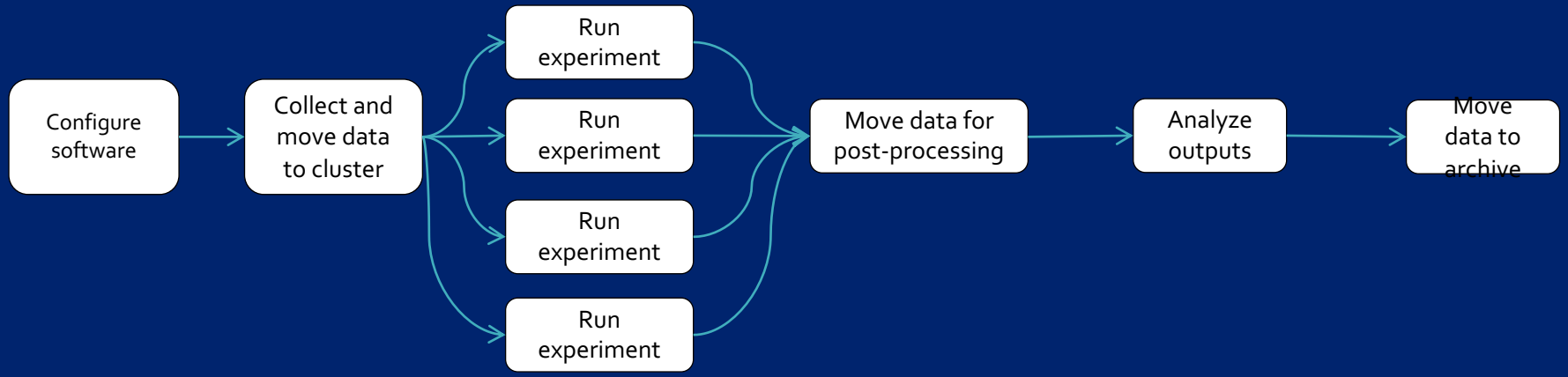
SDSC Summer Institute 2021

Title: Developing Data Science Workflows with Kepler
Instructor: Shweta Purawat
Date: 08/03/2021
Location: Breakout Room

Day 1!

- **To interact with Expanse HPC system:**
  - Launching and managing jobs
  - Managing data on the file system

Pipeline



CHOICES:

- Scripting
- Makefiles
- Workflow systems

Part1: Introduction to Scientific Workflow

Part 2: Scientific Workflow Examples

Part3: Introduction to Kepler

Part4: Kepler Demo

Part1: Introduction to Scientific Workflow

- Define what is a workflow
- Describe three ways workflows empower scientific research
- List out positive impacts that scientific workflows have on society

Real World Example #1

- **Traveling to a new city by air**

Book Tickets → Pack your bags → Drive to Airport → Check In (in person) → Fly

Check In (online) → Drive to Airport

Drive to Airport → Check In (online)

What is a scientific workflow?

A scientific workflow is a **set of computational steps** that scientists use to generate results.

That may involve accessing multiple applications and databases, and processing the data using computationally intensive jobs on high-performance clusters.

Scientific workflows emerged as an answer to the need to **combine multiple Cyberinfrastructure components** in automated process networks.

Computing Today has Many Shapes and Sizes

kubernetes

amazon web services

NVIDIA TESLA

+

Requires software for dynamic coordination and resource optimization

COMPUTING AT SCALE

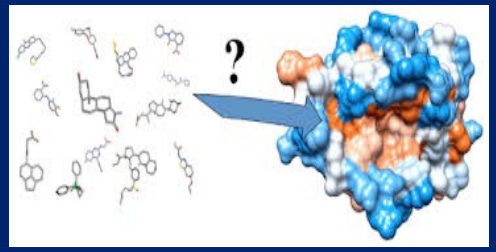
BIG DATA

Workflow Systems

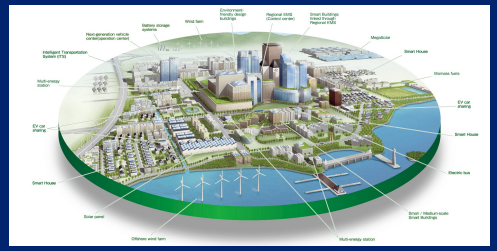
Enables dynamic data-driven applications

Impact of Scientific Workflows

Computer-Aided Drug Discovery



Smart Cities



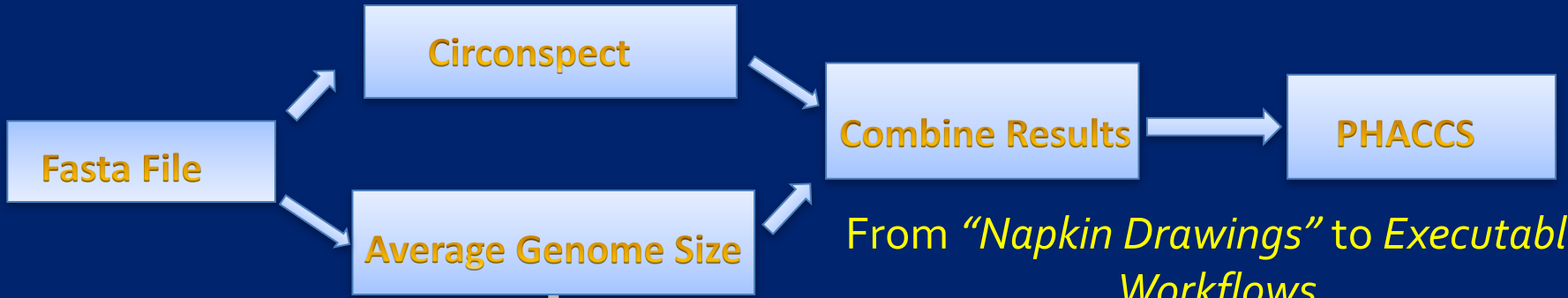
Disaster Resilience and Response

Smart Grid and Energy Management

Personalized Precision Medicine

Smart Manufacturing

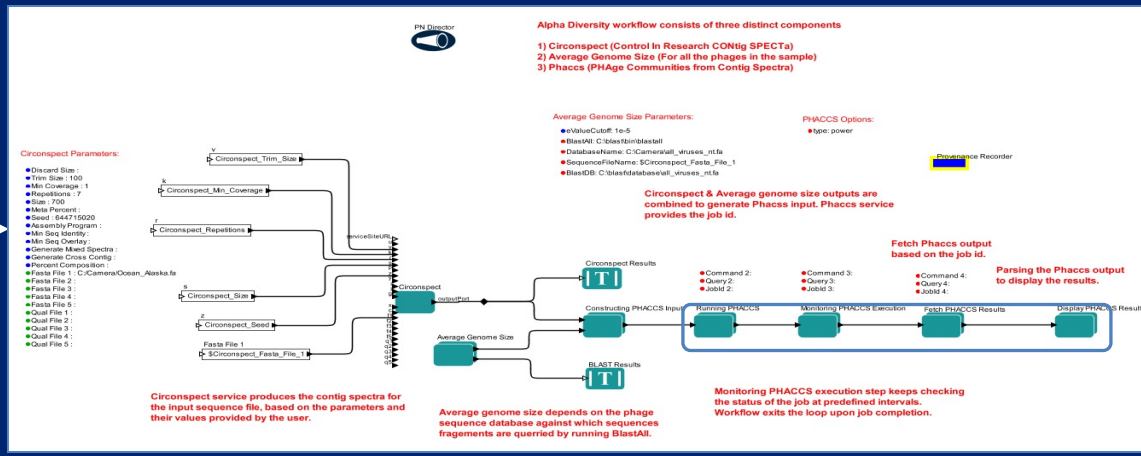
The Big Picture is Supporting the Scientist



From *"Napkin Drawings"* to *Executable Workflows*

Conceptual SWF

Executable SWF



Build Once, Run Many Times…

- The same workflow should support experimental work and dynamic scalability on many platforms
- Scalability based on:
  - data volume and velocity
  - dynamic modeling needs based on various optimization criteria
  - changes in network, storage and computing availability

Accountable Science

- **Scientific experiments involve many:**
  - Data
    - Which data came from which source?
    - Which version of the data?
  - Processes
    - What processes ran in which order?
    - Which libraries were used?
  - Collaborators
    - Who produced what?

Provenance

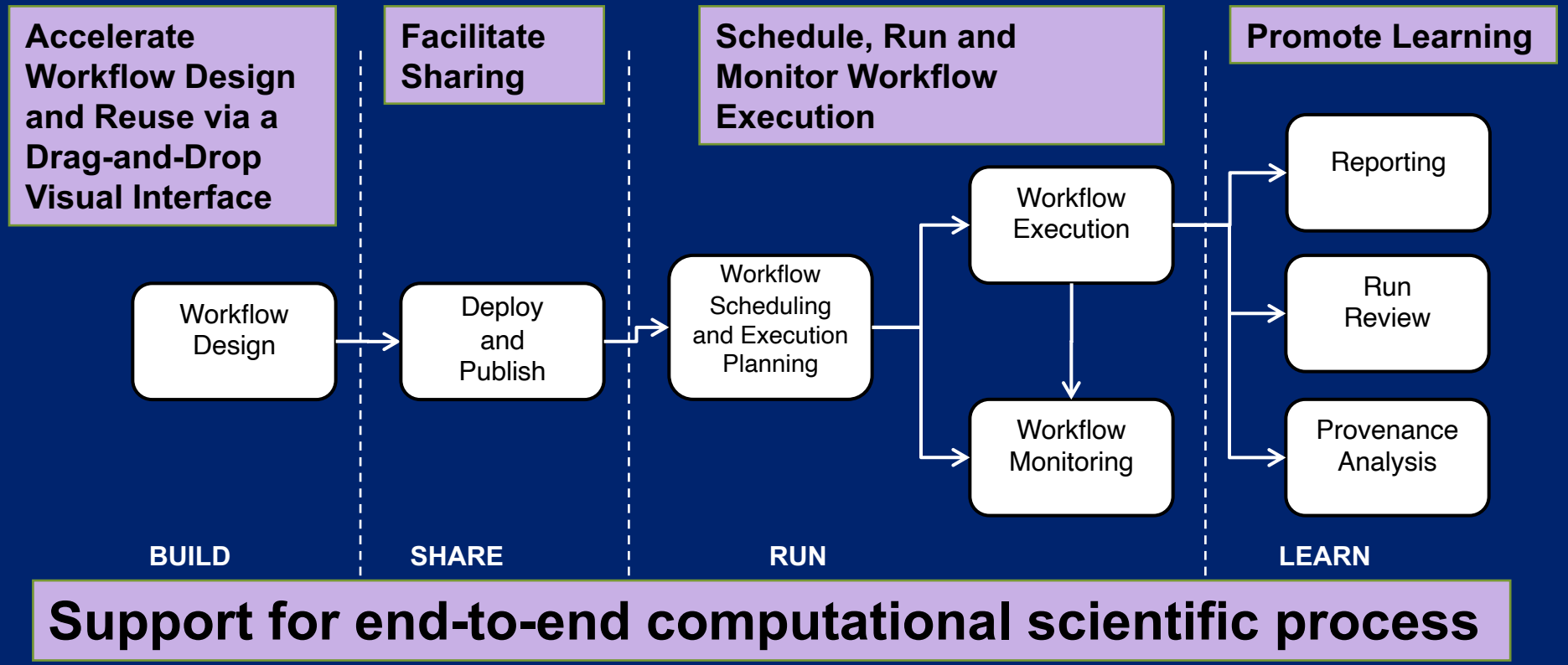
Provenance Helps with Accountability and Reproducibility

