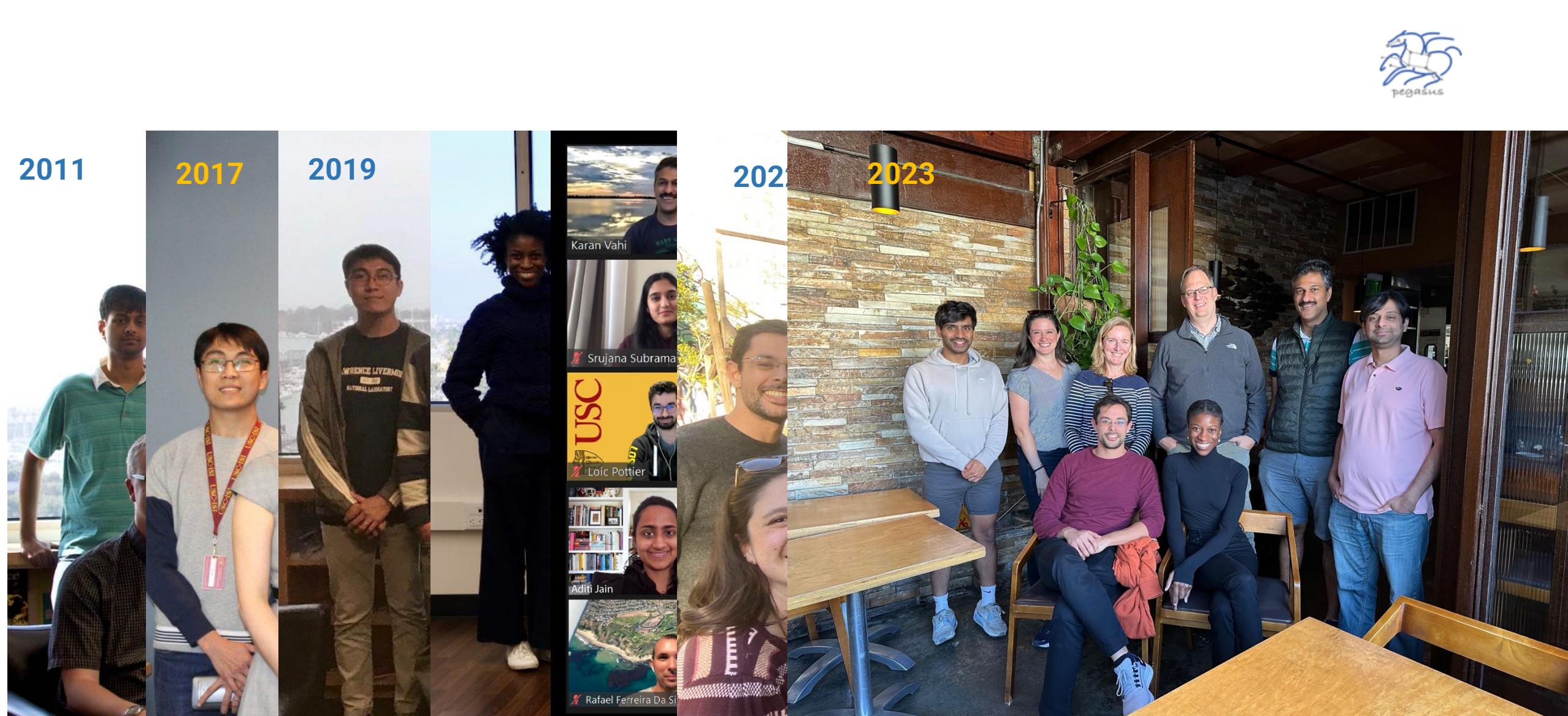


Abstraction and Automation: From Workflows to Intelligent Systems and the Future of Scientific Discovery

Ewa Deelman

University of Southern California, Viterbi School of Engineering
Information Sciences Institute
deelman@isi.edu





SC'17 Denver



SC'18 Dallas

Thank you Caroline and
Nicholas!!
Thank you Peter!



SC'19 Denver



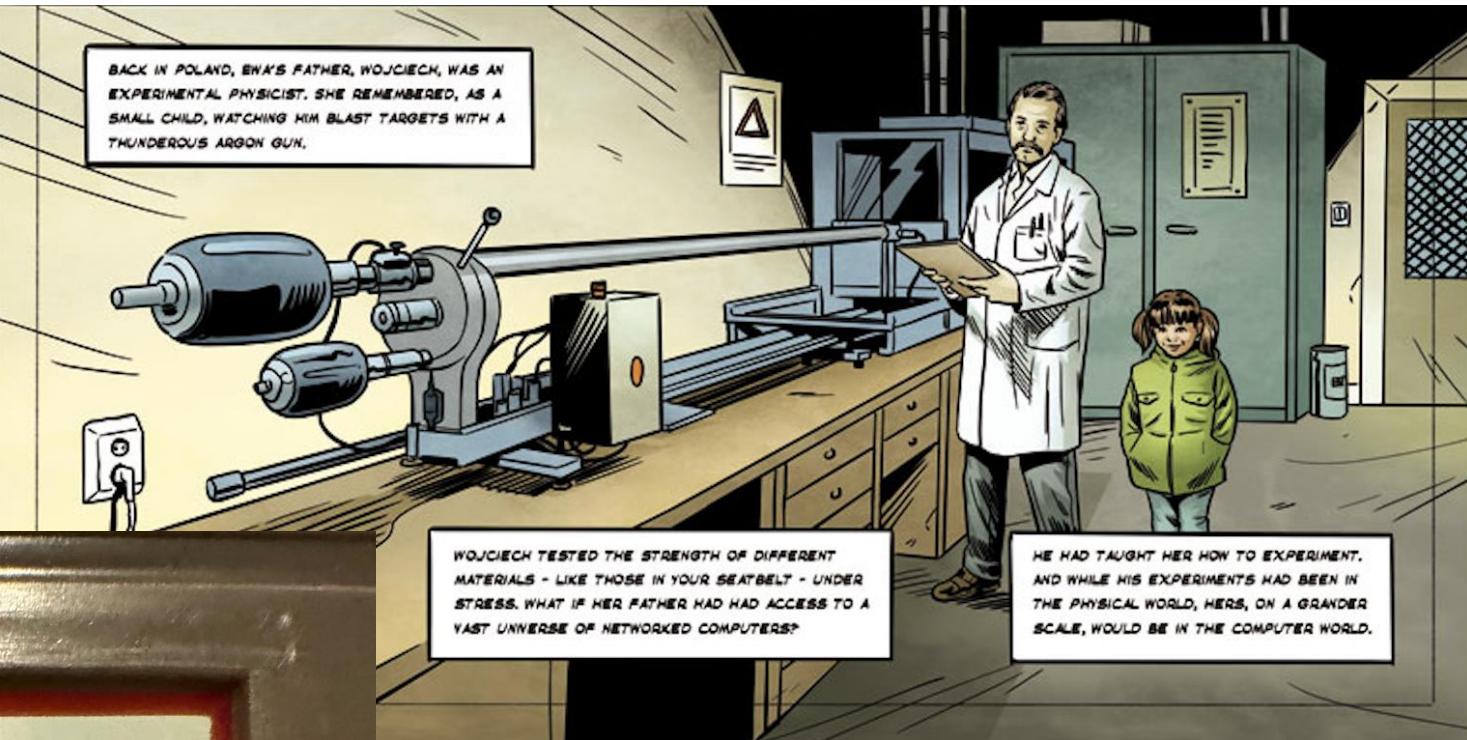
Cornell University
July 2025



Wojciech Nowacki
Nicholas and Caroline Deelman
June 2009



Wojciech Nowacki, c. 1974



Wojciech Nowacki and Ewa Nowacka,
USC Viterbi graphic novel, 4/2018

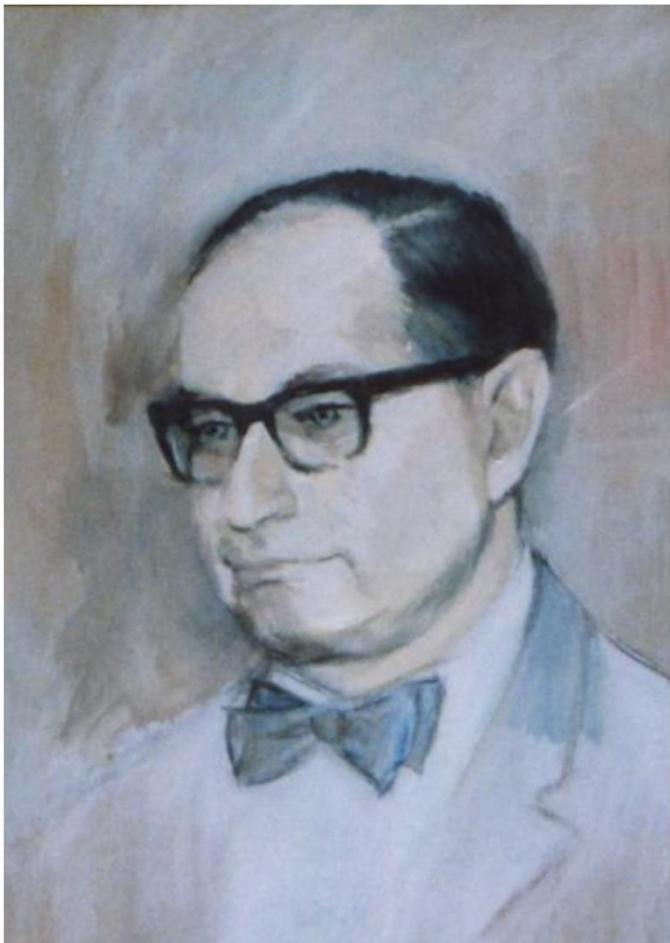


8-Hole Punched Tape

1970s



The Progress of Automation



Sketch of Sidney Fernbach, by his brother Frank Fernbach

- Automating arithmetic calculations
- Automating entire sequences of computations
- Automating data management
- Automating the management of computational resources
- Fernbach abstracted the needs of applications to architect of large-scale computing
- Fernbach turned scientific computing into an automated and repeatable process

