WORKS@20: The Evolution of Automation in Science — The Pegasus Perspective

Ewa Deelman
University of Southern California

The Workshop on Workflows in Support of Large-Scale Science (WORKS06)

in conjunction with HPDC06
Paris, June 20th, 2006

Message from the Chair

Starting in 2008
At SC

WORKS '09: Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science

Proceeding

2009 Proceeding

Conference Chairs: Ewa Deelman, Ian Taylor

Publisher: Association for Computing Machinery, New York, NY, United States

Conference: SC '09: International Conference for High Performance Computing, Networking, Storage and Analysis • Portland Oregon • 16 November 2009

USC Viterbi
School of Engineering
Information Sciences Institute

June 2006

July 2025

WORKS CHAIRS


Ian Taylor


Johan Montagnat


Sandra Gesing


Rafael Ferreira da Silva


Rosa Filgueira


Anirban Mandal


Silvina Caino-Lores


David Abramson
University of Queensland, Australia


Malcolm Atkinson
University of Edinburgh, UK


Michela Taufer
University of Tennessee, USA


Ewa Deelman
University of Southern Californi

Keynote at WORKS'15

USC VITERBI SCHOOL OF ENGINEERING

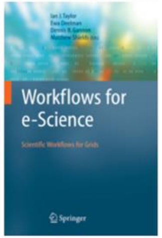
pegasus

Scientific Workflows
Past and Current Issues

Ewa Deelman
USC Information Sciences Institute

http://pegasus.isi.edu deelman@isi.edu

Challenges circa 2006



- Hiding the complexity of the execution environment
  - Include better error descriptions
  - Better fault tolerance
  - Debugging tools
- Real time interaction with workflows
  - inspecting and modifying a running workflow
- Workflow sharing and reuse
  - Workflow and component libraries
- Result validation, verification, reproducibility
  - Provenance provides part of the answer

Challenges ca 2006 cont'd

pegasus

- Workflow composition/editing
  - Hard to compose workflows for a novice
- Workflow compilers
  - Need for late-binding
- Workflow Execution
  - Common engine (or a set of engines)
- Workflow Interoperability

Challenges revisited

- Hiding the complexity of the environment
  - Include better error descriptions
  - Better fault tolerance
  - Debugging tools
- Real time interaction with workflows
  - Is it needed?
- Workflow sharing and reuse
  - myExperiment (tied to a particular workflow system)
- Result validation, verification, reproducibility
  - Much work done in this area (Provenance challenges, OPM, W3C working group)

Keynote at WORKS'15

Challenges revisited

- Workflow composition/editing
  - Semantics-based composition
- Workflow compilers
  - Need for late-binding (yes in Grids, but clouds?)
- Workflow Execution
  - Common engine (or a set of engines)
- Workflow Interoperability
  - EU SHIWA project (www.shiwa-workflow.eu/)
- New Challenges and Opportunities: workflows on the cloud

Progression of Automation

Pegasus
Computation automation

Pegasus AI
Infusing AI techniques

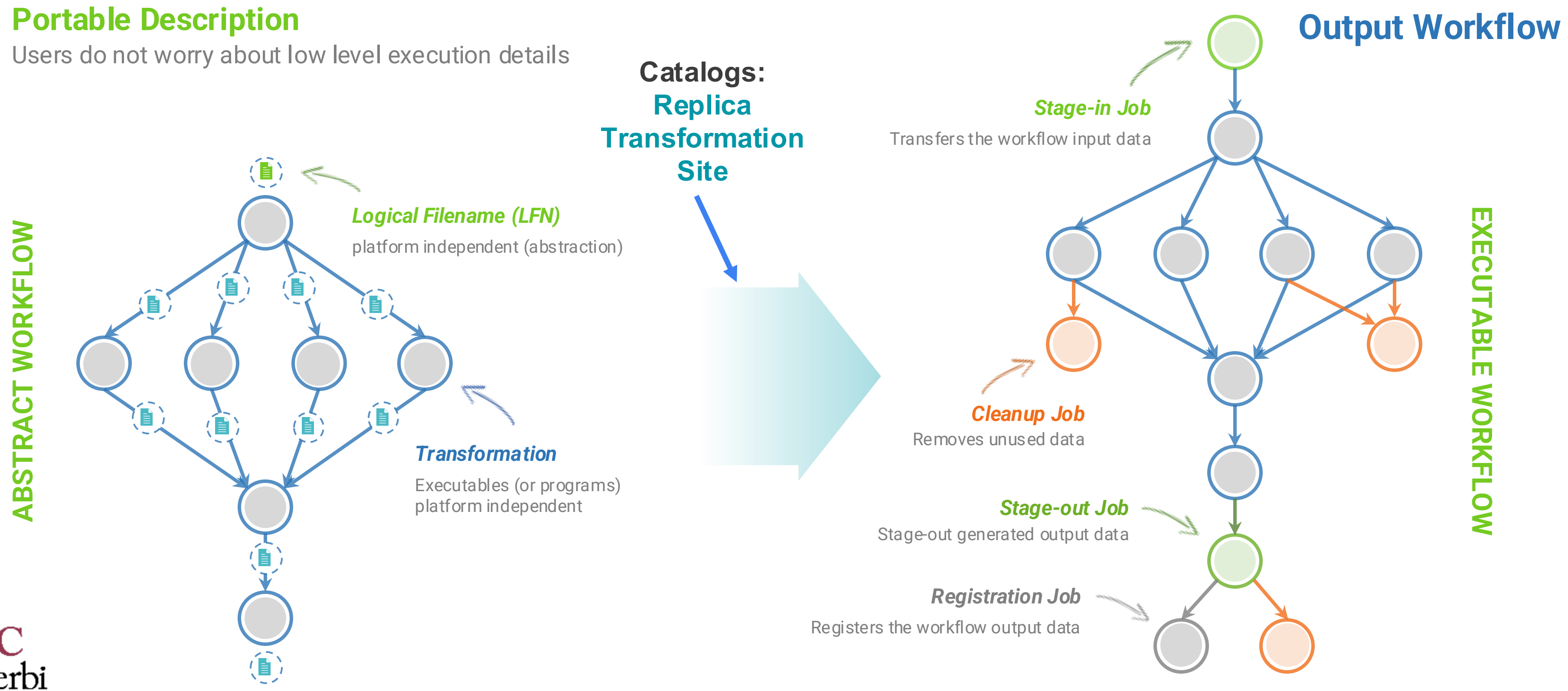
Agentic Workflows
Based on swarm intelligence

