How do I automate my job pipelin
ensure reproducibility?

İlkay ALTINTAŞ. Ph.D.

Chief Data Science Officer, San Diego Supercomputer Center
Division Director, Cyberinfrastructure Research, Education and De
Director, Workflows for Data Science Center of Excellence
Lecturer, Computer Science and Engineering
University of California, San Diego

Welcome to Day 2!

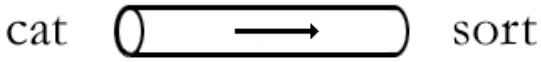
What have you learned so far?

How do you interact with HPC systems?

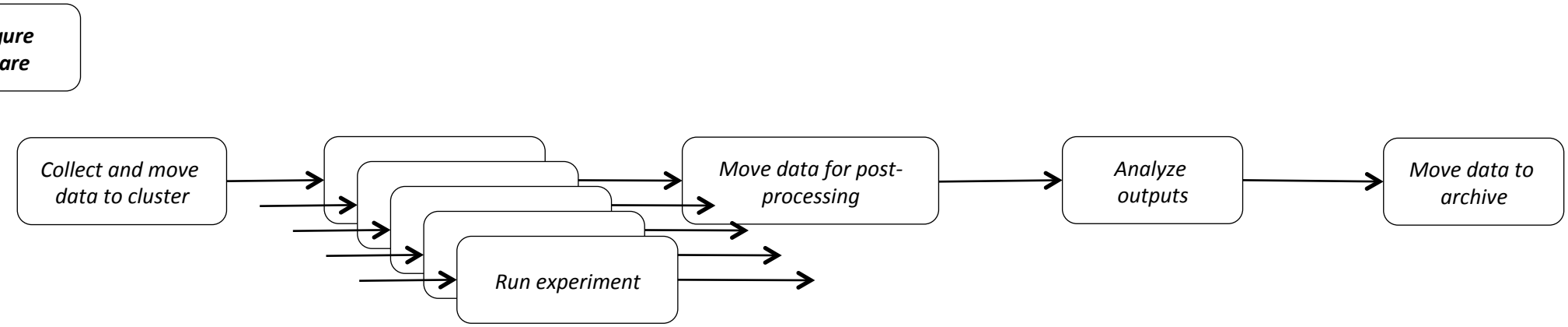
- Launching and managing jobs
- Managing data on the file system

What if you wanted to combine these steps?

What is a Pipeline?



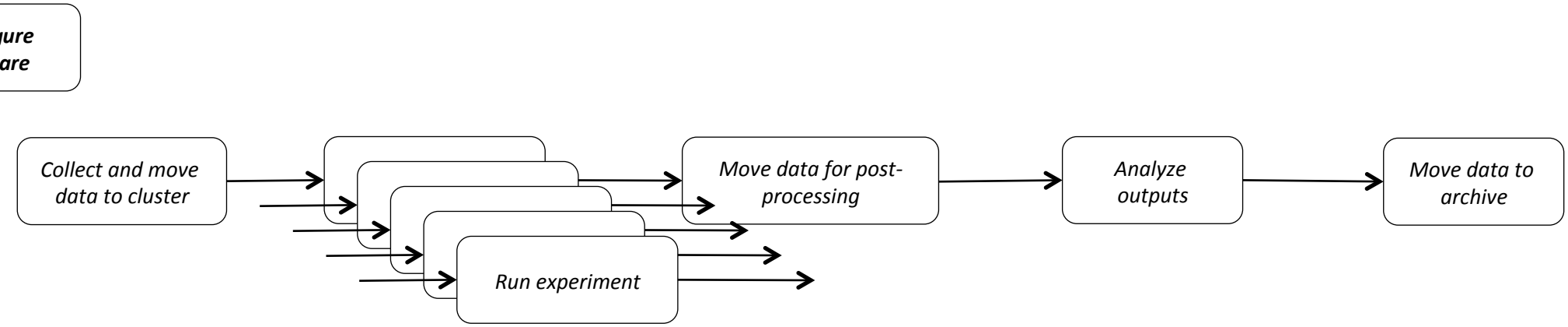
A UNIX pipe provides one-way communication between two processes on the same computer

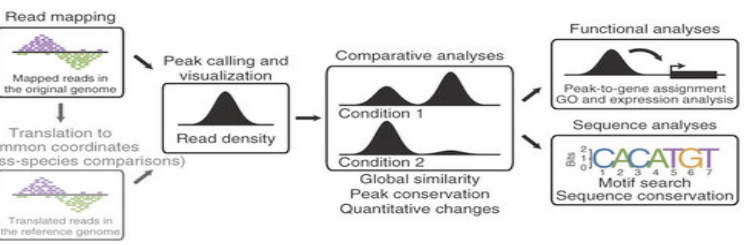


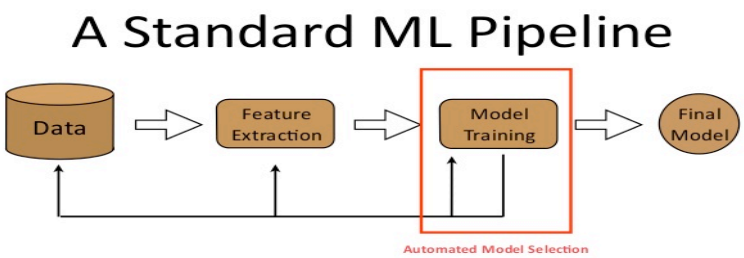
OICES:

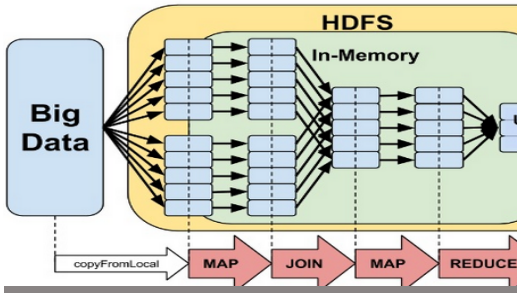
- Scripting
- Makefiles
- Workflow systems

