Use of AI for Scientific Workflows

Ewa Deelman

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

https://pegasus.isi.edu

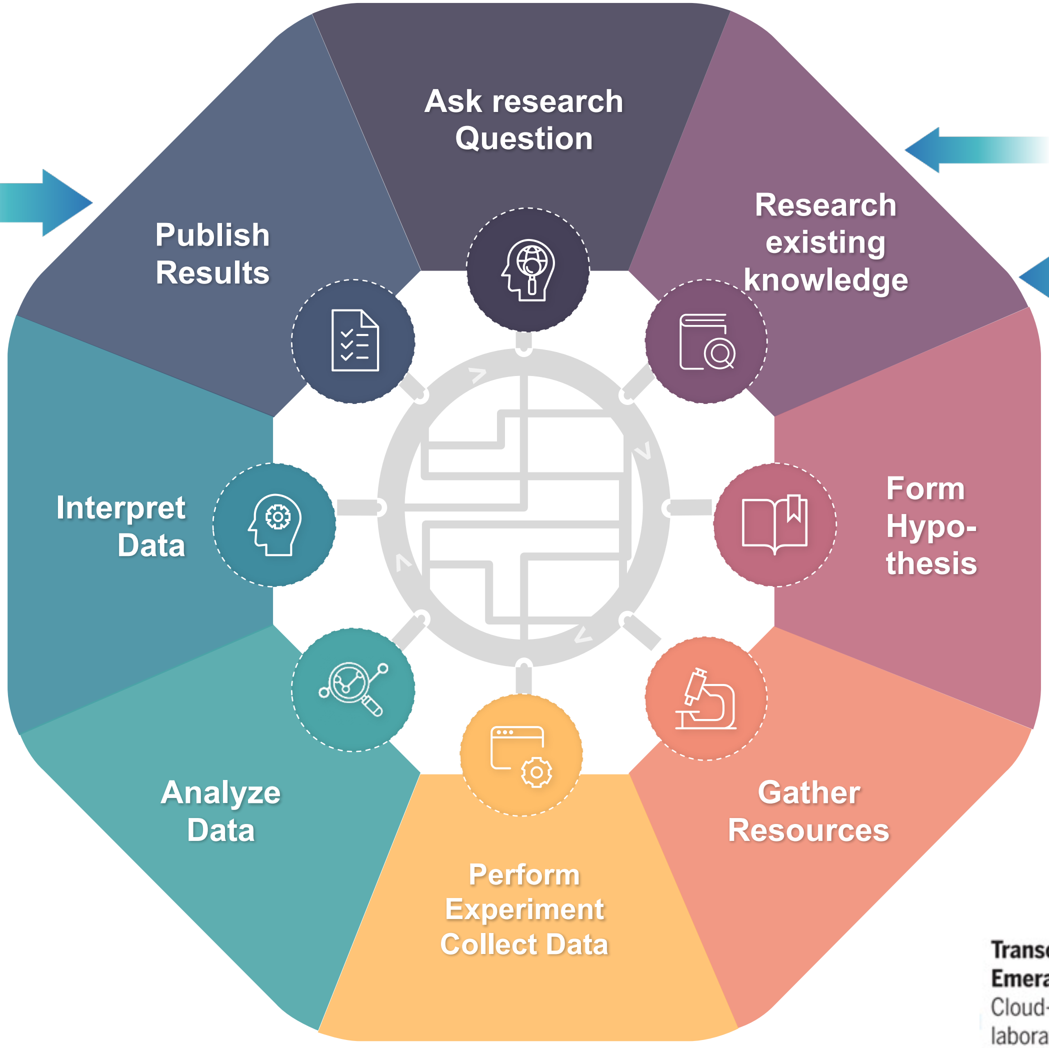
This work is funded by DOE under grant #DE-SC0022328

Exploring a Scientific Question

AI is changing the Science Lifecycle

Citeomatic
A free online tool that reads your paper and predicts what citations are missing.

Nutonian
A software platform that ingests very large data sets and spits out a mathematical theory that explains the patterns in the data.



Semantic Scholar
A search engine that extracts not just words from papers, but graphs and "influential" citations.

Iris.AI
A browsing tool for exploring scientific papers by the concepts that link them.

Zymergen
A company with an AI that tracks thousands of variables while tweaking microbe genomes (see main story, p. 18).

Transcriptic, Emerald Cloud Lab
Cloud-based robotic laboratories for remotely doing automated molecular and cellular biology experiments.

** From the Cyberscientist paper*

NEWS

The cyberscientist

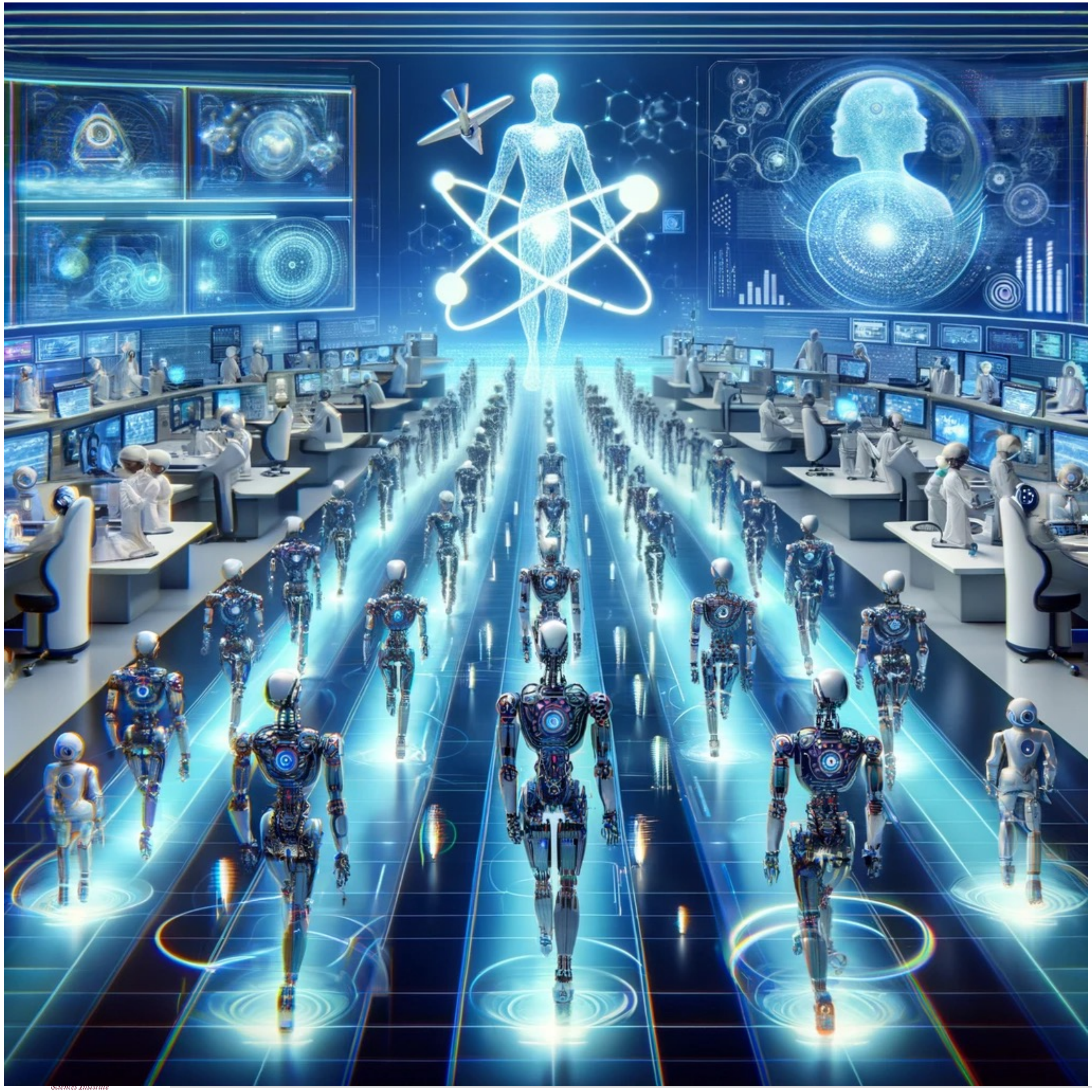
John Bohannon

+ See all authors and affiliations

Science 07 Jul 2017:
Vol. 357, Issue 6346, pp. 18-21
DOI: 10.1126/science.357.6346.18

USC Viterbi School of Engineering Information Sciences Institute

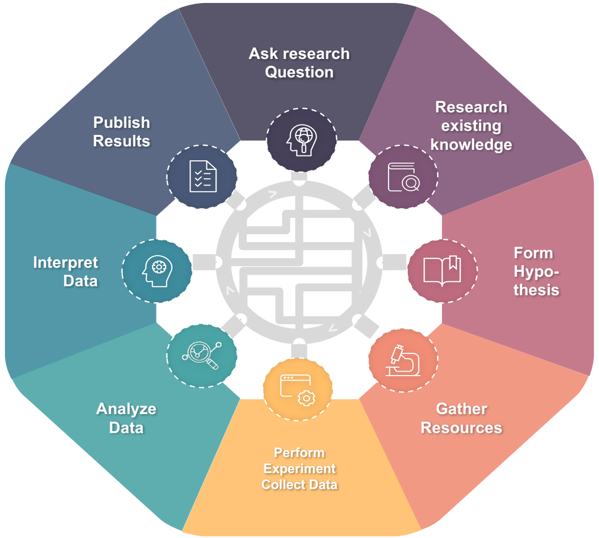
OpenAI ChatGPT 4.0



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

How will we teach science?
How will we share knowledge?
How will the work of scientists look like?

