SDSC Summer Institute 2022

Session 4.2b: A Short Introduction to Data Science and its Applications

Instructor: Shweta Purawat, Subhasis Dasgupta

Date: 08/03/2022

Location: Breakout Room

Day 1 & Day 2 !

- **Expanse HPC System**
- **To interact with Expanse HPC system:**
  - Batch and Interactive computing - Launching and managing jobs
  - Managing data on the file system
- **Parallel Computing Concepts, High Throughput Computing, File processing, Data Management**

- Data Science & Applications
- Data Science Workflows
- Kepler, Data Provenance

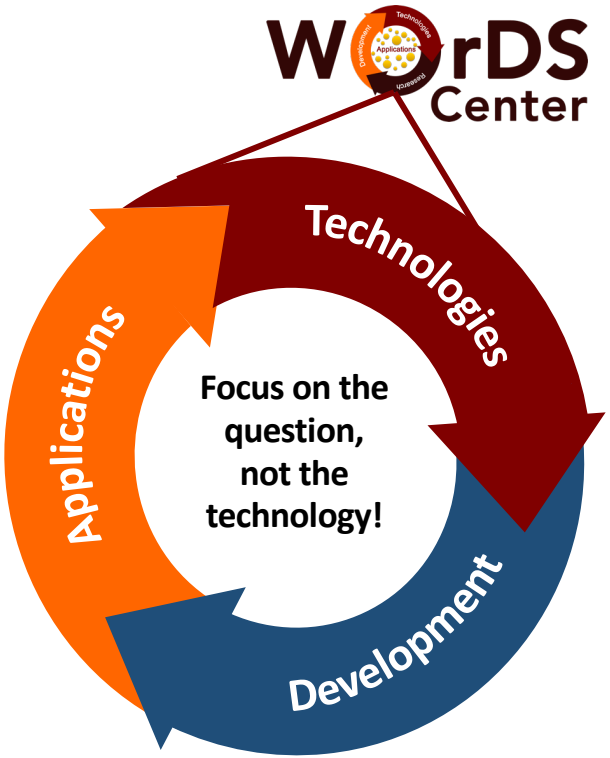
Shweta Purawat

Technical Project Manager & Computational and Data Researcher, SDSC

- Managing Mixed Modeled Data
- Overview of Polystore
- The Awesome Polystore
- Building Knowledge Graph on the top of Polystore

Subhasis Dasgupta

Computational and Data Researcher, SDSC



Workflows for Data Science Center of Excellence at SDSC

http://WorDS.sdsc.edu

Mission:
Methodology and tool development
to enable collaborative workflow-driven data science
and create solution architectures
on top of big data and advanced computing platforms.

Common Theme…

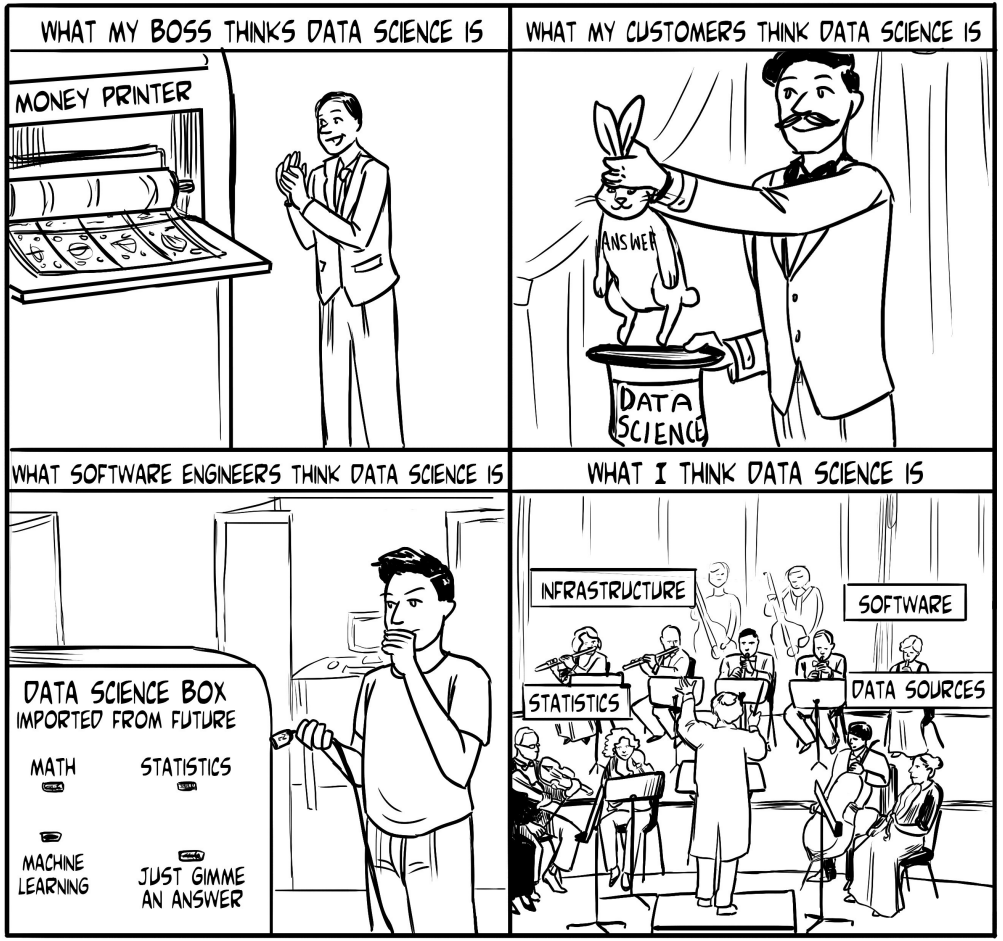
"Big" Data, Computational Science, Data Science, Cyberinfrastructure, and Their Applications

Part1: Introduction to Data Science

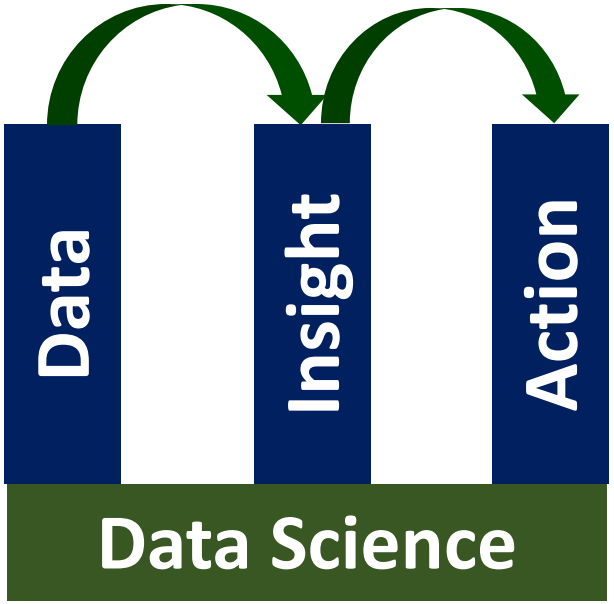
Part2: Data Science Workflows and Kepler

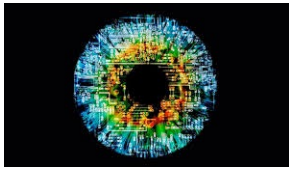
Part3: Data Science Applications

So what is Data Science?

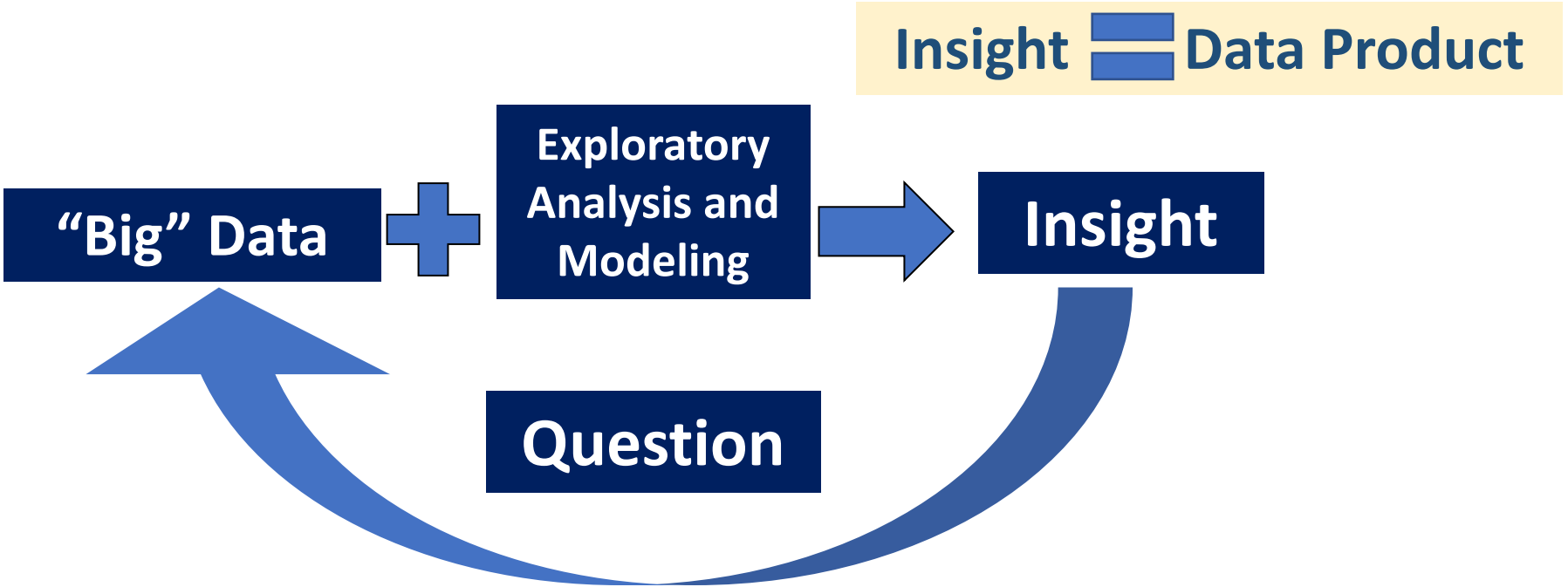


Ultimate Goal of Data Science





How does successful data science happen?

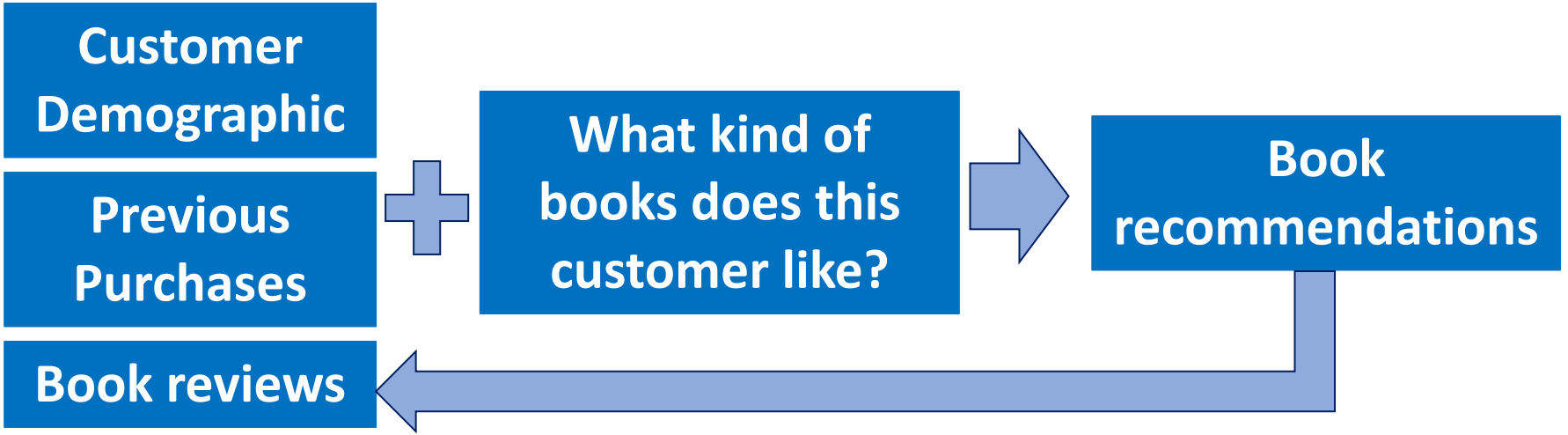


Data Products

"… a product that facilitates an end goal through the use of data"

-- DJ Patil, Former U.S. Chief Data Scientist
in Data Jujitsu

Book Recommendations



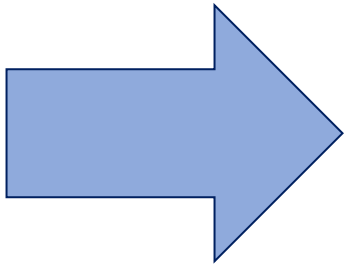
amazon.com

Find Potential Audience for a Book

Model of customer's book preferences

+

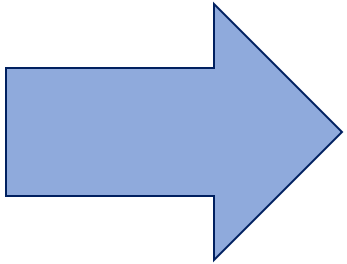
New book information



Who is likely to like this book?

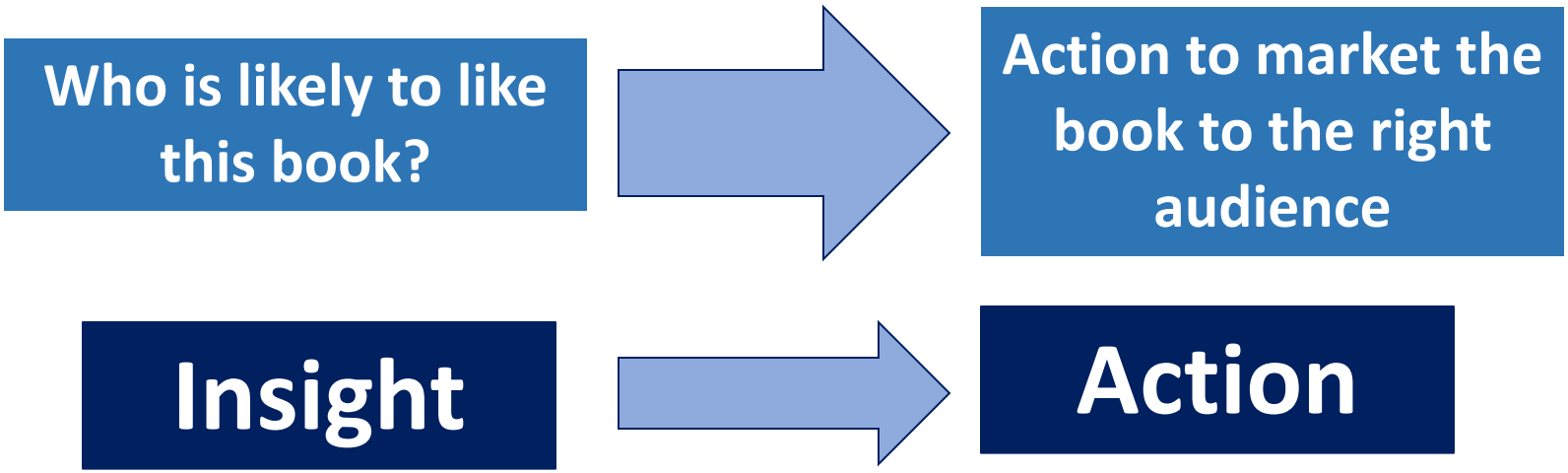
Market a New Book

Who is likely to like this book?