Many different pipelines…

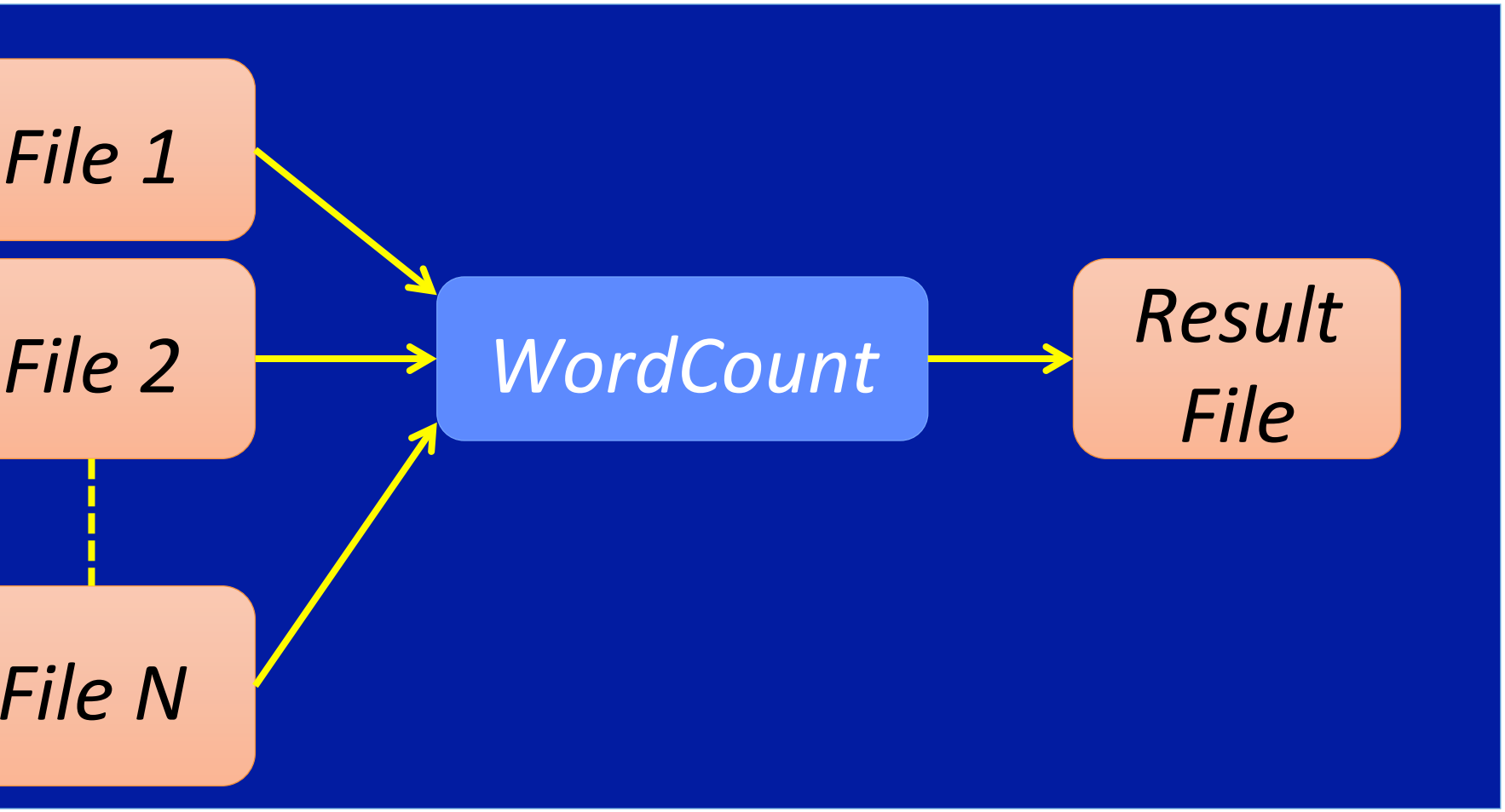


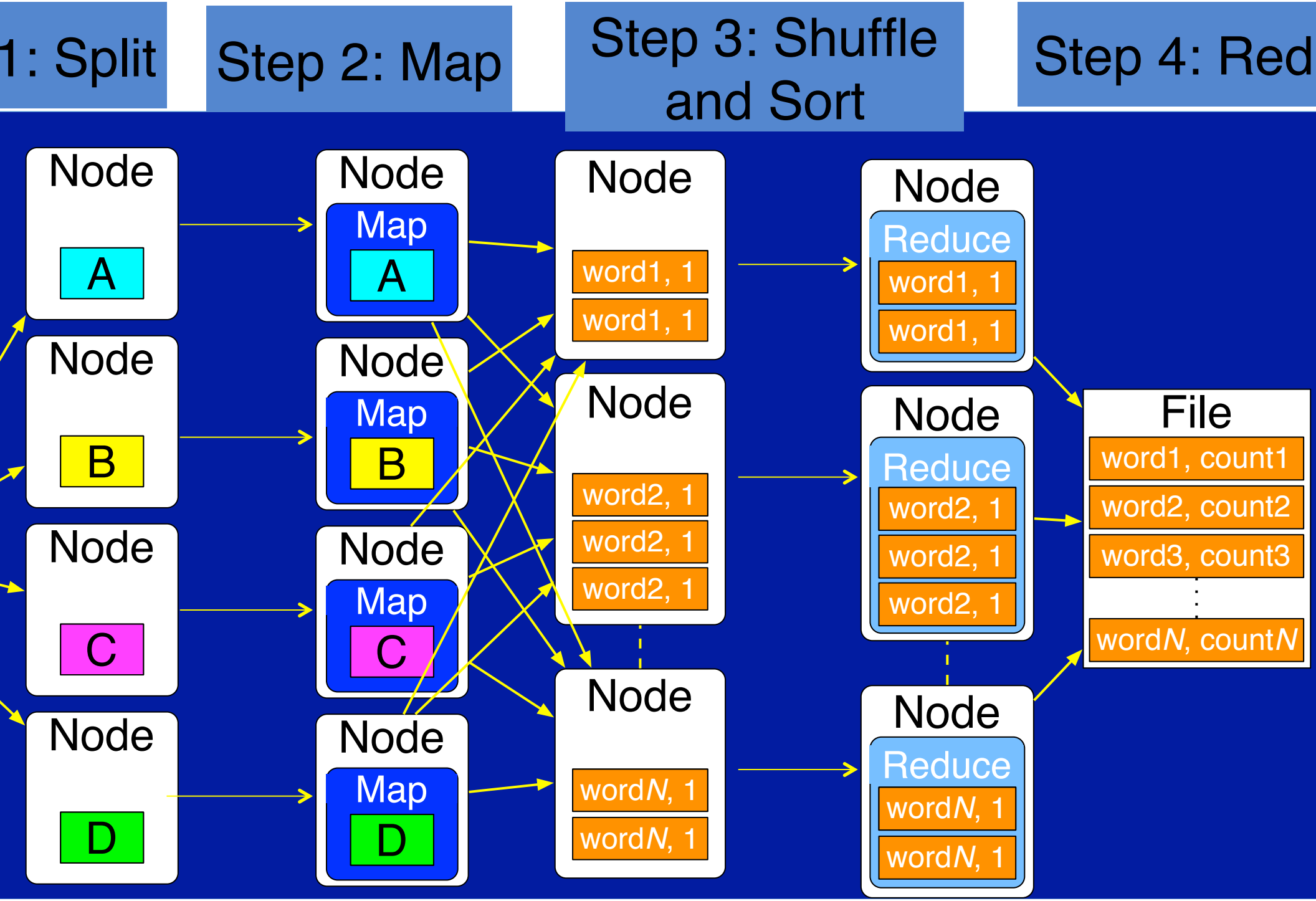


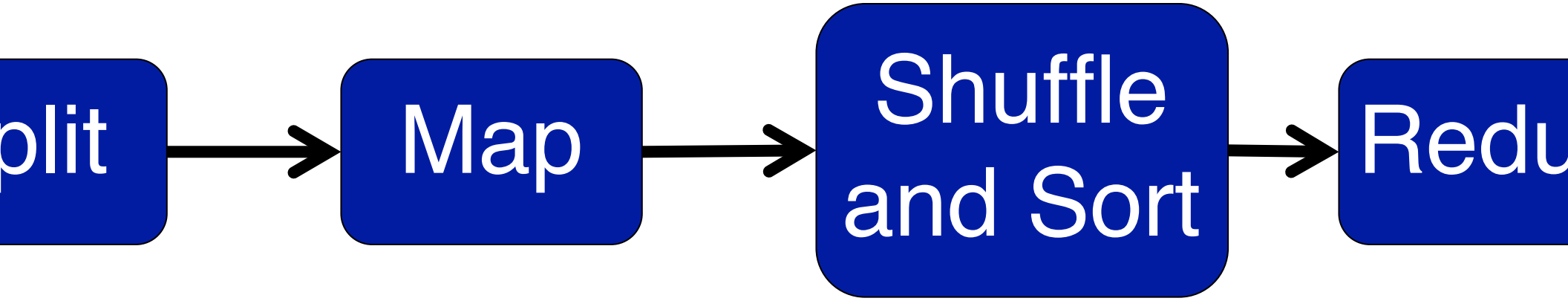




xample MapReduce Application: WordCou



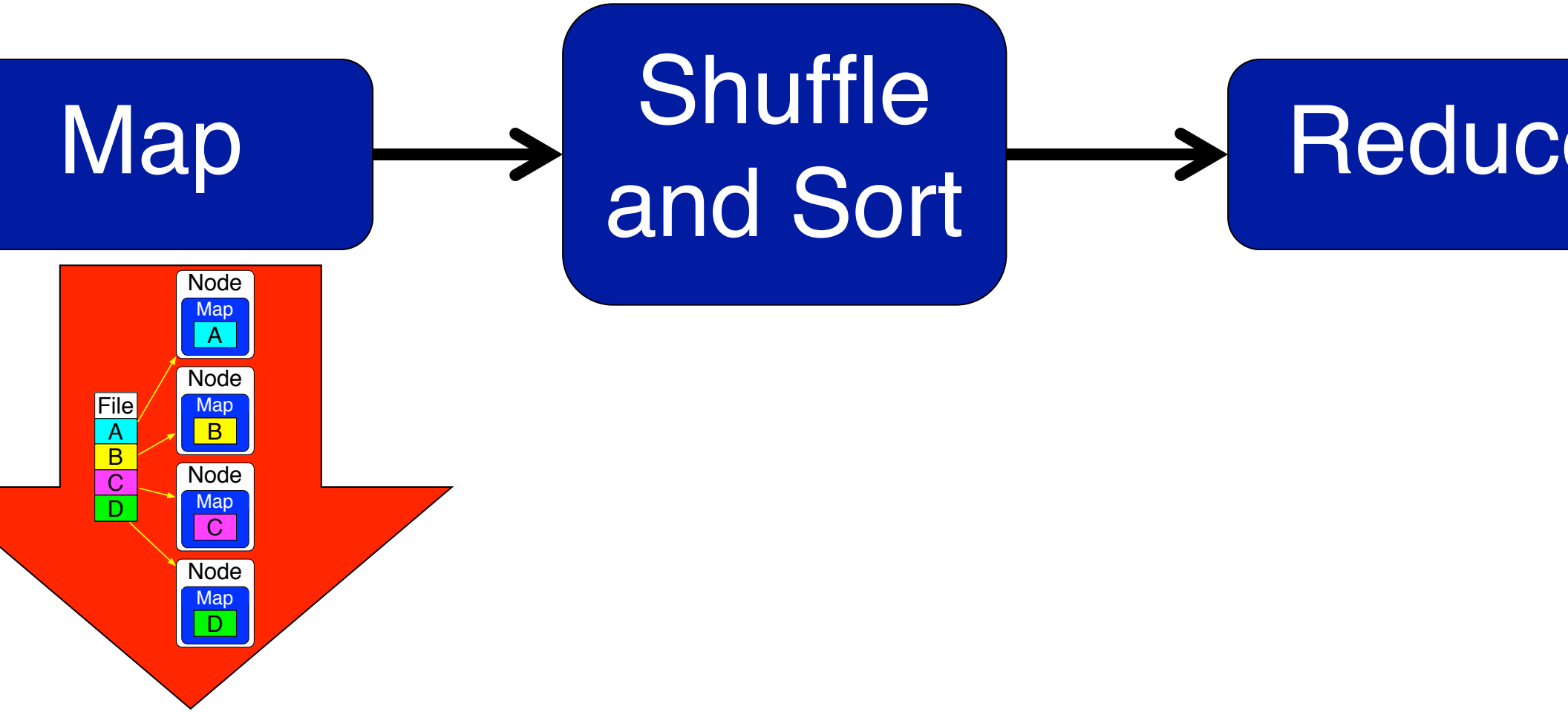

1: Split
Step 2: Map
Step 3: Shuffle and Sort
Step 4: Red
Node
A
Node
B
Node
C
Node
D
Node
Map
A
Node
Map
B
Node
Map
C
Node