Self-driving Labs
Automation of experimental workflows

Resource-independent Specification

Input Workflow Specification YAML formatted

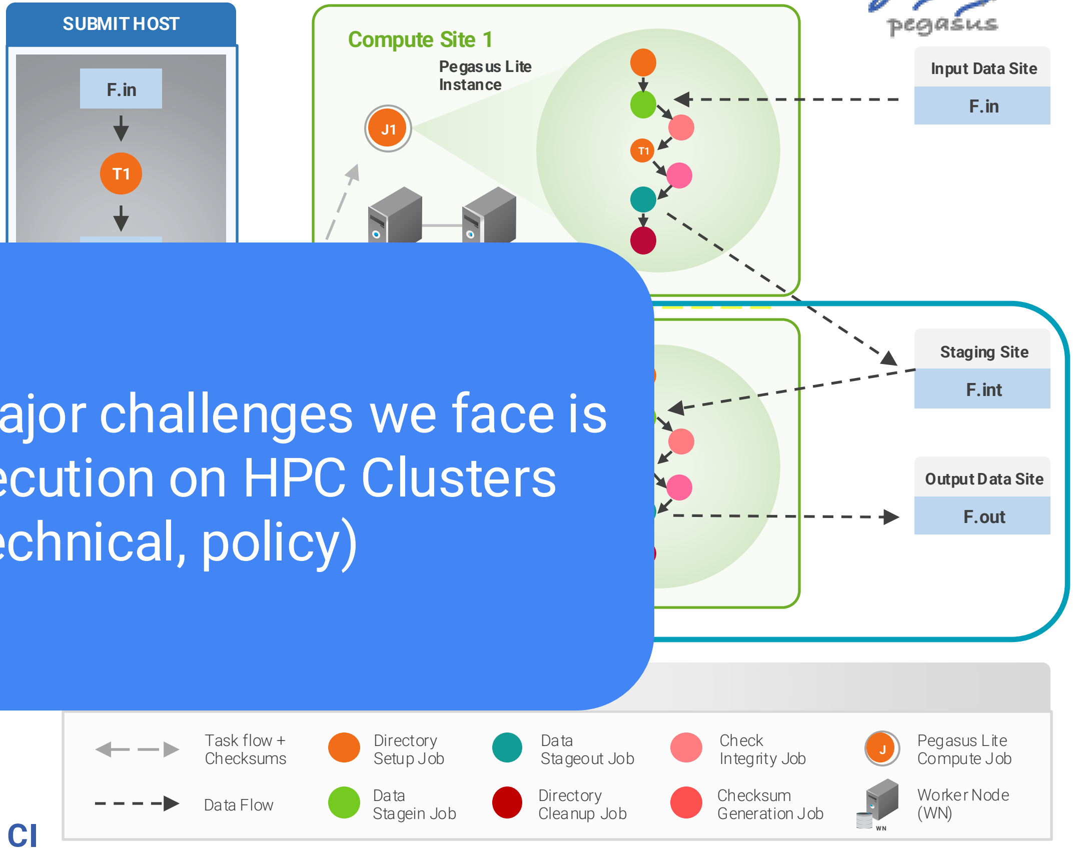
directed-acyclic graphs



USC Viterbi School of Engineering
Information Sciences Institute

Submit locally, run globally

- **Pegasus WMS ==** **Pegasus planner (mapper) + DAGMan workflow engine + HTCondor scheduler/broker**
  - Pegasus maps workflows to target infrastructure more resources)
  - DAGMan manages dependencies and
  - HTCondor is used broker to interface different schedule
- **Planning converts ar into a concrete, exec**
  - Planner is like a compi
  - Optimized performance
  - Provides fault tolerance
- **Can leverage distributed and heterogeneous CI**



One of the major challenges we face is remote execution on HPC Clusters (technical, policy)

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, AWS S3, GCP, 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

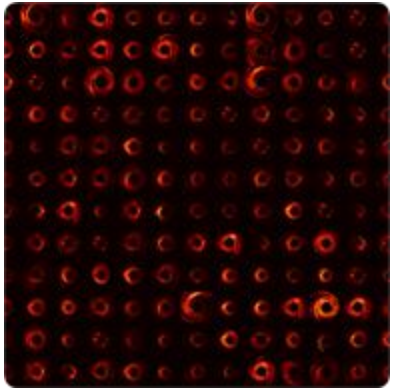
Pegasus keeps track of how the result was obtained: Full Provenance, support for containers

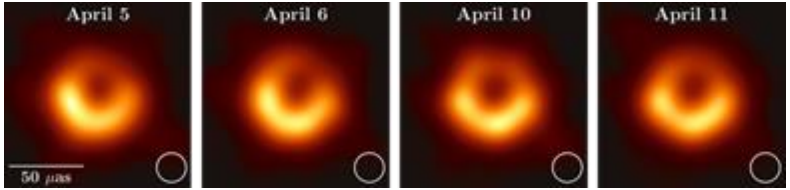
Event Horizon Telescope

Bringing Black Holes into Focus

8 telescopes: 5 PB of data

60 simulations: 35 TB data





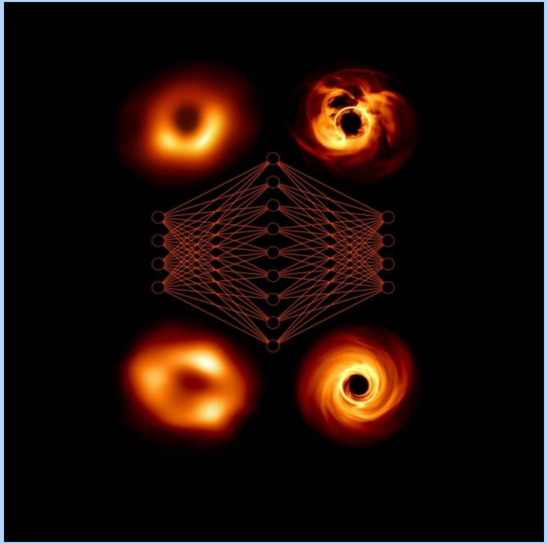
2019

First images of black hole at the center of the M87 galaxy

Improve constraints on Einstein's theory of general relativity by 500x

Michael Janssen (Radboud University, NL)

- trained a neural network with millions of synthetic black hole data sets
- used this and observations to predict that the black hole at the center of our Milky Way is spinning at near top speed

