SCITECH

From Scripts to Workflows: Orchestrating Science with Pegasus

Ewa Deelman

University of Southern California
Information Sciences Institute

deelman@isi.edu

U.S. DEPARTMENT OF ENERGY

NSF U.S. National Science Foundation

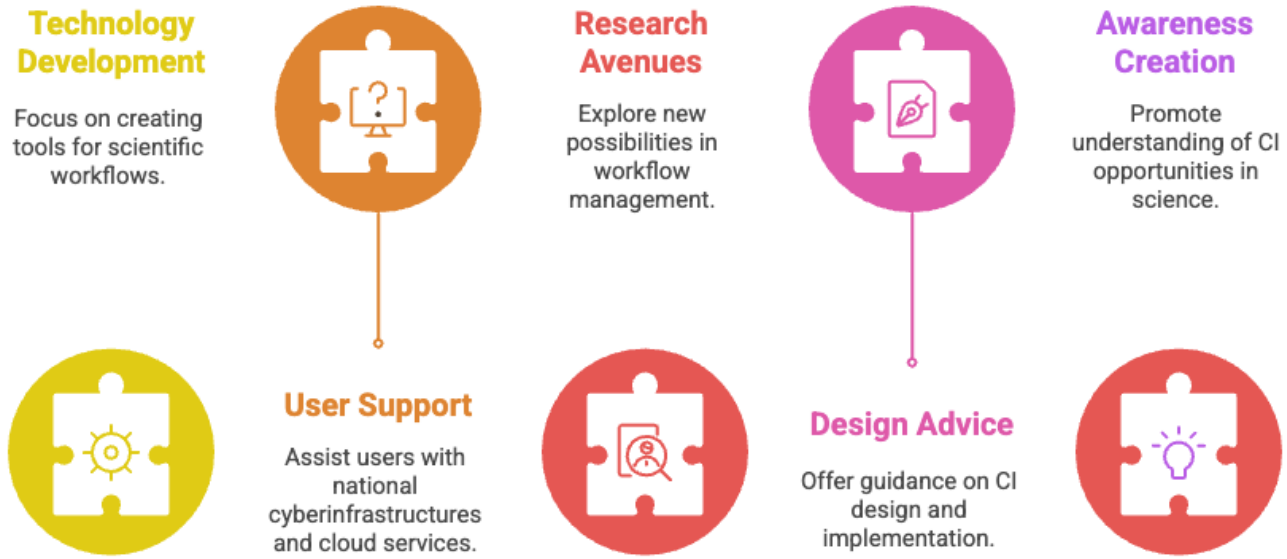
USC Viterbi School of Engineering

SCITECH

SciTech Activities

1. Develop technologies to support scientific workflows
2. Provide user support on national cyberinfrastructures and clouds
3. Research new avenues in the workflow management space
4. Provide advice on the design, implementation and use of Cyberinfrastructure (CI) for the Research Infrastructure lifecycle
5. Create awareness and understanding of opportunities in CI for Science

Marina del Rey, CA

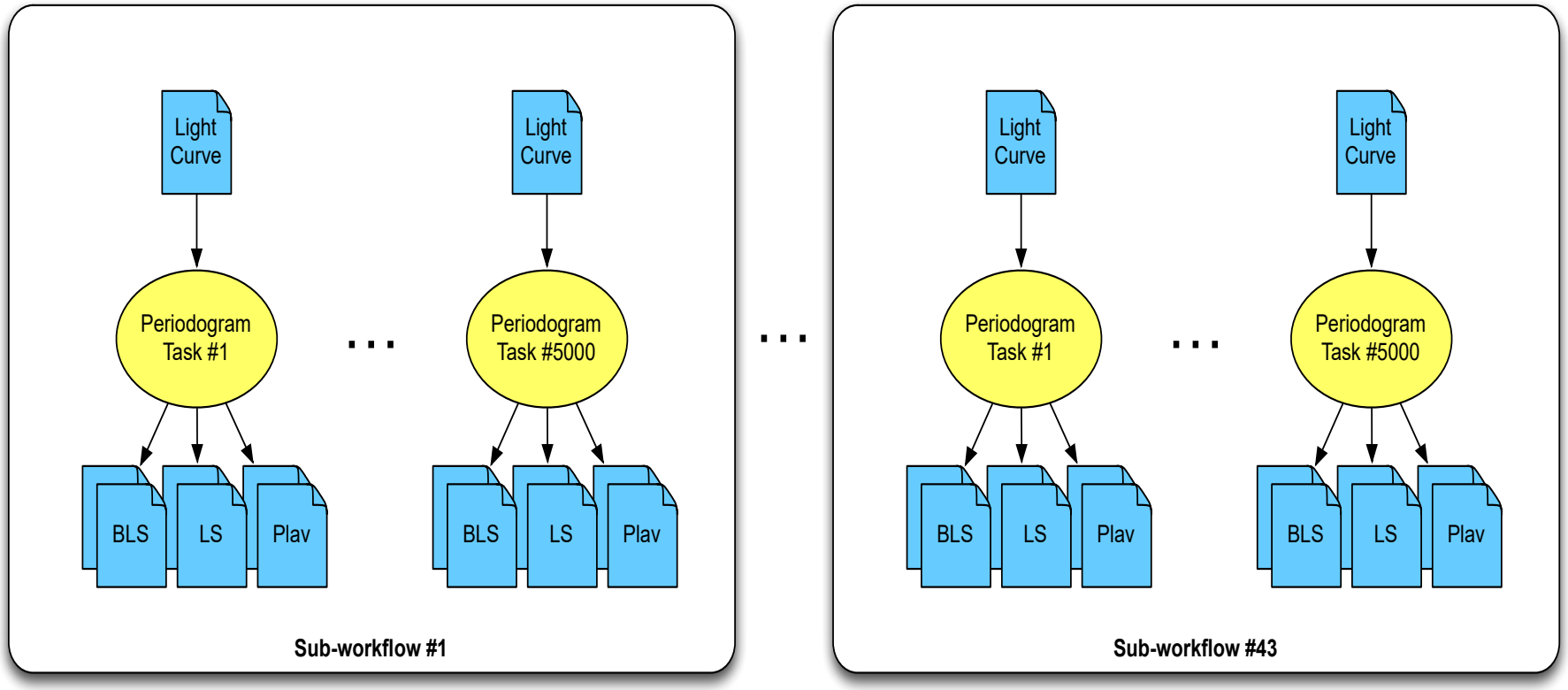


Most projects are collaborative across 4-6 organizations

USC Viterbi School of Engineering

Some workflows are simple!

Bag of Tasks



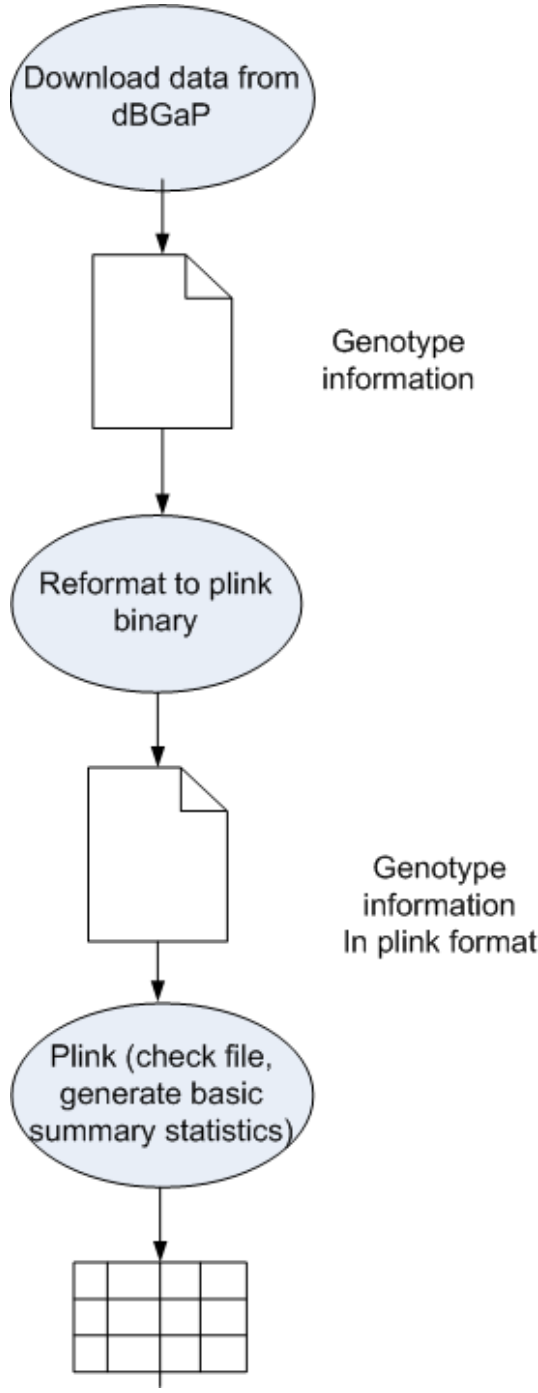
Kepler periodigram workflow

Scripts

- Quick to code and deploy
- Quick to learn and tailor to the local campus resource
- Not very scalable, hard to move between resources, hard to maintain, hard for others to understand, often error-prone

Computational workflows

- Help express multi-step computations in a declarative way
- Can support automation, minimize human involvement
  - Makes analyses easier to run
- Can be high-level and portable across execution platforms
- Keep track of provenance to support reproducibility
- Easier to maintain
- Foster collaboration—code and data sharing
- Take some time investment to learn



Pegasus Workflow Management System, est. 2001

Workflow Challenges Across Domains

- Describe complex workflows in a simple way
- Access distributed, heterogeneous data and resources (heterogeneous interfaces)
- Deal with resources/software that change over time
- Ease of use. Ability to monitor and debug large workflows

Our Focus

- Separation between workflow description and workflow execution
- Workflow planning and scheduling (scalability, performance)
- Task execution (monitoring, fault tolerance, debugging, web dashboard)
- Workflow optimization, restructuring for performance and fault tolerance.