Map
D
Node
word1, 1
word1, 1
Node
word2, 1
word2, 1
word2, 1
Node
wordN, 1
wordN, 1
Node
Reduce
word1, 1
word1, 1
Node
Reduce
word2, 1
word2, 1
word2, 1
Node
Reduce
wordN, 1
wordN, 1
File
word1, count1
word2, count2
word3, count3
wordN, countN


plit
Map
Shuffle
and Sort
Redu

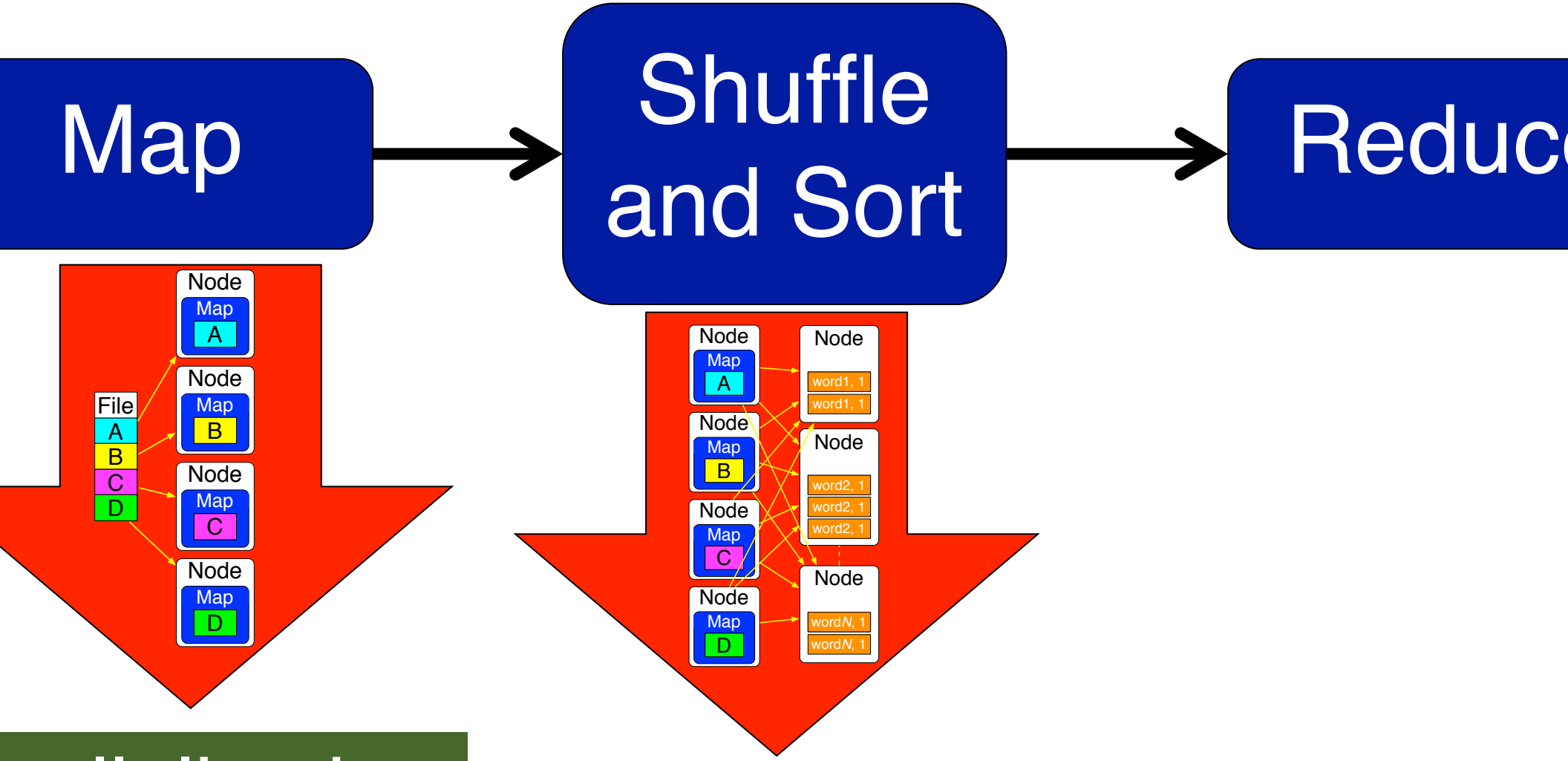
Represents a large
mber of applications.