- Data and Process Provenance
  - Inputs, outputs, intermediate results
  - Workflow: actors, links, parameters, etc.
- Reproducibility with little effort

Collaborate: Save and Share

- **Documentation of all aspects of an analysis**
- **Share with your Team**
  - Final Products
  - Reports
  - Provenance

Workflows are a part of Cyberinfrastructure



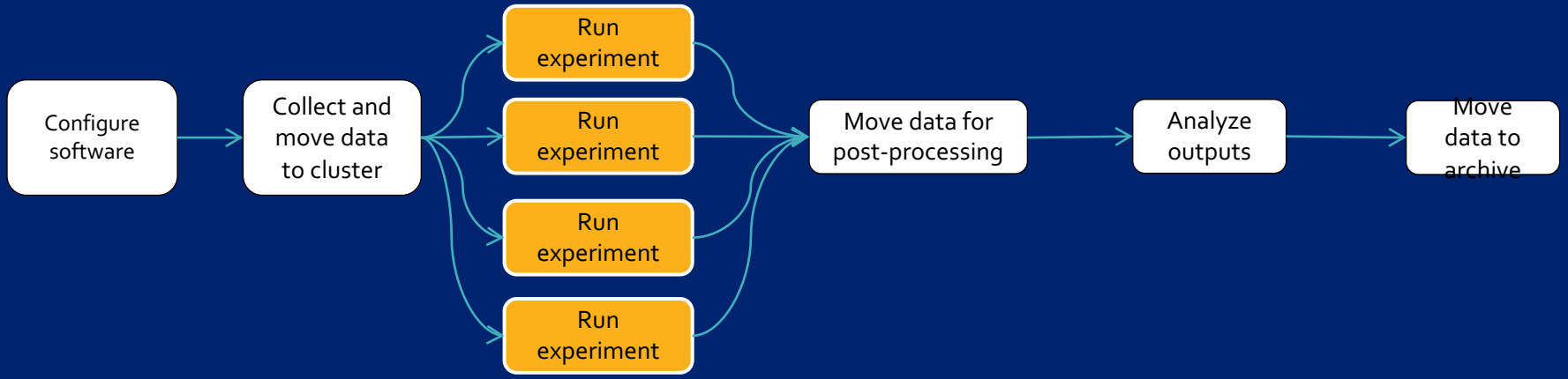
Accelerate Science

- **Speed** : build using Drag-and-Drop Visual Interface
- **Adaptability** : ability to work across multiple systems
- **Integration** : interconnect with other workflows
- **Customization** : declare what needs to be done,
  give freedom of execution
- **Reusability** : a part or entire workflow

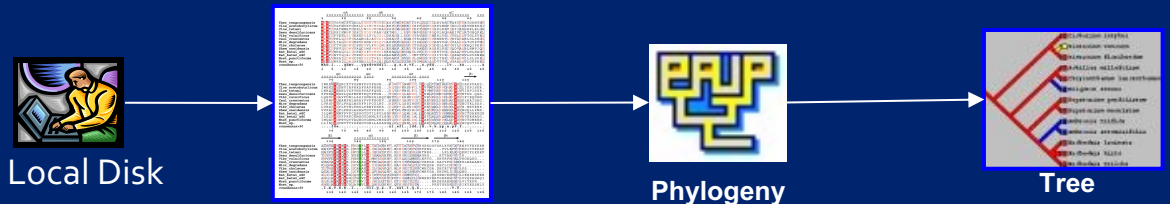
Part 2: Scientific Workflow Examples

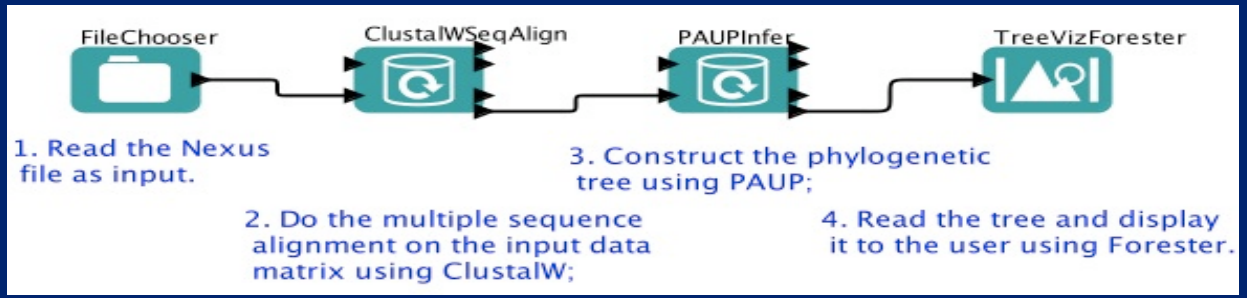
- Describe key metrics useful to distinguish workflows
- List out examples of different workflows types

