Institute for Cyber-Enabled Research and IT Services,
Michigan State University

MSU Cloud Computing Fellows 4th Annual Symposium
05 May, 2023

ICER Hosts

Pat Bills and Mahmoud Parvizi

Schedule

- **10:00am – Welcome and Opening Remarks**
  Brian O'Shea, Director of ICER
- **10:05am** Assessing partisan niche market strategies of media outlets using machine learning techniques
  Suhwoo Ahn, Ph.D. Candidate, Department of Communication
- **10:20am** Using the cloud for on-demand scoring of teacher observations
  Lydia Bradford, Ph.D., Measurement and Quantitative Methods
- **10:35am** A cloud-based web application for generalized virtual metrology of semiconductor manufacturing
  Ritam Guha, Ph.D., Department of Computer Science and Engineering
- **10:50am** Leveraging Azure Cognitive Services for communication research
  Sue Lim, Ph.D. Candidate, Department of Communication
- **11:05am – Break**
- **11:20am** Migrating agricultural data management to a cloud-hosted relational database
  Sarah Manski, Ph.D. Candidate, Department of Statistics and Probability
- **11:35am** Making snakes in the cloud to study diversity
  Miles Roberts, Ph.D. Candidate, Genetics and Genome Sciences
- **11:50am** Optimizing machine learning web application runtime using Azure Web Application Services
  John Salako, MSc., Department of Earth and Environmental Sciences
- **12:05pm** Cloud computing tool for lidar processing
  Meicheng Shen, Ph.D. Candidate, Department of Geography, Environment, and Spatial Sciences
- **12:20pm – Break**
- **12:35pm** Machine learning with unstructured data for social scientists
  Alfred Torsu, MSc., Department of Department of Political Science
- **12:50pm** Scalable Earthquake Catalog Generation: Harnessing cloud computing with Kubeflow on Azure
  Ziyi Xi, Ph.D. Candidate, Department of Computational Mathematics, Science and Engineering
- **1:05pm** Single cell RNA sequencing using the power of cloud
  Arash Yunesi, Ph.D., Department of Computational Mathematics, Science and Engineering
- **1:20pm – Closing Remarks**
  Pat Bills and Mahmoud Parvizi
- **1:30pm – Break**
- **1:45pm – Group Photos**
- **2:00pm – Social Hour**

Project Summaries

Assessing partisan niche market strategies of media outlets using machine learning techniques
Suhwoo Ahn, Ph.D. Candidate, Department of Communication

This study investigates how media outlets take their partisan positions in making political news. I collected 598,998 news articles from 32 media outlets in South Korea. I will analyze them using a supervised machine learning technique based on a cloud computing service. I posit that latecomer news outlets might show stronger partisanship than established ones. I will conclude by highlighting implications for theoretical grounds and methodological advances

Using the cloud for on-demand scoring of teacher observations
Lydia Bradford, Ph.D., Measurement and Quantitative Methods

Building upon previous research (Bradford, 2022) that trained and compared different machine learning methods to score teacher observations from text to a score of 1-4, this cloud computing project creates the building blocks for an on-demand scoring system for observers using the same observation protocol. The trained machine learning model along with the necessary data preprocessing were registered through azure machine learning. A container was established and is shareable for anyone needing to access the machine learning model to score their observations. The container and access to the machine learning model has been successfully tested and is usable for the researchers currently using the PBL observation protocol.

A cloud-based web application for generalized virtual metrology of semiconductor manufacturing
Ritam Guha, Ph.D., Department of Computer Science and Engineering

Due to the ongoing global chip shortage, semiconductor manufacturing has gained massive importance in the last few years. The entire semiconductor industry is trying to scale its manufacturing capacity to cater to the needs of the global population. But currently, it suffers from huge wastage in terms of energy and resources because of the lack of recipe optimization. Semiconductor manufacturing processes are typically very long and span around 3-4 days. At the end of the process, if the product does not meet certain quality requirements, the yields are wasted and the process needs to be restarted which leads to huge wastage. So, Virtual Metrology (VM) has gained tremendous popularity as a supporting tool to optimize energy and resource utilization, thereby improving the efficiency of the manufacturing pipeline. VM refers to the automated estimation of the manufacturing properties using the data sensed from the process without physical metrology operations. We have developed a VM pipeline that takes the sensor data from one-fifth of the process run and virtually simulates the rest of the process and predicts the quality of the yields at the end. So, if the predictions do not lead to the expected properties for the yields, the process can be stopped early and restarted to save energy. As a part of the cloud computing fellowship project, I have designed and trying to deploy a web application linked to our pipeline. The application should accept the first few hours of data from the manufacturing run and estimate the final quality of the yields.