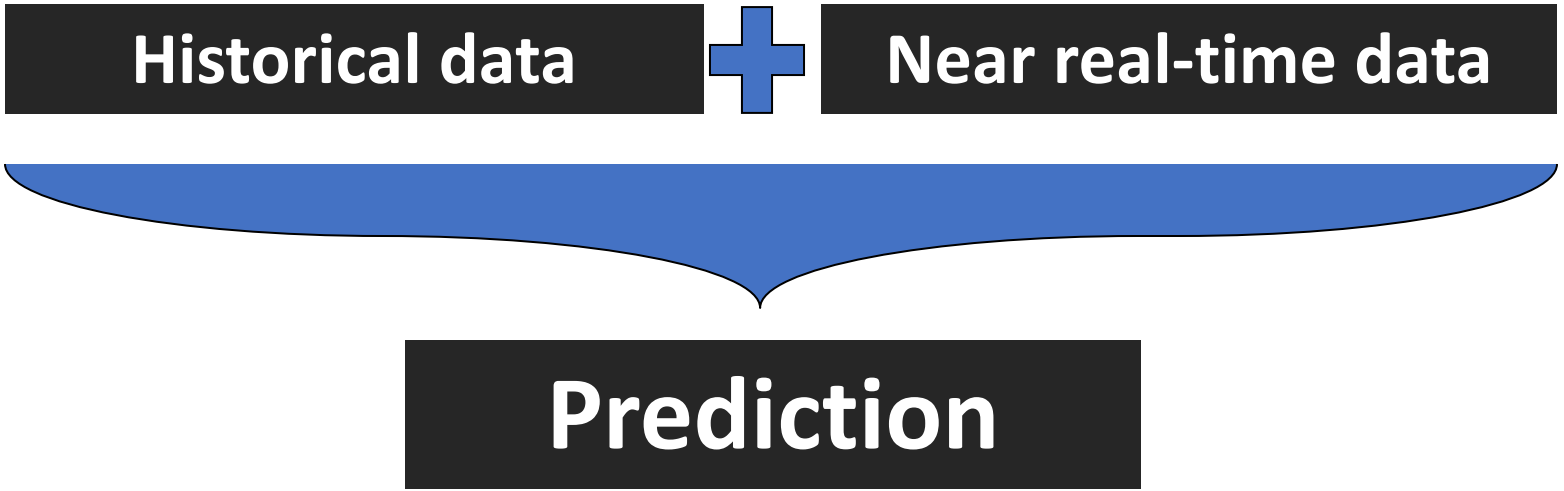


Action to market the book to the right audience

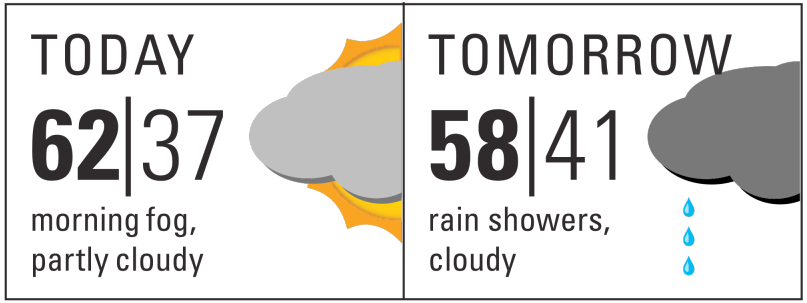
Market a New Book



Actionable Information



Prediction



Action

Systems and models

that help us to

understand data

in order to

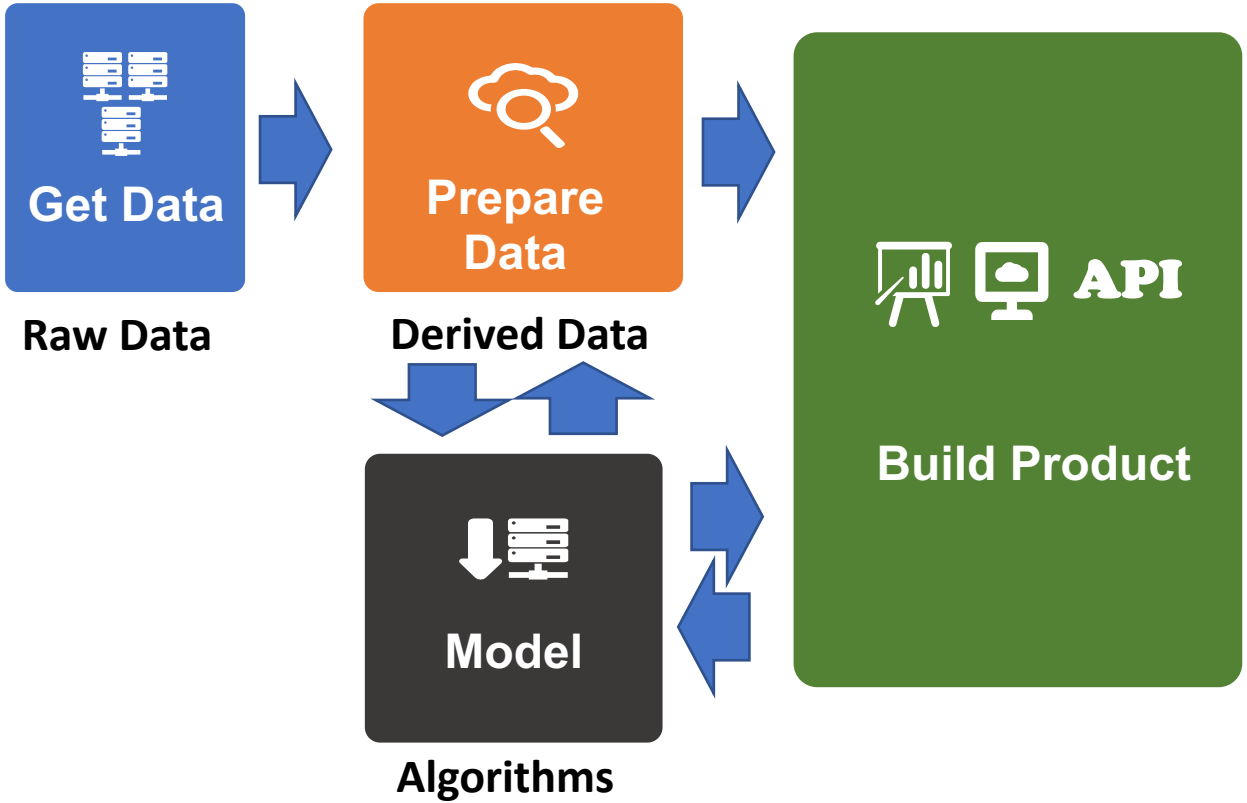
gain insights and

make predictions

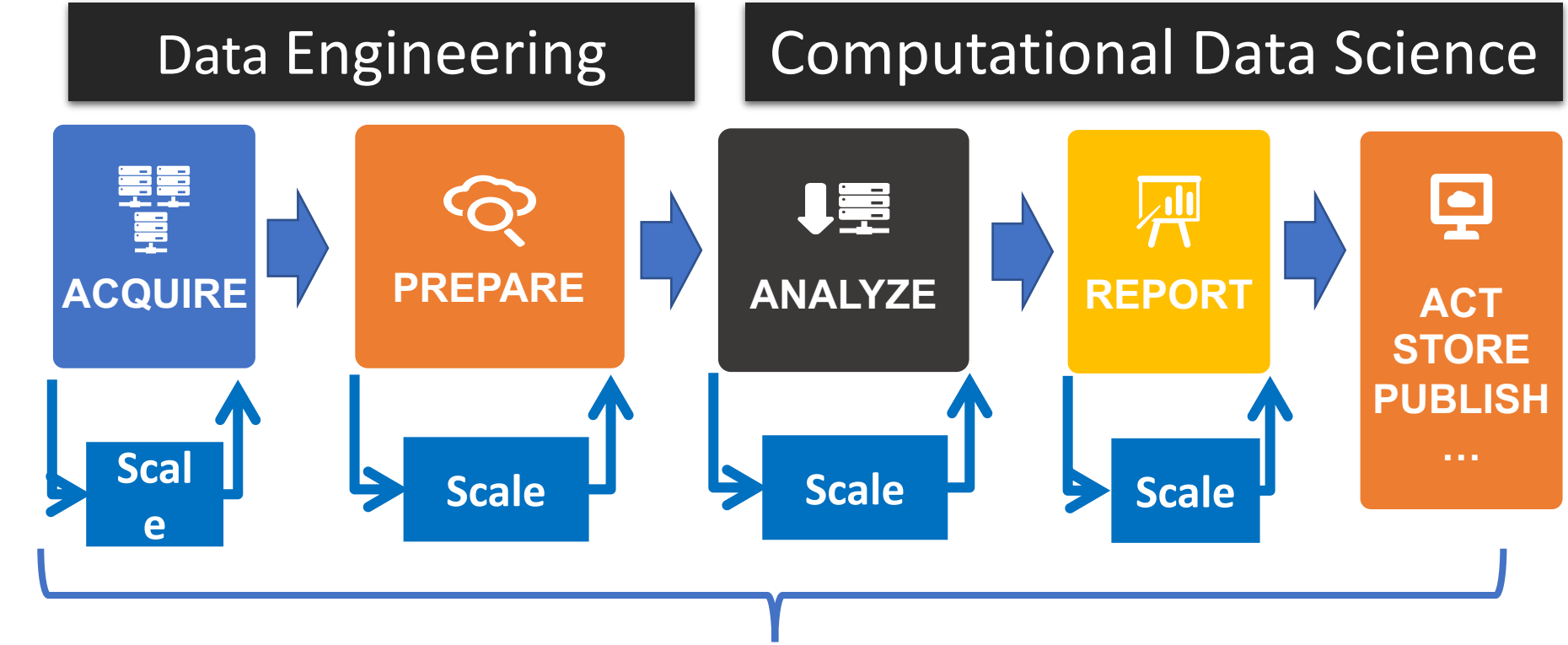
leading to action for impact.

Data Science is “IMPACT” Science!

Going from raw data to a model using data science…



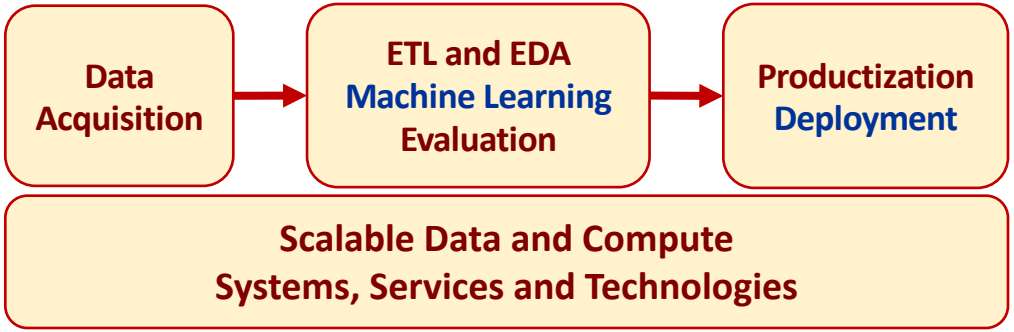
- Visual Dashboards
- Web Interfaces
- Programming Interfaces
- Robotics platforms



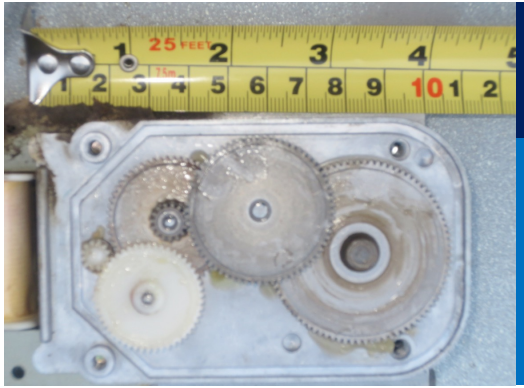
Continuous Iteration, Integration, Programmability, Measurement and Scalability

Scalable Data Science Process

together with



CRISP-DM: Cross-industry standard process for data mining



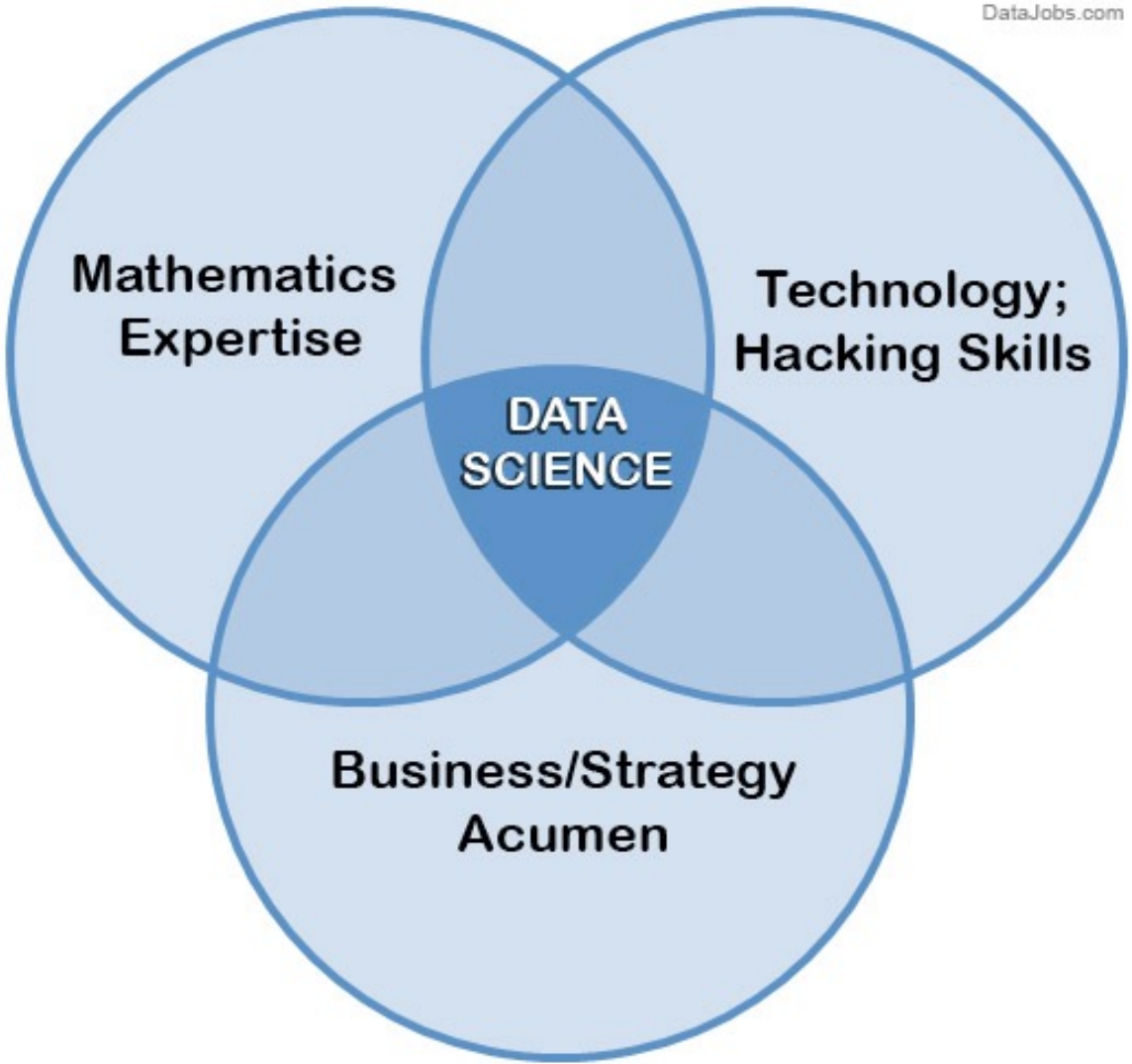
Data Science

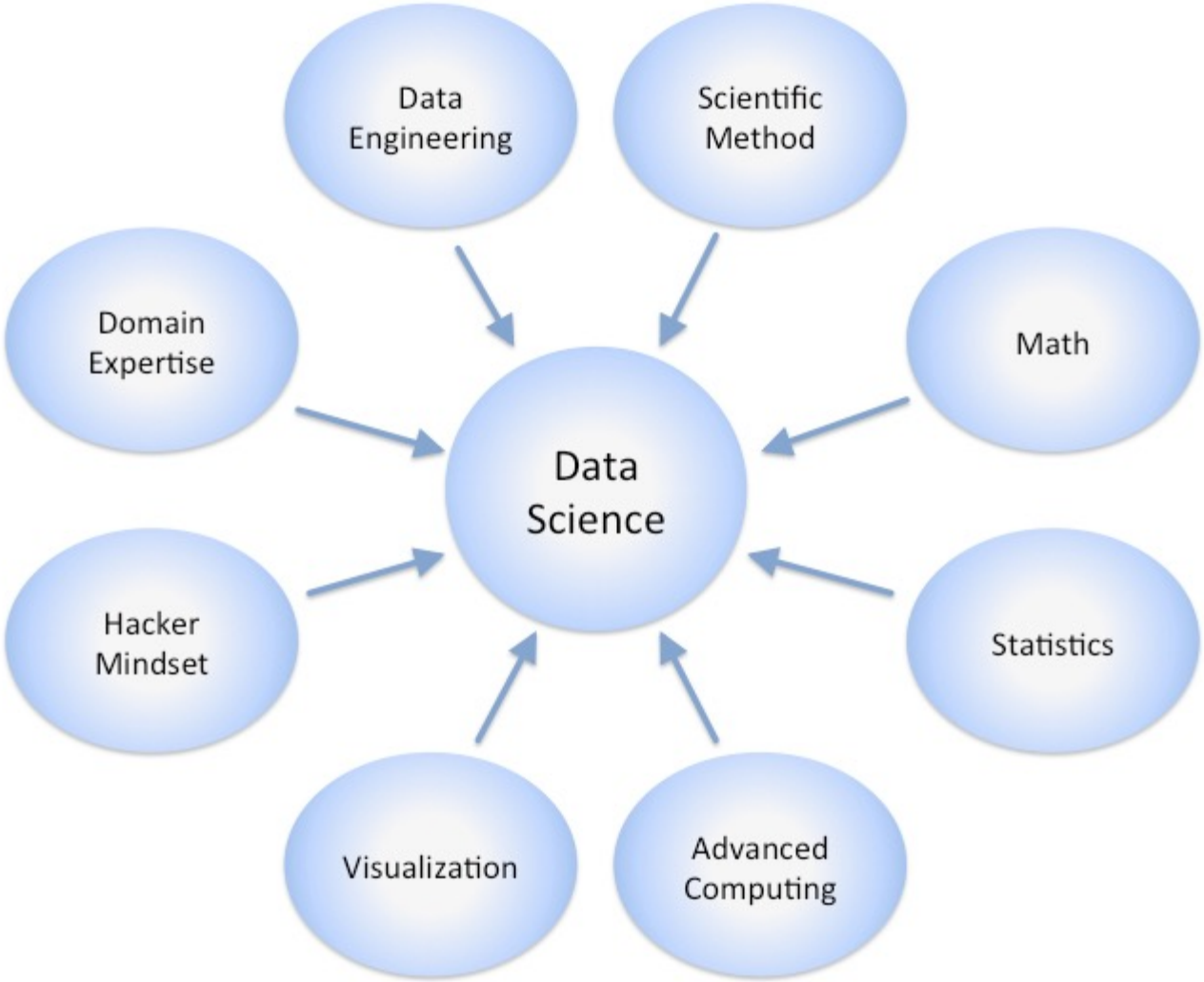
The sum is bigger than the parts!

