

# 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



2011



2017



2019



2021



2023





SC'17 Denver

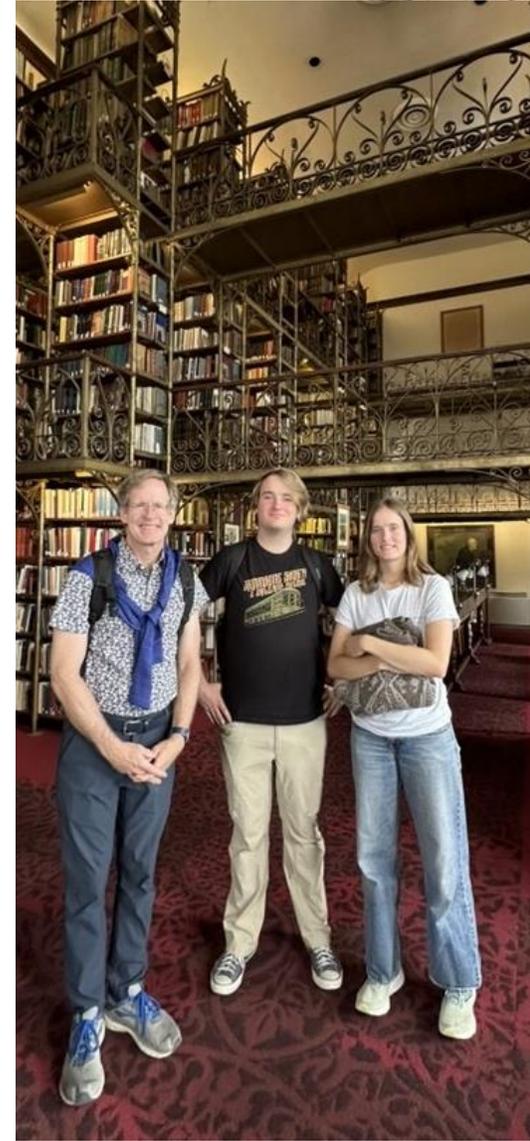
Thank you Caroline and  
Nicholas!!  
Thank you Peter!