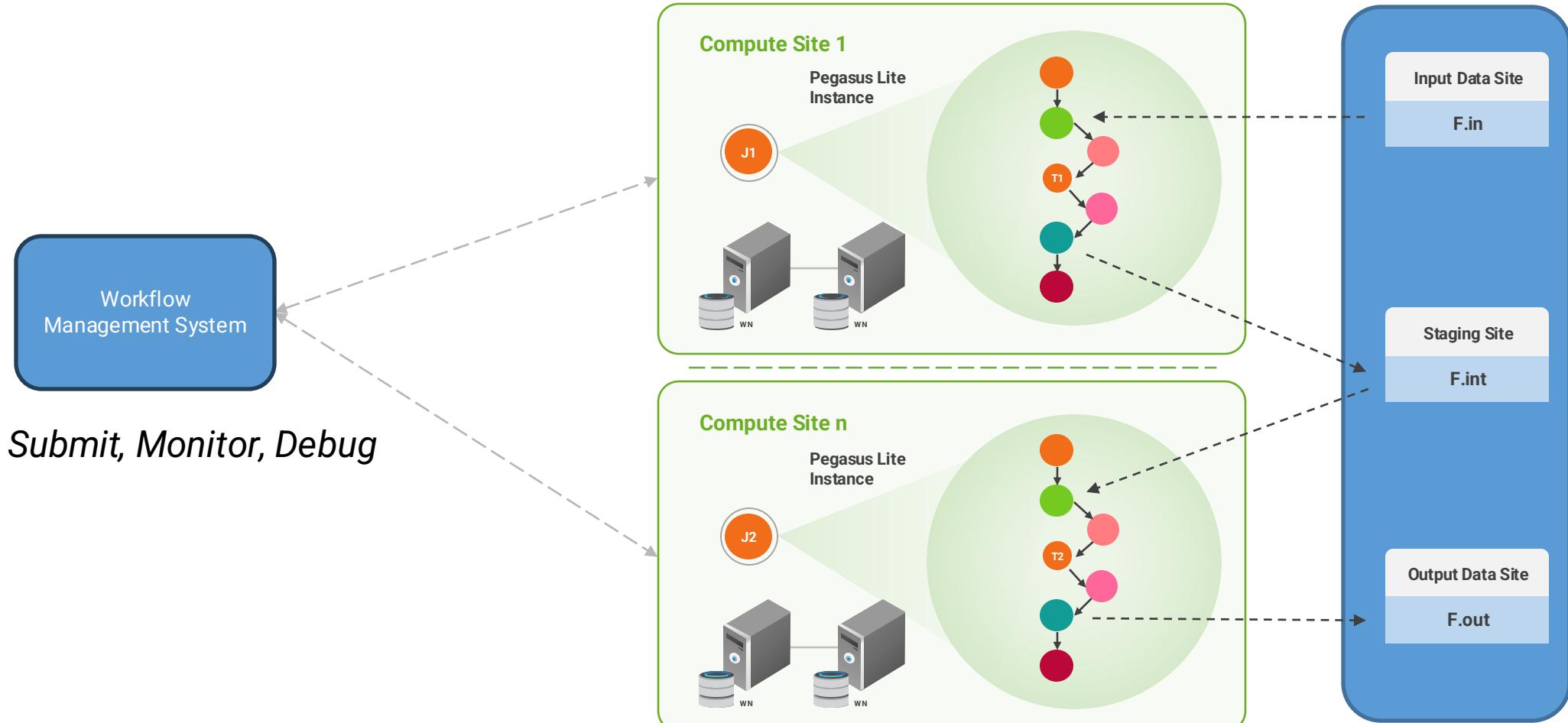


NERSC, Established in 1974 at LLNL
First machine, CDC 6600 1975



Pegasus Workflow Management System 2001

Pegasus' System Model





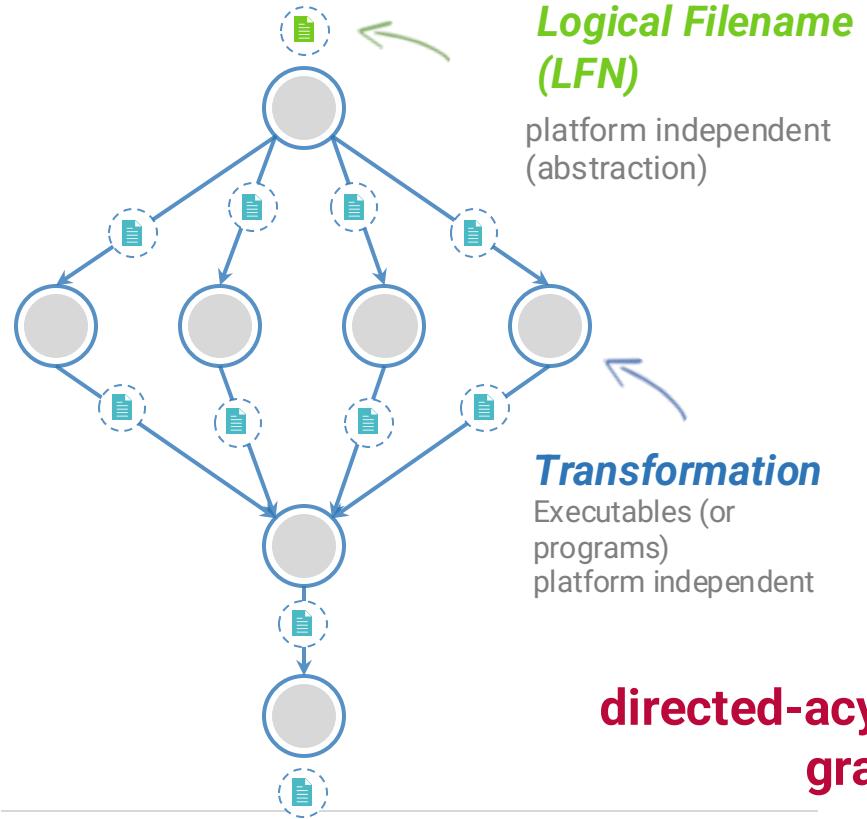
Abstraction: Resource-Independent Specification

Input Workflow Specification

Portable Description

Users do not worry about low level execution details

ABSTRACT WORKFLOW



YAML
formatted

Create Remote Directories

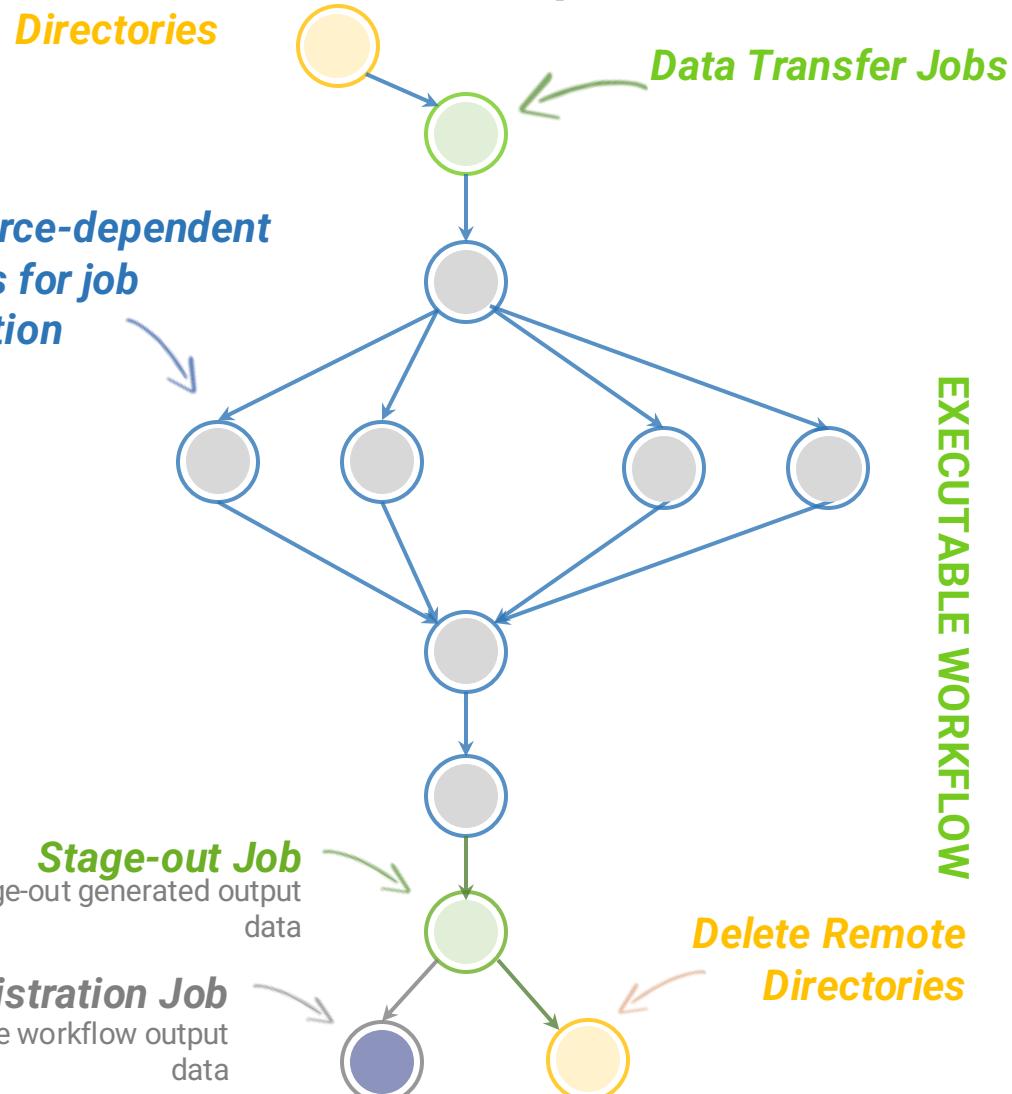
Resource-dependent scripts for job execution

Stage-out Job
Stage-out generated output data

Registration Job
Registers the workflow output data

Delete Remote Directories

Output Workflow



EXECUTABLE WORKFLOW



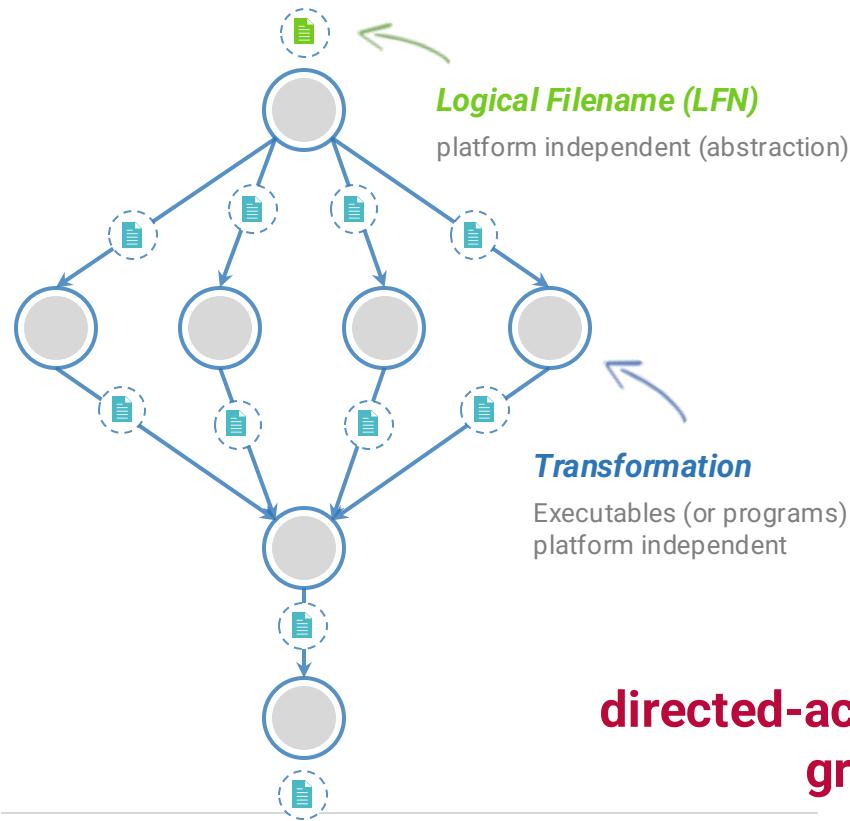
Abstraction: Resource-Independent Specification