Big Data → **Actionable Insight**

Modern Data Science Skills

- Python Programming
- Statistical Analysis
- Machine Learning
- Scalable Big Data Analysis


DataJobs.com
Mathematics Expertise
Technology; Hacking Skills
DATA SCIENCE
Business/Strategy Acumen


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

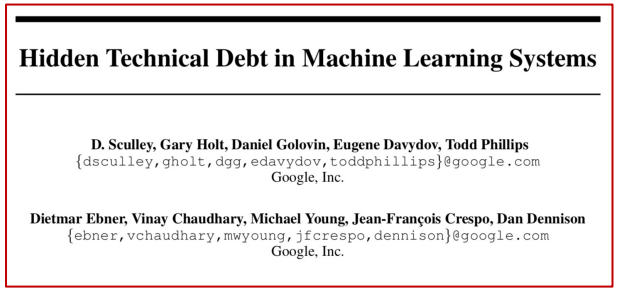
COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

Are data scientists unicorns?



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

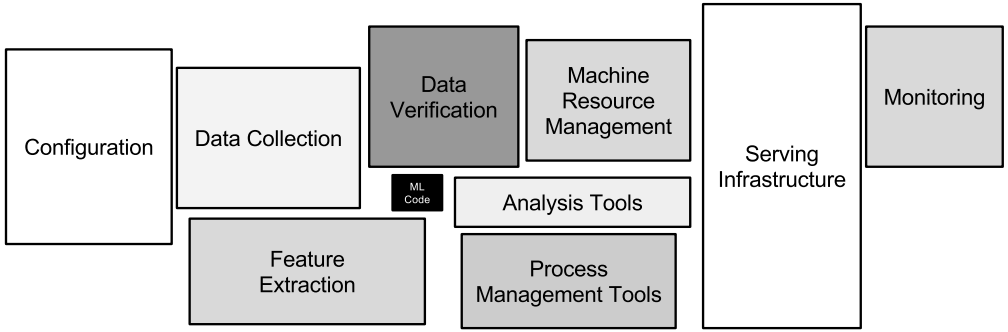


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

The problem is: "solutions are complicated"!

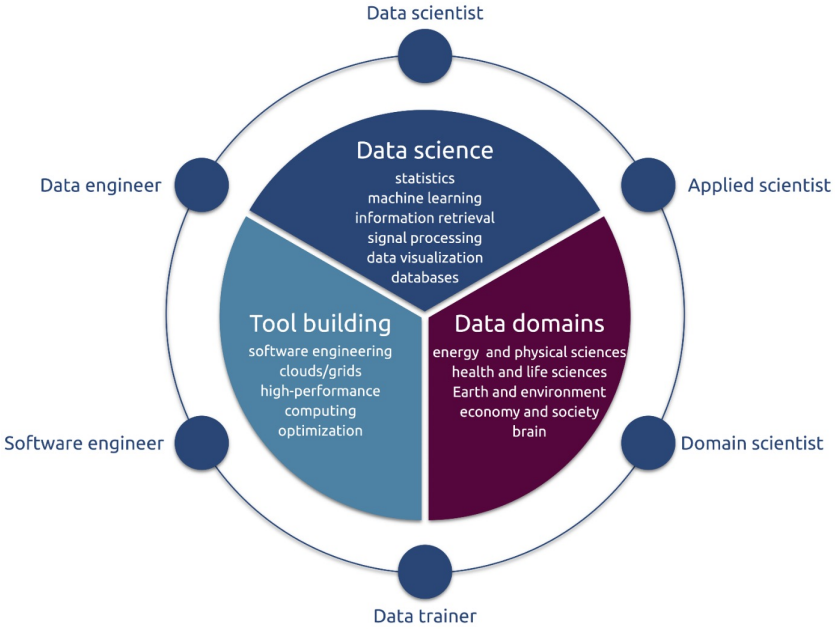
- Heterogenous systems and infrastructure
- Data management
- Machine learning, statistics and analytical methods
- Scalable process management
- Dynamic coordination and resource optimization
- Skilled interdisciplinary team
- Collaborative culture and communication tools

Data science is team sport!

Data Science is “WE” Science!

Data Science Team

- Data engineer
- Data analyst
- Methods expert
- Scalability and operations expert
- Business manager
- Business analyst
- Scientist
- Visualization and dashboard developer
- Solution architect
- Story teller/coordinator
- Project manager



The data science ecosystem: activities and actors
https://medium.com/@balazskegl/the-data-science-ecosystem-678459ba6013

Expertise and skills often overlap, but nobody has it all!

Basic Steps in a Data Science Project

ACQUIRE
- Import raw dataset into your analytics platform

PREPARE
- Explore & Visualize
- Perform Data Cleaning

ANALYZE
- Feature Selection
- Model Selection
- Analyze the results

REPORT
- Present your findings

ACT
- Use them

Data Collection from Diverse Sources

- Databases
  - Relational
  - Non-relational (NoSQL)
- Text files
  - CSV files
  - Text files
- Live feeds
  - Sensors
  - Online Platforms
    - Twitter
    - Live feeds of weather observations

Data Ingestion to Analytics Platform

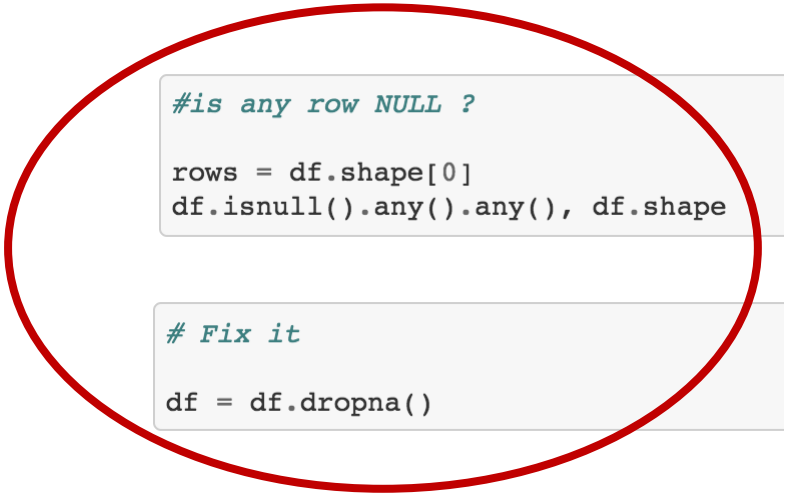
Data Preparation: Explore using Statistics

```
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
id	183978.0	91989.500000	53110.018250	1.0	45995.25	91989.5	137983.75	183978.0
player_fifa_api_id	183978.0	165671.524291	53851.094769	2.0	155798.00	183488.0	199848.00	234141.0
player_api_id	183978.0	135900.617324	136927.840510	2625.0	34763.00	77741.0	191080.00	750584.0
overall_rating	183142.0	68.600015	7.041139	33.0	64.00	69.0	73.00	94.0
potential	183142.0	73.460353	6.592271	39.0	69.00	74.0	78.00	97.0
crossing	183142.0	55.086883	17.242135	1.0	45.00	59.0	68.00	95.0
finishing	183142.0	49.921078	19.038705	1.0	34.00	53.0	65.00	97.0
heading_accuracy	183142.0	57.266023	16.488905	1.0	49.00	60.0	68.00	98.0
short_passing	183142.0	62.429672	14.194068	3.0	57.00	65.0	72.00	97.0
volleys	181265.0	49.468436	18.256618	1.0	35.00	52.0	64.00	93.0
dribbling	183142.0	59.175154	17.744688	1.0	52.00	64.0	72.00	97.0

Data Cleaning

- Why do we need to clean data?
  - Missing entries
  - Garbage values
  - NULLs
- How do we clean data?
  - Remove the entries
  - Impute these entries with a counterpart
    - Ex. Average values of the column
    - Ex. Assign 0, -1, etc



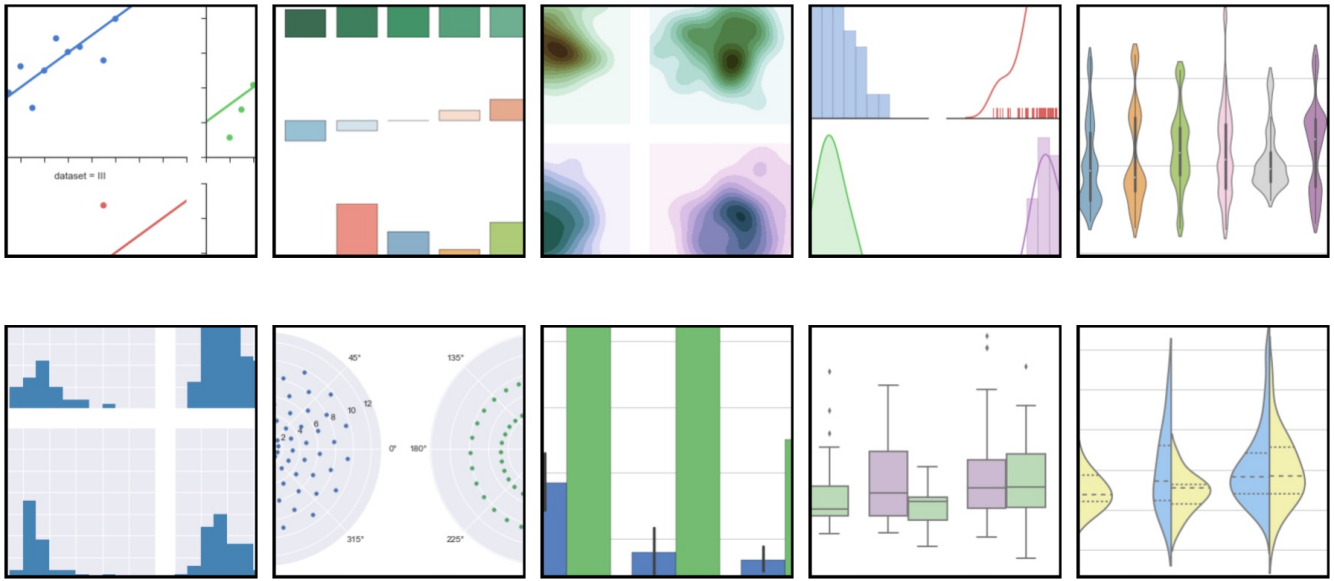
```
#is any row NULL ?

