Provenance

BNL: Kerstin Kleese van Dam
PNNL: Eric Stephan, Todd Elsethagen, Bibi Raju

BROOKHAVEN NATIONAL LABORATORY
a passion for discovery

U.S. DEPARTMENT OF ENERGY | Office of Science

Provenance Definition

- General: Is the chronology of ownership, custody or location of the object. Originally mostly applied to works of art.

- **Computer Science: Provenance is a record that describes the people, institutions, entities, and activities, involved in producing, influencing, or delivering a piece of data or a thing.**
- **https://dvcs.w3.org/hg/prov/raw-file/tip/presentations/wg-overview/overview/index.html**

Popular Provenance Vocabularies

Dublin Core Provenance Task Force

Open Provenance Model

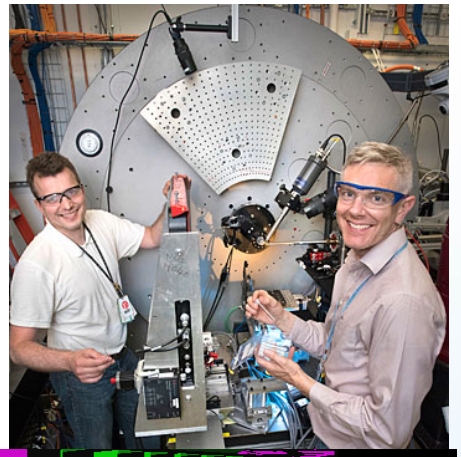
Proof Markup Language Ontology

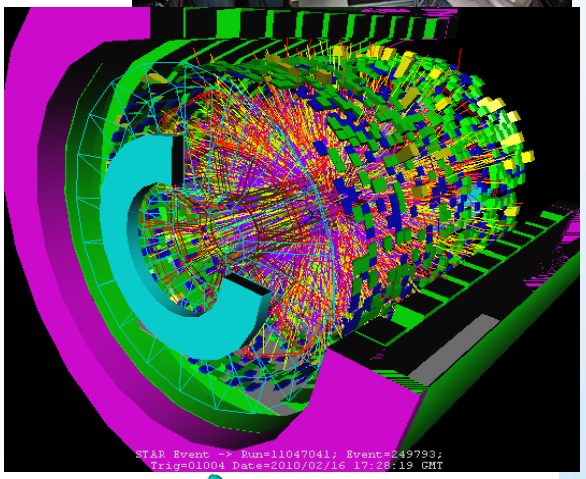
The Provenance Ontology (Prov-O)

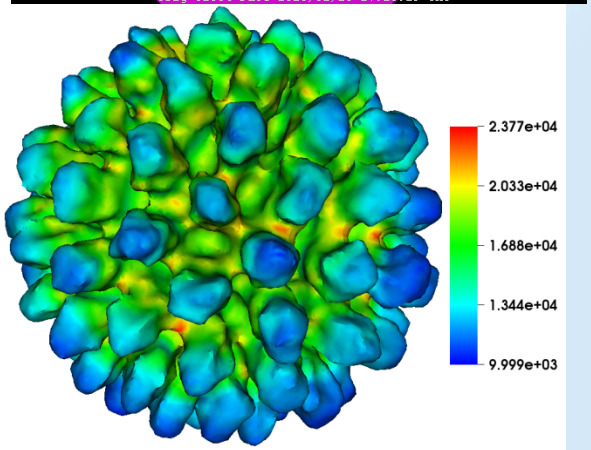
Scientific Research Today

Collaborative and complex, typically has the following characteristics:

- **Involves multiple stakeholders**
- **Leverages multiple tools, algorithms, data products, and sensors**
- **Reliant on highly iterative and potentially repetitive techniques**
- **Steps are difficult to document and are often times committed to memory or notes.**





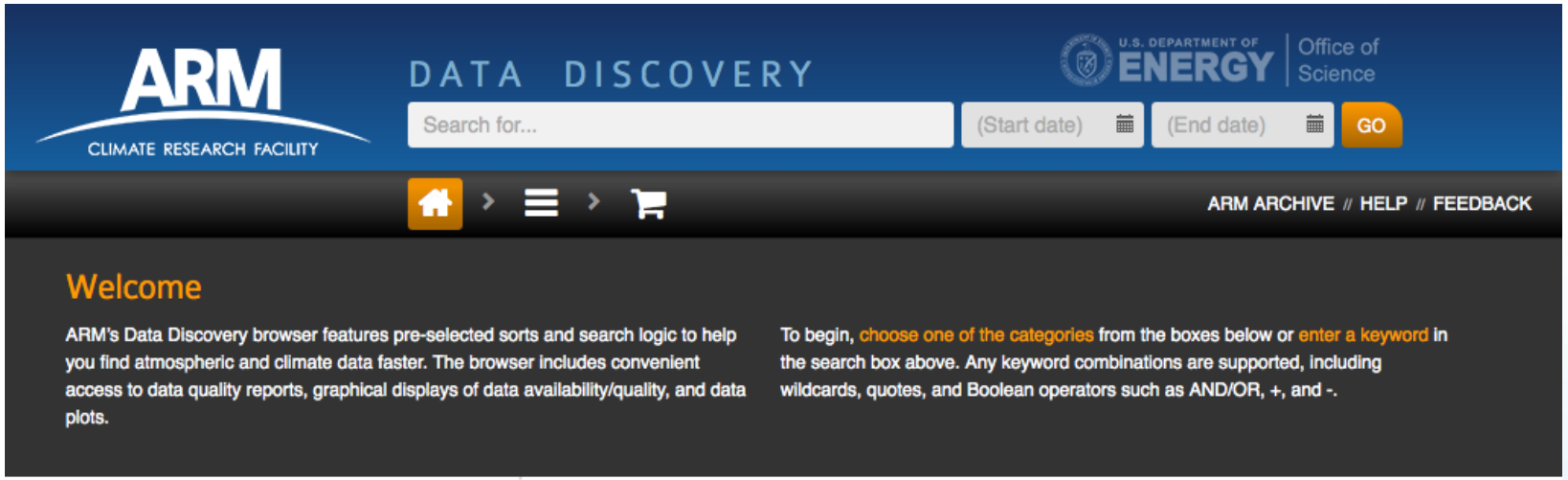


What do we use Provenance for in this context?

- **Result explanation -** How was this data set created, what other data sets is it based on, what tools and algorithms were used to create the data set.
- **Reproducibility -** Provides sufficient information to enable others to reproduce a specific scientific results, utilizing the same tools, algorithms and source data.
- **Performance Analysis** - Explains where a workflow was run, which resources were used and how the the workflow was executed. As empirical study it can show where workflows might be underperforming and why.

RESULT EXPLANATION

Climate Science Example



Data Highlights »

LASSO Alpha 1 Release

The Alpha 1 release is a first look at the large-eddy simulation capability under development for the ARM Climate Research Facility. The release contains 192 simulations spread over five shallow convection cases in 2015, and the intent of the release is to garner feedback from the community.

Modeling Best Estimates

The ARM Best Estimate data products are ARM datastreams specifically tailored to climate modelers for use in the evaluation of global climate models. They contain a best estimate of several cloud, radiation, and atmospheric quantities.

Search by Category »

Aerosols

The effect of aerosols is measured by instrument systems and lidars that provide data on the size distribution, optical properties, scattering, and extinction of aerosols.

Atmospheric State

Surface-based and airborne instruments measure the thermal, moisture, and kinetic properties of the atmospheric (horizontal and vertical distributions) and the atmospheric concentrations of certain radiatively active trace gases (e.g, CO_2 and O_3).

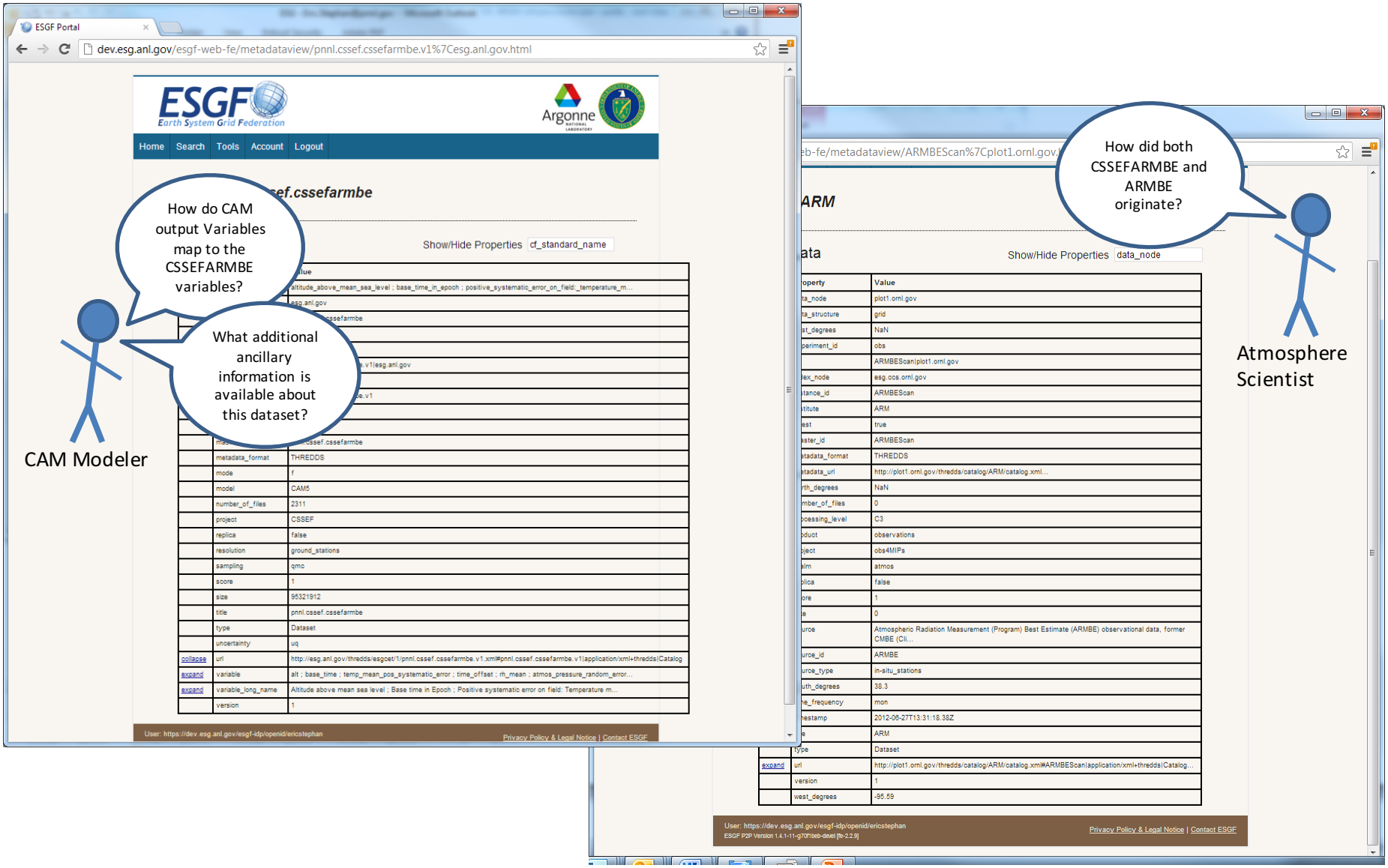
Cloud Properties

Active and passive remote sensing instruments are used to measure the macroscopic properties (horizontal and vertical distributions) of clouds, and the microphysical properties (sizes, shapes, and phases [water or ice]) of the particles that comprise the clouds.