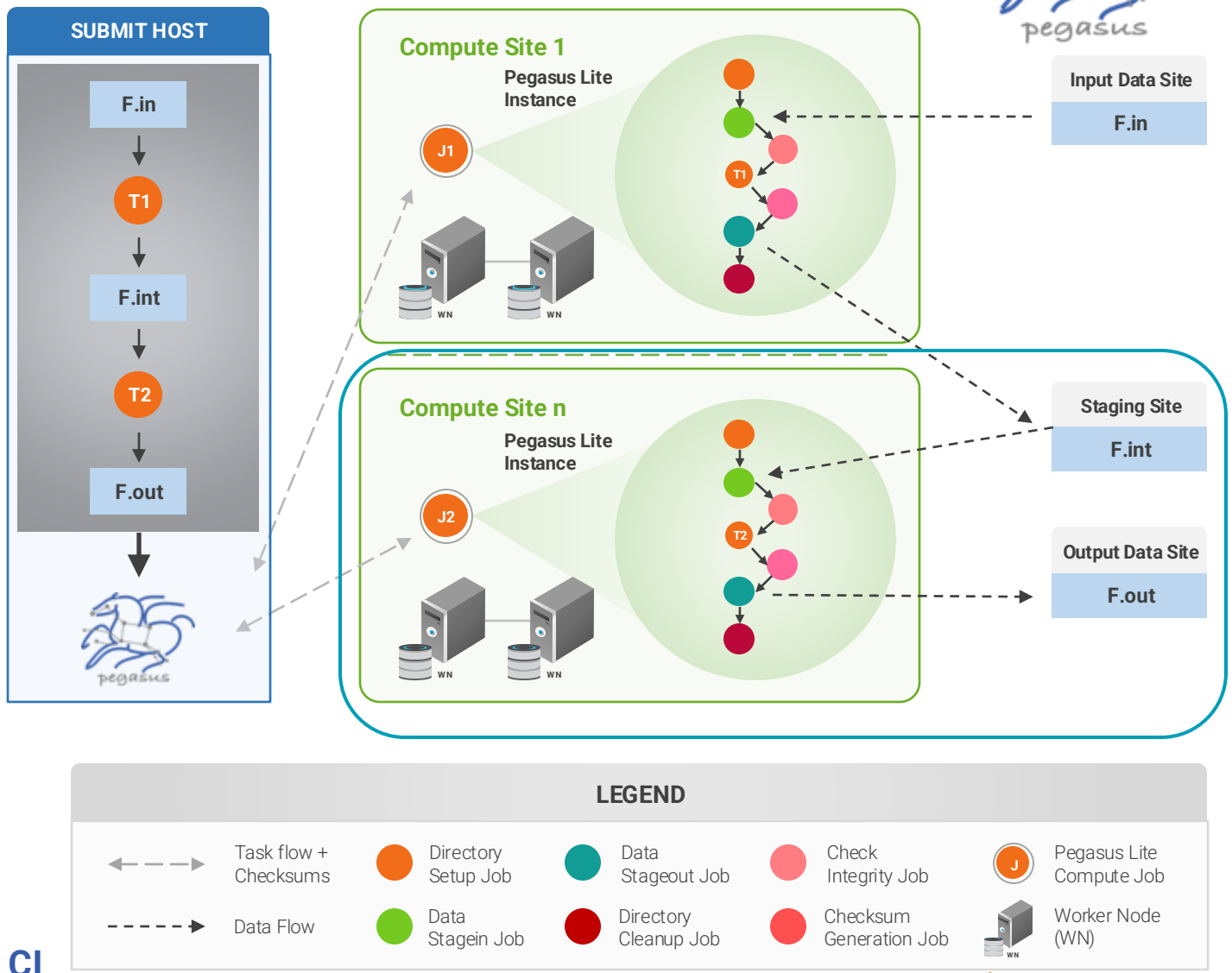
U.S. National Science Foundation

Submit locally, run globally

- **Pegasus WMS == Pegasus planner (mapper) + DAGMan workflow engine + HTCondor scheduler/broker**
  - Pegasus maps workflows to target infrastructure (1 or more resources)
  - DAGMan manages dependencies and reliability
  - HTCondor is used as a broker to interface with different schedulers

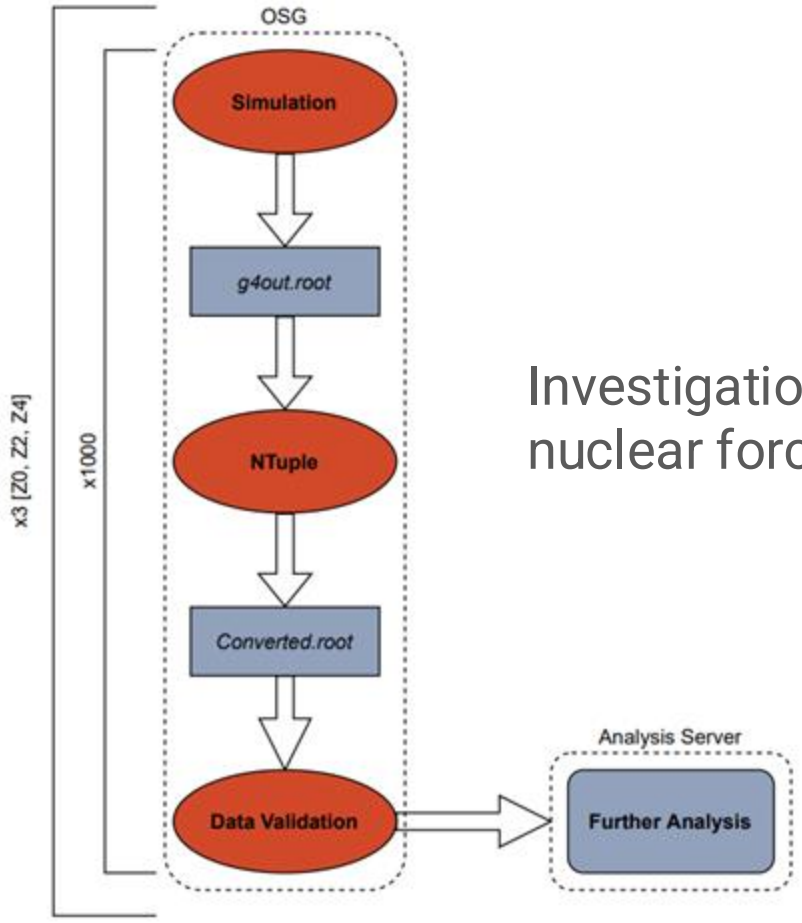
- **Planning converts an abstract workflow into a concrete, executable workflow**
  - Planner is like a compiler
  - Optimized performance
  - Provides fault tolerance

- **Can leverage distributed and heterogeneous CI**



From a proof of concept to winning the 2022 David Swanson Memorial Award

Connor Natzke,
Physics Student
Colorado School of Mines

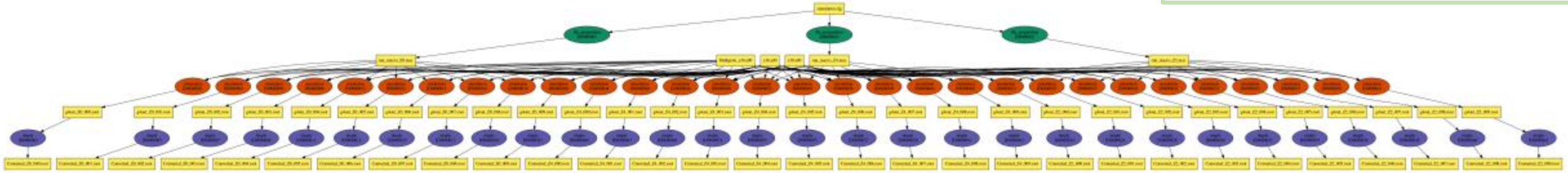


Investigation of the strong nuclear force

Pegasus-based Monte-Carlo simulation, 590,000 jobs, 15 years total wall time, 4 hours wall time on Open Science Grid

Pegasus provided:

1) Automation and scaling
2) Automatic job retries
3) Automation of file transfers
4) Managed disk space at execution sites