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



What kind of shaking can you expect in Southern California in the next 50 years?

Useful information for:

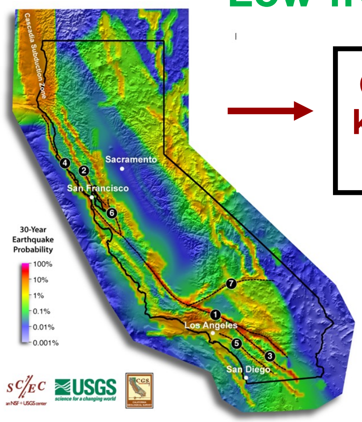
- ✓ Building engineers
- ✓ Disaster planners
- ✓ Insurance agencies

- Explore historical earthquakes
- Locate the seismic faults
- Explore what models of future earthquakes are available
- Decide which areas you want to consider

- Generate a recipe/workflow to answer the question
- Find available compute resources

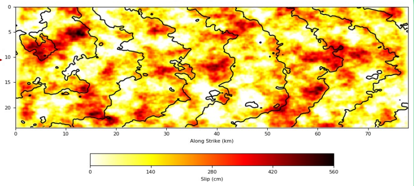
Generating Physics-Based Seismic Hazard Maps of Southern California

Low-frequency CyberShake workflow



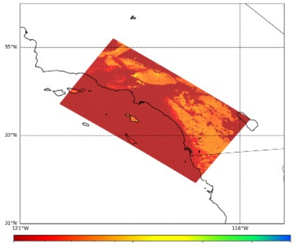
Uniform California Earthquake Rupture Forecast

Graves & Pitarka kinematic rupture generator

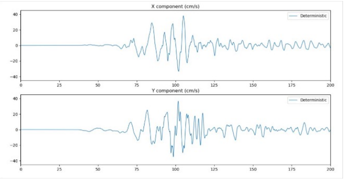


600,000+ events

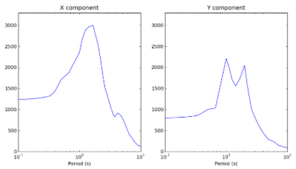
UCVM → **AWP-ODC-SGT wave propagation** → **Low-frequency seismogram synthesis** → **Intensity measures** → **Aggregate data products**



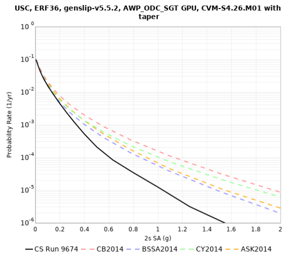
Velocity Mesh



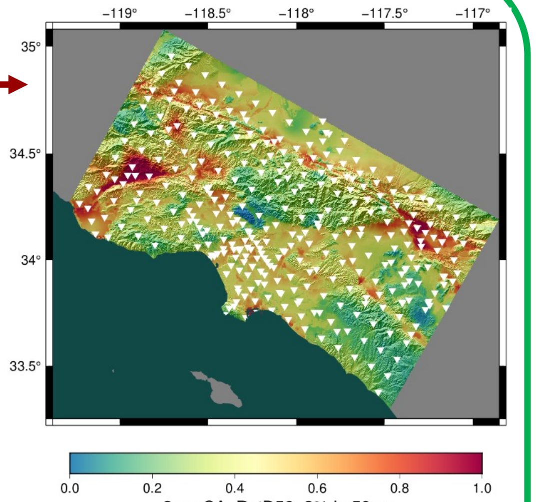
0-1 Hz low-frequency seismograms



RotD50, PGA, PGV



Hazard Curve



Hazard Map

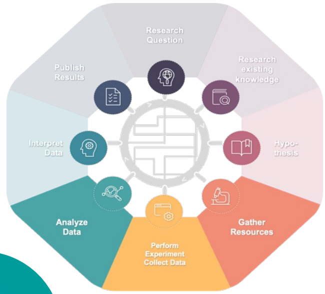
CyberShake Computational Requirements

CyberShake Stage	Number of Tasks	Node-Hours	Output Data
Velocity mesh creation (parallel)	1	10 CPU	300 GB
Wave propagation (parallel)	2	80 GPU	1500 GB
Low-frequency seismogram synthesis (parallel)	1	1000 CPU	38 GB
High-frequency seismogram synthesis (serial)	77,000	1000 CPU	187 GB
Total, 1 site (including small jobs)	77,020	2090	2025 GB
Total, full region	**25.8 million**	**700,000**	**680 TB**

- Large computational and data requirements
- Mix of large parallel CPU and GPU jobs with HTC
- High degree of automation required to support continuous execution

108 days of execution on ORNL's Summit using the the Pegasus Workflow Management System

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)

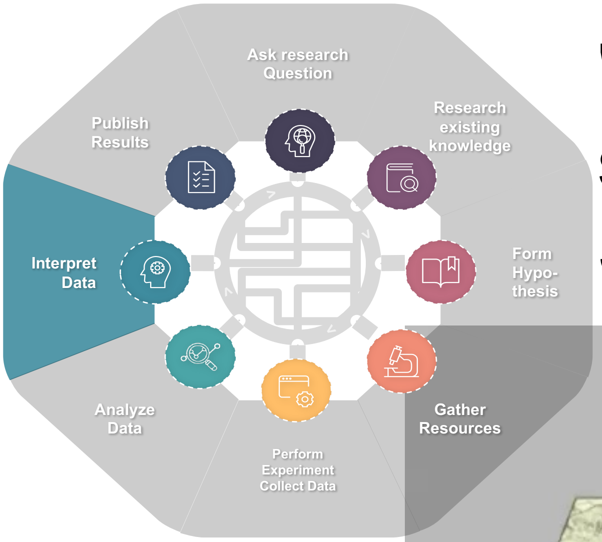
Deals with resources/software that change over time

Ease of use. Ability to monitor and debug large workflows

Our Focus

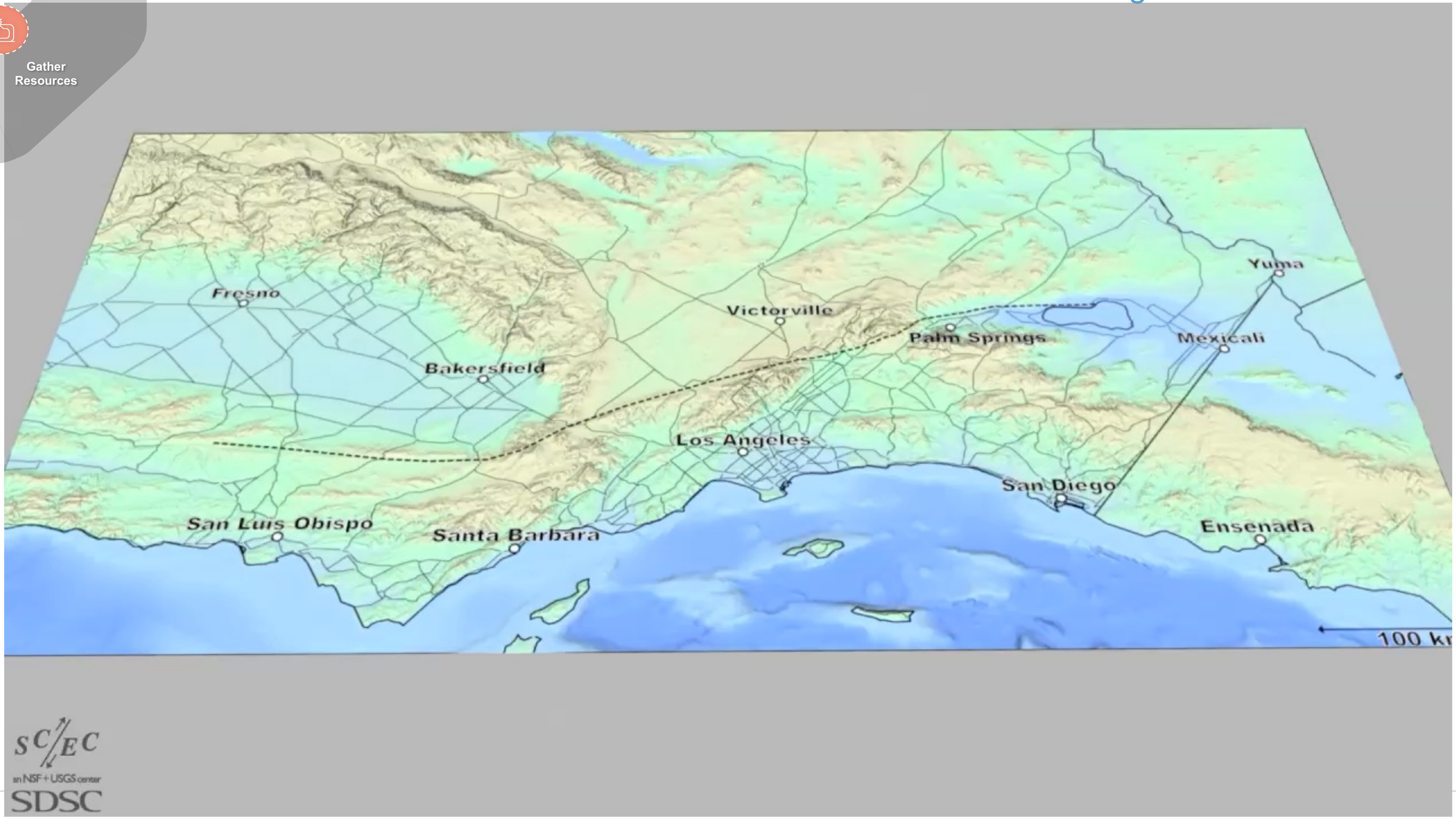
- Separation between workflow description and workflow execution
- Workflow planning and scheduling (scalability, performance)
- Task execution (monitoring, fault tolerance, debugging, web dashboard)
- Workflow optimization, restructuring for performance and fault tolerance.

