A Reproducible Framework Powered by Globus

Tanu Malik, Kyle Chard*, Ian Foster

Computation Institute
University of Chicago and Argonne National Laboratory

Sharing and Reproducibility

Alice wants to share her models and simulation output with Bob

Bob wants to re-execute Alice’s application to validate her analysis

Current Approaches

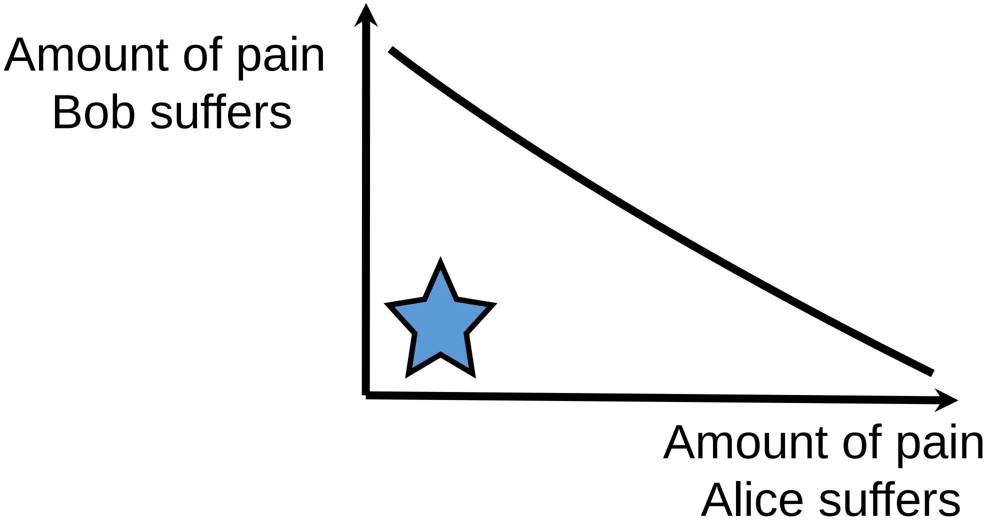
- Compressed archive (zip, tar)
- Metadata encoded in filenames, paths, and README files
- Shared user accounts
- Ad hoc websites with model code, parameters, and data
- Submission to domain or institutional repositories
- Creation and sharing of virtual machines

User Frustrations

“I cannot find the lib.so required to build the model”

“I can’t find input data to run the model”

“I don’t know the parameters used to execute the analysis”



User Frustrations

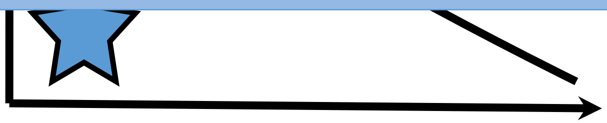
“I cannot find the lib.so required to build the model”

“I can’t find input data to run the model”

“I don’t know the parameters used to execute the

a

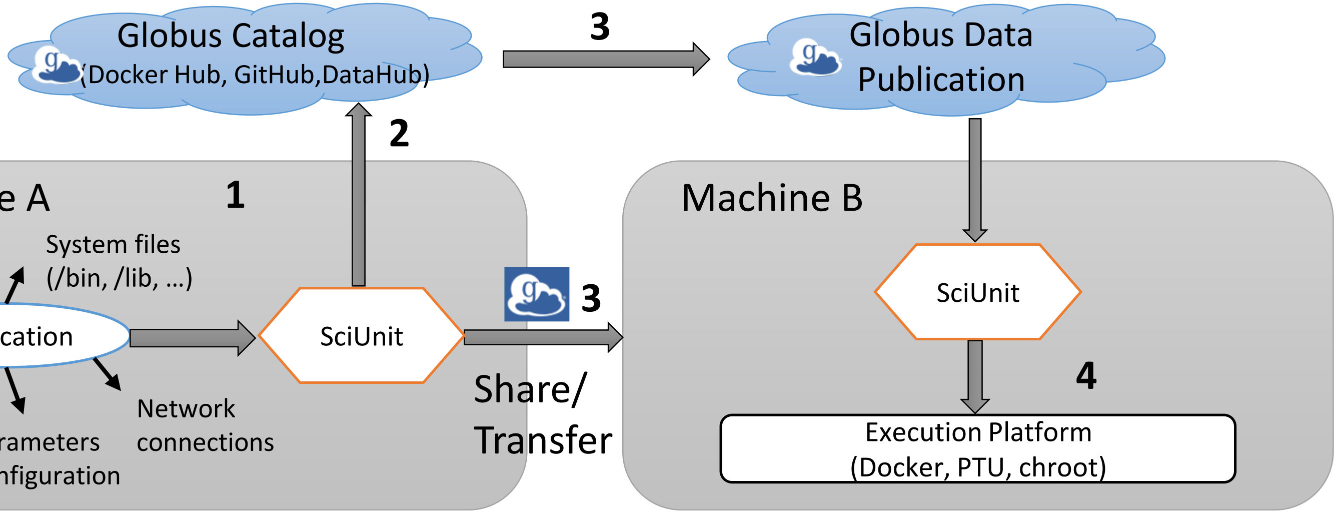
Lack of easy and efficient methods for sharing and reproducibility