Infinite ways to create workflows

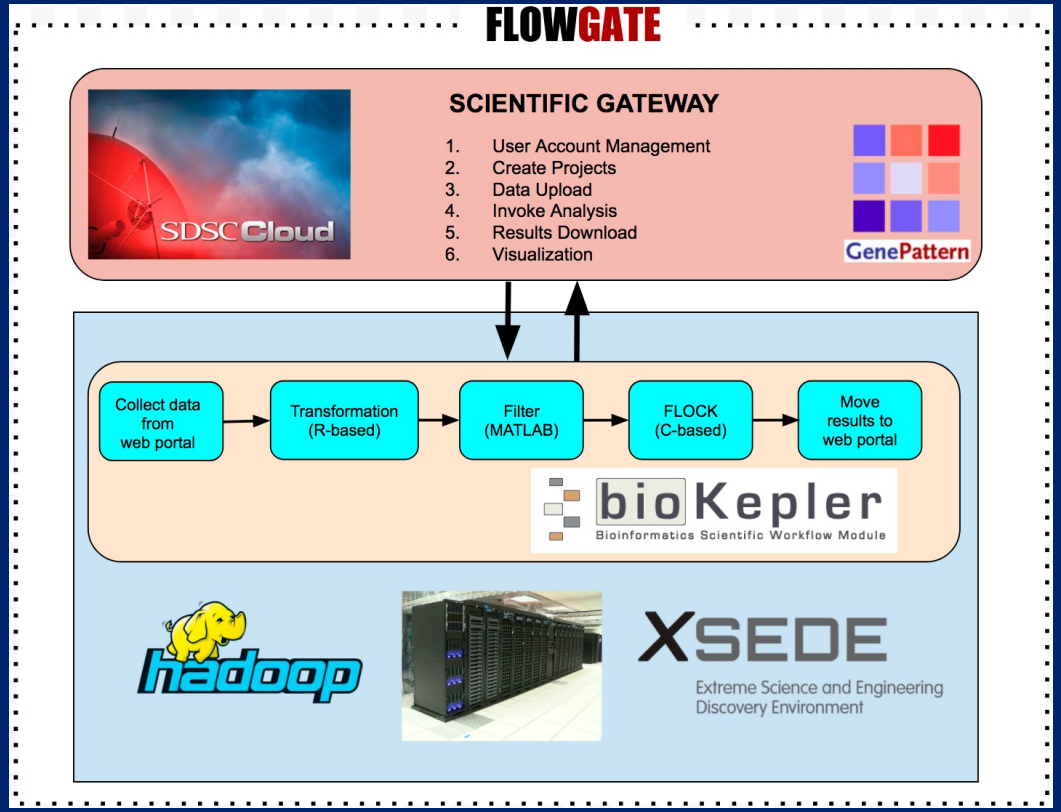


Simple workflows



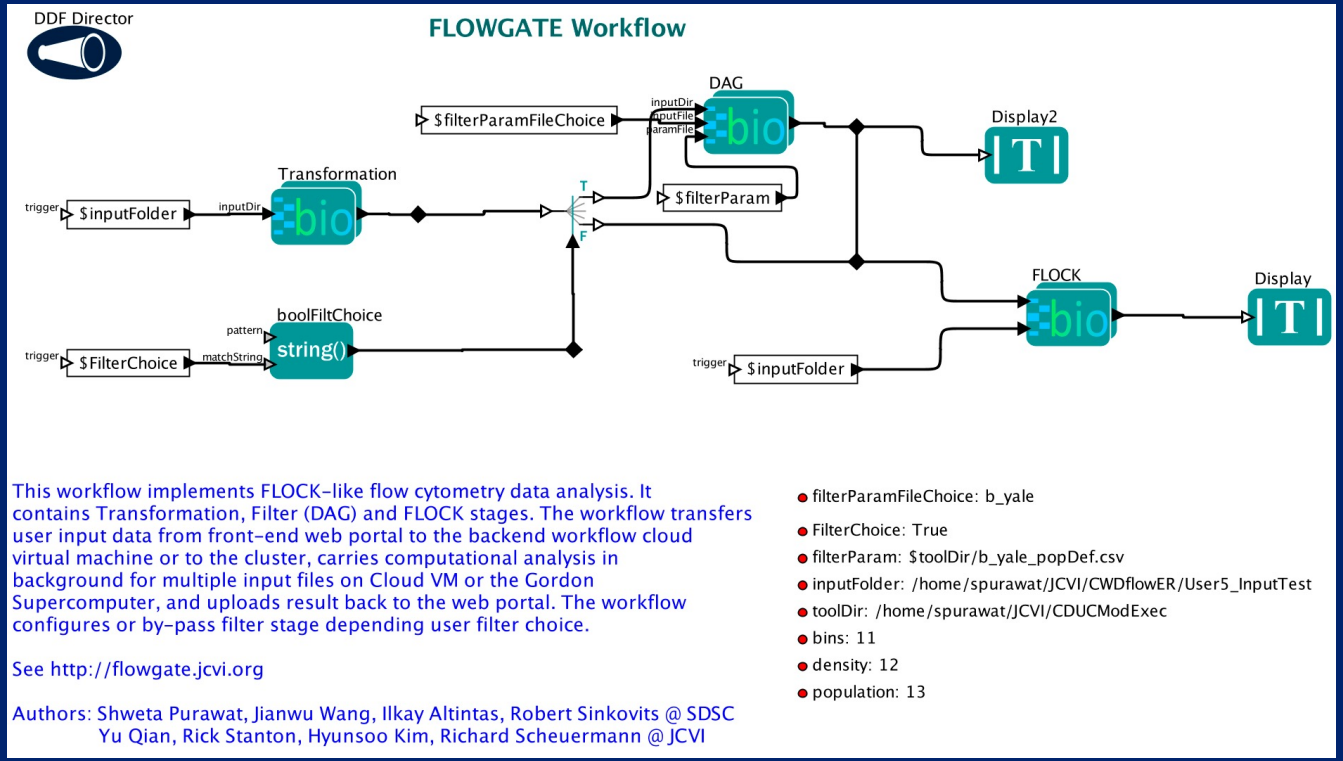


Workflows integrating multiple components



Scalable Web-Based Flow Cytometry Data Analysis

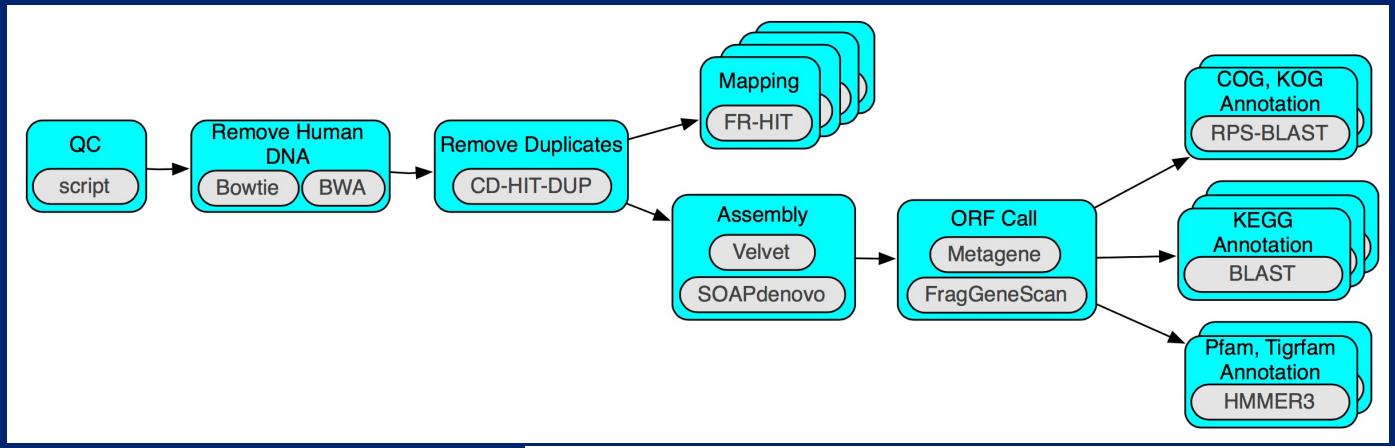
Workflows integrating multiple components

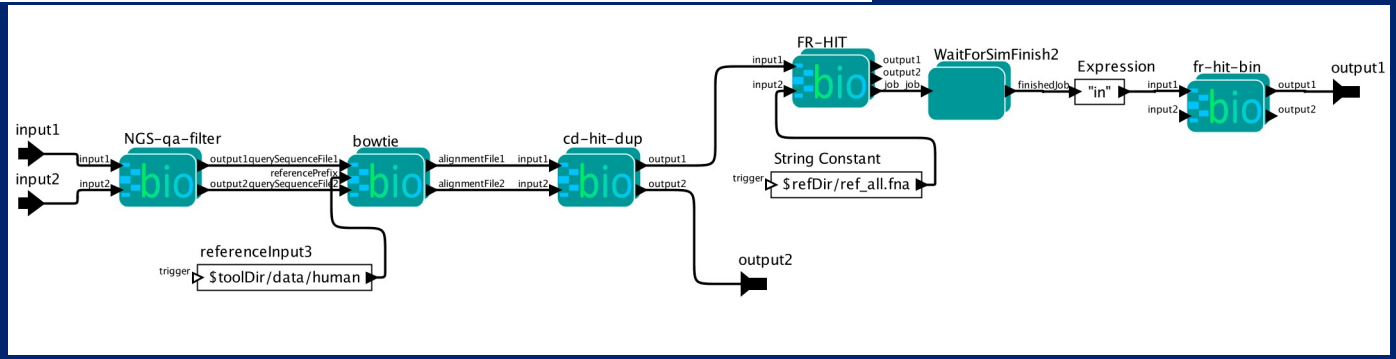


FLOWGATE Workflow:
Scalable Flow
Cytometry
Data Analysis

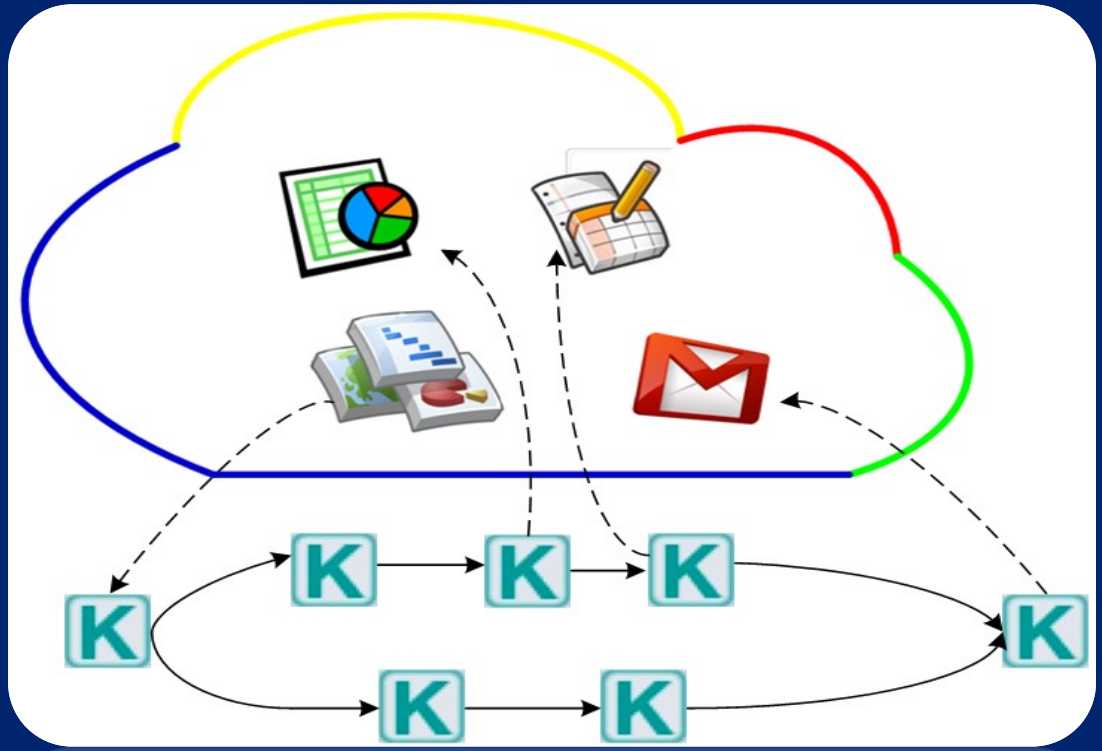
Compute and data intensive workflow

Microbiome Taxonomy and Gene Abundance Workflow (MTGA)



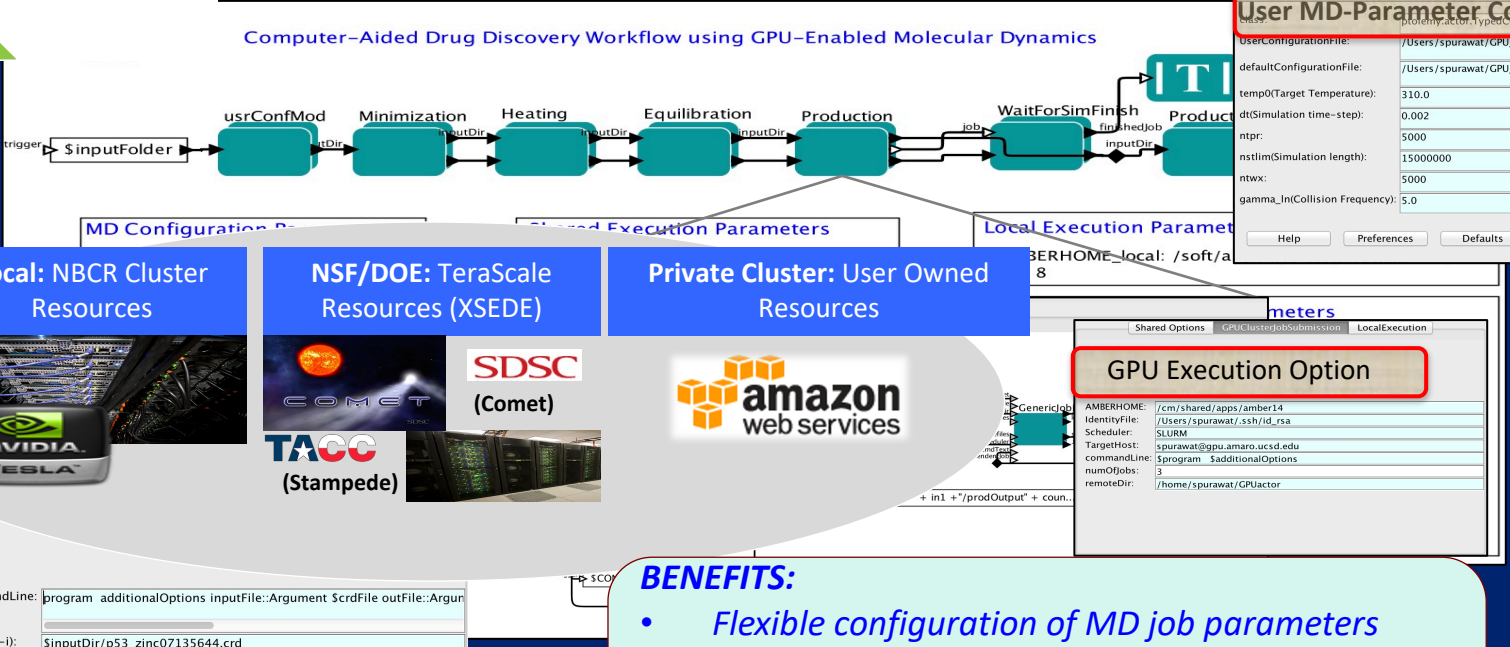


Workflows that Integrate cloud resources



COMPUTE THE CURE

AMBER GPU Molecular Dynamics



Local: NBCR Cluster Resources

NSF/DOE: TeraScale Resources (XSEDE)