Input Workflow Specification

Portable Description

Users do not worry about low level execution details

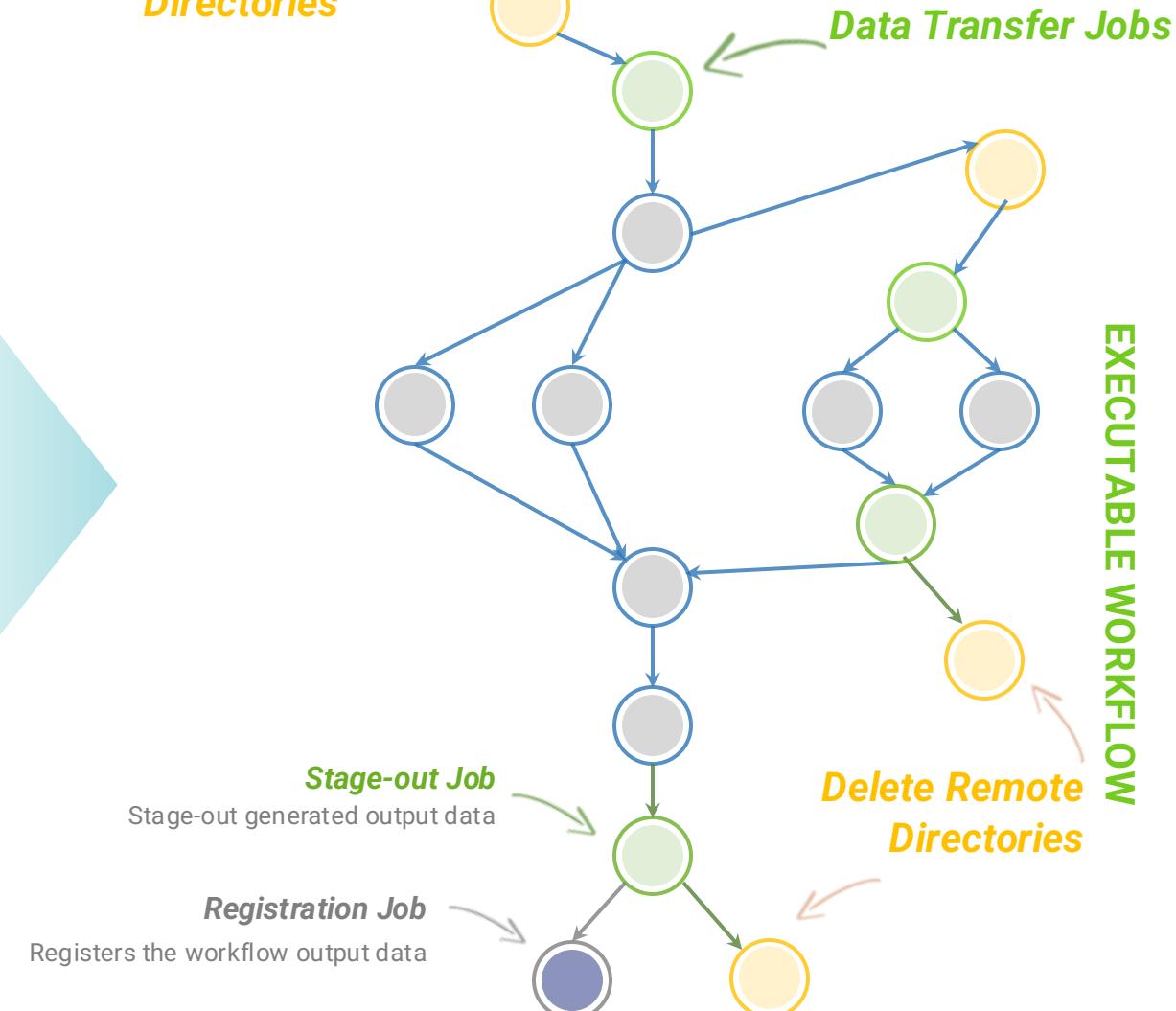
ABSTRACT WORKFLOW



YAML
formatted

Create Remote
Directories

Output Workflow



Supporting Science over Generations of CI

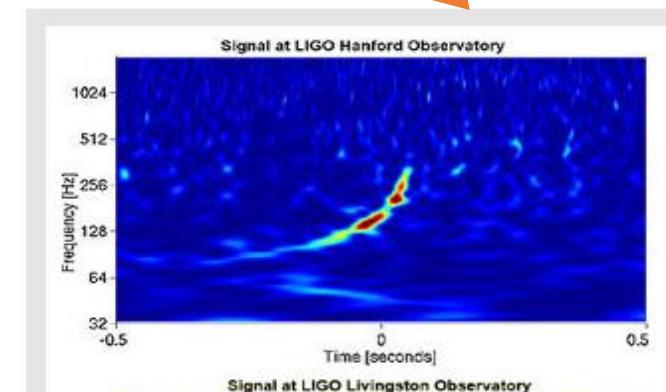
Pegasus working with LIGO
(Laser-Interferometer Gravitational Wave Observatory)



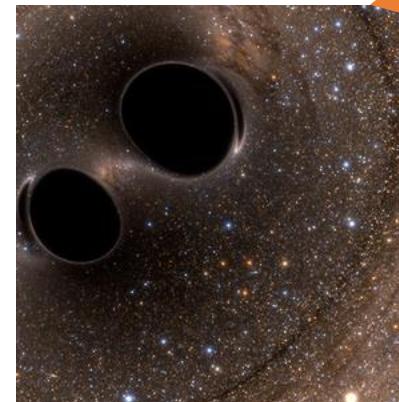
Nobel
Prize



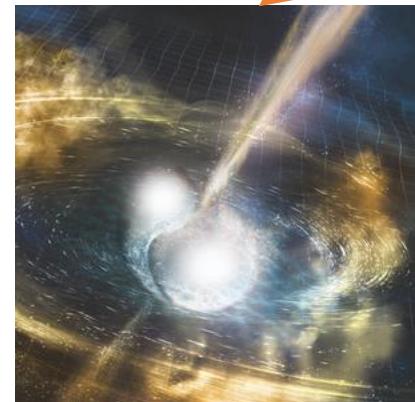
First Pegasus
prototype



Blind injection detection



First detection of
black hole collision



Multi-messenger
neutron star merger
observation

Image credit: LIGO Scientific Collaboration



2003

Compilation and Optimization

Workflow transformations:

- Partitioning to achieve scalability
- Hierarchical workflows for just-in-time planning and loops
- Task clustering to overcome system overheads and increase computation to communication/data transfer
- **Data reuse: for efficiency and fault recovery**
- Data cleanup to reduce the workflow footprint

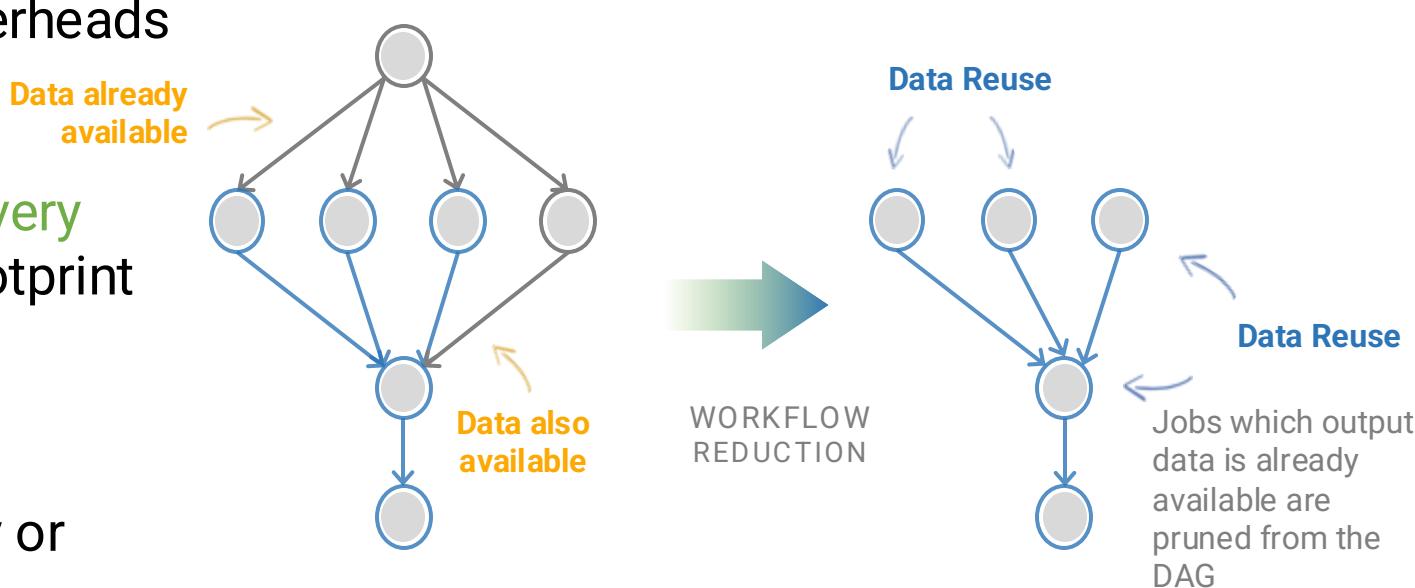
Scheduling:

- Scheduling computational tasks
- Data-aware scheduling (for data locality or reduced data footprint)

Grid-Based Galaxy Morphology Analysis for the National Virtual Observatory

Ewa Deelman
Information Sciences Institute, University of Southern California, Marina Del Rey, CA 90202 (ISI),
deelman@isi.edu

Raymond Plante
National Center for Supercomputing Applications, Champaign, IL 61820
rplante@nscs.uiuc.edu





Compilation and Optimization

Workflow transformations:

- Partitioning to achieve scalability
- Hierarchical workflows for just-in-time planning, loops and conditionals
- Task clustering to overcome system overheads and increase computation to communication/data transfer
- Data reuse: for efficiency and fault recovery
- **Data cleanup to reduce the workflow footprint**

Scheduling:

- Scheduling computational tasks
- Data-aware scheduling (for data locality or reduced data footprint)

Scheduling Data-Intensive Workflows onto Storage-Constrained Distributed Resources

Arun Ramakrishnan¹, Gurmeet Singh², Henan Zhao³, Ewa Deelman², Rizos Sakellariou³, Karan Vahi², Kent Blackburn⁴, David Meyers^{4,5}, and Michael Samidi⁴