Leveraging Azure Cognitive Services for communication research
Sue Lim, Ph.D. Candidate, Department of Communication

With the deployment of ChatGPT in 2022, a significant number of communication researchers have become interested in how to leverage AI for communication research. Azure cognitive services provide opportunities for communication researchers without much programming experience to use AI in their research. In this presentation, I summarize my experience with three different types of cognitive services: computer vision (spatial analysis), Azure's bot service, and OpenAI services. Due to privacy and other constraints with the MSU Azure account, I could not continue the first two attempts . However, the OpenAI service showed promise. The second half of the presentation illustrates potential use cases of OpenAI services for communication researchers.

Migrating agricultural data management to a cloud-hosted relational database
Sarah Manski, Ph.D. Candidate, Department of Statistics and Probability

In developing the largest-scale agricultural modeling effort of its kind encompassing field-level data for nearly one million fields over almost two decades with over 300 associated variables, our data storage structure has outgrown feasibility. For our researchers, using the associated project input and output data is time-consuming, memory intensive, and nearly infeasible for local machines. This project aims to create a comprehensive relational cloud database for all project data to create a reliable, scalable, and efficient storage and query environment.

Making snakes in the cloud to study diversity
Miles Roberts, Ph.D. Candidate, Genetics and Genome Sciences

Biologists are fundamentally interested in studying the diversity of life forms on Earth, which can tell us a lot about how different forms evolved and even how they may change into the future. However, current methods for measuring diversity at the genetic level require synthesizing the inputs and outputs of many pieces of sophisticated software. For my project, I created a workflow to turn raw genetic data into diversity measurements with a single command and ran this workflow in the cloud. In the end, I produced a short tutorial on how to run workflows through Microsoft Azure's Kubernetes Service to help other biologists make their workflows as accessible as possible.

Optimizing machine learning web application runtime using Azure Web Application Services
John Salako, MSc., Department of Earth and Environmental Sciences

In this project, I aimed to optimize the runtime of a machine learning web application I created by utilizing Azure Web Application services. The primary objective was to minimize latency and improve overall performance for end users. By comparing the application's performance on production web application services against the Dev/Test web service plan, I observed a substantial reduction in runtime delay by over 300% when using the production services. Additionally, it was found that the Azure production web services outperformed the free Streamlit cloud alternative, thus providing a more efficient solution for hosting machine learning web applications. In conclusion, leveraging Azure Web Application services for production environments or upgrading to a more robust architecture in Streamlit can significantly enhance the runtime performance of machine learning web applications.

Cloud computing tool for lidar processing
Meicheng Shen, Ph.D. Candidate, Department of Geography, Environment, and Spatial Sciences

More and more observation datasets from the space are available to learn the land surface patterns and related ecological processes. However, processing large datasets are technical, which makes it valuable to develop sharable tools to overcome the technical barrier and advance the scientific understanding. Here, I want to develop a container that can derive canopy structural traits from discrete lidar point clouds.

Machine learning with unstructured data for social scientists
Alfred Torsu, MSc., Department of Department of Political Science

The analysis of unstructured data has long been a time-consuming task for social scientists, requiring extensive reading and analysis to identify core themes and messages. However, recent advancements in cloud computing and the Azure platform have provided pre-built resources for machine learning, enabling the collection of insights from unstructured datasets with greater ease and efficiency. This project highlights the benefits of utilizing these resources, particularly in the analysis of news headlines, presidential speeches, and public policy-related tweets. By utilizing machine learning techniques, social scientists can gain valuable insights from large datasets, allowing for more informed decision-making and policy implementation.

Scalable Earthquake Catalog Generation: Harnessing cloud computing with Kubeflow on Azure
Ziyi Xi, Ph.D. Candidate, Department of Computational Mathematics, Science and Engineering

The demand for accurate and comprehensive earthquake catalogs necessitates efficient and scalable methodologies. In response, we developed a cloud computing workflow using Kubeflow on Azure Kubernetes Service (AKS) to generate earthquake catalogs from archived time-series seismograms. The workflow combines deep learning-based models with traditional geophysical algorithms, providing a scalable and cost-effective solution for processing large volumes of data. Kubeflow on Azure ensures efficient task management through Kubernetes and offers container-based components for easy maintenance. Object storage is utilized for handling archived seismograms and temporary files, while the output earthquake catalog contains information on occurrence time, location, and phase arrival time. Our cloud-based approach demonstrates superior performance compared to traditional supercomputers in terms of efficiency and scalability. By harnessing cloud computing, this project revolutionizes earthquake catalog generation, delivering accessible and scalable solutions for seismic event analysis and contributing to our understanding of Earth's dynamics.

Single cell RNA sequencing using the power of cloud
Arash Yunesi, Ph.D., Department of Computational Mathematics, Science and Engineering

In this project I will demonstrate the power of cloud computing in cancer data analysis, specifically in Single Cell RNA Sequencing. The scalability and possible HIPPA compliance of cloud architecture when hosted in a secure facility enables the safe processing of individual patient data. In near future these results will enable pharmaceutical companies to produce patient-specific drugs for cancer.