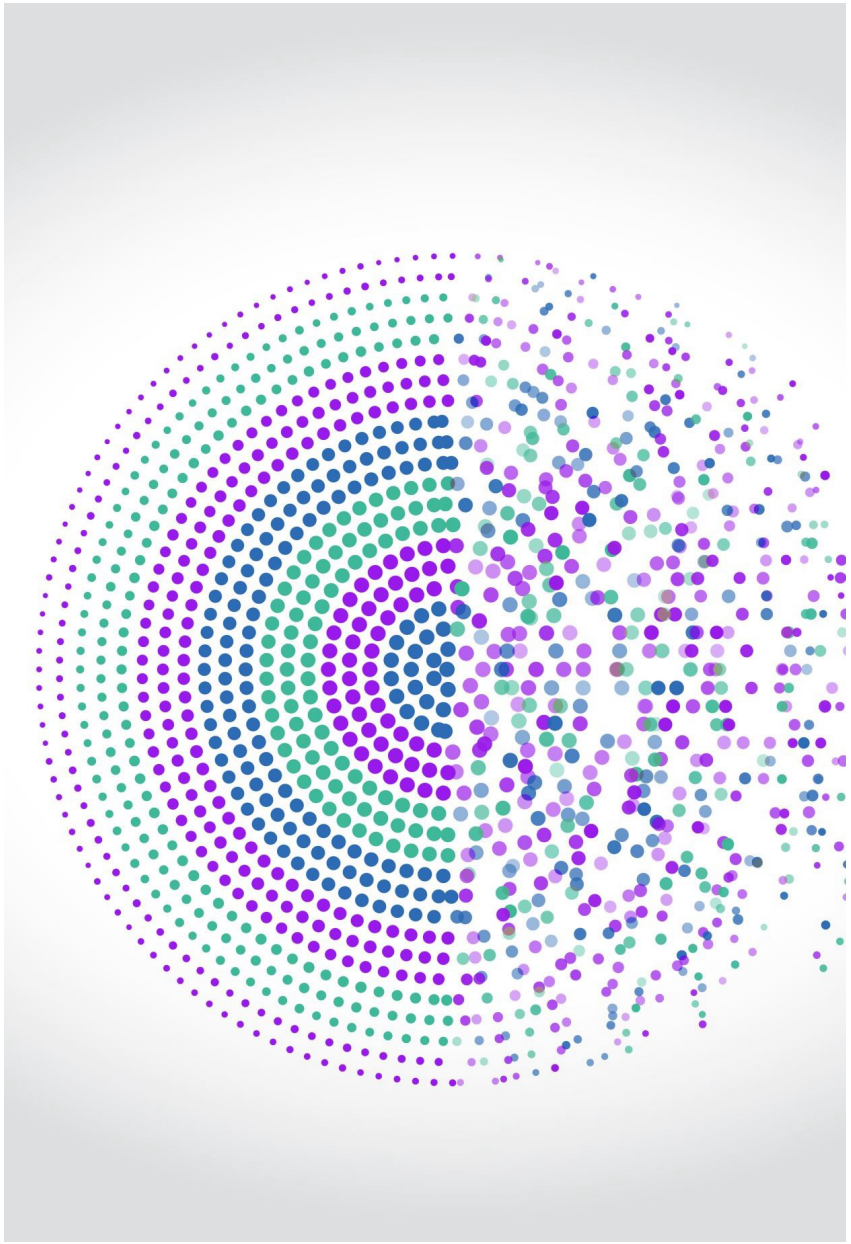
Matthew L. Pergolski

Applied Data Science Capstone Portfolio

Syracuse University, Master of Science in Applied Data Science

Agenda

- Introduction
- Personal Interest in Data Science
- MSADS Program
- Career Goals
- Program Objectives / Outcomes
- Data Science Life Cycle
- Coursework / Curriculum
- Projects
- Conclusion



