/results were learned as a result of the project

### Big Data Analytics: [Quantifying the Impact that features within a home have on Overall Home Price](file:///C:\Users\bryan\_Syracuse%20Jupyter%20Notebook\Big%20Data%20Analytics\IST718_Big_Data_Analytics\Project)

* Learning Objectives Fulfilled:
  1. Yes (something about how we may sure the results or anything that we shared was statistically sound / had a high accuracy)
* For the class’s final project, what:
  1. Problem(s) were solved or were trying to be solved
  2. Tools & techniques were used
  3. Insights/results were learned as a result of the project

### Deep Learning in Practice: [Classifying Pistachio Images by Species](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Classes\Past%20Classes\2023\IST691%20-%20Deep%20Learning%20in%20Practice\IST691-Deep-Learning-in-Practice\Project)

* Learning Objectives Fulfilled:
  1. Yes; collected data through Kaggle, and read it in automatically into Google Colab
  2. No; wouldn’t say that we provided any actionable insight. We could have added a recommendation as to whether the business should incorporate this model or not, and/or provided caveats on how to use it.
  3. Yes; applied a predictive/classification CNN model to take images of pistachios to correctly classify them into 1 of 2 species
  4. Yes; used python to write the code
  5. Yes; recorded presentation of analysis and results of our image classification model, where we shared the accuracy of our best classification model, what the draw backs are of the final, best model, and how it could be further improved.
  6. Yes; when building and training the model we noticed that it was performing well overall, however when looking at the accuracy of the individual classes, one of the pistachio species was very low while the other was nearly perfect. [We noticed that it was learning to only guess one species all the time, and never guess the other pistachio type] [To account for this bias we \_\_\_]

To account for this bias we tried a few different techniques in which the model learned one class very well and the other one not so much, we increase the weight for properly classifying

* For the class’s final project, what:
  1. Problem(s) were solved or were trying to be solved
     + Image Classification for Pistachios, into their correct Species
  2. Tools & techniques were used
     + Python (numpy, keras, sklearn, matplotlib), Google Colab, Tensorboard, CNN, Transfer Learning, Fine Tuning, Data Standardization
  3. Insights/results were learned as a result of the project
     + Bigger, more complex models are not always better. When we tried fine-tuning a prebuilt CNN model, the pre-set weights from when the model was trained originally seemed to result in the model basically overfitting, only choosing to predict one class for all of the images.
     + Transfer learning and fine-tuning a model may be a better option than building a Deep Learning model from scratch.
     + When wanting/needing to use a Deep Learning model, you should start with the simplest model to set a baseline, and then make the model more complex. I think this was something I always think and talk about, but we didn’t do well here. After trying to start with doing transfer learning on a large, prebuilt CNN model and not getting great results, we tried simplifying the model and trying some simple, shallow models to see if we weren’t getting great results because of the data or if it was because of the model. By having that baseline, we were able to see that we actually had something wrong with some earlier code around normalizing the data/images.

# Other Notes & Links:

* [Personal Statement to get into the Applied Data Science Program](file:///C:\Users\bryan\Dropbox\Educational\Grad%20School\Syracuse\Prep%20Program%20Docs\Personal%20Statement%20Final.docx) [4/14/2021]
* [Great example for how final portfolio paper and other submission details should look](https://github.com/SLPeoples/MSADS_Portfolio/blob/master/SamuelPeoplesPorfolioMilestone.pdf)
* [Jazmin’s Github for Portfolio](https://github.com/JazminLogrono/Portfolio/tree/main)