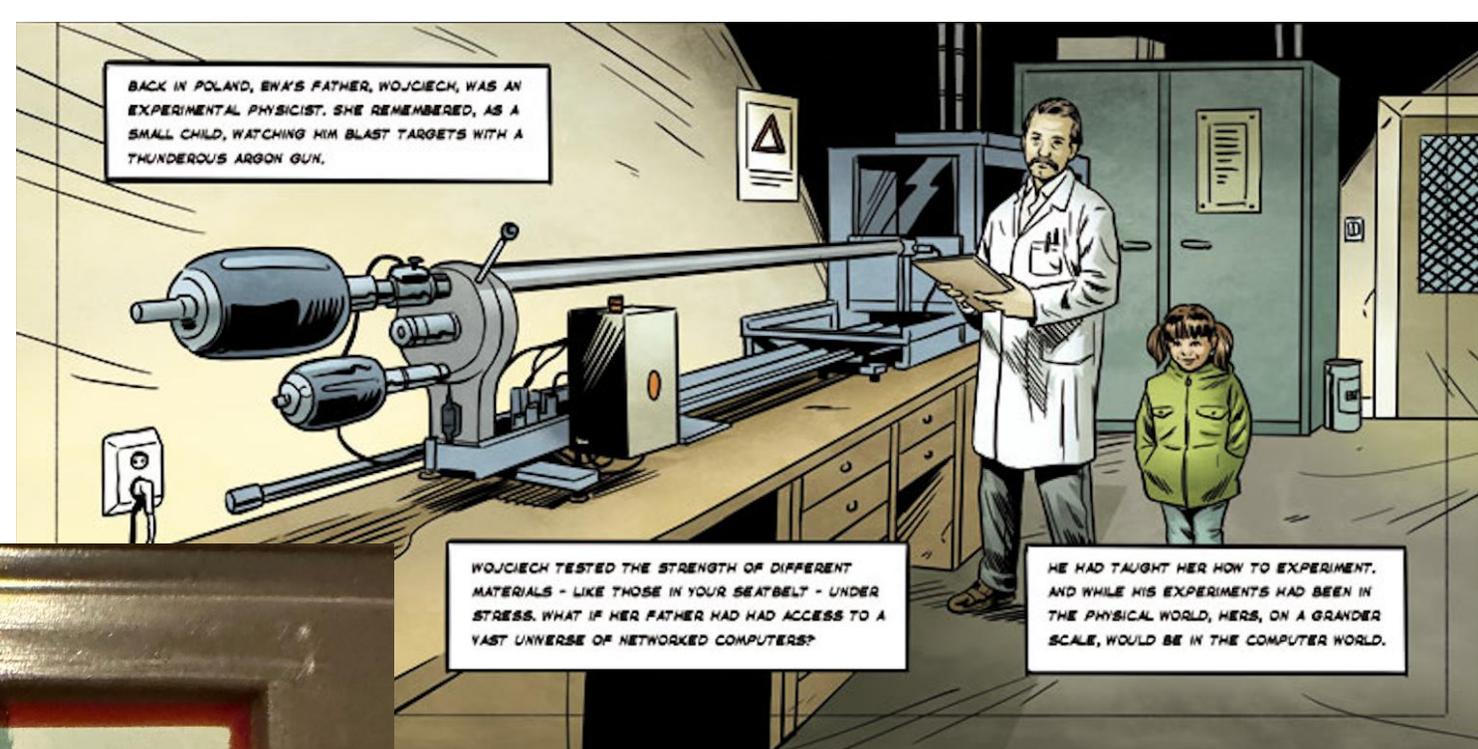
SC'18 Dallas



SC'19 Denver



Cornell University  
July 2025



Wojciech Nowacki  
Nicholas and Caroline Deelman  
June 2009

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



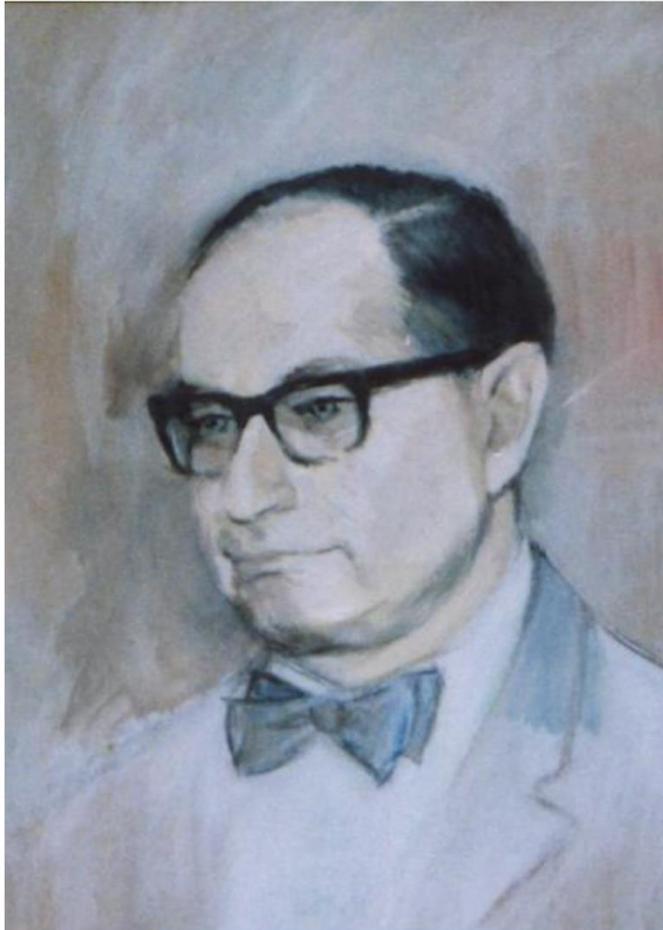
Wojciech Nowacki, c. 1974



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*



# Abstraction: Resource-Independent Specification



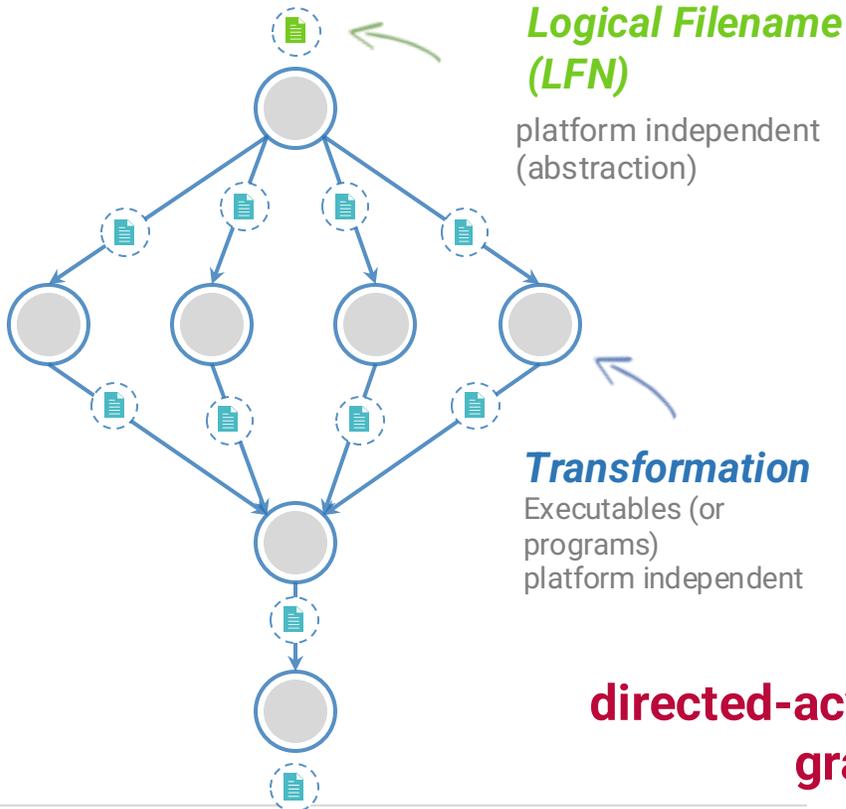
## Input Workflow Specification

**YAML  
formatted**

### Portable Description

Users do not worry about low level execution details

**ABSTRACT WORKFLOW**



**Create Remote Directories**

## Output Workflow

**Resource-dependent scripts for job execution**

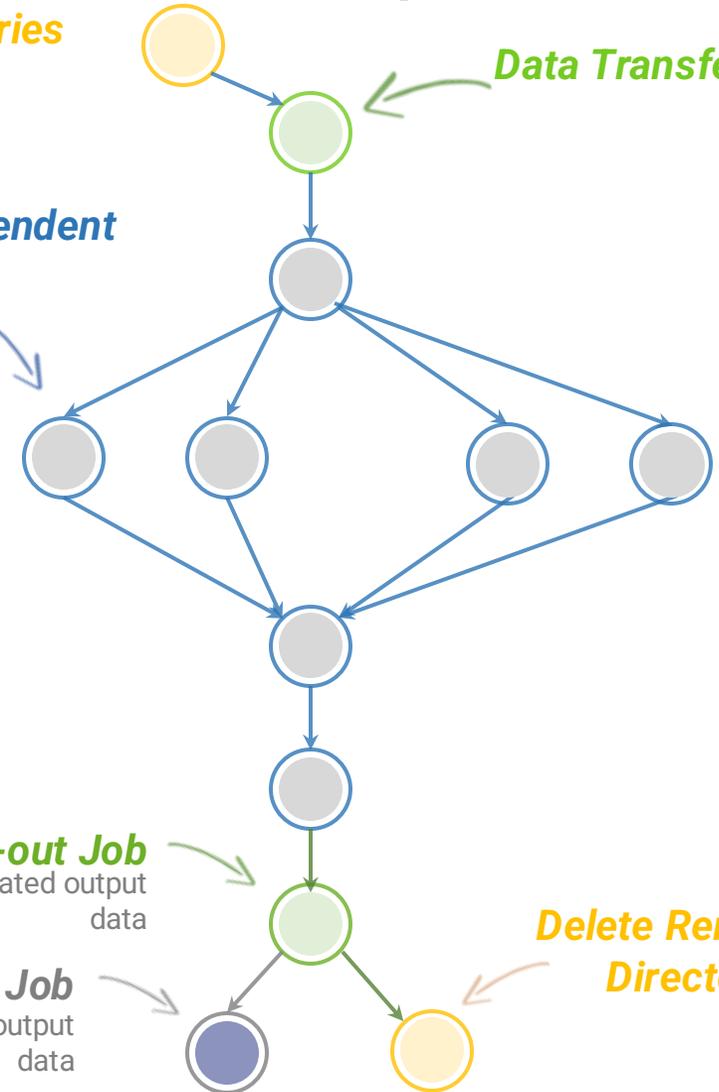
**Stage-out Job**  
Stage-out generated output data

