

# PROJECT PORTFOLIO

APPLIED DATA SCIENCE

JOVITA ANDREWS

## **Table of Contents**

### **Overview**

#### **1 Introduction**

- 1.1 Purpose of Portfolio
- 1.2 Core and Elective Courses Overview
- 1.3 AI Track Selection

#### **2 Program Structure**

- 2.1 Program Learning Outcomes
  - Outcome 1: Data Collection, Storage and Access
  - Outcome 2: Generating Actionable Insights Across Contexts
  - Outcome 3: Visualization and Predictive Modeling
  - Outcome 4: Programming for Insights
  - Outcome 5: Communication of Insights
  - Outcome 6: Ethics in Data Science

#### **3 Core Course Reflections and Projects**

- 3.1 IST 659: Data Administration Concepts and Database Management
- 3.2 IST 686: Quantitative Reasoning for Data Science
- 3.3 IST 687: Introduction to Data Science
- 3.4 IST 707: Applied Machine Learning
- 3.5 SCM 651: Business Analytics
- 3.6 IST 718: Big Data Analytics

#### **4 Elective Course Reflections and Applications**

- 4.1 IST 615: Cloud Management
- 4.2 IST 692: Responsible AI
- 4.3 MBC 638: Data Analysis and Decision Making

#### **5 AI Track Courses**

- 5.1 IST 664: Natural Language Processing
- 5.2 IST 692: Responsible AI

#### **6 Project Descriptions**

- 6.1 NBA Analytics
- 6.2 Movie Reviews Sentiment Analysis
- 6.3 Text-to-Speech Converter
- 6.4 Taxi Fare Predictor
- 6.5 Crime Classification Based on Geographical Location

#### **7 Overall Synthesis**

- 7.1 Summary of Achievement Across Learning Outcomes
- 7.2 Program Reflections and Professional Growth

#### **8 Conclusion**

- 8.1 Key Takeaways and Future Directions
- 8.2 Final Reflections on the Master's Program Journey

## Overview

---

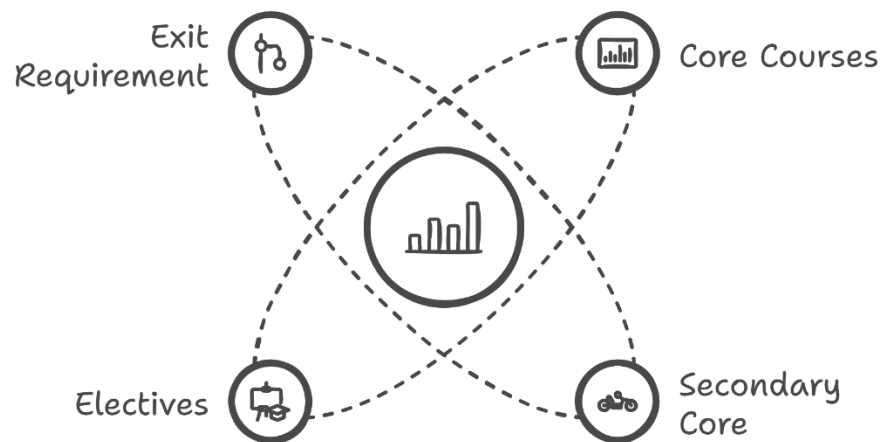
This portfolio reflects my journey from a background in Computer Science and Engineering to a specialized career in Applied Data Science. Before diving into the details, I would like to share why I chose this degree and how iSchool became the launchpad for my data science career aspirations. When considering universities, iSchool stood out as one of the few with a well-defined, structured path tailored to students' career goals. Discovering this degree program felt like finding the perfect match for my professional ambitions.

What particularly drew me to the program was the precision with which each course was designed, organized into specialized tracks like Artificial Intelligence, Big Data Analytics, Business and Data Analytics, Data Pipelines and Platforms, Language Analytics, Project Management, and Visual Analytics. The program's collaboration with the Whitman School of Management offered a unique balance of technical and managerial courses, aligning academic requirements with in-demand job skills. Having fine-tuned my expertise through this degree, I feel well-prepared for entry-level positions in the data science job market—thanks to the comprehensive training provided by the iSchool.