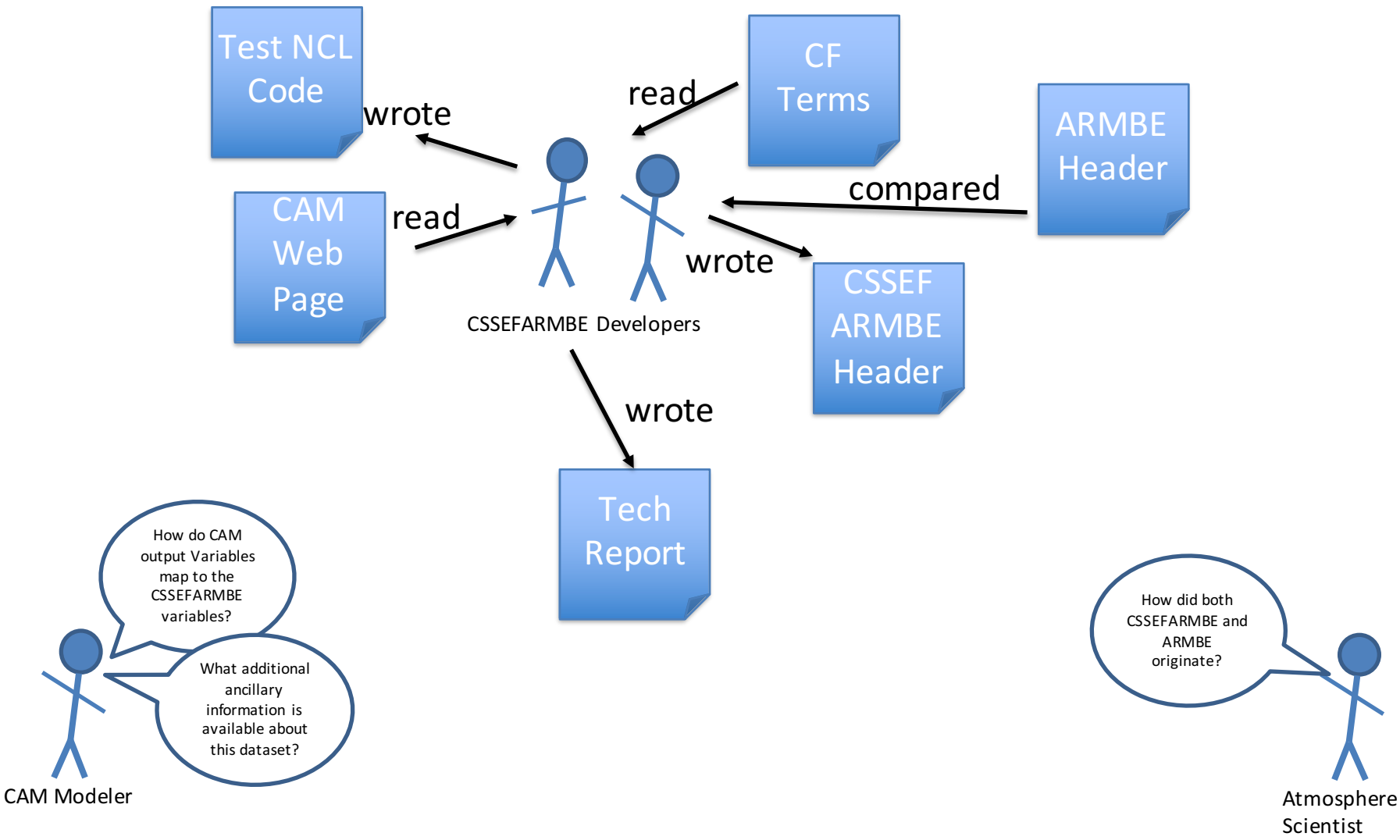
Radiometric

Radiometric measurements provide data on the propagation of electromagnetic energy through the atmosphere. These types of measurements represent the majority of ARM data, and are obtained using various types of active (such as radar and lidar) and passive (such as

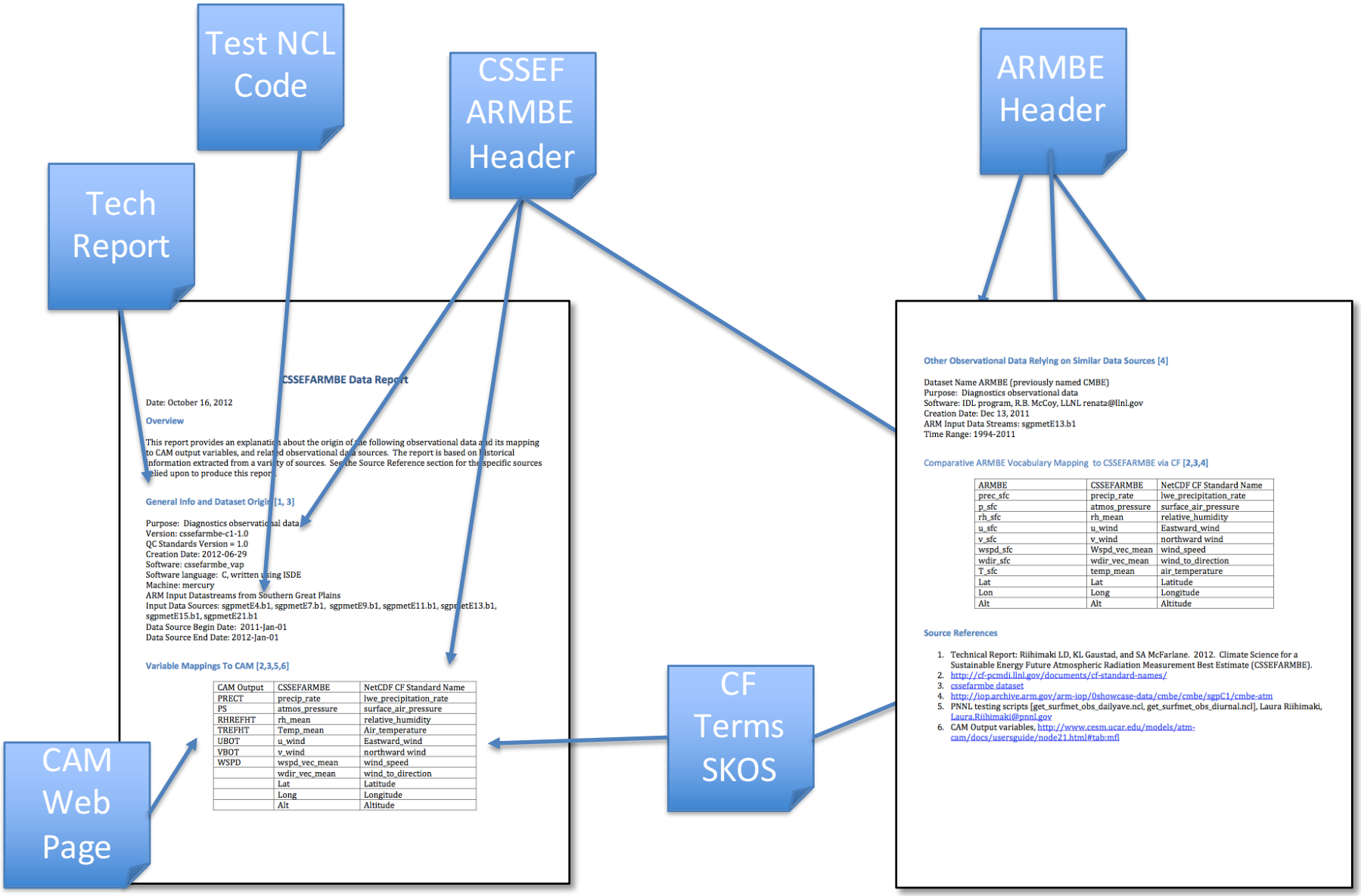
Example: Questions about Climate Diagnostics Dataset