**Registration Job**  
Registers the workflow output data

**Data Transfer Jobs**

**Delete Remote Directories**

**EXECUTABLE WORKFLOW**



# Abstraction: Resource-Independent Specification



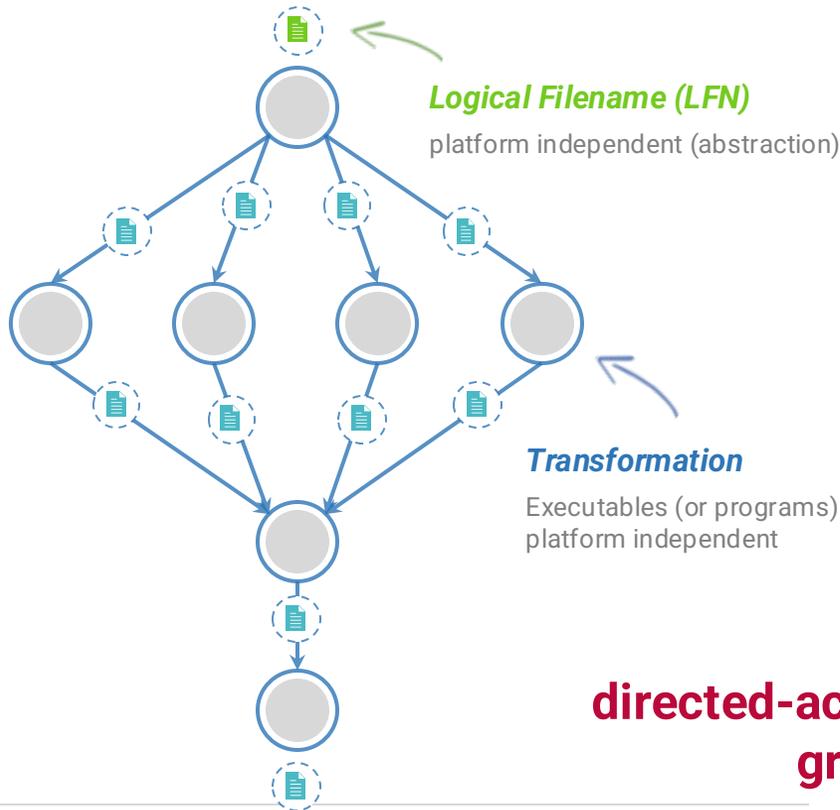
## Input Workflow Specification

**YAML  
formatted**

### Portable Description

Users do not worry about low level execution details

**ABSTRACT WORKFLOW**



**directed-acyclic  
graphs**

**Create Remote  
Directories**

## Output Workflow

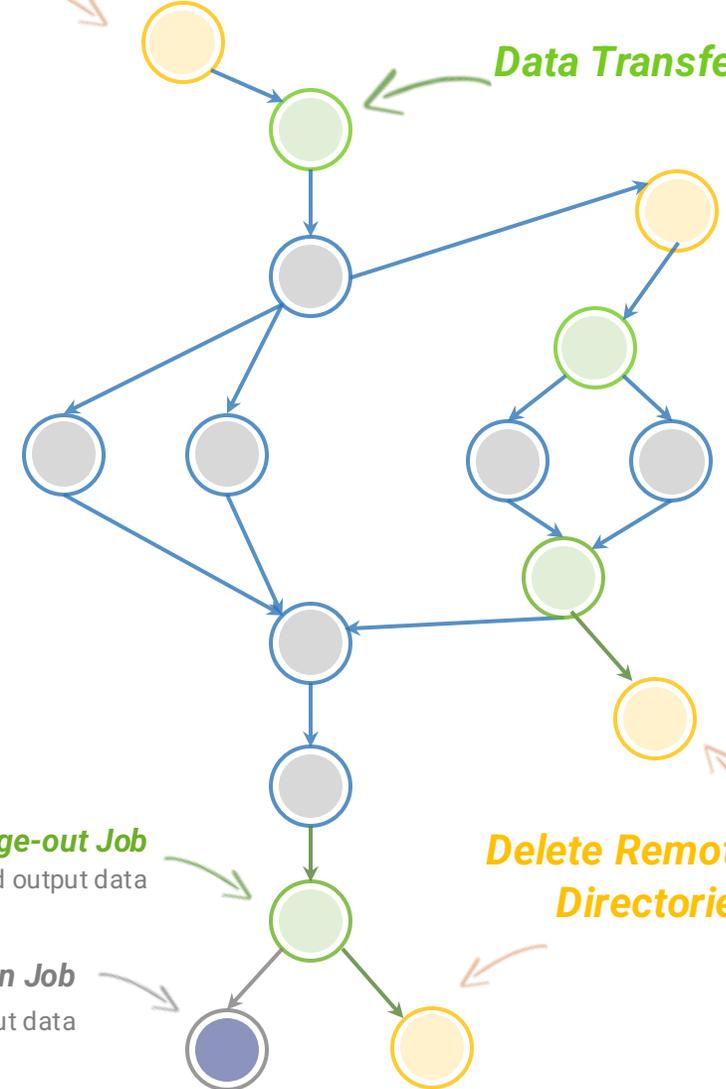
**Data Transfer Jobs**

**Stage-out Job**  
Stage-out generated output data

**Registration Job**  
Registers the workflow output data

**Delete Remote  
Directories**

**EXECUTABLE WORKFLOW**



# LIGO

## Supporting Science over Generations of CI



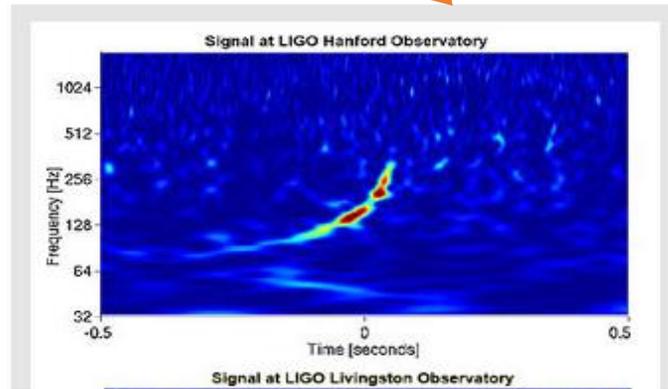
Nobel Prize

### Pegasus working with LIGO

(Laser-Interferometer Gravitational Wave Observatory)



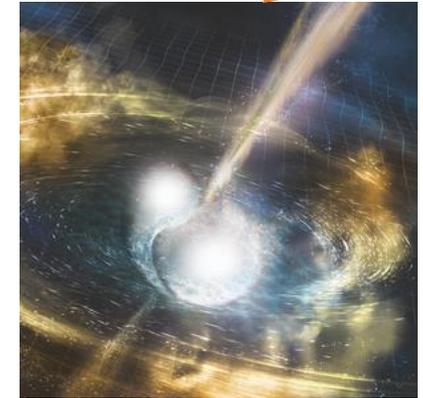
First Pegasus prototype



Blind injection detection



First detection of black hole collision



Multi-messenger neutron star merger observation

Image credit: LIGO Scientific Collaboration

# Compilation and Optimization

2003



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:

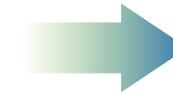
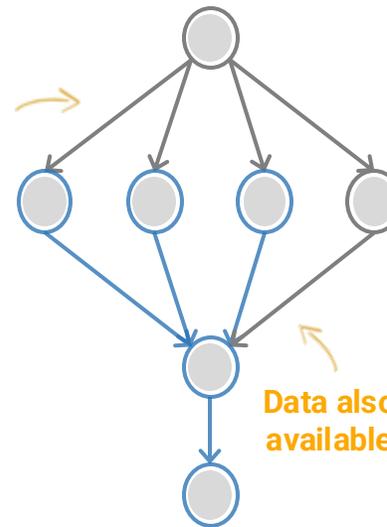
- Scheduling computational tasks
- Data-aware scheduling (for data locality or reduced data footprint)

## 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](mailto:deelman@isi.edu)

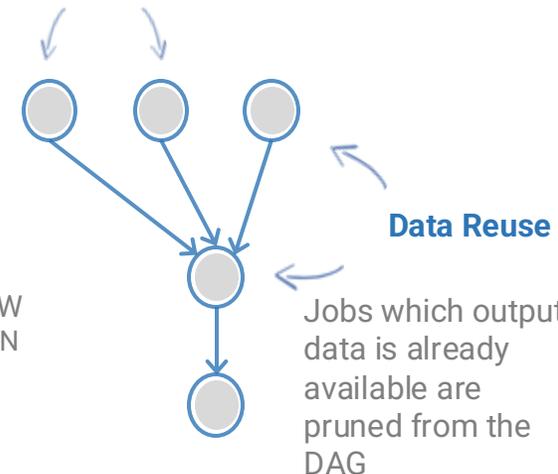
Raymond Plante  
National Center for Supercomputing Applications, Champaign, IL 61820  
[rplante@ncsa.uiuc.edu](mailto:rplante@ncsa.uiuc.edu)

Data already available



WORKFLOW REDUCTION

Data Reuse



*Supercomputing 2003.*

# 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 Ramakrishnan<sup>1</sup>, Gurmeet Singh<sup>2</sup>, Henan Zhao<sup>3</sup>, Ewa Deelman<sup>2</sup>, Rizos Sakellariou<sup>3</sup>, Karan Vahi<sup>2</sup>, Kent Blackburn<sup>4</sup>, David Meyers<sup>4,5</sup>, and Michael Samidi<sup>4</sup>