Big Data Pipelines

Map
Shuffle and Sort
Reduc

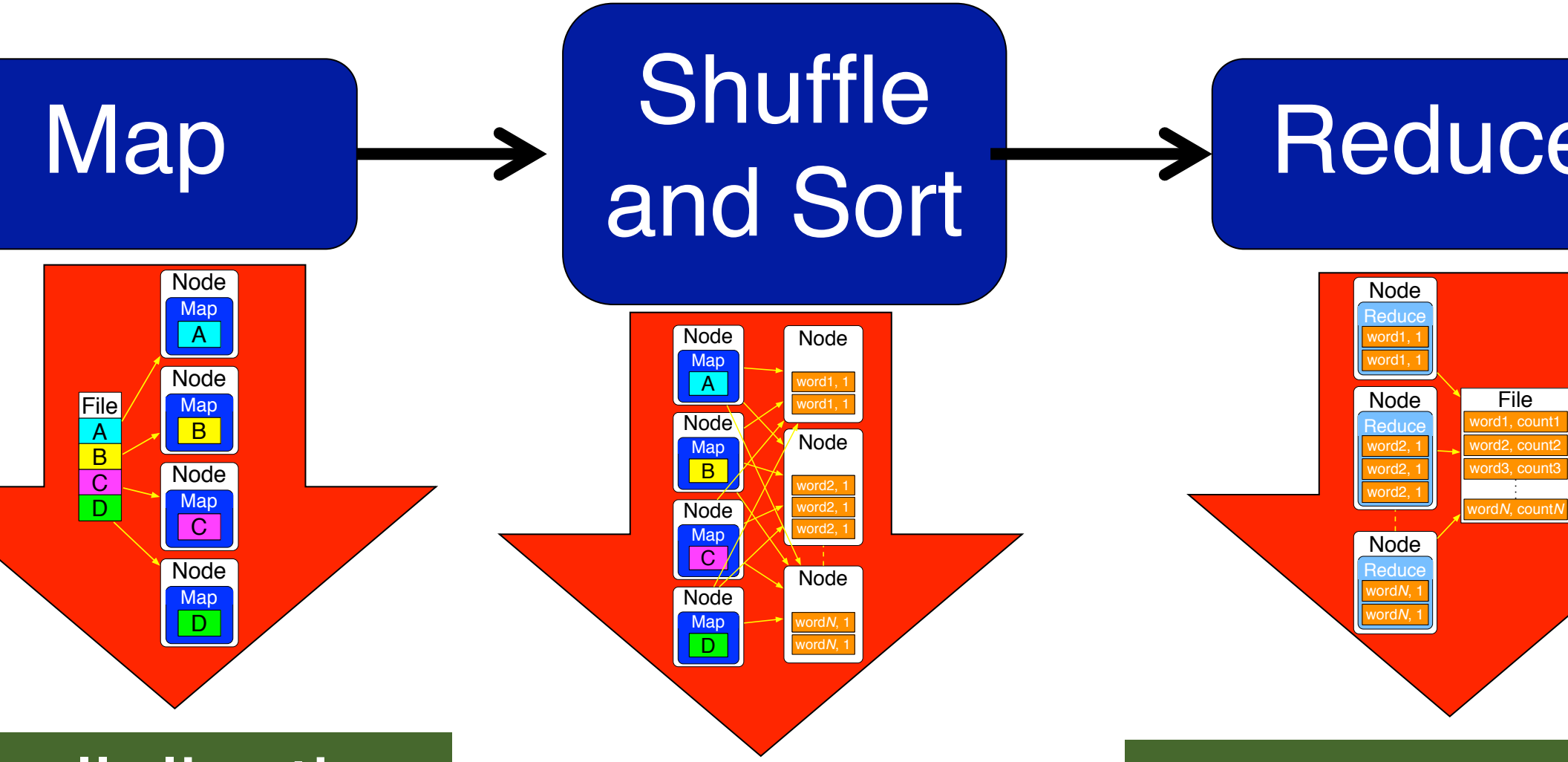


arallelization
ver the input



arallelization
ver the input

Parallelization
data sorting



rallelization
er the input

Parallelization over intermediate data

Parallelizati
over data gro

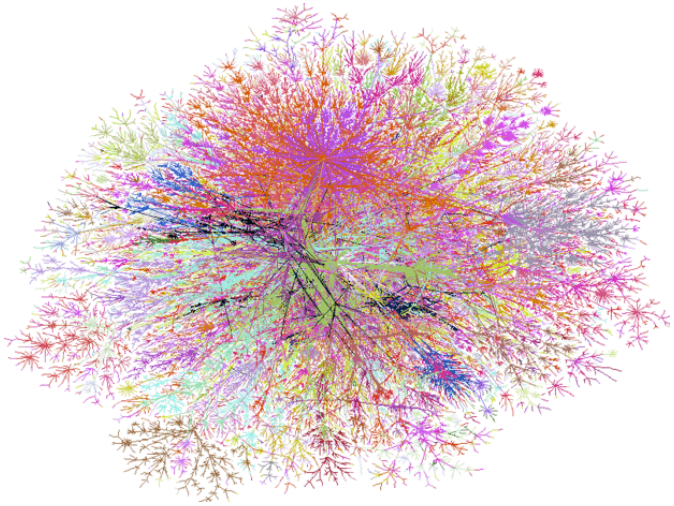
mputing Today has Many Shapes and Si

hadoop

NVIDIA TESLA

amazon web services

COMPUTING AT SCALE



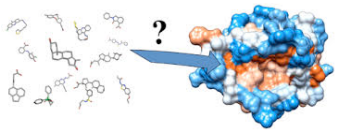
BIG DATA

Requires softw
for dynamic coord
and resource optim

Workflow Systems

Enables dynamic data-driven applications

Computer-Aided Drug Discovery



Smart Cities

Disaster Resilience and Response

anufacturing

Personalized Precision Medicine

Smart Grid and Energy Management

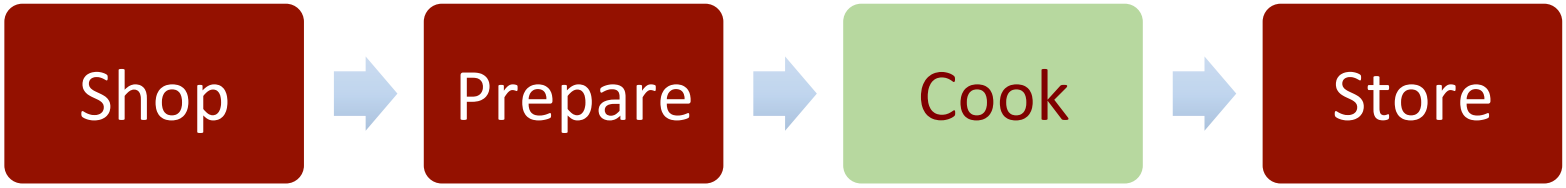
What is a scientific workflow?

*Scientific workflows emerged as an answer to the need to **combine multiple data and computing components** in automated process netwo*

*A scientific workflow is a **series of computational step** that scientists use to generate results. That may involve accessing multiple applications and databases, and processing the data using computationally intensive jobs high-performance clusters.*

So, what is a workflow?

Source:
http://www.fastcodesign.com/1663557/how-a-kitchen-design-could-make-i
easier-to-bond-with-neighbors



et's make pasta this evening!

Shop → Prepare → Cook → Store

30 minutes

30 minutes

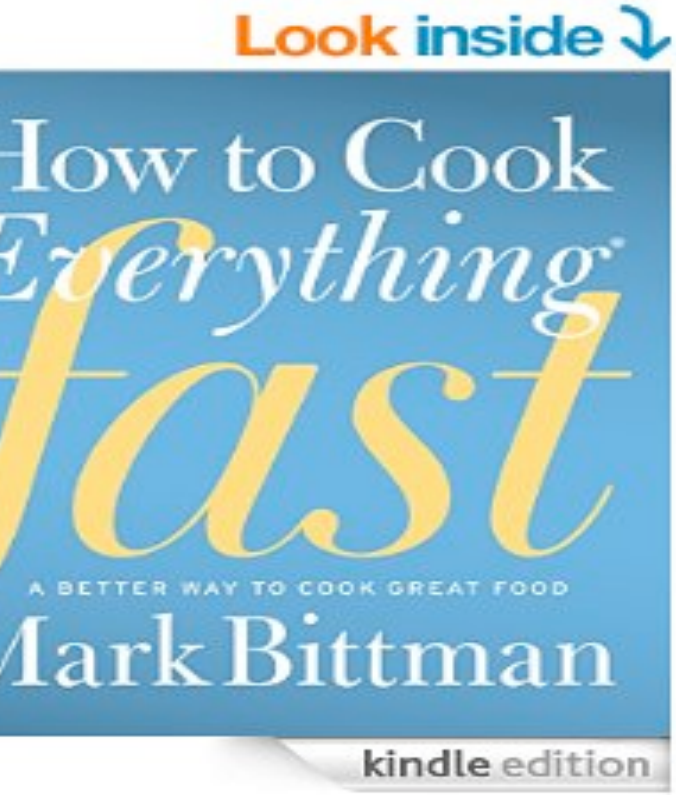
15 minutes

15 minutes

3 minutes

3 minute

How to Cook Everything Fast



"How to Cook Everything Fast is a book of kitchen innovations. ***Time management****— the essential principle of fast cooking— is woven in revolutionary recipes that do the thinking for you. You'll learn how* ***to ta advantage of downtime to prepare vegetables while a simmers or toast croutons while whisking a dressing*** *cook as you read—and let the recipes guide you quickly and easily tow delicious result."*

Image and quote source: amazon.com

What if you have more than one cooks?

MAP

- ***Input:*** *veggies*
- ***User defined function(UDF):*** *chop*
- ***Output:*** *Chopped groups of each kind of veggie*

...

REDUCE

- ***Input:*** *chopped batches for each veggie type*
- ***User defined function(UDF):*** *combine based on veggie type as key*
- ***Output:*** *a bowl of veggies per veggie kind*

nksgiving dinner preparation:
e planning and tasks?

OURTESY OF LOL FOODIE, KING ARTHUR FLOUR, SAVORY SWEET
CIPES, FOOD NETWORK, WHAT WE'RE EATING AND FOODIE TOTS

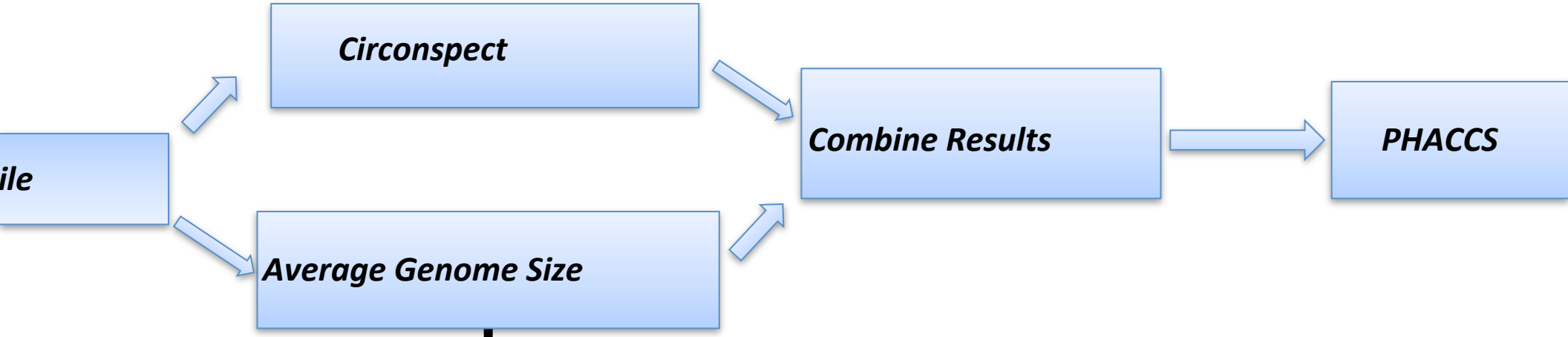
Menu Item	Preparation Time	Cooking Time	Cooling Tim
Turkey	30 minutes	4 hours	15 minutes
Veggies	30 minutes	45 minutes	None
Cranberry Sauce	5 minutes	30 minutes	2 hours
Soup	20 minutes	30 minutes	None
Pie	30 minutes	5 minutes	1 day

- *When do you start cooking?*
- *What order do you cook?*
- *Can you cook some menu items in parallel?*
- *Who cooks what?*
- *…*

- Why are workflows useful in science?
- What are some examples?