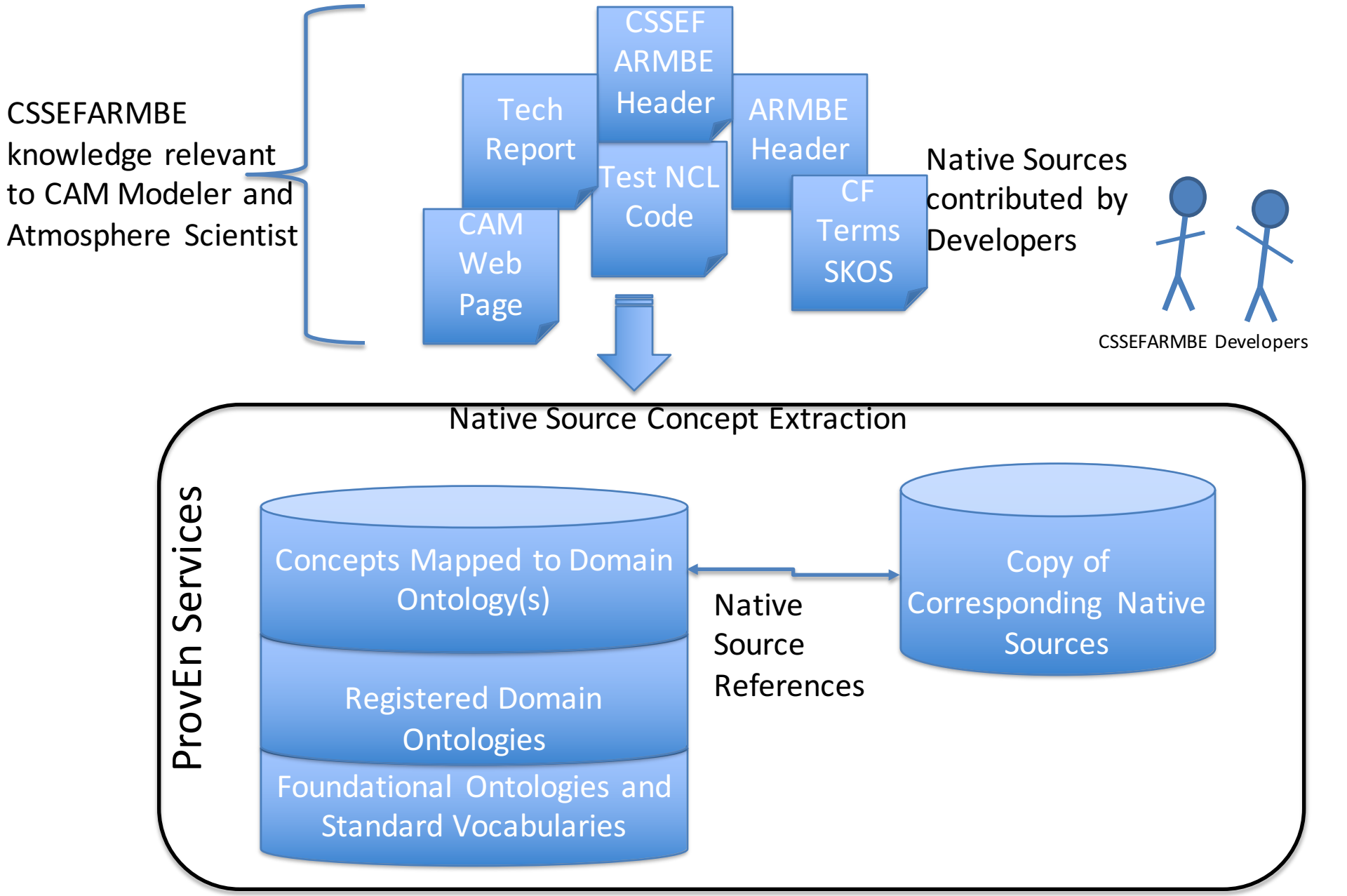
The Knowledge Gap: Needing Answers from Data Producers



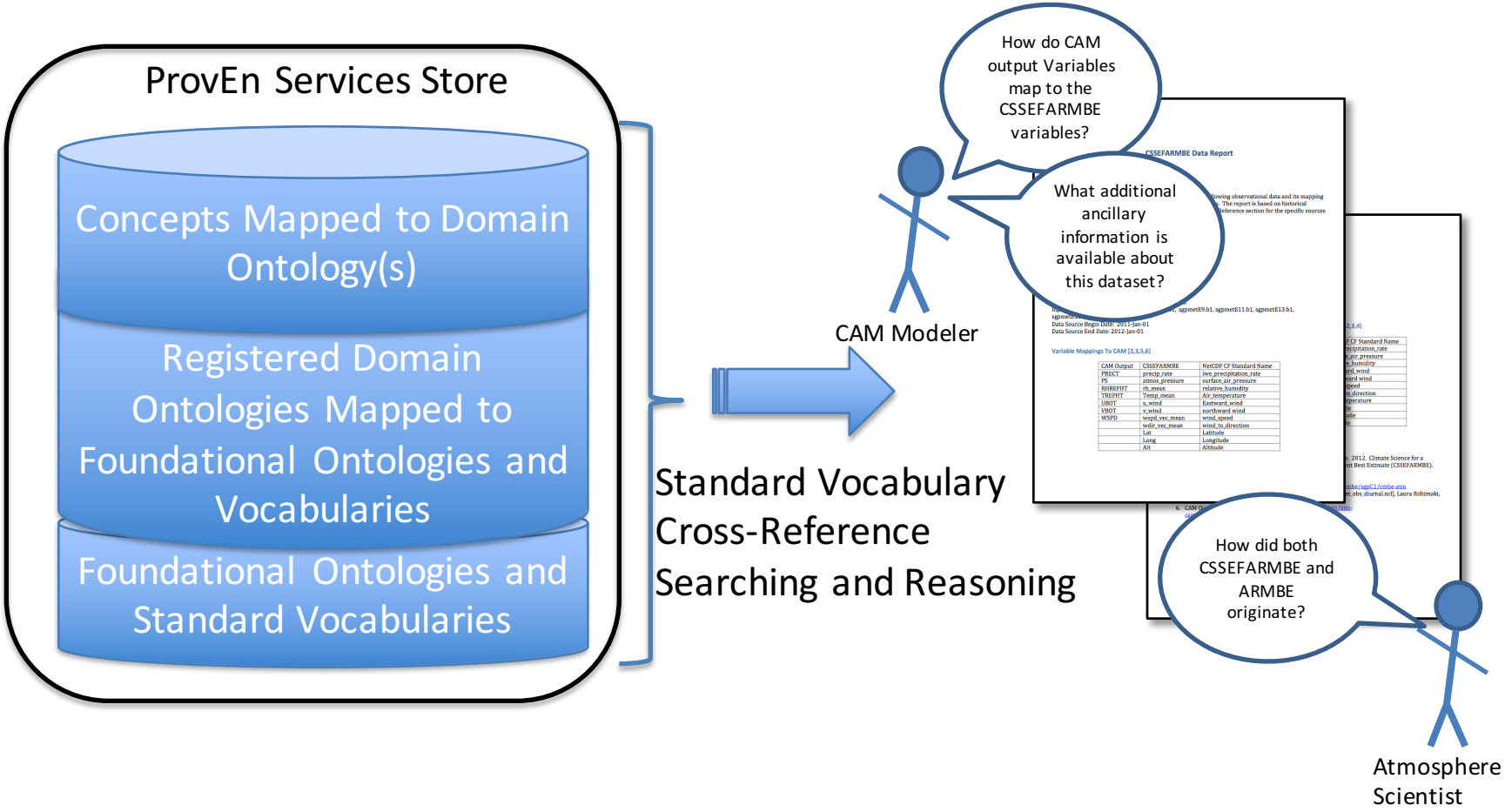
Making the Connection: Example Report Design to Support CAM Modeler and Atmosphere Scientist



Cross-referencing Information with ProvEn Services



ProvEn Services Producing CSSEFARMBE Data Origin Report



Goals of Provenance Environment (ProvEn) Services

- Capture historical information from any native source necessary to describe the origin of the dataset.
- Store this information in a cross-referenced form.
  - Achieved through the use of internationally recognized standards W3C, Dublin Core, CF
- Use this cross-referenced information to provide finished products to different kinds of consumers.

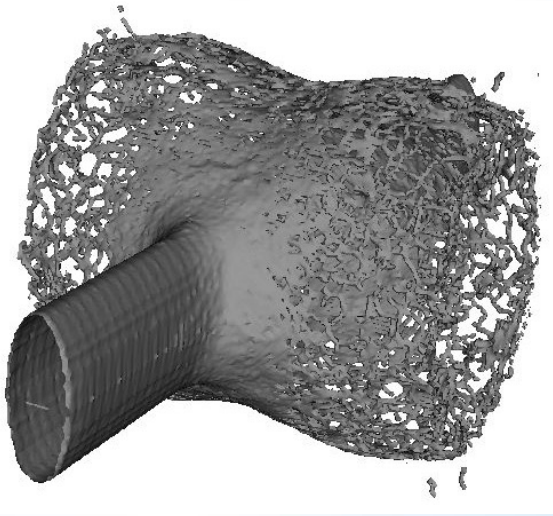
REPRODUCIBILITY

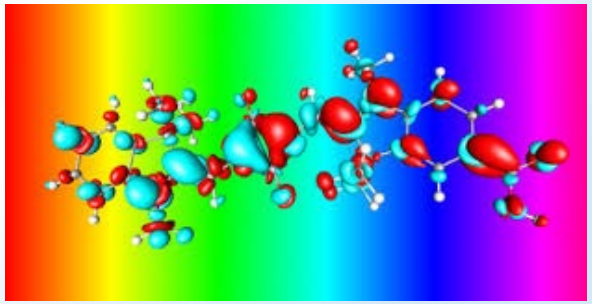
Reproducibility - Definition

- '*the ability of an entire experiment or study to be reproduced, either by the researcher or by someone else working independently. It is one of the main principles of the scientific method and relies on ceteris paribus (other things being equal)*'.
- Different aspects of Computational Reproducibility:
  - Numerical reproducibility i.e. has the algorithm been numerically designed to create the same results if replicated,
  - Experiment Reproducibility i.e. do we have all the information about the simulation to repeat it and
  - Execution Reproducibility i.e. can we recreate the execution environment, execution conditions (including system events) and system architectures.

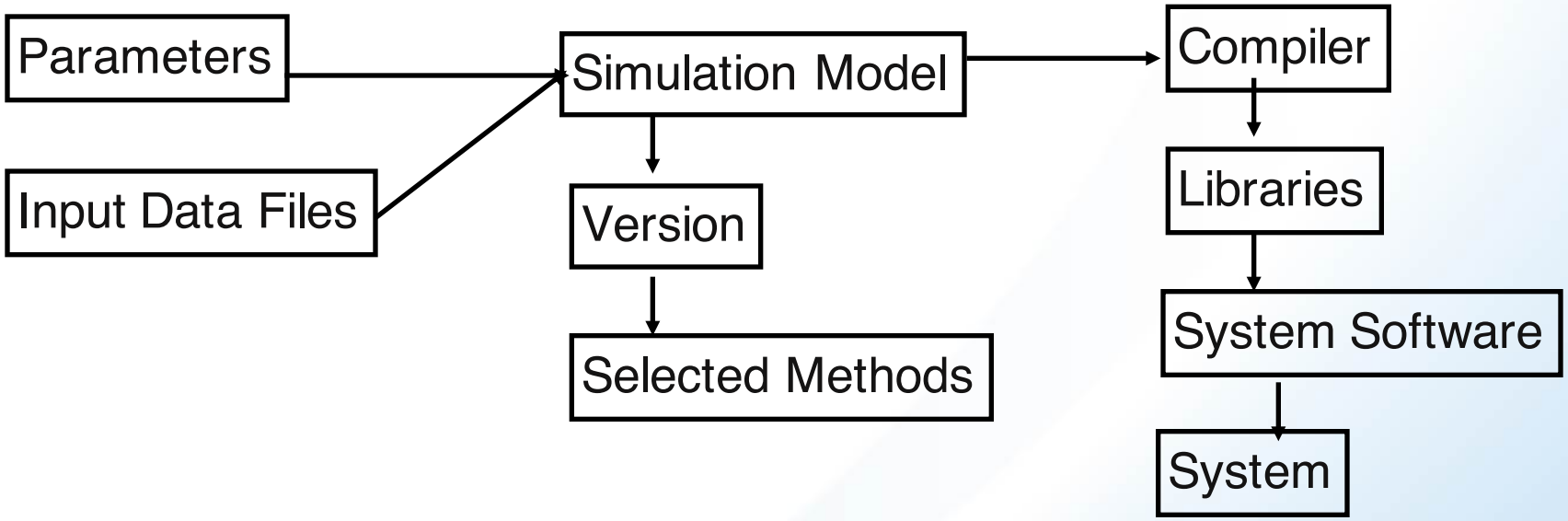
When is Reproducibility particularly needed?