## 1. Introduction

The purpose of this portfolio is to demonstrate my mastery of data science skills and my ability to demonstrate my practical skills into practical – real world scenario applications. Through the coursework completed during my Master of Science in Applied Data Science, I developed competencies in the field of defining a problem, data collection, data cleaning & preprocessing, exploratory data analysis, feature engineering, feature selection, model selection, model training, model evaluation, model validation, model deployment, monitoring and maintenance.

### Overview of Applied Data Science Degree



#### 1.1 Purpose of Portfolio

This portfolio compiles selected projects that illustrate my progress in each program learning outcome, showcasing the knowledge and skills I have gained throughout the program.

#### 1.2 Core and Elective Courses

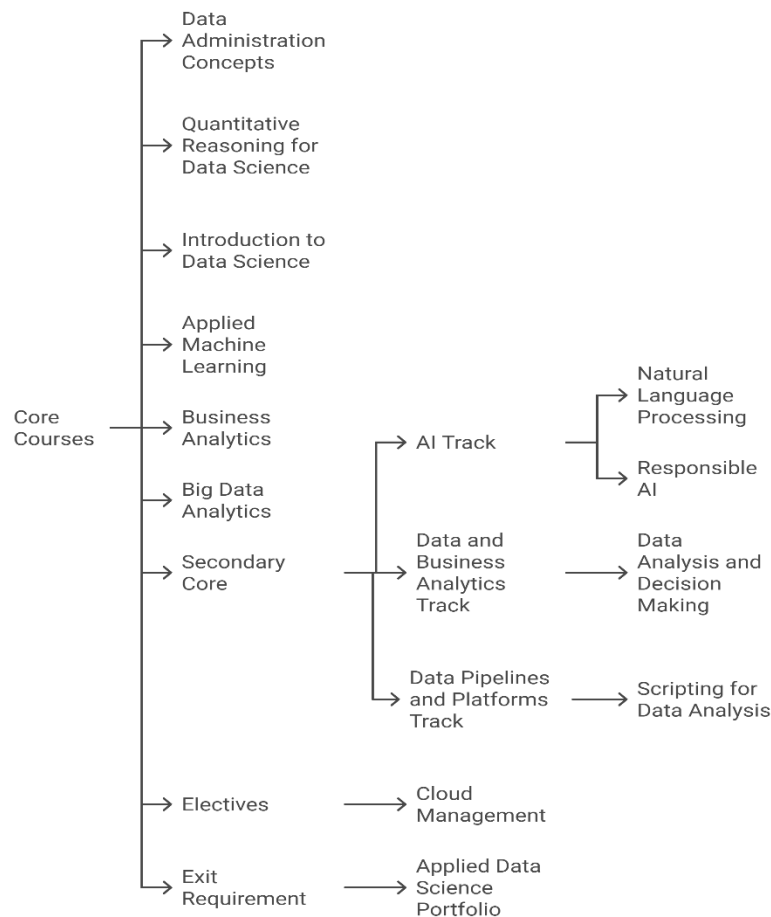
This portfolio includes projects and reflections from core courses, electives, and my AI track, which together contributed to a comprehensive skill set in data science. Core courses, such as **IST 707: Applied Machine Learning** and **IST 718: Big Data Analytics**, provided foundational knowledge in predictive modeling and data handling. Electives like **IST 615: Cloud Management** allowed me to explore cloud-based data solutions, while **MBC 638: Data Analysis and Decision Making** supported my ability to make data-driven business decisions. These courses, combined with projects from the AI track, reflect a balanced approach to both technical and practical aspects of data science.

### 1.3 AI Track

I chose the AI track to deepen my expertise in machine learning and natural language processing (NLP), which are highly relevant to my career interests. Courses like **IST 664: Natural Language Processing** taught me the principles of linguistic analysis and the application of machine learning models for sentiment analysis, which I applied in the **Movie Reviews Sentiment Analysis** project. Additionally, **IST 692: Responsible AI** provided insights into ethical practices in AI, which guided my approach to handling biases and fairness in data models. This focus on AI aligns with my goal of leveraging data science for advanced analytics in fields such as customer experience and behavioral analysis.