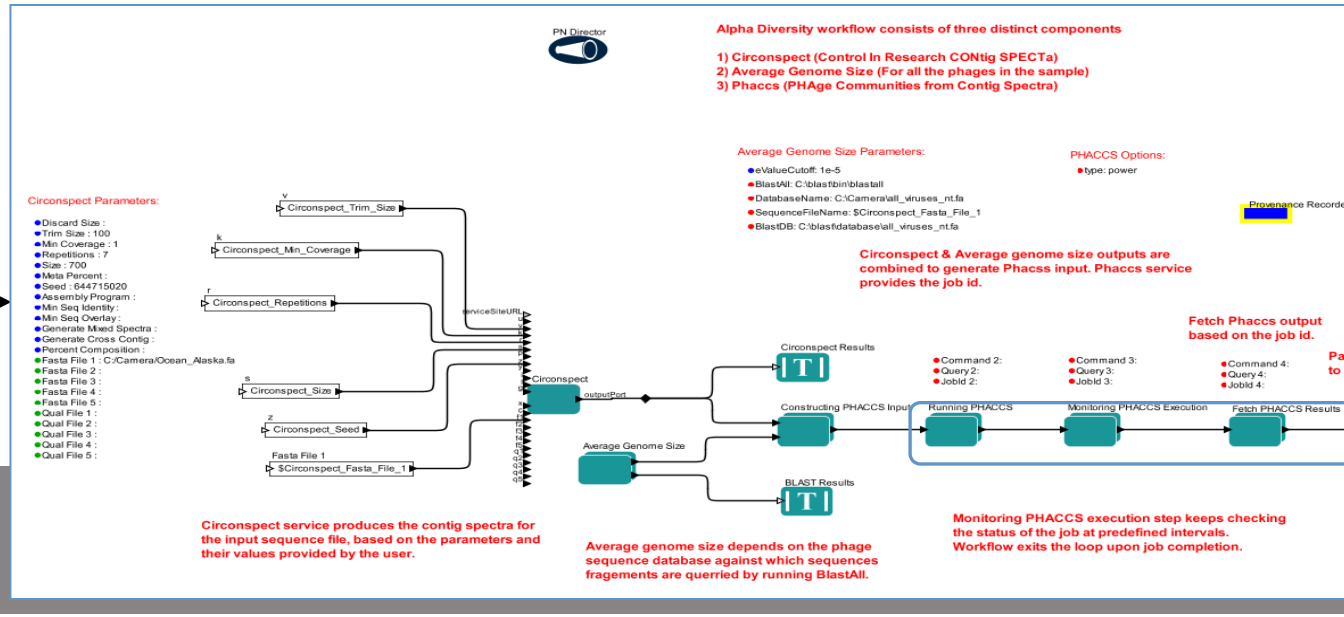
Big Picture is Supporting the Scientist

From "Napkin Drawings" to Executable Workflows

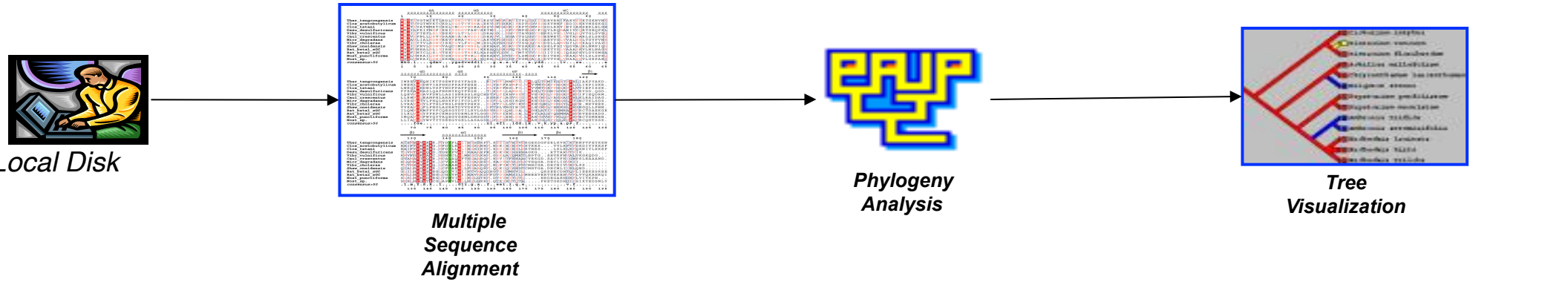


Conceptual SWF

Executable SWF



They can be simple pipelines.



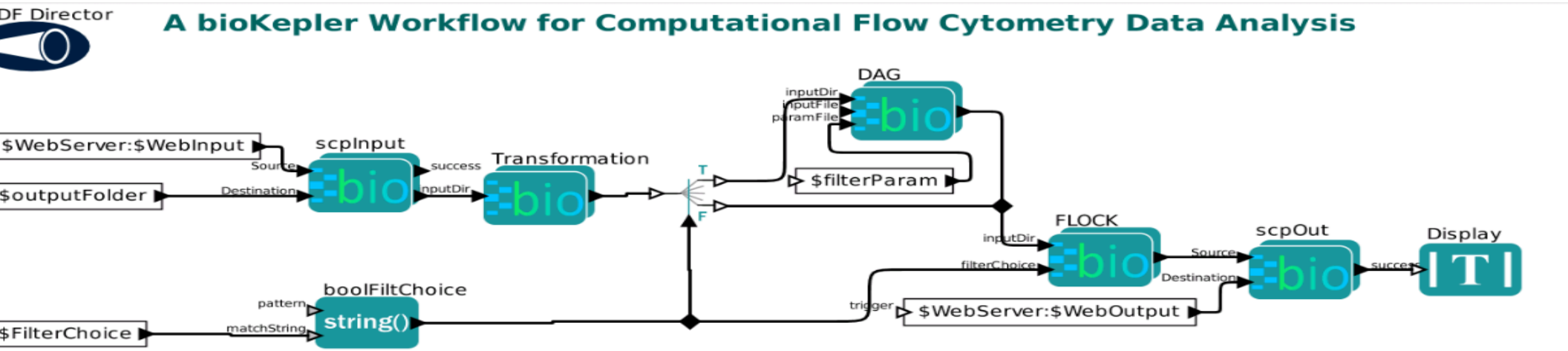
ead the Nexus
as input.

2. Do the multiple sequence alignment on the input data matrix using ClustalW;

3. Construct the phylogenetic tree using PAUP;

4. Read the tree and display it to the user using Foreste

can integrate components to analyze local data.



is workflow implements a computational procedure for automated identification
cell populations from multidimensional flow cytometry data with the FLOCK algorithm.
contains three major steps including Transformation (FCSTrans), Filtering (DAG), and
ustering (FLOCK). The workflow transfers user input data from front-end web portal to the
ckend workflow cloud virtual machine or to the cluster, carries computational analysis
background for multiple input files on Cloud VM or the SDSC Gordon Supercomputer,
d uploads result back to the web portal. The workflow configures or by-pass filter stage
pending user filter choice. The Dynamic Dataflow (DDF) director manages workflow
ecution.

