MS in Applied Data Science Portfolio Reflection

David Caspers
Syracuse University
School of Information Studies

25 March 2025

Overview of Learning Journey

Specializations:

AI & Deep Learning

Language Analytics

Key Learning Goals:

Data Collection & Storage

Data Analysis & Model Development

Predictive Modeling & Visualization

Programming & Data Science Tools

Communication & Decision-Making

AI Ethics & Responsible Modeling

Project Overview

Criteria:

Technical Capability

Real-World Impact

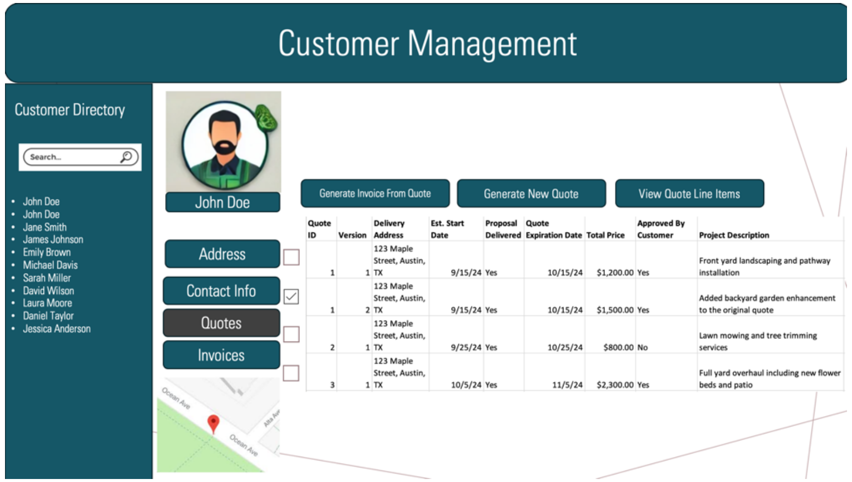
Continuous Learning

Selected Projects

- Database Model for Quote Management System
- Combating Illegal, Unreported, and Unregulated (IUU) Fishing
- Detecting Lung Cancer from Histopathological Images
- Inferring Politician Ideology from Public Statements

Data Collection & Storage Learning Objective

- **Goal:** Developed a MySQL-based system for invoicing and job estimation
- **Technologies Used:** MySQL, SQL, ERD Modeling, Stored Procedures.
- **Key Processes:**
  - Designed relational schema for structured data storage.
  - Implemented automated invoice generation & payment tracking.
  - Optimized queries for fast financial reporting.
- **Learning Outcomes:**
  - Structured method to capture business requirements and translate to technical implementation
  - Hands-on SQL experience
  - Improved data-driven decision-making through reporting tools.



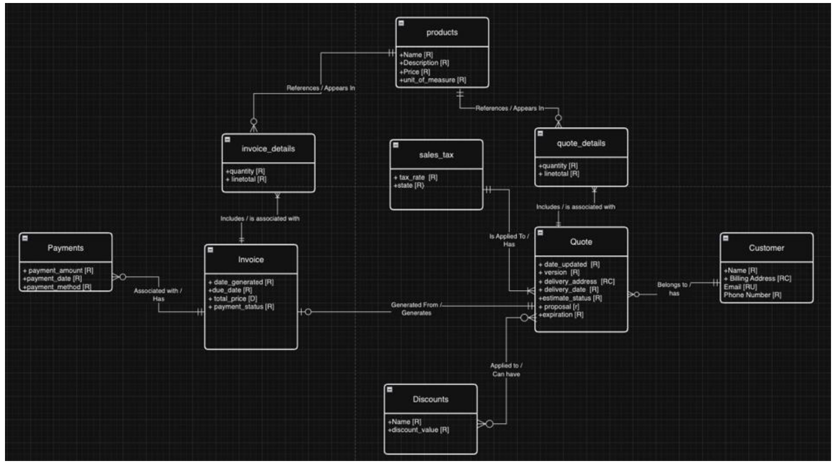
Customer Management

Customer Directory

- John Doe
- John Doe
- Jane Smith
- James Johnson
- Emily Brown
- Michael Davis
- Sarah Miller
- David Wilson
- Laura Moore
- Daniel Taylor
- Jessica Anderson