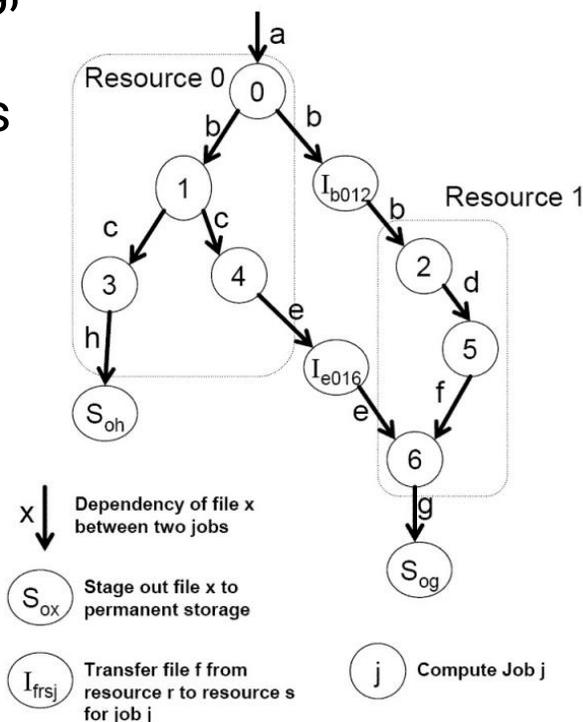


Figure 1. Executable workflow with 7 compute jobs mapped to two resources.

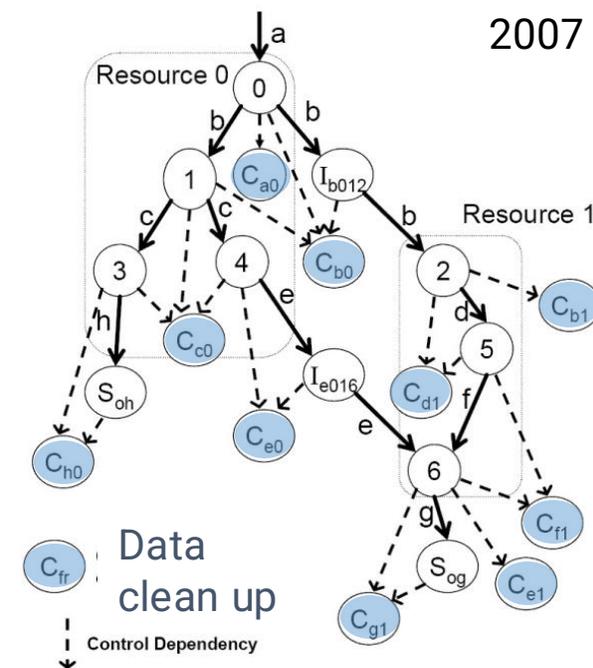
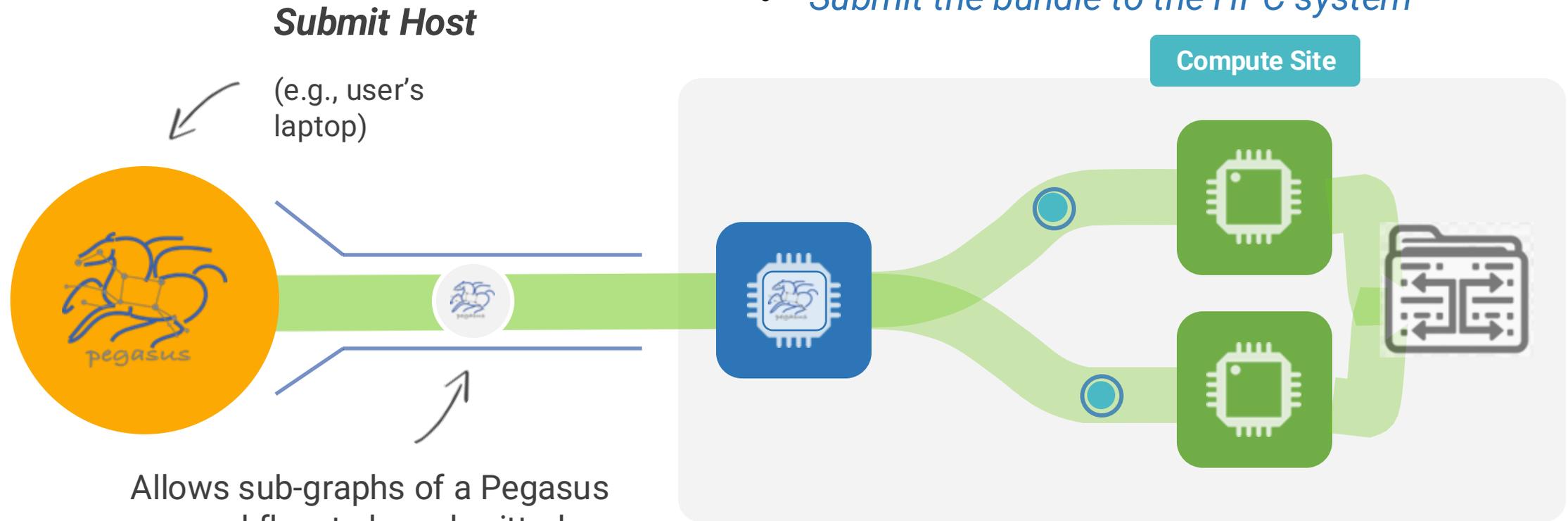


Figure 2. Cleanup nodes added to the executable workflow.

# 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*

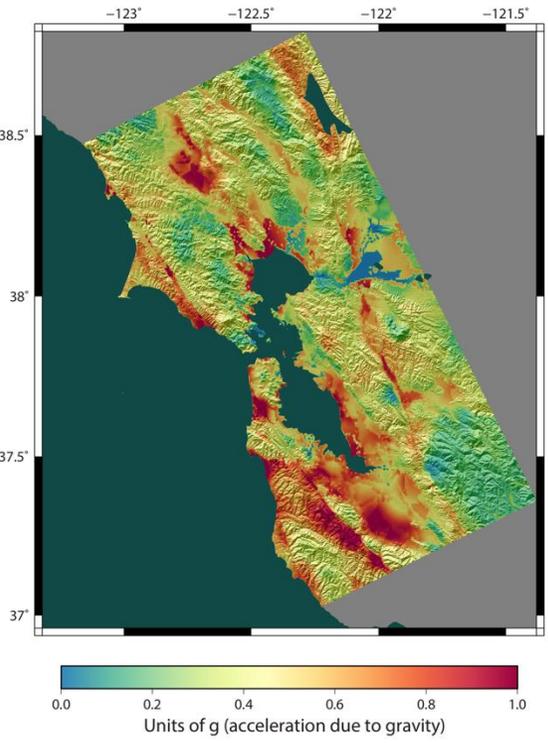


Allows sub-graphs of a Pegasus workflow to be submitted as monolithic jobs to remote resources