USC Viterbi School of Engineering Information Sciences Institute



"M8" simulation models how the ground will shake in a magnitude 8.0 earthquake on the Southern San Andreas Fault.

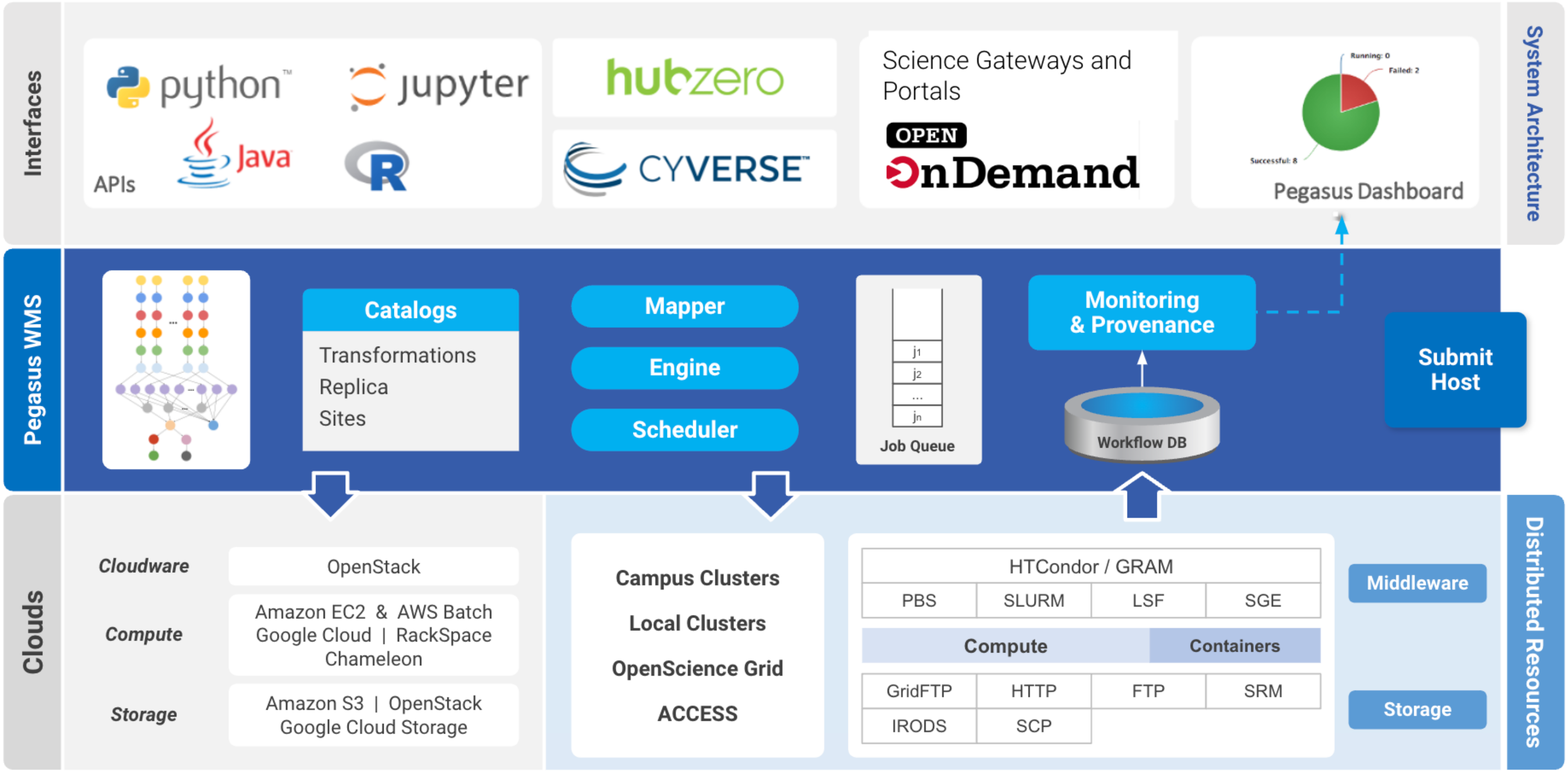
Credit: Southern California Earthquake Center
UC San Diego



Likelihood of M8 earthquake in the next 30 years was in 2% in 2010

USC Viterbi School of Engineering Information Sciences Institute

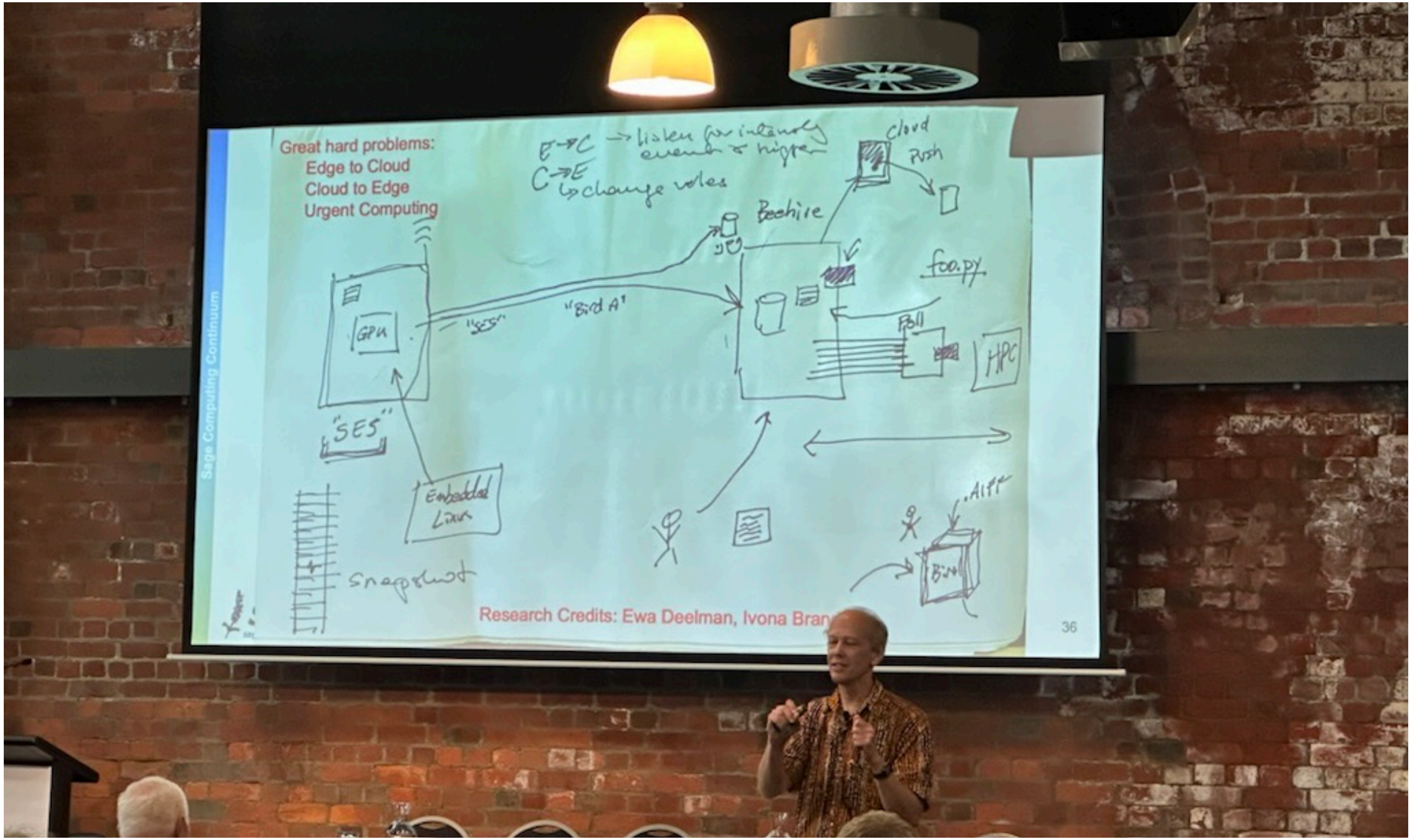
“Up and down” integrations with diverse CI, common languages, and Portal/GUI interfaces