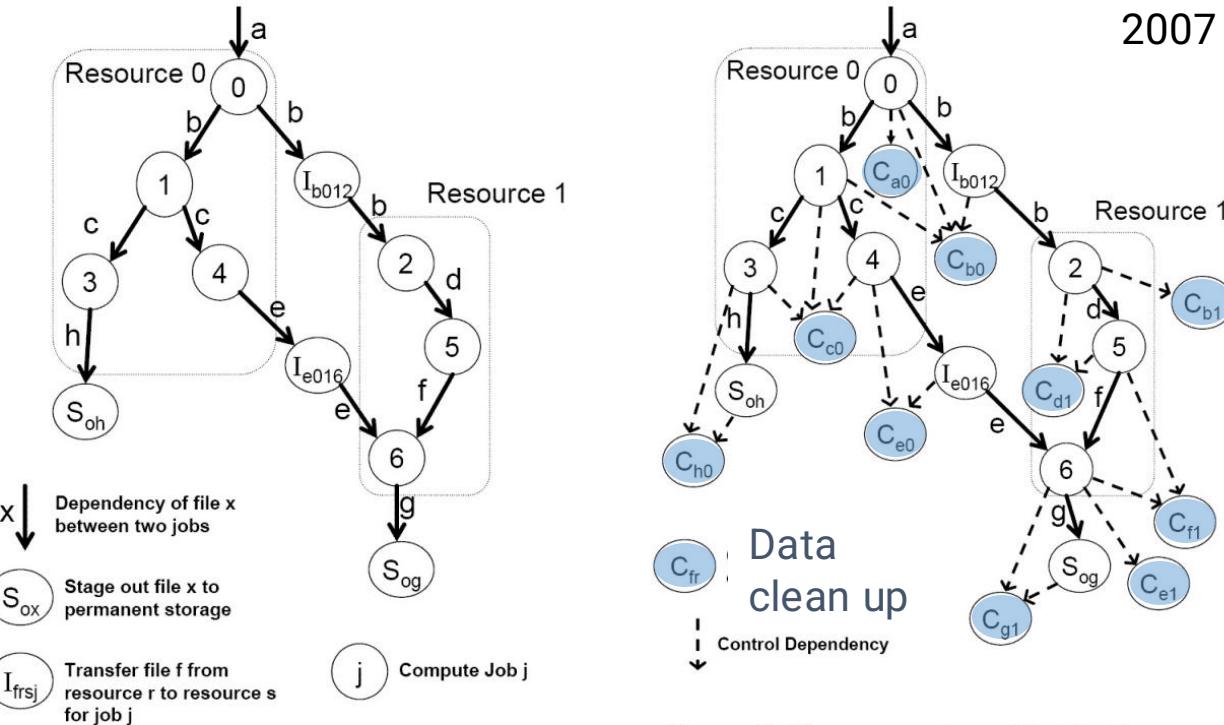


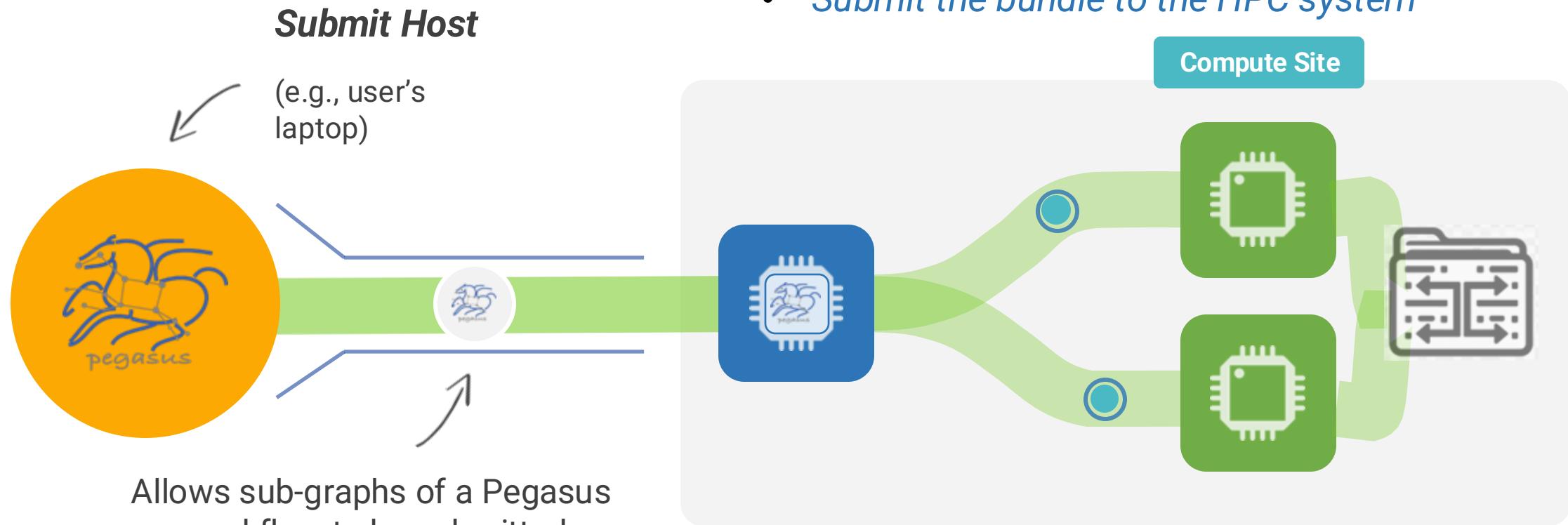
Figure 1. Executable workflow with 7 compute jobs mapped to two resources.

Figure 2. Cleanup nodes added to the executable workflow.

Handling heterogeneous workloads: Running High Throughput Computing jobs on HPC systems...



- Cluster HTC jobs
- Add an MPI based workflow engine
- Submit the bundle to the HPC system

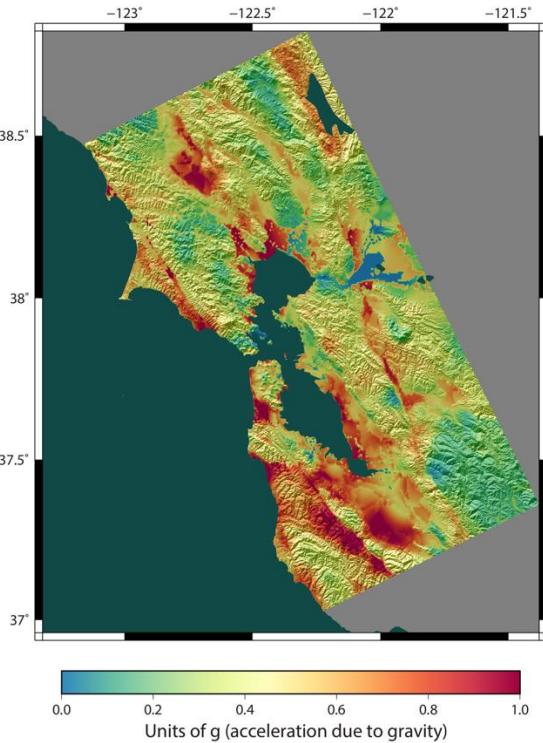


Designed for SCEC, used also by LIGO and others



Pegasus for Seismic Hazard Applications

Two new Northern California hazard models were produced.



Useful information for:

- ✓ Building engineers
- ✓ Disaster planners
- ✓ Insurance agencies

945
Workflows
27,720 jobs

1 PB
Data
Managed by
Pegasus

9M
Files
Staged to USC

45
Days
Duration of 24/7
operation

55K
Frontier
OLCF
Node-hours

125K
Frontera
TACC
Node-hours

2024

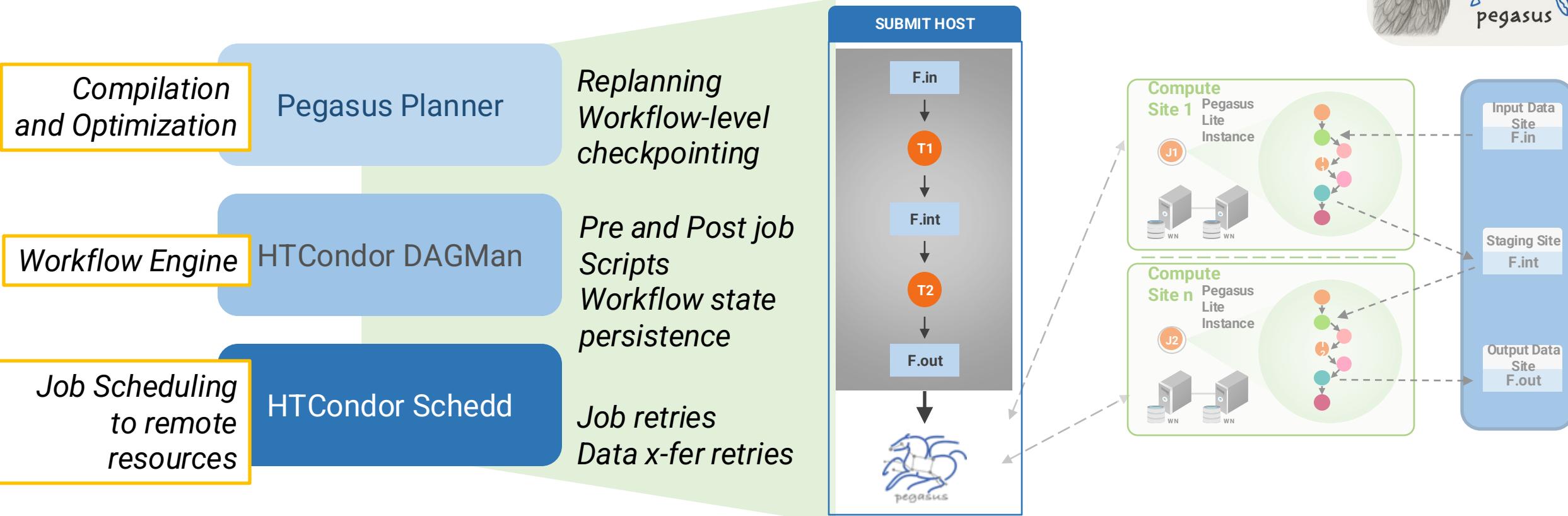
High-water marks were 4130 nodes on Frontier (44%)
and 1029 nodes on Frontera (12%)

Neither used a reservation or priority bump.



