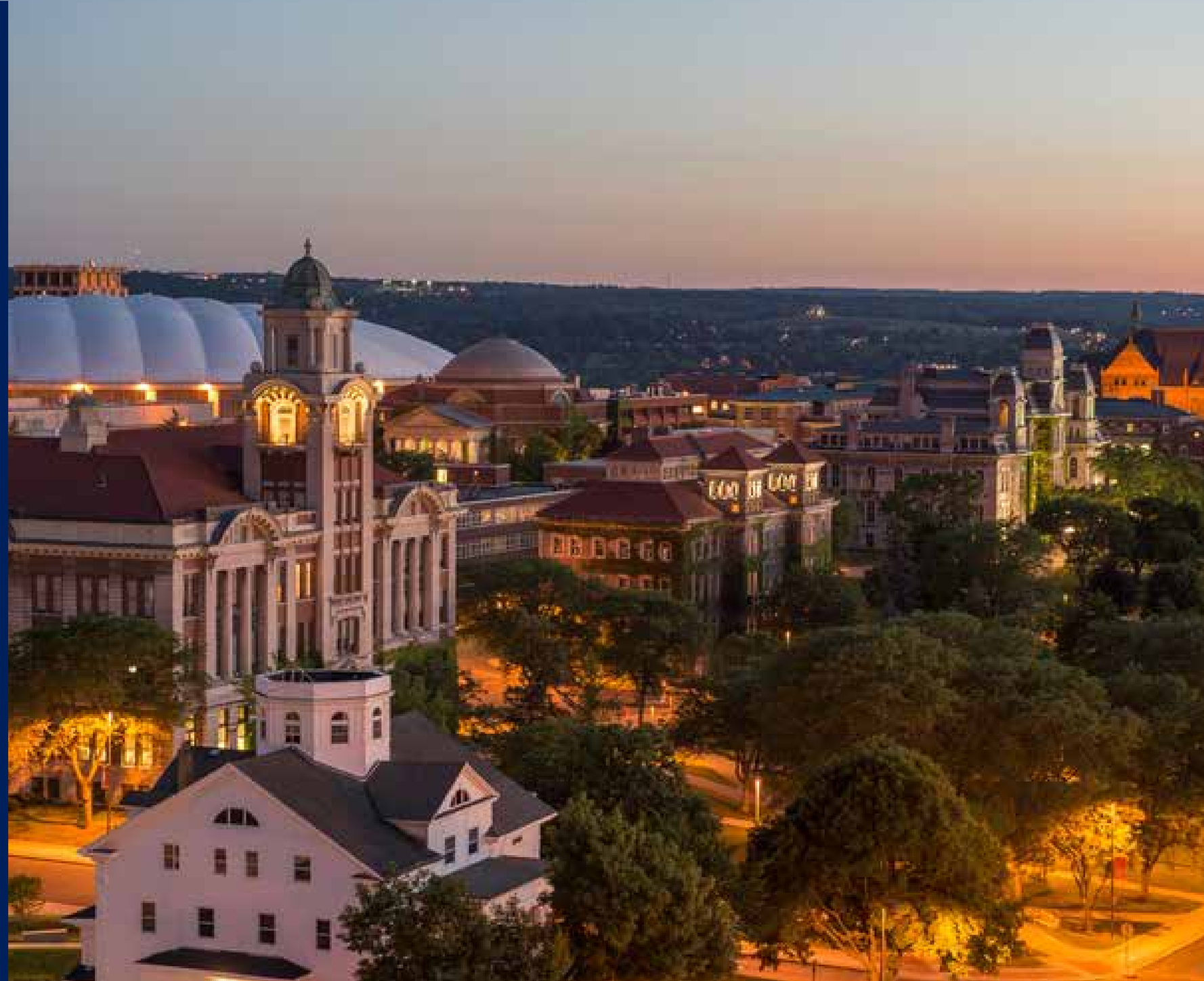
Syracuse University
School of Information Studies

MASTER OF SCIENCE
APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

MARK ROBERTS
SUID: 598273961
HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO
DECEMBER 2021



Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

INTRODUCTION

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

- The Applied Data Science program at Syracuse University's School of Information Studies provides students the opportunity to develop analytic, technical and managerial skills. Students are taught to collect, manage, analyse and develop insights that can be communicated clearly to a broad audience of interested parties.

- Coursework, Reports & Presentations in this Portfolio showcase these skills, and exemplify the intent of the program, in the following courses:

  - Data Administration Concepts & Database Management (IST 659)
  - Data Analytics (IST 707)
  - Scripting for Data Analysis (IST 652)
  - Advanced Database Management (IST 769)

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

THE SEVEN LEARNING OBJECTIVES OF THE PROGRAM

MARK ROBERTS
SUID: 598273961
HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

1. **Describe a broad overview of the major practice areas in data science.**
2. **Collect and organize data.**
3. **Identify patterns in data via visualization, statistical analysis, and data mining.**
4. **Develop alternative strategies based on the data.**
5. **Develop a plan of action to implement the business decisions derived from the analyses.**
6. **Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.**
7. **Synthesize the ethical dimensions of data science practice.**

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 659

Data Administration Concepts & Database Management

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

FINAL PROJECT:

Custom Built SQL Data Base

- Working with Dr. Gregory Block, I created a hypothetical real estate developer data base to manage communities/homes for sale, with relationships to agents, buyers, sellers, culminating in transactions.

- The goal of my project was to sort out some potentially complicated relationships (homes that are part of a community but can be sold individually, entities who could own or buy a house, and agents who could represent buyers or sellers).

