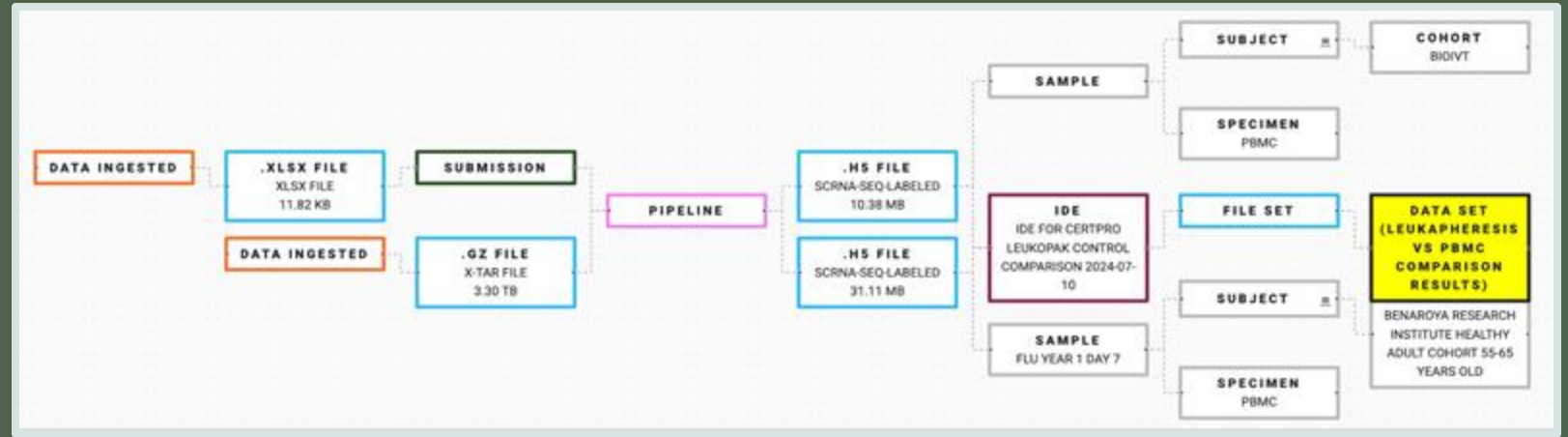




ALLEN INSTITUTE for  
IMMUNOLOGY



# From Reproducible Research to Open Science Dissemination

A Computing Platform-Centric Approach

Paul Meijer

Senior Director, Scientific Software Engineering



# Large Scale Research Presents Transparency and Reproducibility Challenges

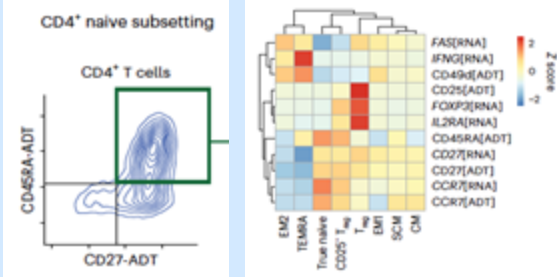
---

1. Store, categorize and find **large amounts of data**
2. Prepare data for transformation and **analysis**
3. Keep track of **complex multi-step** analyses
4. Support interdisciplinary **team collaboration**
5. Enable **transparent** research for **ongoing review**
6. **Share** the data, analysis, and results to the open science community
7. Ensure **analysis reproducibility** of shared results
8. Keep research **available, affordable, and sustainable**

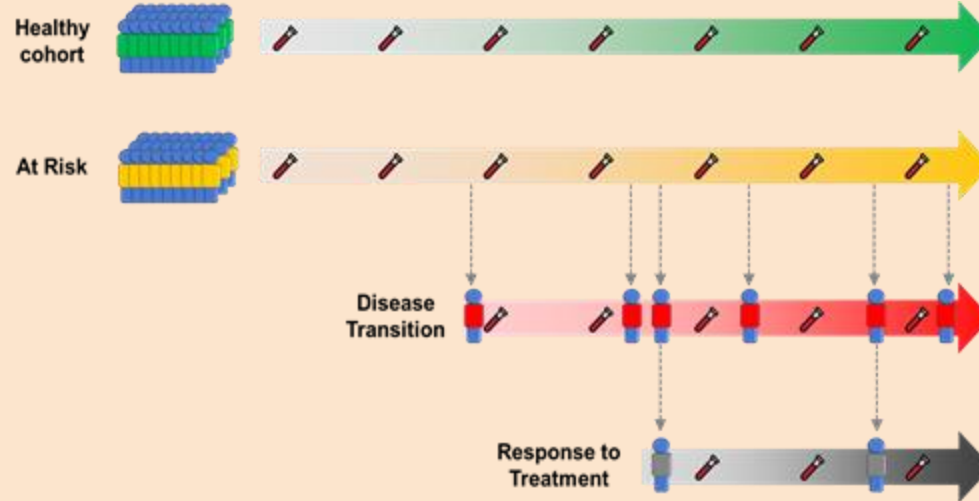


# Use Case: Large-Scale Human Immunology Research...

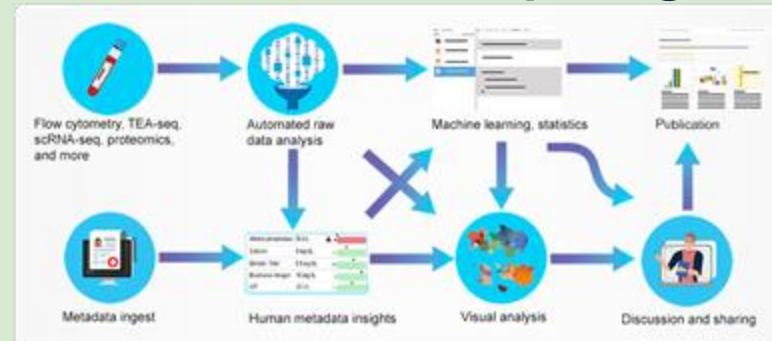
## Wet-bench (validation, follow-up, etc.)