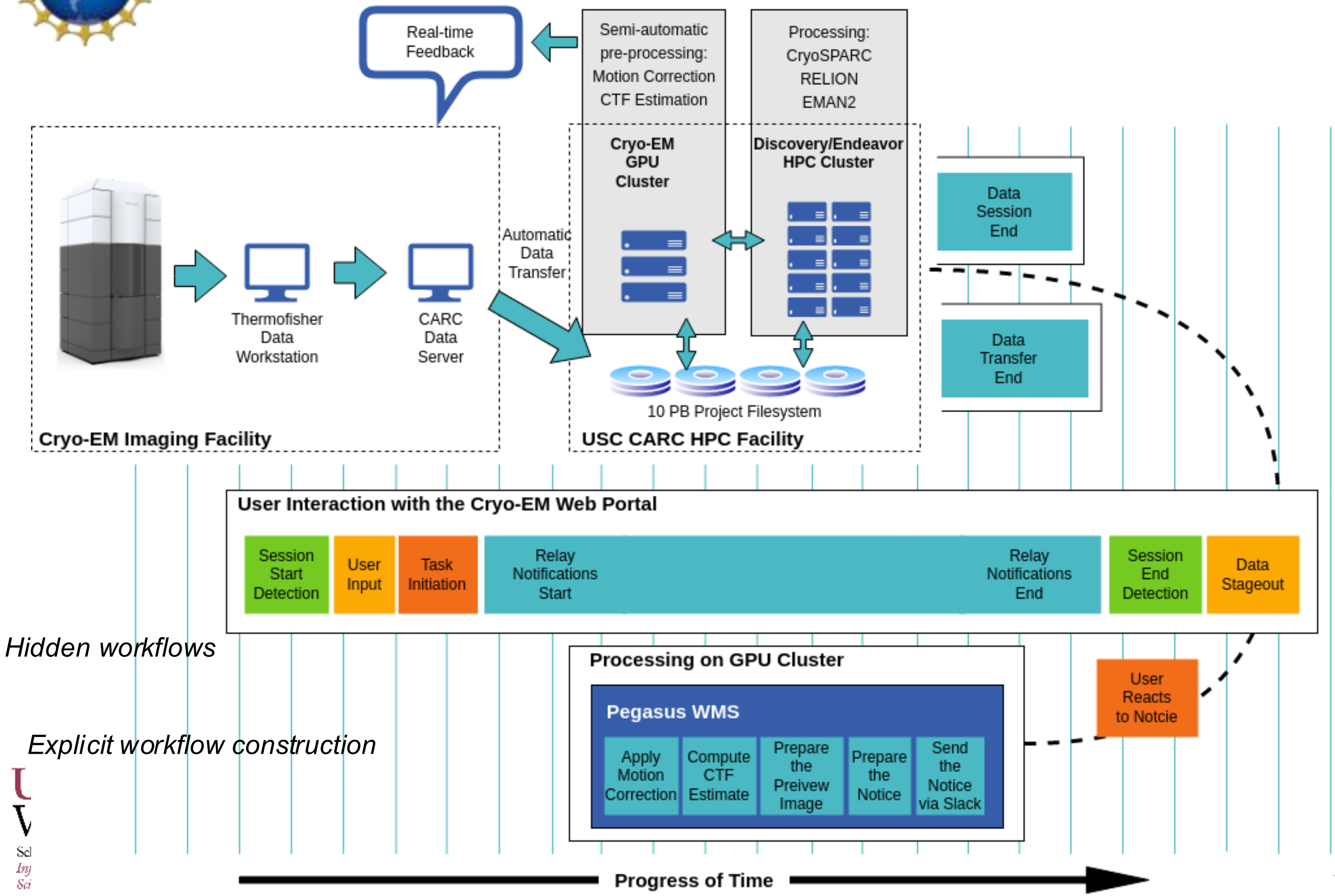


2025

Artist impression of a neural network that connects the observations (left) to the models (right)

Deep learning inference with the Event Horizon Telescope I. Calibration improvements and a comprehensive synthetic data library. By: M. Janssen et al. In: Astronomy & Astrophysics, 6 June 2025.

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**

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

2025 - 2030

Manual Workflows

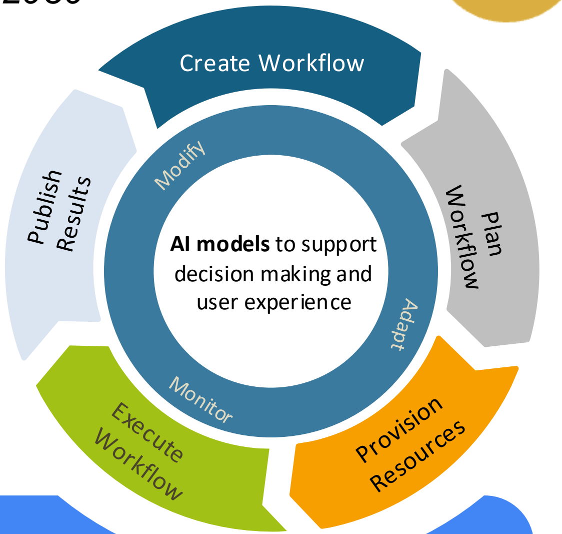
Automated Workflows

AI-Augmented Workflows

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 | UMass Amherst | The University of North Carolina at Chapel Hill

PegasusAI Team

Front row: Komal Thareja, Sai Swaminathan, Michela Taufer, Ewa Deelman, Mike Zink, Ty Anderson, Kin H. Ng
Back row: Michael Sutherlin, Mats Rynge, Karan Vahi, Berent Aldikacti, Ian Lumsden, Micheal Stealey, Kin W. Ng, Dan Scott

USC Viterbi School of Engineering Information Sciences Institute
UMassAmherst
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
THE UNIVERSITY OF TENNESSEE KNOXVILLE

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.

Technical approach: Hierarchical AI Agents, Hybrid Learning Models, Runtime Monitoring & Feedback Loops, Failure Prediction & Resilience Strategies, Adaptive Scheduling, Workflow-Level Summarization, CI-Ready Design

Provenance Capture in Pegasus

Kickstart Process:

Execution Wrapper

Type	Succeeded	Failed	Incomplete	Total	Retries	
Tasks	5	0	0	5	0	5
Jobs	16	0	0	16	2	18
Sub Workflows	0	0	0	0	0	0