MulticoreWorld

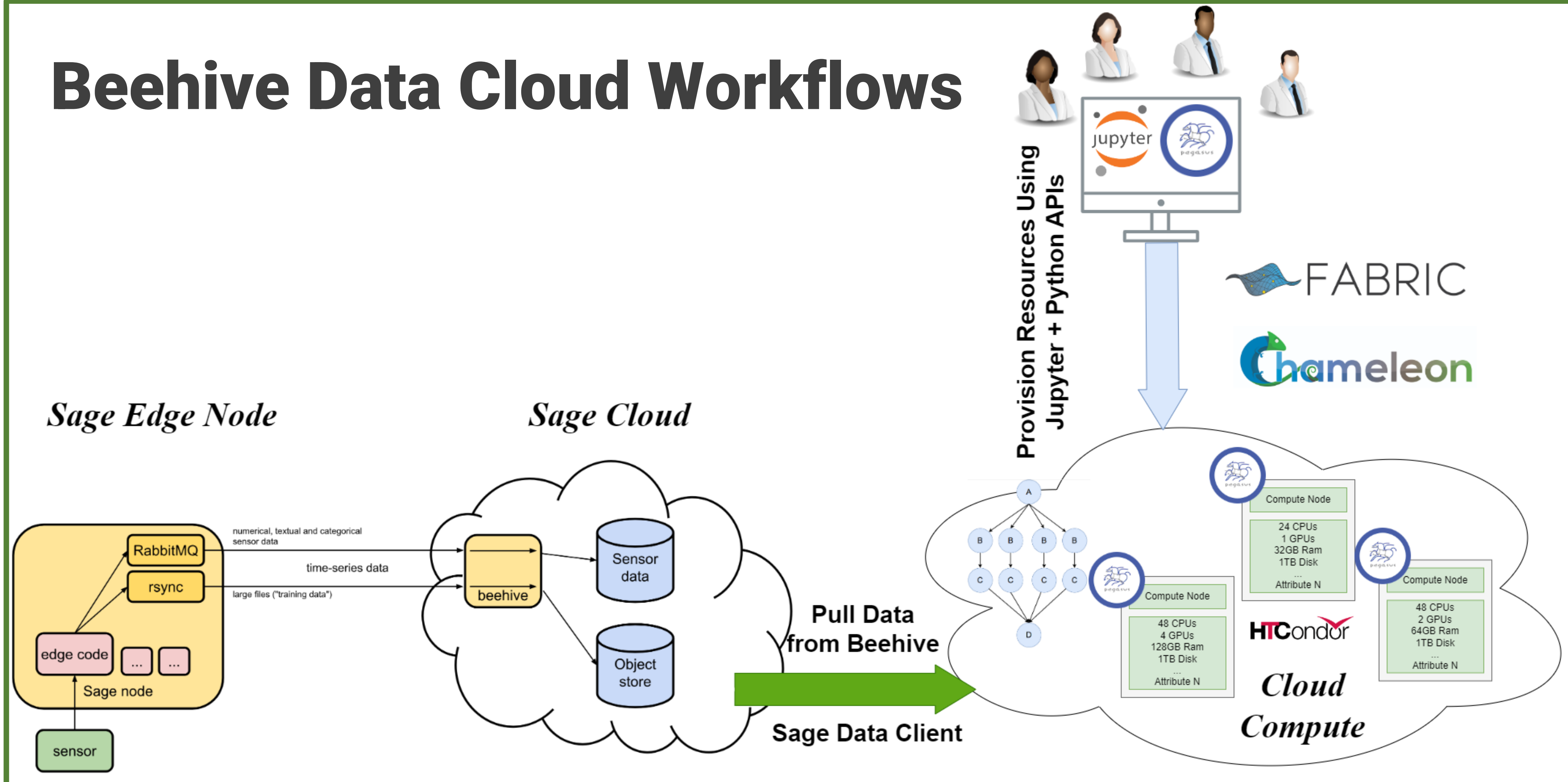
2023

pegasus

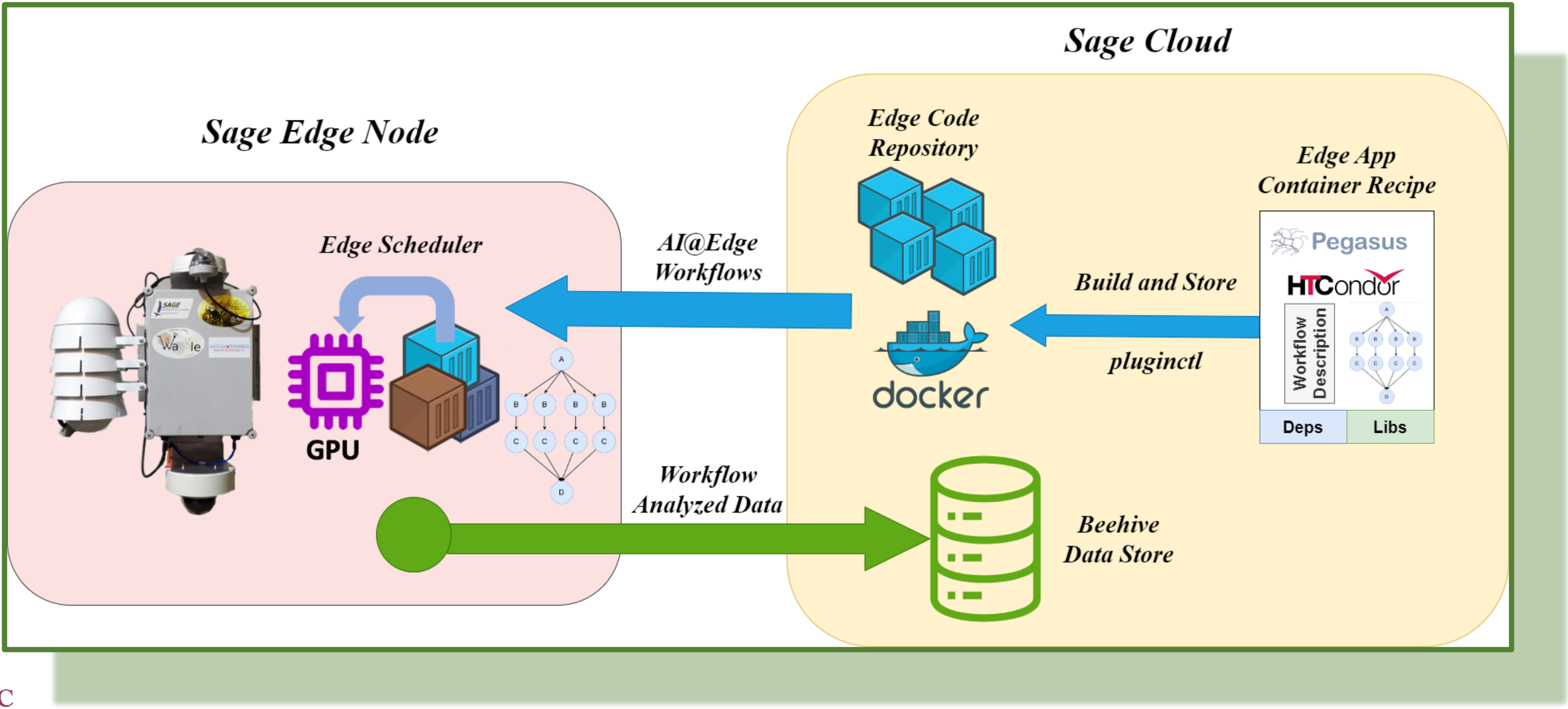


USC Viterbi
School of Engineering
Information Sciences Institute

Beehive Data Cloud Workflows



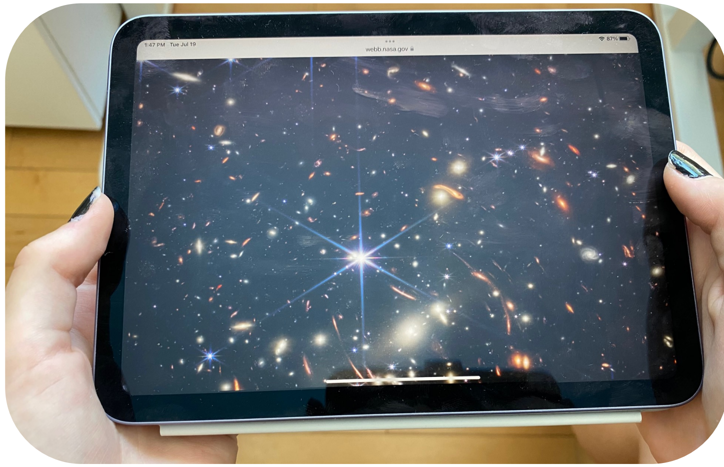
Workflow in a Box



Why is the science lifecycle changing?