SDSC (Comet)

TACC (Stampede)

Private Cluster: User Owned Resources

amazon web services

commandLine: program additionalOptions inputFile::Argument $crdFile outFile::Argun

crdFile (-i): $inputDir/p53_zinc07135644.crd
nxtrst (-r): $inputDir/md5.rst
outFile (-o): $inputDir/md5.out
outnc (-x): $inputDir/md5.nc
prerst (-c): $inputDir/md4.rst
top (-p): $inputDir/p53_zinc07135644.top

Local Execution Option

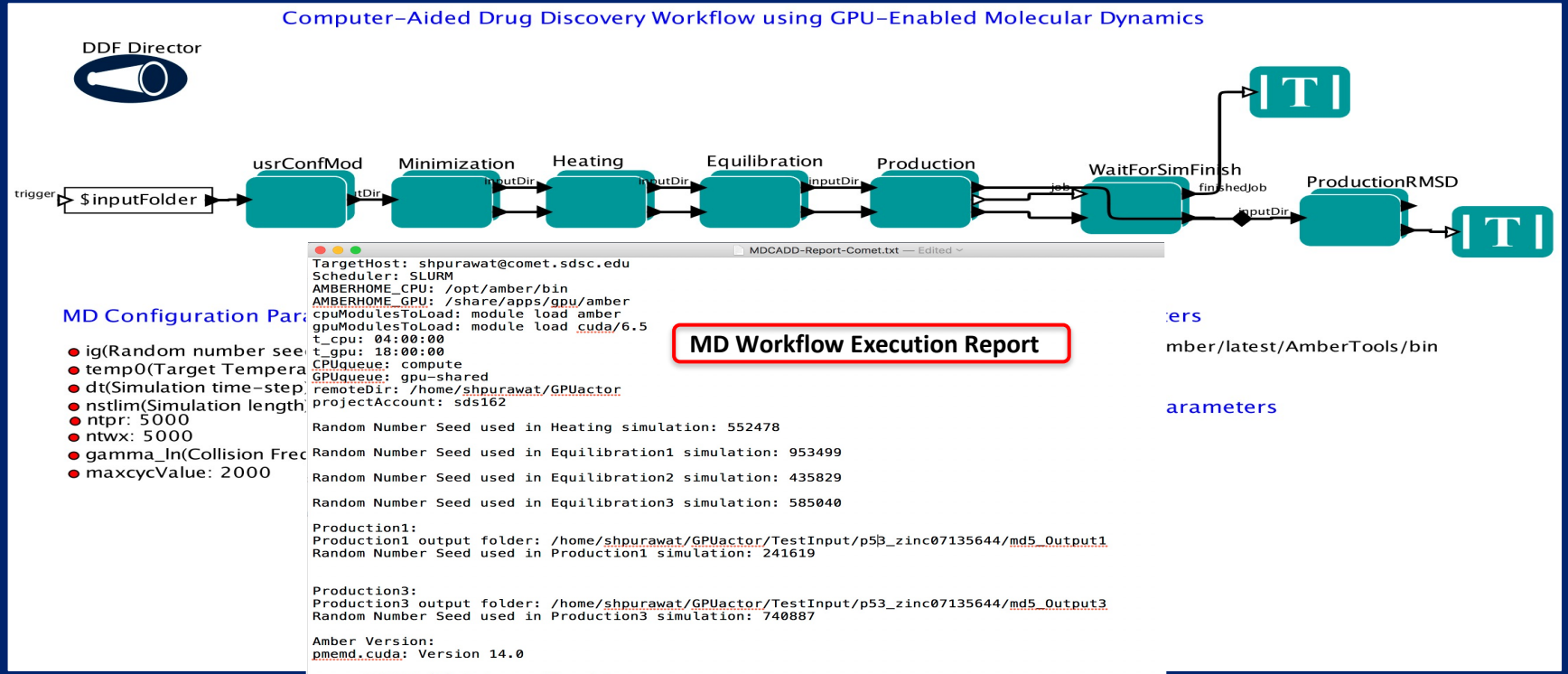
BENEFITS:

- *Flexible configuration of MD job parameters*
- *Scalability at compound level*
- *Computing platform portability*
- *Increased reuse*
- *Provenance*

Parametric execution of each step

Concurrent Data Analysis and Management

MDCADD WF – Workflow Execution Report



HydroFrame

Hydrologic Data Science Workflows and Provenance

Computationally intensive Hydrologic Simulations

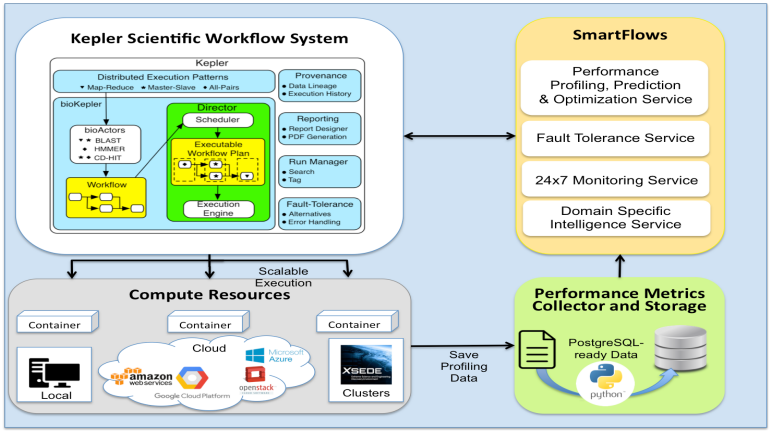
HydroFrame

- Simulation of GW + SW over the ConUS
- forecasting analysis
- climate change projections
- Study water and energy balance

Area of Interest

subset model inputs/outputs

- The Modelers – Integrated Hydrologic Models
- The Analyzers – Custom Analysis Tools
- The Domain Science Educators – Videos, Educational Tools



Quantum Foundry – Quantum Data Hub

Samples

Devices

Characterization

Growth and Material Parameters

Quantum Foundry

Collect

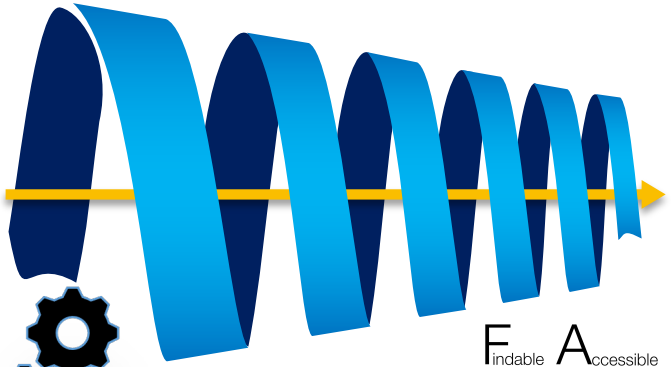
Curate

Manage

Catalog

Integrate

Analyze



Findable Accessible Interoperable Reusable

Scalable Data and Computing Cyberinfrastructure

Search

Query

Visualize

Analyze

Apply

Q-AMASE-i Centers

Create, Process, and Characterize materials for quantum information science

UC San Diego

Collect, curate and manage the foundry data

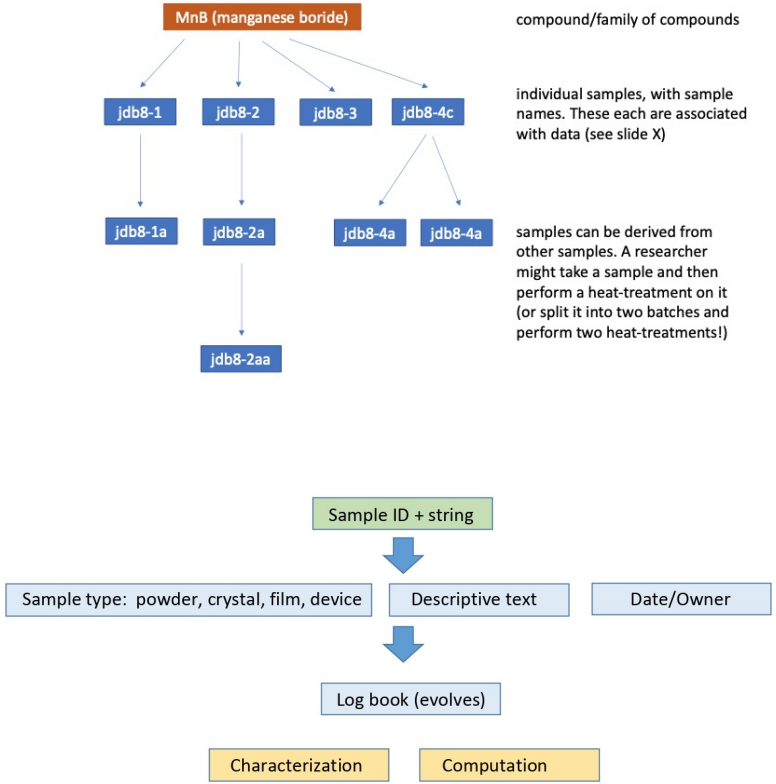
Amplifying the Value of Foundry Data Through Data Science!

