# **Scientific Workflow and Data Management** with the Arvados Platform

# Summary

Reproducibility benefits largely from robust workflow management. The open-source platform Arvados integrates a data management system called "Keep" and the compute management system called "Crunch", creating a unified environment to store and organize data, and run Common Workflow Language workflows on that data. Arvados is multi-user and multi-platform, running on various cloud and high performance computing environments.

Arvados management features including the ability to (1) identify the origin and verify the content of every dataset, track every workflow run, and reliably reproduce any output (2) organize and search for datasets using metadata (3) securely and selectively share your data and workflows (3) efficiently manage data (minimizing storage costs) and (4) efficiently rerun workflows (minimizing time and compute costs).

## **Security and Sharing**

Comply with data protection regulations

• Arvados provides authentication, access and audit controls, data integrity, and transmission security

Collaborate with others by selectively and securely sharing data and workflows

- Private by default
- Read-only, read/write, or manage (to grant permission to others)

vados	plavaround
Dashboard	Share with groups
1.4.10	Only groups you are allowed to access are shown. F a specific group.
	Search
Description	📽 System group
Advanced	📽 Anonymous users
missione	📽 All users

# **Arvados Workflow Management**

- Crunch Workflow Manager:
- Scalable container orchestration system for running CWL workflows
- Designed to maintain data provenance and workflow reproducibility

• Automatically stores complete record of workflow execution in collections

- Inputs/outputs, docker image, logs
- Referenced by content address (portable data hash)

• Reorganization *does not* break references

(Below) An executed workflow in Arvados viewed via the Arvados Workbench. The Workbench web application allows users to interactively access Arvados functionality

	WGS processing workflow sca Variant Caller for WGS data (paired FASTOs) Container Request UUID pirca-xvhdp-yjvxwzjhvlcoap7 Owner Testing (pirca-j7d0g-o4f9f3at4hkqpfc) Created at 6/17/2021, 4:12:14 PM Finished at 6/17/2021, 4:17:42 PM Requesting Container UUID (none)	Variant Caller for WGS data (paired FASTQs) Container Request UUID Dirca-xythdp-yjvxxzjhvicoap7 Created at V17/2021, 4:12:14 PM Finished at V17/2021, 4:17:42 PM Requesting Container UUID			Completed : × Docker Image Portable Data Hash	
	Priority 500 Properties No properties <> Logs 2021-06-17T20:13:04.4214941752 notice: rea	Log Collection	.procs	Output Collectio	n 🖂 ×	
	2021-06-17T20:13:04.421787654Z mem 3404021 2021-06-17T20:13:04.421836736Z notice: rea 2021-06-17T20:13:04.421887942Z notice: rea 2021-06-17T20:13:04.421908185Z cpu 27.2400 2021-06-17T20:13:04.422073254Z net:ens5 12 2021-06-17T20:13:04.422078712Z net:docker0	ding stats from /sys/fs/cgroup/cpuacct/cpuacc ding stats from /sys/fs/cgroup/cpuset/cpuset. user 5.8100 sys 2 cpus 9491 tx 40500569 rx	t.stat			
	CPU Information processor : 0 vendor_id : GenuineIntel cpu family : 6 model : 85	4 #1 SMP Debian 4.19.118-2+deb10u1 (2020-06-0	7) x86_64 GNU/Linux			
Workflow – Steps	model name : Intel(R) Xeon(R) Platinu			<u>San a por a constante da const</u>	$R \equiv 13 \times$	
	• generate-report_5	Status <b>T</b>	Created At 4:16:06 PM	Run Time Oh Om 1s	Actions	
	generate-report_2	Completed	6/17/2021, 4:16:04 PM	0h 0m 2s		

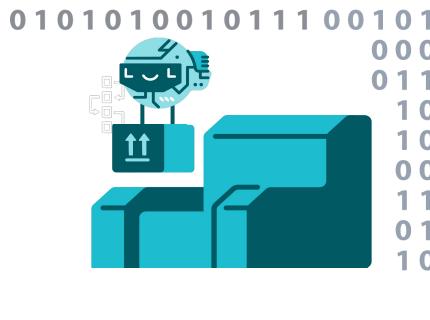
# Why Reproducibility?

*Reproducible* workflows are necessary to:

- Further study or to support scientific claims
- Answer questions from collaborators or regulators
- Fulfill regulatory requirements to retain data



Keep **Storage System** 



By combining both data and workflow management in a single open source system, Arvados can run reproducible, scalable, and portable workflows on large datasets.