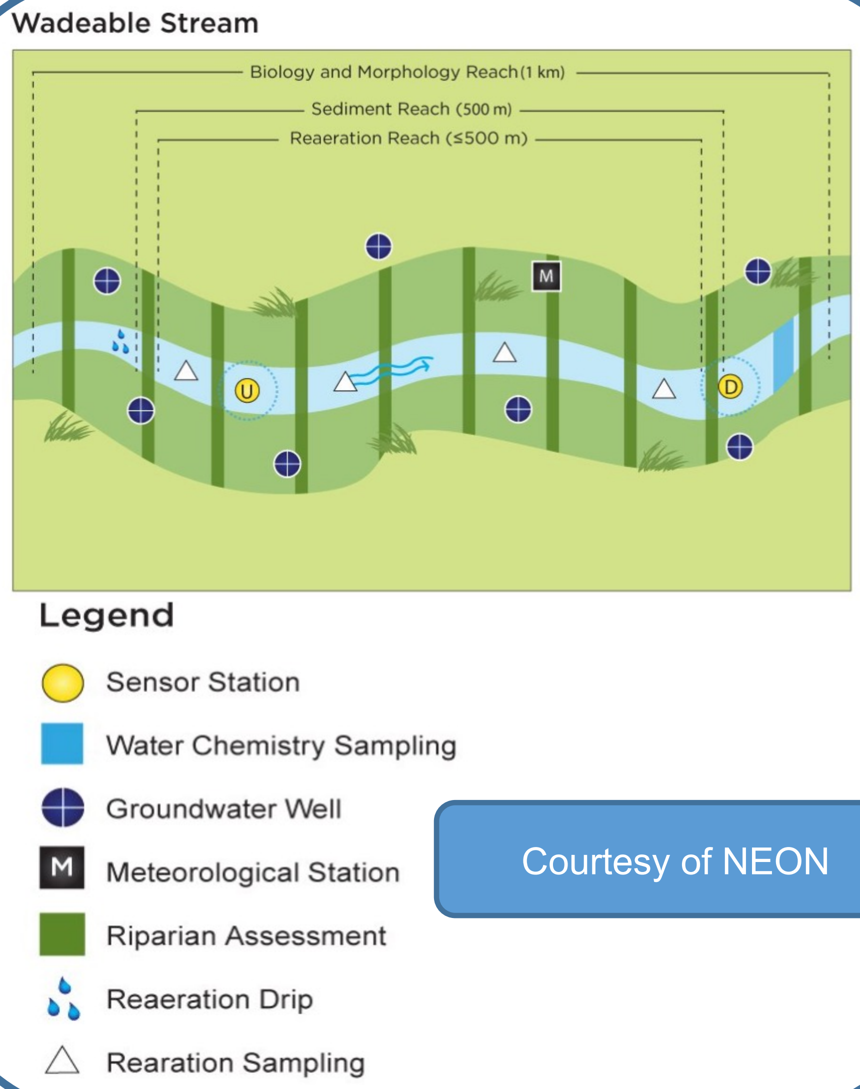
Users' Experiences and Expectations

- Users are often not exposed to complex programming
- Users are not exposed to command-line interfaces
- Users have uneven access to Cyberinfrastructure (even networks)
- Expect easy to use, intuitive interfaces
  - Graphical, conversational, common behavior
- Expect robust systems that are fault tolerant and adaptable
- Want quick response time and/or good information
- Current cyberinfrastructure (CI) is very complex, heterogeneous, and fragmented
- Even simple tasks (remote job submission, monitoring, debugging) are difficult (2-factor authentication)
- Limited support for long-running services and dynamic resource management



James Webb Telescope Image

Means and Methods are changing



Courtesy of NEON

- **More Data and Data Sources**
  - Small and cheap sensor devices
  - Sophisticated instruments (CryoEM ~7TB per experiment)
  - Large-scale data archives: LHC, NASA, (Rubin will produce 20TB/night)
- **Faster Networks**
  - 100Gbps, ESnet6 > 400 Gbps
- **Access to more computing**

Means and Methods are changing

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

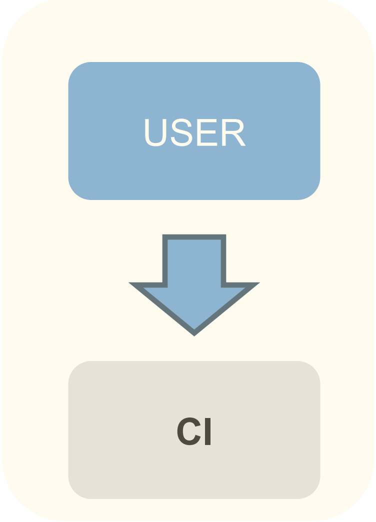
Growing emphasis on Verification/validation/UQ

Growing need for Automation

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

USC Viterbi School of Engineering Information Sciences Institute

What can we do better?



ChatGPT for workflow creation

DE

Create a Pegasus WMS Python workflow generator for a workflow that has 3 steps, first step divides the input data file into 10 pieces, the second step takes each file and does and fft on it, and the final step collects the results.

Pegasus is a workflow management system that allows users to design workflows in abstract terms and execute them efficiently and reliably on complex computing resources. Writing a workflow to divide a file into 10 pieces, apply Fast Fourier Transform (FFT) to each piece, and then collect the results can be done by defining tasks and dependencies between them in a workflow.

Here's an example of how you might set up this workflow using Pegasus WMS with Python:

1. **Step 1:** Divide the input data file into 10 pieces.
2. **Step 2:** Perform an FFT on each piece.
3. **Step 3:** Collect the results.

First, ensure you have the necessary dependencies installed, including Pegasus:

bash Copy code

```bash
pip install Pegasus-wms
```

USC Viterbi School of Engineering Information Sciences Institute

ChatGPT for workflow creation

```
# Create the workflow
wf = Workflow("fft_workflow")

# Create the input file
input_file = File("input.txt")

# Step 1: Divide the input data file into 10 pieces
output_files_step1 = []
for i in range(10):
    output_file = File(f"part-{i}.txt")
    output_files_step1.append(output_file)
    job_split = Job(split_file)\
        .add_args(input_file, i, 10)\
        .add_inputs(input_file)\
        .add_outputs(output_file)
    wf.add_jobs(job_split)

# Step 2: Perform FFT on each piece
output_files_step2 = []
for i, input_file in enumerate(output_files_step1):
    output_file = File(f"fft-{i}.txt")
    output_files_step2.append(output_file)
    job_fft = Job(fft_transform)\
        .add_args(input_file)\
        .add_inputs(input_file)\
        .add_outputs(output_file)
    wf.add_jobs(job_fft)
```

You would need to have corresponding scripts `split_file.py`, `fft_transform.py`, and `collect_results.py` to carry out the splitting, FFT, and collection processes respectively.

Make sure to customize the paths to the transformation scripts and modify the transformations to suit your specific requirements. The above example sets up the structure of the workflow, but the details of each task would depend on the specific operations you want to perform on the da

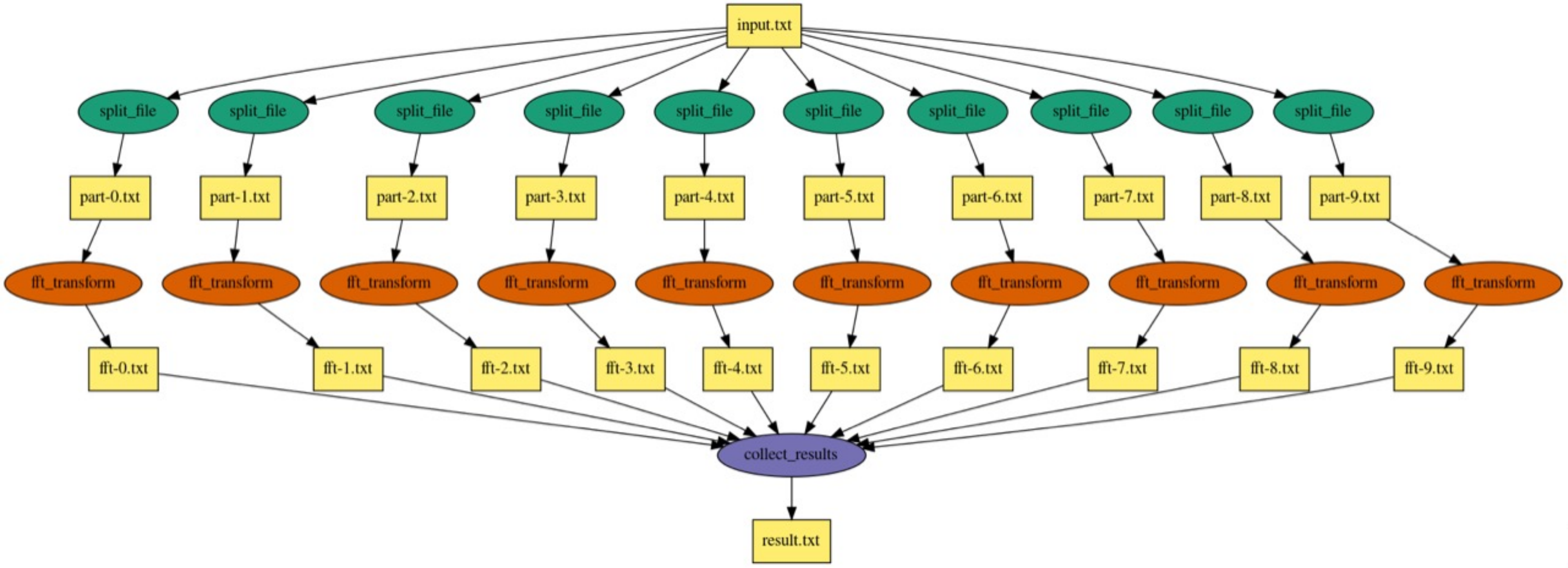
Magic number
You can ask ChatGPT to fix

```
# Define the number of pieces
n = 10 # Change this to the desired value

# Step 1: Divide the input data file into n pieces
output_files_step1 = []
for i in range(n):
    output_file = File(f"part-{i}.txt")
    output_files_step1.append(output_file)
    job_split = Job(split_file)\
        .add_args(input_file, i, n)\
        .add_inputs(input_file)\
        .add_outputs(output_file)
    wf.add_jobs(job_split)

# ... (rest of the code remains the same)
```