rows = df.shape[0]
df.isnull().any().any(), df.shape
```

```
# Fix it

df = df.dropna()
```

Data Visualization



Convey more in less space and time

Use Graphs when possible

http://seaborn.pydata.org/examples/

Analysis and Modeling

- Supervised Learning
- Unsupervised Learning
- Semi supervised Learning

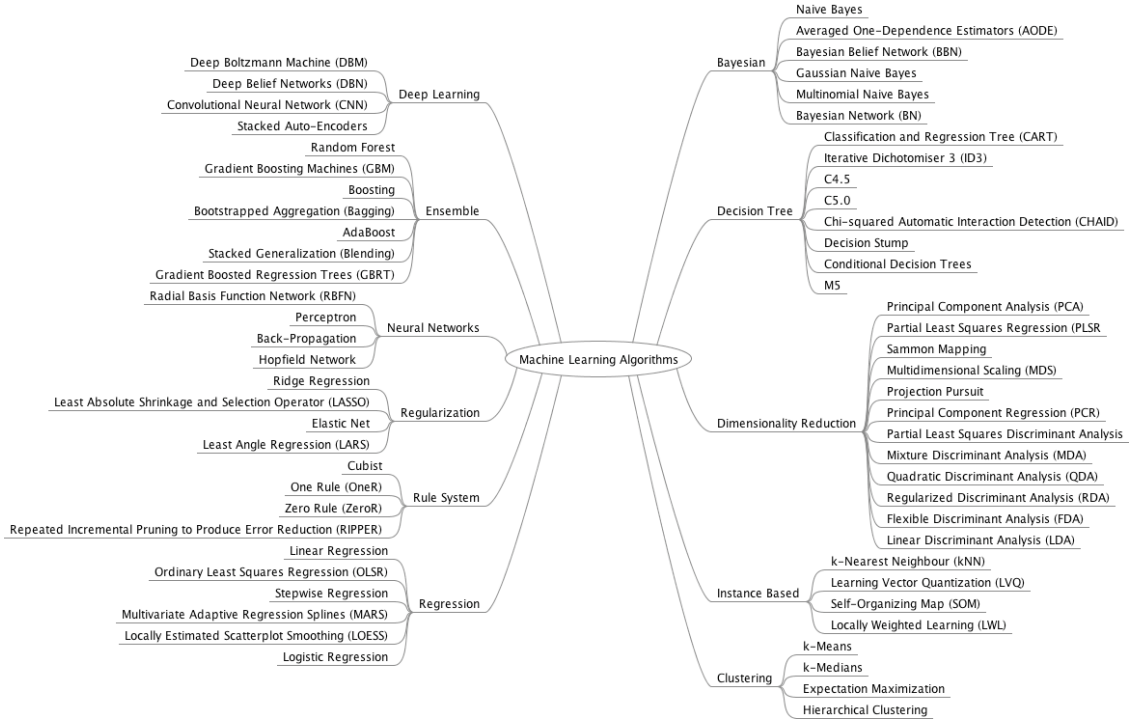
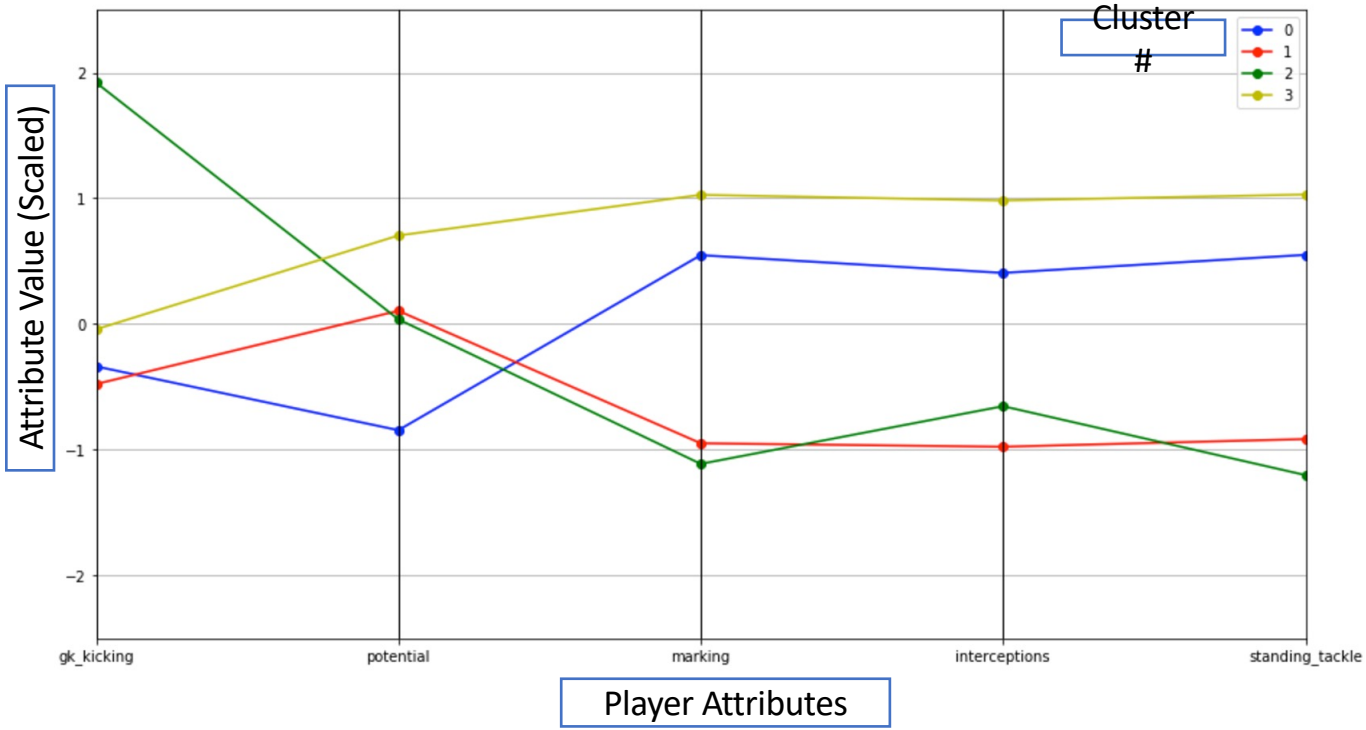


Image Source: https://jixta.wordpress.com/2015/07/17/machine-learning-algorithms-mindmap/

Presenting Data Science Outcomes



Summary

ACQUIRE

PREPARE

ANALYZE

REPORT

ACT

INSIGHTS

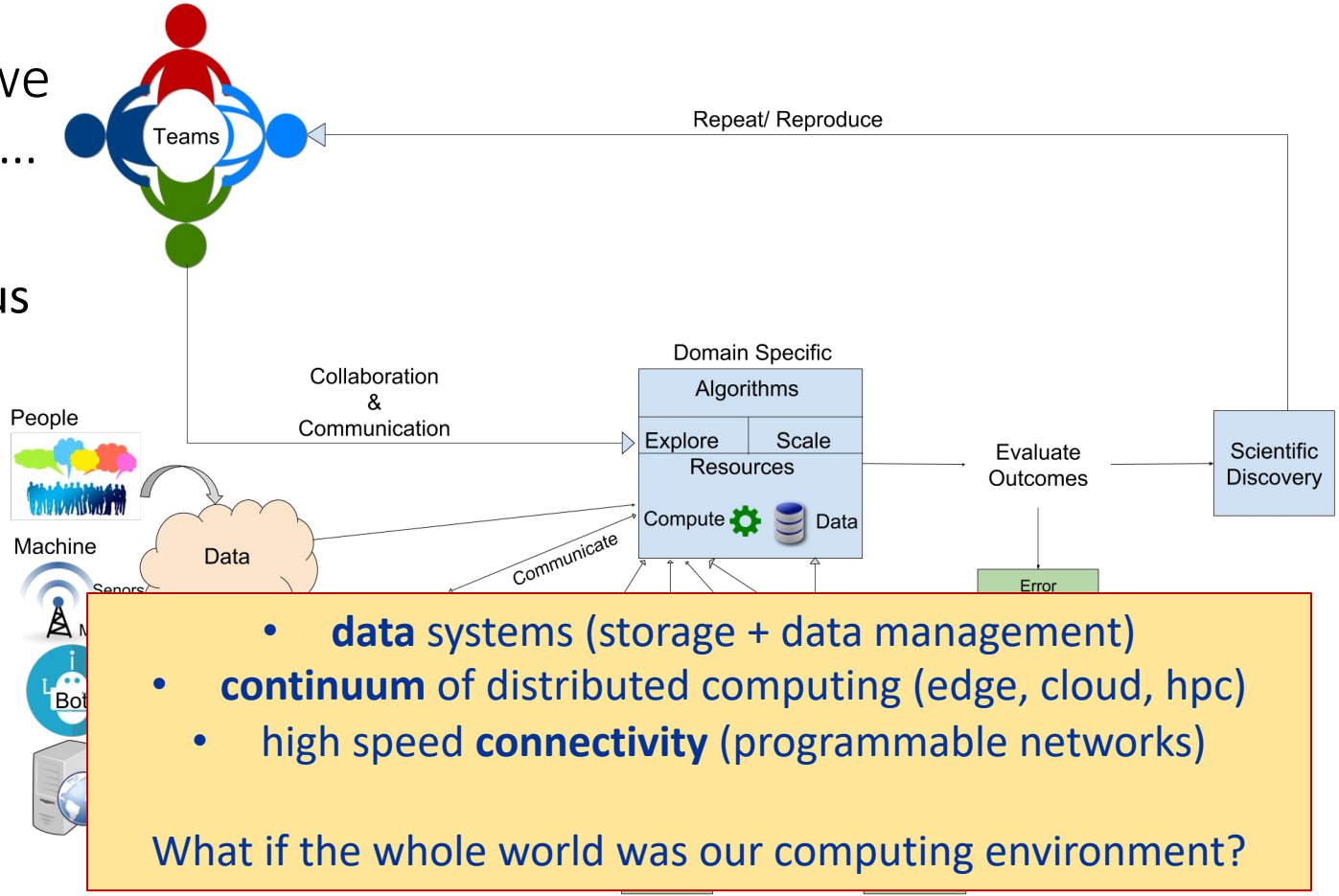
- Better understanding and insights on
  - player strengths
  - enhancing performance
  - critical attributes for a player's performance

ACTIONS

- Coach can design programs that improve these areas in teams

The problems we are solving are …

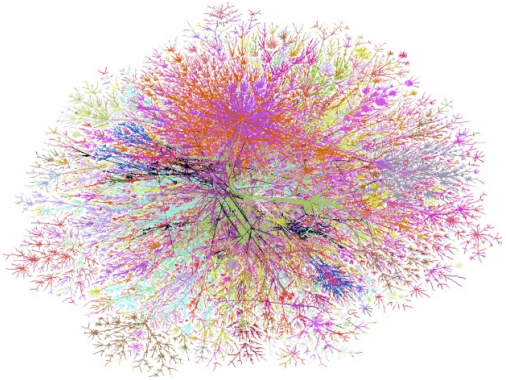
- data-driven
- heterogeneous
- collaborative



- **data** systems (storage + data management)
- **continuum** of distributed computing (edge, cloud, hpc)
- high speed **connectivity** (programmable networks)

What if the whole world was our computing environment?

Big Data combined with Scalable Computing can be very valuable.



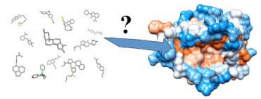
COMPUTING AT DIVERSE SCALES

"BIG" DATA

Enables dynamic data-driven applications

Smart Manufacturing

Computer-Aided Drug Discovery



Personalized Precision Medicine

Smart Cities

Disaster Resilience and Response

Smart Grid and Energy Management

Research Goal:

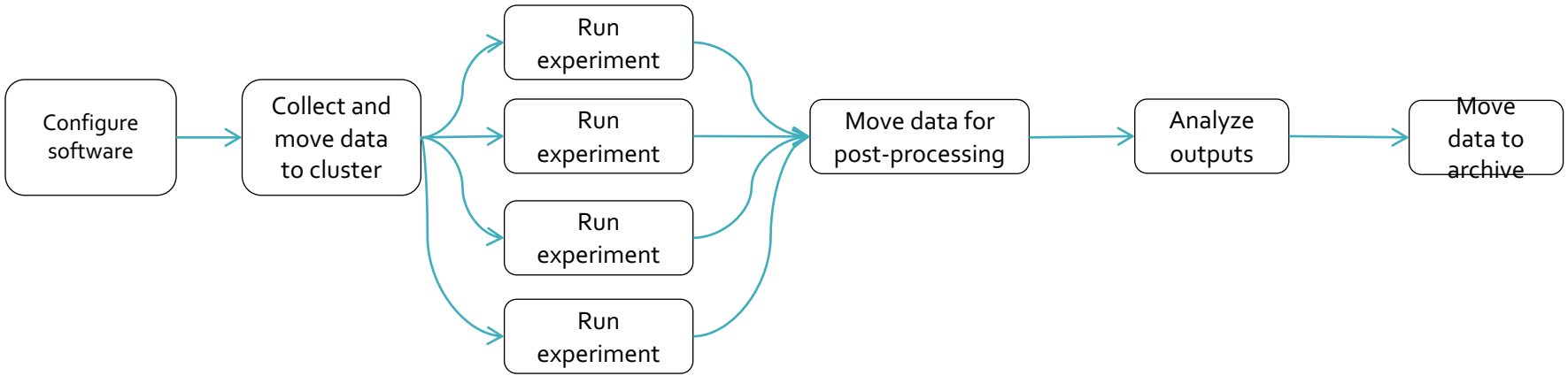
Create a Data Science Solution Ecosystem that Enables Needs and Best Practices

- data-driven
- scalable
- dynamic
- process-driven