- **Publication results are queried -** Can you show how the results presented in the paper where created, so that others can repeat your work and get to the same results?
- **Sharing new methods with a collaborator** - After several trials you found a new way of calculating important properties with a community code, can you show others what you did, so they can do it too?
- **Coordinating and comparing work with collaborators** - you are working in a large collaboration, in which each partner will contribute simulation runs, however the results are diverging more than expected, can you explain what all the key differences are between two runs?





What do we need to capture for a single application?



How can provenance help?

Provenance provides a standard language to describe what happened:

- Classes: Simulation Model, Compiler
- Subclasses: Name, Version, Vendor, Parameters
- Properties: WasCompiledWith

- Simulation Model -> WasCompiledWith -> Compiler
- **NWChem -> WasCompiledWith -> CF90**

This standard format makes it possible to search, analyze, compare , test and summarize the information captured in the provenance records.

Now consider a complex workflow, what do we need to capture?

- Application Details for each Application
- How are the applications connected with each other?
- Where are they executed?
- How do they influence each other?
- Did changes occur during execution?
- Did the workflow management system influence the execution and how?

Traditional Workflow Provenance

Today most workflow management systems have at least a basic level of provenance capture integrated.

The level of detail captured varies however, at minimum they will collect the name of data files, name of applications run and the workflow itself (how the applications where connected)

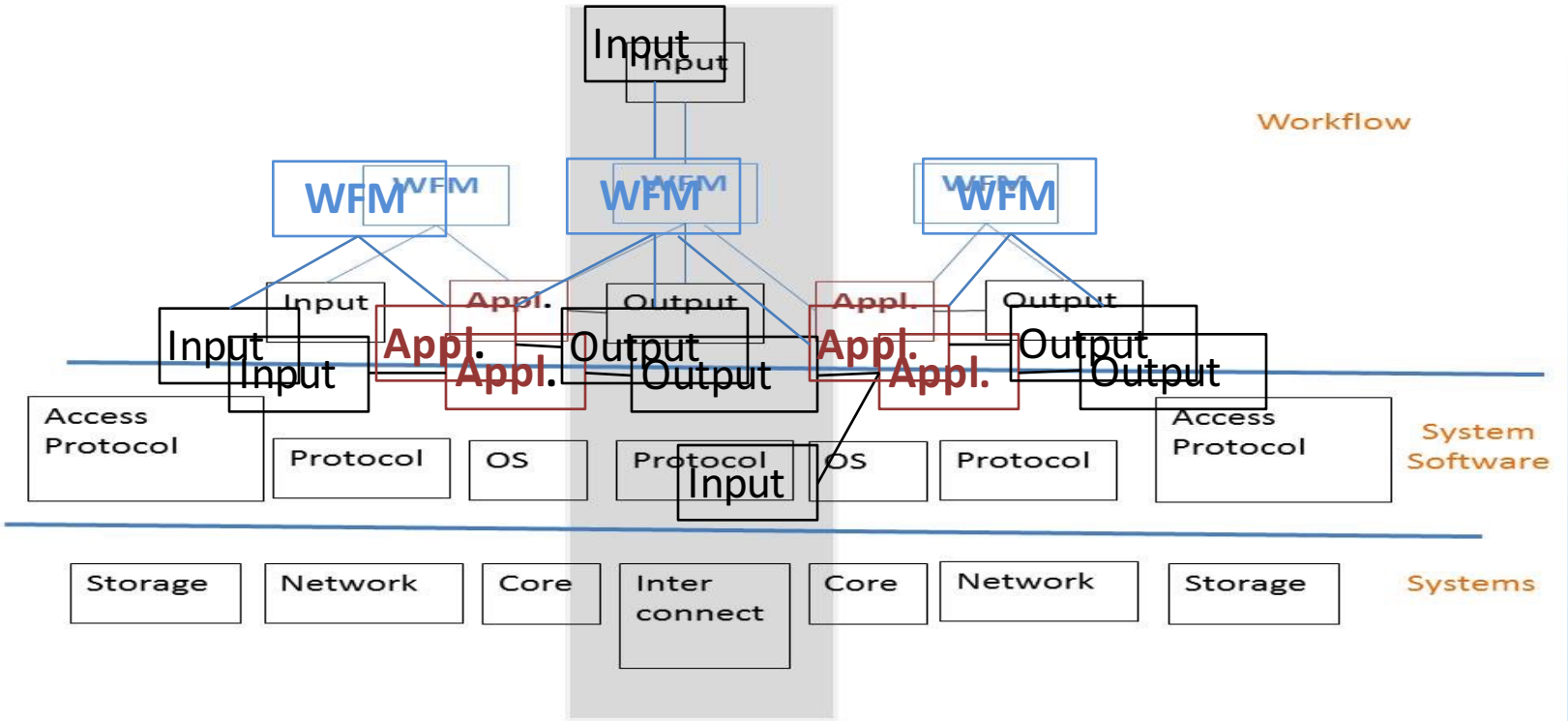
Not all workflow systems capture provenance in a standard provenance format, but rather use their own e.g. VisTrails

Other systems that collect provenance are: Kepler, Pegasus

Special is the Galaxy that captures fine grained user actions too used to create workflows and allows users to roll back actions

Is this level of provenance enough to support reproducibility?