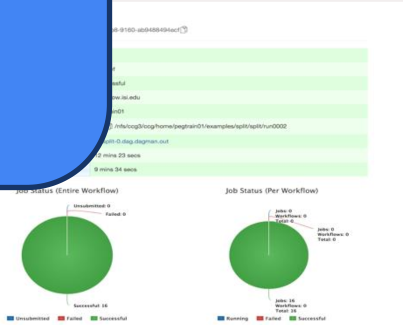
Detailed Runtime Provenance

arguments, and software
versions

Centralized Provenance Database

erying and
on of the
orkflow
story

a flow, identify
es, debug
verify results



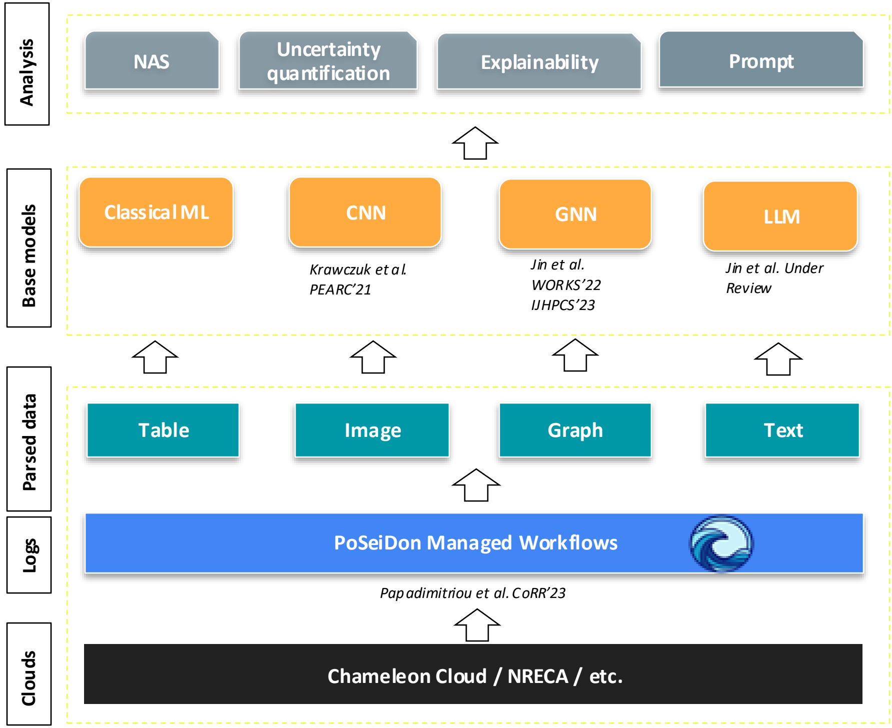
We can use AI to look for anomalies in the executions

We can trace back faulty results

Comprehensive provenance capture transforms complex scientific computations into fully auditable, traceable processes, significantly reducing the burden on researchers for manual record-keeping and enabling unprecedented levels of reproducibility and trust in results.

AI for Execution Anomaly Detection

- **Data processing**: process simulated anomalies on workflows, parse logs as
  - **Tabular** (features as columns)
  - **Image** (Gantt charts)
  - **Graph** (nodes as jobs, edges as dep.)
  - **Text** (sentences describing jobs)
- **Build base models**: supervised / unsupervised learning to identify the anomalies by deep learning
- **Analytics**: improve the performance, quantify uncertainty, provide explanation, etc.



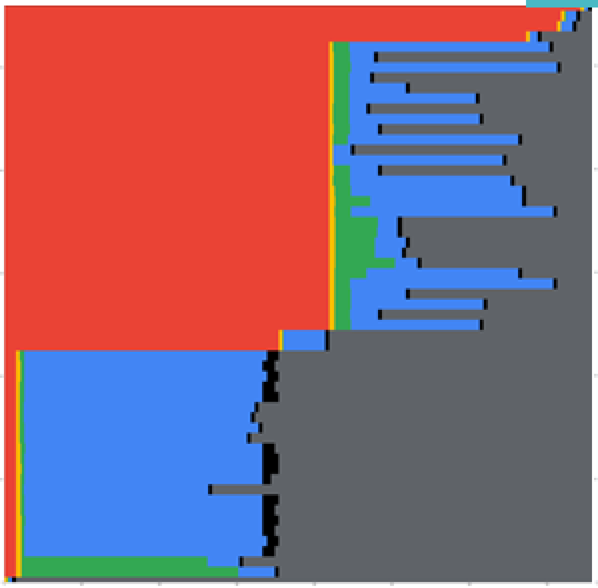
Anomaly Detection Framework

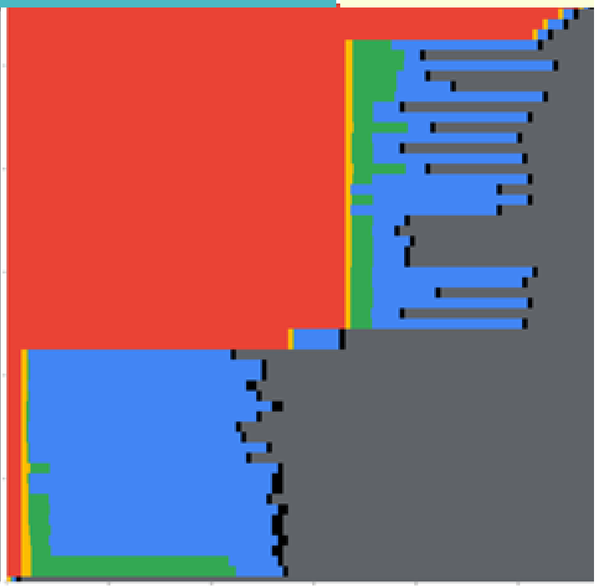
USC Viterbi School of Engineering · renci · Argonne National Laboratory · Berkeley Lab · Oak Ridge National Laboratory · http//poseidon-workflows.org · PoSeiDon

Identifying anomalies and their causes

Gantt Charts: normal execution and different anomalies: hard drive load, network packet loss

Work by Patrycja Krawczuk and George Papadimitriou





normal_1000genome-20200616T174351Z-run0044.png

hdd_50_1000genome-20200610T041238Z-run0006.png

loss_0.5_1000genome-20200520T031010Z-run0017.png

ready_delay | wms_delay | queue_delay | runtime | post_script_delay | finished

90% accuracy on the workflows we trained on

Graph Neural Networks - performance

Available workflows

Single model for multi-workflows

Workflow	Binary				Multi-label
	Accuracy	F1	Recall	Precision	Accuracy
1000 Genome	0.917 ± .014	0.915 ± .019	0.921 ± .009	0.938 ± .010	0.882 ± .006
Nowcast w/ clustering 8	0.768 ± .009	0.715 ± .017	0.778 ± .023	0.768 ± .15	0.792 ± .009
Nowcast w/ clustering 16	0.837 ± .012	0.675 ± .020	0.815 ± .012	0.837 ± .011	0.830 ± .007
Wind w/ clustering casa	0.776 ± .002	0.652 ± .032	0.769 ± .021	0.776 ± .017	0.764 ± .19
Wind w/o clustering casa	0.781 ± .02	0.853 ± .013	0.800 ± .012	0.781 ± .008	0.886 ± .007
1000 Genome (partial anomaly)	1.000 ± .0	1.000 ± .0	1.000 ± .0	1.000 ± .0	1.000 ± .0
ALL	0.836 ± .006	0.878 ± .013	0.886 ± .011	0.856 ± .009	0.877 ± .008

Figure: Graph-level classification

Gantt Chart

Model	Acc.	Recall	Prec.	F1
SVM	0.622	0.622	0.667	0.550
MLP	0.874	0.874	0.875	0.874
RF	0.898	0.898	0.908	0.887
AlexNet	0.910	0.914	0.910	0.910
VGG-16	0.900	0.900	0.900	0.900
ResNet-18	0.910	0.916	0.910	0.910
Our GNN	0.917	0.921	0.939	0.915

SVM: Support vector machines (SVMs)
MLP: Multilayer perceptron with hidden layers (128, 128, 128)
RF: Random forest with maximum depth set to 3.
(AlexNet,…) Gantt Chart: computer vision inspired DNN by generating Gantt charts from node features.