- heterogeneous
- accountable
- reproducible
- interactive
- multidisciplinary
- collaborative

Part2: Introduction to Data Science Workflows

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

Pipeline



CHOICES:

- Scripting
- Makefiles
- Workflow systems

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

COMPUTING AT SCALE

BIG DATA

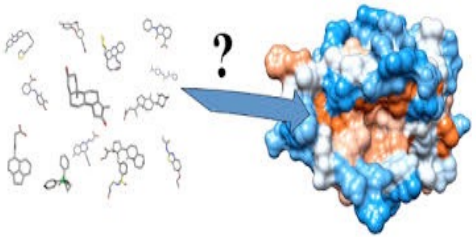
Requires software for dynamic coordination and resource optimization

Workflow Systems

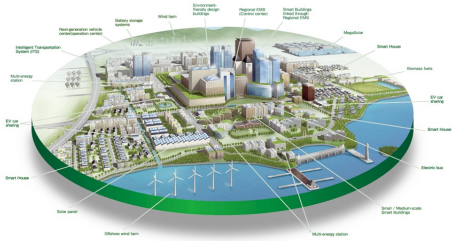
Enables dynamic data-driven applications

Impact of Scientific Workflows

Computer-Aided Drug Discovery



Smart Cities



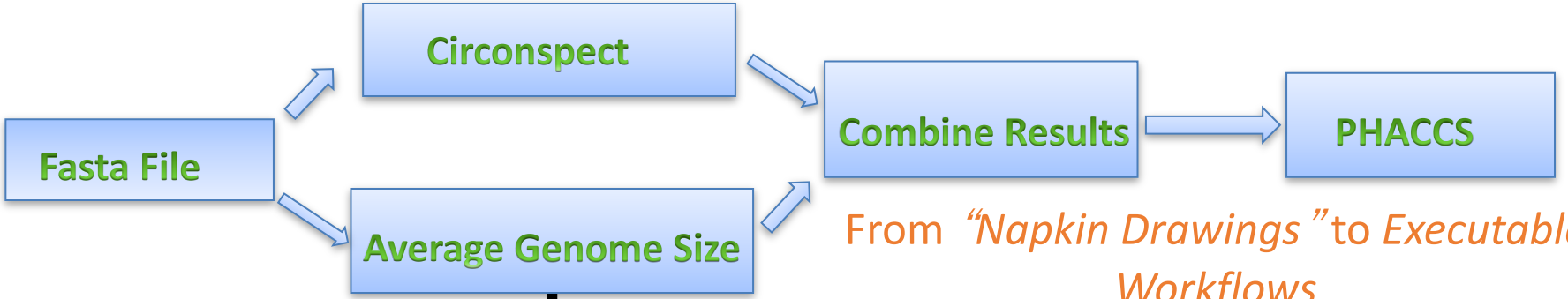
Disaster Resilience and Response

Smart Grid and Energy Management

Personalized Precision Medicine

Smart Manufacturing

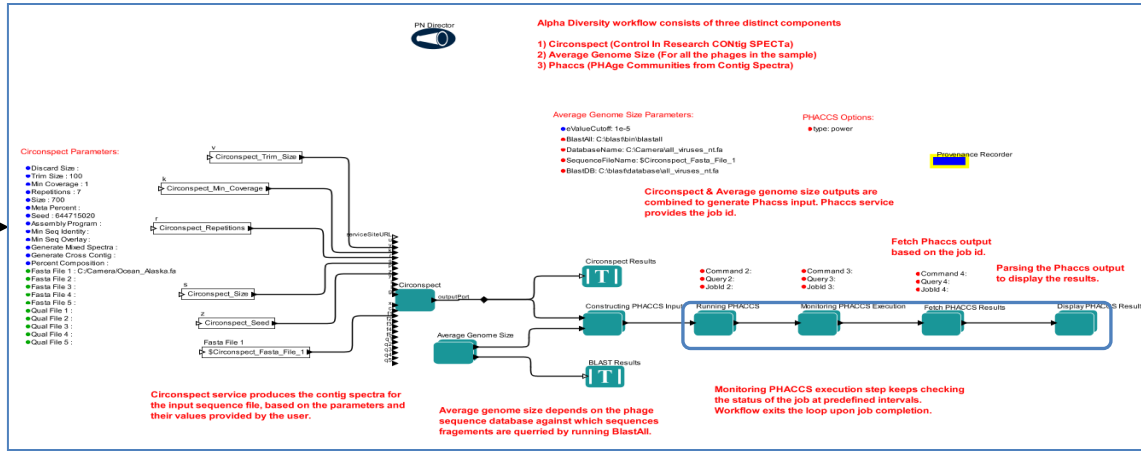
The Big Picture is Supporting the Scientist



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

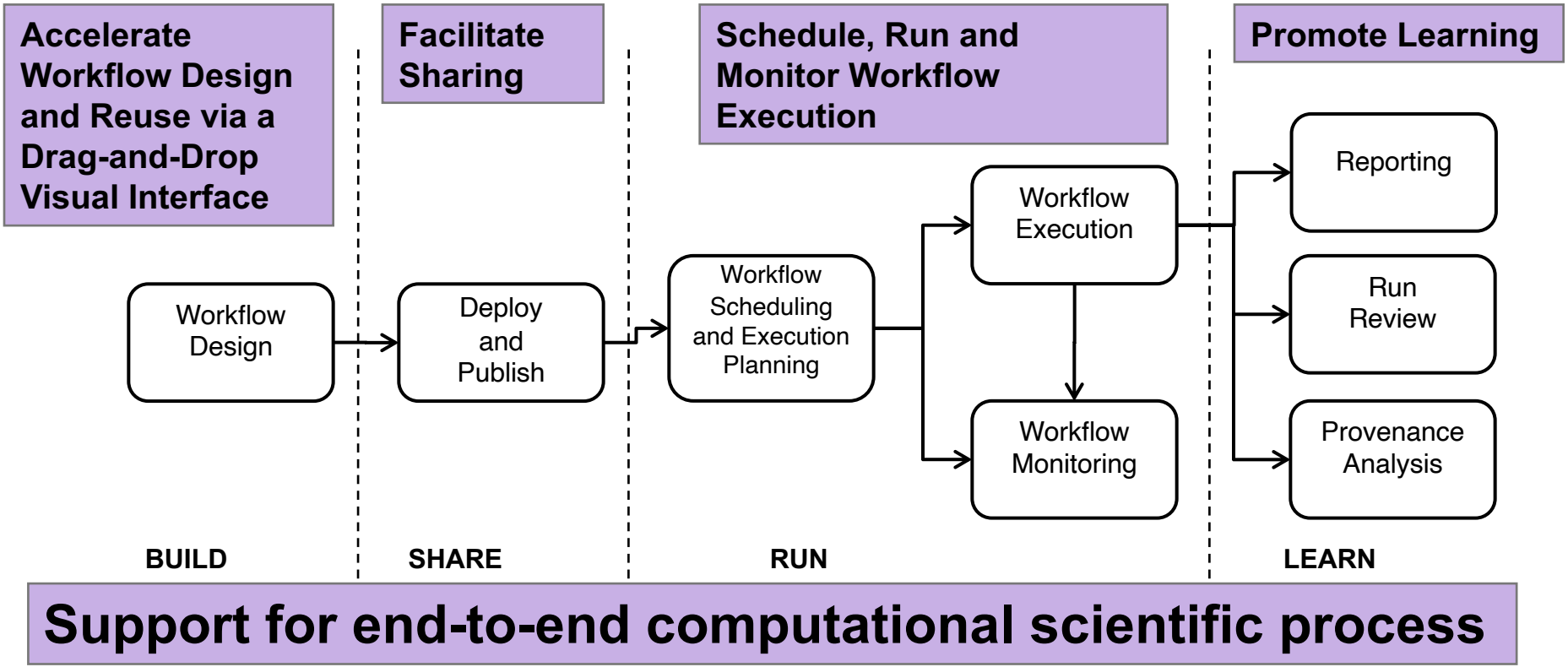
Provenance Helps with Accountability and Reproducibility

- 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

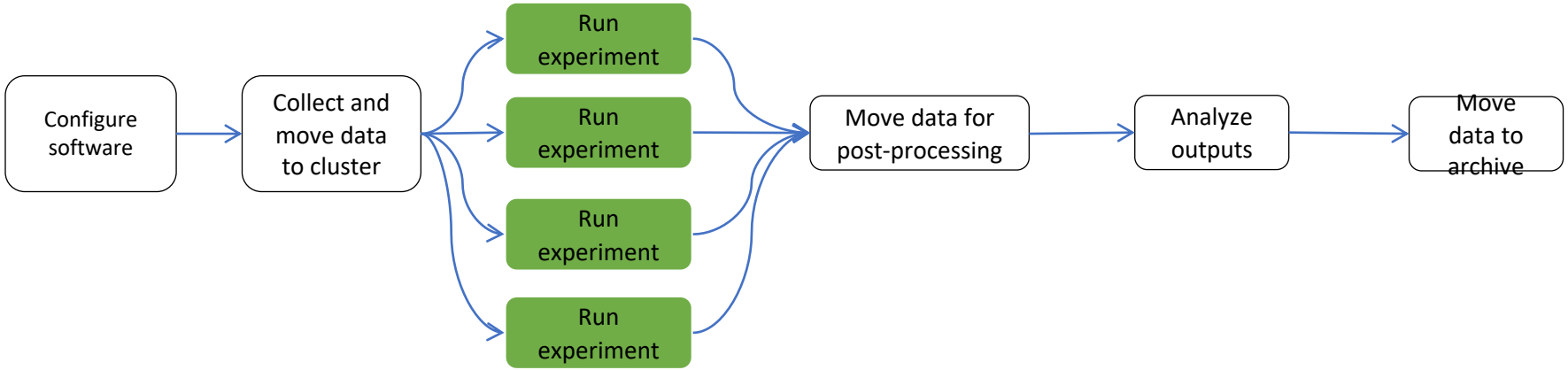
Workflows are a part of Cyberinfrastructure



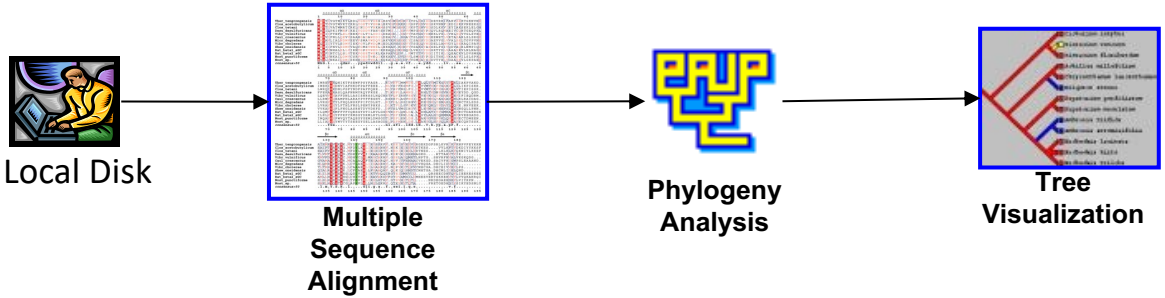
Accelerate Science

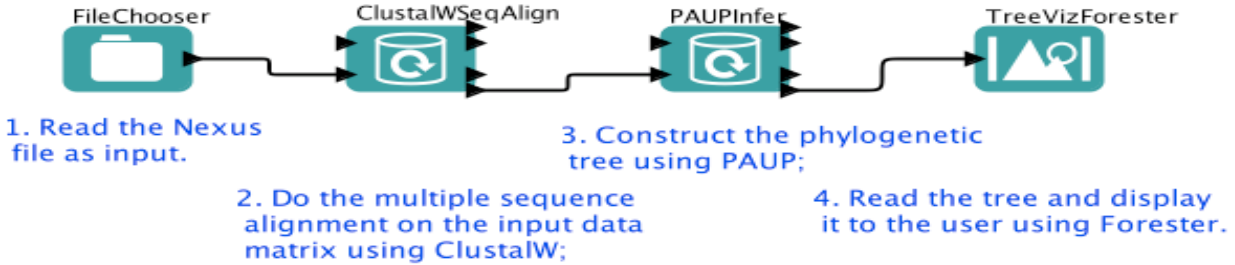
- Speed : build using Drag-and-Drop Visual Interface
- Adaptability : ability to work across multiple systems
- Integration : interconnect with other workflows
- Customization : declare what needs to be done,
  give freedom of execution
- Reusability : a part or entire workflow

Infinite ways to create workflows

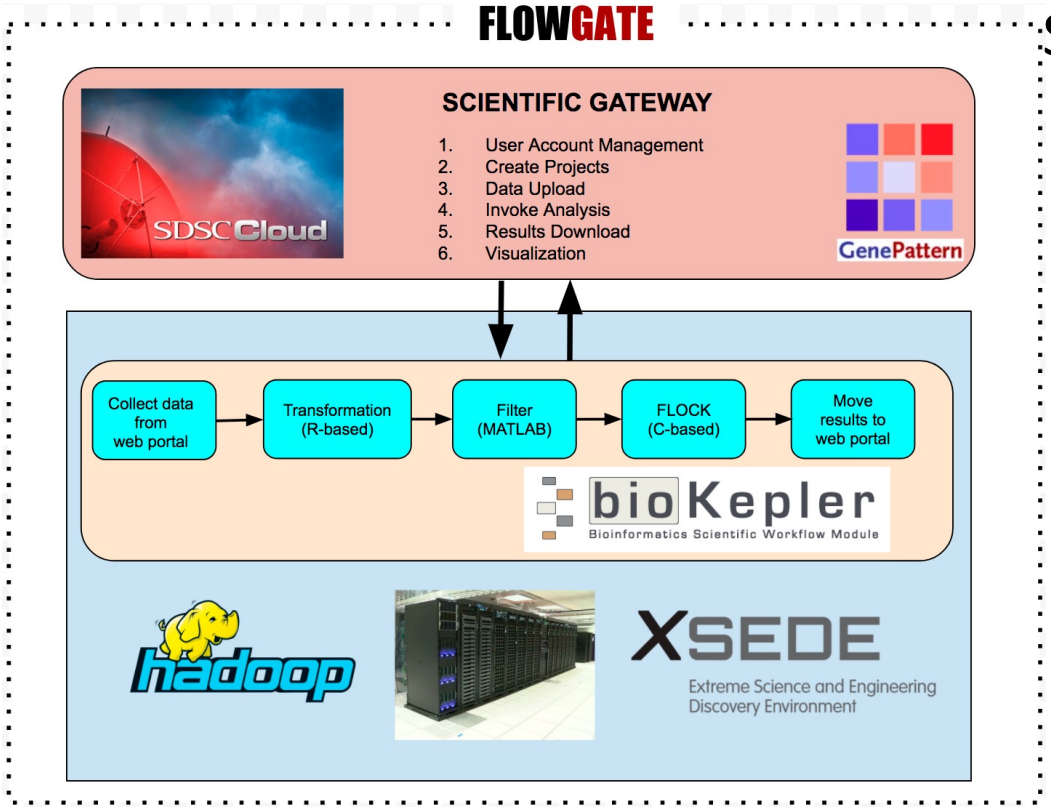


Simple workflows



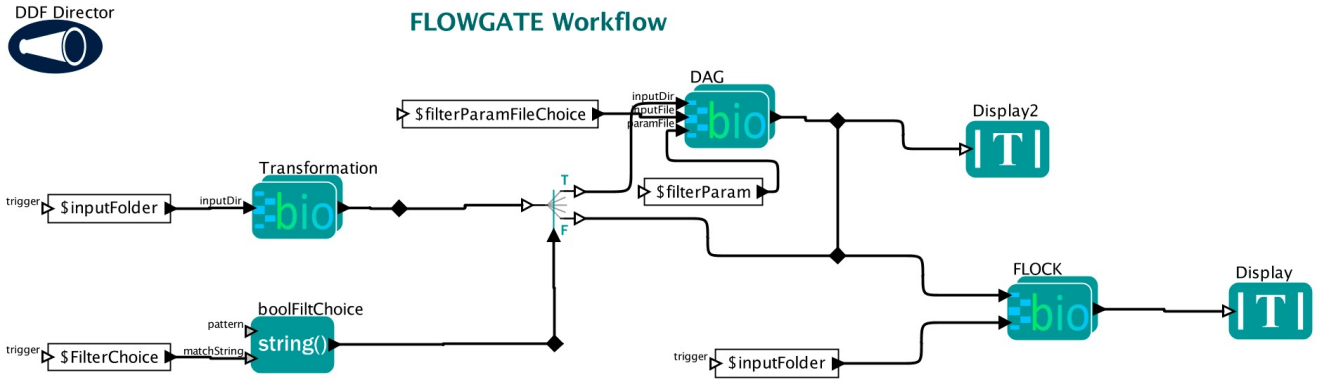