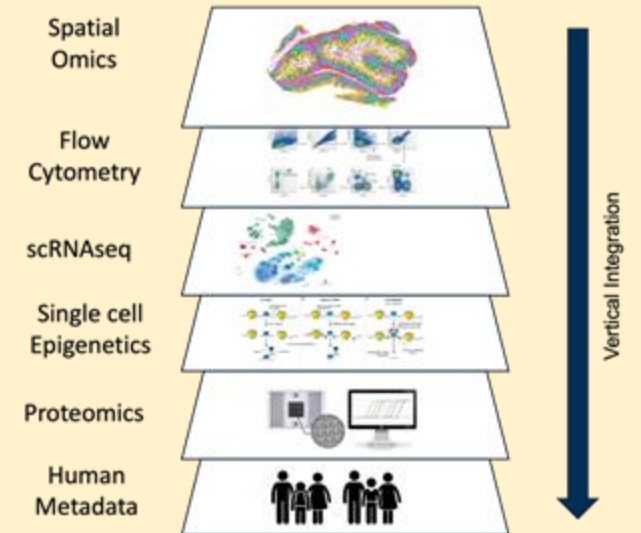
## Longitudinal studies



## Scientific Computing



## Multi-omic data

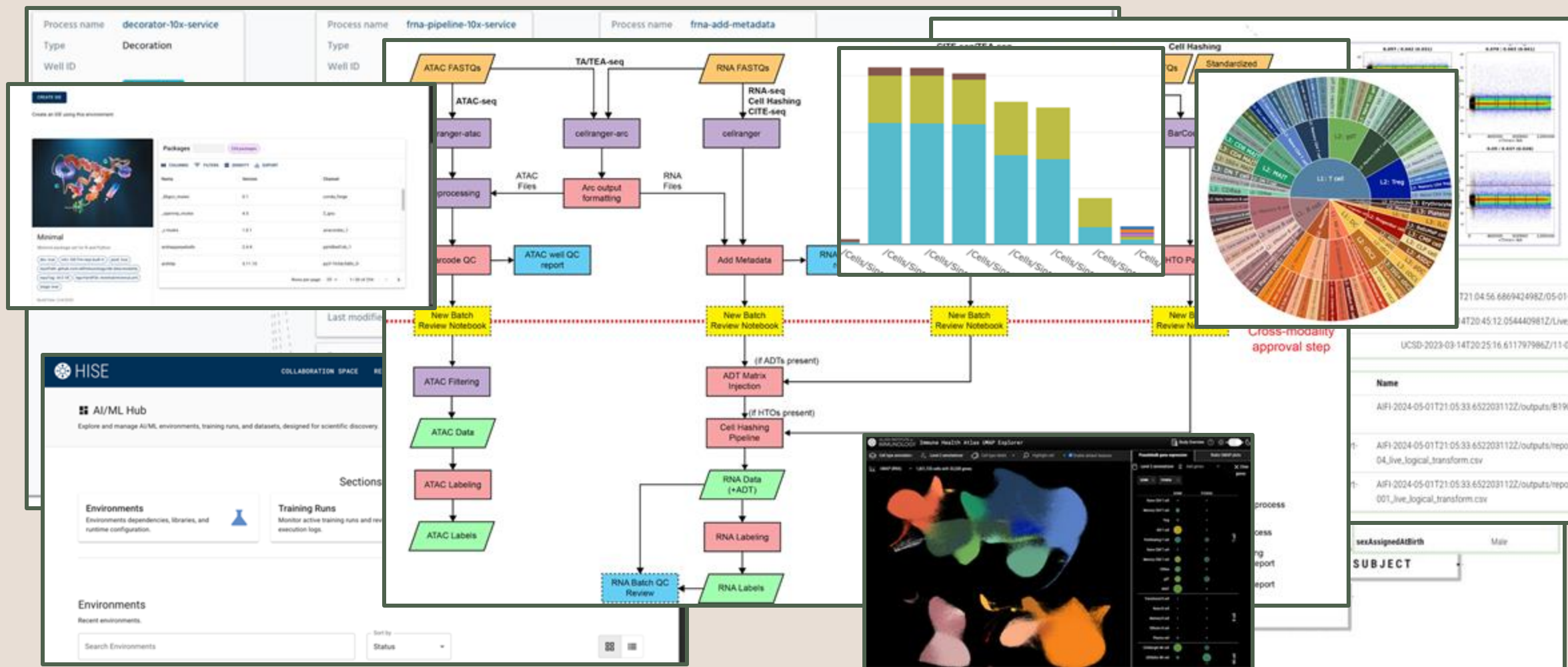






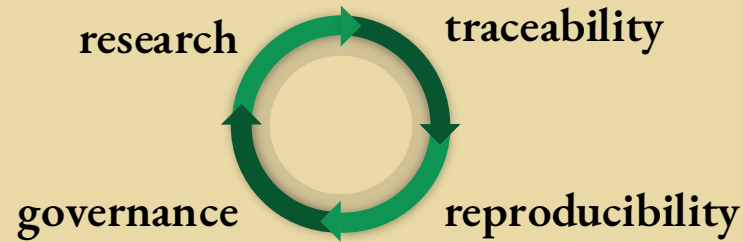
## Research

## Large-Scale Human Immunology



# Our Solution: A Comprehensive Platform Built for Reproducibility and Openness

---



**complex multi-step analyses** → capture research as it unfolds

**transparent ongoing review** → enable exact re-execution of steps

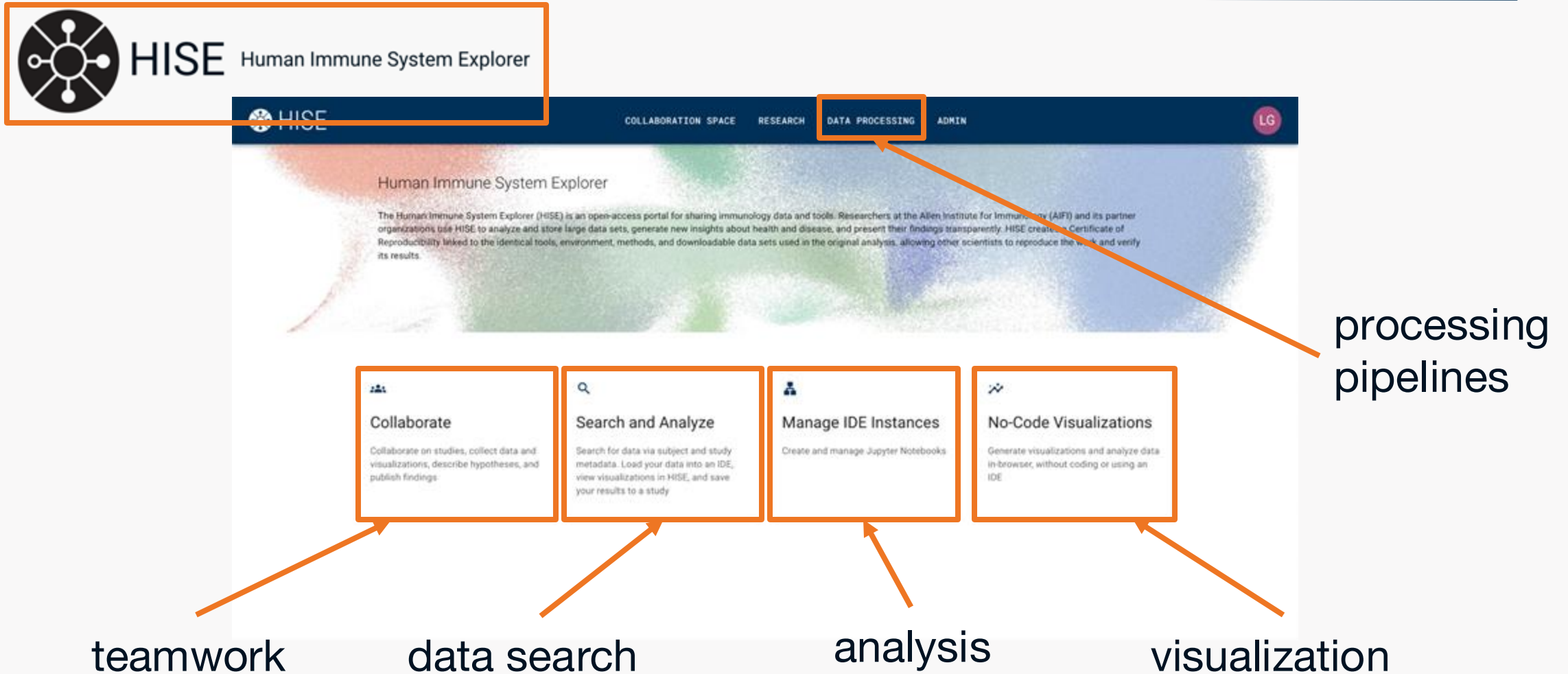
**team collaboration** → trace all tools and transformations

**share data, analysis and results** → directly from the computing platform

**analysis reproducibility** → share the research trace of the results

**available, affordable, sustainable** → ongoing usage guides governance

# A Computing Platform for Large, Complex Research...



## ...and Open Science Dissemination and Interaction



**HISE** Human Immune System Explorer

[explore.allenimmunology.org](https://explore.allenimmunology.org)

## DEEPENING OUR UNDERSTANDING OF THE IMMUNE SYSTEM

High Dimensional Analyses and Computational Insights by Scientists at the Allen Institute for Immunology

DATA APPS    REPRODUCIBILITY    PUBLICATIONS    VISUALIZATIONS    DATA SETS

## Longitudinal Dynamics of Health and Age

File Set Visualization

Claire E Gustafson, Peter J Skene, Ananda W Goldrath, Xiao-jun Li, Troy R Torgerson, Lynne A Becker, Thomas F Burnol, Aishwarya Chander, Ernest M Coffey,

## AIFI Immune Health Atlas

File Set Visualization

Claire E Gustafson, Peter J Skene, Ananda W Goldrath, Xiao-jun Li, Troy R Torgerson, Lynne A Becker, Thomas F Burnol, Aishwarya Chander, Ernest M Coffey, Elizabeth M Domjogh, Jessica Garber

### Systemic Inflammation and Progression in At-Risk Rheumatoid Arthritis

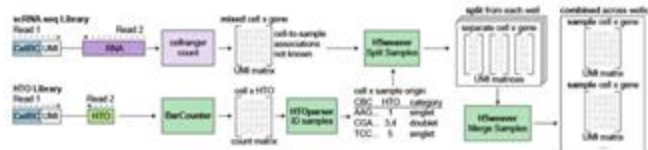
File Set Visualization

