# The Whole Tale:
## Merging Science and Cyberinfrastructure Pathways

Bertram Ludäscher, Adam Brinckman, Kyle Chard, Niall Gaffney, Mihael Hategan, Matthew B. Jones, Kacper Kowalik, Sivakumar Kulasekaran, Bryce Mecum, Jarosław Nabrzyski, Damian Perez, Victoria Stodden, Ian Taylor, Matthew Turk, Thomas Thelen, Craig Willis, Sebastian Wyngaard

*Presented by:*

Craig Willis (willis8@illinois.edu)

National Center for Supercomputing Applications

July 11, 2018

NDS/MBDH Data Science Tools & Methods Workshop






# Motivation and Vision

- Many scientific experiments are difficult, if not impossible, to replicate, verify, and/or reproduce
- The scholarly publication has not kept pace with the changes in science:
  - **Data underpins most research** whether acquired, derived, or obtained from a repository
  - **Computation & software is an integral & inseparable component** via which most research takes place
  - WholeTale aims to **capture and preserve the journey towards discovery rather than just the endpoints**

*Supported by NSF, United States grant 1541450 (CC*DNI DIBBS: Merging Science and Cyberinfrastructure Pathways: The Whole Tale).*

# What is a "Tale"?

- A living publication that preserves all digital scholarly objects and can be shared and replayed
  - Input, intermediate, derived data
  - Software and environment
  - Workflow process
  - Publication narrative
- Captures computational steps and provide compute environment
- Provides unique identifiers to objects
- Publishable research object

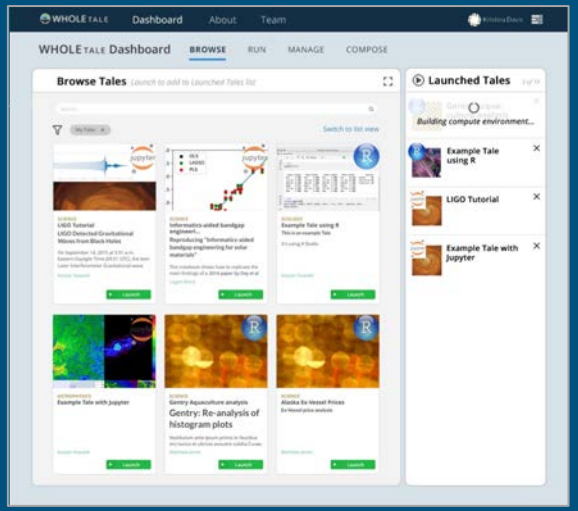
# Community Engagement

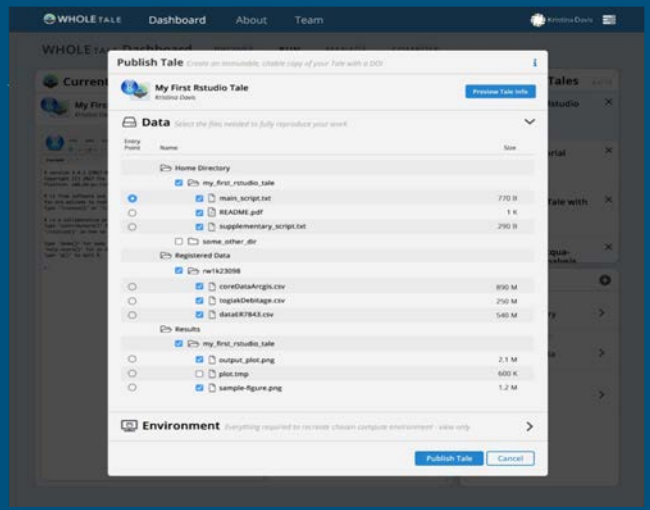
- Cyberinfrastructure and science working groups help drive development
  - Astrophysics, materials science, environmental science, bioinformatics, social sciences, reproducibility, information sciences, education and training
- Internship program
  - Reproducibility of published materials science machine learning models
  - Automated provenance capture when reconstructing environment conditions
  - Understanding infrastructure requirements to promote reproducible research

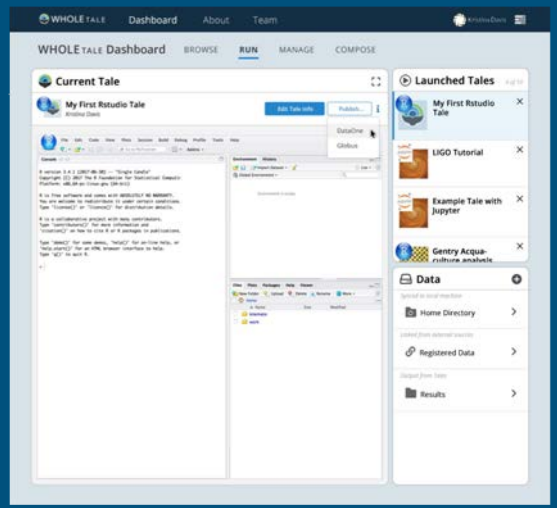
# Community Engagement

- Cyberinfrastructure and science working groups help drive development
  - Astrophysics, materials science, environmental science, bioinformatics, social sciences, reproducibility, information sciences, education and training
- Internship program