Layered Architecture

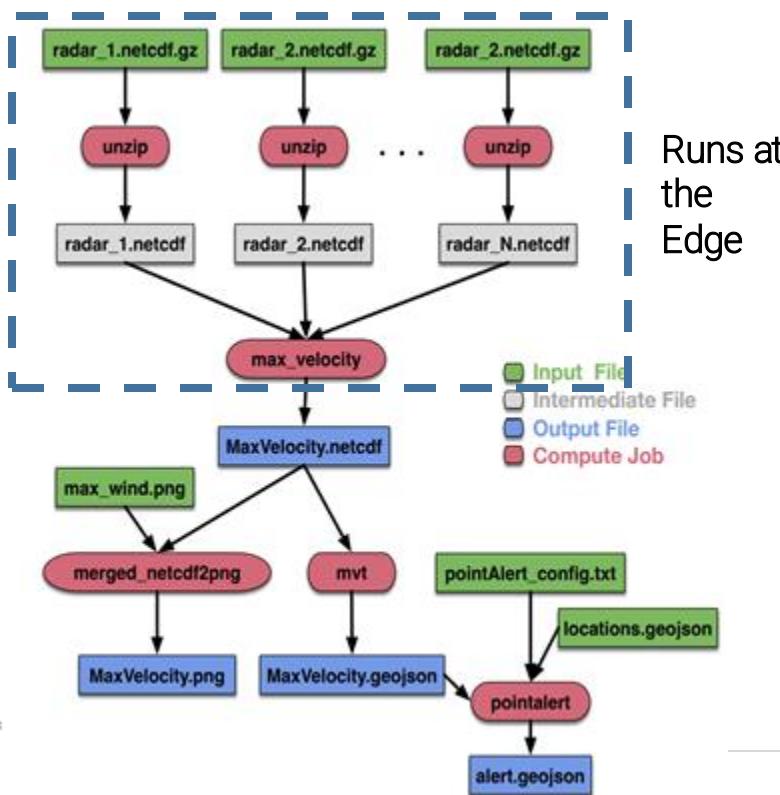
Leveraging proven technologies, building upon layers of resilience



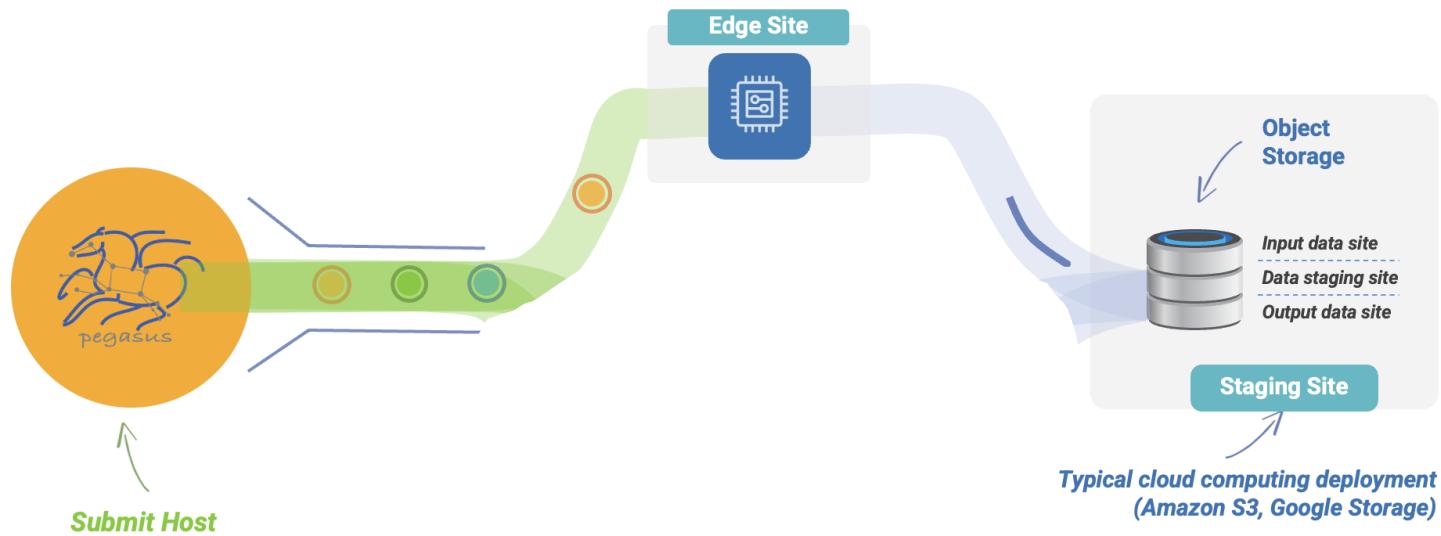
Edge-2-Cloud Applications

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



Distributed Model, Data Footprint Management, Automated Data Staging



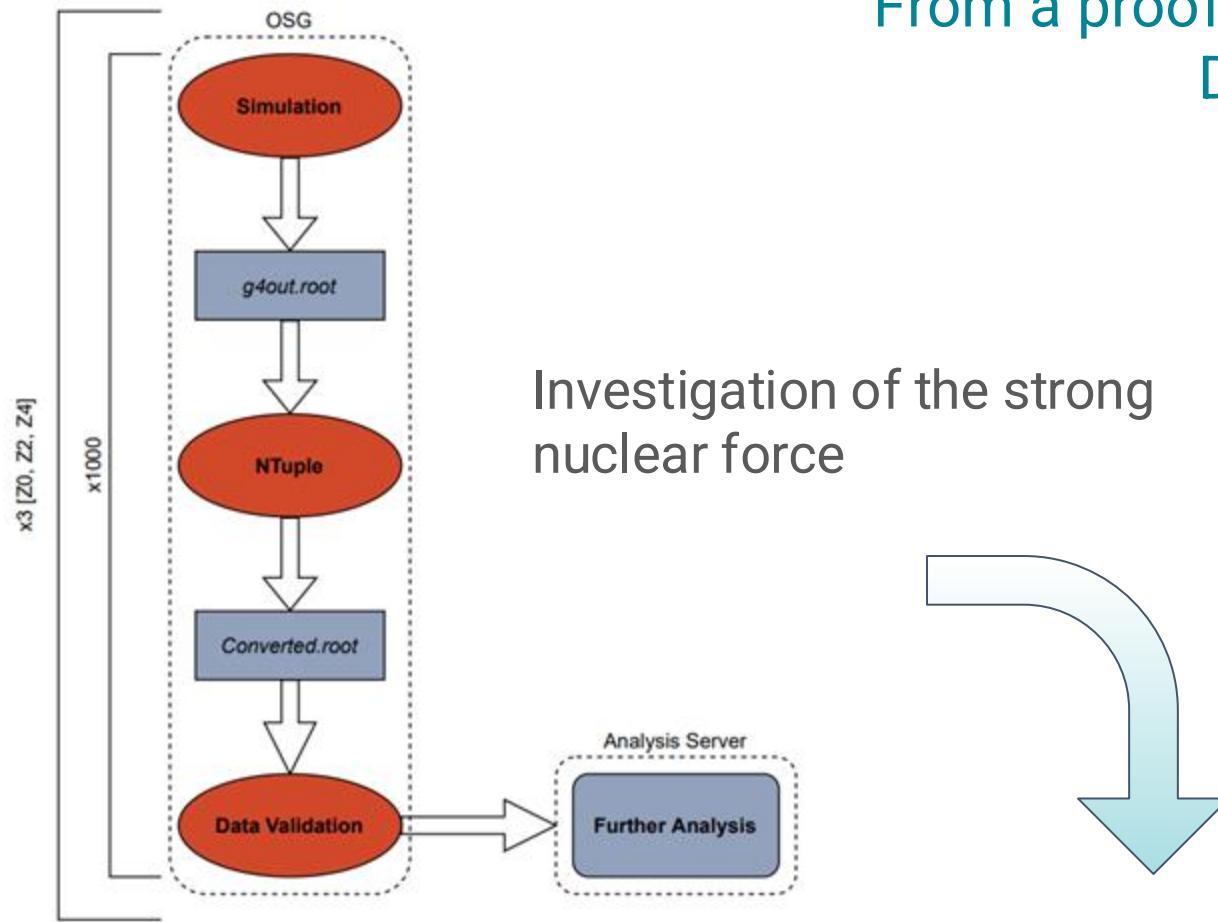
2019

This work is funded by NSF, award #2018074

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



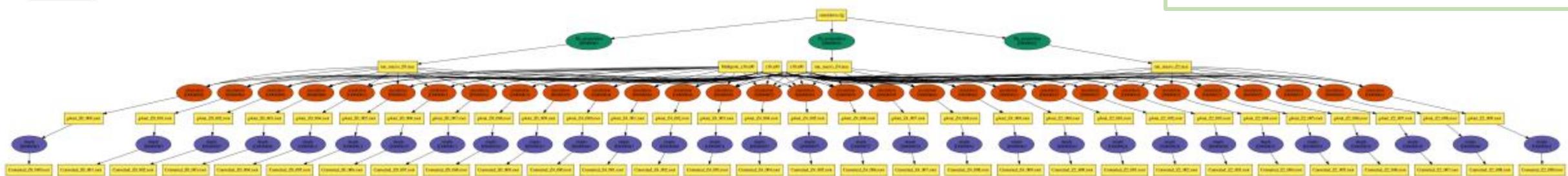
Connor Natzke,
Physics Student
Colorado School of Mines