Ziyuan He, Maria C Glass, Praveena Venkatesan, Marie L Feser, M Kristen Demoruelle, Chelsie L Fleischer, Jessica Garber, Palak C Genge, Qiuyu Gong, Claire E Gustafson, Brian C Hattel, Veronica



## Data Processing Pipeline

After generating sequencing data, we sequenced both the gene Expression and HTO libraries on Illumina sequencing platforms. Multiplexed sequencing data were prepared using a preprocessing pipeline based on 10x Genomics Cell Ranger software (version 3.1.0, Released July 24, 2019), with the 10x Genomics GRCh38.p9 Human Transcriptionome Reference (vrefdata-cellranger-GRCh38-3.0.0, based on the GRCh38 genome reference and Ensembl v98 transcripts annotations, Released Nov. 19, 2018) to generate cell x UMI count matrices for each sample. To identify cells originating from each sample, we utilized software tools for HTO quantification (BioCrunch) and single demultiplexing (HTOPicker, H5Sequencer) to generate separate HDFS-formatted files for each original sample. This process also removes Cell Barcodes that contain HTOs from multiple samples (doublets or multiplets) or with no detectable sample label (no hash).



## Downloads

Human Immune Health Atlas      We offer downloads for multiple types of data related to the Human Immune Health Atlas. See the sub-sections described below to access these datasets.

### Clinical Labs and Metadata

Sample metadata and clinical lab results, including complete blood count, metabolic panels, and lipid panels.

Get LaTeX and MathJax

## scRNA-seq Data and Controls

Our single-cell RNA-seq datasets, along with batch control data and QC reports.

[Get all the news on this](#)

### scRNA-seq Labeling Models

CellType models used in our ["Label Your Own Data" website](#) for both 10x Genomics Universal 3' Gene Expression and 10x Genomics Flex Gene Expression data. We also provide the colorset that we have used for cell type visualization.