Workflow Provenance for Reproducibility



Opening Up Application Black Boxes

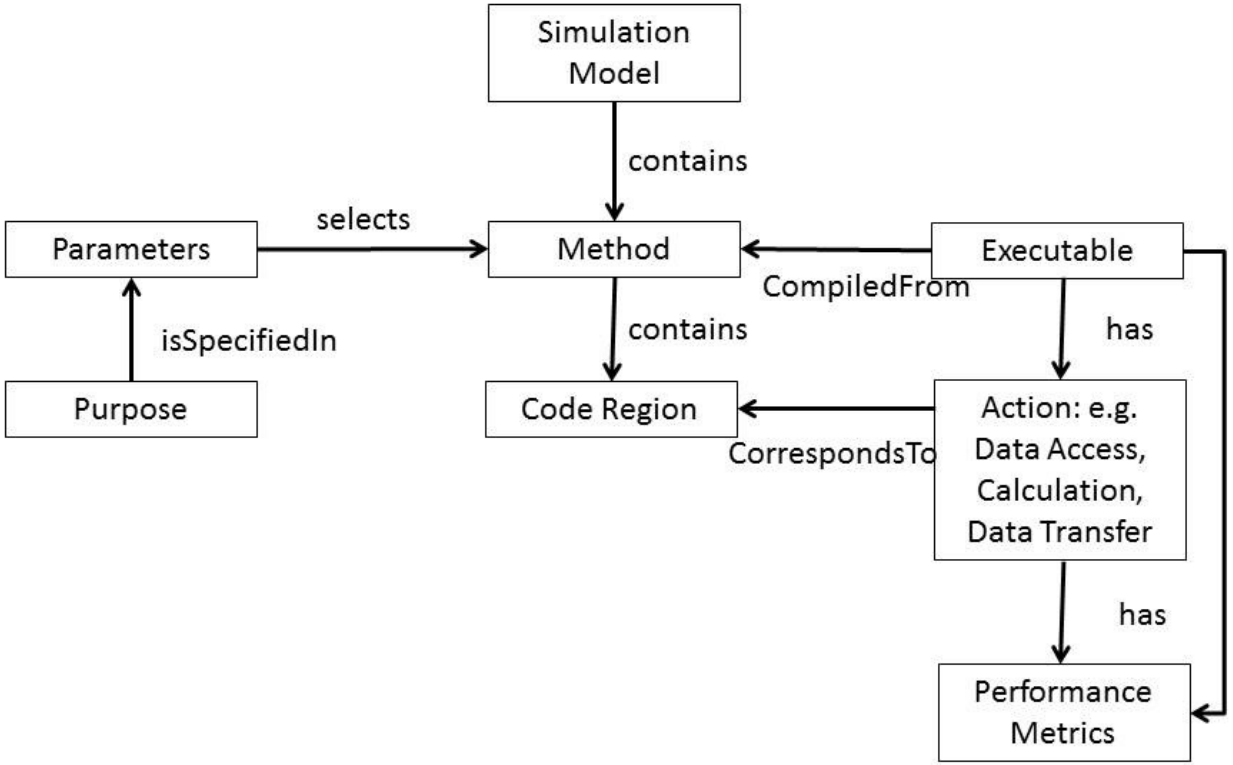


Figure 5: Simulation Model Representation

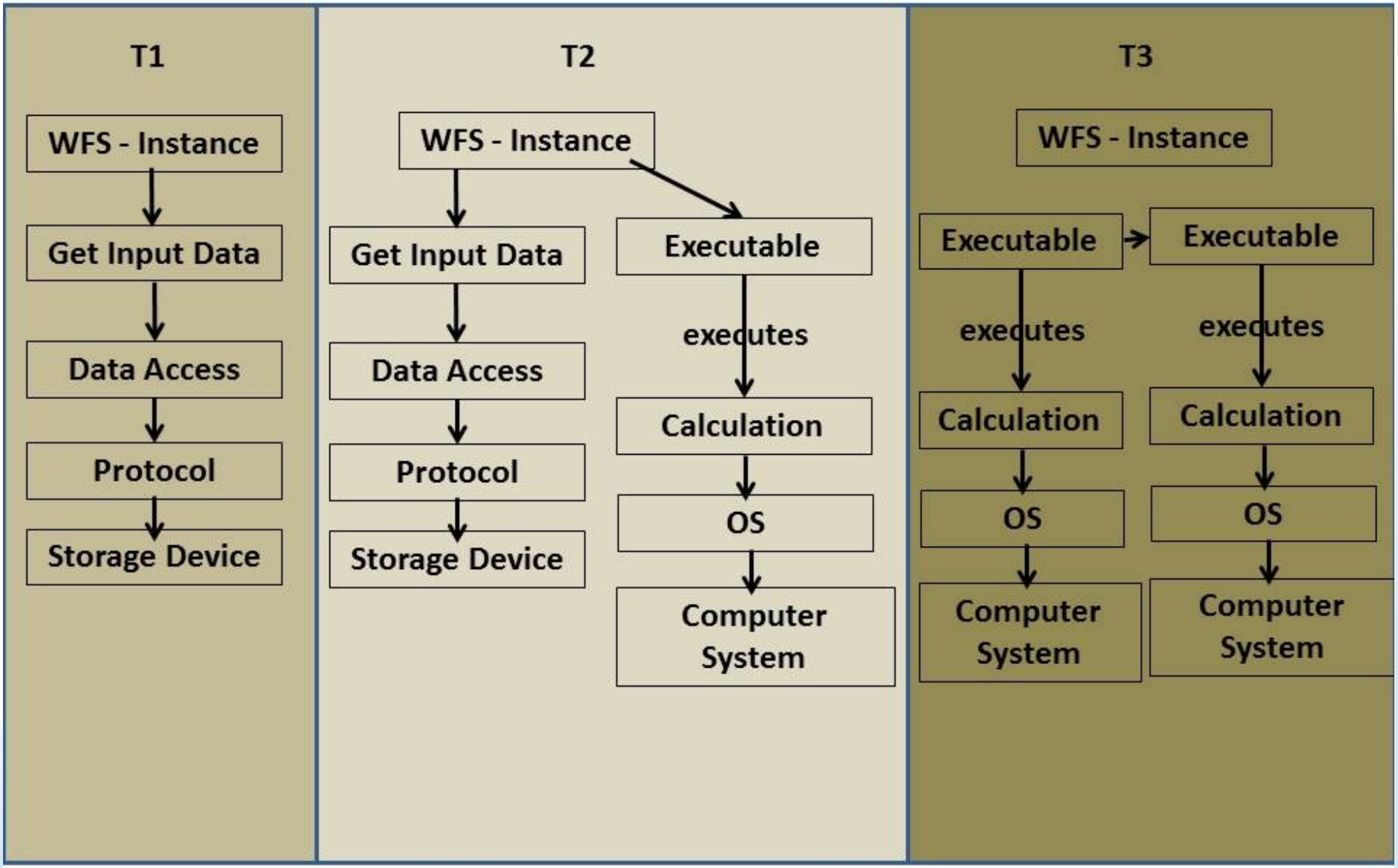


Figure 8: Workflow Evolution Time Series

New OPM based WorkFlow Provenance Model - OPM-WFPP

Extension of OPM, in OWL.

Classes: 59

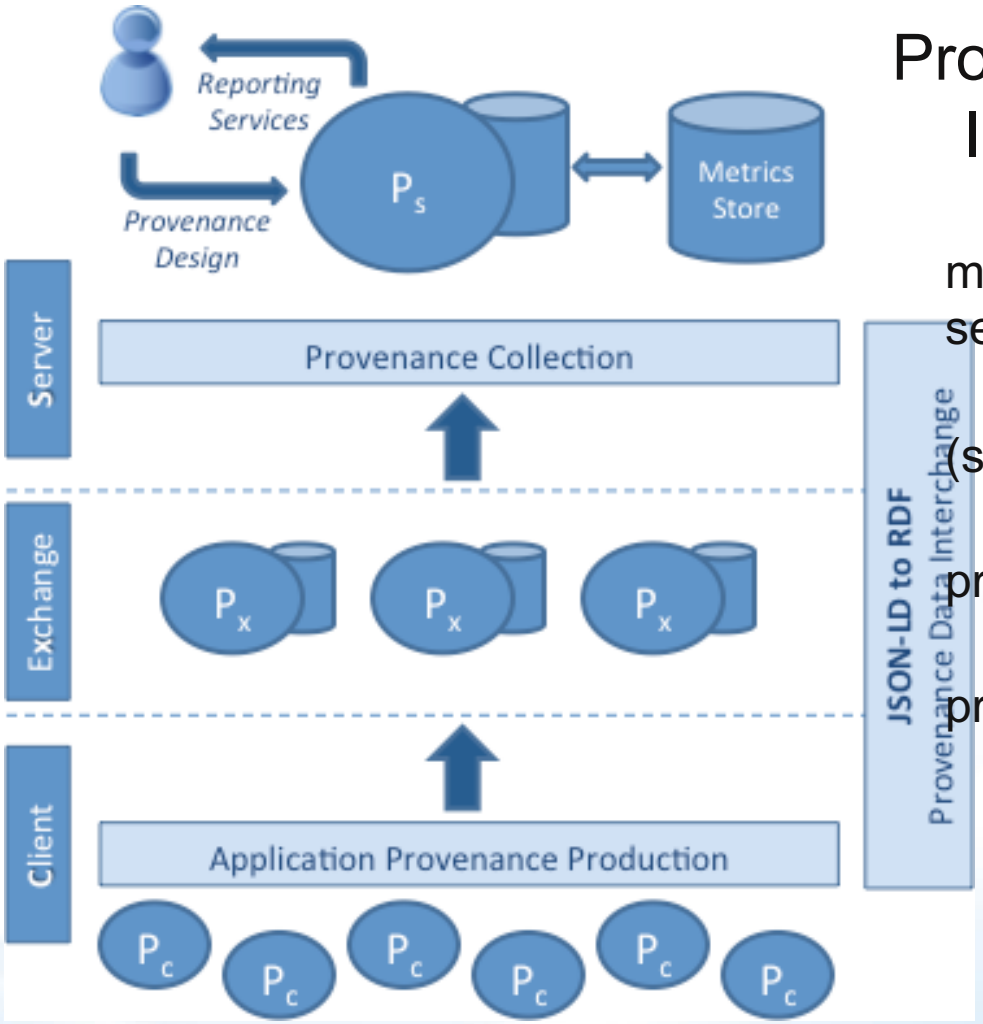
Subclasses: 56

Properties (Relationships): 44

Describes characteristics of: use cases, applications, workflows and system

Captures time series of performance metrics across all levels from system to workflow - implemented as library and collection tools

Provenance Environment (ProvEn) Architecture



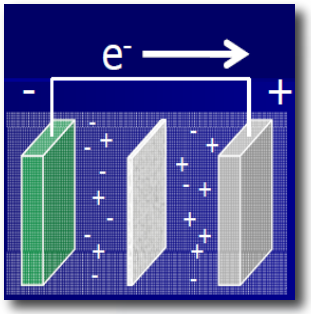
ProvEn Services Infrastructure

- Provenance capture through messaging services and web service APIs
- Server / provenance consumer (semantic information, triple store)
- Client API library / provenance producer
- Time-series client/server (in progress, InfluxDB)

PERFORMANCE ANALYSIS

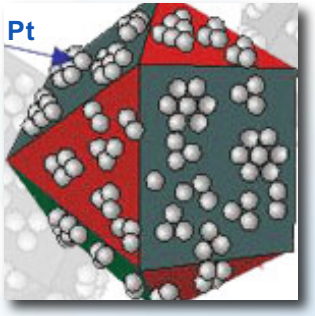
National Synchrotron Light Source II (NSLS-II): Enabling the Nanoscience Revolution

- Brightest synchrotron in the world
  - 10,000 times brighter than original NSLS
  - 2,000 users per year (4,000 in FY17)
  - Running on low-cost NYPA hydropower – significant factor in DOE decision to site NSLS-II at Brookhaven Lab
- Enabling solutions to pressing energy challenges, e.g.:
  - Advanced electrical storage
  - High-temperature superconductors for the electric grid
  - Fuel cells based on nanocatalysts
  - Plant/environment interactions
- Integrated Centers for Energy Science combine:
  - Synergy with the Center for Functional Nanomaterials (CFN), core programs
  - State-of-the-art integrated tools for studies under real world conditions
  - Outreach to university and industry

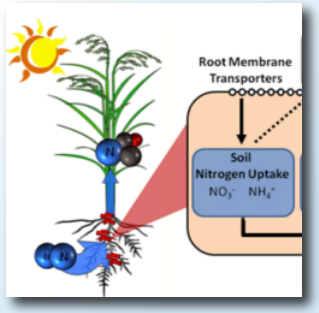


Energy Storage


High Tc Superconductors



Nanocatalysts for Fuel Cells



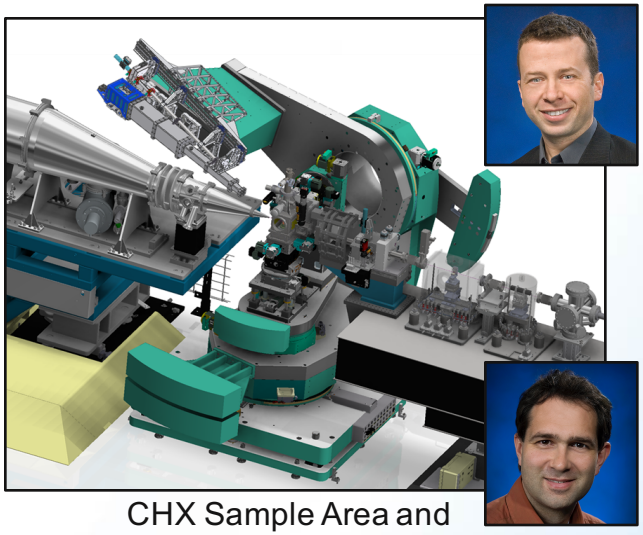
Plant Bioscience

Urgent Facility Requirements

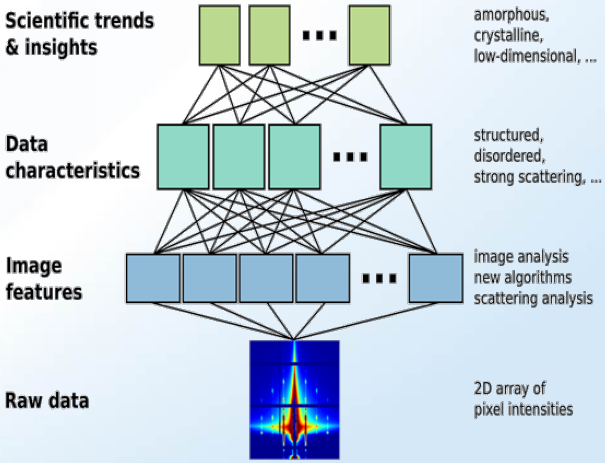
- **Streaming Analysis and Decision Making**
  - **Providing real time scientific information to allow researchers to effectively control the direction of their experiment and optimize its scientific outcome**
- Multi-Source Data Analysis
  - Complex challenges require the use of more than one investigative method (different experiments & simulation) to provide the necessary insights
- Effectively Utilizing Curated Knowledge Repositories
  - Knowledge repositories can provide critical context sensitive background for the interpretation of new results