Reproducibility Requirements

- Automatically solve “dependency hell”
  - E.g., incompatible library versions
- Connect programs with data and capture data flow
  - Which version of the program produced this data?
- Support annotation of human knowledge
  - Sufficient documentation to install and run the program
- Enable reproducibility efficiently and with minimal intervention
  - No change of programming or authoring environments

Reproducibility Framework



- Capture scientific activity
  - Source code, data, environment, including data flow between processes
- Preserve as "SciUnit"
  - Capture as files or as detailed metadata
- Share and distribute
- Re-execute and re-analyze

Components

- “SciUnits”
  - Units of scientific activity/research output
- Metadata catalog
  - A scalable & flexible cloud-based catalog for creating datasets and associating annotations
- Globus services for sharing, transferring, and publishing SciUnits
- Replay capability through native re-execution, Docker or Vagrant
  - Run SciUnits without installation or configuration and metadata information

GeoDataspace

Simplifying **Data** Management for Geoscience Models

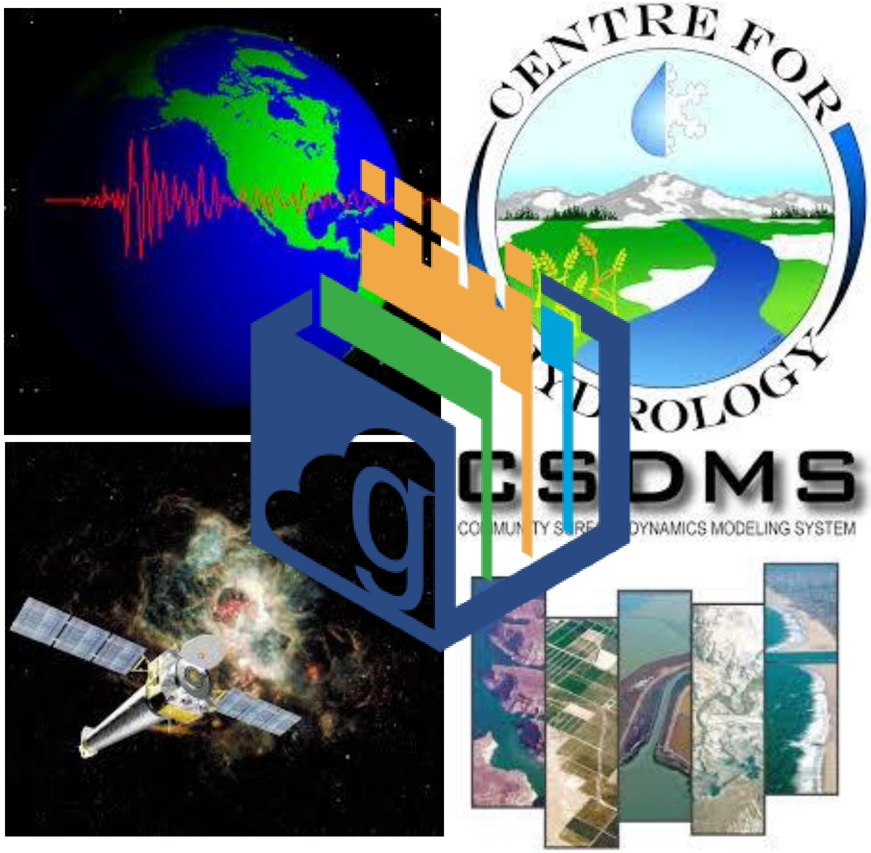
Tanu Malik, Ian Foster, Kyle Chard,

Joseph Baker, Mike Gurnis, Jonathan Goodall, Scott Peckham

Science Drivers

Solid Earth

Hydrology

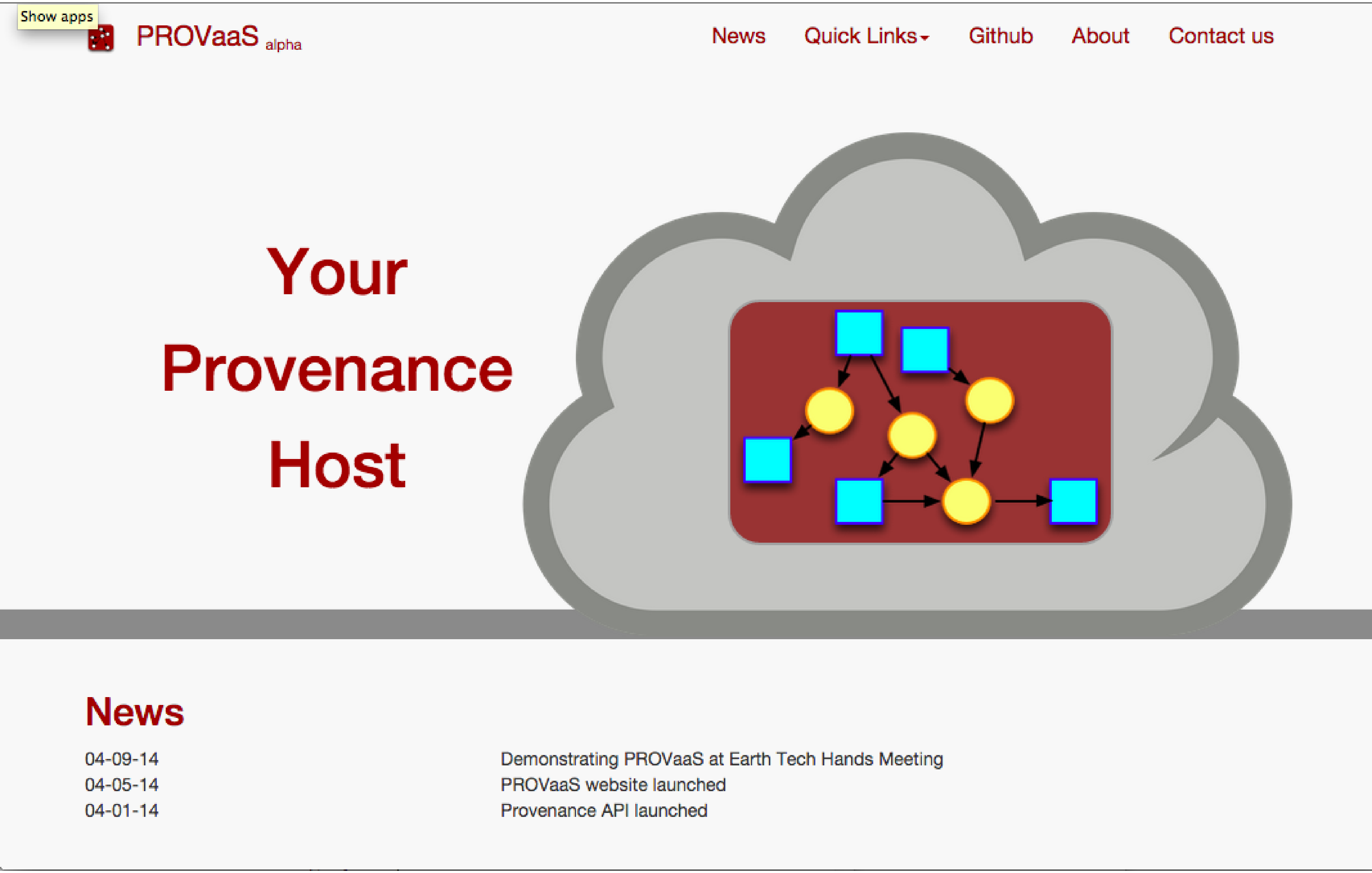


Space Science

CSDMS

Science Usecases

- Solid earth: geounits of 2D and 3D kinematic geoscience models, visualize through GPlates and modifying GPML data files
  - Goal: sharing, preserving, publishing visualization sessions with data
- Space Science: geounits of SuperDARN data with analysis tools as available from the Baker Lab at VT
  - Goal: sharing and publishing geounits
- Hydrology: geounits of iRODS workflows on VIC models
  - Goal: demonstrate end-to-end reproducibility with iRODS



Your Provenance Host

News

04-09-14	Demonstrating PROVaaS at Earth Tech Hands Meeting
04-05-14	PROVaaS website launched
04-01-14	Provenance API launched

Thank You!

NSF

DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA

ALFRED P. SLOAN FOUNDATION
1934

EARTHCUBE