Get Labeling Solutions



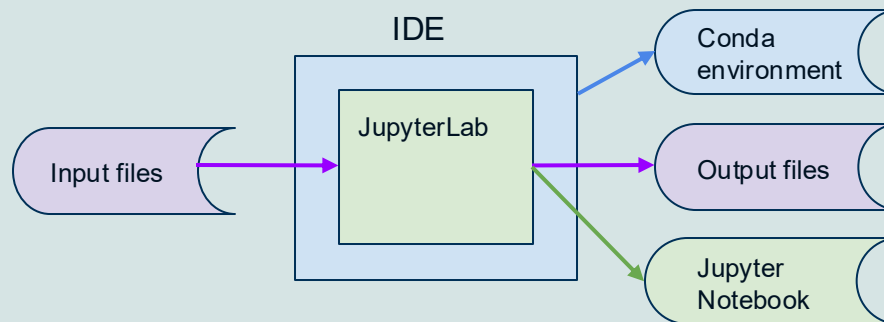


# How it Works: Trace-Driven Architecture Tracks all Data and Transformations

# An analysis platform designed for traceability



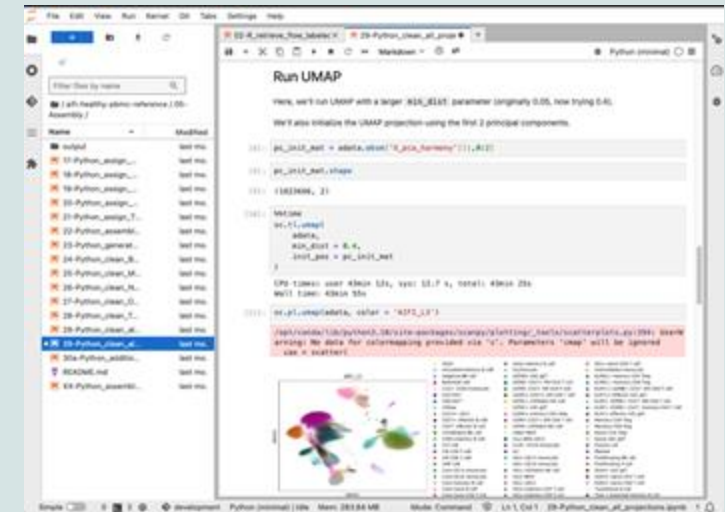
## Register data, code, and analysis environment details



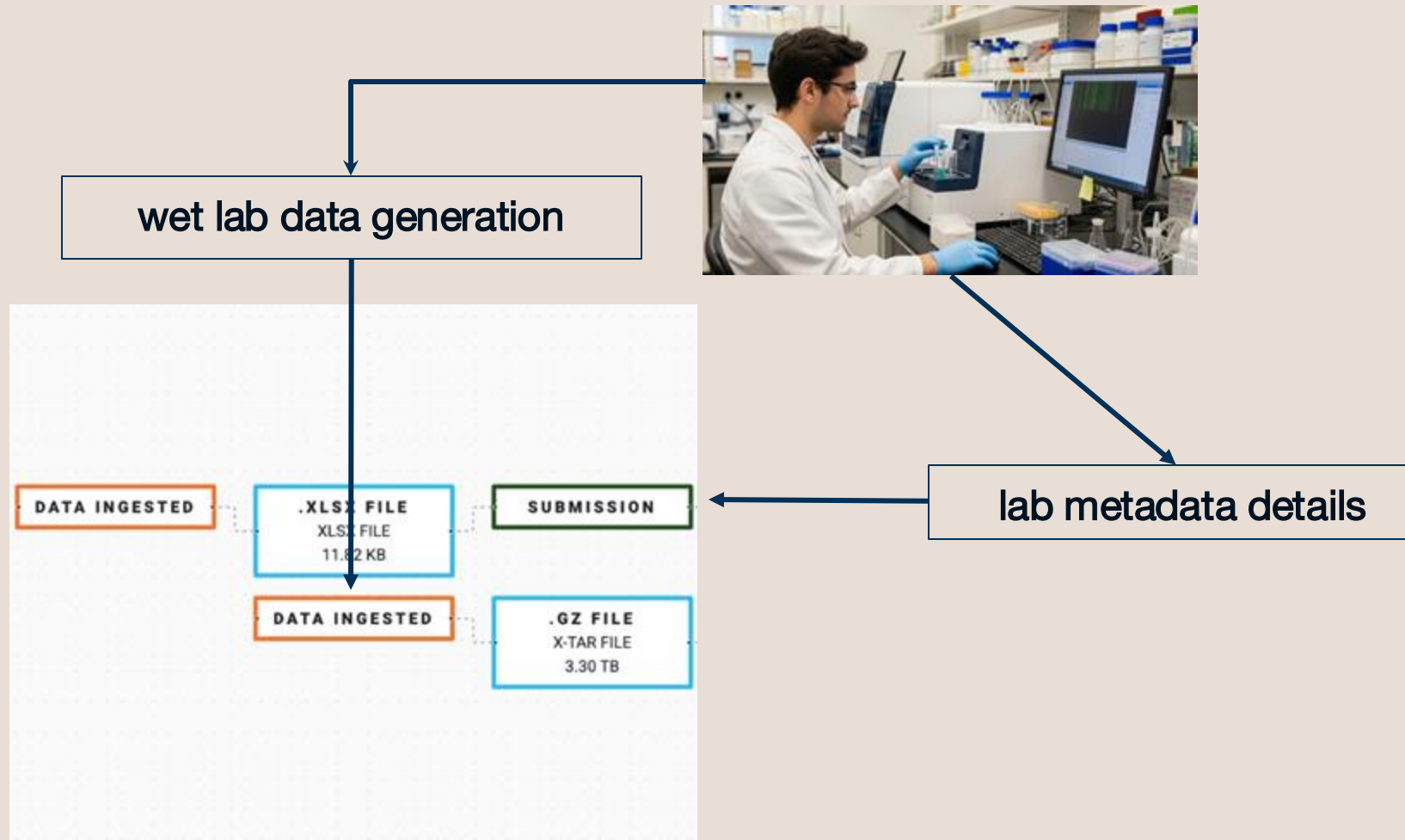
## Incrementally build a graph showing each step