Typical data associated with each sample

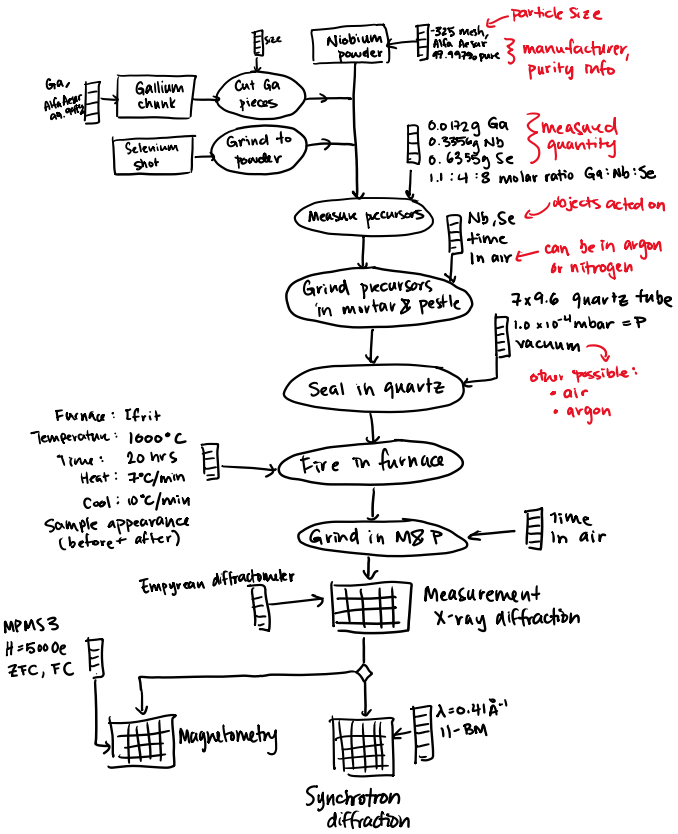
jdb8-1

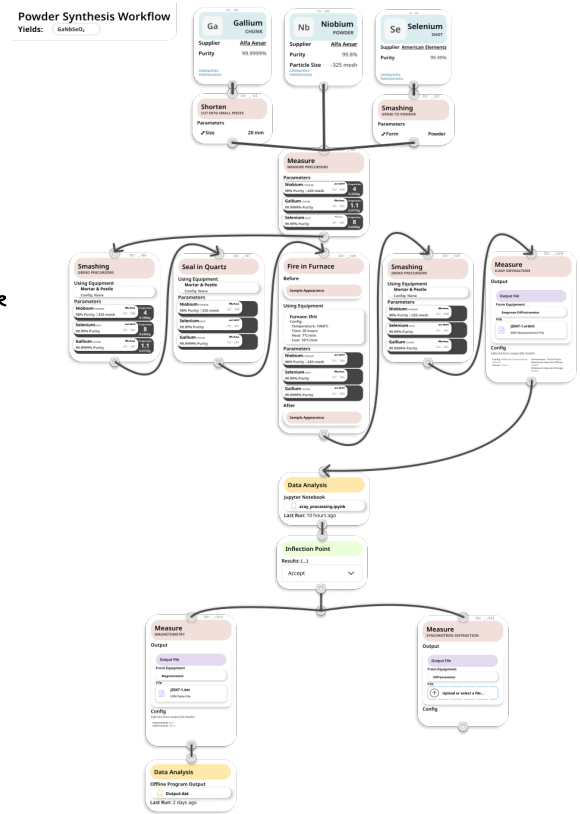
- lab notebook page(s), notes, etc.
- picture of the sample
- X-ray diffraction dataset (jdb8-1.xrdml)
  Analysis:
  - jdb8-1.inp (input file for a program we use to analyze)
  - jdb8-1.cif (file that contains the crystal structure)
  - jdb8-1_rietveld.pdf (figure showing the analysis)
- Magnetic measurements (jdb8-1_MT.dat)
  - jdb8-1_magnetic_data.ipynb (code used to analyze)
  - jdb8-1_MT.agr (qtgrace file used to make the figure)
  - jdb8-1.pdf (figure)
- etc., etc.

Basic data structure



Workflow for the Lab Notebook Process

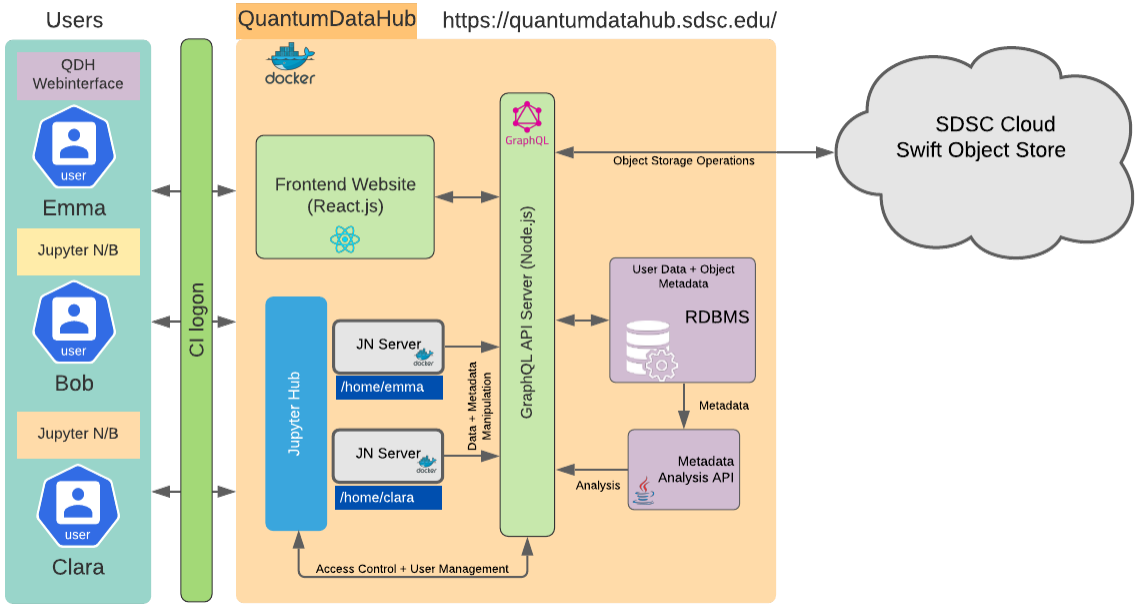




Represent the experiment process as DAG.

Quantum Data Hub Architecture

https://quantumdatahub.sdsc.edu/



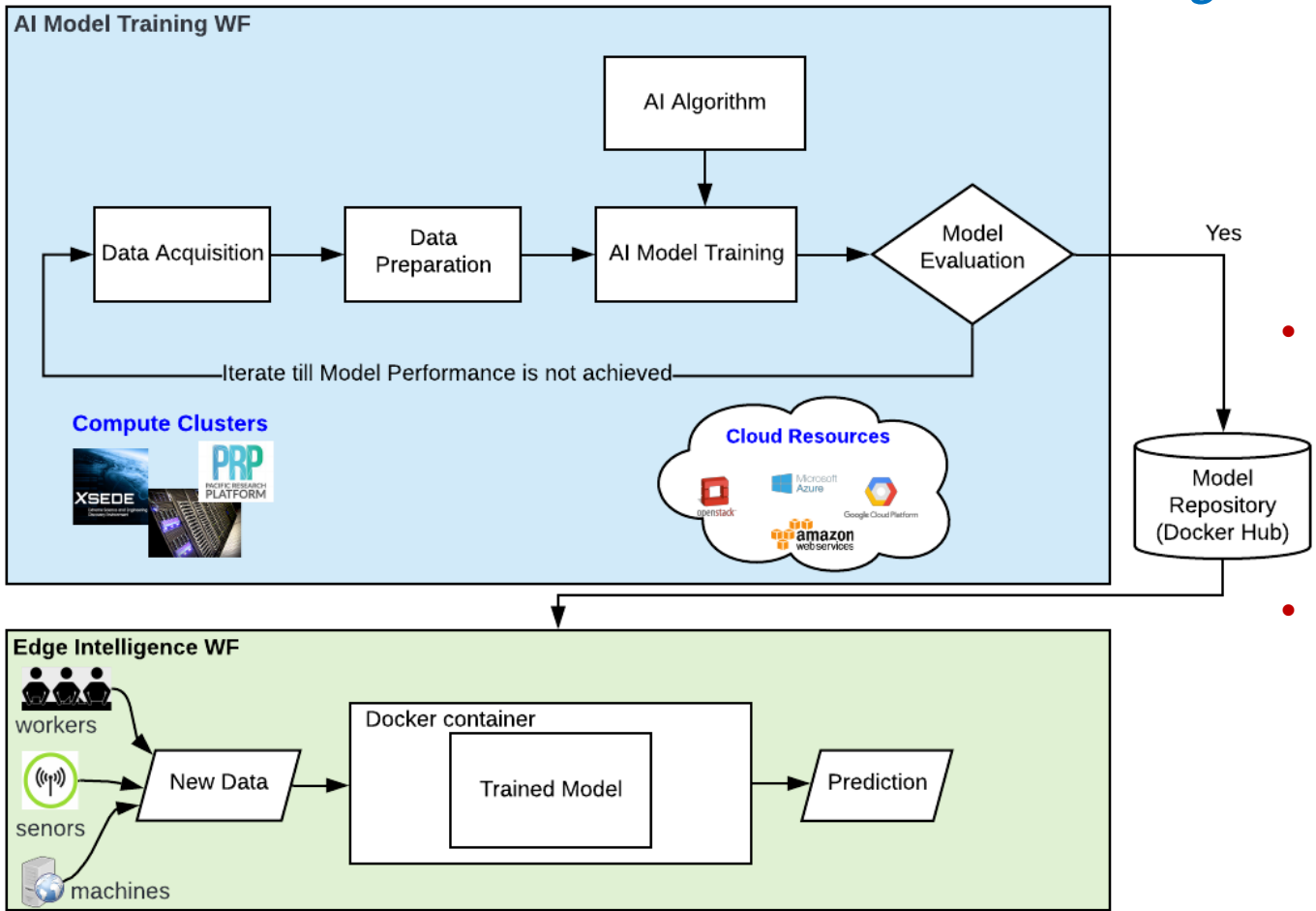
User Interface - React, Node.js

Quantum Data Hub (QDH) API

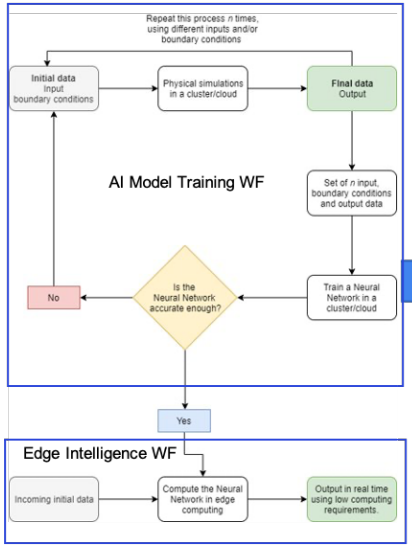
JupyterHub - Simple and Intuitive Jupyter Notebook Exposing QDH APIs