- Client communication is paramount in developing business rules which are then modelled and refined into logical/normalized model. Data tables were populated using SQL Server Management Studio. Custom views and stored procedures were created as well as ODBC connection to Microsoft Access and Excel to produce reports which answered data questions, such as inventory of homes the client has available, sales that have occurred and a detail of sale transactions.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 659

Data Administration Concepts & Database Management

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

LEARNING OUTCOME:

- Describe a broad overview of the major practice areas in data science:
  - How data organization and access impacts the needs of actual clients, and not a theoretical stance. A data base must work for its intended user.
- Collect and organize data:
  - Envision working with a client, developing business needs, and creating a plan to organize, create and populate data tables.
- Develop alternative strategies based on the data:
  - Finding an efficient way to deal with a variety of 'people' types (owner/buyer/seller/agents) using connecting tables to identify roles that people can play in the data base.
- Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization:
  - From creation of stakeholders list, data glossary and conceptual and, later, logical/normalized model, it is important for the data practitioner to be able to communicate clearly the direction undertaken. Further, data bases need to produce actionable insights, through aggregations, views and reports. A data base design must have utility as a core principle and communication as a guide to each facet of that utility.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 652

Scripting for
Data Analysis

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

FINAL PROJECT:

Data Analysis of the Squirrel Census of NYC

- Group Project with Katie Haugh and Sandy Spicer, under the guidance of Dr. Deborah Landowski

- Analyze the dataset from an 11-day squirrel census was conducted in Central Park in New York City. 323 volunteers surveyed the 843-acre park for squirrels, assigned shifts in the morning and afternoon, each volunteer was assigned 1 hectare (100-by-100-meter blocks) per shift to conduct their observations for 20-25 minutes. Data on the squirrel observations, a second set on the hectares and external references to weather data were used.

- Our team obtained the data, performed necessary scrubbing operations, exploration/visualization and answered four questions, one on the physical characteristics, one on location density of squirrels, another on weather patterns vis-à-vis sightings and one on squirrel vocalizations/behaviors.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 652

Scripting for
Data Analysis

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

LEARNING OUTCOME:

- Collect and organize data:
  - As part of the project description, two datasets needed to be used. The squirrel set gave us access to data on their behavior and location, providing a challenge of organization.
- Identify patterns in data via visualization, statistical analysis, and data mining:
  - From early experimentation, we discerned a few patterns regarding 1. Places squirrels were sighted, 2. Weather conditions, 3. Physical description of squirrels and 4. Behavior/ Vocalizations. We centered our data analysis around these four insights gained from initial data mining.
- Develop a plan of action to implement the business decisions derived from the analyses:
  - We sought to utilize scripting learned in the course materials to utilize our data to answer our business questions, providing a detailed analysis of our findings.
- Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization:
  - Our project was centered around Business Questions, i.e. what would our ‘Client” like to know about squirrels. Our analysis focused on being able to address each of our four chosen business questions and communicate all the necessary steps to answer them.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 707

Data
Analytics

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

FINAL PROJECT:

Data Analysis of the 2019
Boston Marathon Dataset

- Under the direction of Dr. Gregory Block, I focused my project on the 2019 Boston Marathon Results dataset. I accessed the data from a CSV file posted on Boston Athletic Association (which manages the event) website www.baa.org.

- Data acquisition, cleaning, data mining and other exploratory data analytics procedures (EDA) lead to forming a question I sought to solve for: could we predict which runners were able to match race day performance to qualifying time (a separate reference table).

- Modeled prediction tests using Decision Tree (DT), Support Vector Machine (SVM) and Random Forest (RF). I performed checks on the different models with accuracy and kappa comparisons as well as n-fold cross-validation techniques.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 707

Data
Analytics

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

LEARNING OUTCOME:

- Collect and organize data:
  - Boston Marathon has several built-in categories (age, gender, location), but it's a ranked/ordered dataset. Care had to be taken to randomize the data set lest classification models be skewed.
- Identify patterns in data via visualization, statistical analysis, and data mining:
  - Data mining for additional categories (generation, region) and use of auxiliary/reference data source, framed the dependent categorical value for predictive modeling.
- Develop a plan of action to implement the business decisions derived from the analyses:
  - A robust combination of data acquisition, cleaning, EDA plus domain knowledge, allowed me to construct the plan, develop and evaluate three classification models.
- Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization:
  - Project report features a step-by-step narration of process, steps taken to clean data, insights derived from EDA, and a full evaluation of model performance with suggestions on additional data which could improve models.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 769

Advanced
Database Management

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

COURSEWORK

- 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.

- This course provides tour of relational, document, key-value, columnar, and streaming database systems through the lens of the CAP theorem. We will explore the strengths and weaknesses of various database systems in the relational, Hadoop, and noSQL spaces. Where possible you will experience these systems first-hand as to gain an understanding of how they can be used to address complex, big data challenges.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

IST 769

Advanced
Database Management

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO

LEARNING OUTCOME:

- The lectures (both live and asynchronous) and particularly the lab work provided me the opportunity for me to deepen my understanding of SQL, going into advanced topics not covered in IST659, such as transactions, views, stored procedures, temporal tables and security measures. The coverage of this material reinforced that the data scientist must be a steward of an organization's data, its most precious resource, and become familiar with ways to ensure its security. Digital threats are more the norm now than ever before. Also, organizations need to be concerned with scale and useability. SQL may not always be the solution. Exploration of the CAP theorem, and exploration of ETL/NoSQL programs like Hadoop, MongoDB and others provided an invaluable view into ways to structure data to meet data needs that may not be satisfied by RDBMS.

Syracuse University
School of Information Studies

MASTER OF SCIENCE APPLIED DATA SCIENCE

PORTFOLIO MILESTONE

Thank you.

MARK ROBERTS

SUID: 598273961

HTTPS://GITHUB.COM/DODEKETE/MSADS_PORTFOLIO