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 ===============================

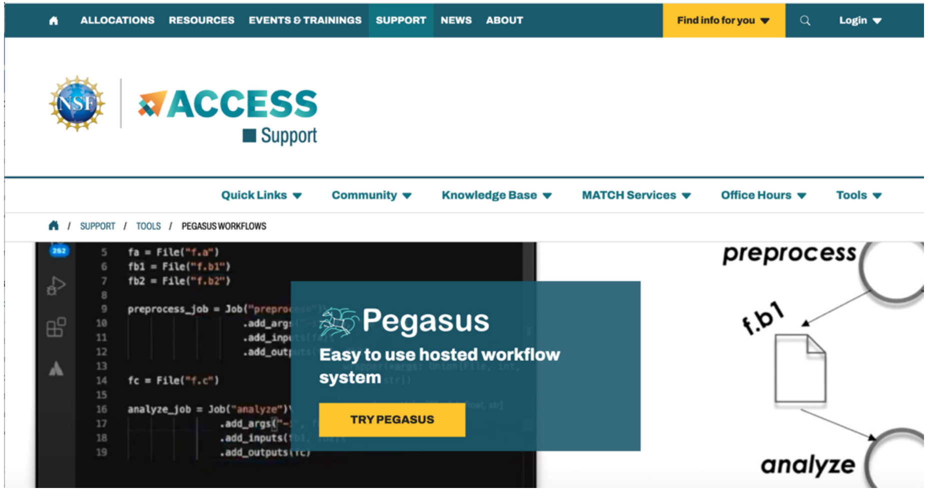
The workflow failed due to a missing input file. The job `stage_in_local_local_0_0` encountered an error:
"Expected local file does not exist: /path/to/Alices_Adventures.txt"

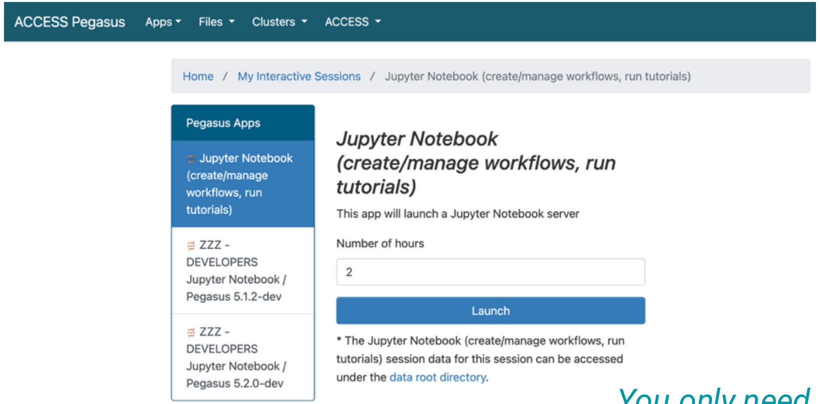
Root Cause:
- The required input file is missing from the specified path.
- This prevents the transfer process from completing, causing the workflow to fail.

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.

To Try Production or Dev Pegasus





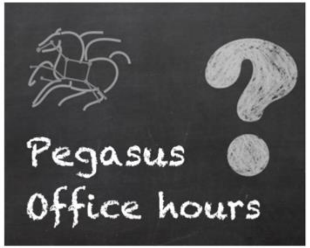
You only need a free ACCESS account

- 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:



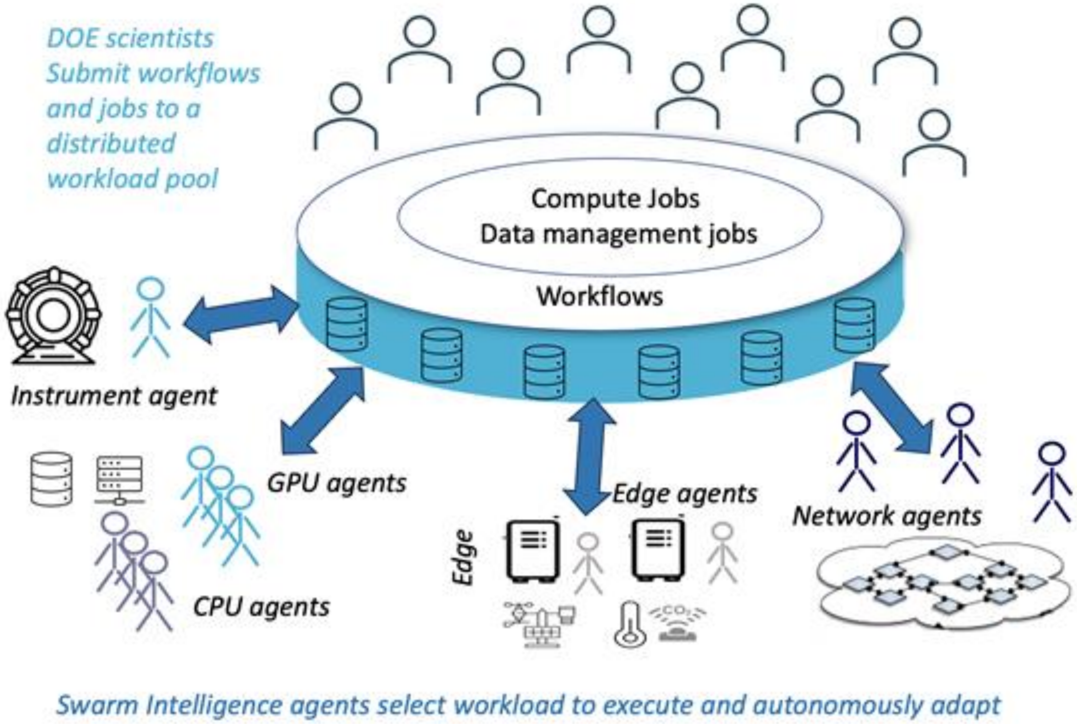
- Tutorial walkthrough First Friday of the month
- http://pegasus.isi.edu

We can help you get started!