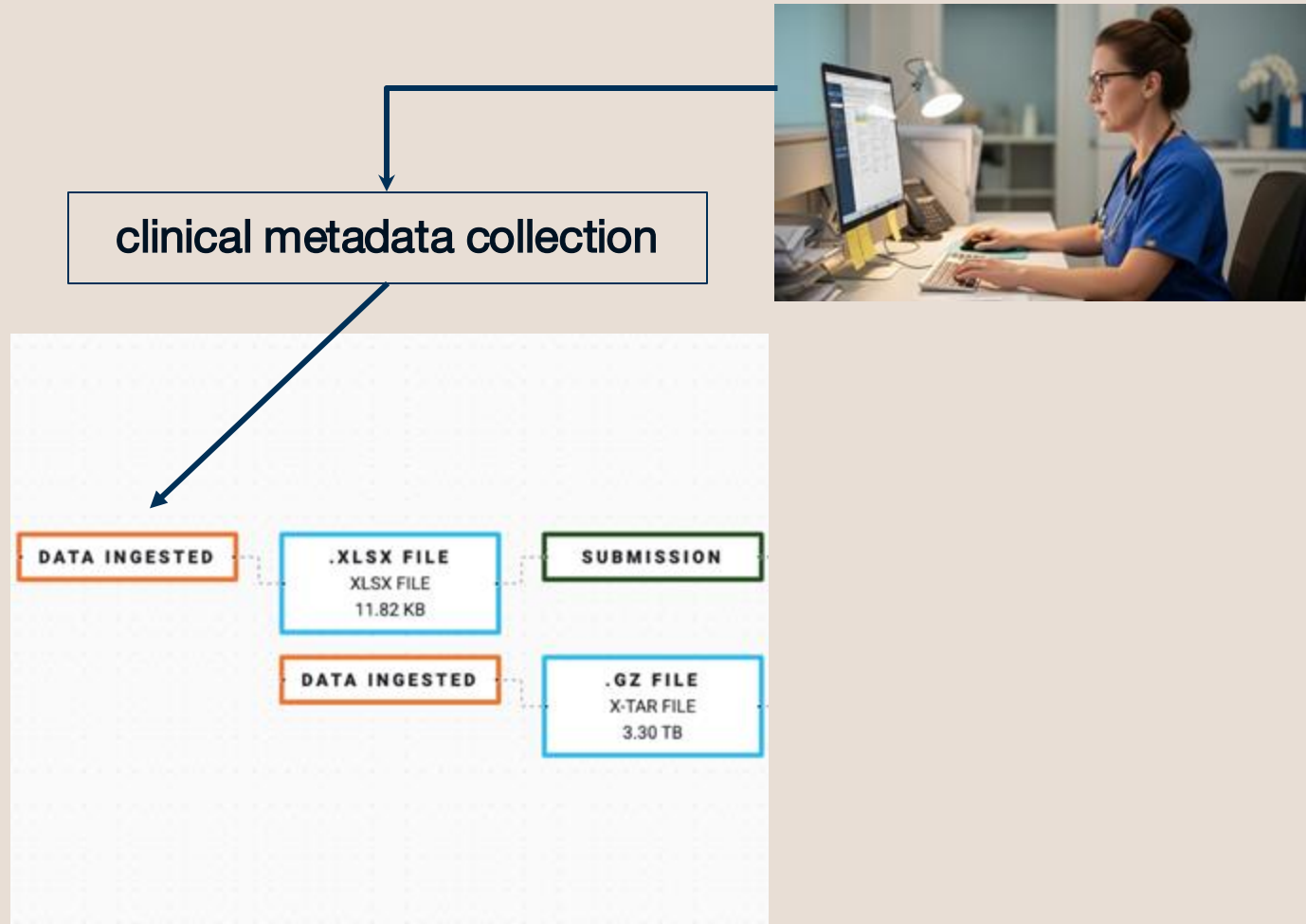
Publish data, code, and tools for interactive inspection and reproducibility



# Proactively Capture the Research Trace



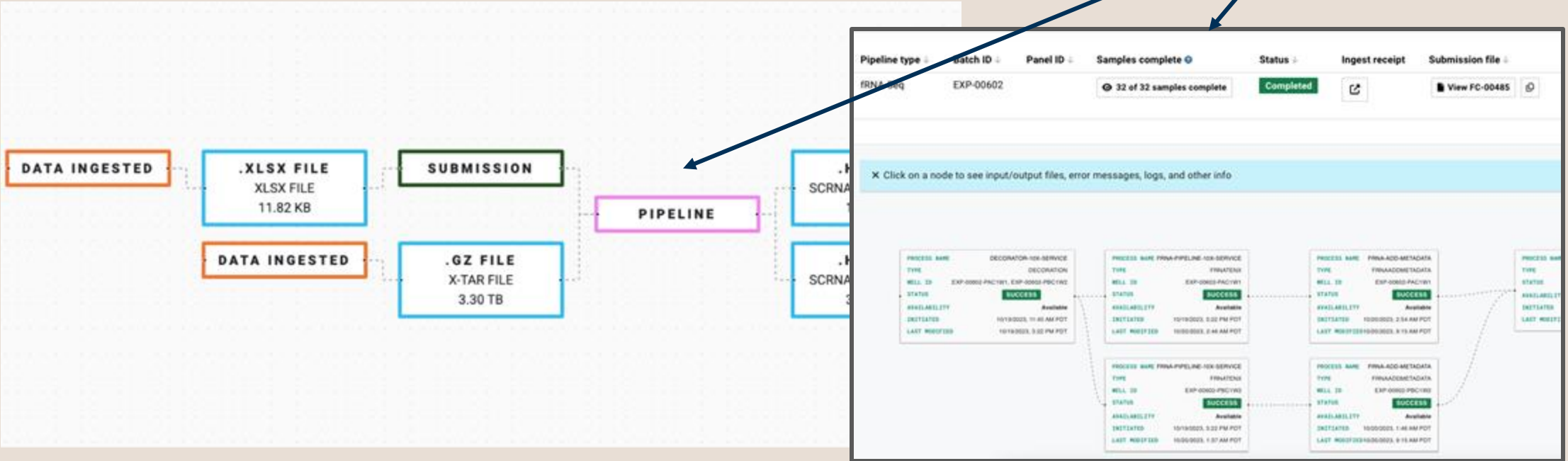
# Proactively Capture the Research Trace