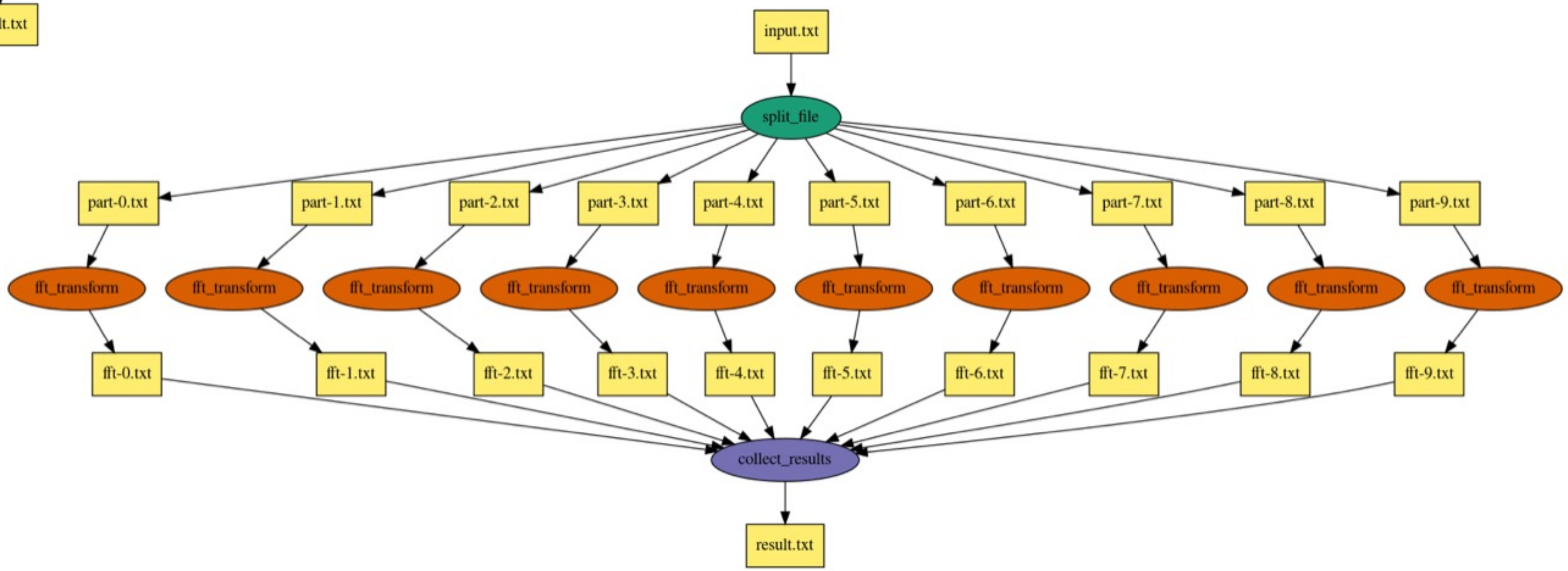
ChatGPT for workflow creation



ChatGPT

Karan Vahi, M.S.

SENIOR COMPUTER SCIENTIST

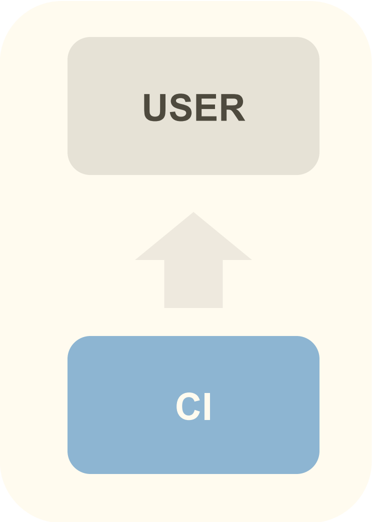


USC Viterbi School of Engineering Information Sciences Institute

What can we do better? Can we use ML to make our systems "smarter"/more autonomous

PoSeiDon

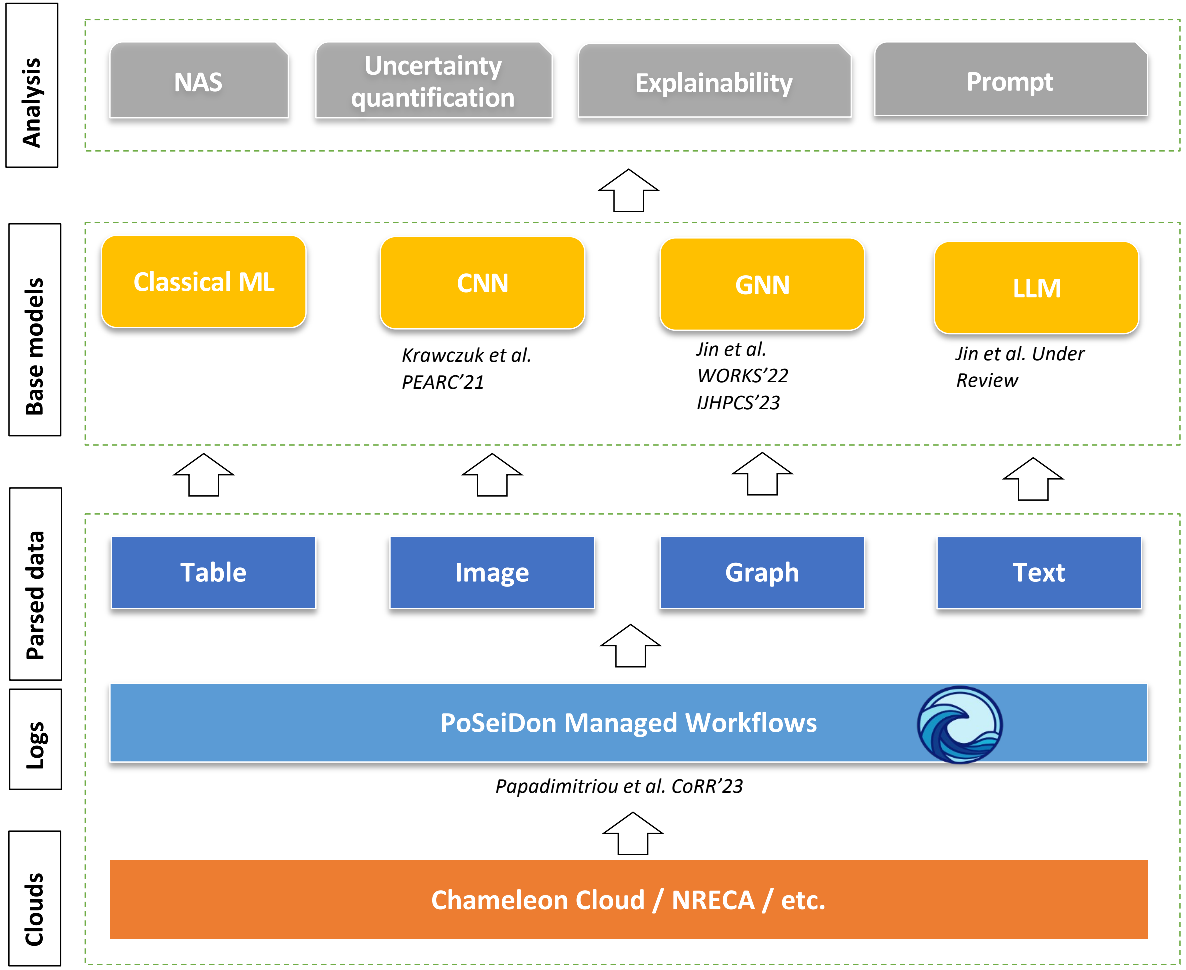
- **Anomaly detection**
- **Anomaly/error classification and attribution**
- **Better workflow adaptation based on failures and anomalies**
- Challenges:
  - Collect enough (quality data, richness, balanced class representation)
  - Enough labeled data, need to augment data
  - Structure (normalize, scale, transform) the data in a way that is amenable to the application of current techniques (or develop new ones)
  - Select the appropriate ML algorithms or architectures
    - DL hyperparameter optimization (learning rate, #epochs, hidden layers, activations functions..)



This work is funded by DOE under grant #DE-SC0022328

Anomaly Detection Framework

- **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.

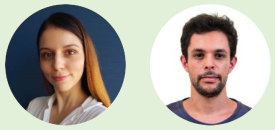


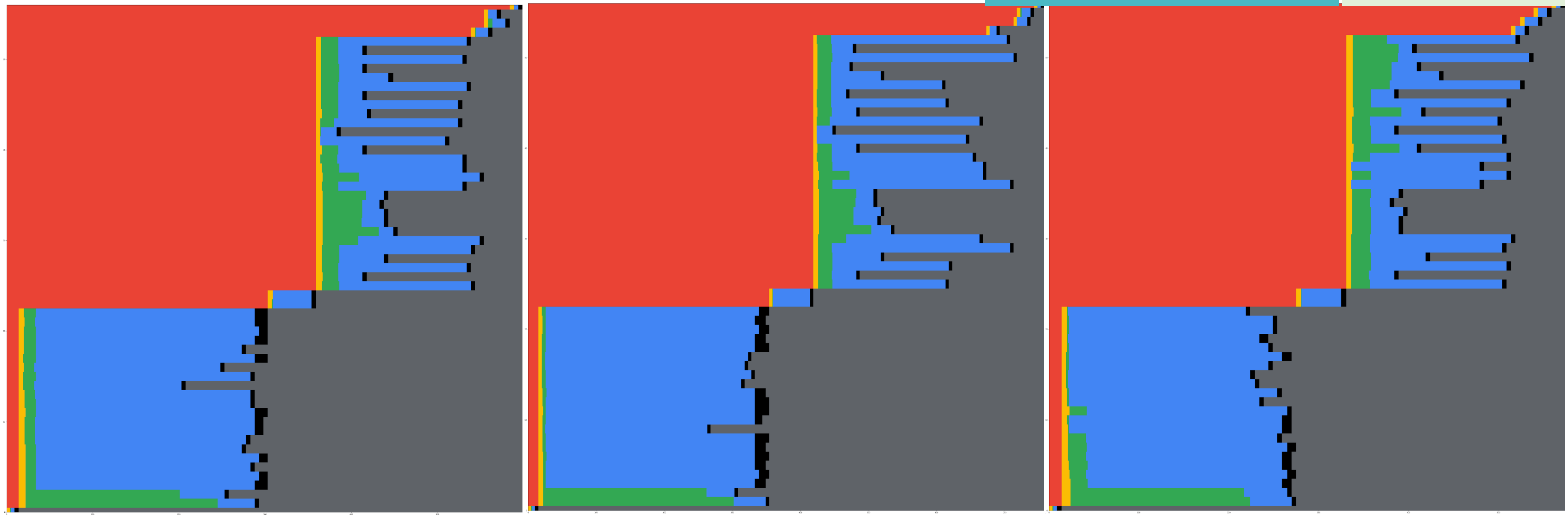
Anomaly Detection Framework

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

Robust Execution: Anomaly Detection and Classification using Graph Neural Networks (GNN)