Funded by NSF under Grant # 2138286

SWARM: Scientific Workflow Applications on Resilient Metasystem

SWARM aims to improve resilience by employing multi-agent approach



Funded by DOE: DE-SC0024387

2023- 2028

USC Viterbi School of Engineering, renci, Argonne National Laboratory, Berkeley Lab, Oak Ridge National Laboratory, UC Riverside

SWARM team

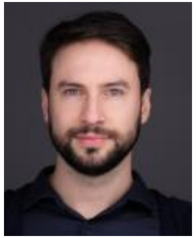

Ewa Deelman, Ph.D.
USC


Prasanna Balaprakash, Ph.D.
ORNL


Anirban Mandal, Ph.D.
RENCI


Krishnan Raghavan, Ph.D.
ANL


Franck Cappello, Ph.D
ANL


Jean Luca Bez, Ph.D
LBNL


Imtiaz Mahmud, Ph.D.
LBNL


Zizhong Chen, Ph.D.
UCR


Erik Scott
RENCI


Hongwei Jin, Ph.D.
ANL


Cong Wang, Ph.D.
RENCI


Komal Thareja
RENCI


Shixun Wu
UCR


Hong-Jun Yoon, Ph.D.
ORNL

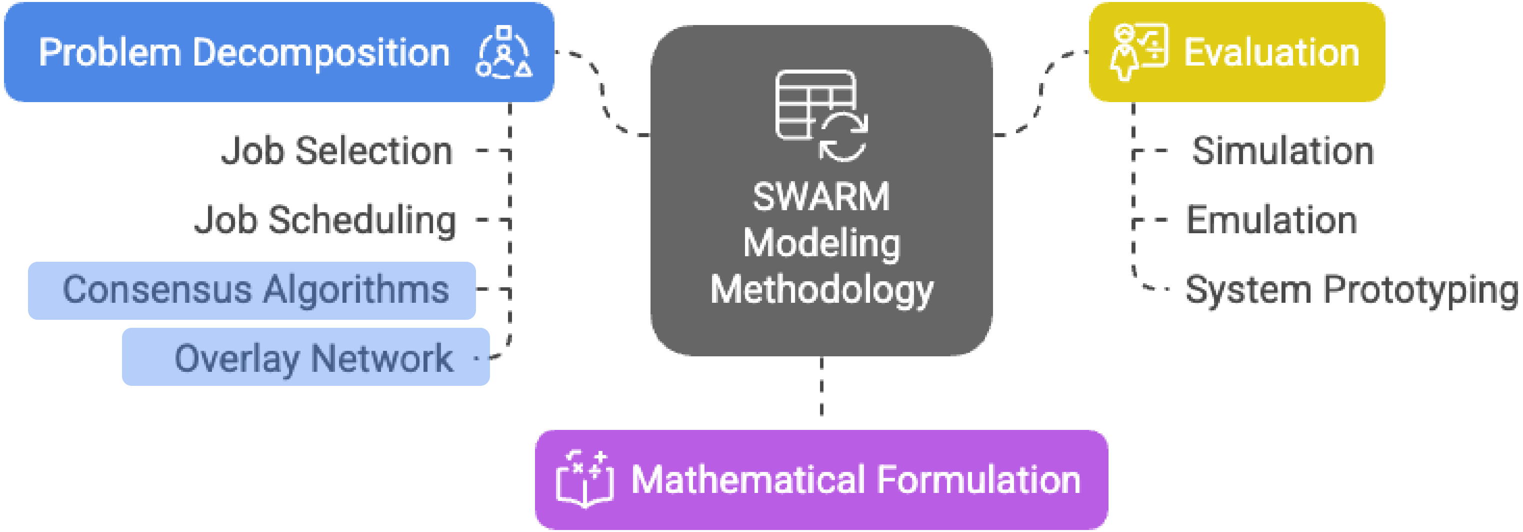

Sheng Di, Ph.D.
UCR


Suman Raj
USC


Aiden Hamade
ORNL


Prachi Jadhav
ORNL

SWARM Methodology



SWARM Consensus Algorithm for Job Selection

Anirban Mandal
Komal Thareja
RENCI

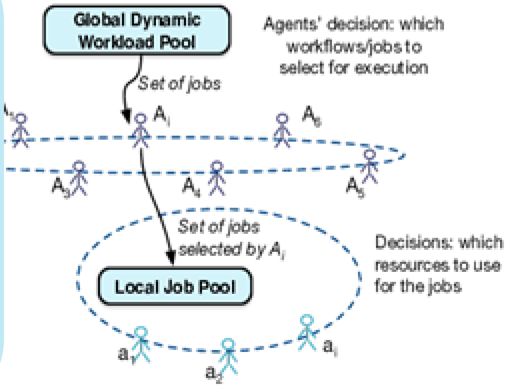
Our Approach: Multi-Agent Systems (MAS) for Resilient Job Selection

- Globall
- ***Nove***
  consi
- ***Gree***
  Toler
  resilie
- Conse
  -
  -
  - *Commit:* a quorum of confirmations finalizes the decision
- All agents communicate with all other agents

Assumptions:

1. Each agent knows the capabilities and workload of other agents and can compute their job selection
2. Each agents communicates with each other

Relaxing the assumption:
Agents can learn each other's capabilities over time, and potentially anticipate their selections



$$h_i = \begin{cases} \text{current load}_i + \\ \text{feasibility}_i \times \text{projected load}_i, & \text{feasible job} \\ \infty & \text{infeasible job} \end{cases}$$

- **Improved scheduling latency by 63.5 %**
- **Improved idle time by 63.8 % compared to PBFT**

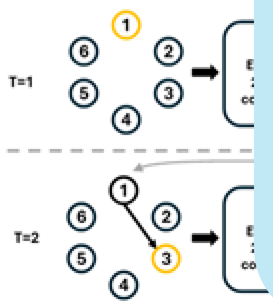
SWARM Overlay Network

Franck Cappello
Shixun Wu
ANL, UCR

- **Motivation**
  - Existing membership protocols use logical ring, not considering underlying **physical latency.**
  - Consensus on membership is upper bounded by the **diameter of the overlay topology.**
- **Challenge:** Degree-constrained diameter minimization is an NP-hard problem.
- **Our Contributions**
  - Diame… …sed overlay const…

Multi-agent systems introduce many security challenges

- Identity and trust
- Tool and capability abuse
- Content-borne attacks (agents leak secrets)
- Data & model integrity (corrupt facts enter the system)
- More efficient DoS



Our Node Selection with Deep Q-Network.

…ibuted sites

Fabric Testbed: https://portal.fabric-testbed.net/

Action: Selecting the next node to connect.
Reward: Reduction in network diameter between consecutive steps, with an additional latency penalty/bonus to encourage low-latency links
Q-function: A neural network estimates the expected future reward of connecting the current node to candidate node

K-Ring constructed by DGRO outperforms Chord, Nearest Neighbour, Rapid, Perigee.