Workflows integrating multiple components



Scalable Web-Based Flow Cytometry Data Analysis

Workflows integrating multiple components



FLOWGATE Workflow:
Scalable Flow
Cytometry
Data Analysis

This workflow implements FLOCK-like flow cytometry data analysis. It contains Transformation, Filter (DAG) and FLOCK stages. The workflow transfers user input data from front-end web portal to the backend workflow cloud virtual machine or to the cluster, carries computational analysis in background for multiple input files on Cloud VM or the Gordon Supercomputer, and uploads result back to the web portal. The workflow configures or by-pass filter stage depending user filter choice.

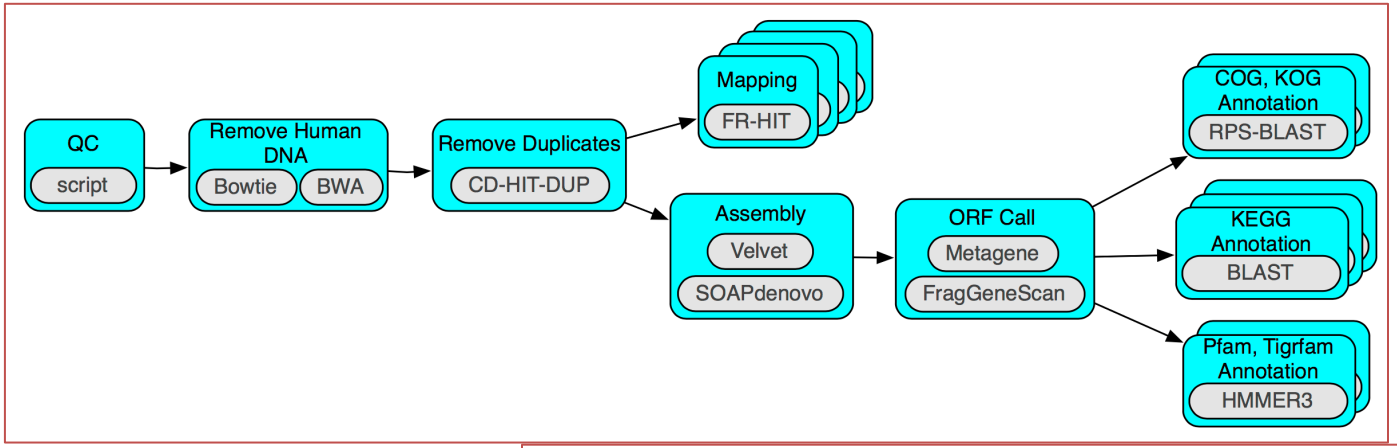
See http://flowgate.jcvi.org

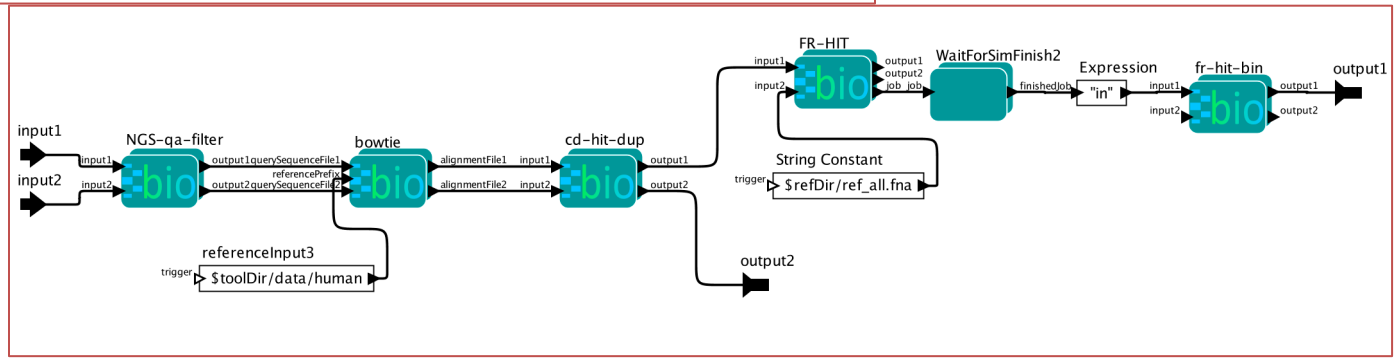
Authors: Shweta Purawat, Jianwu Wang, Ilkay Altintas, Robert Sinkovits @ SDSC
Yu Qian, Rick Stanton, Hyunsoo Kim, Richard Scheuermann @ JCVI

- filterParamFileChoice: b_yale
- FilterChoice: True
- filterParam: $toolDir/b_yale_popDef.csv
- inputFolder: /home/spurawat/JCVI/CWDflowER/User5_InputTest
- toolDir: /home/spurawat/JCVI/CDUCModExec
- bins: 11
- density: 12
- population: 13

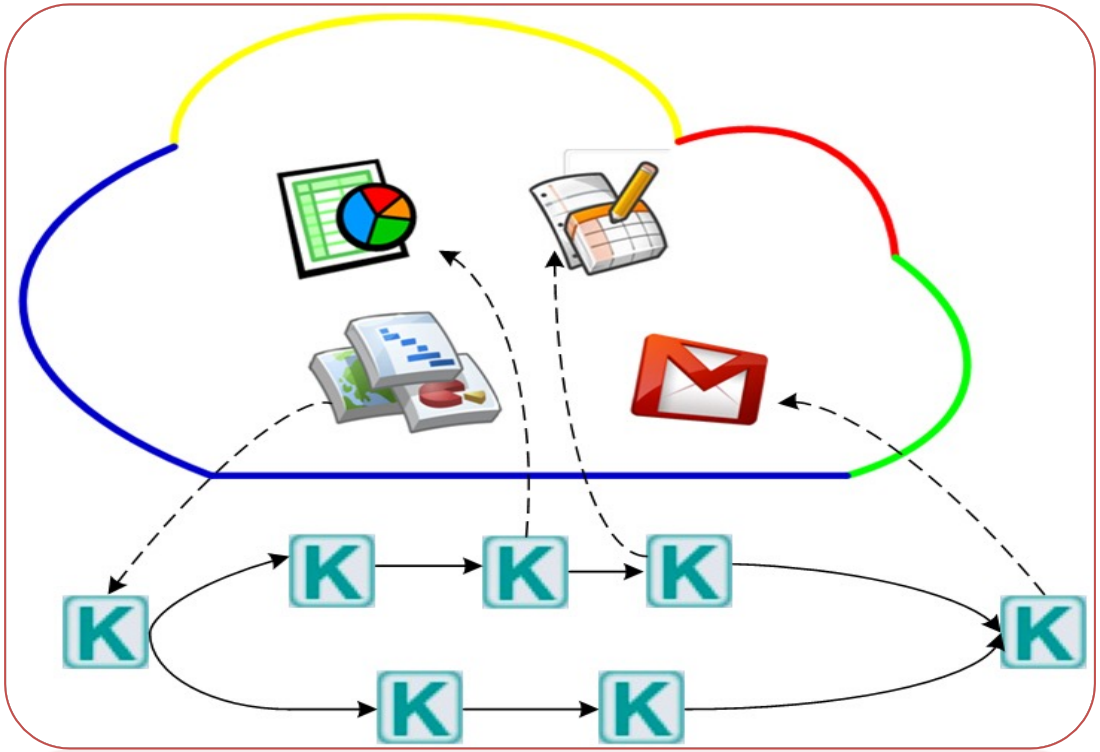
Compute and data intensive workflow

Microbiome Taxonomy and Gene Abundance Workflow (MTGA)





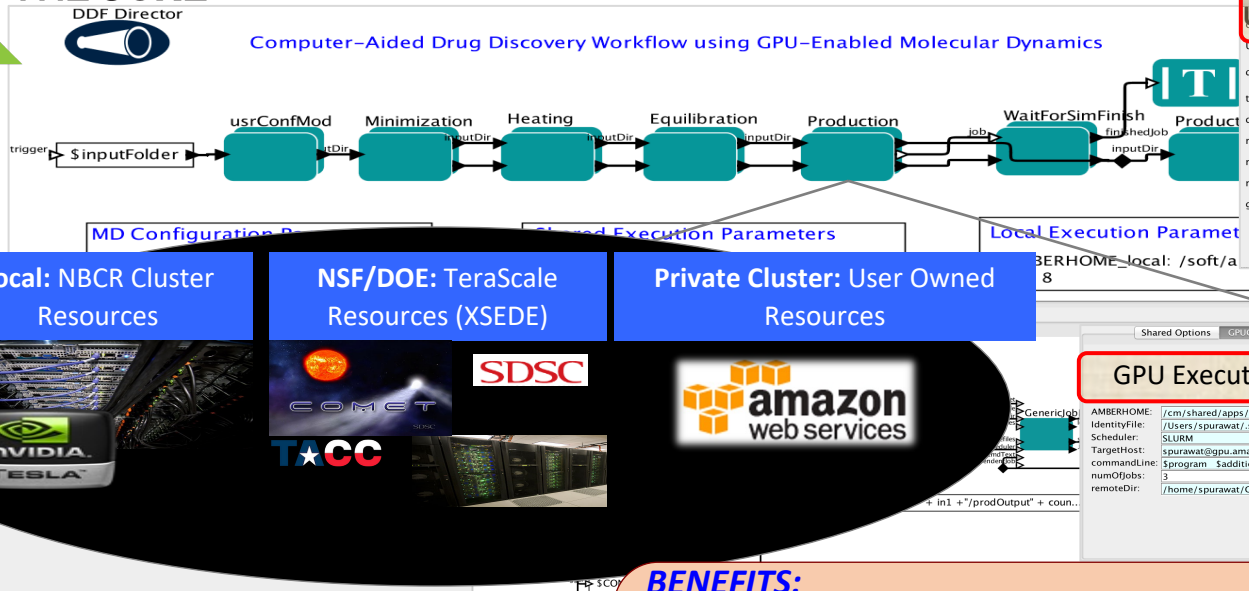
Workflows that Integrate cloud resources

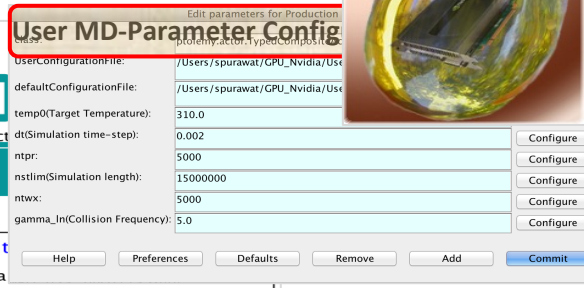


AMBER GPU Molecular Dynamics

COMPUTE THE CURE

DDF Director





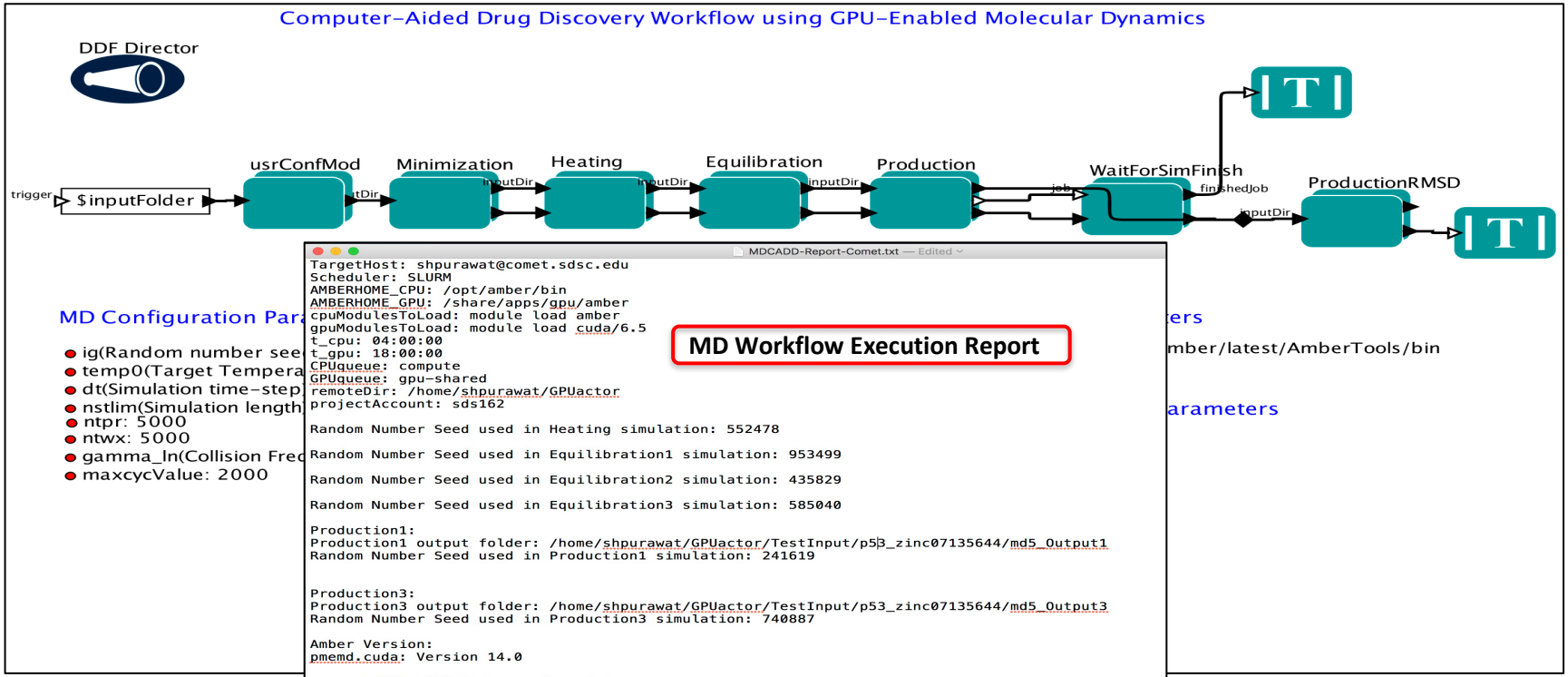
BENEFITS:

- *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

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

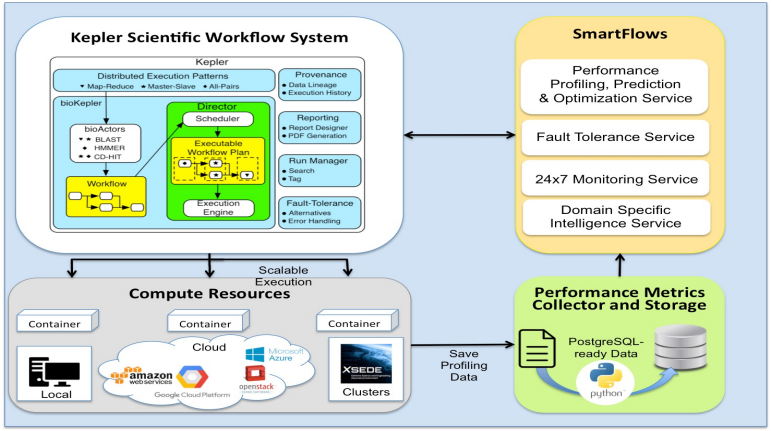
Area of Interest