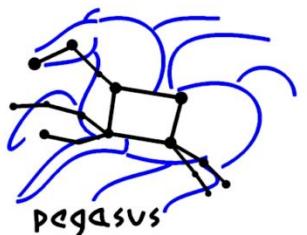


Investigation of the strong nuclear force

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

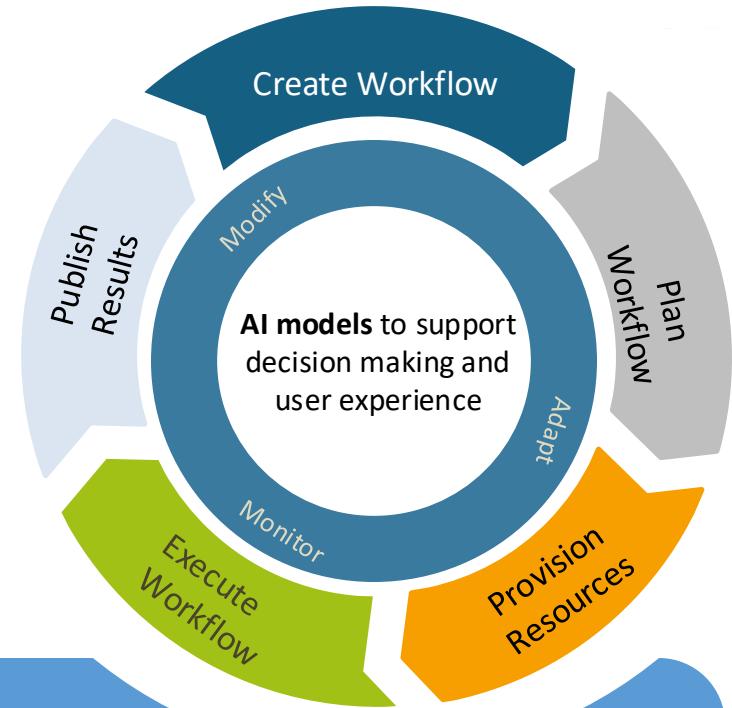
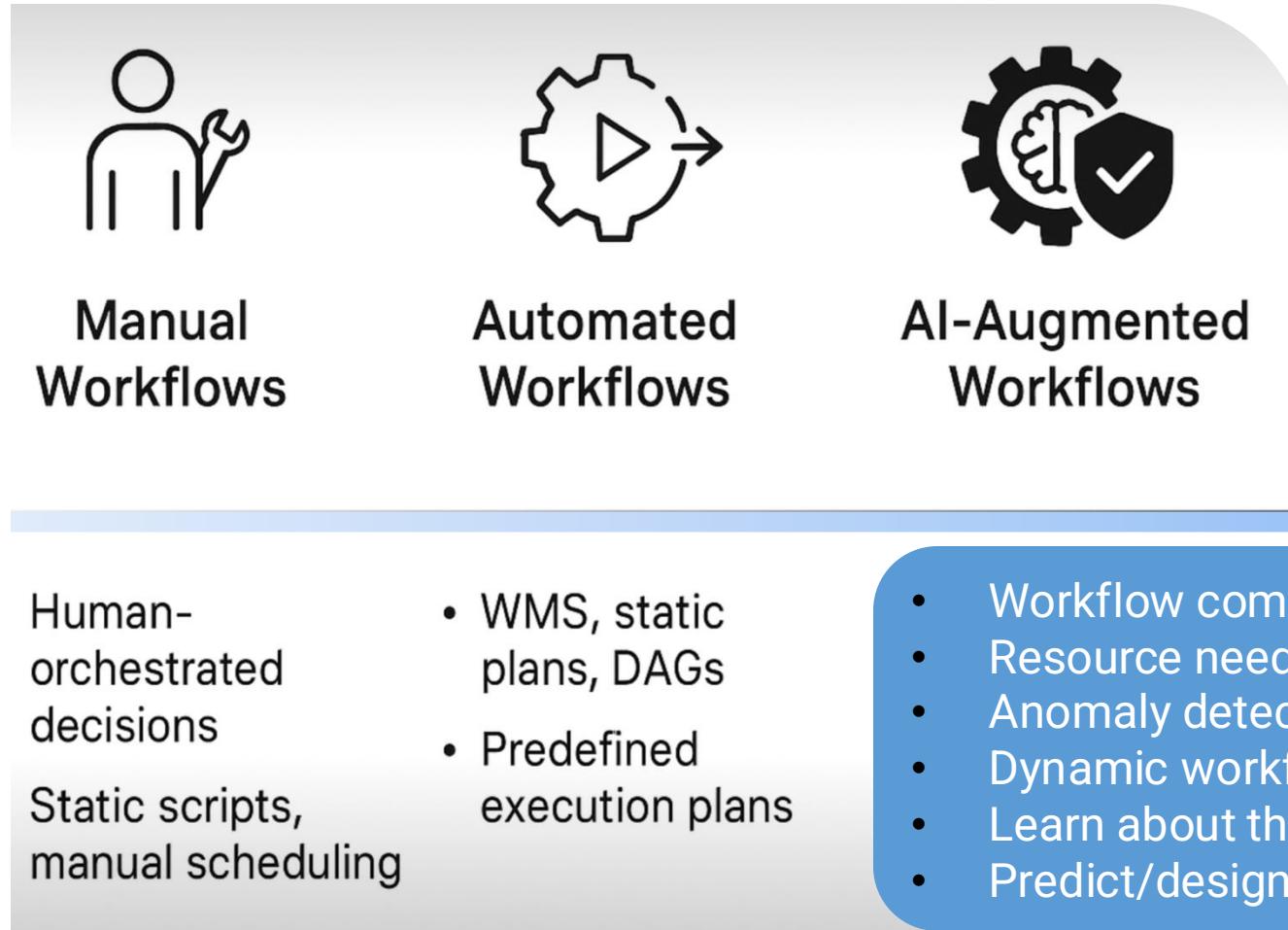
- Pegasus provided:
- 1) Automation and scaling
 - 2) Automatic job retries
 - 3) Automation of file transfers
 - 4) Managed disk space at execution sites