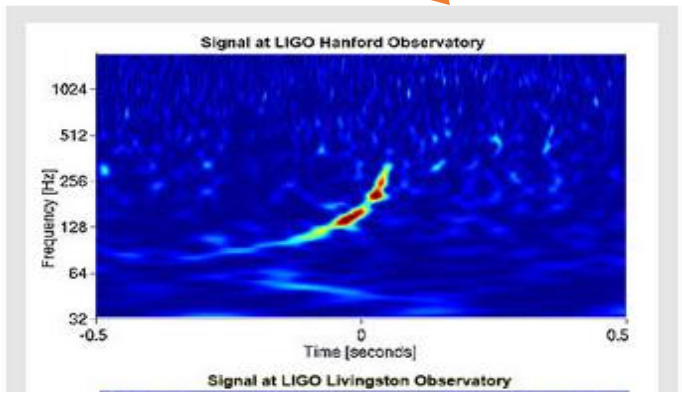
Workflows Can Move with Time

Pegasus use in the Laser Interferometer Gravitational Wave Observatory

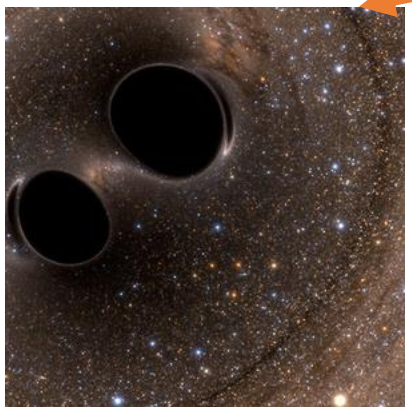
Nobel Prize

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

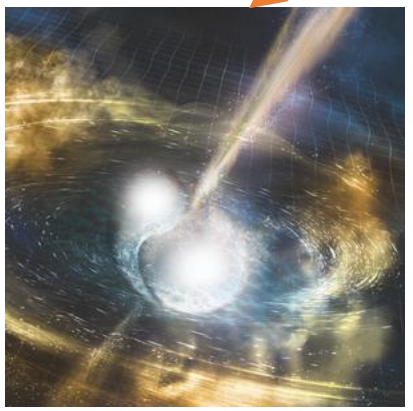
First Pegasus prototype



Blind injection detection



First detection of black hole collision



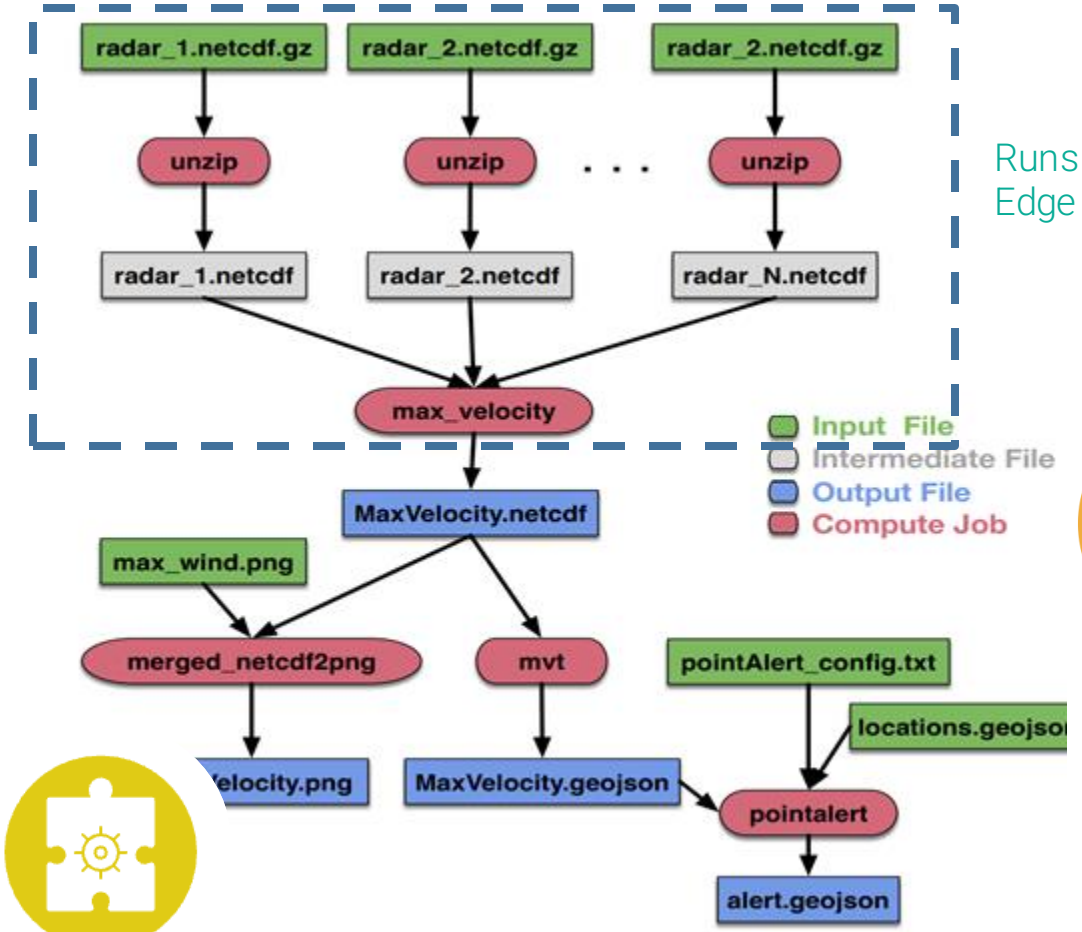
Multi-messenger neutron star merger observation

Edge-2-Cloud Applications

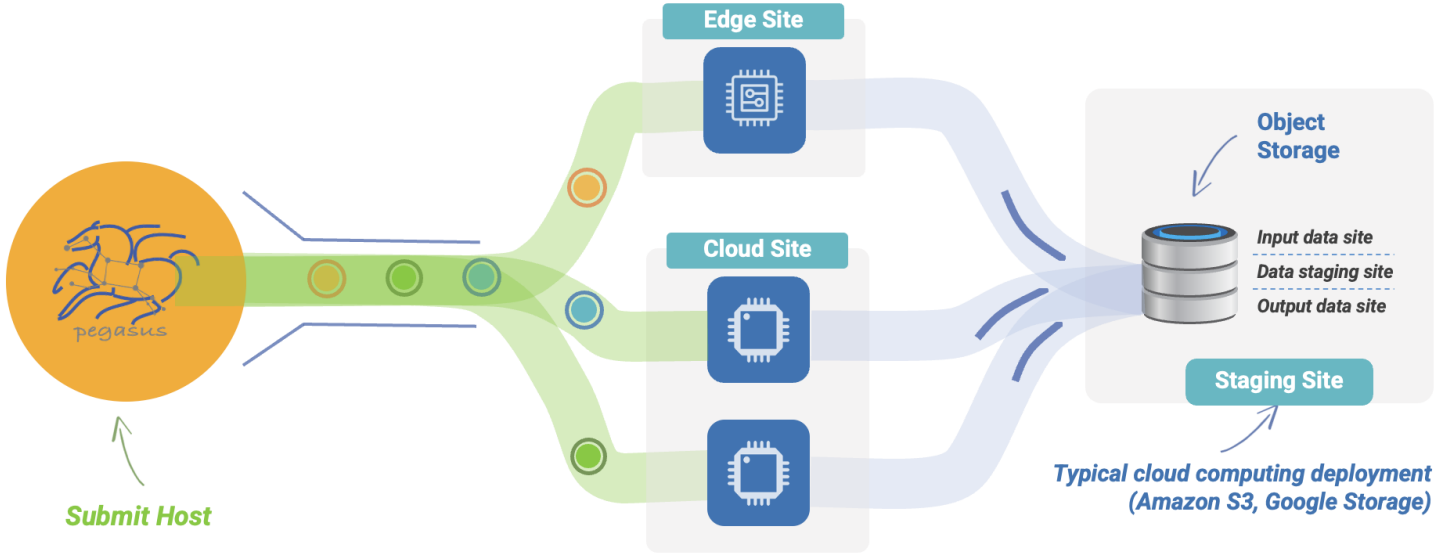
CASA: Collaborative and Adaptive Sensing of the Atmosphere

- Has deployed a network of short-range Doppler radars
- Compute and data repositories at the edge, close to the radars
- Use on demand cloud resources to scale up their computations