# Proactively Capture the Research Trace



automated analysis and verification



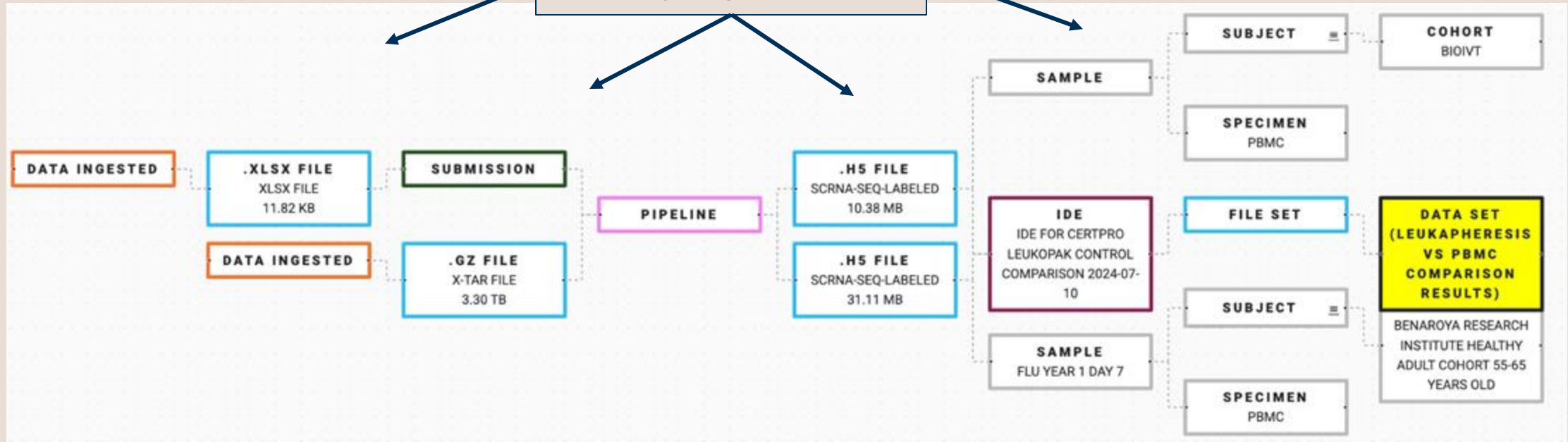




# Proactively Capture the Research Trace



ongoing review



# Published Research Trace: Certificate of Reproducibility

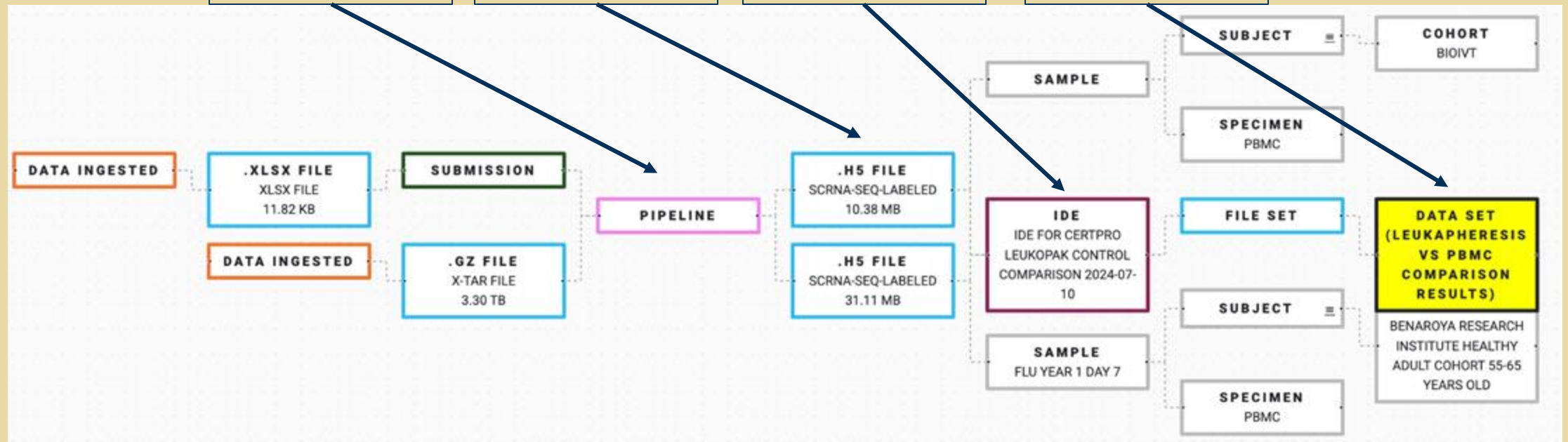


rerun pipelines

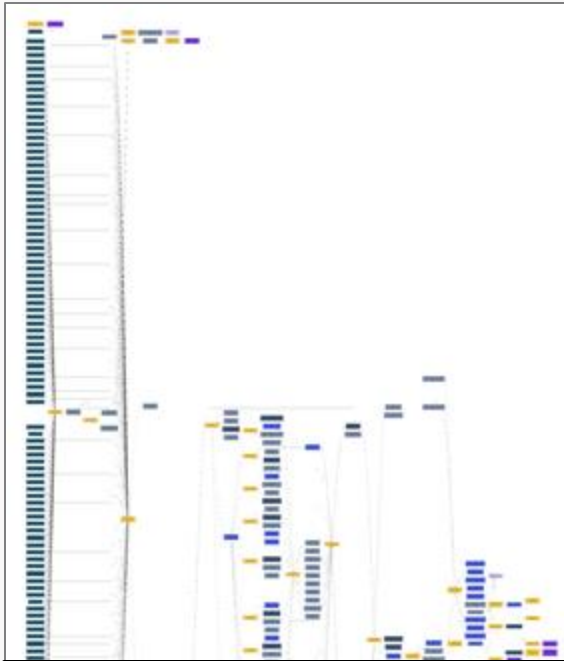
download data

verify analysis

interpret results



# Certificates of Reproducibility: The Reality



Certificate trace from the  
Human Immune Health Atlas

92 automated pipeline runs  
47 analysis steps  
182 output files  
results for 108 wet lab samples

More in our paper in Royal Academy Open Science:

Meijer P, Howard N, Liang J, Kelsey A, Subramanian S, Johnson E, *et al.*

Provide proactive reproducible analysis transparency with every publication.

R Soc Open Sci. 2025;12: 241936. doi:10.1098/rsos.241936

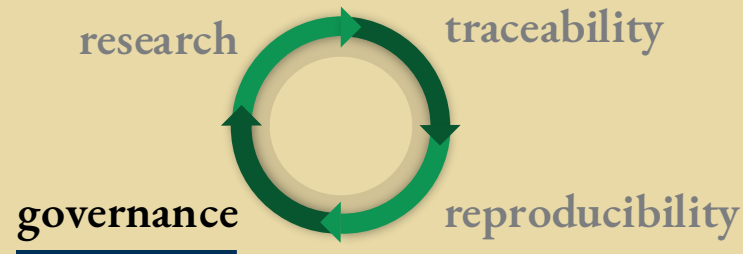
<https://tinyurl.com/repro-article>





# Our Solution: A Comprehensive Platform Built for Reproducibility and Openness

---



- ✓ complex multi-step analyses → capture research as it unfolds
- ✓ transparent ongoing review → enable exact re-execution of steps
- ✓ team collaboration → trace all tools and transformations
- ✓ share data, analysis and results → directly from the computing platform
- ✓ analysis reproducibility → share the research trace of the results