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

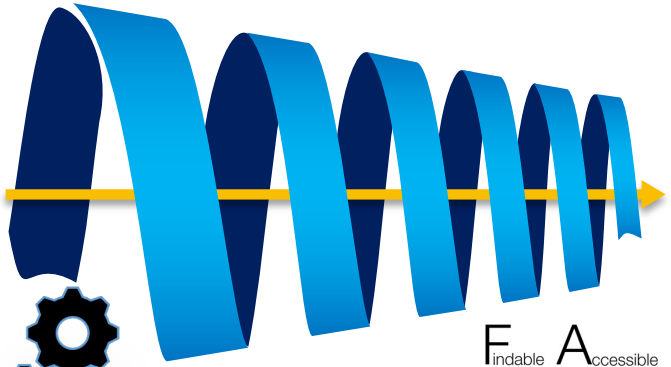
Curate

Manage

Catalog

Integrate

Analyze



Findable Accessible Interoperable Reusable

Scalable Data and Computing Cyberinfrastructure

Search

Query

Visualize

Analyze

Apply

Q-AMASE-i Centers

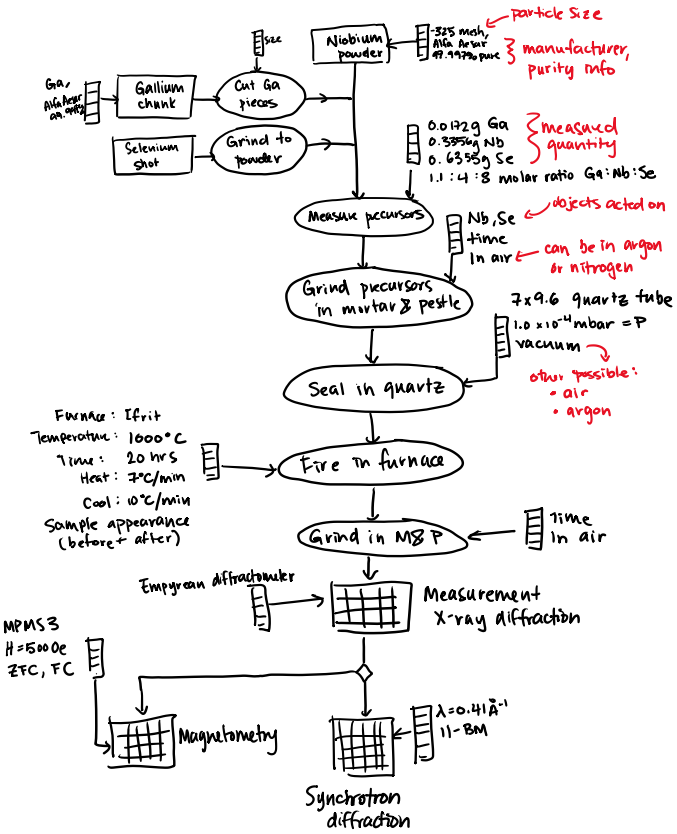
Create, Process, and Characterize materials for quantum information science

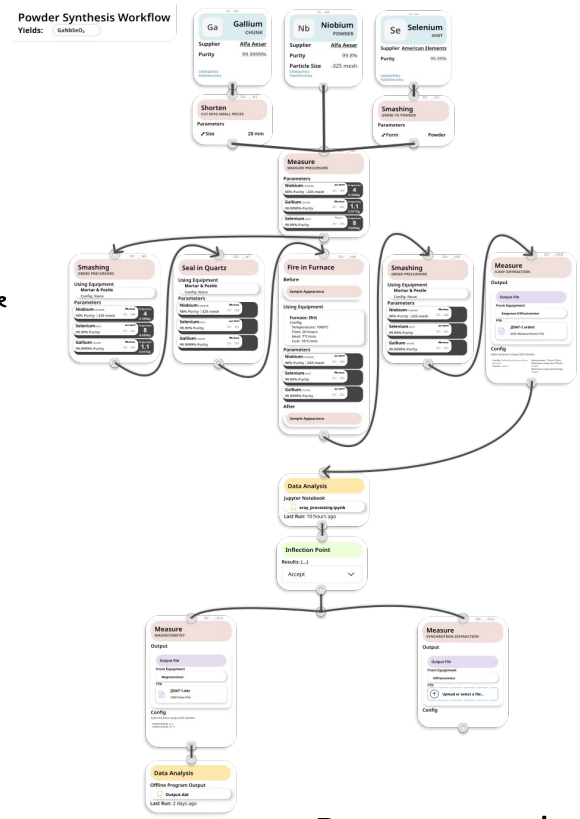
UC San Diego

Collect, curate and manage the foundry data

Amplifying the Value of Foundry Data Through Data Science!

Workflow for the Lab Notebook Process

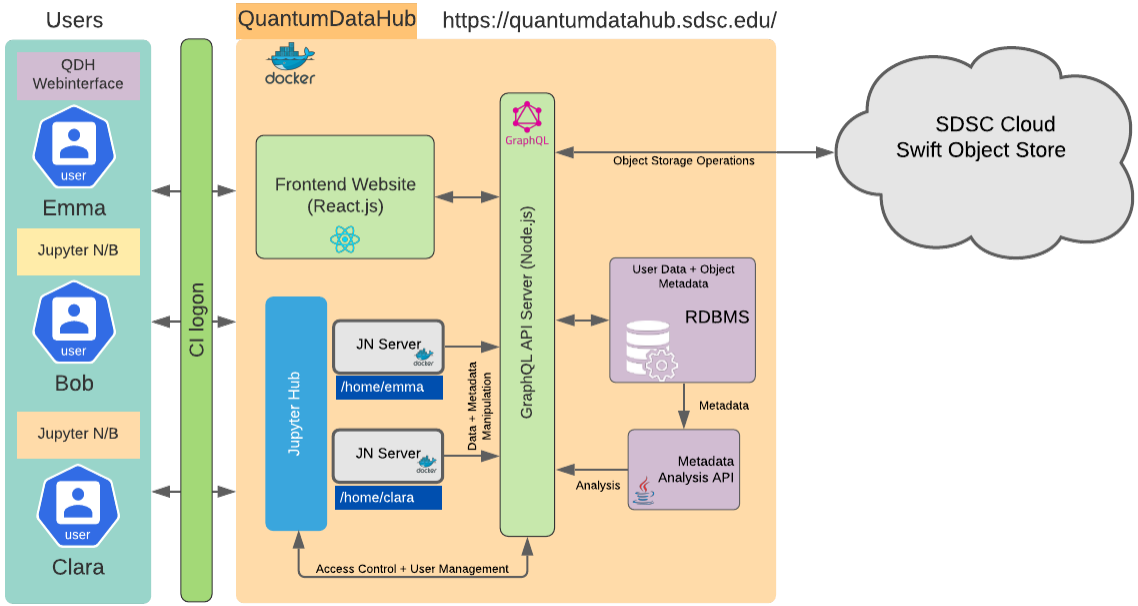




Represent the experiment process as DAG.

Quantum Data Hub Architecture

https://quantumdatahub.sdsc.edu/



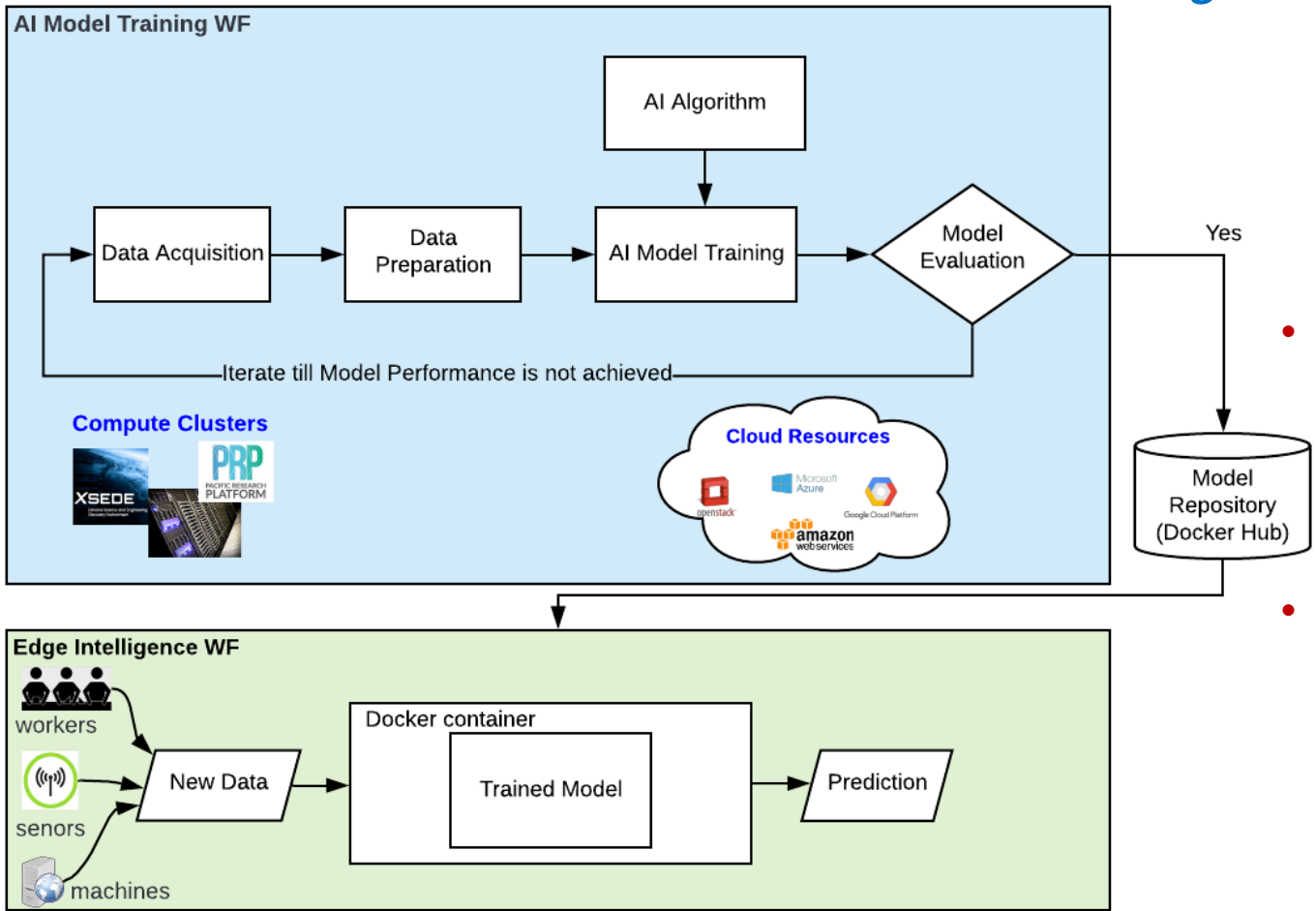
User Interface - React, Node.js

Quantum Data Hub (QDH) API

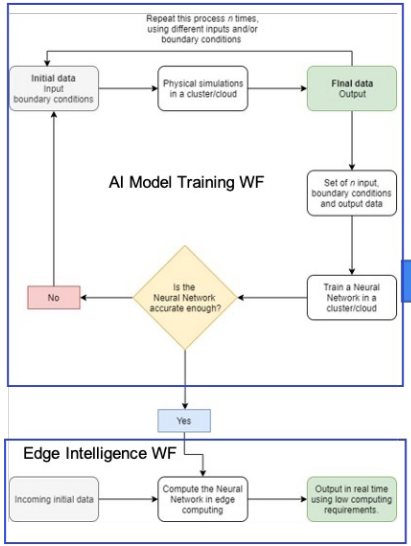
JupyterHub - Simple and Intuitive Jupyter Notebook Exposing QDH APIs

Database – User management, Activity log and Quantum Metadata

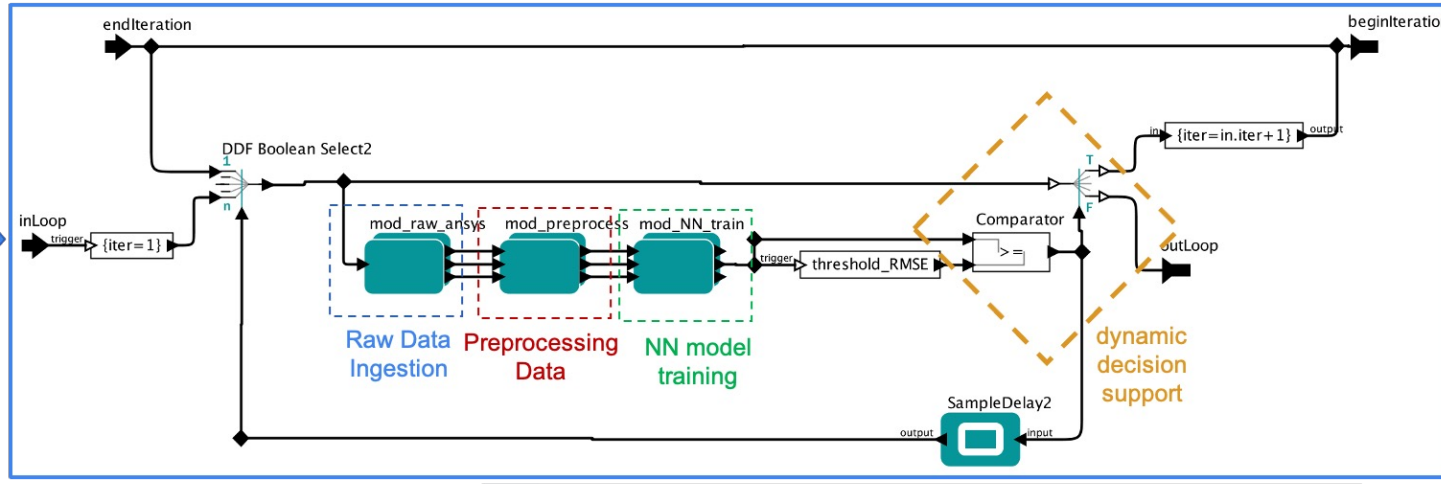
Smart Manufacturing - AI Workflow Reference Architecture for Advance Manufacturing



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

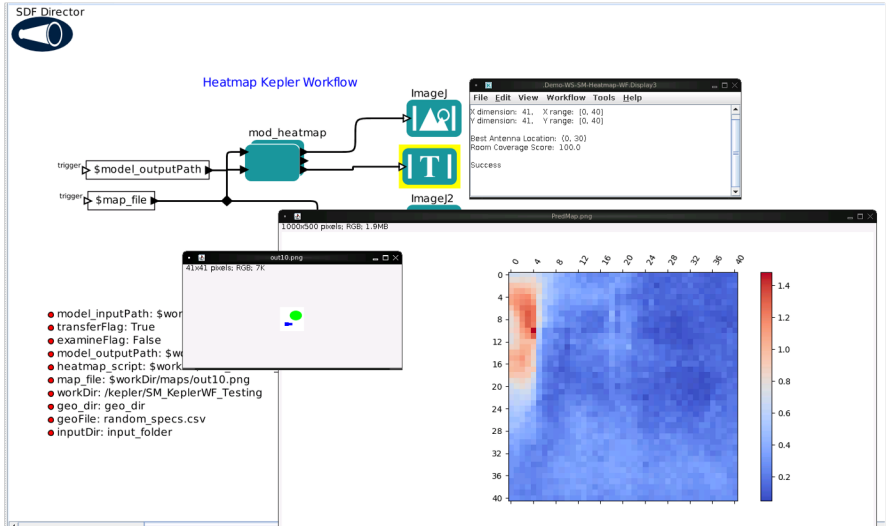


Conceptual flow diagram

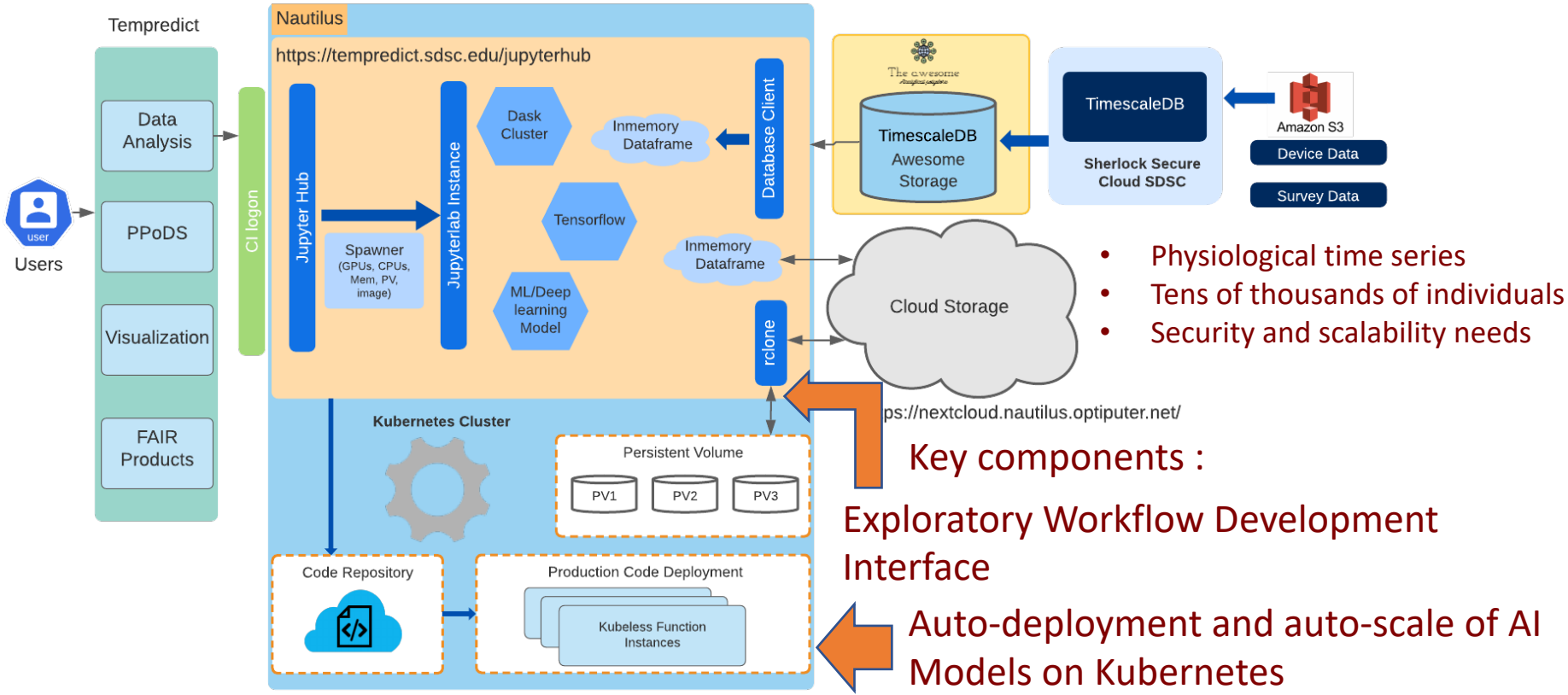


The CNN Model Training Workflow for WiFi Received Signal Strength Intensity (RSSI)

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



Tempredict: 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

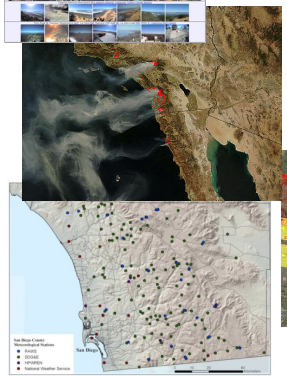
WIFIRE

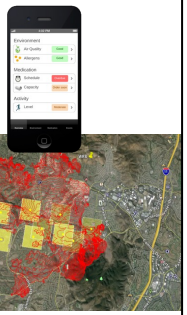
Monitoring

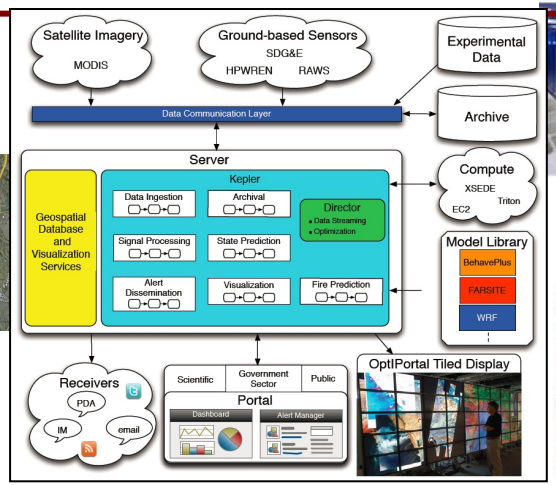
Visualization

