# SYRACUSE

## M.S. IN APPLIED DATA SCIENCE
## JUSTIN PATE PORTFOLIO

SUID: 445414259
jdpate@syr.edu

GRADUATION DATE: WINTER 2020

### Introduction

Thank you for taking the time to review my portfolio for the Masters of Science in Applied Data Science at Syracuse. This has been one of the most challenging yet most rewarding processes I have ever experienced. My exposure to data science was minimal prior to the start of the program. My professional experience was limited to a data analytics and analysis approach. My previous experience gave me familiarity with relational databases and basic coding in various languages, but I would have never imagined that it did not really scratch the surface of the Data Science Envelope. The program has been an eye opener and has brought a greater understanding to the world of data, and for that, a greater understanding to actual human behavior.

### Professional Background

I received my bachelors while overseas in the Air Force from American Military University. I have to say that the Syracuse online learning platform that includes the live lectures, breakout sections and phenomenal asynchronous material is impressive and effective. My current profession is in the Aerospace manufacturing environment. My focus is on precision measurement which is a unique form of data collection in itself.

Precision measurement in the aerospace fields is the constant monitoring of machine accuracy and product conformance through various equipment. This aspect of analysis has allowed me to solve complex manufacturing problems and provide solutions.

The additional knowledge that I have gained from the Data science education from Syracuse can be applied in the same platform and predict these variations rather than simply waiting for a monitored trigger once a product is approaching an out of tolerance condition or a machine's recorded data measurements are starting to drift near a zone of non conformance.

### Overview

If I were to sum up the major contributor to data science, then I would focus on the "Human" element. In most exercises

throughout the program, the goal was to find patterns in human behavior or outcomes that were driven by human decision making. We analyzed the Titanic data set in almost every programming course. The primary contributor's there boiled down to demographic make up, but the data set is a reflection of human behavior at the time of data collection. Another driver is natural language processing.

I am sitting in a coffee shop as I write this portfolio, and because of what I have learned in this program, I am thinking about how much data could be collected right now just based on conversations happening around me. Could I analyze the likes and dislikes of the people sitting next to me? Could I predict how many children they have and where they live based on those likes and dislikes? Could I tell you the car that Random person number one owns and where they work and probable income?

## Learning Goals contained in Educational Projects

Outlined below are the learning objectives achieved through the program. Each item is mastered through project based learning.

1. **Major Practice areas** contained in project overview are:
    a. Text mining, Geo data, predictive modeling through training and test methods, clustering analysis, time prediction.
    b. Image recognition is another topic covered in depth during the program.
2. All projects contain an element of **Collecting and Organizing** data.
    a. Not all data is ready to use or easy to get.
    b. Part of a data scientist skill has to be preparing and cleaning the dataset.
3. **Identify patterns through visualization.**
    a. A picture is worth a 1000 words
4. If the original plan for analysis fails, show an **alternative strategy** based on the failure.
5. Using the results of the analysis, effectively communicate intelligent **business decisions** to the stake holder that will help in a path towards the desired goal.
6. Who is the audience for each **method of communication**?
    a. IT organization
    b. Business Leader/Manager
    c. Shop floor professional
    d. Etc…
7. Monitoring of any type of data has specific **ethical boundaries**.

## Project base learning

One of the largest strengths of the Data Science Program at Syracuse is the project based learning. Below is a list of projects that reflect both a general knowledge of data science and some specifics regarding analysis and recommendations that are good examples of real world application.

### Wing Measure Data:

https://github.com/jdcpate/Syracuse_Portfolio/tree/master/IST_687_WING_MEASURE_SUB

In this project, the goal was to identify dimensional features from an actual build of an airplane wing that are correlated. The data collection methods are through precision measurement equipment which is then analyzed to perfect 3d engineering models. The results were uploaded to a SQL database. We wanted to see if any upstream features affected the downstream variations and if those variations were predictable.

The Wing Measure data was one of the first final projects that allowed the demonstration of applied knowledge to my actual real world experience in data analysis. This was REAL data that was used. There was an obvious masking of specific products and customer's for confidentiality reasons, but the data was real. In this example, it was easy to see the additional difficulties that real world data would present. It is not the perfectly organized examples that are sometimes given through the Kaggle data competitions. I was forced to spend a lot of time on cleaning up the data really honing in on the best variables to use.

### Methods:

This project was written in Rscript. As this was real world data, the clean up methods included only complete records and searching to eliminate outliers. The final method choses for prediction was neural network which performed the best through multiple nodes representing the build process and estimating outcomes.

**Learning Objectives Completed:** Data Cleanup, Collecting and organizing real world data, Model Training and Testing, alternative strategies, Business Decisions, Well Defined audience of Manufacturing Engineers.

### IST-707_Employee_Attrition_Analysis

https://github.com/jdcpate/Syracuse_Portfolio/tree/master/IST_707_EMP_ATTRITION_PROJECT

The employee Attrition analysis project was another fantastic example of taking the knowledge learned and applying it to a real world issue that is relatable to any professional. In this project the goal was to analyze and predict the employees that would quit a job after only a short time and the employees that were more likely to stay "long term" as defined by the project definitions.

### Methods:

Using R script and Model training and testing this project dataset through trial and error with Naïve Bayes, Random forest and finally coming to the conclusion that Kmeans clustering was the best algorithm for this project.