## 2. Program Learning Outcomes

### List of Courses Taken:



### Primary Core (15 credits)

IST 659 Data Administration Concepts and Database Management

IST 686 Quantitative Reasoning for Data Science

IST 687 Introduction to Data Science

IST 707 Applied Machine Learning

SCM 651 Business Analytics

IST 718 Big Data Analytics

### Secondary Core

IST 664 Natural Language Processing

IST 692 Responsible AI

MBC 638 Data Analysis and Decision Making

IST 652 Scripting for Data Analysis

IST 615 Cloud Management

### Exit Requirements

IST 782 Applied Data Science Portfolio

## 2.1 Program Learning Outcomes

- **Outcome 1:** *“Collect, store, and access data by identifying and leveraging applicable technologies”*

In **IST 659: Data Administration Concepts and Database Management**, I learned SQL and database design principles, which provided a strong foundation in data organization and management. This was further reinforced in **IST 615: Cloud Management**, where I applied these concepts in a cloud environment. For the **Text-to-Speech Converter** project, I used AWS Lambda and AWS Polly for efficient data processing and storage, showcasing my skills in handling and accessing data within a cloud-based infrastructure.

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

In **SCM 651: Business Analytics**, I developed skills to identify patterns in business data and draw insights to guide strategic decisions. Similarly, **IST 686: Quantitative Reasoning for Data Science** emphasized statistical reasoning, which I applied in the **NBA Analytics** project. Here, I used statistical analysis to determine player performance metrics, generating insights relevant to team management and strategy. In **MBC 638: Data Analysis and Decision Making**, I further developed my ability to interpret data for real-world applications, enhancing my ability to make informed decisions based on data.

- **Outcome 3:** *“Apply visualization and predictive models to help generate actionable insight”*

In **IST 707: Applied Machine Learning**, I gained hands-on experience with predictive modeling, using techniques like decision trees and clustering. For the **Crime**

**Classification** project, I used KMeans clustering to identify crime patterns in Los Angeles, which helped in visually representing high-risk areas for public safety. Additionally, **IST 718: Big Data Analytics** focused on visualizing large datasets, which I applied in the **Taxi Fare Predictor** project, where I used Spark and PySpark to handle and visualize trip and weather data for fare prediction.

- **Outcome 4:** “Use programming languages such as R and Python to support the generation of actionable insight”

*In **IST 664: Natural Language Processing**, I utilized Python libraries such as NLTK and TensorFlow for text processing and sentiment analysis. For the **Movie Reviews Sentiment Analysis** project, I implemented models like Naive Bayes and BERT, which helped in accurately classifying movie review sentiments. In **IST 652: Scripting for Data Analysis**, I used Python to analyze NBA statistics, further honing my programming skills to generate insights for team management decisions.*

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

**IST 687: Introduction to Data Science** introduced me to various techniques for communicating data insights, which I applied in several projects. For example, in the **Taxi Fare Predictor** project, I presented fare prediction results to stakeholders in a clear and understandable format, making complex data accessible to non-technical users. **SCM 651: Business Analytics** further enhanced my communication skills, focusing on data storytelling for business insights.

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

In **IST 692: Responsible AI**, I explored ethical issues in AI and data science, such as bias mitigation and transparency. These principles guided my approach to the **Movie Reviews Sentiment Analysis** project, where I considered bias in training data and model transparency to improve fairness in sentiment classification. Ethical considerations were also incorporated in the **Crime Classification** project, where I ensured sensitive data was handled responsibly to protect privacy.

### 3. Core Course Reflections and Projects

- **IST 659 - Data Administration Concepts and Database Management:** Reflect on database management skills, SQL programming, and how this course supported data storage and retrieval (Learning Outcome 1).
- **IST 686 - Quantitative Reasoning for Data Science:** Highlight statistical analysis, R programming, and insights generated for quantitative reasoning (Learning Outcomes 2, 4).
- **IST 687 - Introduction to Data Science:** Discuss introductory data science techniques, data wrangling, and communication of insights (Learning Outcomes 2, 5).

- **IST 707 - Applied Machine Learning:** Detail machine learning model-building and optimization skills, particularly in real-world applications (Learning Outcomes 3, 4).
- **SCM 651 - Business Analytics:** Emphasize business decision-making skills developed through visualization and data mining (Learning Outcomes 2, 3, 5).
- **IST 718 - Big Data Analytics:** Discuss big data management, use of Apache Spark, and handling of large data volumes (Learning Outcomes 1, 3).