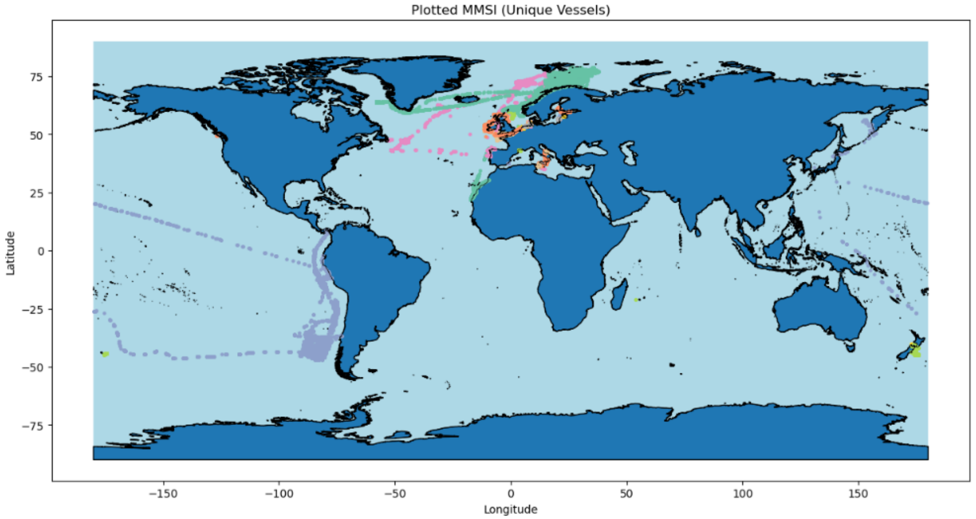
John Doe

	Quote ID	Version	Delivery Address	Est. Start Date	Proposal Delivered	Quote Expiration Date	Total Price	Approved By Customer	Project Description
☐	1	1	123 Maple Street, Austin, TX	9/15/24	Yes	10/15/24	$1,200.00	Yes	Front yard landscaping and pathway installation
☑	1	2	123 Maple Street, Austin, TX	9/15/24	Yes	10/15/24	$1,500.00	Yes	Added backyard garden enhancement to the original quote
☐	2	1	123 Maple Street, Austin, TX	9/25/24	Yes	10/25/24	$800.00	No	Lawn mowing and tree trimming services
☐	3	1	123 Maple Street, Austin, TX	10/5/24	Yes	11/5/24	$2,300.00	Yes	Full yard overhaul including new flower beds and patio



Combating Illegal, Unreported, and Unregulated (IUU) Fishing

- **Goal:** Develop a machine learning-based system to detect illegal fishing activity using AIS geospatial data.
- **Technologies Used:** Python, Scikit-learn (Random Forest, SVM, DBSCAN).
- **Key Processes:**
  - Cleaned AIS data (removed errors, handled missing values)
  - Applied clustering (DBSCAN, K-Means) and classification (Random Forest, SVM, Naïve Bayes)
  - Used time-series validation to prevent data leakage
- **Learning Outcomes:**
  - Hands-on experience with real-world noisy datasets
  - Strengthened ML & geospatial analysis skills
  - Applied feature engineering for time-series data

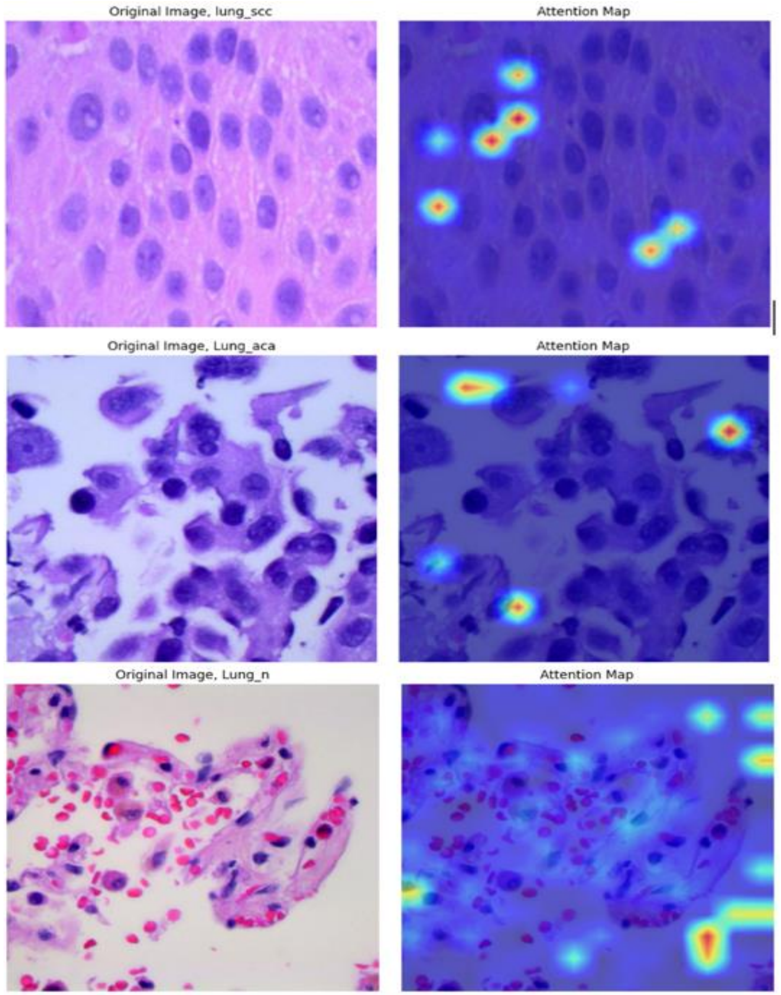


Top Performing Model (Random Forest)