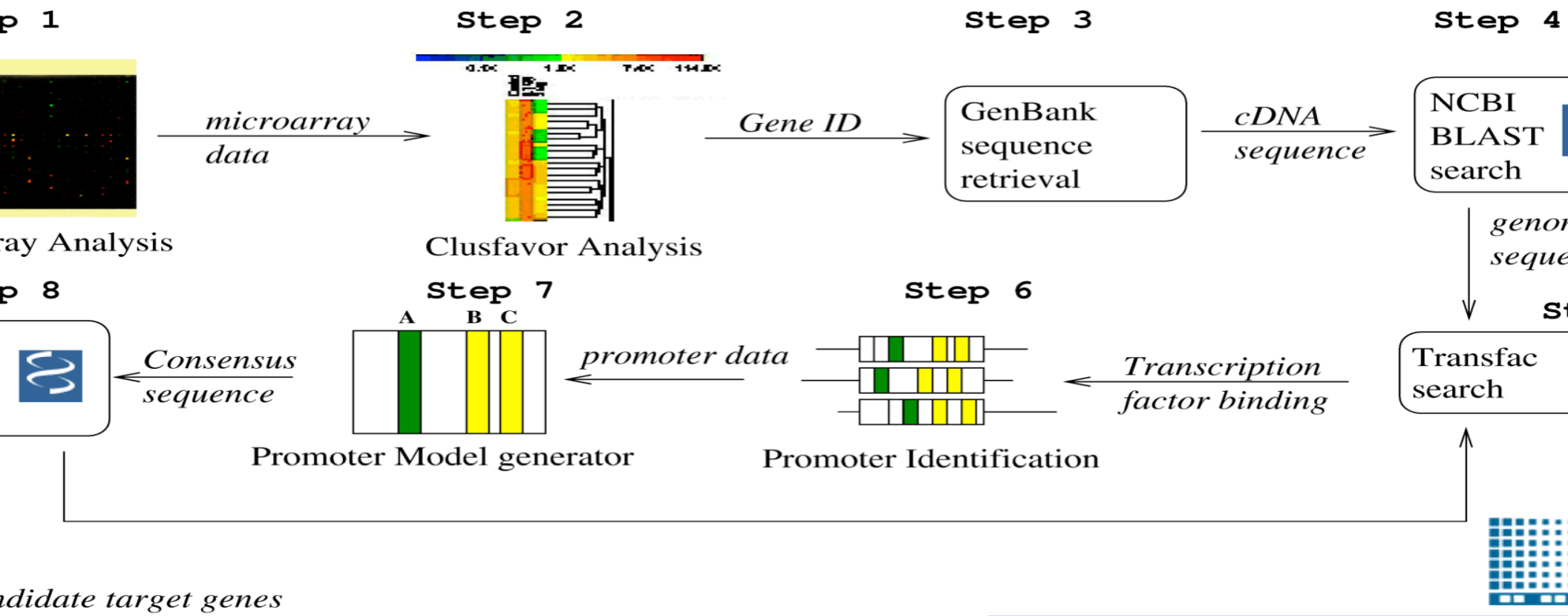
e http://flowgate.jcvi.org

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

ril-25-2014

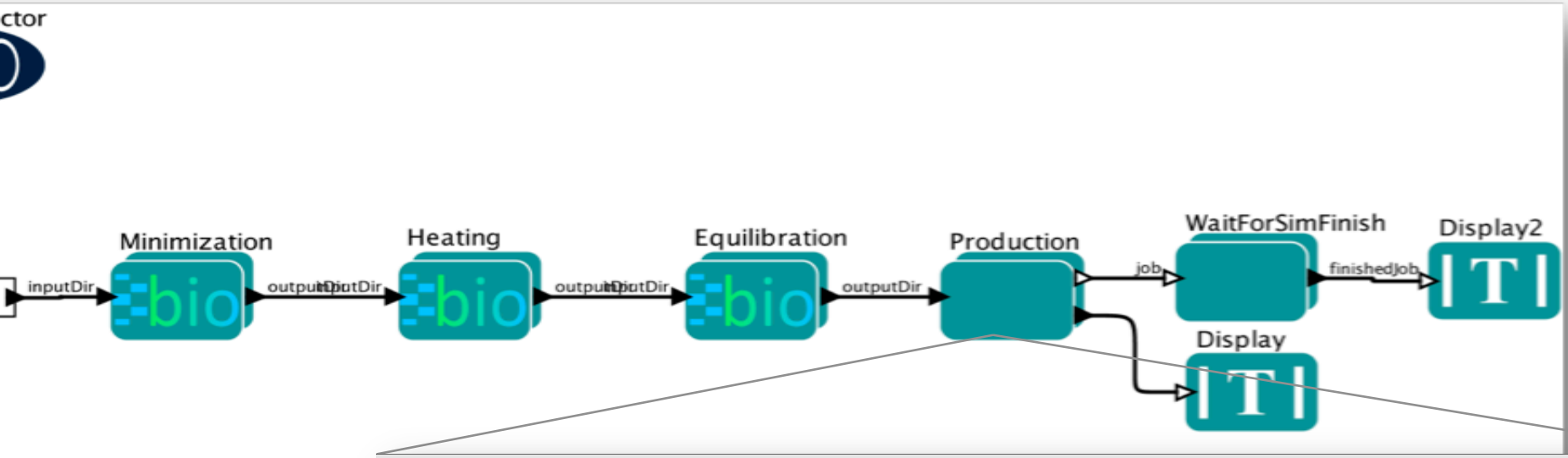
- FilterChoice: True
- filterParam: $toolDir/b_filter_popDef.csv
- outputFolder: /home/spurawat/JCVI/CWDflow
- toolDir: /home/spurawat/JCVI/CDUCModExec
- WebInput: /export/CDUCWebInput/User1_Inp
- WebOutput: /export/CDUCWebOutput
- WebServer: ubuntu@132.249.230.30

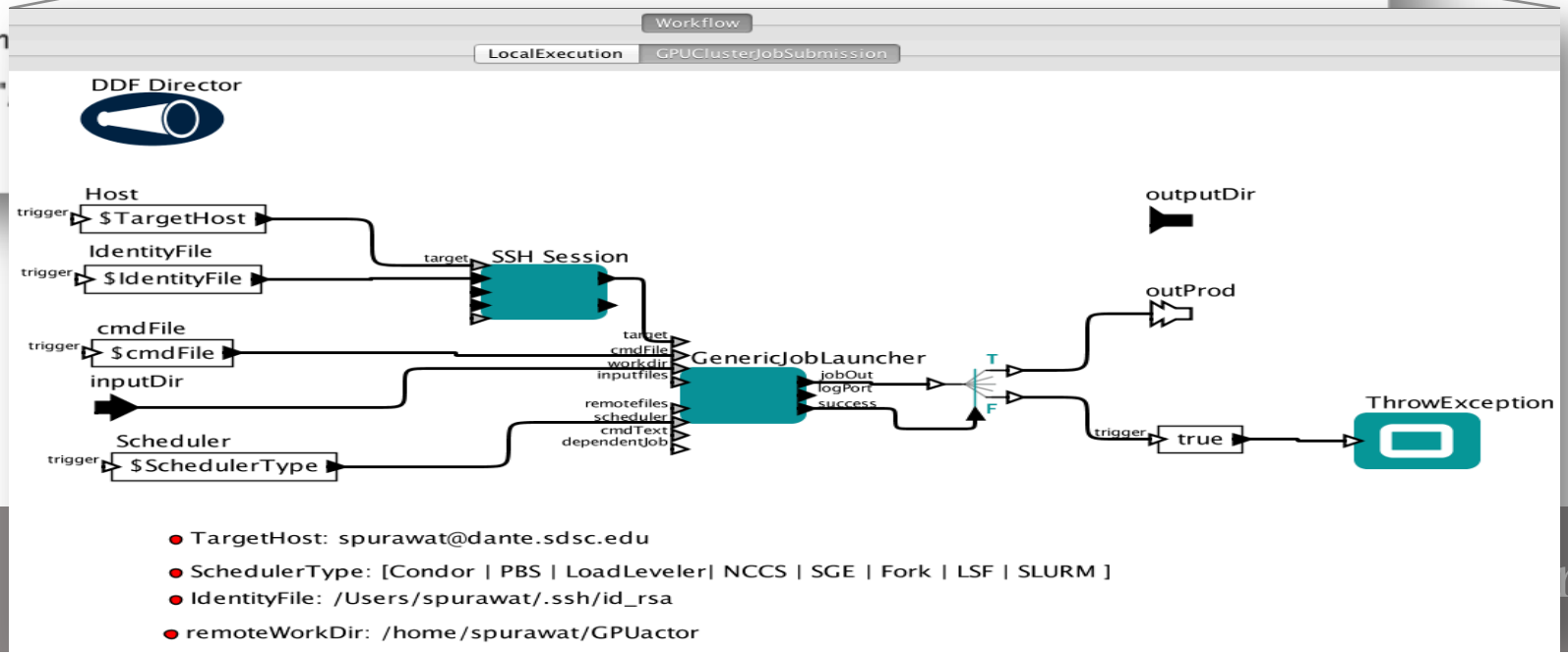
can integrate online tools and databases using com
ramming structures.