Cloud Labs and Self-Driving Labs

CloudLab at CMU

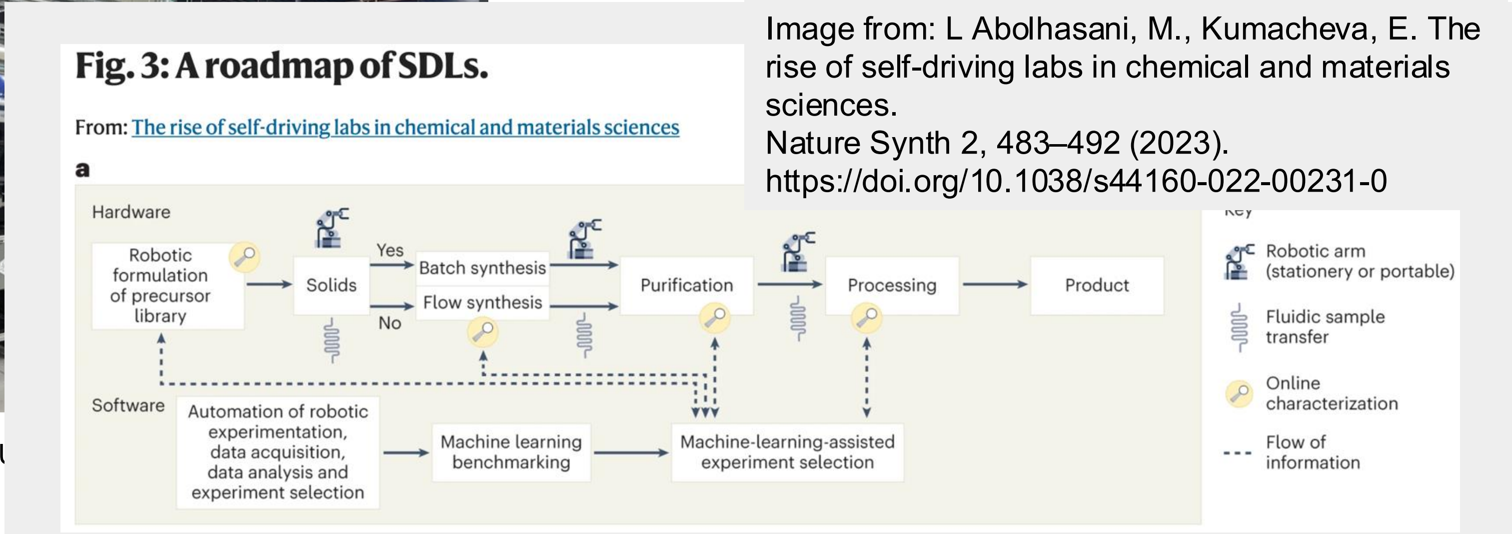


Image from: L Abolhasani, M., Kumacheva, E. The rise of self-driving labs in chemical and materials sciences.
Nature Synth 2, 483–492 (2023).
https://doi.org/10.1038/s44160-022-00231-0

Computational Workflow Systems for Automated Labs

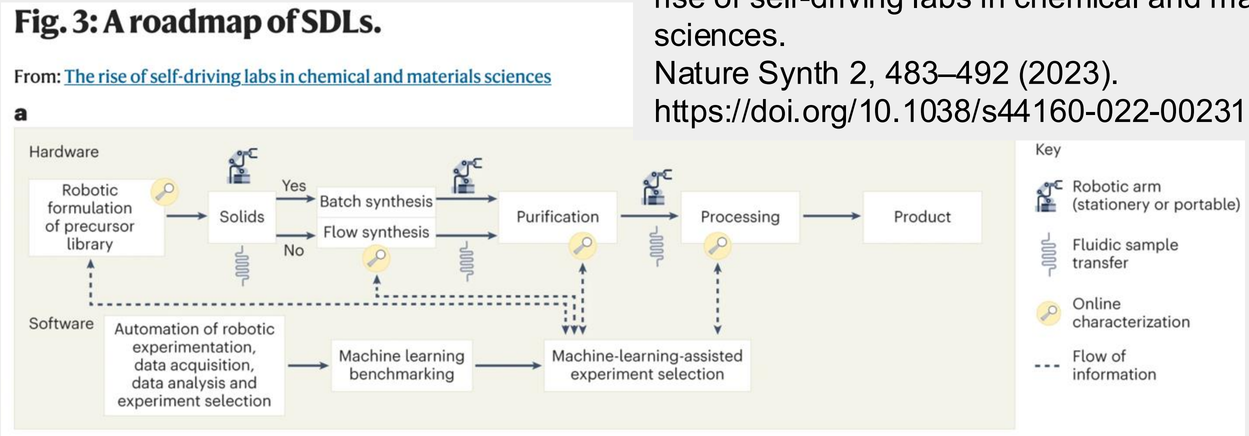


Image from: L Abolhasani, M., Kumacheva, E. The rise of self-driving labs in chemical and materials sciences.
Nature Synth 2, 483–492 (2023).
https://doi.org/10.1038/s44160-022-00231-0

Computational Workflow Management System

- Predict results of reactions and check whether safe, already performed and data is available, ...
- Run ahead of the experimental workflow and re-evaluate predictions
- Assimilate other relevant data along the way
- Collect and annotate intermediate and final data
- Collect and analyze data about failures
- Further process the results and deposit in community repositories

SWARM for Scientific Workflows at an Automated Lab

Instrument

Agent

Checks instrument status, Checks data quality, triggers pre-processing

Edge Cluster

Agent

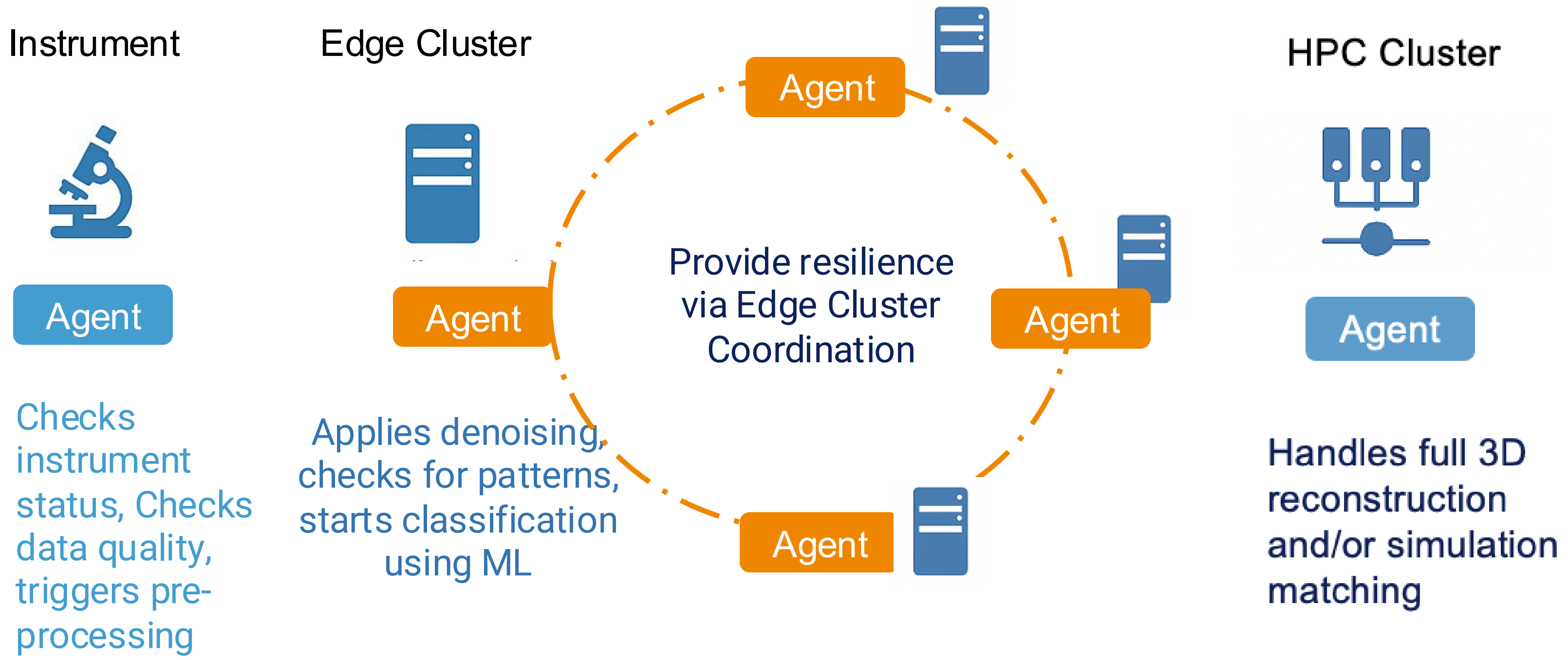
Applies denoising, checks for patterns, starts classification using ML