http://www.casa.umass.edu/



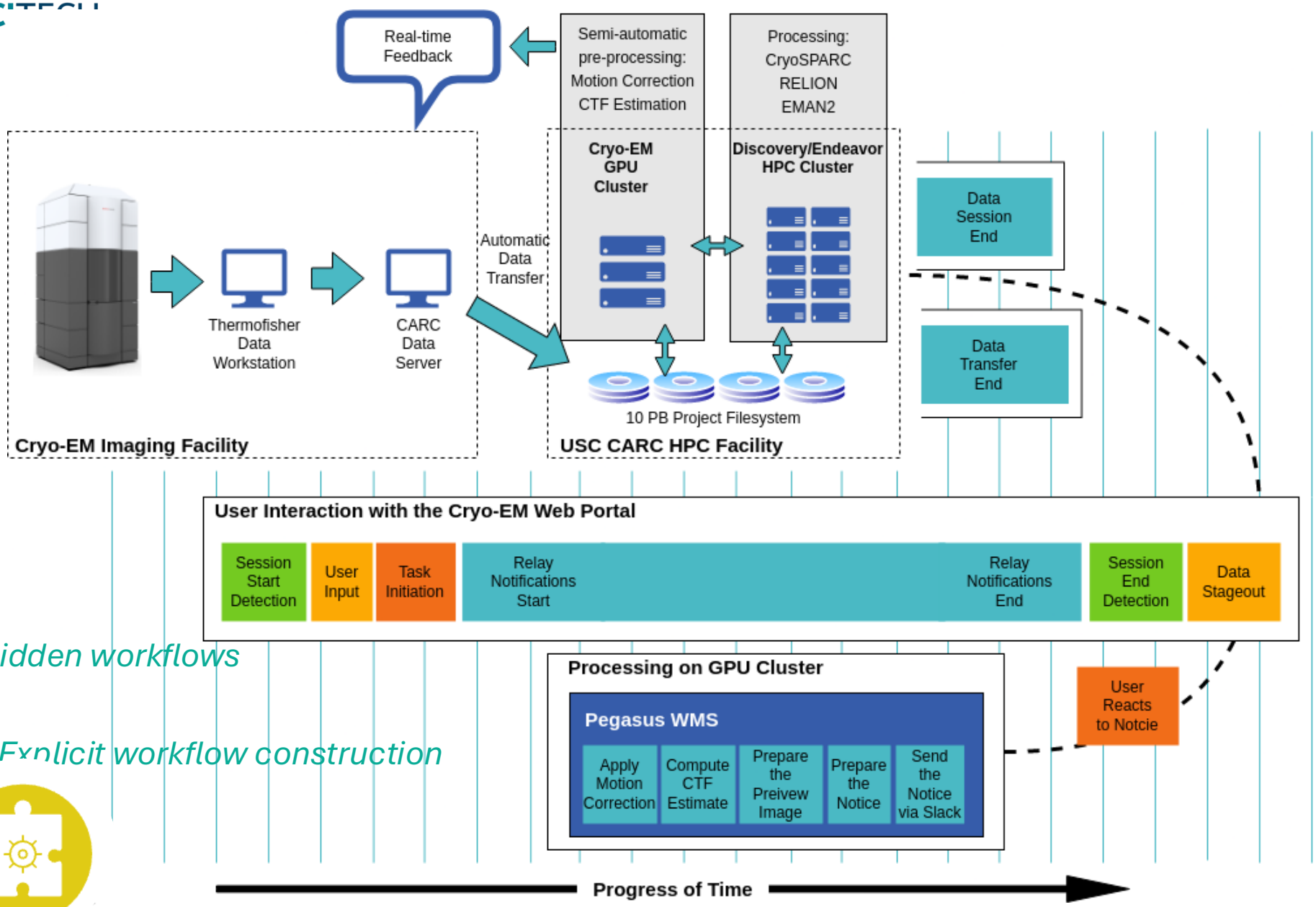
Runs at the Edge



Decoupling of compute and storage

U.S. National Science Foundation

Processing instrument data in real time



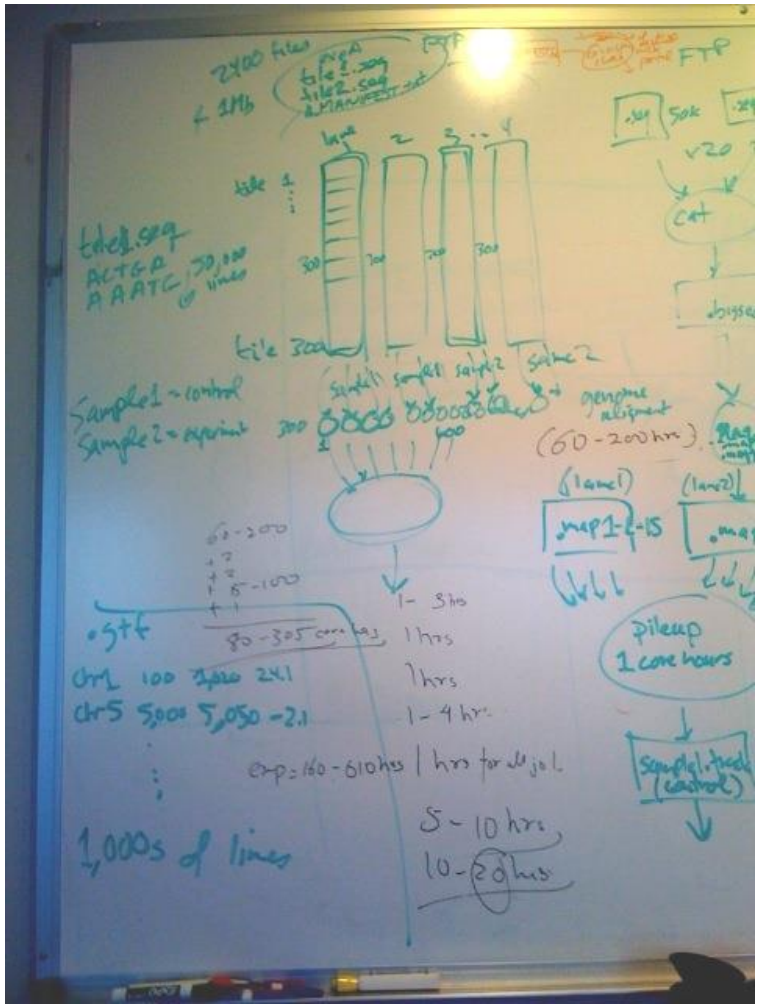
- Totally hidden from the user
- Curated, pre-defined workflow
- Automated data transfers
- Automation of pre-processing
- Quick feedback during experiments
- Used in production at USC

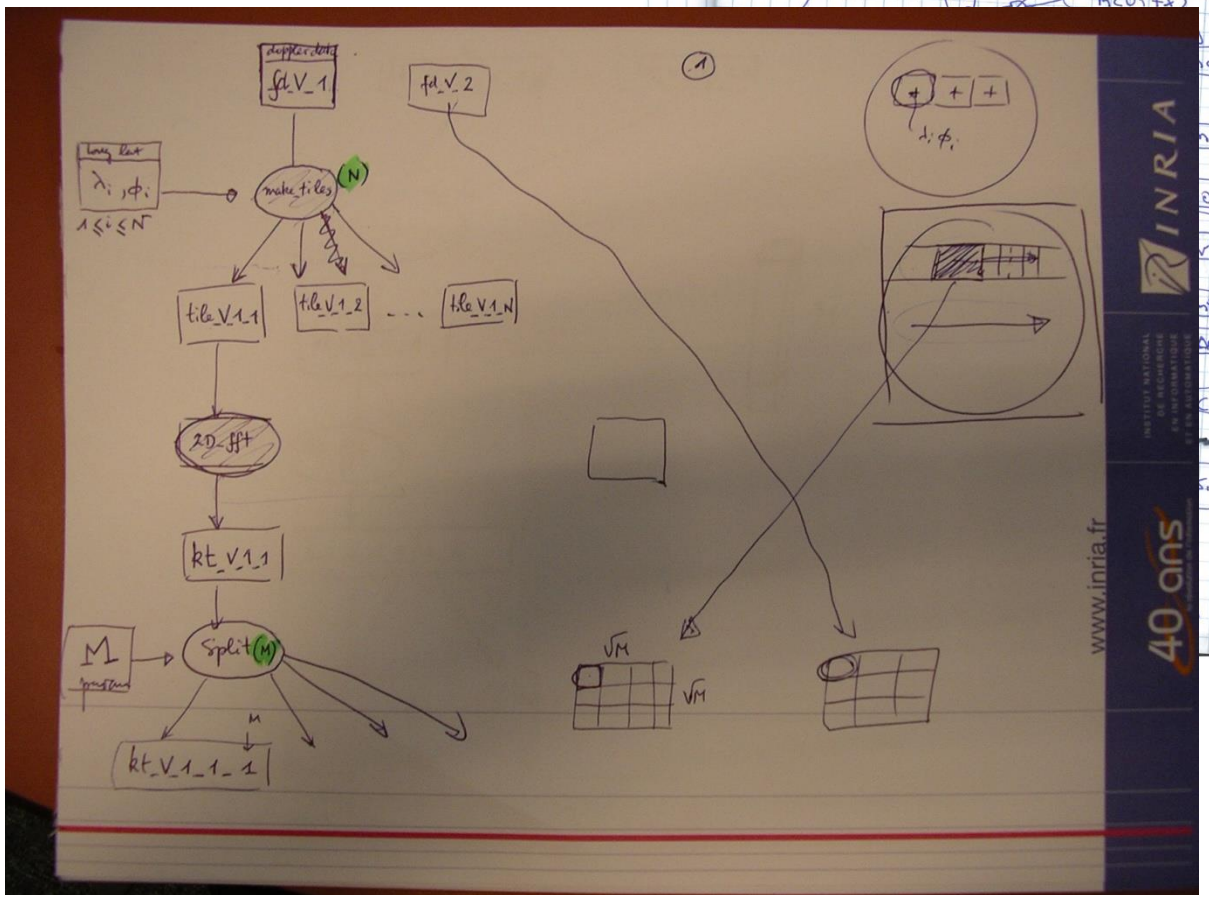
Hidden workflows

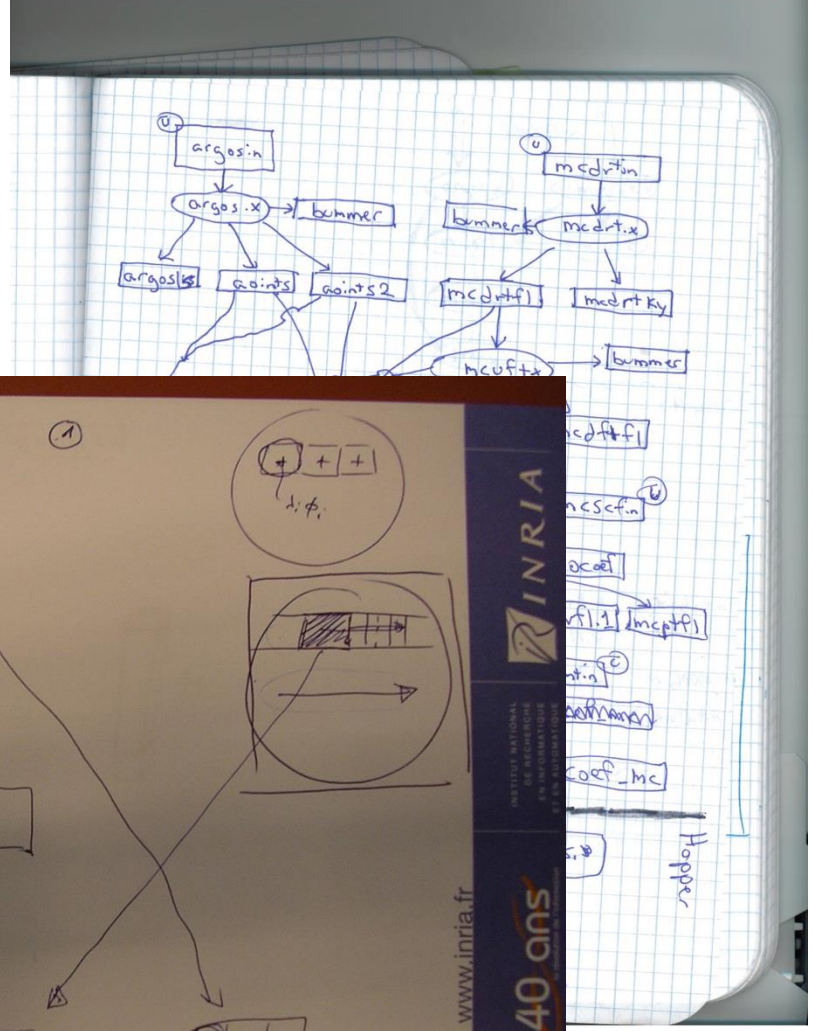
Explicit workflow construction

Pegasus
Workflow Management System
https://pegasus.isi.edu /

How do workflows start?