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

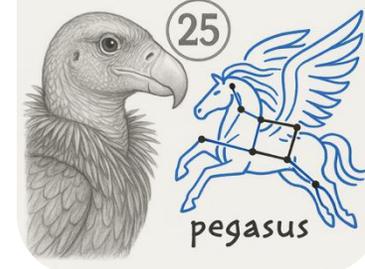
**55K**  
Frontier  
OLCF  
Node-hours

**125K**  
Frontera  
TACC  
Node-hours

2024

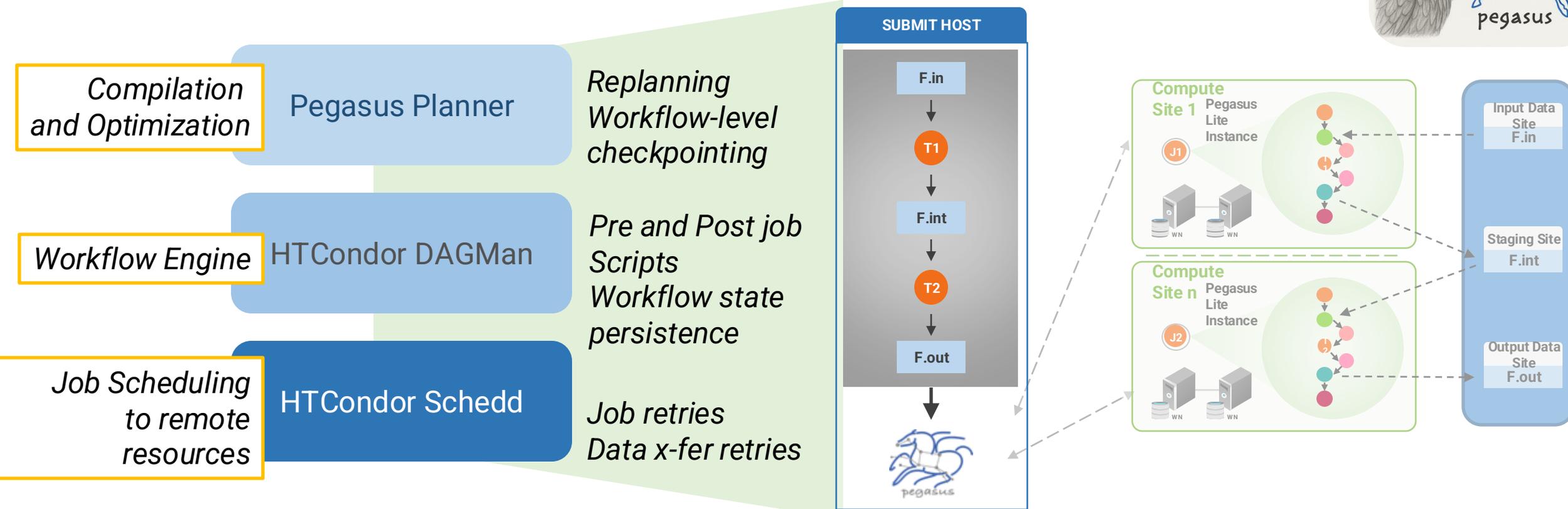
High-water marks were 4130 nodes on Frontier (44%)  
and 1029 nodes on Frontera (12%)

Neither used a reservation or priority bump.



# Layered Architecture

Leveraging proven technologies, building upon layers of resilience



Long-standing collaboration with Miron Livny and



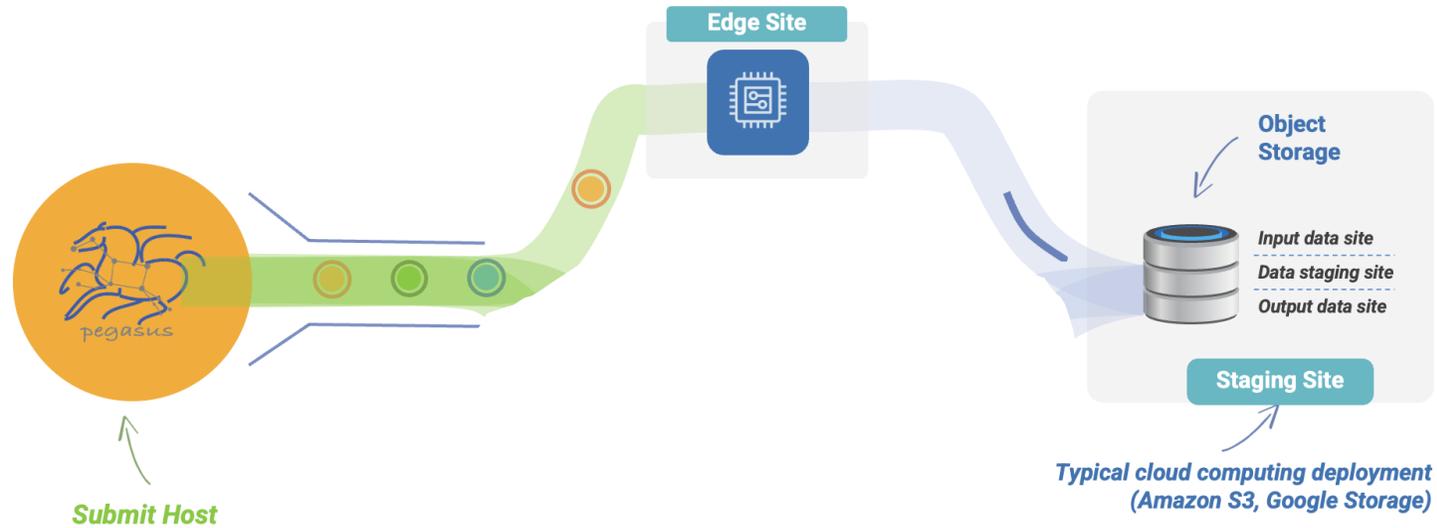
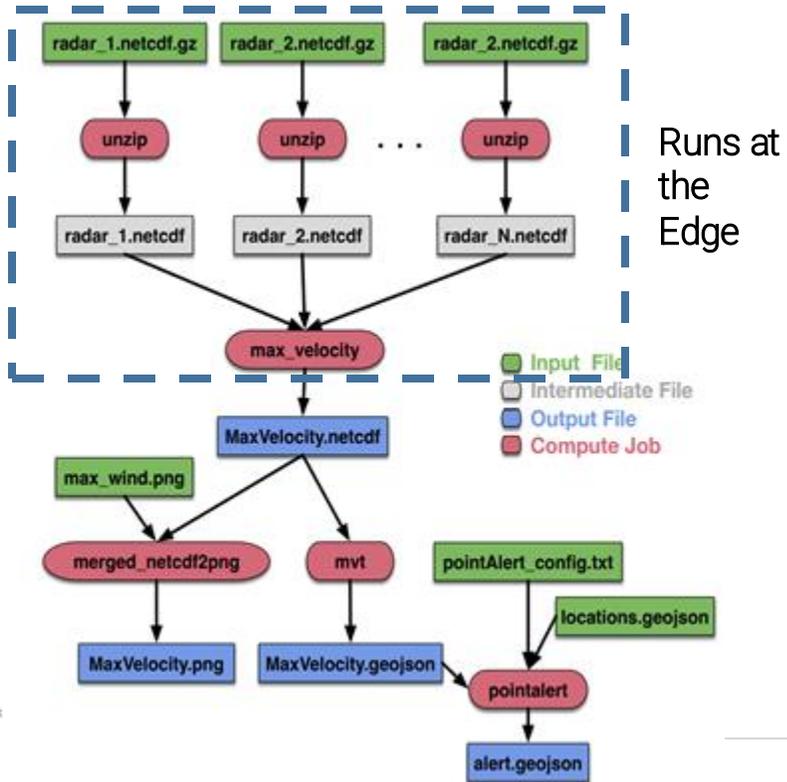
# Edge-2-Cloud Applications