Supported by NSF Grant #2513101

2025 - 2030



Make AI-ready monitoring data available to the community



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



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
```

Mats is adding a Pegasus AI assistant
to help interpret the errors.



Mats Rynge
SENIOR COMPUTER SCIENTIST



Latest development version of Pegasus has an AI assistant



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

The workflow failed due to:

**Expected local file

Root Cause:

- The required input file was not found.
- This prevents the training process from starting.

Next Steps:

1. Verify the file exists.
2. Ensure the file path is correct.
3. Resubmit the workflow.

The remaining unsubmitted workflows will be retried. This error will enable the AI to learn from it and improve future executions.

Challenges of using AI technologies

- Accuracy
- Deployment hosted versus local deployments
- Dealing with network problems
- Data privacy
- Resources usage and costs



Simple Input Validation Opportunities



Prompt obtained from a discussion with LLM

You are a Pegasus workflow validator. Perform **structural & consistency checks** on a Pegasus abstract workflow.



sites.yml

File



montage-workflow.yml

File

Perform a A) DAX Integrity 1) Every job.id is unique and has a job.name. 2) For every job.name, there must be a matching transformation in the Transformation Catalog (TC). - If TC tracks namespace/version, prefer exact match on (namespace, name, version). 3) Every file the jobs **consume** (uses[].type|link == "input") must have at least one PFN in the Replica Catalog (RC). 4) If jobDependencies is present: - All referenced parent/child IDs must exist among jobs.

Here's the result of A) DAX Integrity on your uploaded workflow.

1) Job IDs & names

PASS. All 472 jobs have a unique `job.id` and a `job.name`. No duplicates or missing names detected. montage-workflow

2) Transformation Catalog coverage

PASS. Every job's `name` resolves to a transformation present in the TC. In this workflow, neither the jobs nor the TC specify `namespace / version`, so matches are by **name-only** (exact `(namespace, name, version)` match not applicable). No jobs without a TC entry were found. montage-workflow

3) Input files have PFNs in the Replica Catalog

PASS. All source inputs (inputs that are not produced by any job) have at least one PFN in the RC.

- Source inputs identified: 62 LFNs (e.g., `region-oversized.hdr`, the `*-images.tbl` / `*-raw.tbl` / `*-projected.tbl` / `*-corrected.tbl` / `*-stat.tbl` tables, and the external `poss2ukstu_*/*.fits` files).
- RC provides file:// PFNs for local tables and http:// PFNs for external FITS imagery; no missing PFNs found.
(Intermediate files produced by one job and consumed by another were excluded from this check.) montage-workflow

4) `jobDependencies` integrity (if present)

PASS. `jobDependencies` is present; all referenced parent/child IDs exist among defined jobs. No missing parents or children were found. montage-workflow



Simple Input Validation Opportunities

Here's the quick, human-friendly readout of the **DAX Integrity** checks we ran:

- **Jobs are well-formed**

Every job has a unique `id` and a `name`.

Why it matters: prevents planner confusion and duplicate execution.

- **Transformations are covered**

Each job's `name` matches a transformation in the Transformation Catalog. (No

re used.)

ontainer

simplify some more

ob has at least

Sure — super short version:

- **Jobs:** all unique IDs and names — **OK**
- **Transforms:** every job maps to a known transformation — **OK**
- **Inputs:** all required input files have a location in the RC — **OK**
- **Dependencies:** all parent/child links point to real jobs — **OK**

can you sum
understanda

ents/children

e executed as a

Bottom line: The workflow's structure looks solid and ready to plan/execute.

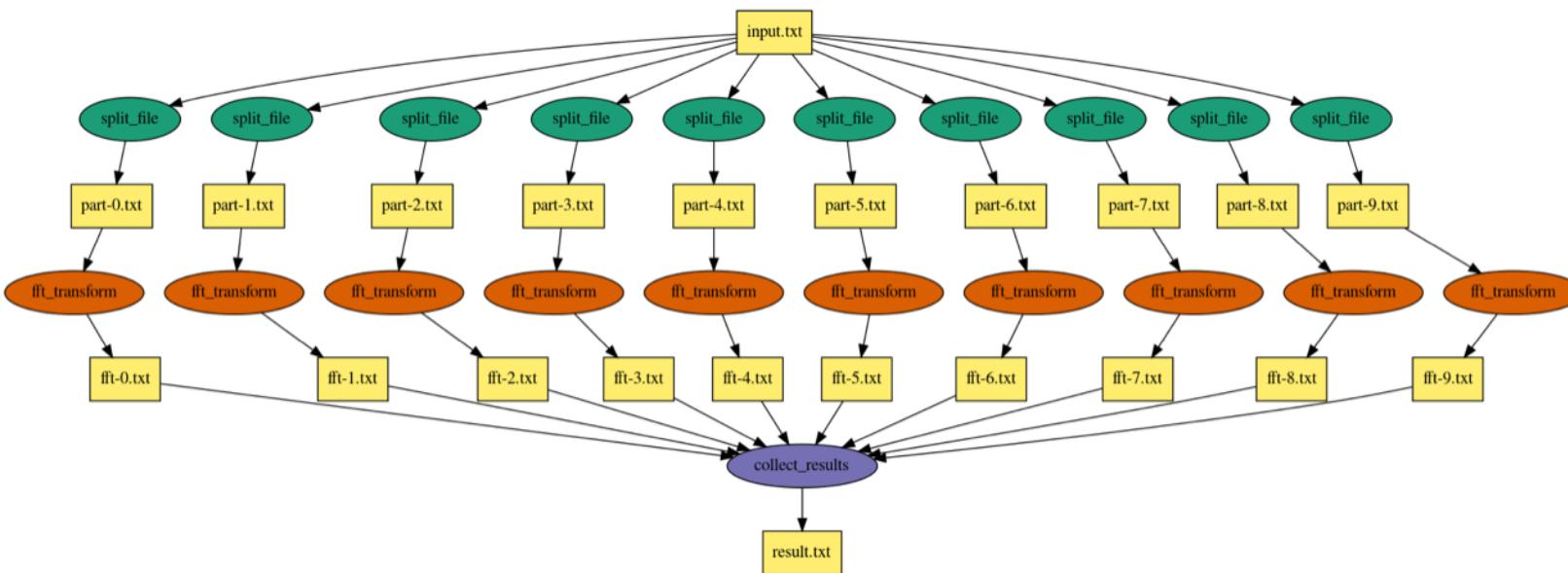
transformations exist for all jobs, required inputs have locations, and dependency references are valid. If you'd like, I can now run the **Dependency Graph Analysis** (cycles/dangling/dead-ends) and **Metadata Validation** or generate a compact JSON report you can archive.



Abstraction and Intent

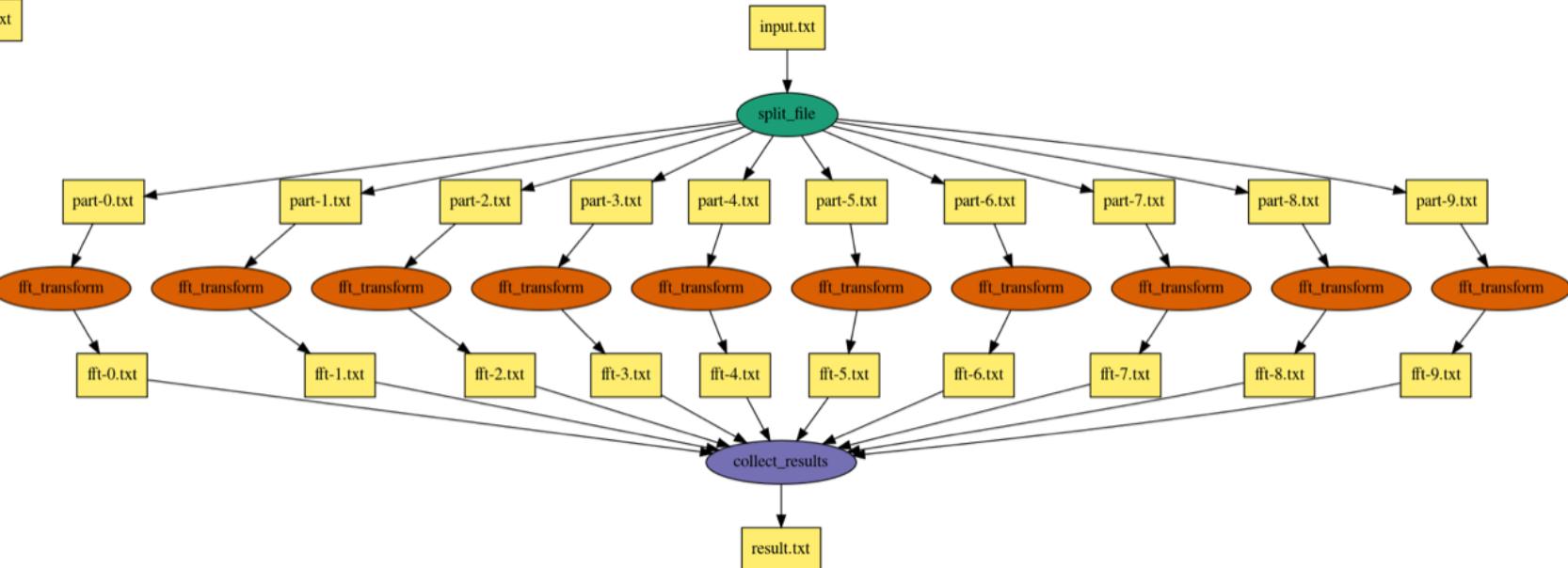


ChatGPT



in

Karan Vahi
SENIOR COMPUTER SCIENTIST



Balance between Creation and Validation and Verification

- More black boxes in the methods—not only AI, lots of software available for reuse and repurposing
- More AI-based methods are being used to analyze and synthesize data, chose solutions, etc..
- LLMs used to generate codes and connect codes

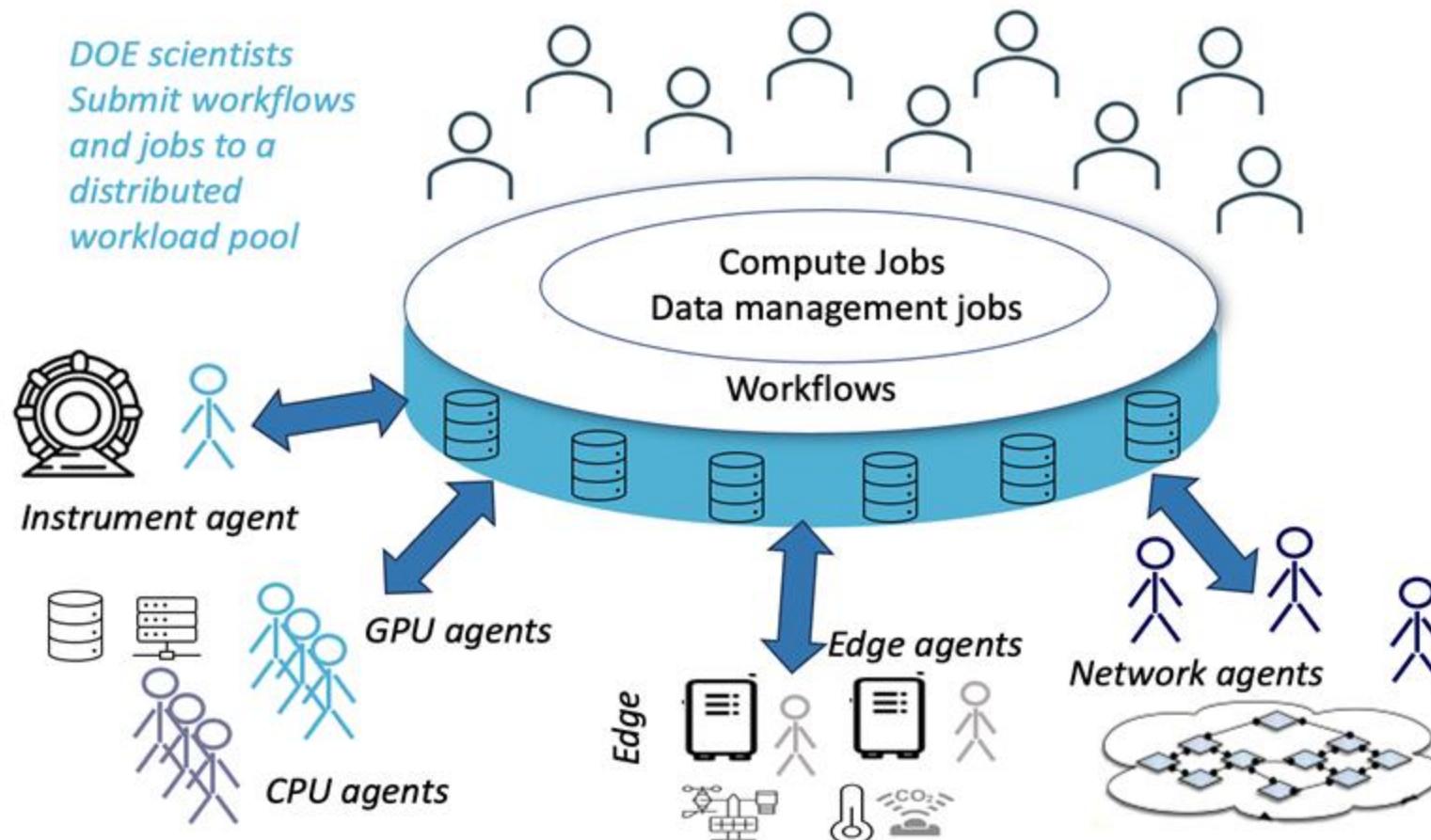
* Methods change over time as AI models learn

Growing emphasis on verification methods and validation of results

Growing need for Automation

- Easily replicate our own work
- Easily reproduce others' finding
- Meta-analysis: Need be able to easily re-run the same analysis with different data and parameters (potentially a large number of times)
- And more

SWARM: Scientific Workflow Applications on **Resilient** Metasystem



Intersection of Agentic AI
and Distributed and High-
performance computing

- Heterogenous agents
- Inter-agent communications
- Resilient strategies
- Consensus
- Communication patterns

Swarm Intelligence agents select workload to execute and autonomously adapt