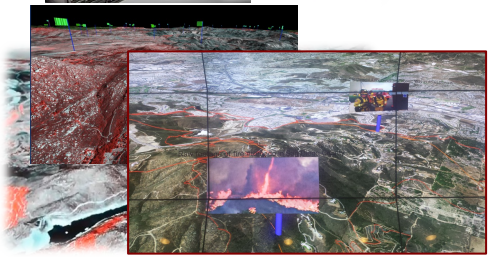
Fire Mapping

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









Multimodal wild-fire related data

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

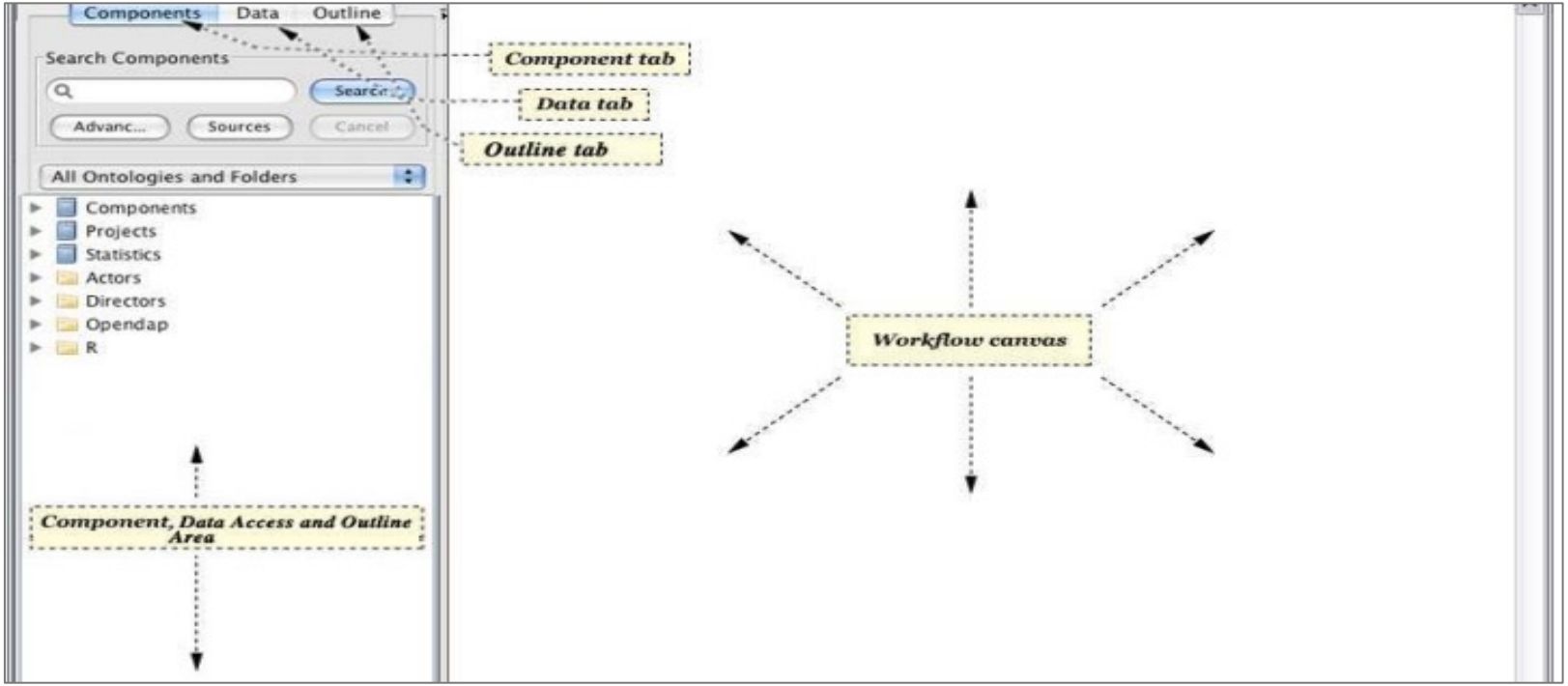
Graphical Workflow Systems

-Toolboxes with Many Tools-

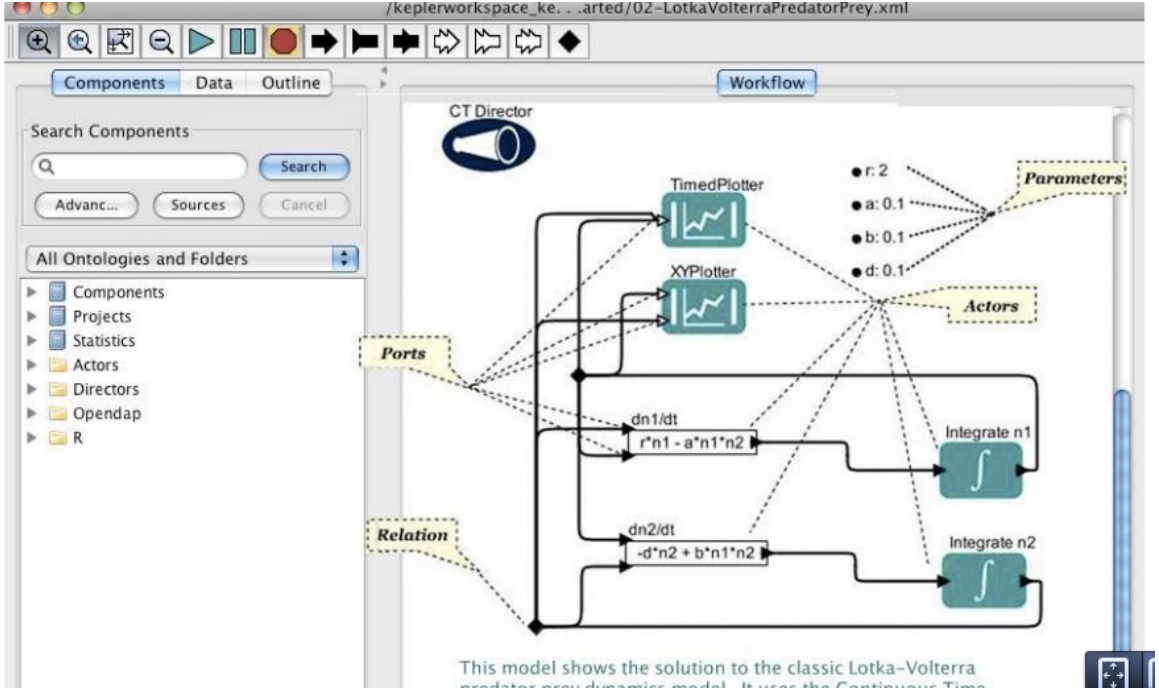
- Data
  - Search, database access, IO operations, streaming data in real-time…
- Compute
  - Data-parallel patterns, external execution, …
- Network operations
- Provenance and fault tolerance

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

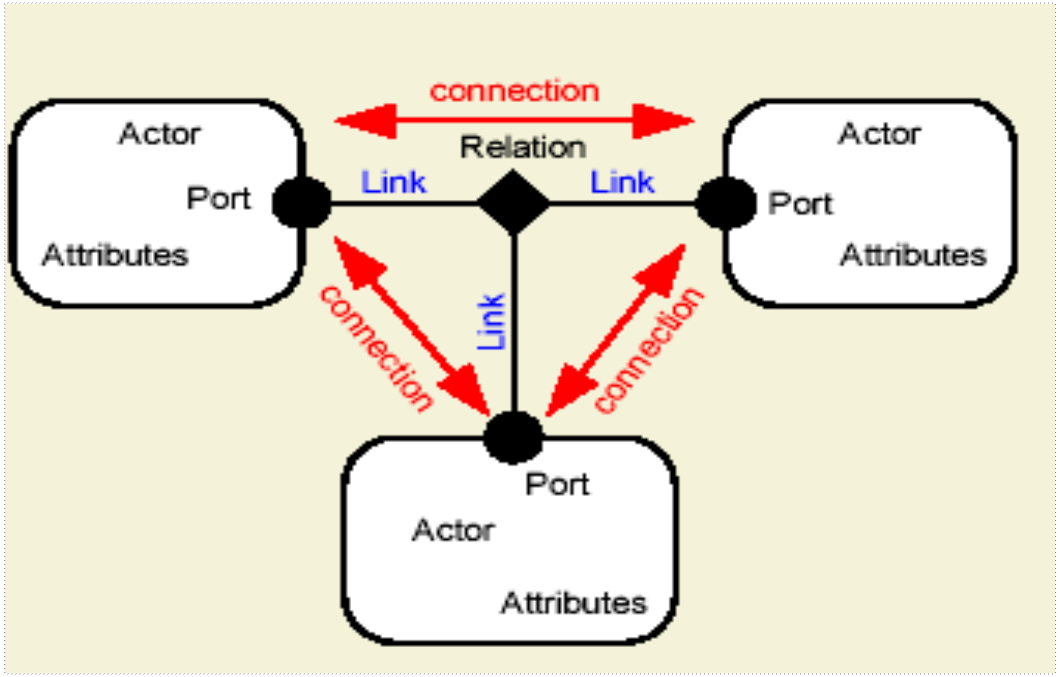
Kepler's UI


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

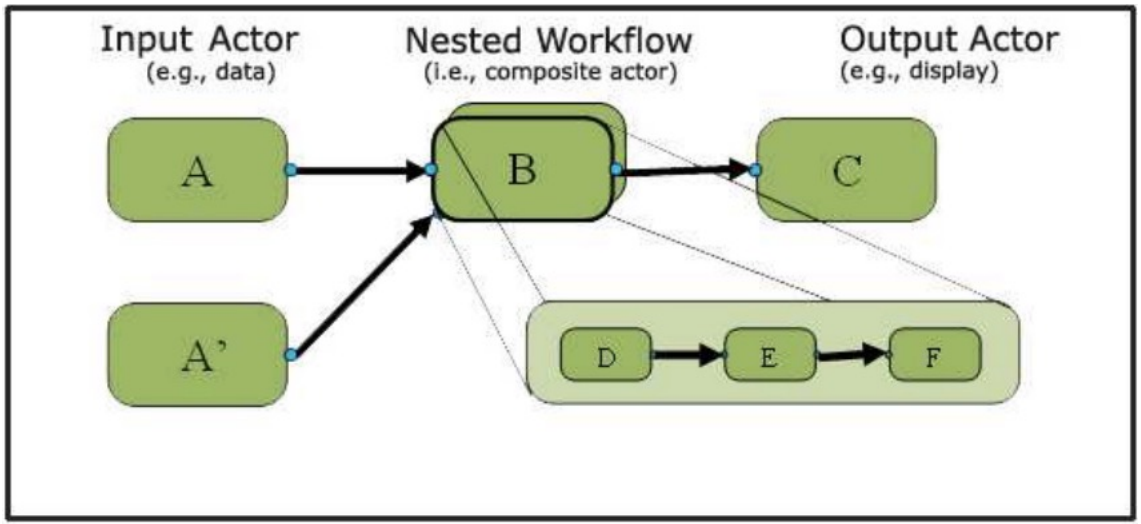


Actors are the Processing Components



Actor-Oriented Design

Kepler Actor Types

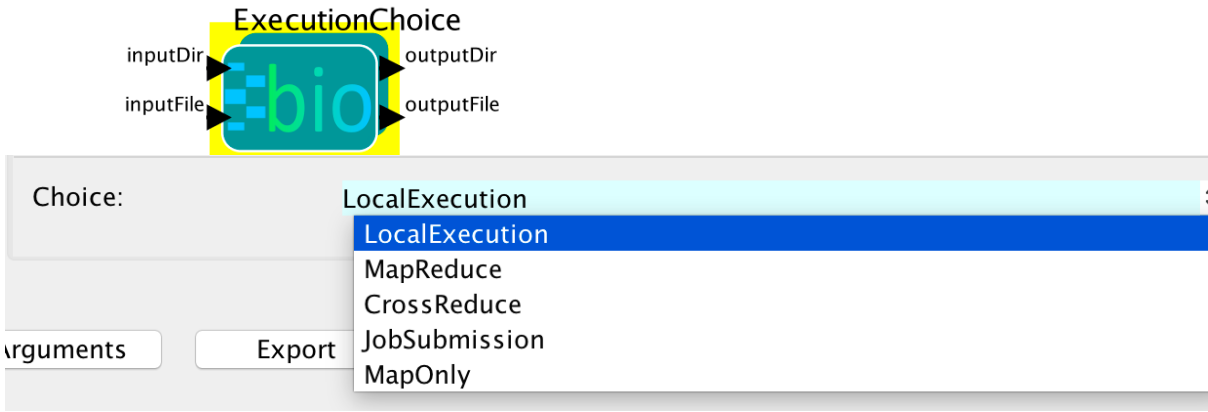


Some actors in place for…

- 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
  - **Local** execution
  - **Hadoop** execution
  - **EC2** execution
  - **Remote job** execution



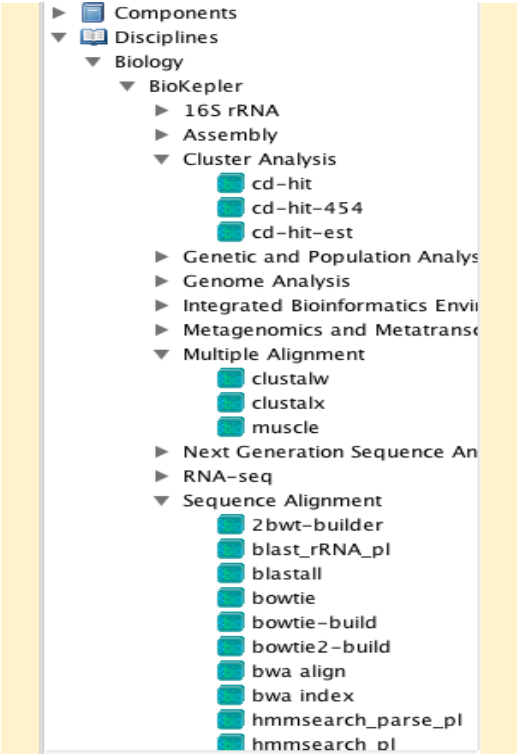
- Useful for heterogeneous execution requirements

bioKepler

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

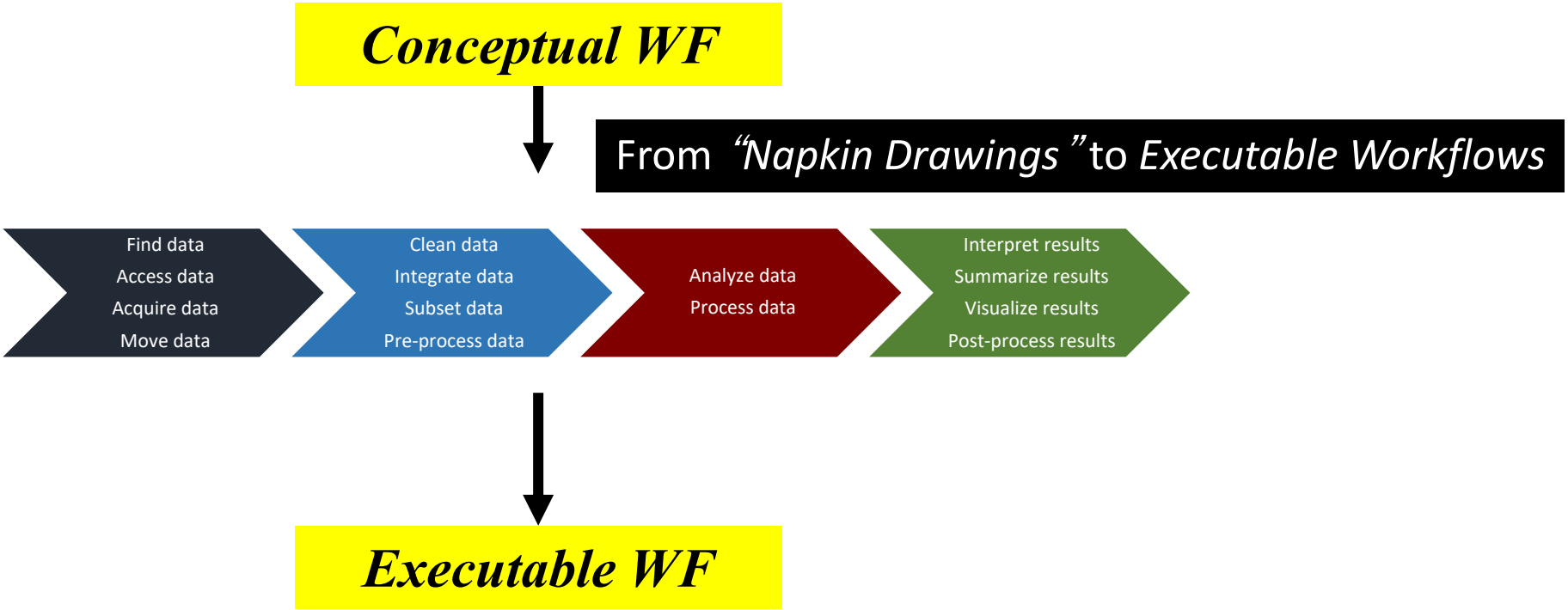
https://www.biokepler.org/install-biokepler-1.2

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



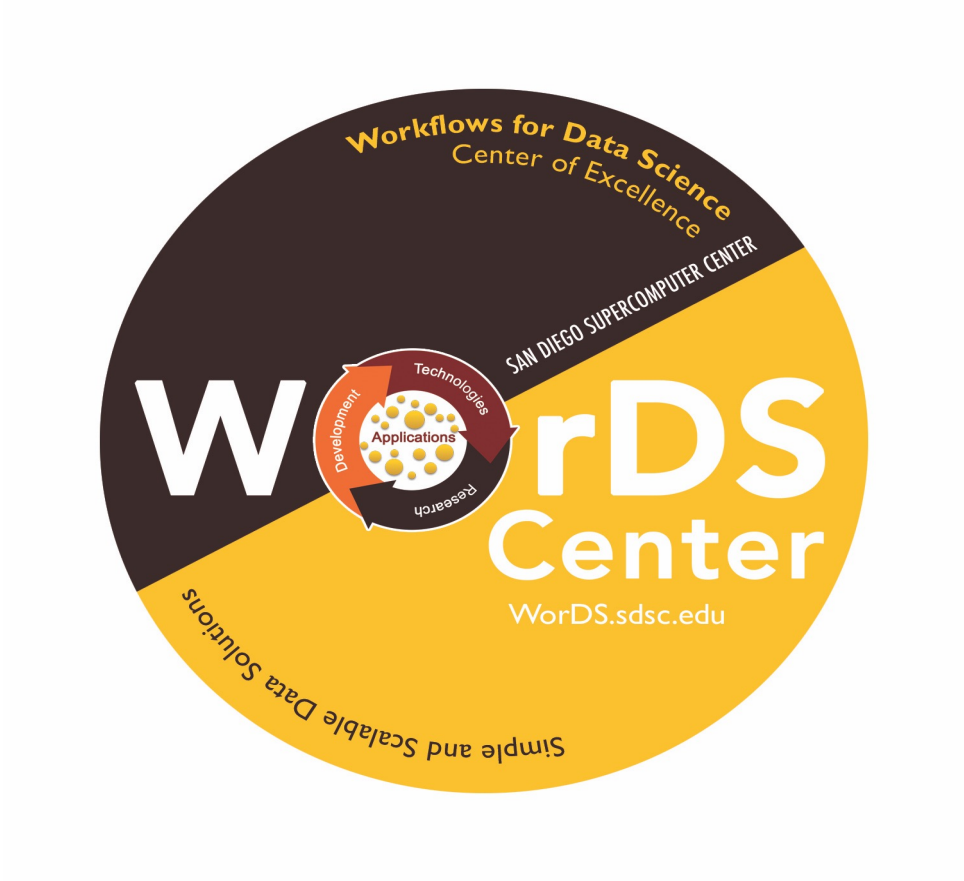
Section Summary:

- 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

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!

How do we find the connections?

Such Data Science Applications requires, proper use of Data Management

Finding the connections between the different kinds of datasets leads to interesting discoveries.