PEARC19 Workshop
Humans in the Loop
Monday, July 29, 2019 8:30a – 5:30p

Content Agenda

Presenter: Brian D. Voss, Pervasive Technology Institute at Indiana University
Presentation Abstract: *Humanware – The Critical Importance of People in Cyberinfrastructure (and the Cloud).* This session will discuss the history of the role humans have played in the deployment of information technology, with a focus on the increasing need for investments in 'humanware' in supporting the use of cyberinfrastructure by researchers. The need is becoming even more important as research moves from traditional, on-campus facilities to cloud providers; navigating vended solutions, improving effectiveness and cost efficiency, and facilitating the ability to diversify use based upon the offerings and costs of cloud services. The Humanware project, managed by the Pervasive Technology Institute at Indiana University, is exploring this topic as well as illuminating the potential return on investment (ROI) of the cloud compared to traditional cyberinfrastructure.

Presenter: J. Eric Coulter, Indiana University
Presentation Abstract: *Virtual Clusters in the Jetstream Cloud: A story of elasticized HPC.* We discuss our work providing resources for batch computing via the Jetstream cloud, in the form of SLURM clusters. While these are mainly used by science gateways, there have been a few used in the more traditional commandline manner. The flexible nature of these has also lent itself well to educational work, and has provided the basis for a very successful series of tutorials and workshops. This paper discusses the technical evolution of the Virtual Cluster product, and gives an overview of the science enabled. We discuss the challenges in supporting an ecosystem of these virtual clusters, and in supporting research on cloud resources in general.

Presenter: Kristopher Ezra, Purdue University
Presentation Abstract: *System-of-Systems Analytics Leveraging Azure and the Discrete Agent Framework.* The Discrete Agent Framework (DAF) is a tool for rapidly prototyping system-of-systems simulations at medium fidelity which leverages agent-based modeling. The design of the DAF--which features libraries of modular, composable elements--is aimed at creating simulation environments for Monte-Carlo analysis to provide decision support and metric design for features of interest to problem stakeholders. Because of the need for Monte-Carlo analysis, cloud-computing infrastructure is a logical supplement to SoS analytics using the DAF. This presentation explores the generation of an example single-target tracking model and describes the outcomes, lessons learned, and enabling infrastructure provided by Azure.

Presenter: Eletheria (Ria) Kontou, University of North Carolina Chapel Hill
Presentation Abstract: *Transportation Research with Cloud Resources: Uncovering the Impacts of Ridesourcing Use on Road Crashes.* Improving road safety and setting targets for reduction of traffic-related crashes and deaths is highlighted as part of the United Nation's sustainable development goals and multi-national vision zero efforts around the globe. The advent of shared transportation services, such as ridesourcing, expands mobility options in cities and may impact road safety outcomes. In this empirical work, we analyze the effects of ridesourcing use on road crashes, injuries, fatalities, and driving while intoxicated offences in Travis County Texas. Our approach leverages real-time ridesourcing volumes from this region to explain variation in road safety outcomes. Machine learning techniques, such as linear and Poisson panel data models, are deployed to examine whether the use of ridesourcing is significantly associated with road crashes and other safety outcomes.

Additional covariates capturing socio-demographic characteristics and overall travel exposure serve as controls. Our results suggest that for a 10% increase in ridesourcing use, we expect a 0.26% decrease in road injuries and a 0.37% decrease in driving while intoxicated offences in Travis County in Texas. On the contrary, ridesourcing exposure is not found associated with road crashes and fatalities at a 0.1 significance level. This study augments existing work because it moves beyond binary indicators of ridesourcing presence or absence and analyzes patterns within an urbanized area rather than metropolitan-level variation. Contributions include developing a data-rich, cloud-based approach for assessing the impacts of ridesourcing use on our transportation system's safety. Our case study can serve as a template for other US cities. Our findings provide feedback to policymakers by clarifying the associations between ridesourcing and traffic safety, while helping identify sets of actions to achieve safer and more efficient shared mobility systems.

Presenters: Kate Keahey, Argonne National Laboratory; and Jason Anderson, University of Chicago
Presentation Abstract: *Operational Lessons from Chameleon.* Chameleon is a large-scale, deeply reconfigurable testbed built to support Computer Science experimentation. Unlike traditional systems of this kind, Chameleon has been configured using an adaptation of a mainstream open source infrastructure cloud system called OpenStack. We show that operating cloud systems requires both more skill and extra effort on the part of the operators - in particular where those systems are expected to evolve quickly - which can make systems of this kind expensive to run. We discuss three ways in which those operations costs can be managed: innovative monitoring and automation of systems tasks, building "operator co-ops", and collaborating with users.

Presenter: Richard Knepper, Cornell University
Presentation Abstract: *Red Cloud and Aristotle: campus clouds and federations.* Campus cloud resources represent significant resources for research computing tasks, with the caveat that transitioning to cloud contexts and scaling analyses is not always as simple as it might seem. We detail Red Cloud, Cornell's campus research cloud, and some of the work undertaken by the Center for Advanced Computing (CAC) to help researchers make use of cloud computing technologies. In 2015, Cornell CAC joined with two other universities to develop the Aristotle Cloud Federation, composed of separate campus cloud resources and data sources, supporting a range of science use cases. We discuss the lessons learned from helping researchers leverage both of these science cloud resources as well as leveraging other research cloud infrastructure and transitioning to public cloud.