PoSeiDon

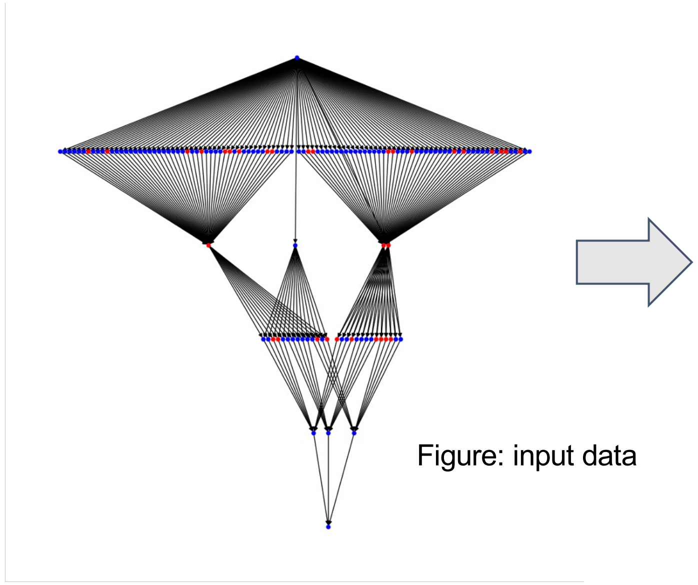


Figure: input data

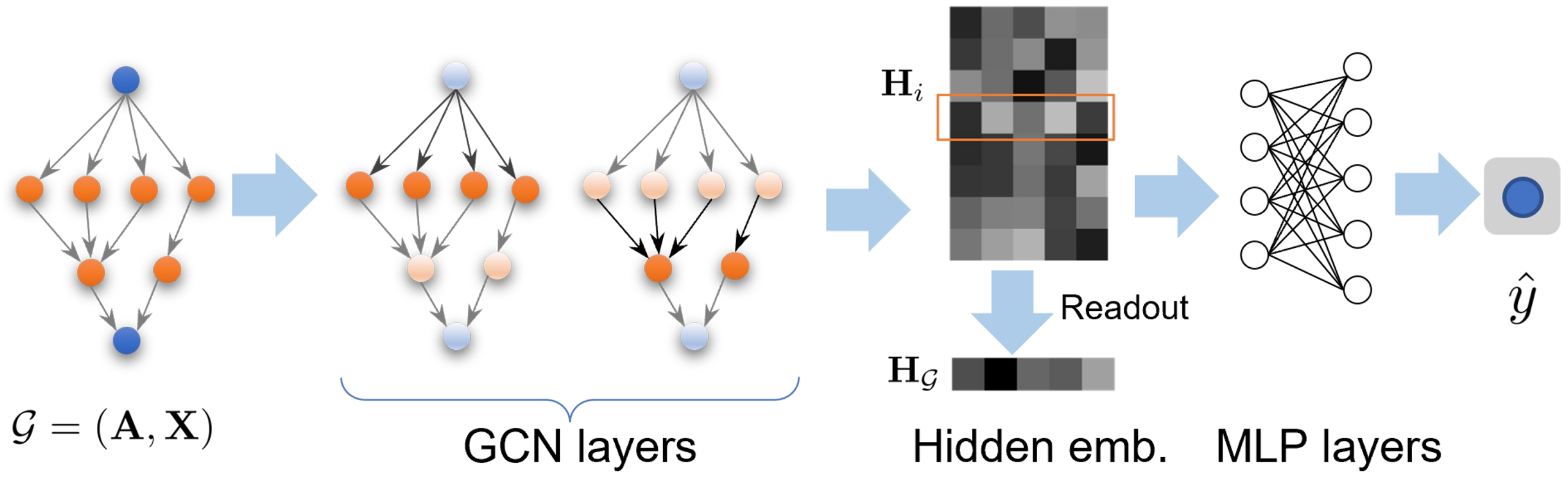


Figure: Graph neural networks architecture

Input: directed acyclic graphs (DAGs) represent normal and anomaly workflows

Output: the normal/anomaly labels for workflow-level (entire graph) and job-level (single node)

- Our GNN models achieved 25% improvement accuracy over conventional methods for anomaly detection.
- We achieve 2-4 times faster training time when compared with conventional machine learning models.
- Developing explainable AI methods to explain anomalies in the workflow performance.

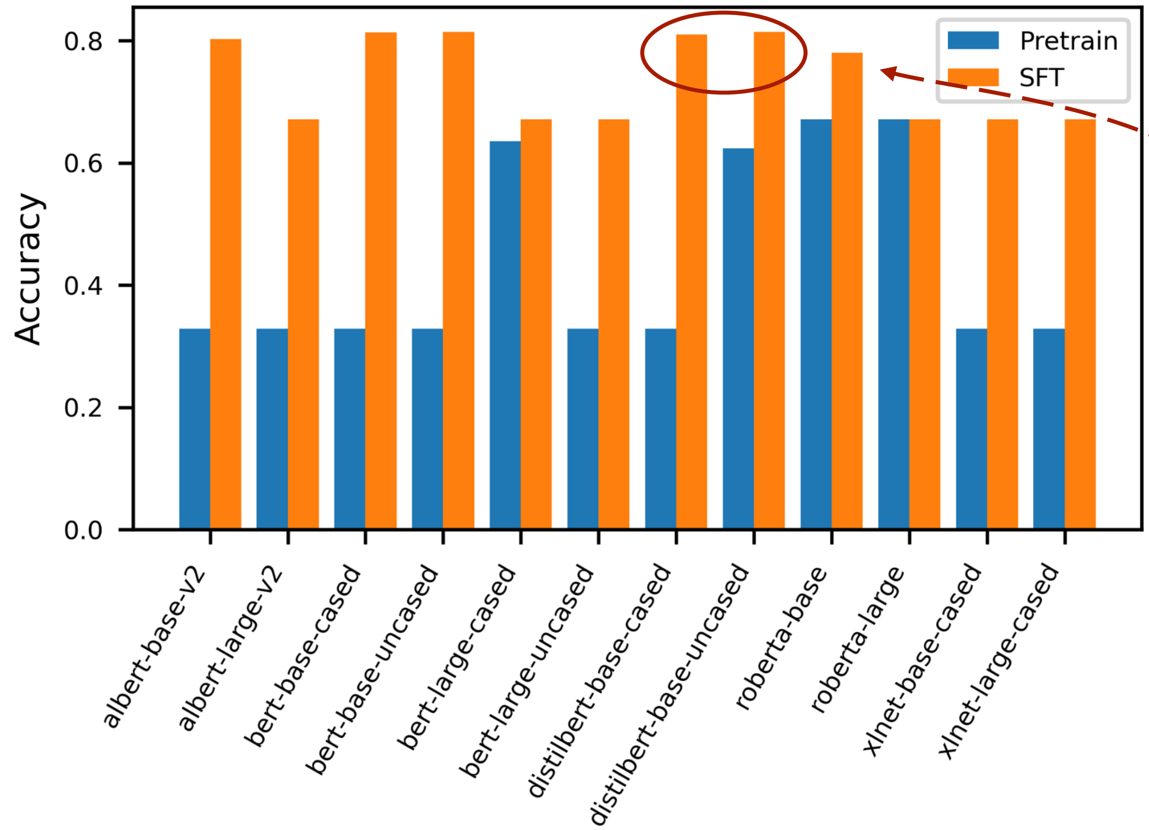
USC Viterbi School of Engineering · renci · Argonne National Laboratory · Berkeley Lab · Oak Ridge National Laboratory