## **Arvados Data Management**

Keep Storage System combines:

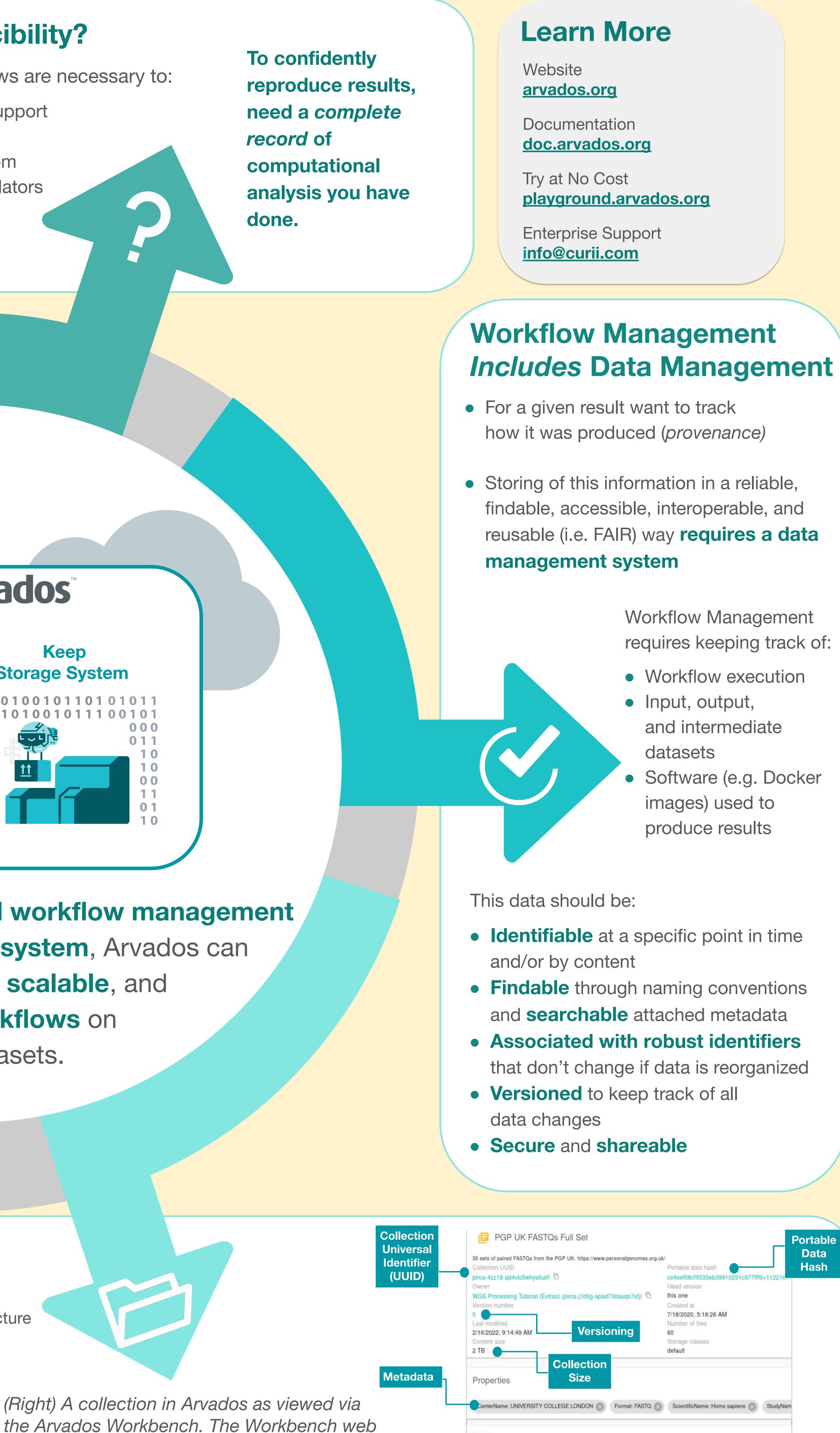
- Content addressing and distributed storage architecture
- Collections contain set of files (dataset)
  - Organized into shareable "Projects"
  - Add and query metadata
  - Keep history of changes
  - Associated with multiple identifiers: content address, database UUID, name

Arvados functionality

# Peter Amstutz

Tom Clegg, Lucas Di Pentima, Stephen Smith, Ward Vandewege, Alexander (Sasha) Wait Zaranek, Sarah Wait Zaranek

# Curii Corporation, info@curii.com



application allows users to interactively access

ERR2122553\_1.fastq.gz

# Listing

Portable Data Hash 7/18/2020, 5:18:26 AM Number of files Storage classes default

• **Identifiable** at a specific point in time • **Findable** through naming conventions and **searchable** attached metadata • Associated with robust identifiers that don't change if data is reorganized

### Input, output, and intermediate datasets • Software (e.g. Docker images) used to produce results

findable, accessible, interoperable, and reusable (i.e. FAIR) way **requires a data**