Introduction

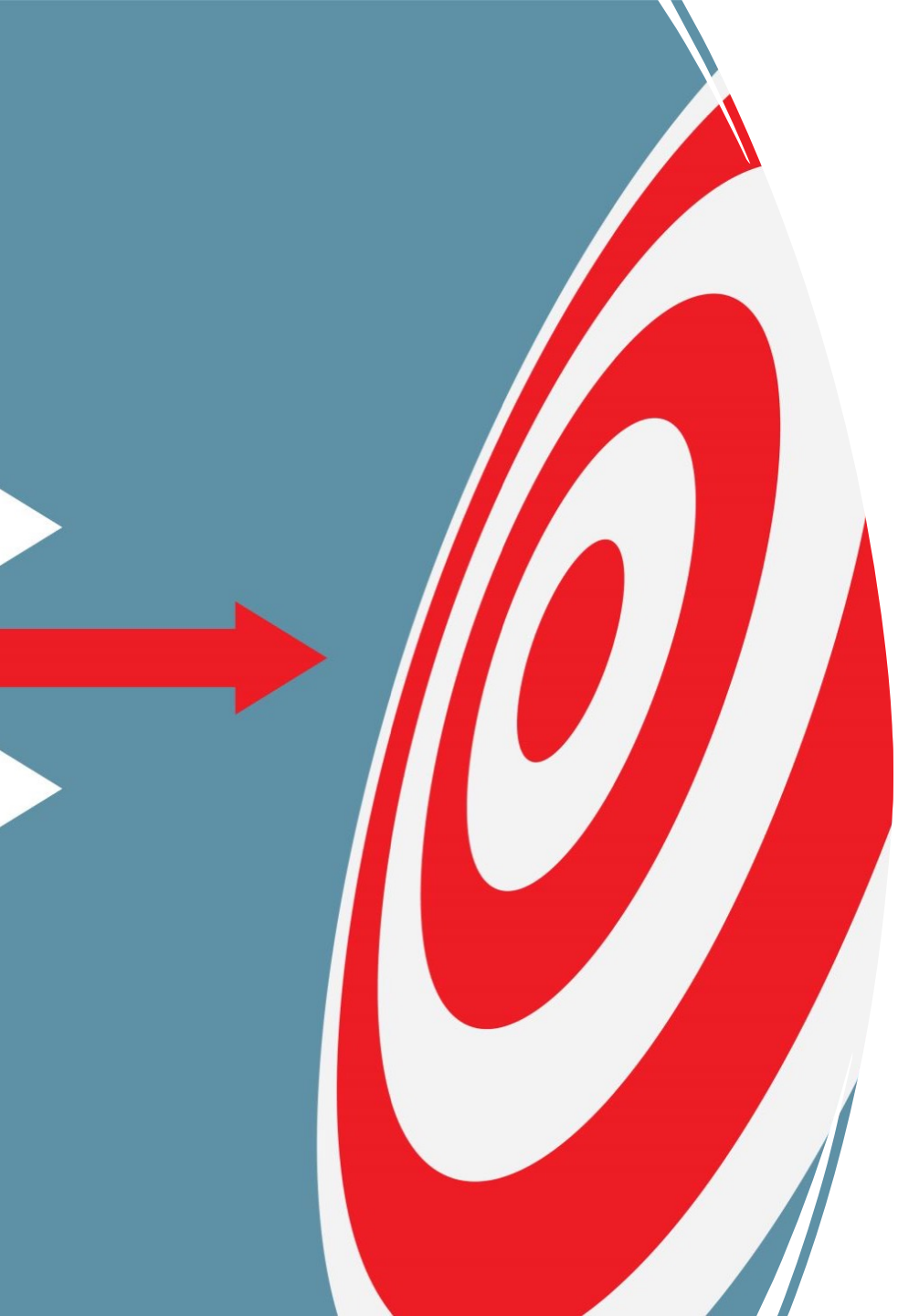
- Background
  - Undergraduate degree: University of Wisconsin-Eau Claire
  - Industry: Aerospace
- Transition to data science driven by a desire for more advanced automation and problem-solving

Personal Interest in Data Science

- Initial interest in Excel and VBA, but quickly recognized limitations
- Discovery of Python and its potential for automation
- COVID-19 pandemic provided an opportunity to explore data science