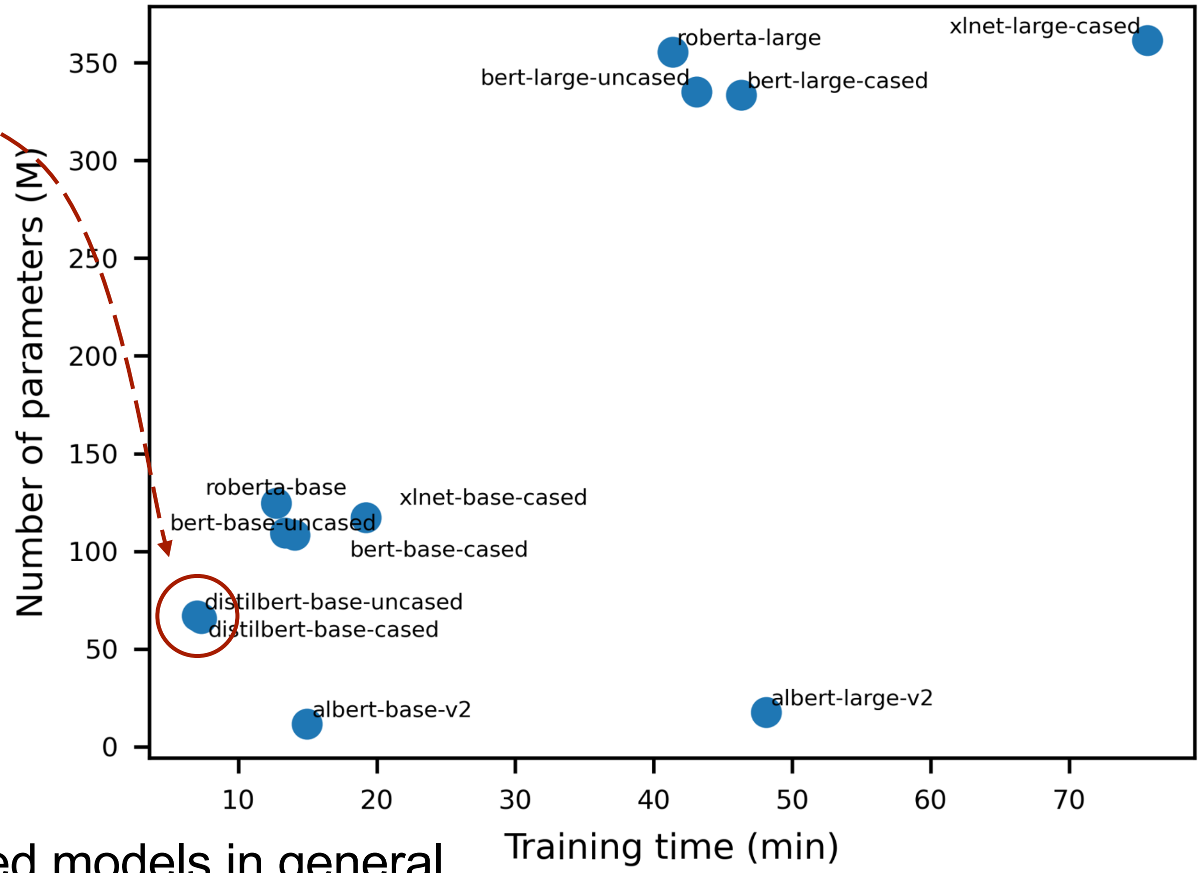
http://poseidon-workflows.org

This work is funded by DOE under grant #DE-SC0022328

Performance of Different LLM Models

PoSeiDon





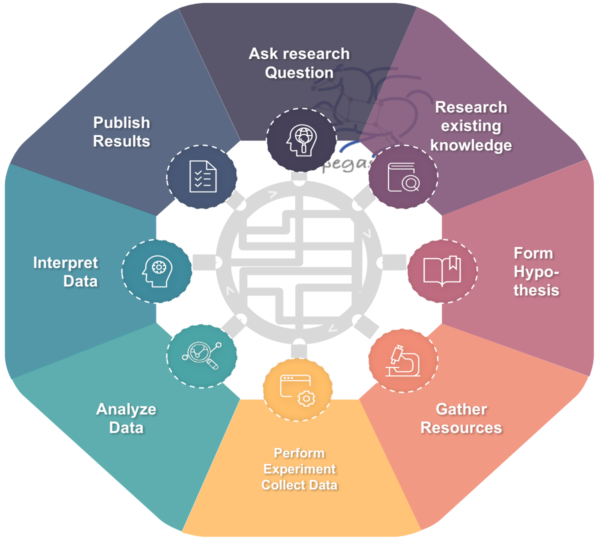
- SFT models are better than pretrained models in general
- Larger model does not necessarily mean better performance

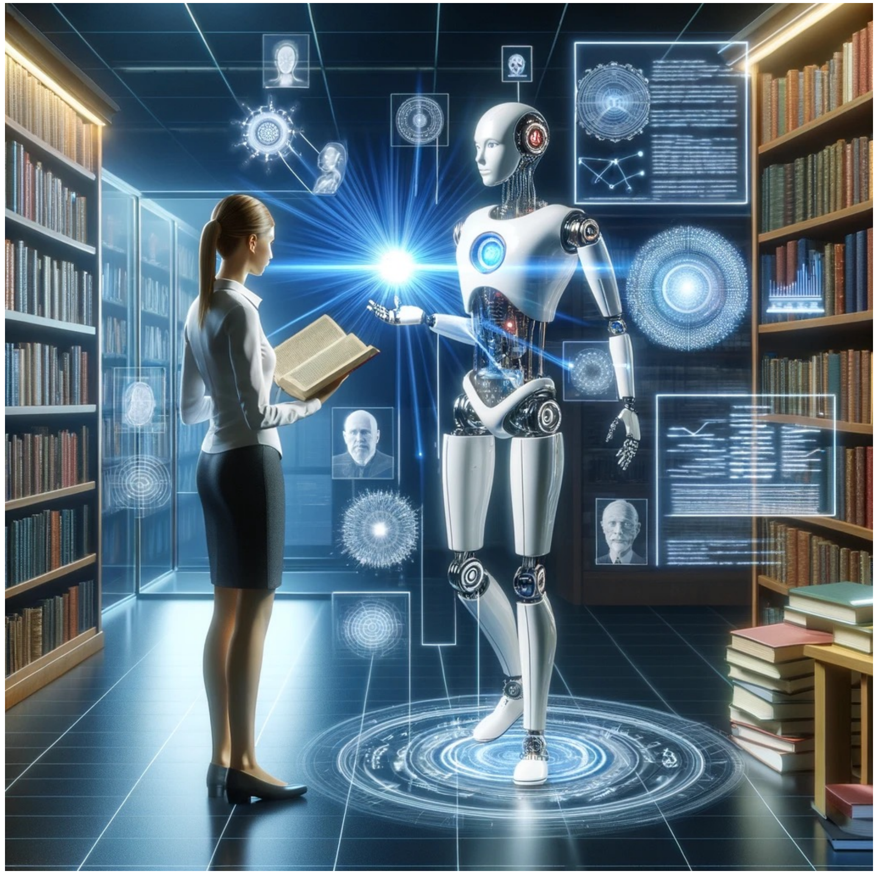
USC Viterbi School of Engineering · renci · Argonne National Laboratory · Berkeley Lab · Oak Ridge National Laboratory

http://poseidon-workflows.org

This work is funded by DOE under grant #DE-SC0022328

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





ChatGPT-4 *create an image of AI agent learning from publications and delivering information to a researcher*

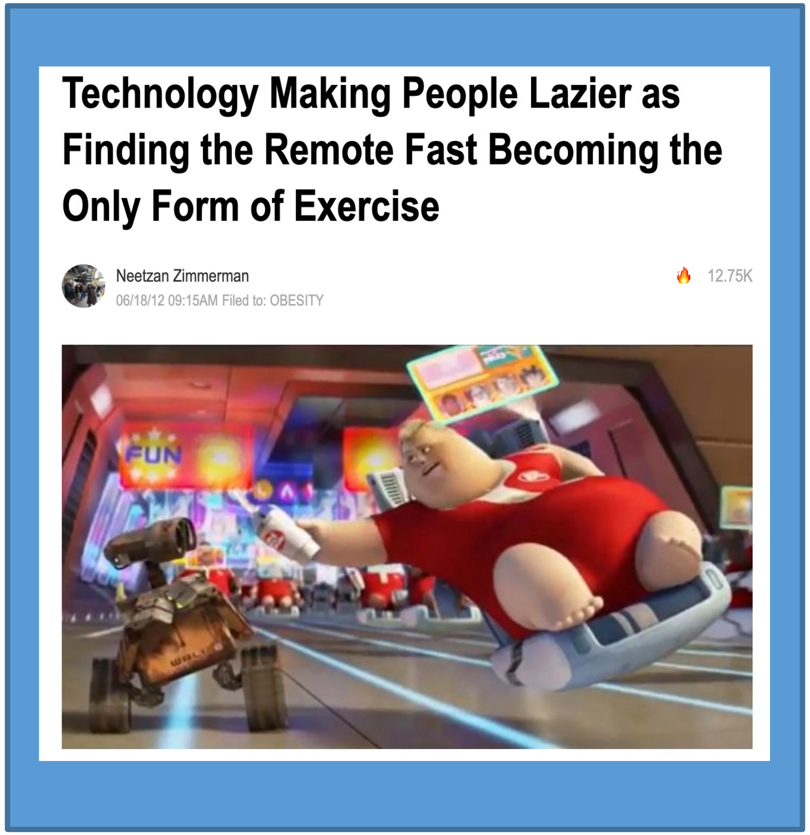
How will the scientific lifecycle look like in 10-20 years?

How will we teach science? What skills?
How will we share knowledge?
How will the work of scientists look like?

Will we still need scientists ?

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

- Seeing both sides of an issue
- Being open to new evidence that disconfirms ideas
- Reasoning dispassionately
- Demanding that claims be backed up by evidence
- Deducing and inferring conclusions from available facts – what are the facts?
- Solving problems



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

Neetzan Zimmerman 12.75K

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

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

Studies show that you cannot teach critical thinking in the abstract
The process of thinking is intertwined with the content of thought (domain knowledge)

USC Viterbi School of Engineering Information Sciences Institute

and curious

"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