**available, affordable, sustainable → ongoing usage guides governance**

# Available, Affordable, and Sustainable Research

## What Does That Mean?

---

### Data

#### Available

Keep data that is used for new studies or by the open science community

#### Affordable

Archive all other data - abandoned analysis paths, data that has lost interest

#### Sustainable

Evaluate data retention against regeneration via modern methods

### Tools

#### Available

Containerize tools and algorithms so they continue to run

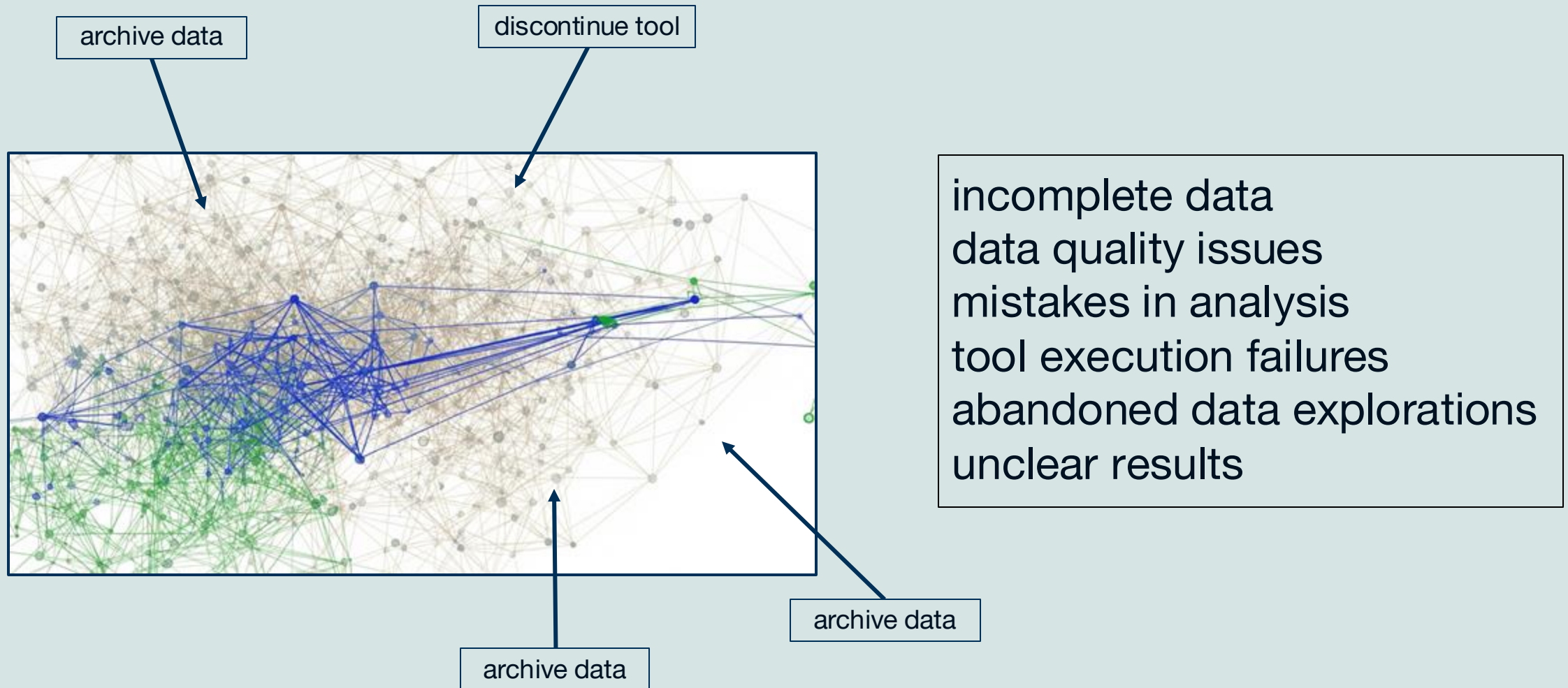
#### Affordable

Discontinue tools along abandoned research traces

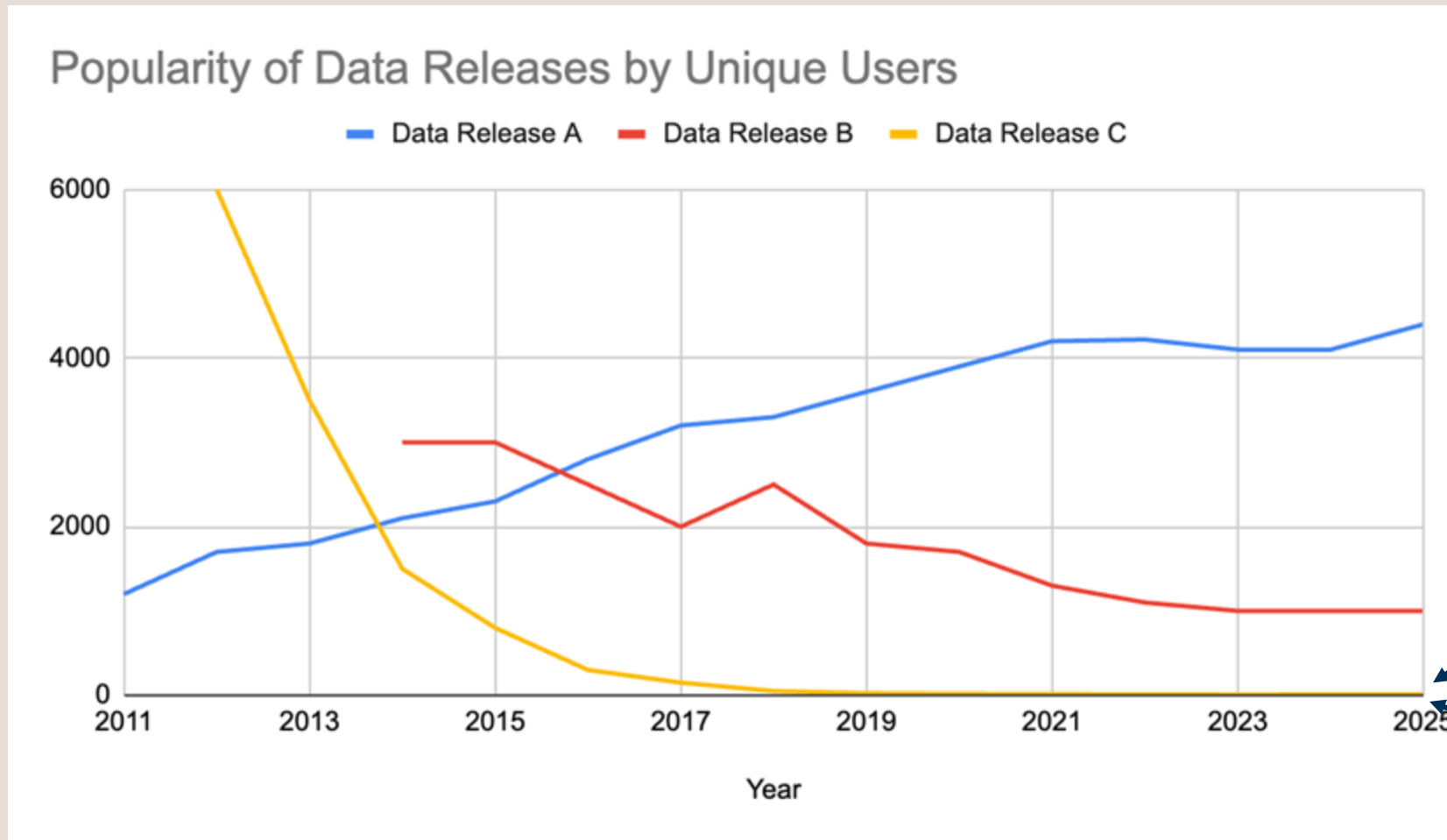
#### Sustainable

Manage and upgrade containerized tools and algorithms with continued usage

# Not All Research Traces Are Published



# Not All Data Releases Remain Relevant



consider data  
archival savings

evaluate cost of  
tool maintenance



# Research Traces Inform Sophisticated Data and Tool Governance Policies

---

Data age and size are arbitrary metrics for governance

Research traces reveal the true relevance of data and tools