#### 4. Elective Course Reflections and Applications

- **IST 615 - Cloud Management:** Reflect on cloud services, virtualization, and hybrid cloud management skills (Learning Outcome 1).
- **IST 692 - Responsible AI:** Highlight ethical considerations in AI, aligning with program learning outcomes on ethics and responsible data use (Learning Outcome 6).
- **MBC 638 - Data Analysis and Decision Making:** Emphasize skills gained in data analysis to support business decision-making (Learning Outcome 2).

#### 5. AI Track Courses

- **IST 664 - Natural Language Processing:** Explain the course's focus on text data processing and application of Python for NLP tasks (Learning Outcomes 2, 4).
- **IST 692 - Responsible AI:** Reinforce how ethical AI practices align with data science responsibilities (Learning Outcome 6).

#### 6. Project Description

##### 6.1 NBA Analytics

- **Course:** IST 652 – Scripting for Data Analysis
- **Project Goals:** The primary goal of this project is to analyze NBA player and team data to predict award winners, including MVP, championship teams, and Defensive Player of the Year, based on performance metrics such as points per game, rebounds, and efficiency ratings.
- **Technologies:** Python, 'nba\_api' for data collection, Pandas for data manipulation, and visualization libraries (Matplotlib, Seaborn).
- **Actionable Insight Generated:** The project identified key performance factors influencing player impact and team success, providing insights valuable to NBA team managers. For example, statistical patterns in player efficiency highlighted attributes that contribute significantly to team wins.
- **Individual Contribution:** I was responsible for data extraction, statistical analysis, and developing predictive models using regression techniques. Additionally, I created visualizations to communicate insights on player performance and team dynamics.
- **Program Learning Outcomes Demonstrated:**



- **Outcome 1:** Data collection, storage, and accessibility through API integration and data organization.
- **Outcome 2:** Generating actionable insights by identifying trends in player and team success metrics.
- **Outcome 3:** Visualization and Predictive Modeling to Project Potential award winners
- **Outcome 4:** Programming skills in Python for data manipulation and analysis
- **Outcome 5:** Communicating findings through clear and informative visualizations

## 6.2 Movie Reviews Sentiment Analysis

- **Course:** IST 664 - Natural Language Processing
- **Project Goals:** This project aimed to classify movie reviews based on sentiment (very negative to very positive) to understand customer sentiment and support applications like recommendation systems and customer feedback analysis.
- **Technologies:** Python, NLP libraries (NLTK, TensorFlow), models including Naive Bayes, Decision Tree, and BERT for deep learning, with TF-IDF for feature extraction.
- **Actionable Insight Generated:** The BERT model achieved the highest accuracy, demonstrating its effectiveness in capturing nuanced sentiment distinctions. This insight supports the use of BERT in scenarios that require detailed sentiment analysis, such as customer feedback systems.
- **Individual Contribution:** I implemented the BERT model, processed and visualized sentiment data, and evaluated model performance using metrics like accuracy, F1 score, and precision.
- **Program Learning Outcomes Demonstrated:**
  - **Outcome 2:** Generating insights across contexts by identifying patterns in customer sentiment.
  - **Outcome 3:** Visualization and modeling with various machine learning techniques for classification tasks.
  - **Outcome 4:** Programming in Python for NLP and deep learning.
  - **Outcome 6:** Ethical considerations in handling biases in sentiment data and ensuring fairness.

## 6.3 Text-to-Speech Converter

- **Course:** IST 615 – Cloud Management
- **Project Goals:** The objective was to create a cloud based text-to-speech converter using AWS services, providing an accessible audio solution for diverse audiences, including those with visual impairments.
- **Technologies Used:** AWS Lambda, AWS Polly for text-to-speech conversion, and optionally AWS S3 for data storage.
- **Actionable Insight Generated:** This project demonstrated the advantages of serverless architecture in scaling and cost efficiency, making it a feasible approach for real time applications like text-to-speech conversion.

- **Individual Contribution:** I designed the serverless architecture, implemented Lambda functions, and configured the API Gateway for seamless user access. Additionally, I optimized IAM for secure access and data management.
- **Program Learning Outcomes Demonstrated:**
  - **Outcome 1:** Data collection, storage, and accessibility in a cloud-based environment.
  - **Outcome 2:** Generating insights on the benefits of serverless architecture for accessibility solutions.
  - **Outcome 4:** Programming and automation within AWS for data processing and storage.
  - **Outcome 5:** Communicating insights by ensuring accessibility for diverse user groups.