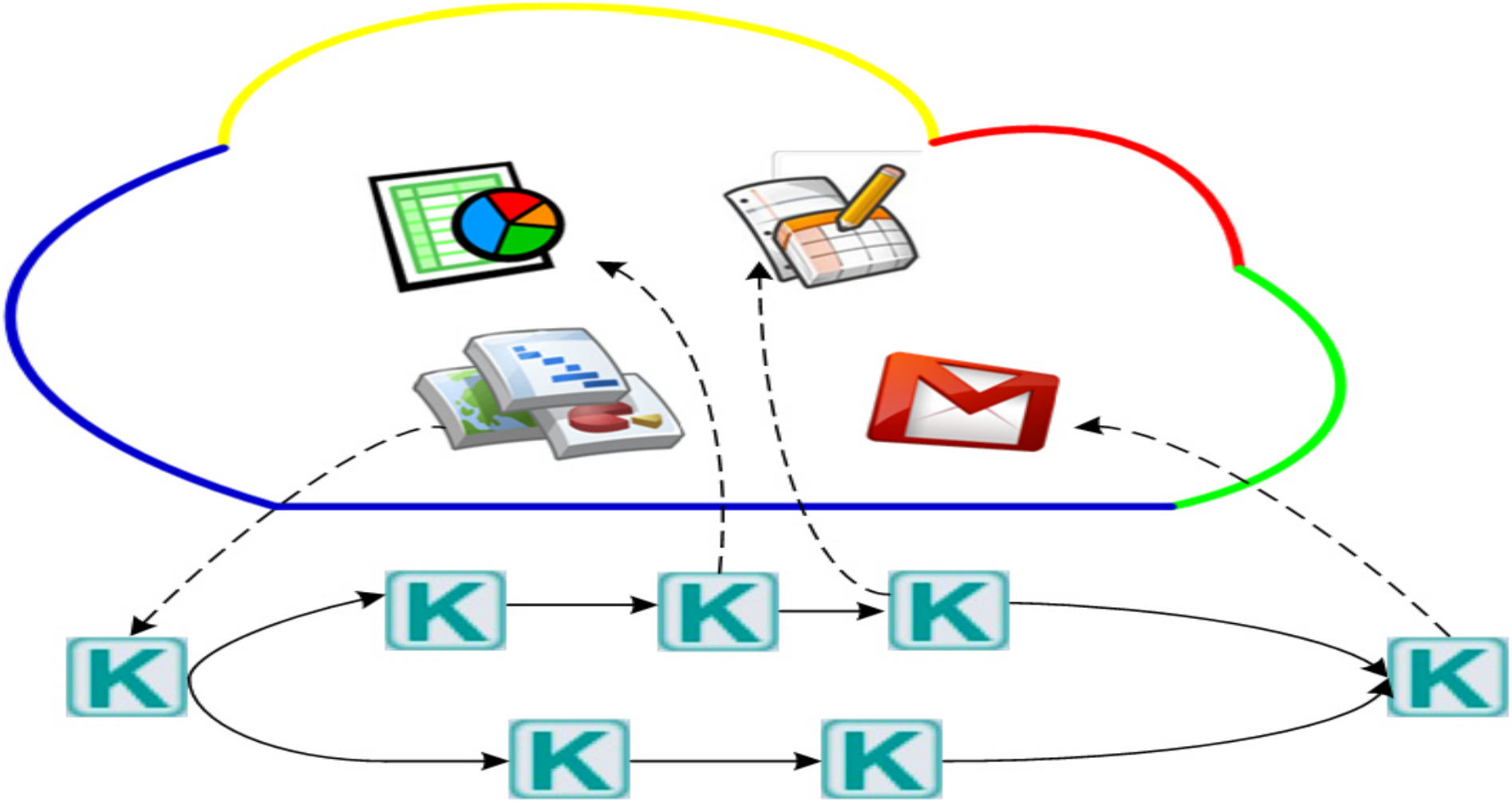
Source: Matt Coleman (LLNL)

ey can have distributed concurrent analysis and dat anagement steps.

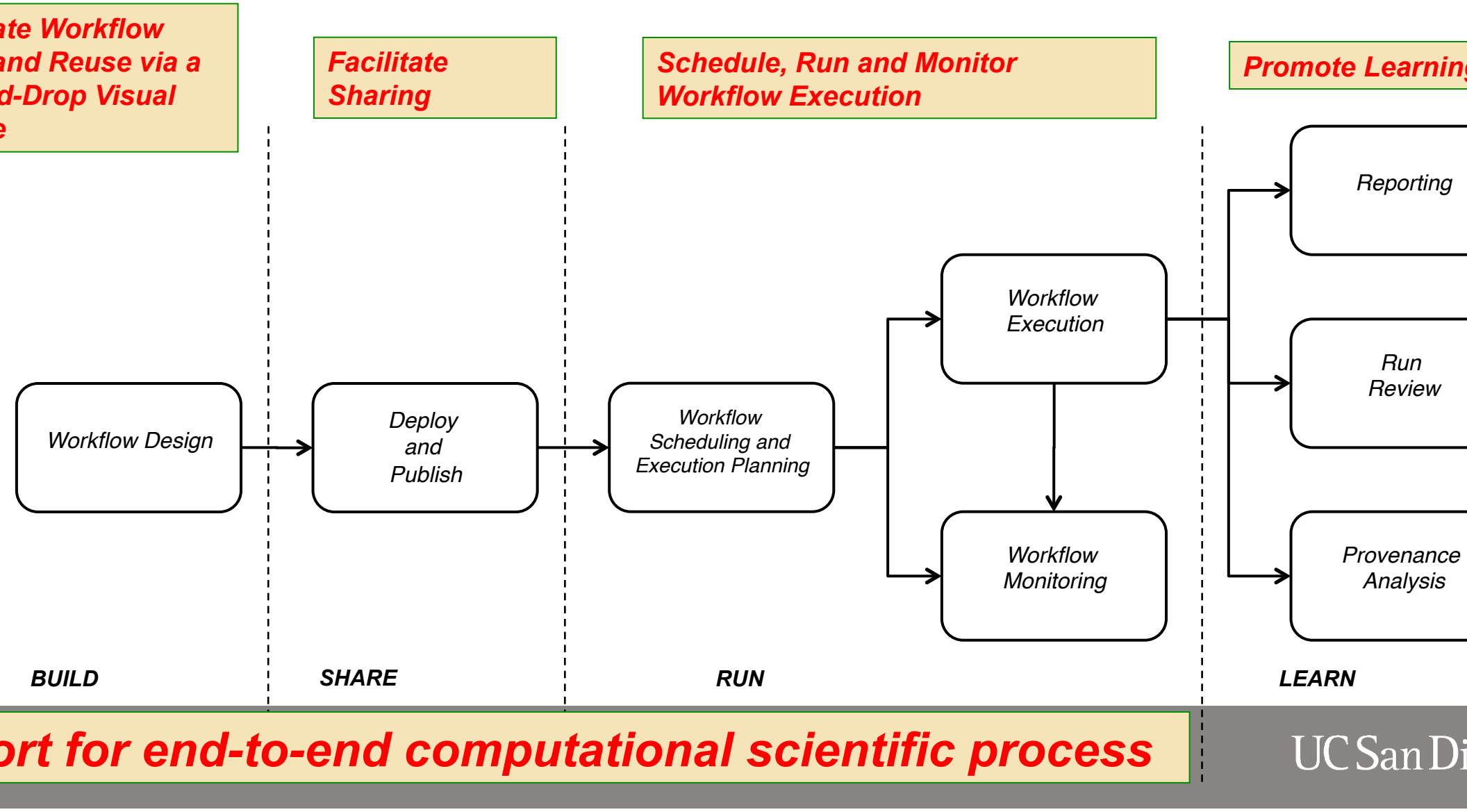




ey can integrate Cloud resources.