**Learning Objectives Completed:** Data Cleanup, Collecting and organizing real world data, Model Training and Testing, alternative strategies, Business Decisions, Well Defined audience of Human Resources Department.

### IST-718_MANUFACT_MANPOWER

https://github.com/jdcpate/Syracuse_Portfolio/tree/master/IST-718_MANUFACT_MANPOWER

Estimating the fluctuations of manpower is important to any company. This is especially true in a manufacturing environment.

In the Big data final project, an analysis actual human resources time card data revealed useful information. Using the time based series Prophet model in python, we were able to accurately predict a less than expected decline in manpower during the summer months with increases prior to holiday months. An additional layer that would affect profitability during this time is the decrease in quality due to the back filling of manpower.

**Methods:**

This was a Python based project. The Primary goal of this project was to estimate Manpower and quality fluctuations by time, the primary model used is the Prophet model in Python. A unique library used for data clean up in this project was MSNO. This library is focused around finding the completeness of each column and rows and then producing built in visualizations that show the data gaps.

**Learning Objectives Completed:** Data Clean up, Text mining to identify Which hours were associated with non-conformance work. Collecting and organizing real world data, Model Training and Testing, alternative strategies, Business Decisions, Well Defined audience of Industrial Engineers and Manufacturing Supervision.

### IST-719_VIZATHON_ONONDAGA_LAKE_WATER_QUALIYT

https://github.com/jdcpate/Syracuse_Portfolio/tree/master/Onondaga_Lake_Water_Quality

In this project, the student was challenged to produce a poster in the Adobe illustrator using only R visuals and a dataset that the entire class was not familiar with. This project forced the students to quickly familiarize themselves with the data and produce meaningful results to tell a visual story while under the pressure of time. This project forced the use of GGplot and some other basic R visualizations. It also allowed the students to showcase their adobe illustrator skills which was a challenge in itself.

**Methods:**

This project was done in RScript and Adobe illustrator. The main visualization library was GGPlot using the export to PDF command from RStudio to import the Vector image into Adobe Illustrator.

**Learning Objectives Completed:** Visualization and Data cleanup

### IST-719_Final Poster Submittal

https://github.com/jdcpate/Syracuse_Portfolio/tree/master/P3_IST-719_POSTER_FINAL_SUBMITTAL

Even though, this class was taken in Term 5 for me, I looked at it as a good class to take at the end of the data science program. It forces the student back into the world of R studio and the class is more than just a data analysis class. You are now learning how to tell a comprehensive visual story with what you have learning through your entire experience in the data science program. The presentations of the final poster in this class were some of the most impressive I had seen through the entire program.

**Methods:**

This project was done in RScript and Adobe illustrator. The main visualization library was GGPlot using the export to PDF command from RStudio to import the Vector image into Adobe Illustrator. The primary audience for this project would be Human resources management. The visualization methods taught in this class were applied using the spiral method for poster story telling.

**Learning Objectives Completed:** Data Cleanup, Collecting and organizing real world data, Model Training and Testing, alternative strategies, Business Decisions, Well Defined audience of Human Resources Employees.

## MAR-653_SOLAR PANEL SALES

https://github.com/jdcpate/Syracuse_Portfolio/tree/master/P1_SOLAR_PANEL_PROJECT

The solar panel sales project was a fantastic combination of predictive modeling and using GGPLOT for mapping the dataset for a visual representation of data exploration and results. This project involved a large dataset of solar panel installation by geographical scans from a satellite image. The goal of the project was to identify based on local demographics and poverty levels where a good market for future solar panel sales would be.

**Methods:**

This project was done in RScript. Simple data clean up was performed to only include complete records. Kmeans clustering was the final decided upon method after experimenting with both linear regression and Naïve Bayes.

**Learning Objectives Completed:** Data Cleanup, Collecting and organizing real world data, Model Training and Testing, alternative strategies, Business Decisions, Well Defined audience of Marketing managers of the Solar Panel Industry.

## IST-652_CRIME DATA

https://github.com/jdcpate/Syracuse_Portfolio/tree/master/P4_IST-652_CRIME_DATA

IST-652 was a fantastic introduction to python programming methods. By the end of the course, I was able to create dynamic visualizations as well as functions to map any data set that contained a latitude and longitude. In this course, I was brave enough to show my python code in my final presentation and give a live mapping demonstration.

**Methods:**

This project was done in Python Colab. Multiple functions were written using Folium maps. Sampling had to be done as there were over 1.5 Million records that would crash the mapping API. There are user inputs to allow for a city selection to be mapped by the crime data.

**Learning Objectives Completed:** Data Cleanup, Collecting and organizing real world data, Mapping Geographic data, writing Functions, sampling data, Well Defined audience of Law Enforcement agencies and Potential Home Buyers.

Term 6 Learning Objectives:

In my final term, I look forward to learning about privacy and information security in the introduction to Information Security class. I will also become more familiar with complex text mining in the Text mining course.

## CONCLUSION:

The Syracuse Masters in Applied data science program has given the students a well rounded exposure to the majority of the data science tools, methods and scope that exist in this modern, data rich world. The program has also touched on ethical boundaries that should not be violated during analysis. The world is still entitled to privacy despite what some marketing companies might think. Through these new skill sets, the student is now equipped to succeed further in any profession or to cross over into a data science focused career path. Thank you Syracuse!