Pegasus provides tools and APIs to create workflows using Python, Java, R and support Jupyter Notebooks

SCITECH

Challenge	Pegasus' Solution
Staging data	Automated data transfer to and from computations
Different storage systems	Pegasus can talk a number of protocols, including HTTP, FTP, Globus Online, HTCondor and others
Small workflow tasks	Pegasus can cluster tasks together for more efficient execution
Limited storage (edge)	Pegasus analyzes the workflow and cleans up data no longer needed
Failures during execution	Job retries, trying different data sources, workflow-level checkpoint, rescue DAGs
Have a full workflow, but some data was already computed	Pegasus can re-use that data and run only the necessary jobs
Don't know what happened during the execution	Pegasus has tools for analyzing workflow performance and help debug them, pinpointing errors

Pegasus keeps track of how the result was obtained: Full Provenance, support for containers

SCITECH

NSF | ACCESS

ACCESS CI provides researchers with:

High-Performance Computing (HPC) : NCSA Delta, Purdue Anvil, TACC Frontera, SDSC Expanse, PSC Bridges2, …

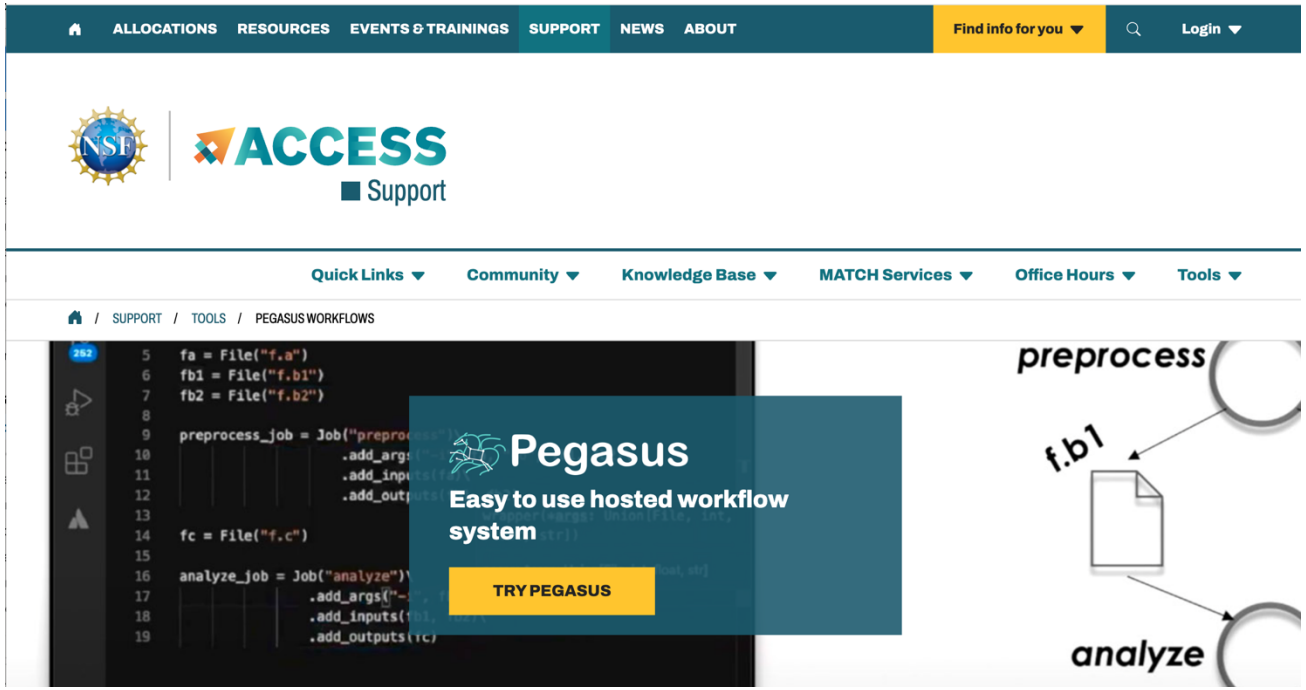
High-Throughput Computing (HTC): OSPool (Open Science Pool)

Cloud & Interactive Platforms: IU Jetstream2 for on-demand VMs and analysis

Data & Storage Services: Shared storage, data movement, OSN (Open Storage Network)

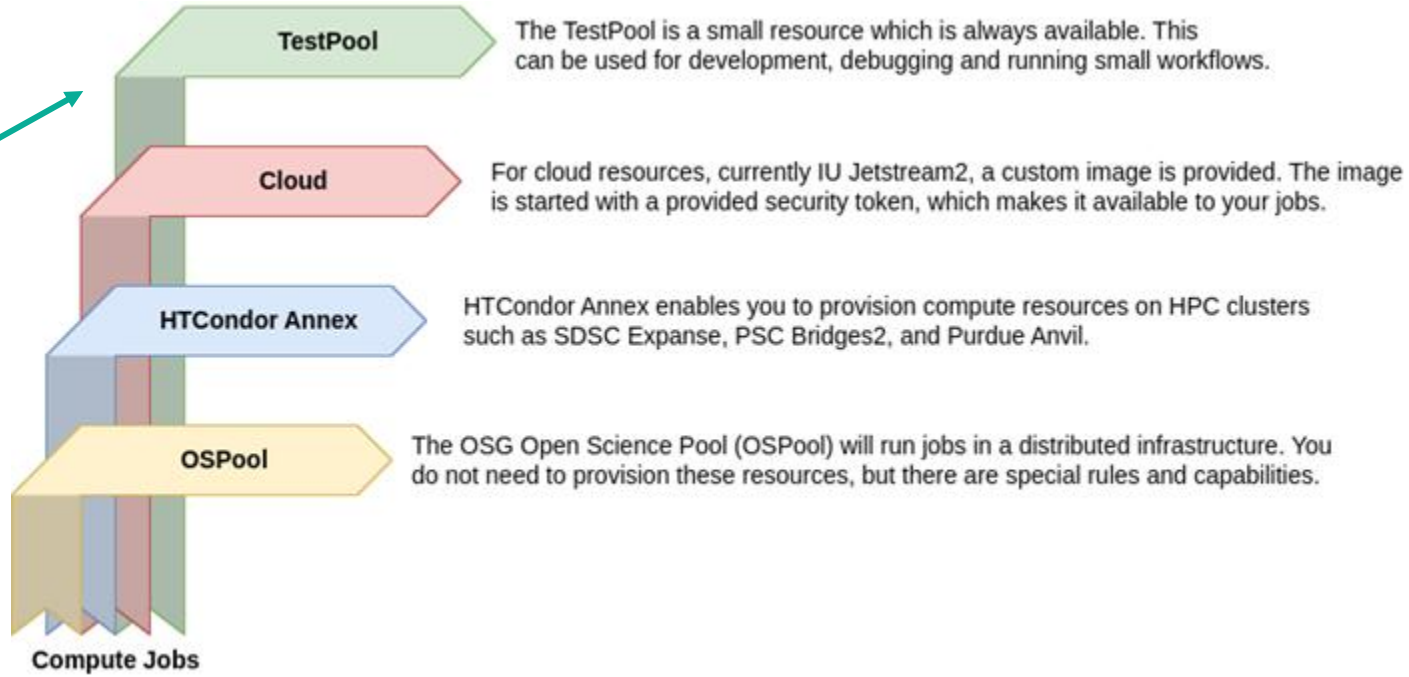
Networking & Middleware: Fast research network connections and software services