## collection

- data ingest
- data transformation
- tool generation

## storage

- data persistence
- tool containerization

## utility

- data usage
- data archival
- data deletion
- tool execution
- tool upgrade
- tool removal

# A Holistic View Of Sustainable Research

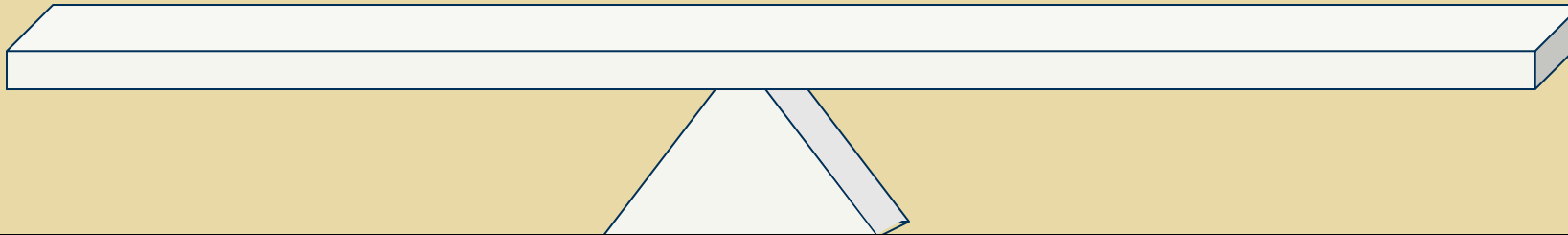
---

## Enable Original Research

New hypotheses  
Innovative data generation techniques  
Novel algorithmic approaches

## Support Open Science

Reproducibility and verification  
Availability of reference data sets  
Tool democratization



## More in our paper in the Harvard Data Science Review:

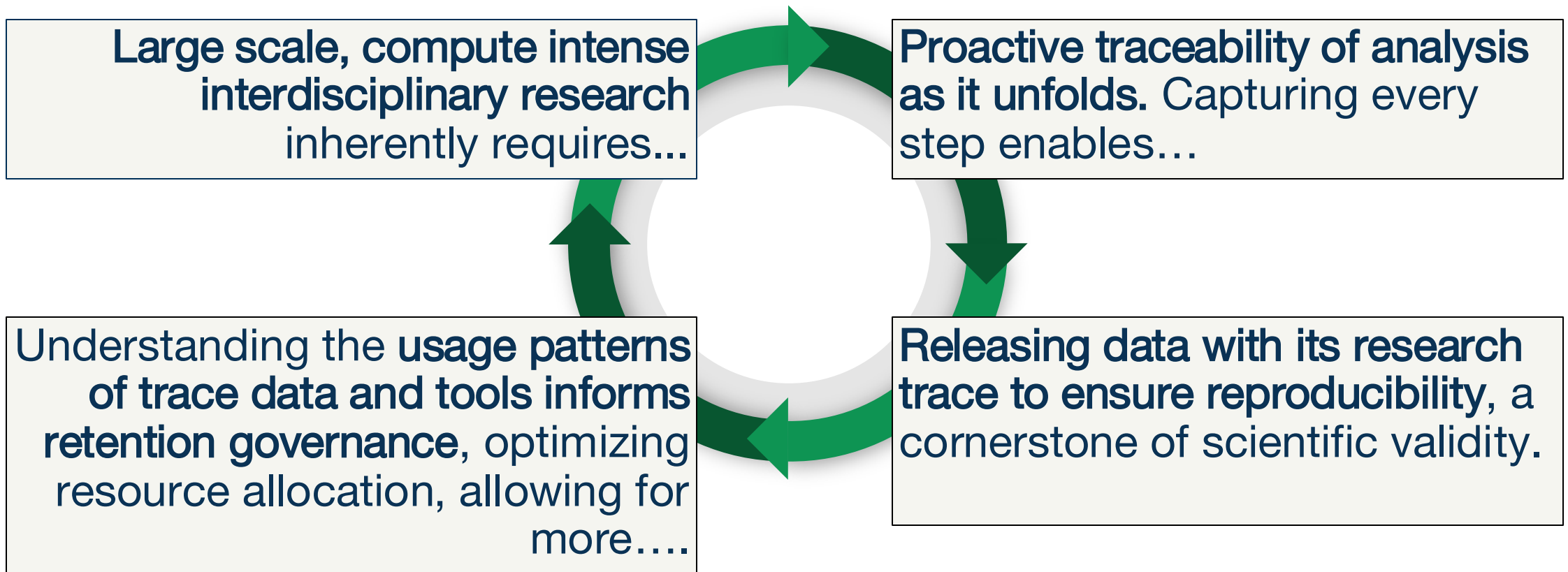
Meijer, P. et al. (2025). Research Lifecycle Management: Using Analysis  
Reproducibility Research Software to Define Contextual Data Governance Policies.  
Harvard Data Science Review, 7(3). doi:10.1162/99608f92.08da1513.

<https://tinyurl.com/reproGov>



# In Summary: Transparent Interdisciplinary Analysis Drives Reproducible and Sustainable Research

---







# THANK YOU

We wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.



ALLEN INSTITUTE *for*  
IMMUNOLOGY