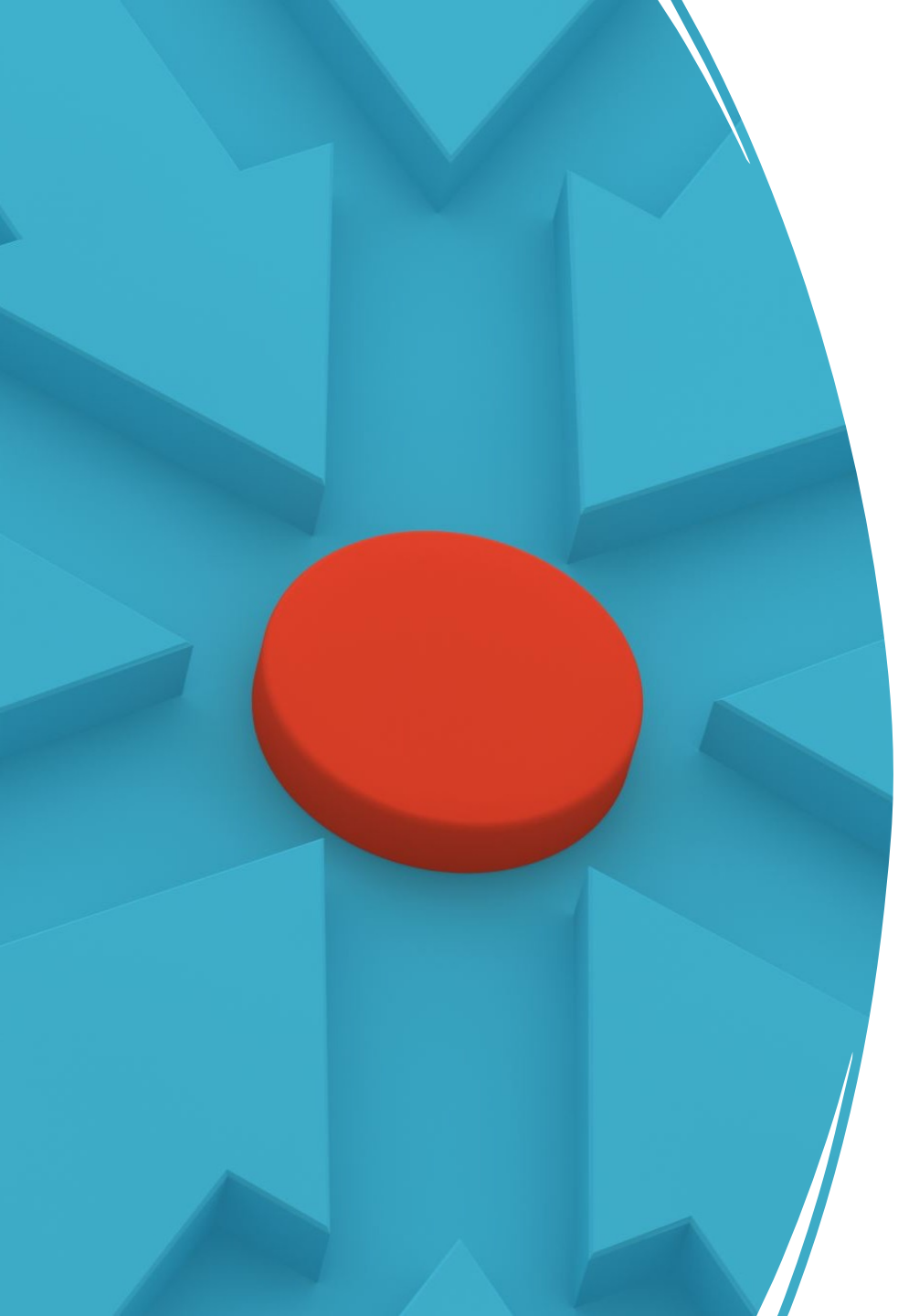
Master of Science in Applied Data Science Program

- Emphasis on practical, real-world applications of data science
- Available tuition assistance from employer
- Preparedness for the program
  - Undergraduate courses
  - Programming courses via Coursera, Data Camp, etc.



Career Goals

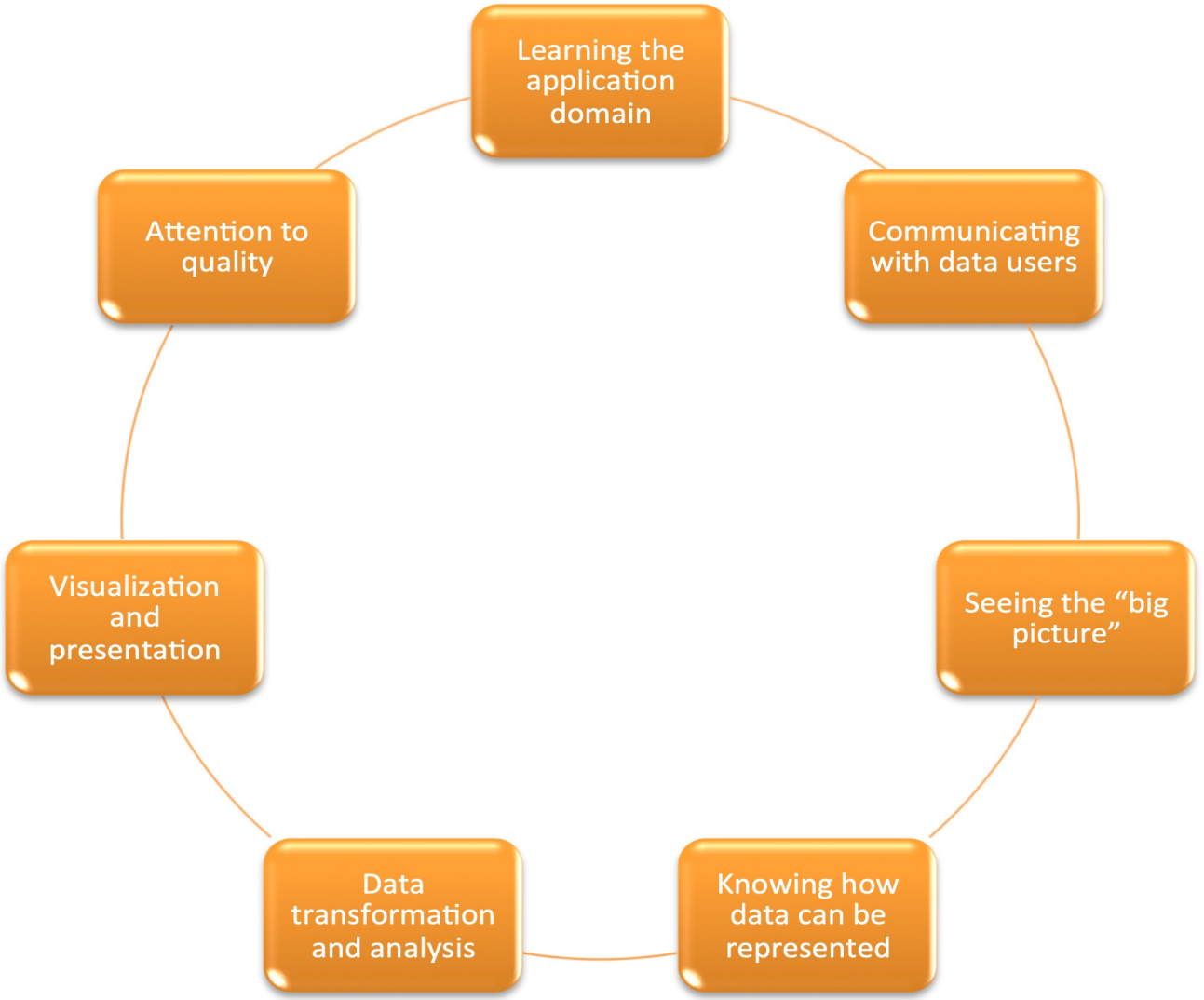
- Desire to master fundamental and advanced data science topics
- Commitment to lifelong learning through online platforms and personal projects
- Current position as a Data Scientist (Level III) obtained during the program