Funded by DOE: DE-SC0024387
2023- 2028

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



Chandreyee Bhowmick, Ph.D.
ORNL



Hongwei Jin, Ph.D.
ANL



Komal Thareja
RENCI



Fred Sutter, Ph.D.
ORNL



Erik Scott
RENCI



Shixun Wu
UCR



Hong-Jun Yoon, Ph.D.
ORNL



Sheng Di, Ph.D.
UCR

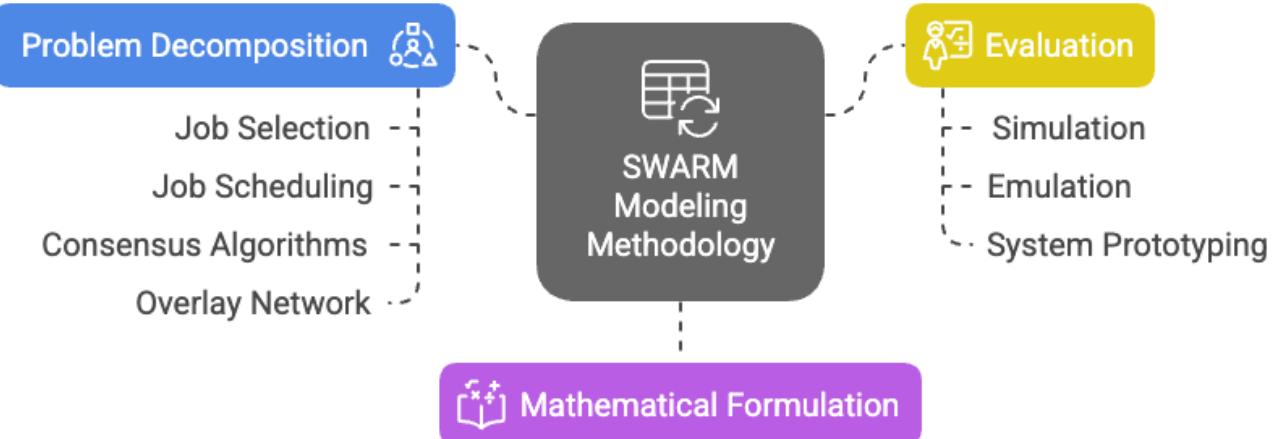


Suman Raj
USC



Prachi Jadhav
ORNL

SWARM Methodology & Findings



Greedy algorithms perform well for consensus-based job selection when compared to traditional consensus protocols

Swarm intelligence algorithms do not give good solutions for traditional problems like job scheduling

With sophisticated prompting, LLMs can be used for job scheduling and can make multi-criteria decisions, criteria can be dynamic

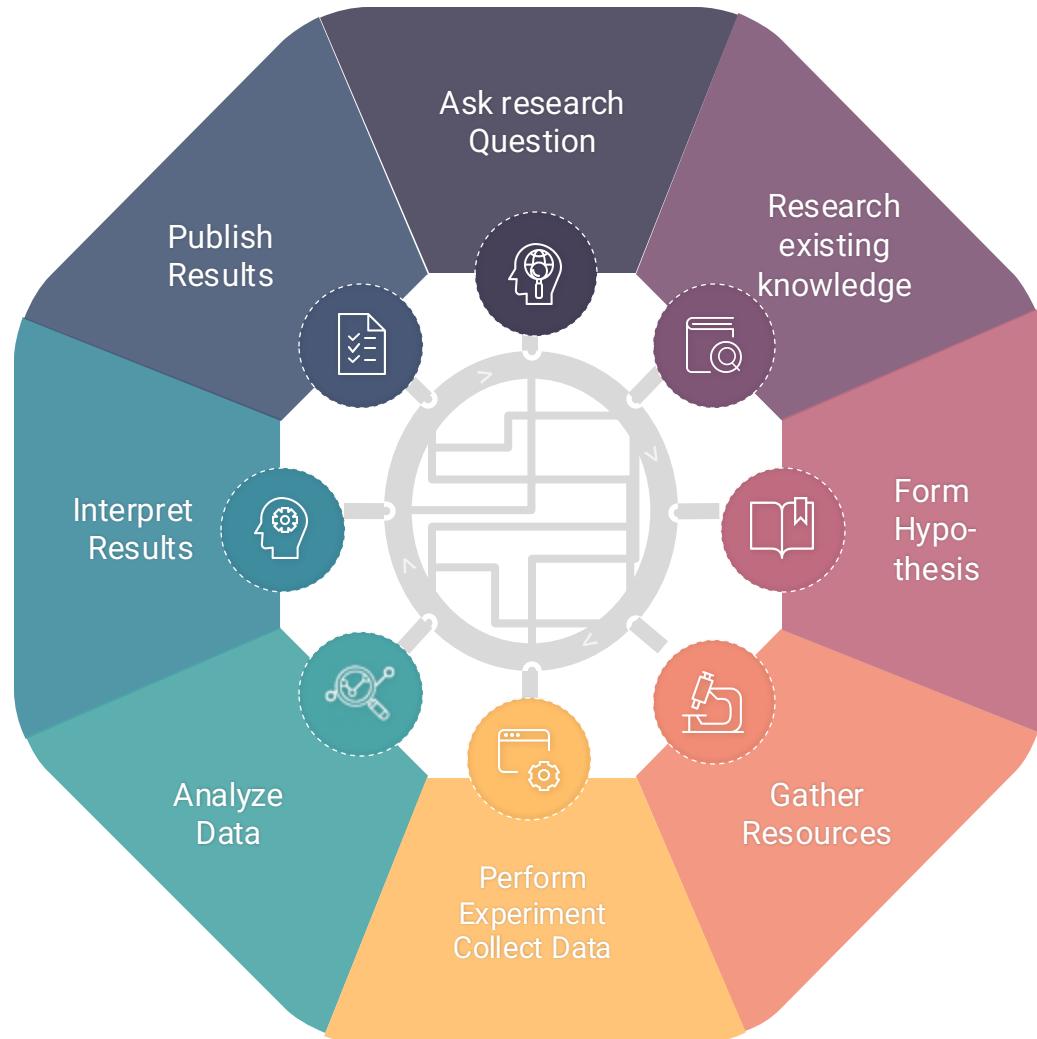
Ring building algorithms enhanced by Q-learning improve the diameter of network overlays over traditional methods

Agentic frameworks can be used for job scheduling and data management using tools for information gathering and job submission

Pushing the boundaries of abstraction and automation in the scientific lifecycle

Potential:

- Enhance productivity
- Overcome our limited time, attention, cognition
- Enhance objectivity

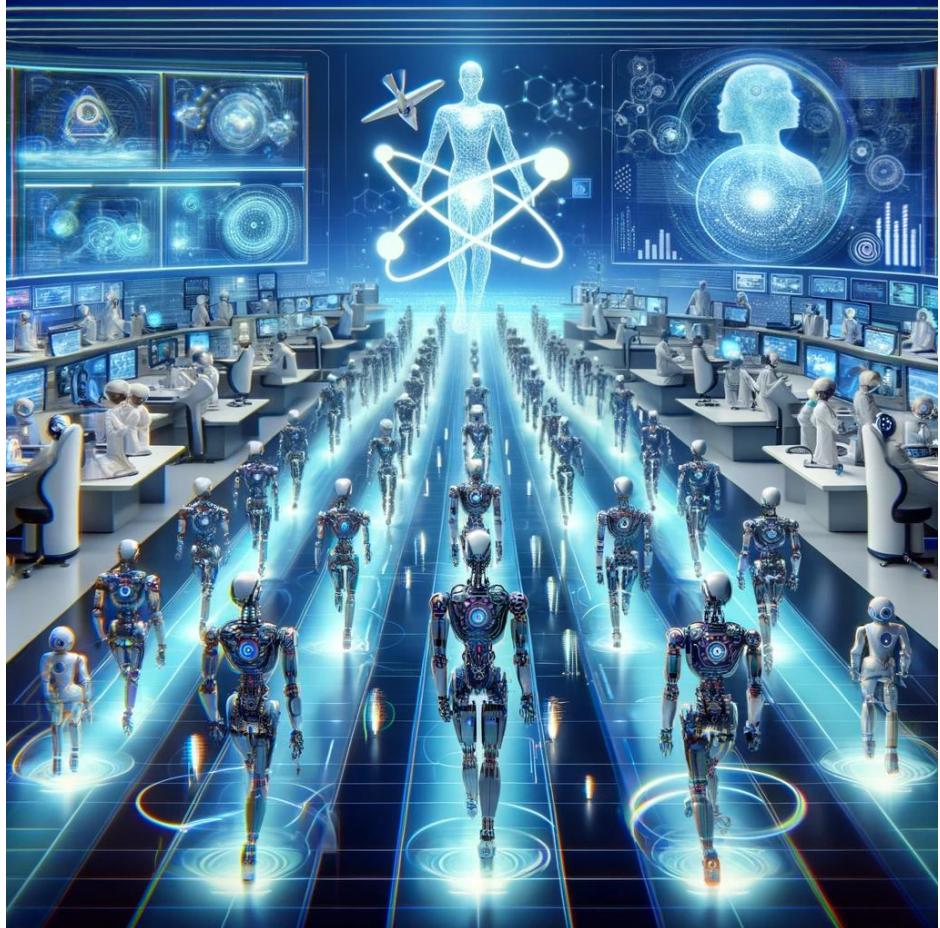


Downsides of Cognitive Automation

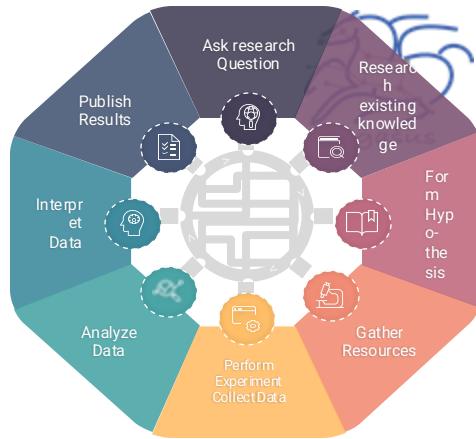
- Leads to skill degradation
- Reduction in critical thinking skills

* "This could result in a workforce that is highly efficient, yet potentially less capable of independent problem-solving and critical evaluation."

Impact on Science: AI and Automation Change the Science Workforce Landscape



ChatGPT-4: “draw a revolution in the scientific lifecycle organized by ai robots”



How will the scientific lifecycle look like in 10-20 years?

- How much of it will be automated?
- Will we still publish papers?
- Are there better ways of expressing ideas and gaining knowledge
- What happens to human collaborations?
- How will we teach science?
- What skills will we need?



Impact on Society: Can we maintain/enhance critical thinking skills?

“Ability to analyse, evaluate, and synthesise information to make reasoned decisions <...> fundamental cognitive skill essential for academic success, professional competence, and informed citizenship” *

- Problem-solving
- Decision-making
- Reflective thinking
- Deep thinking

Technology Making People Lazier as Finding the Remote Fast Becoming the Only Form of Exercise



Neetzan Zimmerman
06/18/12 09:15AM Filed to: OBESITY

12.75K

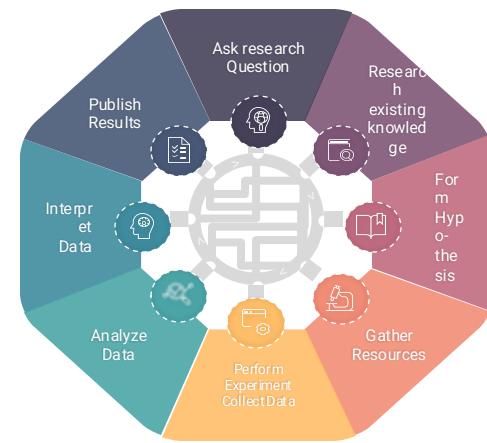


<https://gawker.com/5919185/technology-making-people-lazier-as-finding-the-remote-fast-becoming-the-only-form-of-exercise>

* Gerlich, M. 'AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking"
Societies 2025, 15, 6.

Summary thoughts on AI automation

- Tremendous opportunities to improve systems, make new **scientific discoveries**
- We need methodologies to **quantify and verify, validate, and curate AI solutions**
- How do we **curate knowledge**, how we exchange ideas and findings?
- How we improve productivity while **maintaining human collaborations and satisfaction?**
- How to use AI to improve education and **enhance critical thinking skills?**





and inquisitive

**“To be creative, you have to dislike
being bored. And I don’t think a
computer will ever feel bored”**

Kai Lars Polsterer
Heidelberg Institute for Theoretical Studies

Thank you!

Alan Chalker
Alana Romanella
Anirban Mandal
Andrew Pasquale
Ann Chervenack
Alicia Esquivel
Amanda Randles
Shishir Bharati
Angela Murillo
Boleslaw Szyniszowski
Bruce Hendrickson
Brian Tierney
Bruce G Berriman
Carl Kesselman
Cassandra Hayes
Charles Vardeman
Christina Hoffa
Chris Carothers
Daniel Reed
Daniel Gunter
David Abromson

Gurmeet Singh
Hamza Safri
Harshita Sahni
Henri Casanova
Hongwei Jin
John Foster
Ian Taylor
I. Santana-Perez
Igor Sfiligoi
Petri
Jukka Anttila
Faldin
Imtiaz Mahmud
Ivona Brandic
Jack Wells
Jack Dongarra
James Blythe
James A Knowles
James Howison
James Blythe
James C Browne

Ken Kennedy
Kent Blackburn
Kerk Kee
Kesheng Wu
Komal Thareja
Krishnan Raghavan
Laura Christopherson
Laura Pearlman
Laurens Versluis
Lavanya Ramakrishnan
Loic Pottier
Louis H Ziantz
Luc Moreau
Maciej Malawski
Mahedi Hasan
Malcolm P Atkinson
Manish Parashar
Marcia McNutt
Marcus Thiebaux
Maria A. Pano

Nicole Virdone
Nona Nersisyan
Norbert Meyer
O. Okunloye
Orcun Yildiz
Patricia J Tamm
Patrick R Brady
Patrycja Krawczuk
Paul Groth
Paul Ruth
Paulo Llava
Pawel Gepner
Pawel Zuk
Philip Maechling
Prachi Jadhav
Prasad Calyam
Prasanna Balaprakash
Rafael F. Da Silva
Rajiv Mayani
Rajive L Bagrodia
Raphael Polze

Ruth Pordes
Ryan Mitchell
Ryan Tanaka
Sandeej Gupta
Saralyn Giesing
Sarowar Hossain
Satoshi Matsuoka
Scott Koranda
Scott Denehy
Scott Carlson
Sebastian Seung
Sepideh Azarnoosh
Shava Smallen
Shelley Knuth
Shishir Bharathi
Shixun Wu
S. Chandrasekaran
Silvina Caino-Lores
Simon Miles
Sonali Jain
Vipin Kumar

Thomas Caraco
Thomas Fahringer
Todd Miller
Tom Peterka
Tomasz Osinski
Tony Hey
Trilce Estrada
Tristan Glatard
Tu Mai Anh Do
Umut Catalyurek
Veronique E Taylor
Veronique Taylor
Valerio Pascucci
Veronika Nefedova
Vickie Lynch
Victoria Stodden
Vikram Adve
Vikram Gazula
Viktor Prasanna
Vincenzo De Maio
Vipin Kumar



U.S. National Science Foundation



Turning Discovery Into Health



U.S. DEPARTMENT OF
ENERGY

