Sharing Research through replicable Data Science Environments

Project Logistics:

Mentors: Sri Krishnamurthy email: s.krishnamurthy@neu.edu; email: ;
Min-max team size: 4-6
Expected project hours per week (per team member): 6-8
Will the project be open source: no

Preferred Past Experience:

- Job handling / concepts of distributed computing Required

Linux basics Required
Python, Mean stack, MongoDB Valuable
Docker and HPC scheduling Required
Web technologies for analytics Valuable

Project Overview:

Background:

Researchers typically share their research through forums like https://arxiv.org/. In the past decade, there has been a growing interest in replicable research projects to address the so called replication and reproducibility crisis (https://en.wikipedia.org/wiki/Replication_crisis). In a recent Nature paper (https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970), it was quoted that more than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. QuSandbox (www.qusandbox.com) is our effort to address the reproducibility problem by enabling creation of replicable data science environments. Our technology leverages Docker, many open source and proprietary technologies and the cloud to embed code and replicable data science environments to enable researchers, students and practitioners to create and share projects that can be replicated and shared.

Project Specifics:

In this project, we aim to prototype an environment customized for researchers to be able to share their research. For example, think of a researcher creating a new algorithm for text analysis. Let us say a draft paper is published on arxiv and the software is available on github. Our goal is to create a platform where in addition to github, the researcher can share the code and the environment leveraging QuSandbox's research cloud environment.

Some Technologies you will learn/use:

Working on challenges encountered by researchers and data science practitioners
Designing applications for the cloud
Using cutting edge technologies for cloud automation
Docker,Python and Distributed computing