Expert Support: Consulting and training to optimize codes, workflows, and data



Get an account on ACCESS and visit the Support Page

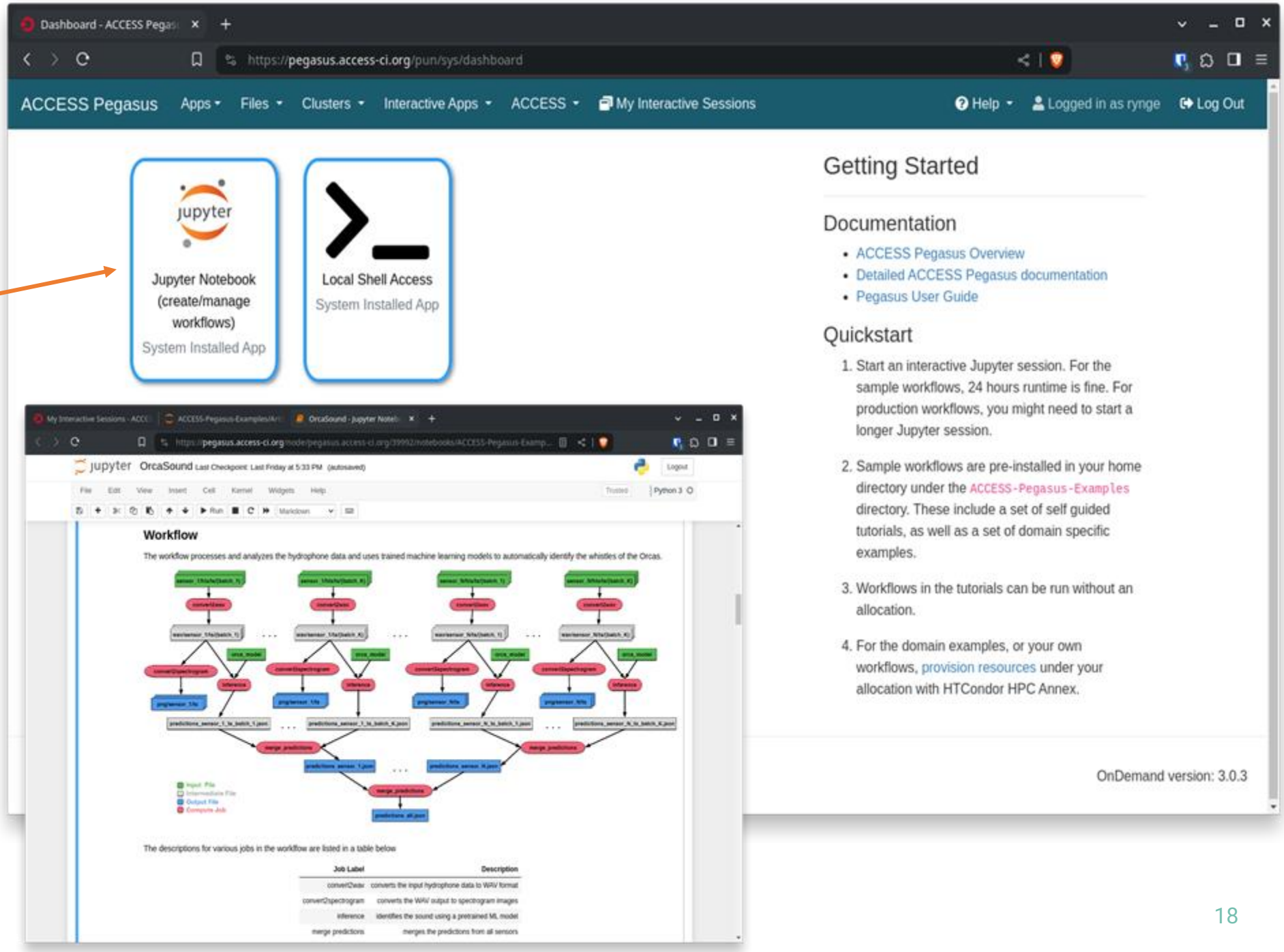
You don't need an allocation to try Pegasus



SCITECH

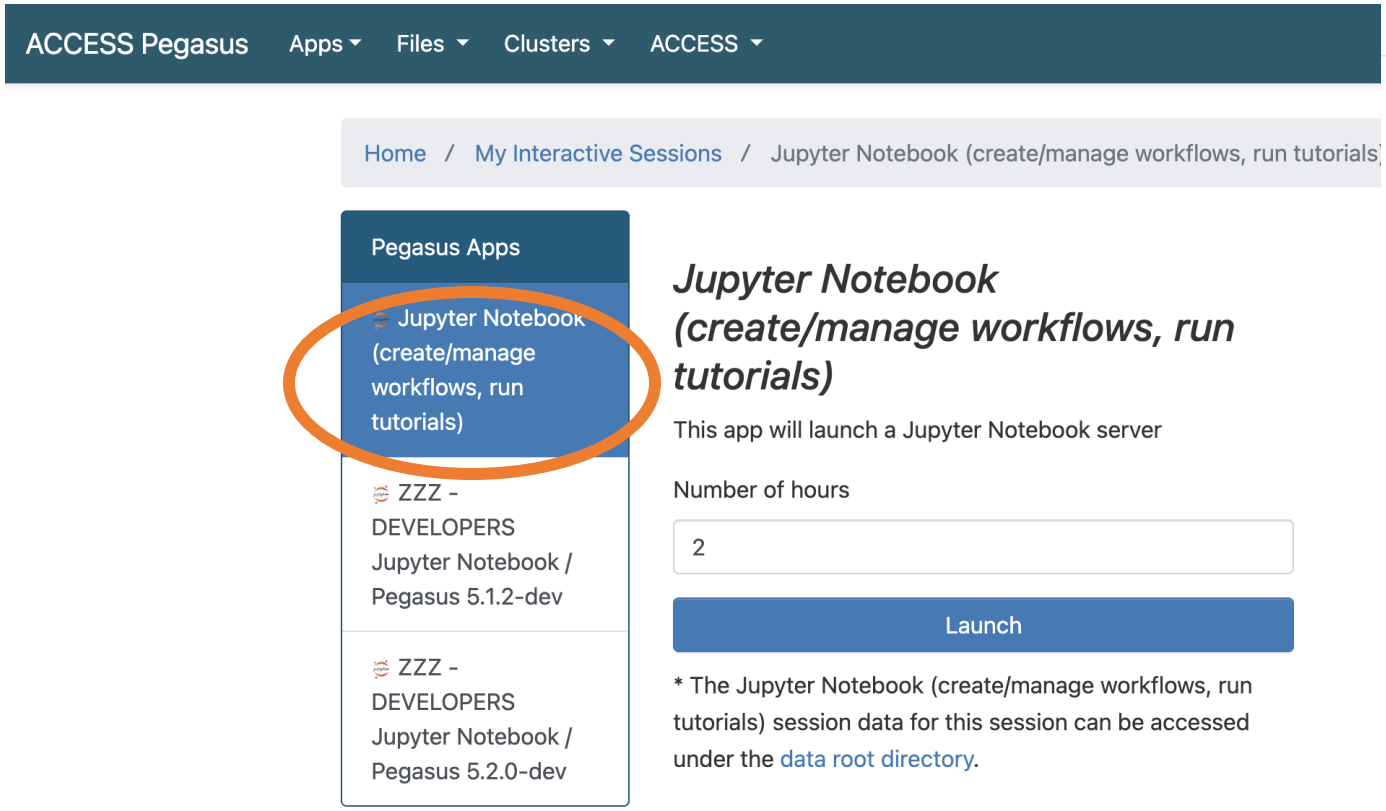
ACCESS Pegasus

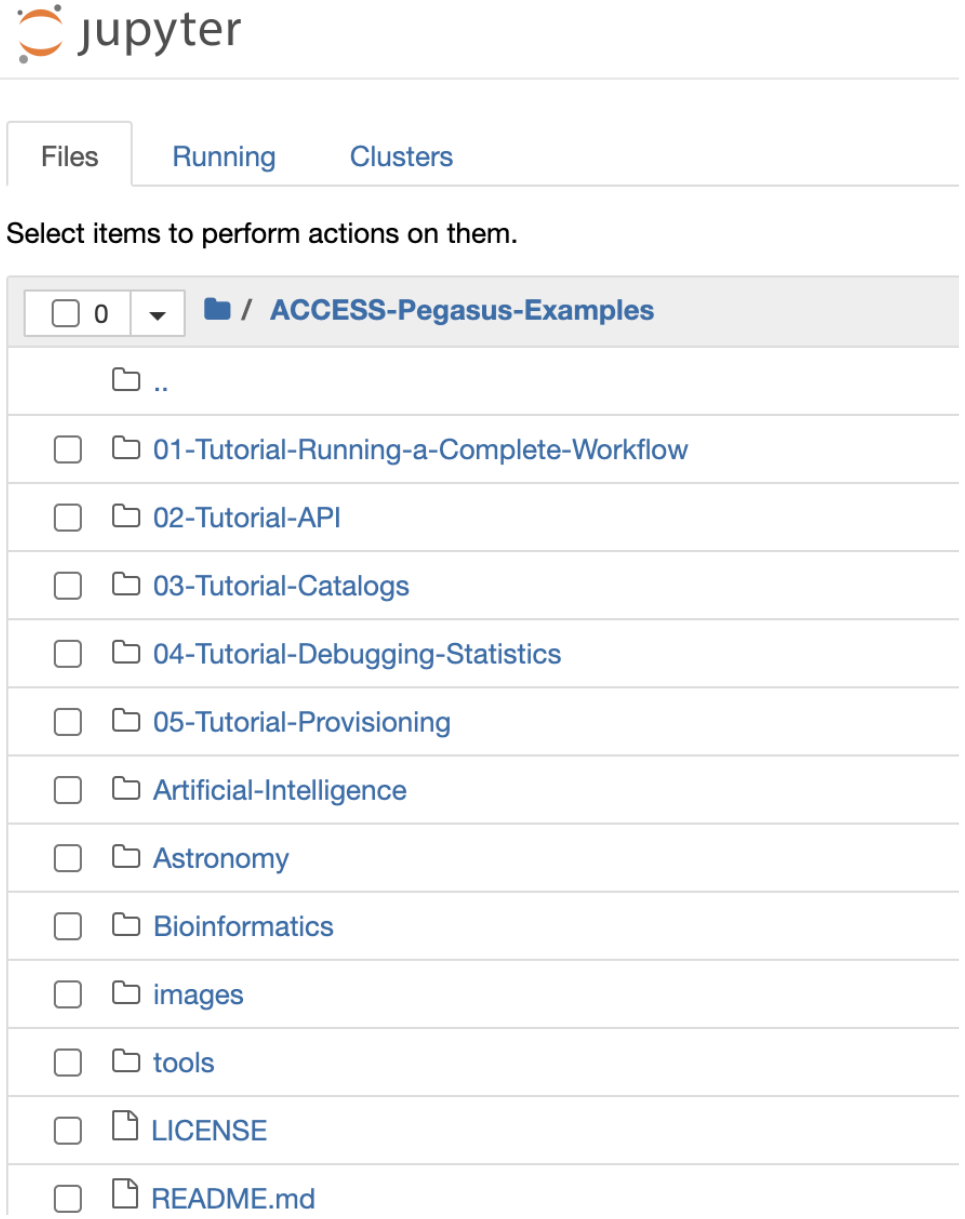
Leveraging Open OnDemand and Jupyter Notebooks to to broaden access to Pegasus' capabilities