Data Driven Discovery at NSLS-II Beamlines

- Scientific Goal
  - Enable reliable, real time, data-driven steering of experiments
- Achievement
  - New holistic approach to streaming data analysis and decision support
    - Integrating results from 7 projects, including ASCR-funded research
    - Applying results to Coherent Hard X-Ray (CHX) beamline at NSLS-II
    - Workflow provides streaming statistics, machine learning and visual analytics for decision support
- Impact
  - Enable data-driven steering of experiments to optimize their scientific outcomes
  - With 4.5 GB/s sustained data rates, CHX is a good test ground for higher rate instruments, such as HXN (1 – 5 TB/s in burst)
  - Solutions applicable to other beamlines and light sources, CFN, eRHIC, exascale simulations, Electric Power Grid, Observational Sensor Networks



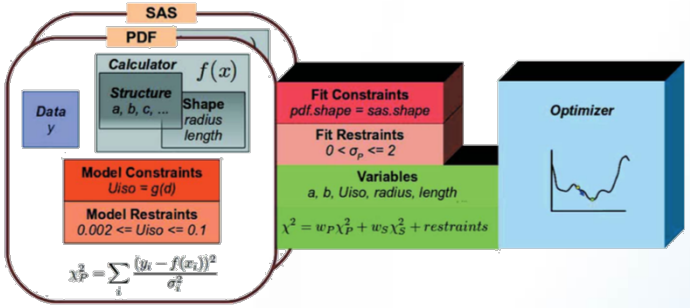
CHX Sample Area and Scientists



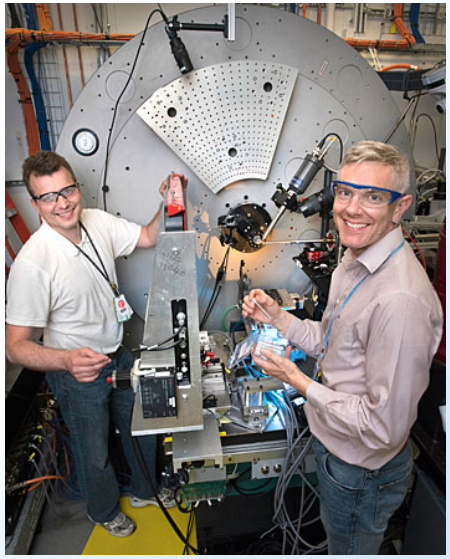
Workflow at CHX

Integration of Data Analysis and Theory in Support of Scientific Discovery

- Scientific Goal
  - Operational integration of multi-modal experimental and numerical modeling results
- Achievement
  - Software DiffPy-CMI - determines stable structures for complex materials using results from different experimental modalities in combination with *ab initio* and atomistic scale modeling (MatDeLab, NWChem)
- Impact
  - DiffPy-CMI supported 14 materials studies since its first release
    - 2014 release downloaded ~1300 times (several hundred users)
    - 2016 release downloaded ~130 times (~50 users)
  - In use at NSLS-II XPD beamline
  - Similar approaches under development, e.g., at the NSLS-II ISS beamline to interpret *in operando* device performance behavior; also of interest to NSLS-II FMX, AMX, and NYX beamlines



www.diffpy.org

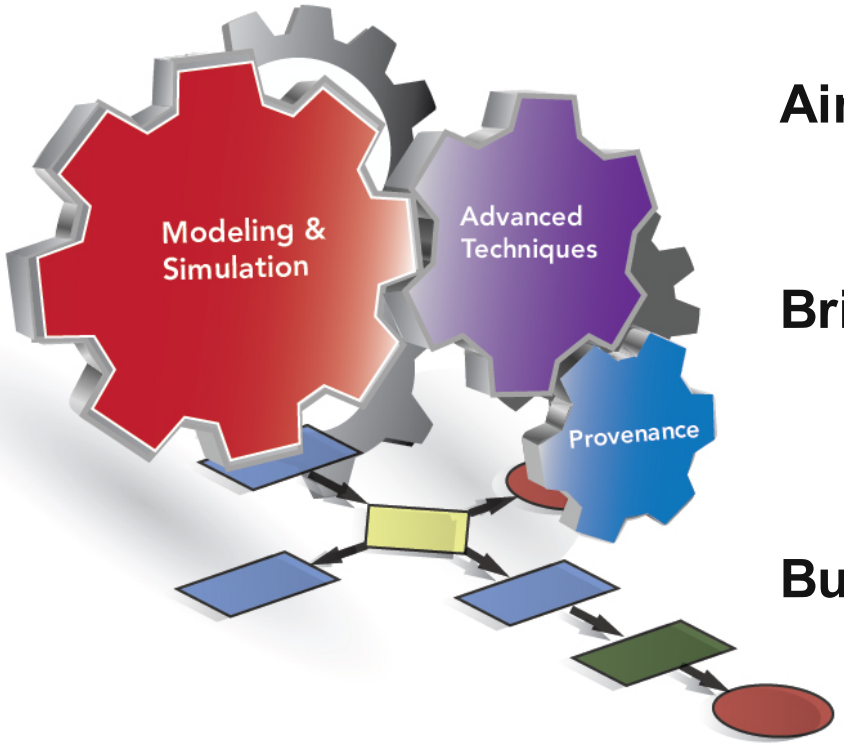


Results from complex modeling study recently published in *Nature Communications*

Challenges in in-situ experimental analysis

- **Goal** - Provide enough targeted information to the scientists, early enough, to enable them to take critical decisions on steering of the data taking and its analysis
- **Critical characteristics**:
  - Speed, Accuracy, Completeness (incl. background, prediction)
  - Information selection and representation
  - Different programing languages, programming models, heterogenous data, computing and networking infrastructure
  - **Essential - Reliable in Time Result Delivery**

DOE ASCR - Integrated End-to-End Performance Prediction and Diagnosis for Extreme Scientific Workflows



- **Aim to provide an integrated approach to the modeling of extreme scale scientific workflows**
- **Brings together researchers working on modeling / simulation / empirical analysis, workflows and domain scientists**
- **Builds upon existing research much of which has focused to date on large-scale HPC systems and applications**

Explore in advance – Design-space exploration & Sensitivity Analyses

Optimize at run-time – Guide execution based on dynamic behavior

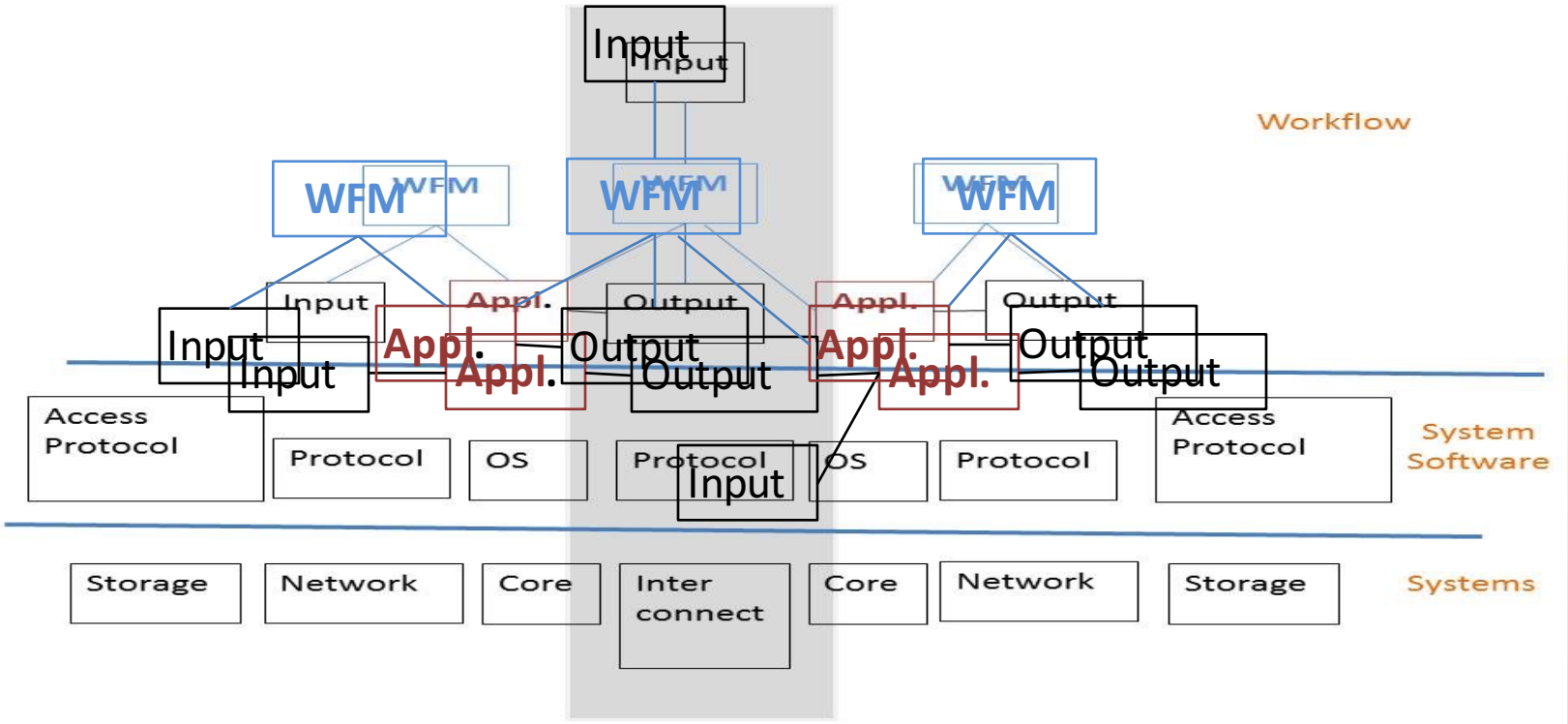
Expanding Provenance: Empirical Information Gathering

Today we only have hypothesis on what causes the variability in workflow performance or how performance could be improved

ASCR IPPD Project will use provenance to capture empirical performance information from workflows and systems to:

- Collect quantitative performance information to investigate workflow performance variability, degradation, sensitivity and impact
- Provide empirical data backed assessments of particularly prevalent performance bottlenecks and sources of performance variability
- Provide a record of performance changes over time that can be correlated with changes to applications, workflows and systems

