Portable, Reproducible Analysis with Arvados

Peter Amstutz
<peter.amstutz@curoverse.com>
BOSC 2015
Need for Computational Reproducibility

• Scientific method is premised on reproducibility
• Record the process leading to results
• Collaboration/confirmation of results or adoption of techniques by 3rd parties
What is Arvados?

• [https://arvados.org](https://arvados.org)
• Platform for managing and processing data at scale
• FOSS (AGPLv3 / Apache 2.0)
• Commercially supported by [Curoverse](https://curoverse.com)
Content addressable storage “Keep”

**Key Properties**
- Immutable
- Self-certifying
- Versioned
- Manifests allow reorganization

1. Read files
2. Split into blocks
3. Checksum each block
4. Store in Keep
5. Write manifest
Repeatable computation
“Crunch”
Portable computation

- Recursively copy pipelines, input files, git repositories, docker images
- Copy between cluster/cloud/workstation
- Run on target system with a single click
- Verify results using content addresses
Collaboration and Sharing

• Share data, git repositories, and pipelines
• Share with users, groups, or general public
• Run shared pipeline on own data
• Run own pipeline on shared data
Next generation: Common Workflow Language

• Developing and supporting cross-vendor standard for pipelines

• Goal to provide cloud-based “app store” for bioinformatics

• Come back for Michael Crusoe's talk about CWL at 2:00pm