CASA: Collaborative and Adaptive Sensing of the Atmosphere



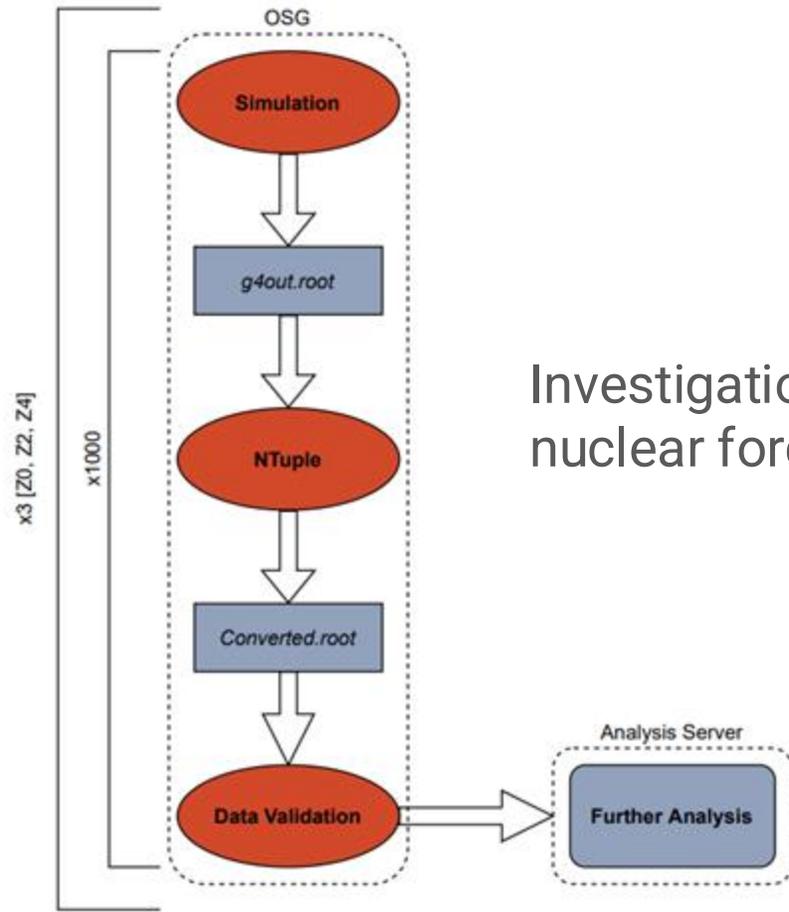
- Has deployed a network of short-range Doppler radars
- Compute and data repositories at the edge, close to the radars
- Use on demand cloud resources to scale up their computations

## Distributed Model, Data Footprint Management, Automated Data Staging



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

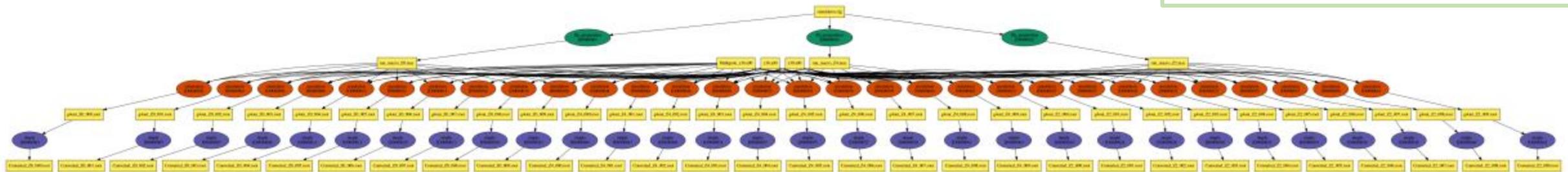
Connor Natzke,  
Physics Student  
Colorado School of Mines



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



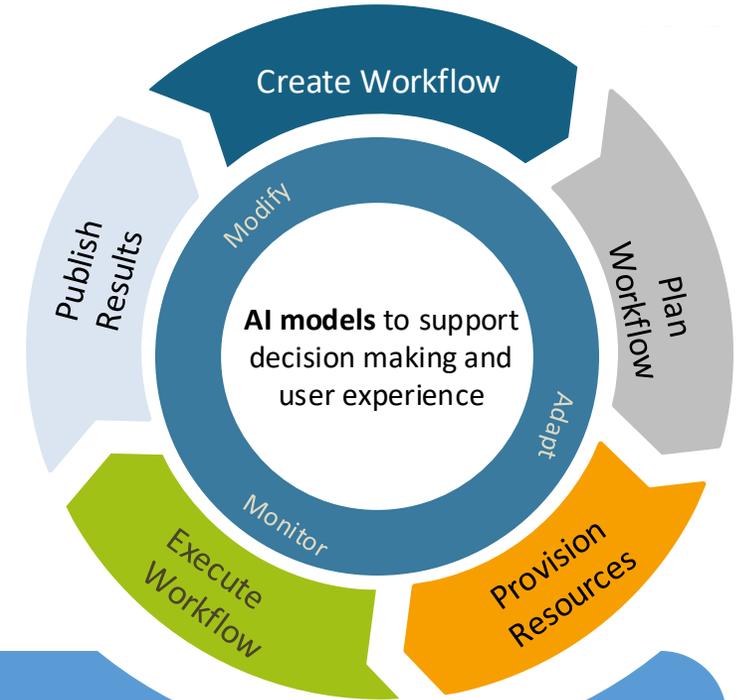
Manual Workflows



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



# 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





# 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 the following error: **Entered an error:**  
**\*\*"Expected local file**

**\*\*Root Cause:\*\***

- The required input file does not exist
- This prevents the transfer of data

**\*\*Next Steps:\*\***

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

The remaining unsubmitted workflow will be terminated. This error will enable further execution.

### 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



ChatGPT 5.0

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.)

container

simplify some more

job 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**

nts/children

be executed as a

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

and named,

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.