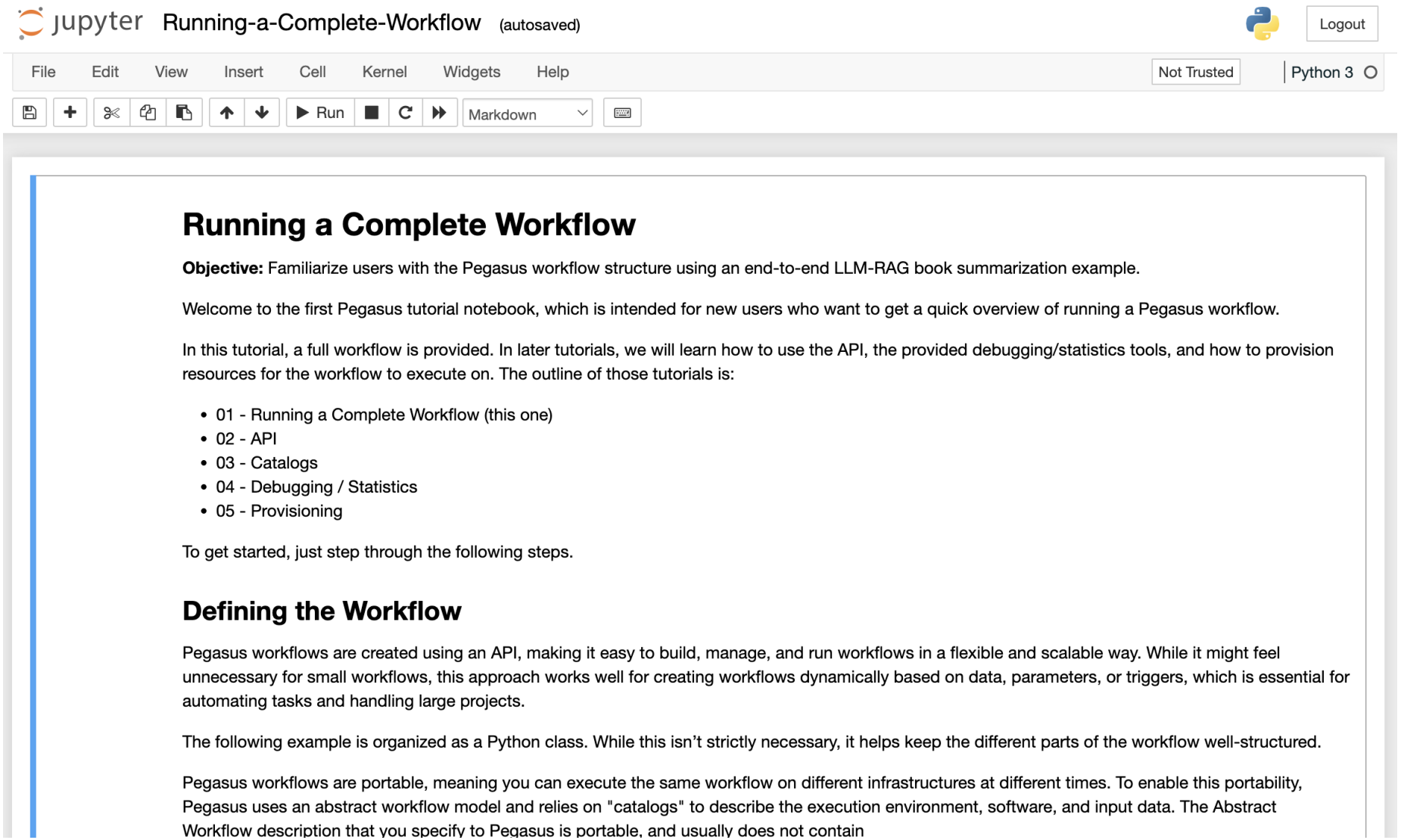
USC Viterbi School of Engineering

Jupyter-Notebook Self-Paced Tutorials and Examples



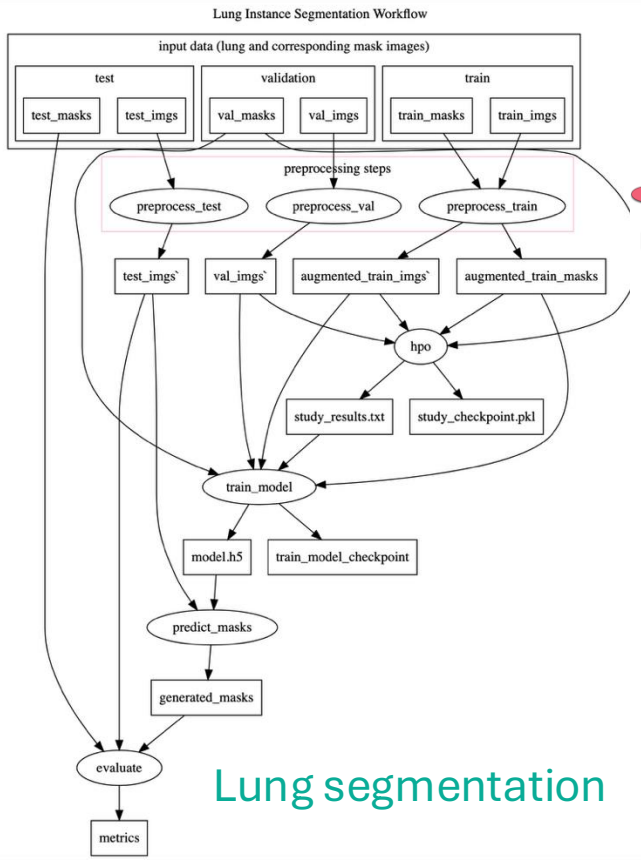


Jupyter-Notebook based Self-Paced Tutorials and Examples



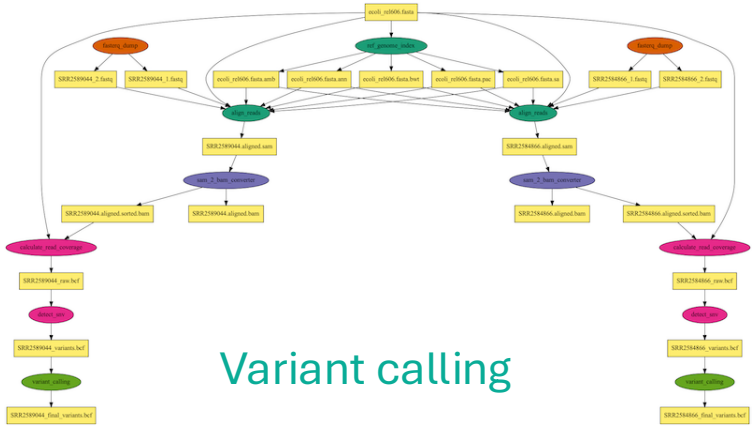
SCITECH

NSF | ACCESS Support

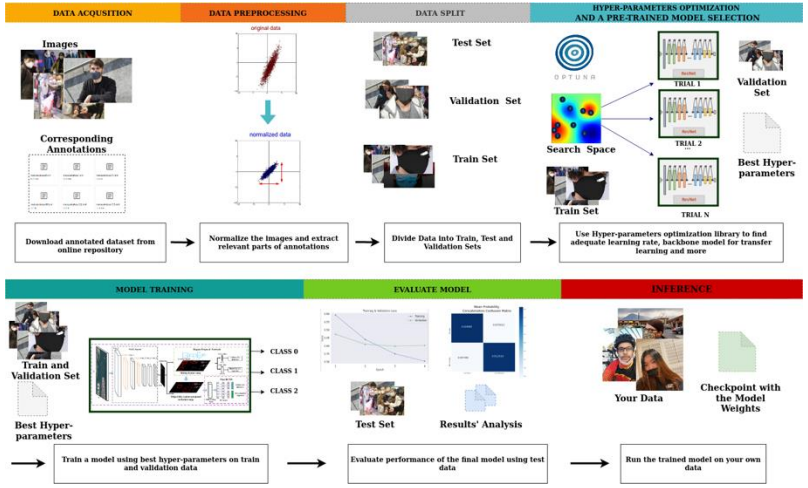


Lung segmentation

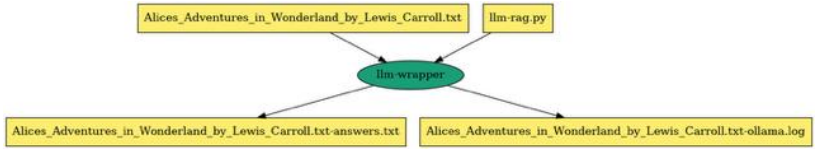
Orca Sound



Variant calling



Mask detection



LLM + RAG

AI Workflows

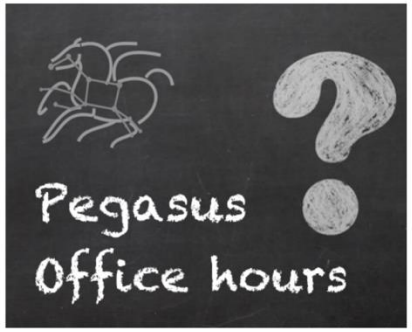
Engage with us!