HPC Cluster

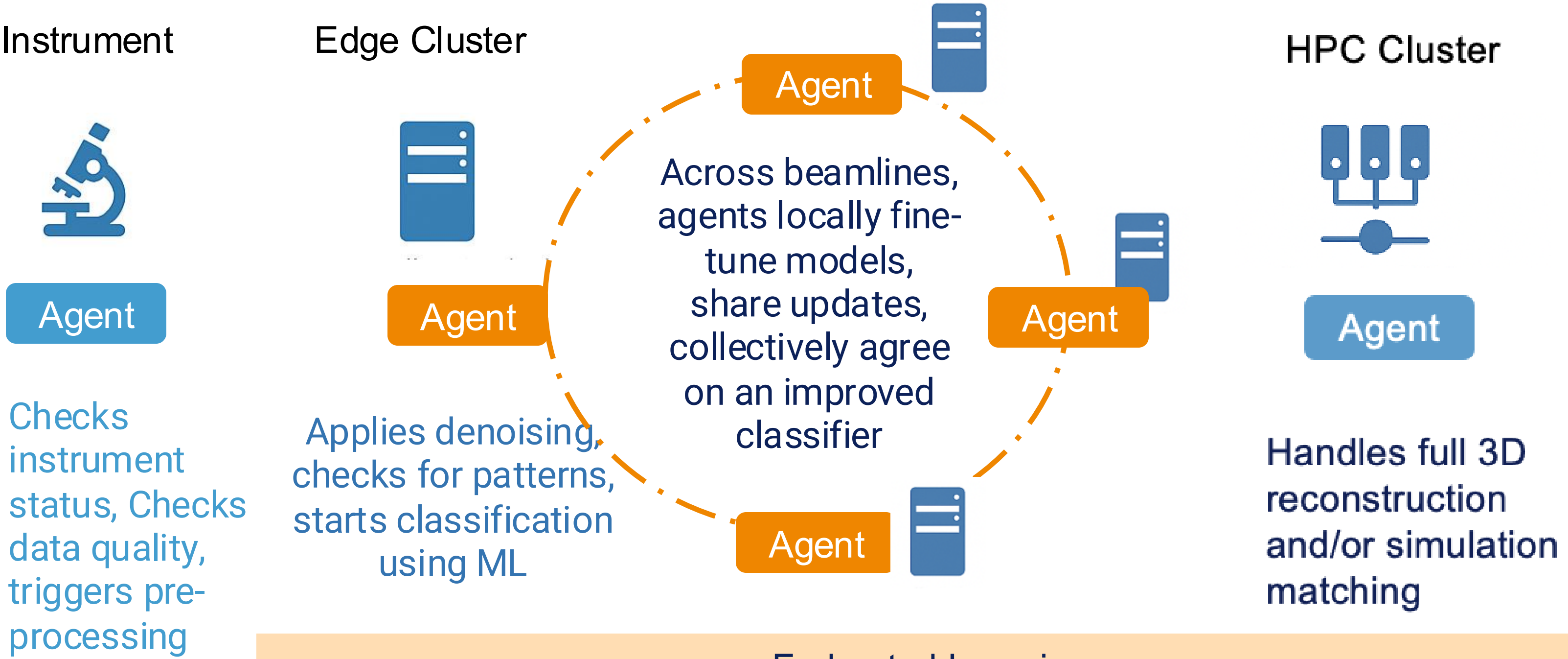
Agent

Handles full 3D reconstruction and/or simulation matching

SWARM for Scientific Workflows at an Automated Lab

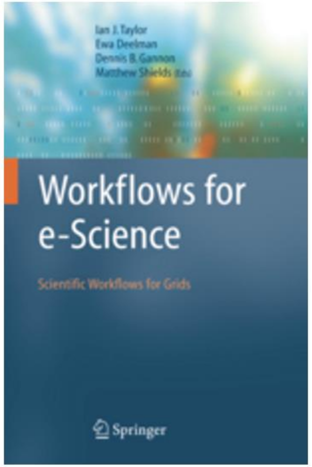


SWARM for Scientific Workflows at a User Facility



Federated Learning:
Local model training, peer-to-peer exchange of updates, decentralized consensus

Challenges circa 2006



- Hiding the complexity of the execution environment
  - Include better error descriptions
  - Better fault tolerance
  - Debugging tools
- Real time interaction with workflows
  - inspecting and modifying a running workflow
- Workflow sharing and reuse
  - Workflow and component libraries
- Result validation, verification, reproducibility
  - Provenance provides part of the answer

USC
Viterbi
School of Engineering
Information Sciences Institute

Challenges ca 2006 cont'd

- Workflow composition/editing
  - Hard to compose workflows for a novice
- Workflow compilers
  - Need for late-binding
- Workflow Execution
  - Common engine (or a set of engines)
- Workflow Interoperability

Conclusions

Automation enables significant science breakthroughs

- AI is bringing significant opportunities for automation
  - We can improve the entire CI stack, all the way up to workflows and applications
  - We need to deal with issues of correctness, efficiency (performance, **resource costs**) – simple methods may be better in some cases
  - We need AI curation, verification and validation methods
  - Cybersecurity risks are increasing, but cybersecurity methods can improve as well
- Agentic Frameworks can benefit from traditional CS methods, increase cybersecurity risks, they can take unpredictable actions
- Automation of physical experimentation can generate more ideas for experiments
  - Challenges are similar to challenges with CI but additional safety issues come into play

USC Viterbi School of Engineering Information Sciences Institute

U.S. DEPARTMENT OF ENERGY

http://pegasus.isi.edu
https://pegasusai.io/
https://swarm-workflows.org/