New skills acquired during summer research

- My knowledge in software preservation was very limited in the beginning of the summer. My only saving grace was that I had command line experience (helpful since most of the software was linux-based) and basic exposure to big data. I spent most of my time Googling and asking questions. I learned how to spin up VM’s using Vagrant and VirtualBox. I learned to use the replication tools, CDE and Reprozip. Furthermore, I spent a lot of time working with Docker and understanding how to containerize it. Working with all of these tools contributed to my knowledge of Linux (and Windows since I’m running Windows).

Project Summary

Scientists can almost never recreate the same computational experiment twice which often leads to large gaps in knowledge. Reproducibility in computational sciences is hard due to the numerous dependencies between software, data and hardware, and lack of good tools to define computational steps (workflow tools) each experiment requires. In order for scientist A, Alice, to send her experiment to scientist B, Bob, Bob must install and run the executable, the exact version of pip, various software packages, and a version control software as well to deal with file installation issue/corruption, etc. Along the way something will break causing the experiment to fail to execute.

This is a critical issue for DASPOS scientists who need to preserve their experiments for future scientists to examine when the technology is better. Several software technologies, such as Reprozip and CDE are available to package the experiment and rerun the experiment in a sandboxed environment. These programs capture the provenance (data, workflow and environment) by tracking executables of the experiment via p-trace or systemtap. However, there are several drawbacks to these technologies. They are oftentimes difficult to install and use, only operate in Linux environments and are kernel dependent.
Due to kernel dependencies, the package will fail to execute on Bob’s side unless it is an exact match to the Alice’s kernel environment.

A proposed solution to eliminating kernel dependency is to use Docker. Docker is a light, portable container which allows to run other applications (ie Reprozip, CDE) inside Linux containers (LCX). This allows the dependency to move from the virtual machine specific kernel to Docker - a much more feasible solution to save.

**Publications (papers/posters/presentations):**
- Poster presentation at the Summer Undergraduate Research Symposium (SURS) at the University of Notre Dame on August 1, 2014