Correlating Performance Relevant Information and Metrics



Opening Up Application Black Boxes

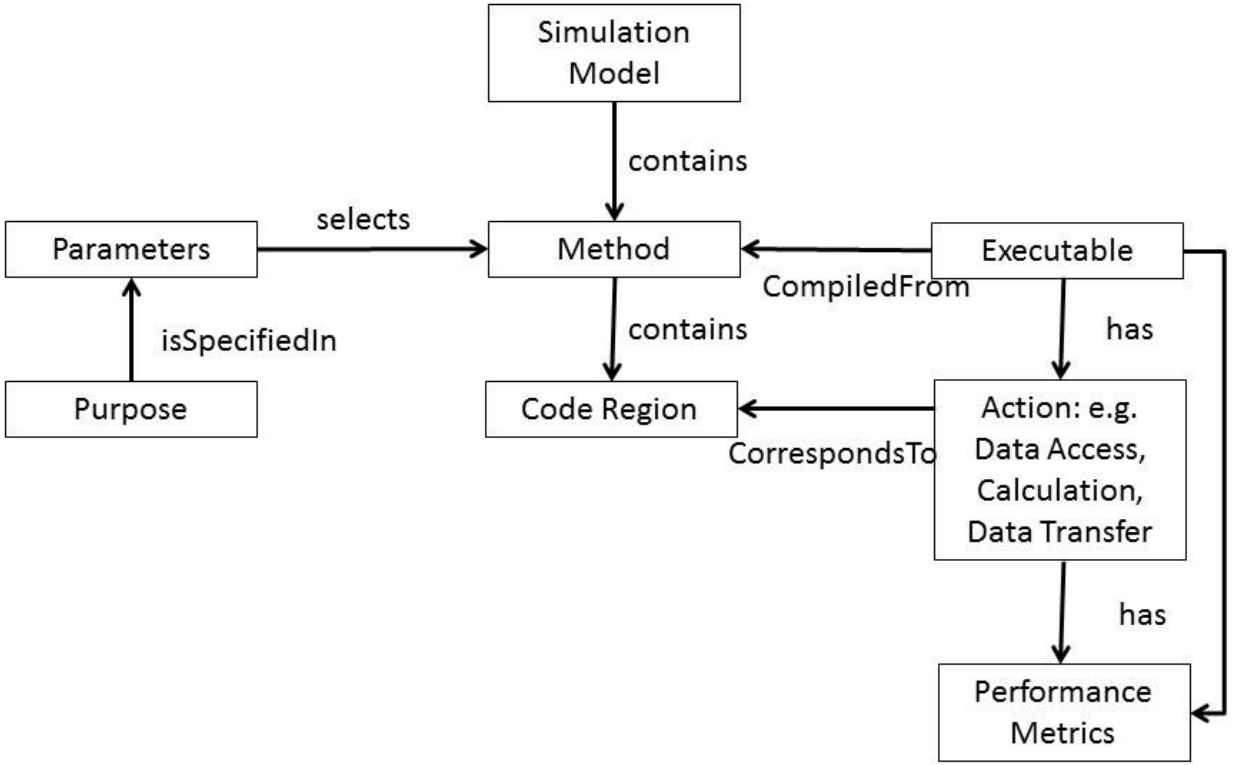


Figure 5: Simulation Model Representation

Representing Parallelism (2)

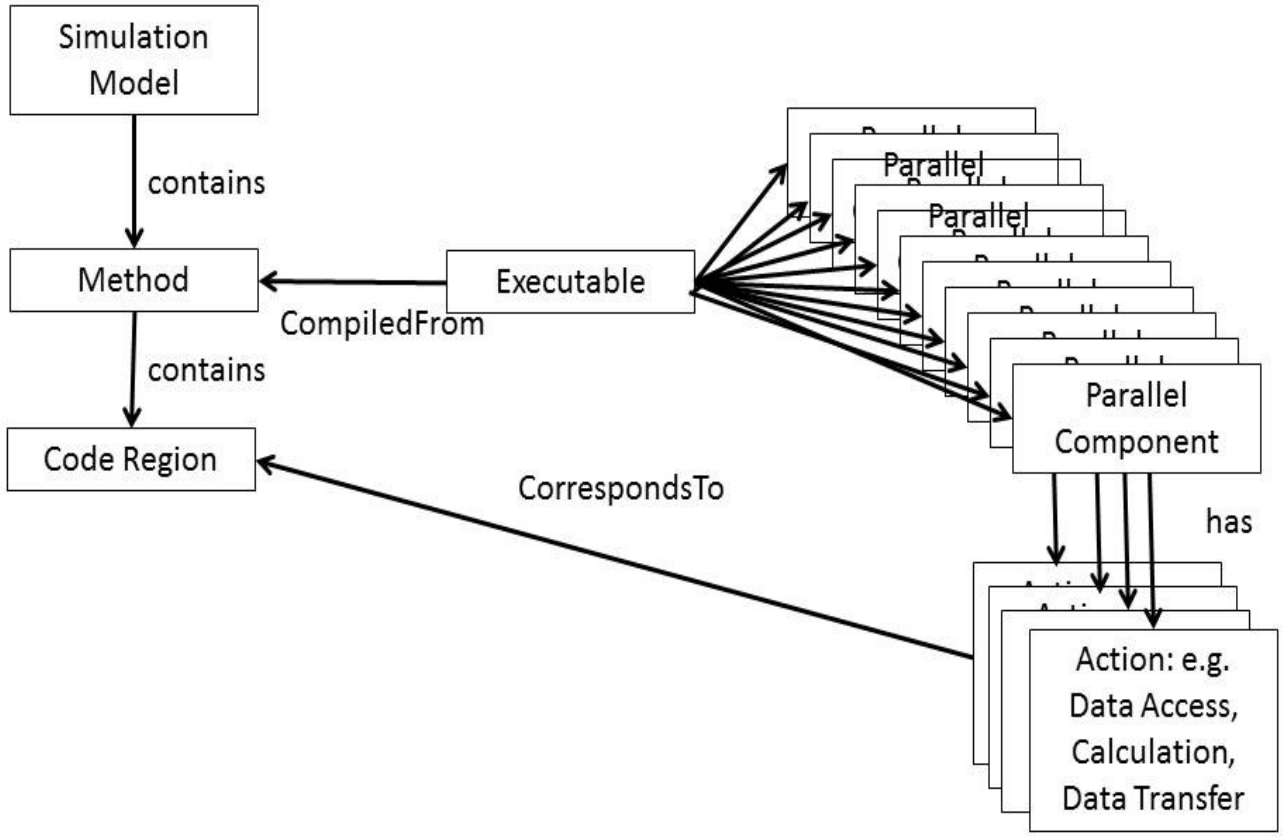


Figure 4: Representations of Parallel Programs

Capturing Interdependencies (3)