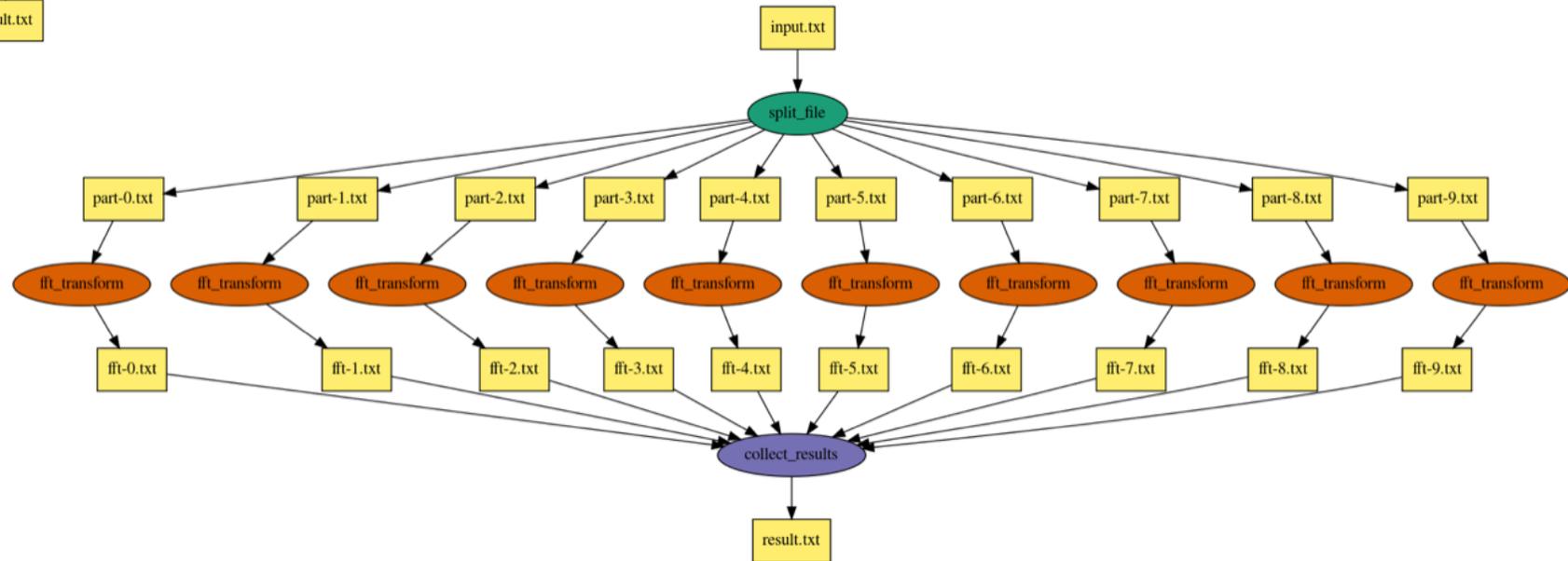
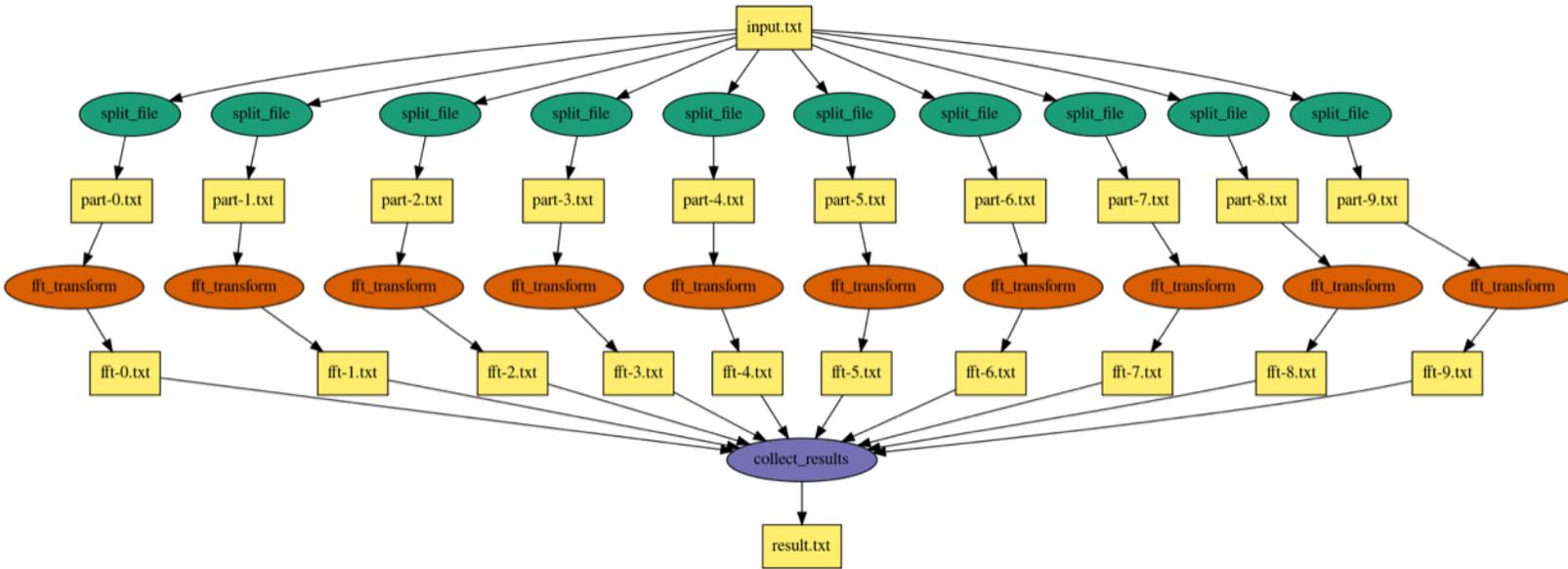


can you sum  
understanda

# Abstraction and Intent



ChatGPT



**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

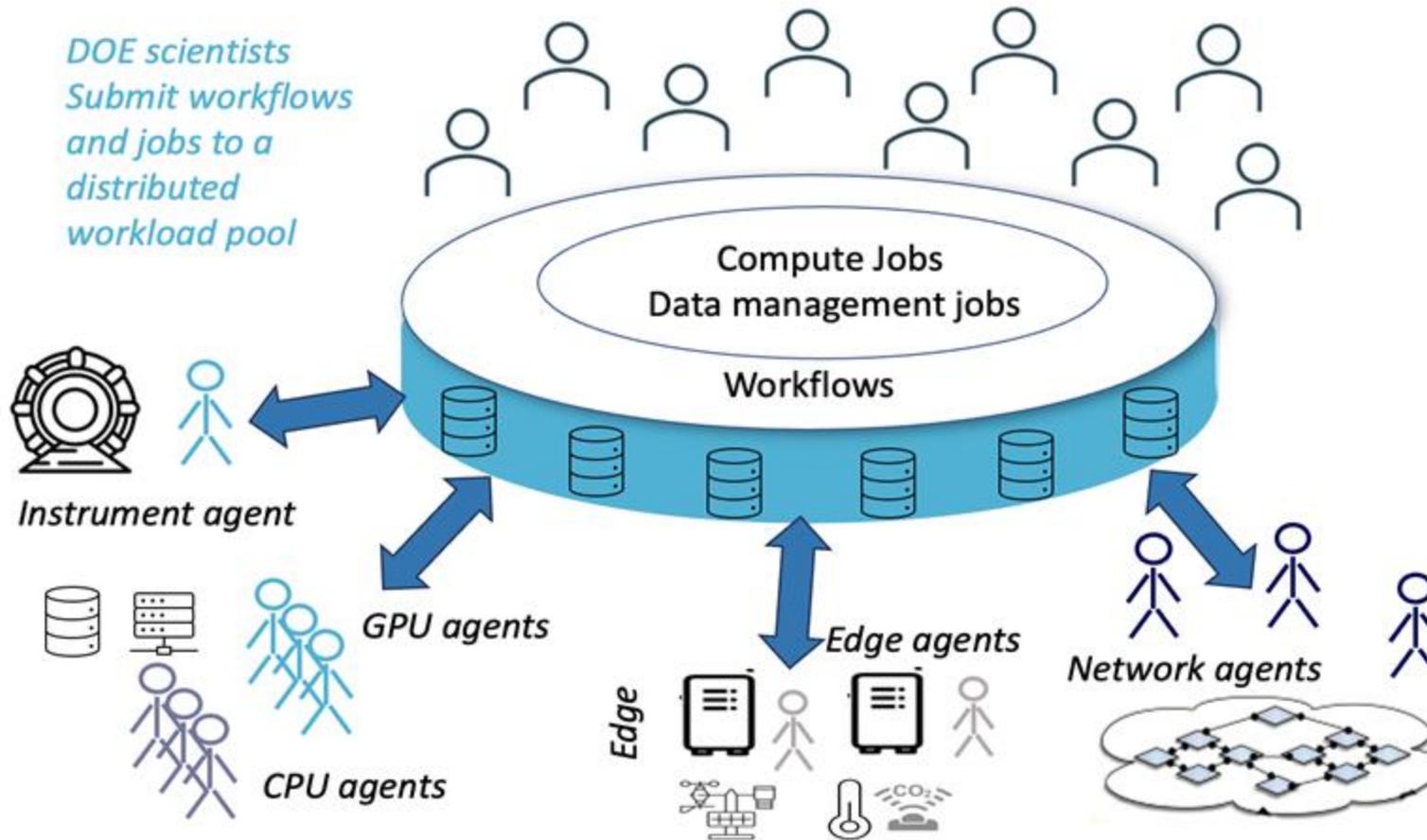
\* **Methods change over time as AI models learn**