- **Accuracy**: 87.69%
- **Precision**:
  - Non-Fishing: 0.88
  - Fishing: 0.87
- **Recall**:
  - Non-Fishing: 0.91
  - Fishing: 0.80
- **Top 5 Features**:
  - **Lag Speed**: 31.00%
  - **Speed**: 21.54%
  - **Lag Distance from Shore**: 16.94%
  - **Distance from Shore**: 8.80%
  - **Latitude**: 6.65%

Detecting Lung Cancer from Histopathological Images

- **Goal:** Develop a deep learning model to classify lung tissue into adenocarcinoma, squamous cell carcinoma, or normal tissue using histopathological images.
- **Technologies Used:** TensorFlow, Keras (CNN, Vision Transformer), transfer learning, hugging face pretrained models
- **Key Processes:**
  - Preprocessed 15,000+ images (normalized, resized, augmented for training).
  - Trained CNN & Vision Transformer models to compare accuracy and interpretability.
  - Used attention heatmaps to highlight critical regions for model explainability.
- **Learning Outcomes:**
  - Advanced deep learning model application with focus on model interpretability for clinical adoption
  - Gained experience in handling memory intensive datasets efficiently.



Visualized Attention Maps

Inferring Politicians' Political Ideology

- **Goal:** Use NLP to determine a politician's ideological stance. Rank politicians by analyzing opinion-based rhetoric.
- **Technologies Used:** TensorFlow, Keras (CNN, Vision Transformer), transfer learning, hugging face pretrained models
- **Key Processes:**
  - Scraped **37,000+ statements** from VoteSmart API.
  - Fine-tuned **DistilBERT** for opinion classification.
  - Mapped embeddings onto an **ideological spectrum** for ranking.

Learning Outcomes:

- Strengthened in **data acquisition & handling large-scale text data**.
- Applied **ranking & predictive modeling** beyond standard classification.

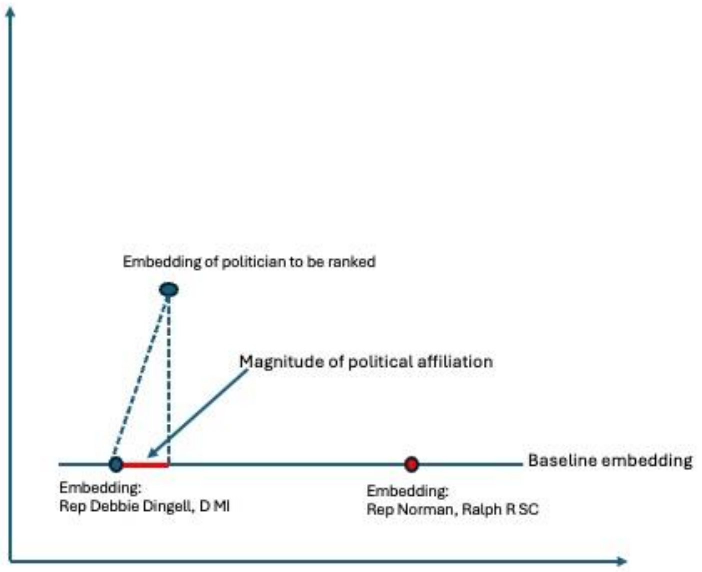


Illustration of Ranking Methodology

Cancer Types:	Transformer Model Performance Statistics (Top Performing Model)			
	Precision	Recall	F-1	Support
ACA	0.91	0.99	0.95	1002
Normal	1.00	1.00	1.00	992
SCC	0.99	0.91	0.95	1006

Reflection on Growth & Future Development

- **Program Impact:** Developed a strong foundation in applied data science.

- **Remaining Areas for Growth:**
  - Cloud ML deployment & MLOps
  - Scaling big data solutions
  - AI Ethics & Explainability

- **Next Steps:**
  - Ongoing Learning through courses, certifications, and research.
  - Practical Experience applying skills in real-world projects.

Thank you!

caspersdavid@gmail.com / dcaspers@syr.edu