#### 6.4 Taxi Fare Predictor

- **Course:** IST 718 - Big Data Analytics
- **Project Goals:** The goal was to predict NYC taxi fares by analyzing historical trip and weather data to identify fare-influencing factors and provide insights for resource optimization in urban transportation.
- **Technologies Used:** PySpark, Spark SQL for data processing, Google BigQuery for data retrieval, and machine learning models (Linear Regression, Random Forest, Decision Tree).
- **Actionable Insight Generated:** The model revealed that fare amounts were significantly influenced by factors like pickup/drop-off zones and weather conditions, providing actionable insights for pricing adjustments and operational improvements.
- **Individual Contribution:** I was responsible for data cleaning, feature selection, and training regression models, as well as evaluating model performance with RMSE.
- **Program Learning Outcomes Demonstrated:**
  - **Outcome 1:** Managing and processing large datasets with Spark.
  - **Outcome 2:** Generating insights into fare dynamics in urban transportation.
  - **Outcome 3:** Visualization and predictive modeling for fare prediction.
  - **Outcome 4:** Scripting in Python and PySpark to handle data at scale.
  - **Outcome 5:** Communicating model insights effectively to non-technical stakeholders.

#### 6.5 Crime Classification Based on Geographical Location

- **Course:** IST 707 - Applied Machine Learning
- **Project Goals:** This project aimed to classify and visualize crime types across Los Angeles neighborhoods to support public safety agencies and inform residents about localized crime trends.
- **Technologies Used:** Python, KMeans clustering, Folium for geospatial visualization, and data preprocessing libraries.

- **Actionable Insight Generated:** The analysis revealed crime hotspots and helped identify crime patterns by location, allowing agencies to focus resources on high-risk areas and provide residents with information about neighborhood safety.
- **Individual Contribution:** I was responsible for data preprocessing, clustering, and creating geospatial visualizations to map crime patterns, supporting insights for crime prevention.
- **Program Learning Outcomes Demonstrated:**
  - **Outcome 1:** Data handling and organization for geospatial analysis
  - **Outcome 2:** Generating actionable insights for public safety and resource allocation.
  - **Outcome 3:** Visualization of crime clusters to highlight high-risk areas.
  - **Outcome 4:** Programming in Python for clustering and data visualization.
  - **Outcome 5:** Ethical considerations in handling and presenting sensitive crime data.

## 7. Overall Synthesis

### 7.1 Summary of Achievement:

The Master's in Applied Data Science program has equipped me with a robust skill set that aligns with the program's learning outcomes. **Core courses** like **IST 707 (Applied Machine Learning)** and **IST 687 (Introduction to Data Science)** provided foundational skills in data wrangling, machine learning, and data communication, while projects like **Crime Classification** reinforced predictive modeling and ethical considerations. **Electives**, including **IST 615 (Cloud Management)** and **MBC 638 (Data Analysis and Decision Making)**, expanded my expertise in cloud computing and decision frameworks, demonstrated through projects like **Text-to-Speech Converter**. The **AI track** deepened my knowledge in machine learning and ethics, where projects like **Movie Reviews Sentiment Analysis** helped me apply NLP techniques and address AI biases collectively contributed to achieving program learning outcomes.

### 7.2 Program Reflections

The program has seamlessly blended theory and practical applications, preparing me for real-world data science challenges. Key experiences included managing large datasets in **Taxi Fare Predictor** using PySpark and refining NLP models in **Movie Reviews Sentiment Analysis** through extensive tuning. These challenges strengthened my skills in data processing, predictive modeling, and cloud-based solutions, equipping me for versatile, data-driven roles.

## **8. Conclusion**

### **8.1 Key Takeaways and Future Directions**

- I gained a comprehensive understanding of the data science life cycle, technical tools (eg: AWS, PySpark, TensorFlow), and ethical data science. These insights were reinforced through projects like NBA Analytics and Movie Sentiment Analysis.
- I aim to expand my knowledge in deep learning, particularly in computer vision and advanced NLP, with potential applications in healthcare and social media. Additionally, I plan to further develop cloud- native data science skills for scalable, distributed systems.

### **8.3 Final Reflections on the Master's Program Journey**

My journey through this program has been transformative, combining in-depth coursework with hands-on projects across diverse data science domains. I leave with a deep appreciation for the role of data and its ethical implications, well-prepared to make meaningful impact in the field.