Presenter: Josiah Leong, Indiana University
Presentation Abstract: *Analyzing a Large Neuroimaging Dataset with Cloud-based Tools.* The size and complexity of neuroimaging data is growing. Researchers are also building more sophisticated methods to analyze the data. While the increased complexity of research methodology reflects advances in neuroscience, the complexity also creates challenges for researchers to faithfully disseminate data, methods, and results. Recent advances in cloud hardware and software can help researchers to face the challenge. Parallelized computing resources can help researchers to analyze large datasets. Virtual software environments can help researchers to share tools and ensure scientific reproducibility. Our project tested the utility of cloud-based tools for analyzing a large neuroimaging dataset. The Adolescent Brain Cognitive Development study is longitudinally collecting neuroimaging data in 11,000 participants for a decade. The data are large (30 terabytes), and are larger after creating data derivatives. The dataset includes several types of neuroimaging data, including structural anatomy and connectivity, and functional activity. I combined the Brain-Life platform (brainlife.io) with Microsoft Azure resources to analyze the data. I compared methods to visualize and measure specific structural brain connections. The measurements that I derived may relate to important brain functions and behaviors, such as impulsivity and risk taking.

Presenter: John Mulligan, Rice University
Presentation Abstract: *Digital Humanities Application Incubation in the Cloud.* Applications that allow humanities researchers to leverage networked assets are relegated mostly to the domain of the commercial web applications: Google Drive and/or Box for backing up photographs, but in ways that do not allow researchers to organize and operate on their assets in the ways that give their datasets the rich complexity that humanities data are (in)famous for. The general use case explored in this project is the incubation of digital humanities applications in the cloud. The purpose of this incubation is experimental: it was chiefly conducted with a view to ascertain how far the infrastructure of cloud technology is adapted to the purposes of humanistic research. In order to see if researchers could take advantage of cloud's *availabilization* of scalable, redundant resources, we developed a custom web application frontend and asset management backend that enables storage, retrieval, organization, and enrichment workflows on networked assets.

Presenter: Dan Sholler, University of California, Berkeley
Presentation Abstract: *"Invisible Work" as a Lens for Understanding Humanware's Role in Research Cloud Computing: Evidence from an interview-based study.* Published discussions of cloud computing's promise place technological capabilities and financial benefits front-and-center. The attention to these outcomes, however, leaves assessment of the "invisible" work placed on human actors relatively unexplored. I report on 45 interviews with researchers and support staff and reveal two commonalities in the "invisible" labor required to conduct cloud research via private vendors: absorbing the time costs of learning new skills for cloud computing and managing billing for cloud projects. I argue for continued documentation of invisible labor to ensure that costs are understood and shared appropriately among vendors, universities, and researchers.

Presenter: Yongwook (Song) Song, University of Kentucky
Presentation Abstract: *A Use Case of Humanware and Cloud-based Cyberinfrastructure: Time-series Data Classification Using Machine Learning.* Humanware, the human component of cyberinfrastructure, is focused on understanding and developing the human expertise needed to support computationally-based research with the goal of maximizing efficiency, productivity, and return on investment associated with cyberinfrastructure. In this workshop session, we present an example use case describing the humanware challenges associated with leveraging cloud-based cyberinfrastructure to implement a machine learning software framework (1-D CNN) that classifies ambiguous time-series data sets. Our project demonstrates that collaboration between researchers and cyberinfrastructure experts significantly advanced our empirical research efforts and maximized the return on investment by utilizing a cost-efficient cloud-based cyberinfrastructure.

Presenter: Gregor Von Laszewski, Indiana University
Presentation Abstract: *Human in the Loop Virtual machine Management on Comet*. The Comet petascale system is an XSEDE resource with the goal of serving a large user community. The Comet project has served a large number of users while using traditional supercomputing as well as science gateways. In addition to these offerings, comet also includes a non traditional virtual machine framework that allows users to access entire virtual clusters instead of just focusing on individual virtual machines. The virtual framework is based on the reuse of the high performance computing scheduler that governs the rest of the machine. However, to access and manage it user input is required. In this paper, we summarize the support supported by and are supported for efforts of human in the loop-for-cloud as part of the computing activities on comet. This includes a discussion of how to get access, how to use the system, how to obtain support and what lessons we learned form the operation of this facility for users.

Presenter: Derek Weitzel, University of Nebraska, Lincoln
Presentation Abstract: *Enabling Microsoft OneDrive Integration with HTCondor.* Accessing data from distributed computing is essential in many workflows, but can be complicated for users of cyberinfrastructure. Creating an easy to use data distribution method can reduce the time researchers spend learning cyberinfrastructure and increase its usefulness. Microsoft OneDrive is a online storage solution providing both file storage and sharing. A barrier to using services such as OneDrive is the credential management necessary to access the service. Recent innovations in HTCondor have allowed the management of OAuth credentials to be handled by the scheduler on the user's behalf. In this presentation, I will focus on providing an easy to use data distribution method utilizing Microsoft OneDrive. Additionally, I will compare it to measurements of data distribution methods currently used on a national cyberinfastructure, the Open Science Grid.

Presenter: Nuyun (Nellie) Zhang, Georgia Institute of Technology
Presentation Abstract: *Towards Cloud Research Support.* Public clouds have great potential to advance research by providing unlimited and on-demand computing resources, but there are gaps that need to be addressed in order to make this happen. In this paper, we try to identify the gaps in supporting research by public Cloud within higher education. However, there is very little research or existing work focusing on this topic. We have to conduct a number of in person consultations and email surveys towards researchers, research support engineers in IT and Public Cloud providers to understand the gaps. This paper tries to provide some preliminary results about the findings in how to better support research in the cloud and proposes some future work.

The Workshop will end with a general discussion (Q&A, audience comments, etc.) among workshop attendees.