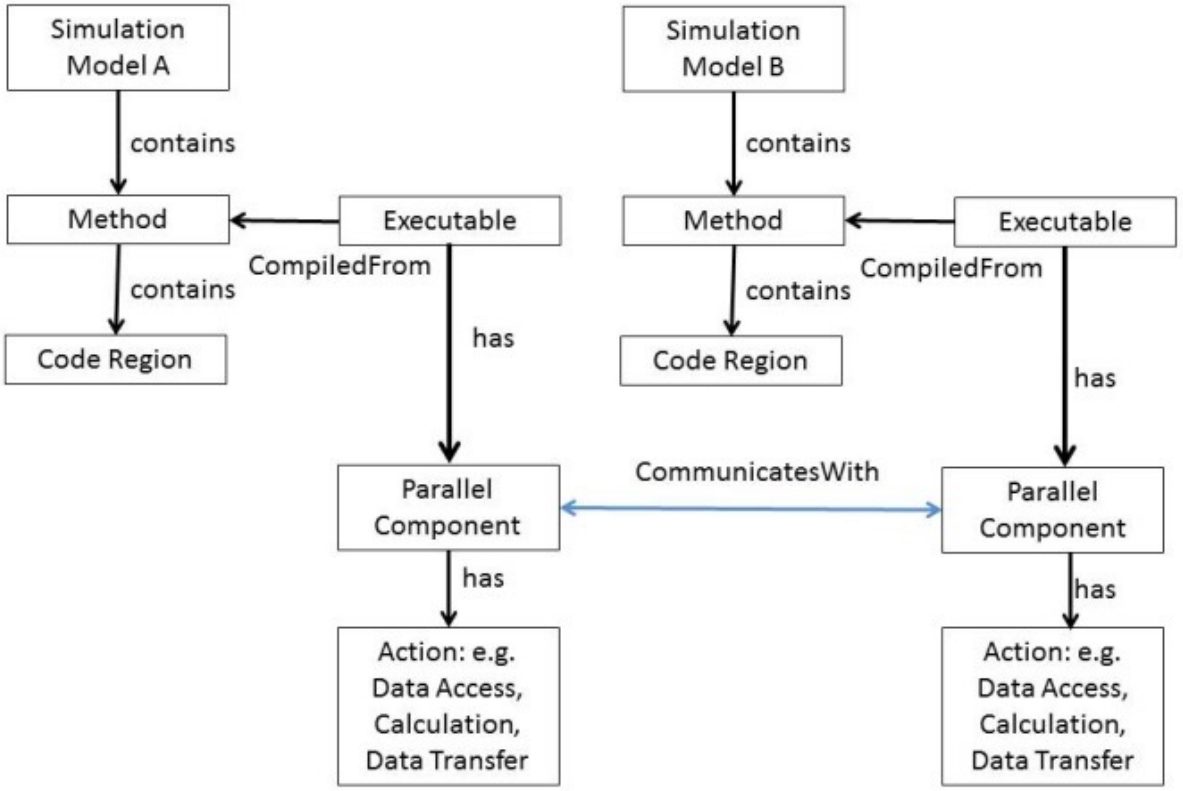


Figure 5: Capturing Interdependencies between Parallel Components at Runtime

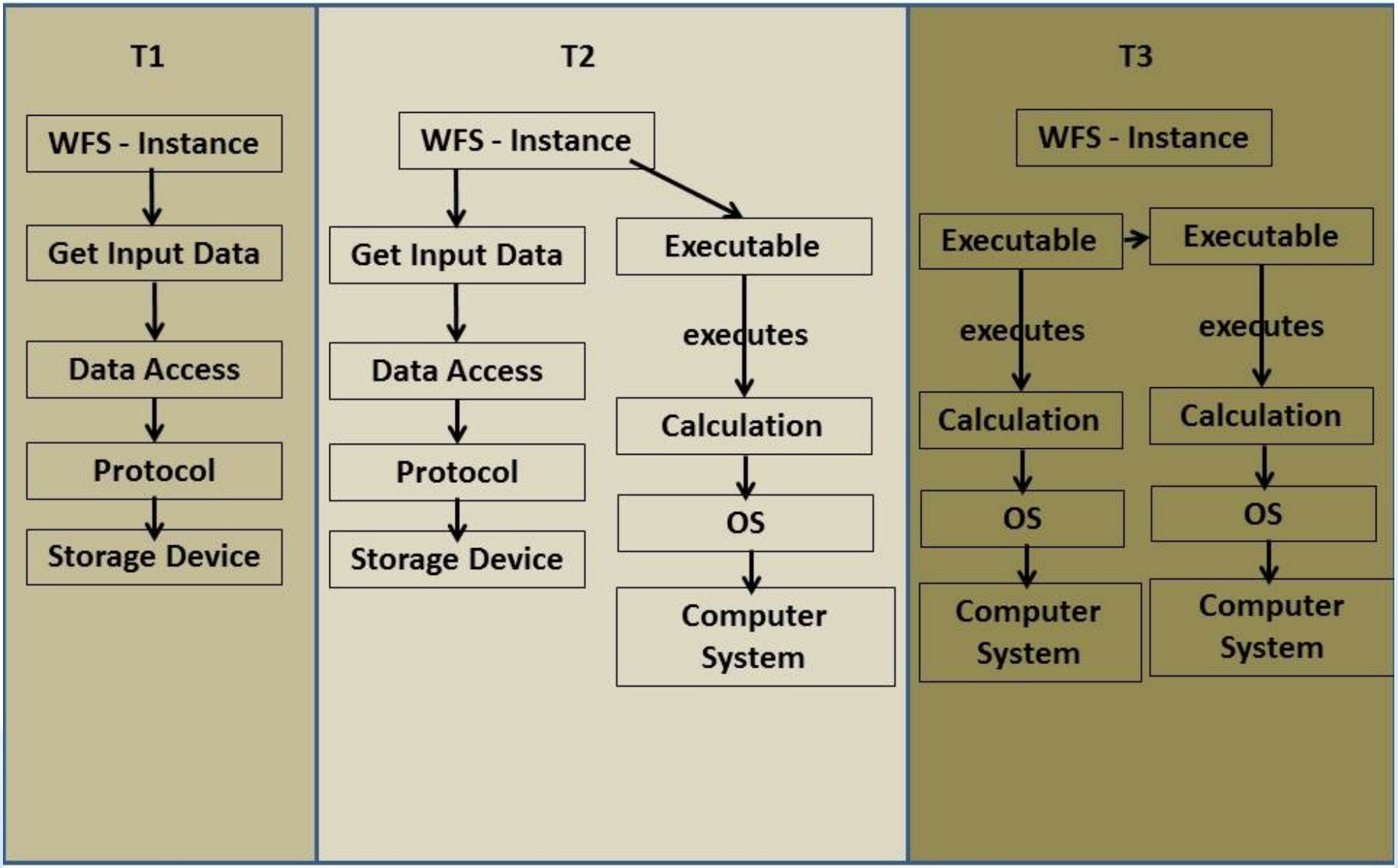
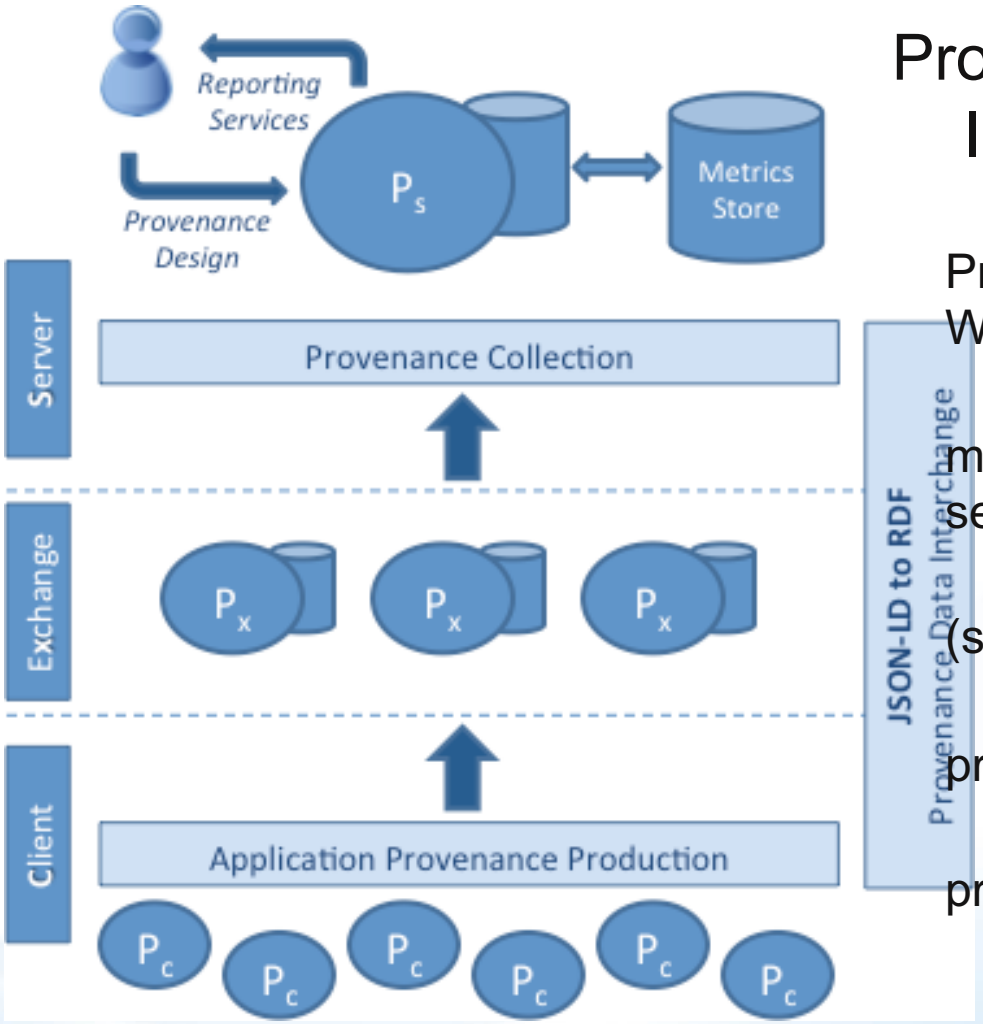


Figure 8: Workflow Evolution Time Series

Provenance Specific Characteristics and Metrics

- Workflow Script Performance Determining Characteristics (incl. e.g. number of tasks)
- Workflow Script Instance Performance Metrics (incl. some outlined in [9],[10])
- Code Region Performance Metrics to be collected for each call to the code region, for each core, (selected list of metrics informed by [9], [10])
- Computer System Performance Characteristics (incl. [17])
- Computer System Performance Metrics (as collectable by e.g. SYSSTAT [18])
- Wide Area Network Performance Characteristics
- Wide Area Network Performance Metrics
- Interconnect Performance Characteristics
- Interconnect Performance Metrics
- Storage System Performance Characteristics
- Storage System Performance Metrics

Provenance Environment (ProvEn) Architecture

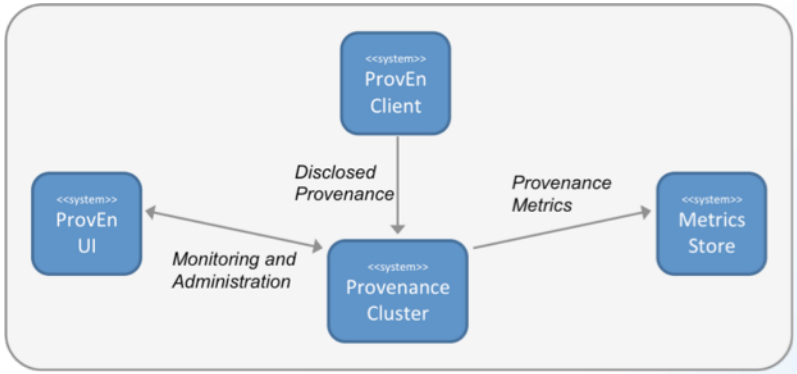


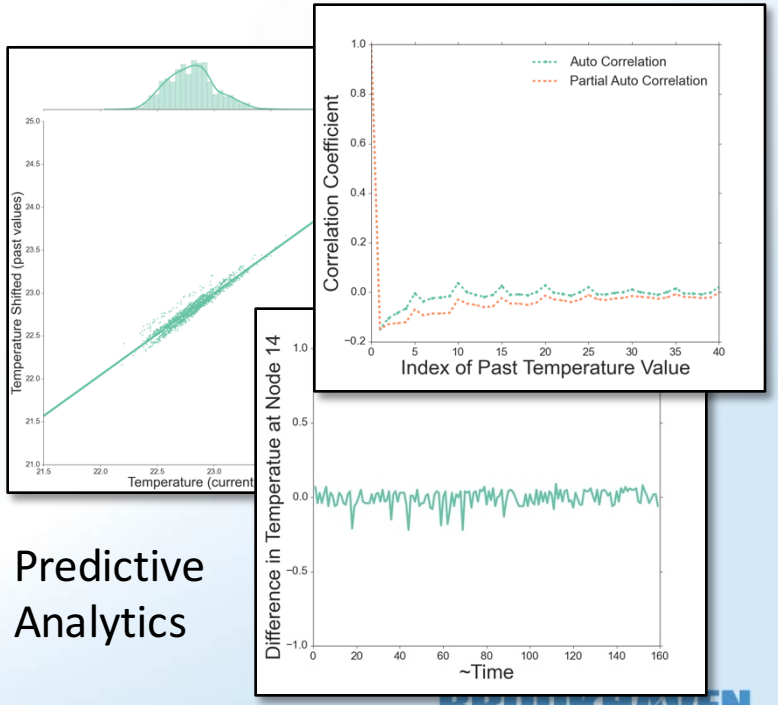
ProvEn Services Infrastructure

- New Workflow Performance Provenance Ontology (OPM-WFPP)
- Provenance capture through messaging services and web service APIs
- Server / provenance consumer (semantic information, triple store)
- Client API library / provenance producer
- Time-series client/server (in progress, InfluxDB)

ProvEn Overview

- Provenance Environment (ProvEn) - A Provenance production and collection framework.
- Provides services and libraries to collect provenance produced in a distributed environment
- ProvEn Client API aids in the production of provenance from client applications
- The following types of provenance are collected:
  - Time series-based information from a system/host perspective
  - Performance metrics tracking from an application/workflow perspective
- ProvEn enables building of accurate Machine Learning models by capturing detailed footprints of large-scale execution traces.
- ProvEn will support identification of sources of performance variability in streaming analysis workflows, and provide runtime guidance to resource allocation systems.





Predictive Analytics

What Can You Search?

Example Activity	Search Capability
Workflow Tracing and Diagnostics	Graph traversal
Verifying of metric occurrence	Pattern matching, query by example
Workflow stage completion verification	Transitive closure, inferencing
Ranges of metrics	Value and term based searches
Extracting provenance for single workflow	Named graph
Translating two workflow graphs into common representation	Reconstruct graph into new form
Detecting differences/similaries in processing	Cross-referencing

Questions?

Kerstin Kleese van Dam, kleese@bnl.gov