Database – User management, Activity log and Quantum Metadata

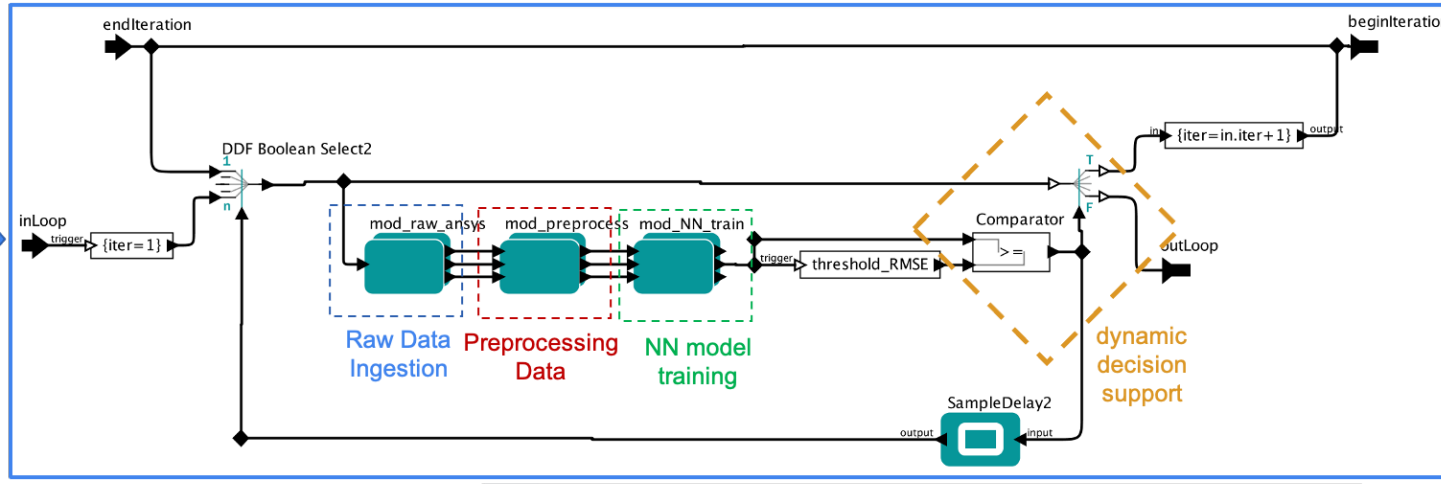
Smart Manufacturing - AI Workflow Reference Architecture for Advance Manufacturing



- Training the model on historical data - Computationally intensive task
- Deploying the trained model in manufacturing floor for real time prediction.

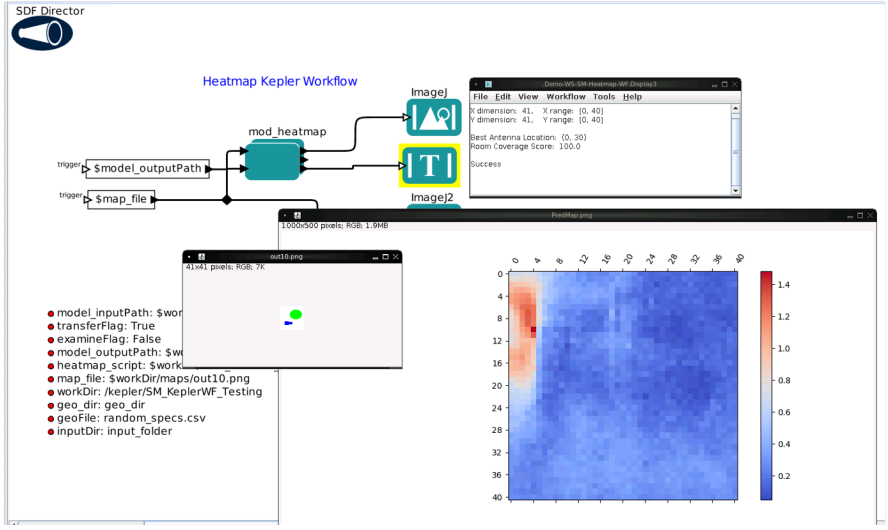


Conceptual flow diagram

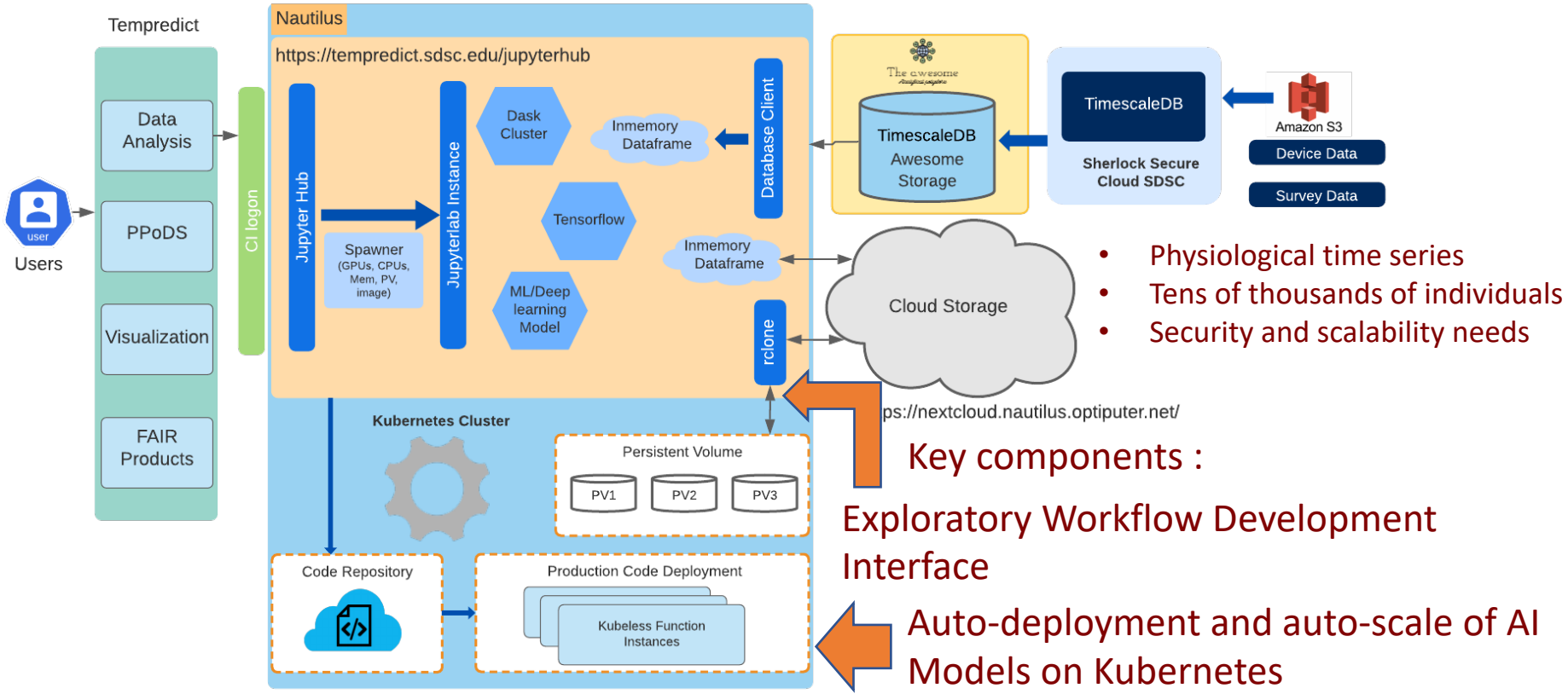


The CNN Model Training Workflow for WiFi Received Signal Strength Intensity (RSSI)

The Edge Intelligence Workflow for WiFi Received Signal StrengthIntensity (RSSI) deployment



A Big Data System for Scalable Exploration and Monitoring of Personalized MultimodalData for COVID-19



- Physiological time series
- Tens of thousands of individuals
- Security and scalability needs

Key components :

Exploratory Workflow Development Interface

Auto-deployment and auto-scale of AI Models on Kubernetes

Using Workflows and Cyberinfrastructure for Wildfire Resilience

- A Scalable Data-Driven Monitoring and Dynamic Prediction Approach -

Real-time sensors

Weather forecast

Landscape data

Fire perimeter

Big Data

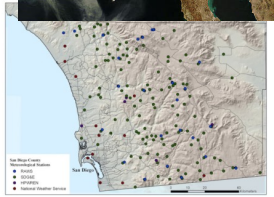
WIFIRE

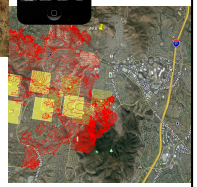
Monitoring

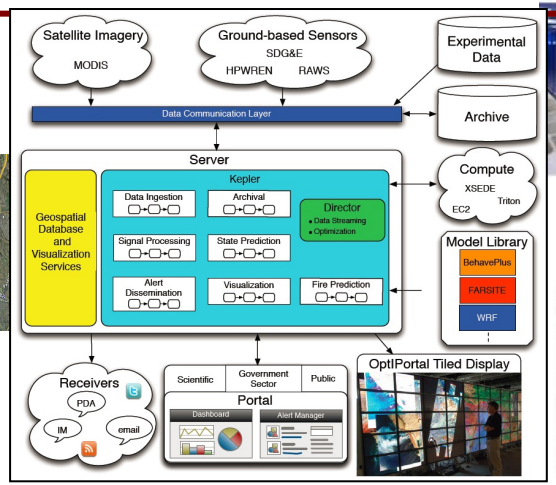
Visualization

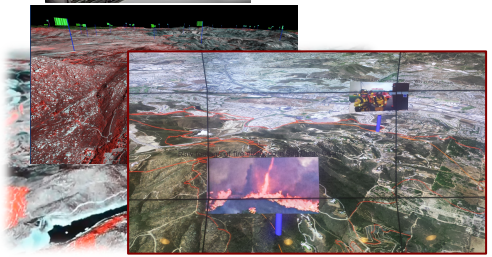
Fire Mapping

- predicts in real time the spread of wild fires
- actionable insights support firefighting on the ground.









