The Paul J. ’69 and Kathleen M. Severino Center for Technological Entrepreneurship

invites you to the seminar

“Reproducible Data Science in the Cloud”

Thursday, October 26, 2017

4:00 p.m. to 5:00 p.m. – Ricketts 211

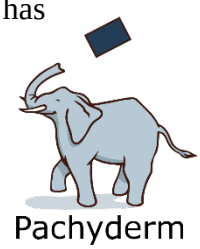
Refreshments served at 3:30 p.m.

Rensselaer Polytechnic Institute, Troy Campus

Daniel Whitenack, Ph.D., Data Scientist, Pachyderm

Summary: Despite the many amazing applications of statistics, machine learning, and visualization in industry, many attempts at doing "data science" are anything but scientific. Specifically, data science processes often lack reproducibility, a key tenet of science in general and a precursor to having true collaboration in a scientific (or engineering) community. In this session, I will discuss the importance of reproducibility and data provenance in any data science organization, and I will provide some practical steps to help data science organizations produce reproducible data analyses and maintain integrity in their data science applications. I will also demo a reproducible data science workflow in the cloud that includes complete provenance explaining the entire process that produced specific results. The workflow will run on modern infrastructure (Kubernetes and Docker) which, in addition to being reproducible, allows it to be language/framework agnostic and scale as needed.

Speaker Bio: Daniel Whitenack (@dwhitena) is a Ph.D. trained data scientist working with Pachyderm (@pachydermIO). Daniel develops innovative, distributed data pipelines which include predictive models, data visualizations, statistical analyses, and more. He has spoken at conferences around the world (GopherCon, JuliaCon, PyCon, ODSC, Spark Summit, and more), teaches data science/engineering at Purdue University (@LifeAtPurdue) and with Ardan Labs (@ardanlabs), maintains the Go kernel for Jupyter, and is actively helping to organize contributions to various open source data science projects.



Co – Sponsored by:

Rensselaer Center for Open Source (RCOS)

Center for Supply Networks and Analytics

IDEA Rensselaer Institute for Data Exploration and Applications

Rensselaer SEVERINO CENTER for TECHNOLOGICAL ENTREPRENEURSHIP