Program Objectives and Outcomes

- Collecting, storing, and accessing data
- Creating actionable insights in various contexts (societal, business, political)
- Applying visualization and predictive models
- Using programming languages like R and Python
- Communicating insights to diverse audiences
- Applying ethical considerations in data science practices

Data Science Life Cycle



Coursework and Curriculum

- IST 687 | Introduction to Data Science
- IST 772 | Quantitative Reasoning in Data Science
- IST 707 | Applied Machine Learning
- IST 659 | Data Administration Concepts and Database Management
- IST 652 | Scripting for Data Analysis
- IST 769 | Advanced Big Data Management
- IST 664 | Natural Language Processing
- SCM 651 | Business Analytics
- IST 718 | Big Data Analytics
- IST 736 | Text Mining
- IST 691 | Deep Learning In Practice

IST 718: Big Data Analytics

Developed an image classification model to distinguish between normal and pneumonia-affected chest X-rays.

- Dataset consisted of 5,863 chest X-rays, each categorized as either normal or pneumonia

The project utilized two deep learning models:

- ResNet: This model architecture is known for its ability to effectively learn features from images. ResNet models with varying layers (18, 50, and 152) achieved accuracies between 81% and 85%
- VGG-16: This model is another widely used deep learning architecture for image classification. Fine-tuning and increasing the number of epochs for the VGG-16 model resulted in a training accuracy of 97.9% and validation accuracies similar to ResNet

IST 736: Text Mining

Developed a text mining application to help prospective students explore course syllabi for the Master of Science in Applied Data Science program at Syracuse University.

The application leverages several text mining techniques:

- Term frequency analysis: This technique allows the application to identify the most frequent terms in each syllabus, providing insights into the key concepts covered in each course
- Document similarity search: This technique enables the application to compare syllabi based on their content, allowing prospective students to find courses that are similar to those they are interested in
- Topic modeling: This technique allows the application to uncover the underlying themes and topics covered across the entire curriculum

Prospective students can use this information to understand the breadth and depth of the program's offerings.

IST 691: Deep Learning in Practice

This project involved creating an interactive application that classifies bird species based on audio recordings. The application uses artificial neural networks and evaluates different training strategies to optimize performance:

- Learning rate optimization: This technique involves finding the optimal learning rate for the neural network, which can significantly impact its performance
- Early stopping: This technique helps prevent the model from overfitting the training data by stopping the training process when the model's performance on a validation set starts to decrease
- Stratified sampling: This technique ensures that the training and validation sets have a balanced representation of each bird species, improving the model's ability to generalize to unseen data


Business Strategy
Domain Knowledge
Statistics
Pattern Recognition
Neurocomputing
Visualisations
Business Analyse
Communicatio
Machine Learning
AI
Data Science
Databases & Data Processing
KDD
Problem Solving
Data Mining
Presentation
Inquisitiveness

Conclusion