- Slack channel
- Email: pegasus-support@isi.edu
- Office hours every Friday

Office Hours

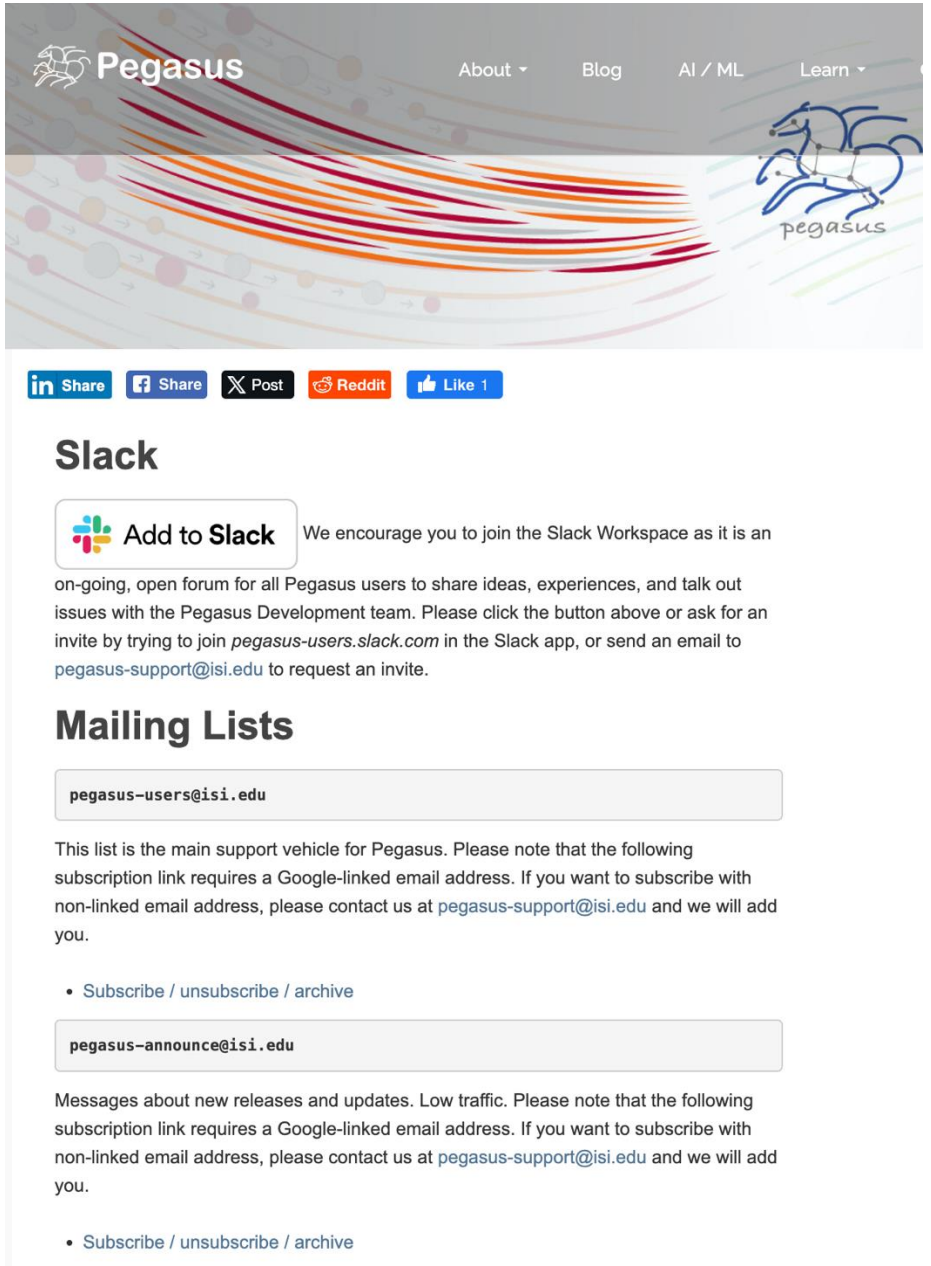
Join the Pegasus team every Friday for virtual office hours at 11 AM Pacific / 2 PM Eastern.

Do you have questions about workflows or need guidance on organizing and implementing them? Join our weekly office hours – designed to support both new and experienced users in learning and engaging with Pegasus. Here's what to expect:



ACCESS

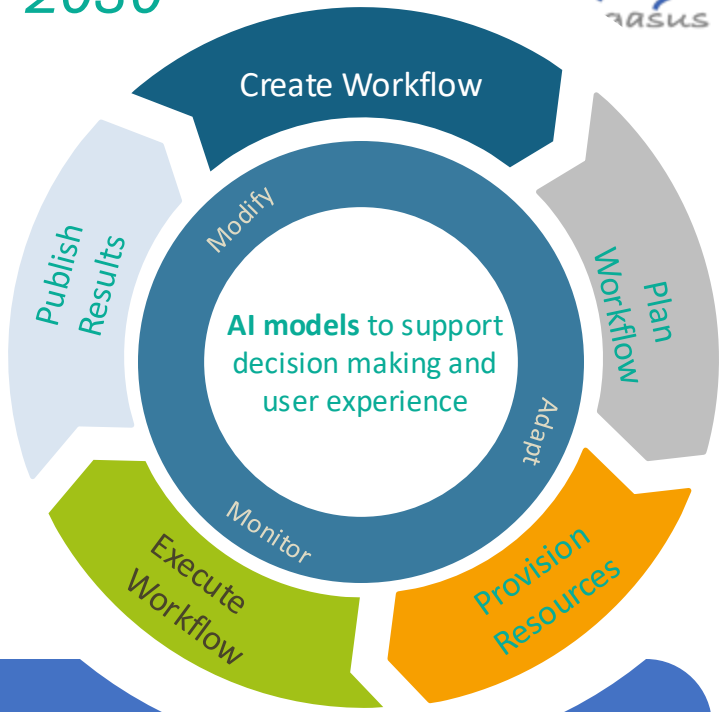
- Tutorial walkthrough First Friday of the month
- http://pegasus.isi.edu



We can help you get started!

USC Viterbi School of Engineering

2025 - 2030



Human-orchestrated decisions

Static scripts, manual scheduling

- WMS, static plans, DAGs
- Predefined execution plans

- Workflow composition
- Resource need and performance prediction
- Anomaly detection
- Dynamic workflow execution adaptation
- Learn about the workloads and systems
- Predict/design systems to serve the workflows

USC University of Southern California | THE UNIVERSITY OF TENNESSEE KNOXVILLE | UMassAmherst | THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

PegasusAI Plans

- **Intelligent Resource Planning**: Uses machine learning models to predict resource needs and optimize workflow execution.
- **Adaptive Workflow Management**: Detects anomalies and performance issues in real time, automatically adjusting plans or alerting users.
- **Human-in-the-Loop Design**: Guides researchers through AI-augmented tools for workflow creation, monitoring, and debugging.
- **Scalable Across CI**: Supports execution on HPC, cloud, and edge platforms, enabling flexible deployment and broad applicability.
- **AI-Ready Data Generation**: Provides curated datasets and trained models to advance AI for scientific computing and CI research.

Next major version of Pegasus will have AI assistant

Pegasus Analyzer

Current version of analyzer pinpoints and highlight failures. Example:

```
2025-09-25 19:10:16,985    INFO:  /bin/cp -f -R -L '/home/rynge/ACCESS-Pegasus-Examples/04-Tutorial-Debugging-Statistics/bin/llm-rag.py' '/home/rynge/ACCESS-Pegasus-Examples/04-Tutorial-Debugging-Statistics/scratch/rynge/pegasus/llm-rag-books/20250925T190853+0000/./llm-rag.py'
2025-09-25 19:10:16,991   ERROR:  Expected local file does not exist: /home/rynge/ACCESS-Pegasus-Examples/04-Tutorial-Debugging-Statistics/inputs/Alices_Adventures_in_Wonderland_by_Lewis_Carroll.txt
```

We are adding a Pegasus AI assistant to help interpret the errors.

Next major version of Pegasus will have AI assistant

========== Pegasus AI Analysis ==========

The workflow failed due to a missing input file. The job `stage_in_local_local_0_0` encountered an error: **"Expected local file does not exist: /path/to/Alices_Adventures.txt"**

Root Cause:
- The required input file is missing from the specified path.
- This prevents the transfer process from completing, causing the workflow to fail.

Next Steps:
1. Verify the file exists at the specified path.
2. Ensure the file path in the workflow configuration matches the actual location.
3. Resubmit the workflow after resolving the file issue.

The remaining unsubmitted jobs (7 total) likely depend on this staged file, so fixing this error will enable further execution.

Pegasus Workflow Management System

Automates the execution of scientific workflows across CI

- Provides a portable workflow description
- Automates repetitive and time-consuming tasks
- Reduces human error
- Increases productivity

Data Management

Pegasus handles data transfers, input data selection and output registration by adding them as auxiliary jobs to the workflow

Error Recovery

Pegasus handles errors by retrying tasks, workflow-level checkpointing, re-mapping and alternative data sources for data staging

Provenance Tracking

Pegasus allows users to trace the history of a workflow and its outputs, including information about data sources and softwares used

Heterogeneous Environments

Pegasus can execute workflows in a variety of distributed computing environments such as HPC clusters, Amazon EC2, Google Cloud, Open Science Grid or ACCESS

USC Viterbi School of Engineering

CI Compass Fellowship Program (CICF)

For Undergrads! and Faculty mentors

Goal: Broaden student participation in CI research, development, deployment, and operations

Virtual Spring Program

- Free to undergraduate students. Possibility of course credit.
- *Technical Skills Component*: Students are taught technical skills relevant to CI.
- *Data Lifecycle Component*: Students research MFs and the data lifecycle to understand the importance and context of MFs, and the related data and CI. They present their results at the end of the Spring Program.

(Optional/Invited) Summer Program

- We collaborate with MFs to provide CI-related summer projects for some of our student fellows.
- In-person or virtual, depending on the MF/project.
- Students are paid for their participation.

https://ci-compass.org/

Major Facilities are large-scale NSF-funded science projects: Telescopes, Research Vessels, Ecological and Ocean Observatories

SCITECH

More Information About the Projects:

pegasus

Pegasus coming to Laguna

- Pegasus: http://pegasus.isi.edu
- Pegasus support: https://pegasus.isi.edu/contact/
- PegasusAI: http://pegasusai.io
- ACCESS Support: https://support.access-ci.org/
- CI Compass: https://ci-compass.org

NAIRR AI Workshop at USC January 22, 2026

https://nairrpilot.org/

pegasus@isi.edu

U.S. DEPARTMENT OF ENERGY

Big Thanks to DOE and NSF for the support!

NSF U.S. National Science Foundation

USC Viterbi School of Engineering