Multimodal wild-fire related data

Section Summary:

- Workflows can be compute intensive and/or data intensive
- Integrate multiple components using workflow engines, and achieve true synergy of collaborative efforts
- Run your workflow on multiple computing platforms such as GPU clusters, HPC clusters, Cloud etc.
- Workflow Reports provide a detailed summary of jobs
- Kepler allows you to scale your workflows without any impact on performance.
- Kepler provenance module records execution history for reproducibility

Part 3: Introduction to Kepler

- Define what is Kepler
- Identify terminologies used in Kepler
- Identify commonly used features of Kepler

Kepler is a Scientific Workflow System

- A cross-project collaboration
  … initiated August 2003
- Kepler 2.5 - Current stable version
- Frequent module release updates
- Builds upon the open-source Ptolemy II framework

www.kepler-project.org

Ptolemy II: A laboratory for investigating design

KEPLER = "Ptolemy II + X" for Scientific Workflows

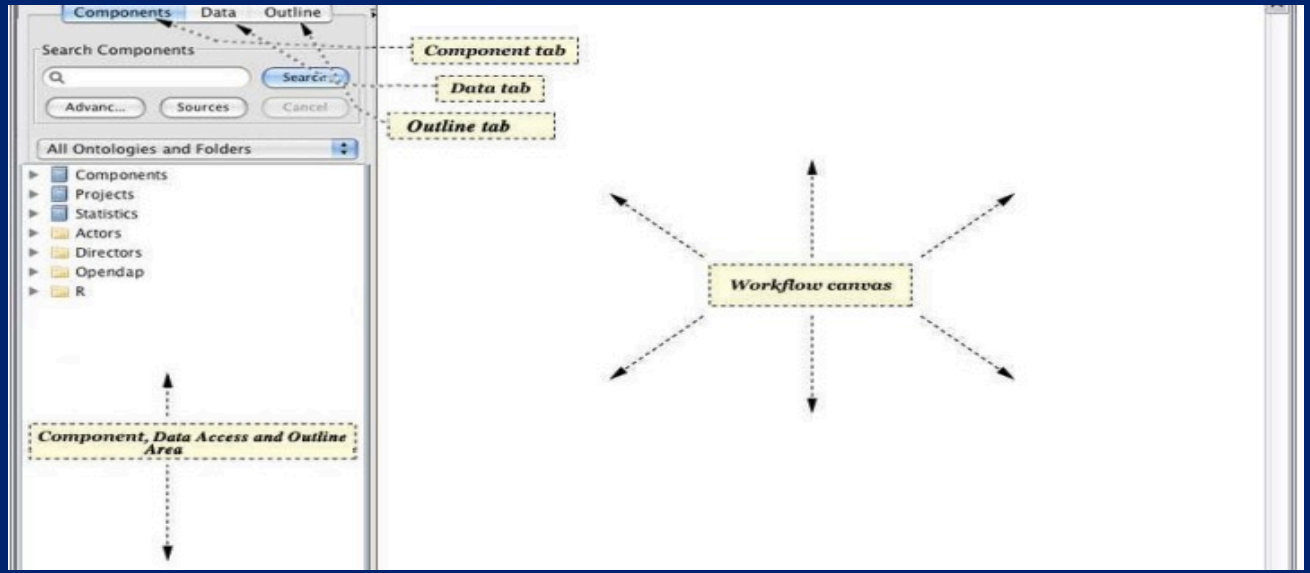
Graphical Workflow Systems

-Toolboxes with Many Tools-

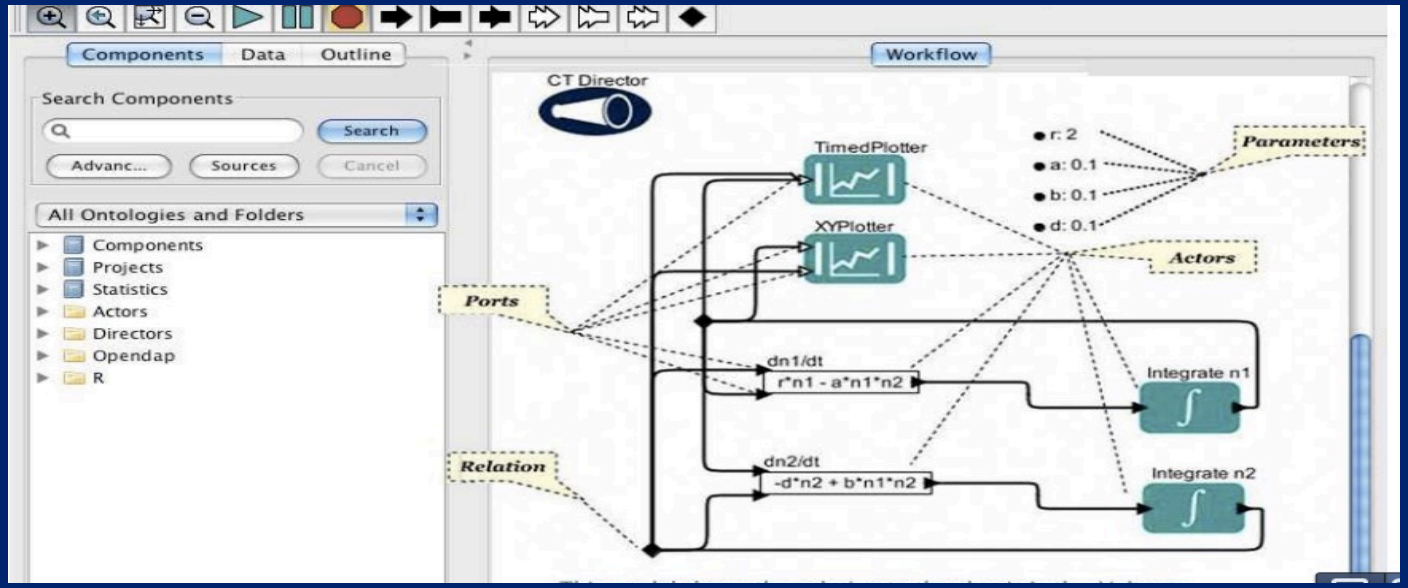
- Data
  - Search, database access, IO operations, streaming data in real-time…
- Compute
  - Data-parallel patterns, external execution, …
- Network operations
- Provenance and fault tolerance

Need expertise to identify which tool to use when and how!
Require computation models to schedule and optimize execution!

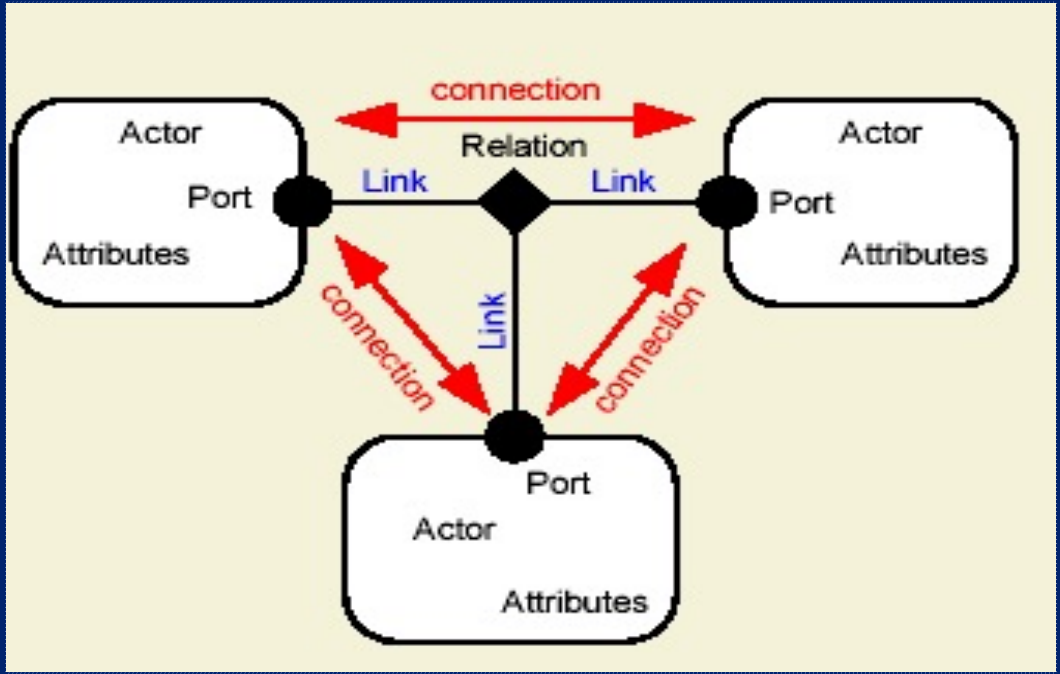
Kepler's UI


Components
Data
Outline
Search Components
Search
Advanc...
Sources
Cancel
All Ontologies and Folders
Components
Projects
Statistics
Actors
Directors
Opendap
R
Component tab
Data tab
Outline tab
Workflow canvas
Component, Data Access and Outline Area