**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 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



*Swarm Intelligence agents select workload to execute and autonomously adapt*

Intersection of Agentic AI  
and Distributed and High-  
performance computing

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

*Funded by DOE: DE-SC0024387  
2023- 2028*

# SWARM team



Ewa Deelman, Ph.D.  
USC



Prasanna Balaprakash, Ph.D.  
ORNL



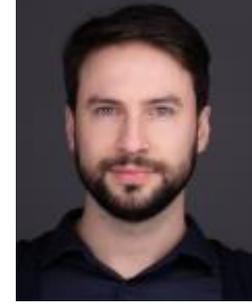
Anirban Mandal, Ph.D.  
RENCI



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

Fred Sutter, Ph.D.  
ORNL

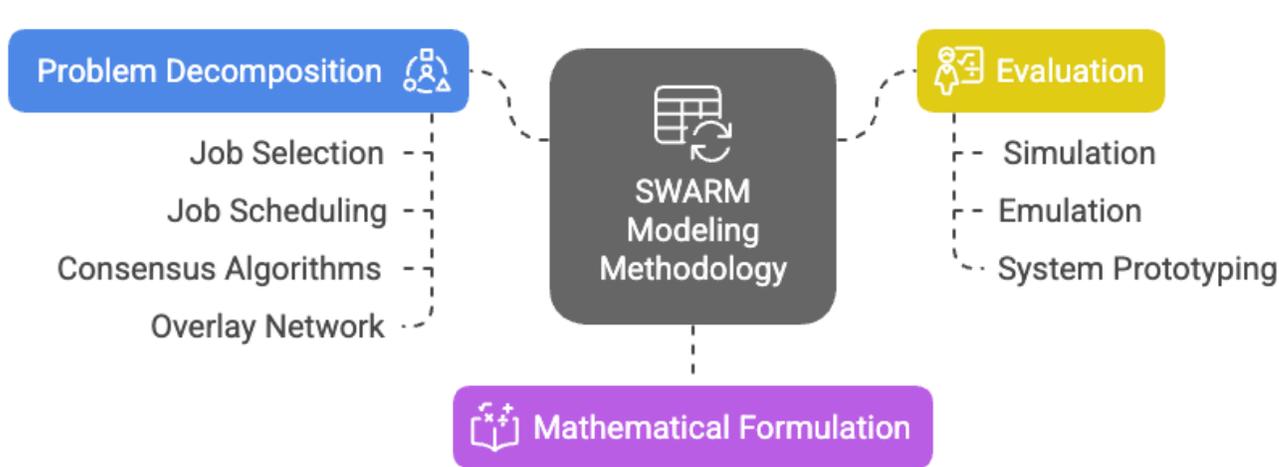
USC Viterbi  
School of Engineering

Argonne  
NATIONAL LABORATORY



OAK RIDGE  
National Laboratory

renci



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

*Utilizing Simulations, the NSF Fabric testbed, and DOE systems at ANL, LBNL, and ORNL*

# Pushing the boundaries of abstraction and automation in the scientific lifecycle

## Potential:

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

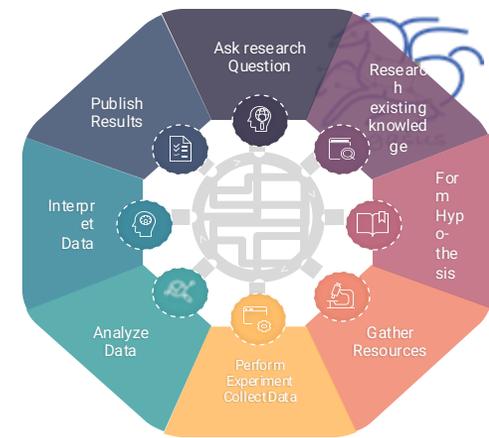


## Downsides of Cognitive Automation

- 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



## 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?

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

# 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

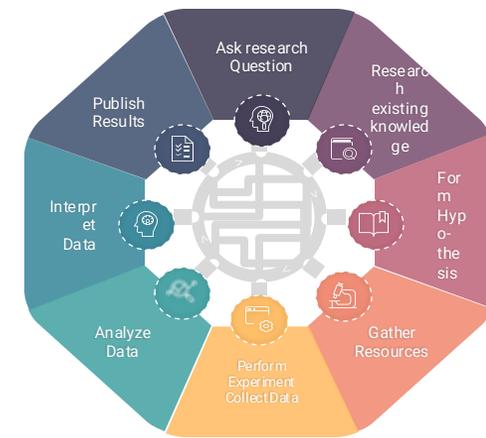
12.75K



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

# 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	Gurmeet Singh	Ken Kennedy	Nicole Virdone	Ruth Pordes	Thomas Caraco
Alana Romanella	Hamza Safri	Kent Blackburn	Nona Nersisyan	Ryan Mitchell	Thomas Fahringer
Anirban Mandal	Harshita Sahni	Kerk Kee	Norbert Meyer	Ryan Tanaka	Todd Miller
Andrew Pasquale	Henri Casanova	Kesheng Wu	O. Okunloye	Sandeep Chandra	Tom Peterka
Ann Chervenak	Hongwei Jin	Komal Thareja	Orcun Yildiz	Samuel Gupta	Tomasz Osinski
Alicia Esquivel	John Foster	Krishnan Raghavan	Patricia J Teague	Sandeep Kesring	Tony Hey
Amanda Randles	Johan Taylor	Laura Christopherson	Patrick R Brady	Sarowar Hossain	Trilce Estrada
Shishir Bharati	I. Santana-Perez	Laura Pearlman	Patrycja Krawczuk	Satoshi Matsuoka	Tristan Glatard
Angela Murillo	Igor Sfiligoi	Laurens Versluis	Paul Groth	Scott Koranda	Tu Mai Anh Do
Boleslaw Szymanski	Jacopo Petri	Lavanya Ramakrishnan	Paul Ruth	Scott D. Srinivasan	Udit Catalyurek
Bruce Hendrickson	Jukka Aittinen	Lassina Sarr	Paula Jajava	Scott C. Brumby	Verie E Taylor
Brian Tierney	Paolo Falda	Loic Pottier	Pawel Gepner	Sebastian Schuster	Verie Taylor
Bruce G Berriman	Imtiaz Mahmud	Louis H Ziantz	Pawel Zuk	Sepideh Azarnoosh	Valerio Pascucci
Carl Kesselman	Ivona Brandic	Luc Moreau	Philip Maechling	Shava Smallen	Veronika Nefedova
Cassandra Hayes	Jack Wells	Maciej Malawski	Prachi Jadhav	Shelley Knuth	Vickie Lynch
Charles Vardeman	Jack Dongarra	Mahedi Hasan	Prasad Calyam	Shishir Bharathi	Victoria Stodden
Christina Hoffa	James Blythe	Malcolm P Atkinson	Prasanna Balaprakash	Shixun Wu	Vikram Adve
Chris Carothers	James A Knowles	Manish Parashar	Rafael F. Da Silva	S. Chandrasekaran	Vikram Gazula
Daniel Reed	James Howison	Marcia McNutt	Rajiv Mayani	Silvina Caino-Lores	Viktor Prasanna
Daniel Gunter	James Blythe	Marcus Thiebaux	Rajive L Bagrodia	Simon Miles	Vincenzo De Maio
David Abramson	James C Browne	Maria A. Rana	Raphael Balza	Sonal Jain	Vinay Kumar



U.S. National Science Foundation



U.S. DEPARTMENT OF ENERGY



National Institutes of Health

Turning Discovery Into Health