# User Interface



# Upcoming Workshop

- Whole Tale Workshop: Tools and Approaches for Publishing Reproducible Research
- When: September 13-14, 2018
- Where: Big Ten Center, Chicago
- What:
  - Discuss initiatives to conduct, track, remix, and share research
  - How can Whole Tale make workflows easier and more reproducible?
- Who:
  - Domain scientists, computer scientists, infrastructure developers

# **COLDFRAME**:

## A Scalable Framework for Collaborative, Data-Intensive Research and Education

*Nurturing collaborative data-intensive research in adverse conditions*

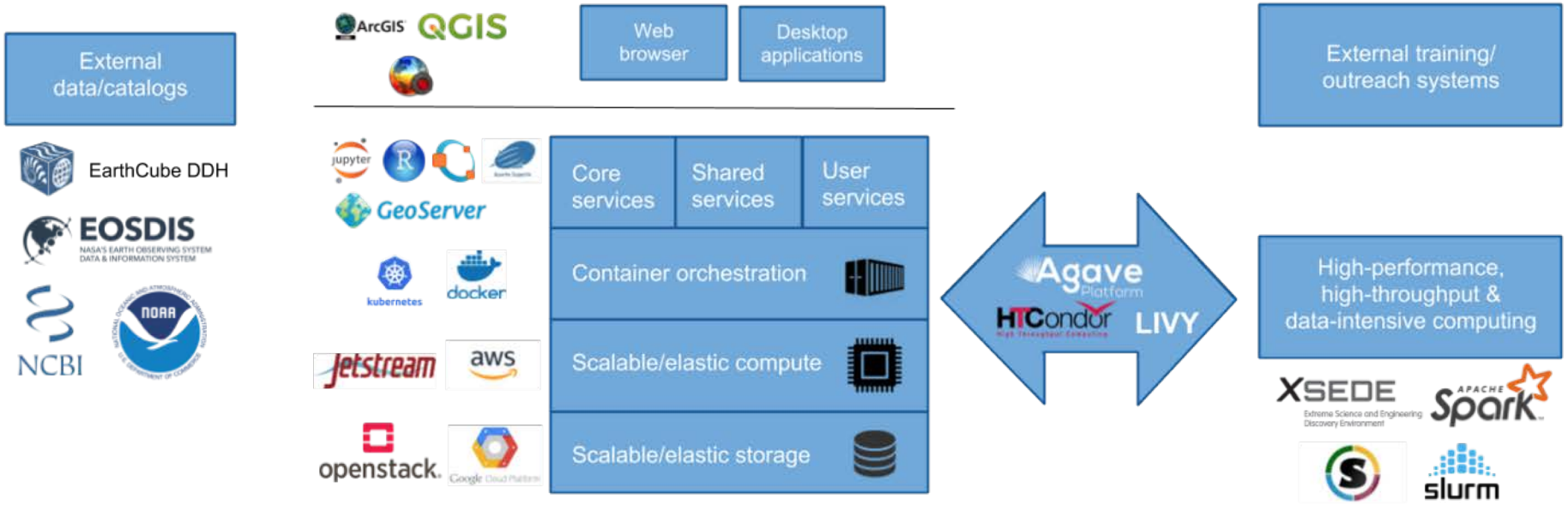
# Problem

- Data intensive science is increasingly collaborative
  - Diverse sets and levels of expertise
  - Geographically distributed teams
- Teams face challenge of enabling collaboration
  - Technical infrastructure
  - Rapidly evolving computational/research data management systems
  - Common patterns/architectures and tools
- NSF big idea: Harnessing the Data Revolution
  - Integrated data and computational infrastructure to accelerate data-intensive research and workforce development.

# Coldframe

- Nascent initiative
- Framework to enable research teams to easily provide collaborative access to large research datasets and specialized computational resources
- Address needs identified through extensive work in the EarthCube, National Data Service (NDS), and Big Data Hub (BDHubs) communities
- Bridging the gap between interactive analysis and execution on cloud, High Performance Computing (HPC) and High Throughput Computing (HTC) resources
- Supports outreach and education on the system used for research

# High-level architecture



# Communities/Drivers

- EarthCube Coral Reef Science and Cyberinfrastructure Network (CRESCYNT)
  - Enabling collaborative, multi-disciplinary analysis of coral reef bleaching events
- High-throughput phenomics and genomics (TERRA-REF)
  - Supporting collaborative access to 2PB reference dataset and tools on a variety of resources
- Data science education for non-R1 institutions
  - Consortial/cooperative models for data science/CI education
- Access and sharing of fused satellite data (Terra Fusion)
  - How can they provide access to 3PB fused dataset for collaborators and future reuse?
- Computational astrophysics (Moesta)
  - How can they provide access to 300TB simulation for analysis/visualization?

# Key features

- Multi-scale deployment
- Data ingest
- Access control and sharing
- Scale-out compute support
- Image preservation (research)
- External compute support
- External data management
- Community-created catalogs