Basic components and terminologies

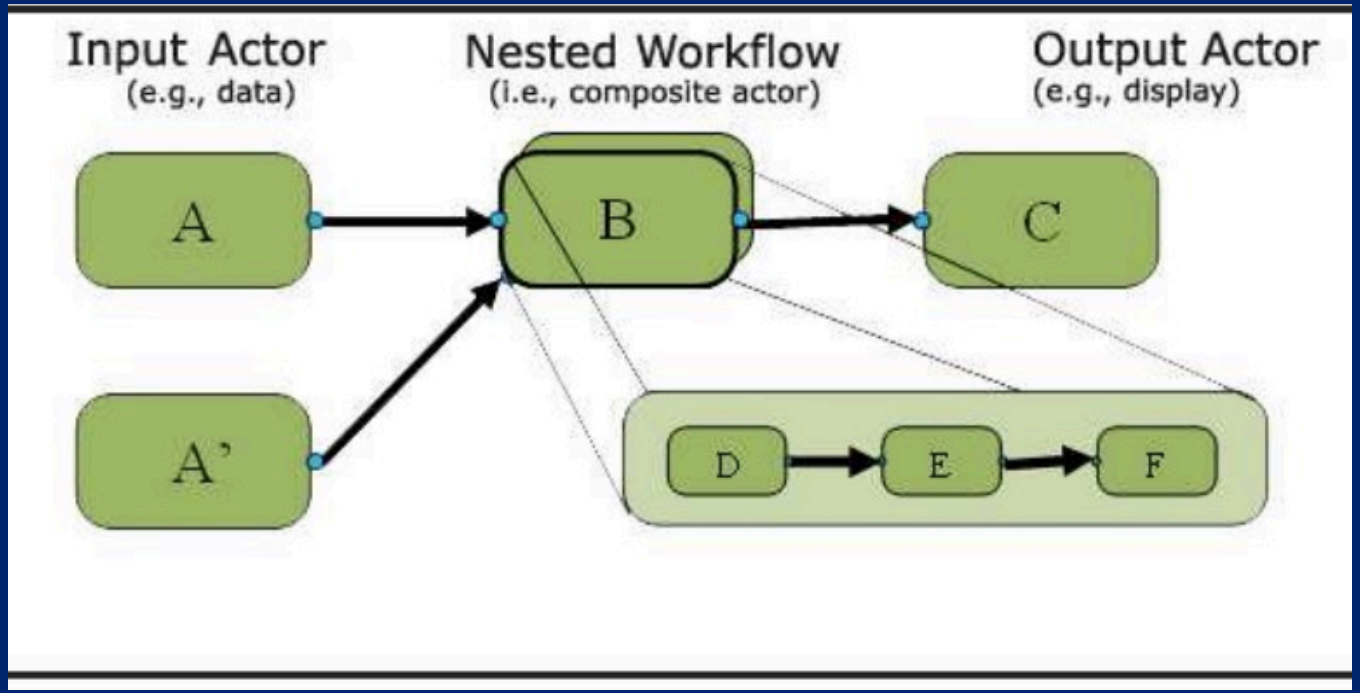


Actors are the Processing Components



Actor-Oriented Design

Kepler Actor Types

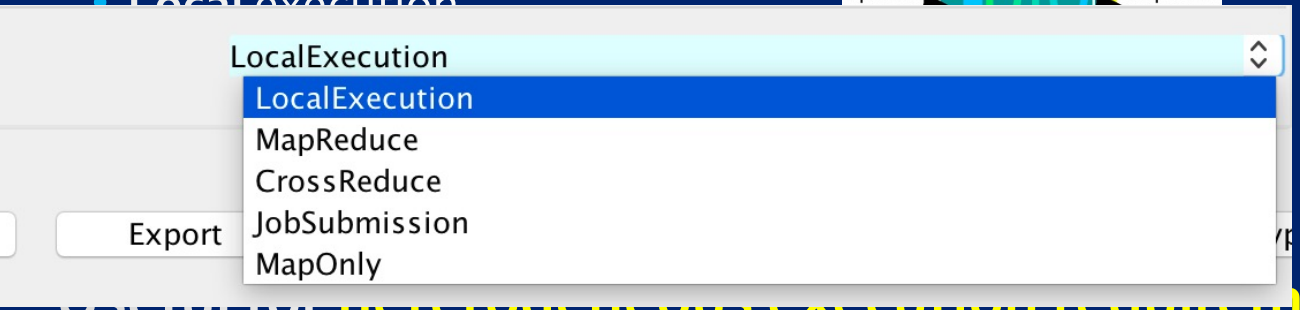


Some actors in place for…

- Command Line wrapper tools (**local execution**, **ssh**, **scp**, ftp, etc.)
- Generic Web Service Clients for SOAP and **REST**
- A suite of **cloud computing** actors for VM instantiation and management
- Job management actors for **HPC, GPU, SGE** and other commodity clusters
- Customizable **RDBMS** query and update
- Distributed data parallel patterns, e.g., Map, Reduce, Cross
- **Hadoop**, Stratosphere, and **Spark** integration
- iRODS support
- Native **R** and **Matlab** support
- Communication with external workflow engines, e.g., **KNIME**
- Communication with sensor data loggers through actors and services
- Imaging, Gridding, Vis Support
- Textual and Graphical Output
- Integration with **Jython, JavaScript, Java, JRuby**
- …more generic and domain-oriented actors…

Workflow Execution across Multiple Environments

- Execution Choice Actor: Multiple types of executions within one workflow



- Useful for heterogeneous execution requirements

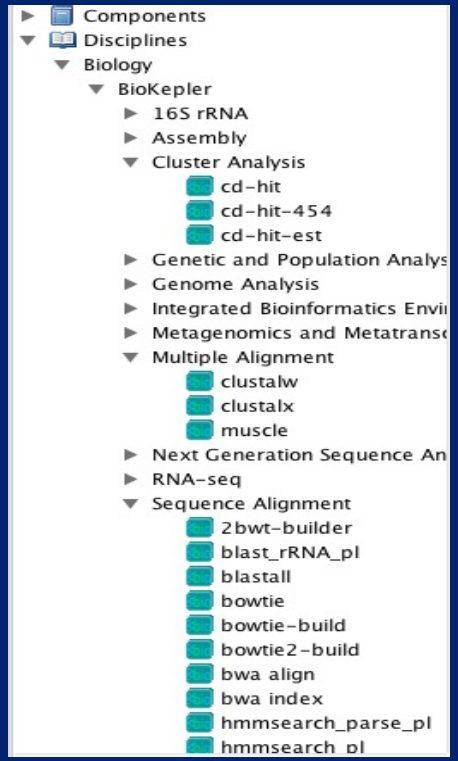
bioKepler

A Comprehensive Bioinformatics Scientific Workflow Module for Distributed Analysis Of Large-Scale Biological Data

https://www.biokepler.org/install-biokepler-1.2

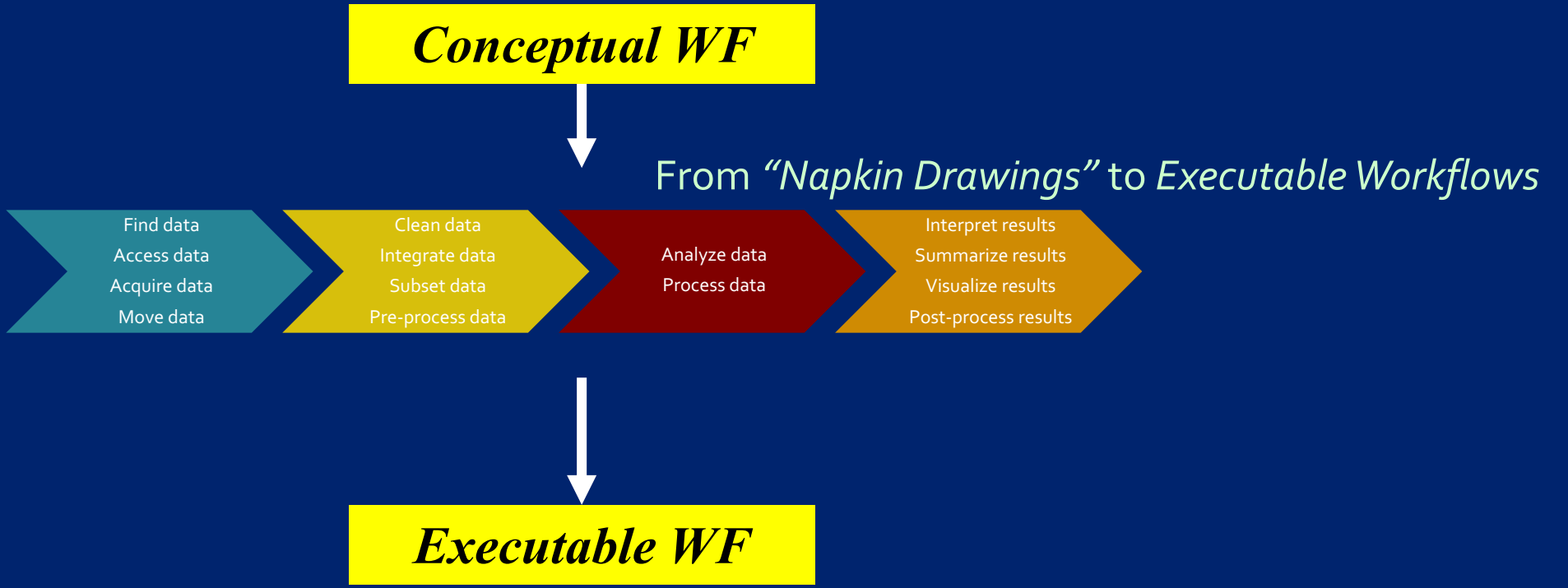
bioKepler 1.2 released in 2015

bioActors



- Alignment: BLAST, BLAT
- Profile-Sequence Alignment: PSI-BLAST
- Hidden Markov Model: HMMER
- Mapping: Bowtie, BWA, Samtools
- Multiple Alignment: ClustalW, Muscle
- Clustering: CD-HIT, Blastclust
- Gene Prediction: Glimmer, Genescan, Fraggenescan
- tRNA prediction: tRNA-scan, Meta-RNA
- Phylogeny: FastTree, RAxML

The Big Picture is Supporting the Scientist



Part4: Kepler Demo!

Downloading and Installing Kepler

- **Java 1.8 or later is a prerequisite to run Kepler 2.5**
- **Download Kepler installer for Windows, Mac, or Linux-based platforms.**

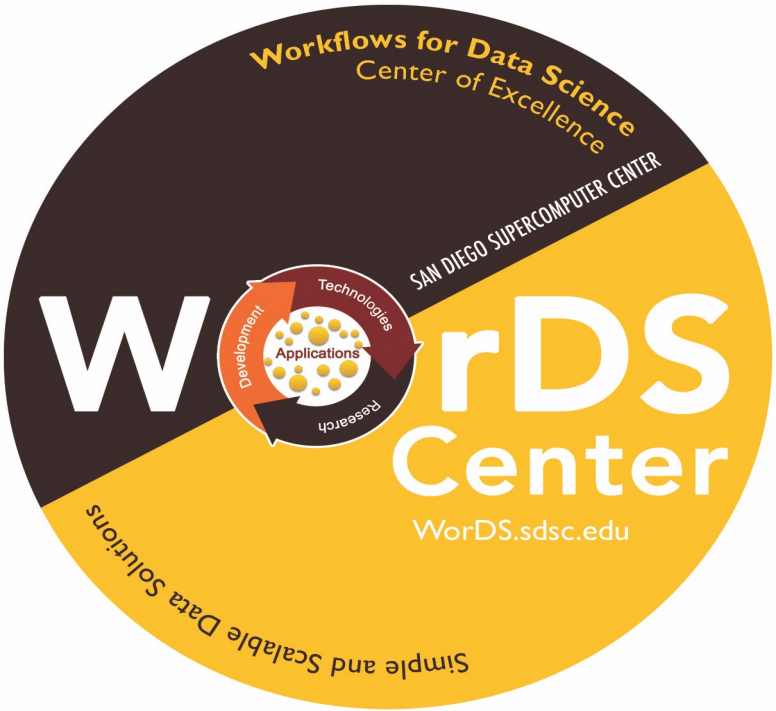
  https://kepler-project.org/users/downloads.html

- **Add-on Modules can be added to Kepler to enhance its functionality. Example: bioKepler, Provenance, Reporting.**

Useful Links

- https://kepler-project.org/users/downloads.html
- https://kepler-project.org/users/documentation.html
- https://words.sdsc.edu/sites/default/files/biokepler/userguide.html
- http://words.sdsc.edu
- https://github.com/words-sdsc/Jupyter_Kepler_Integration
- https://words.sdsc.edu/publications

Questions?



Ilkay Altintas, Ph.D.
ialtintas@ucsd.edu

Shweta Purawat
shpurawat@ucsd.edu

Thank you!