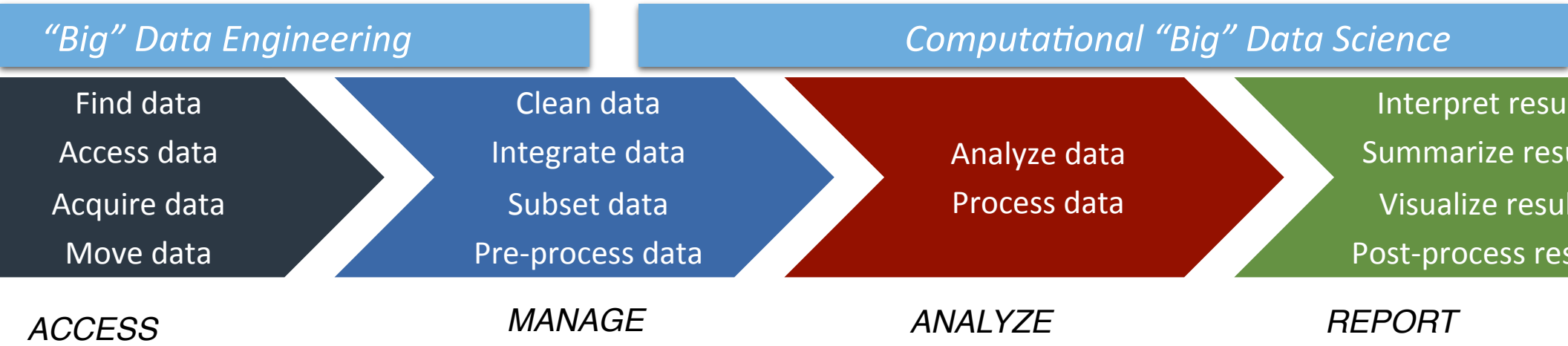


orkflows are a Part of Cyberinfrastructure

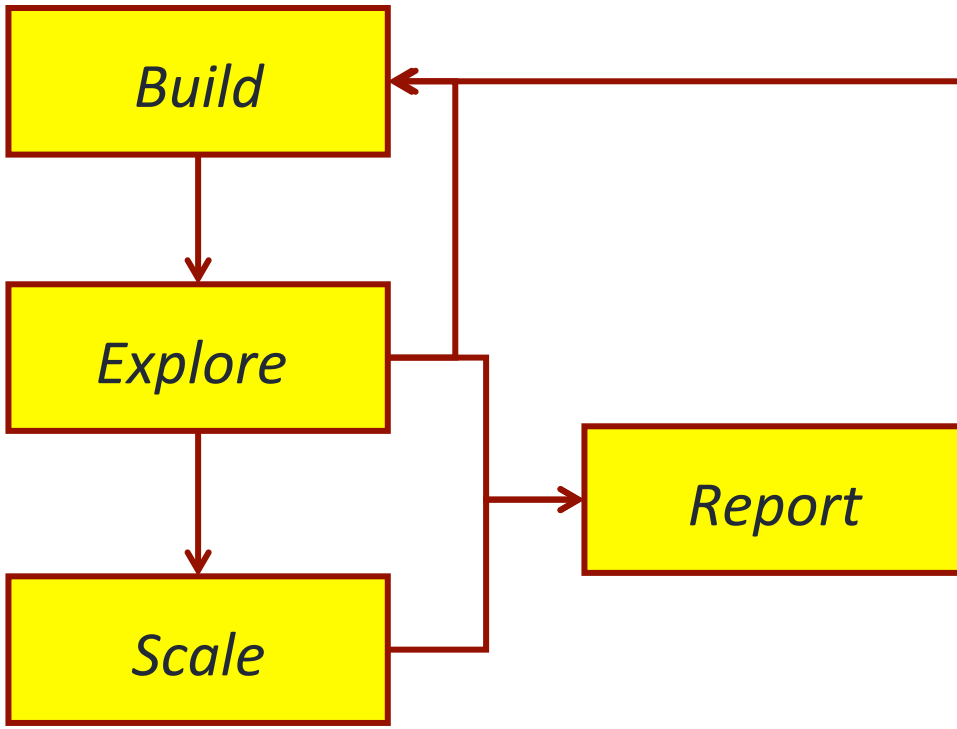


ort for end-to-end computational scientific process

UC San Di



lany ways to look t the process for PC and Big Data… ot every step is utomatable!



Requirements are similar for man domains

-- with slight variations --

tating and Accelerating XXX-Info or
o-XXX Research using Scientific Workflows

lmportant Attributes

Assemble complex processing easily

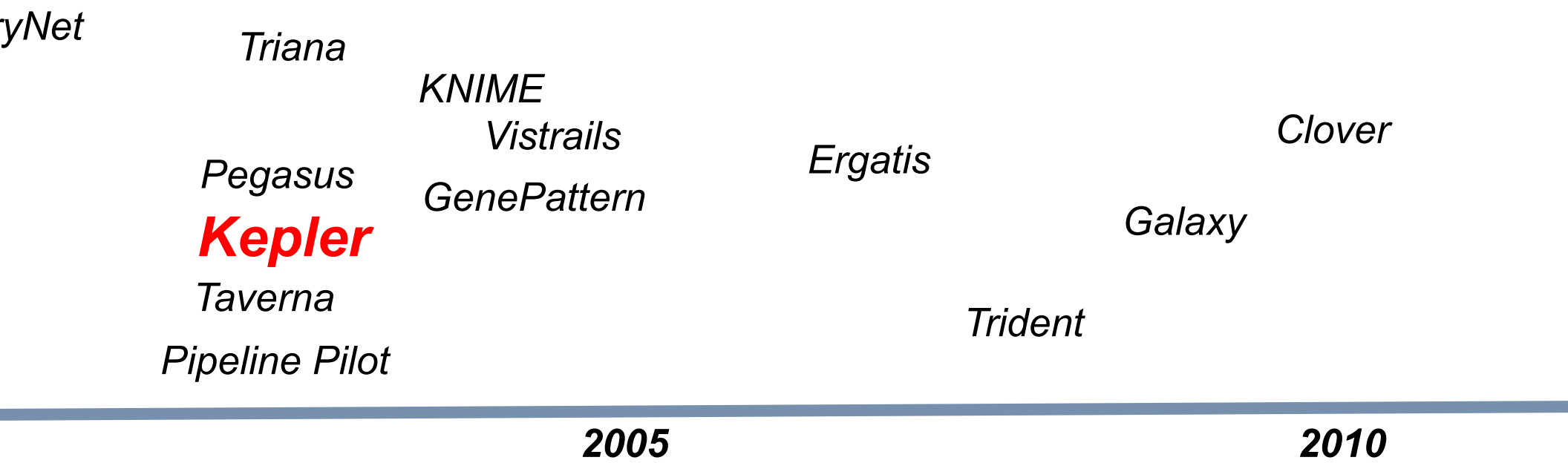
Access transparently to diverse resources

Incorporate multiple software tools

Assure reproducibility

Build around community development model

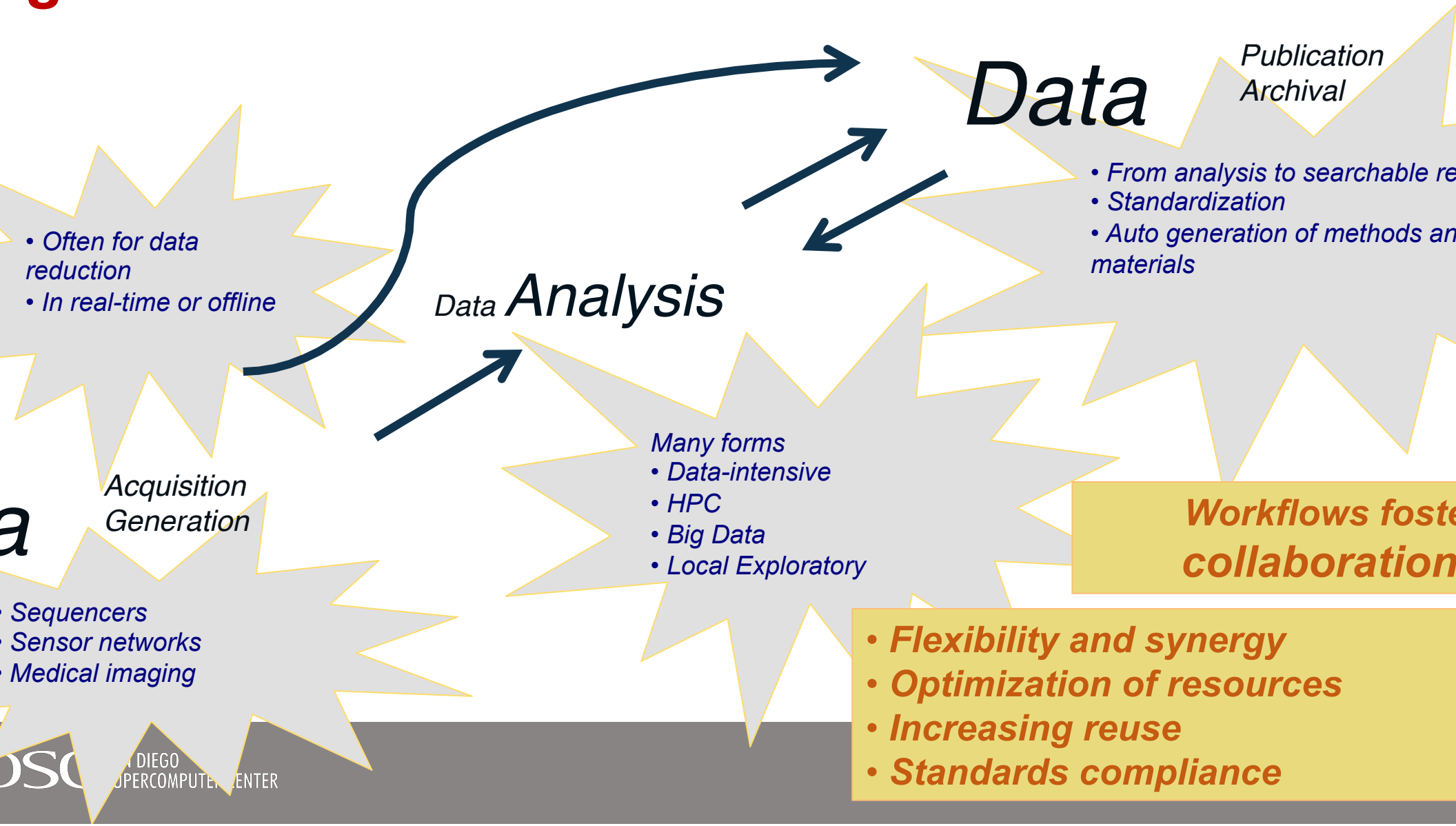
y Scientific Workflow Systems



Kepler

- A diverse library of scientific components and usecases
- Transparent support for multiple workflow engines
- Used by many communities, specialized gateways and individuals

kflows are Used in These Diverse Scenarios in ogical Sciences



Workflows foste collaboration

- **Flexibility and synergy**
- **Optimization of resources**
- **Increasing reuse**
- **Standards compliance**

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!

SDSC Experts Work on Every Part of the Process:

many application as:

- *Molecular dynamics*
- *Hazards resilience*
- *Metro data science*
- *Climate*
- *Bioinformatics*
- *Network analysis*
- …

- "Big" data managemen
- "Big" data engineering
- "Big" data systems
- Statistics
- Graph analytics
- Machine learning
- Workflow managemen
- Data-intensive and high performance computin
- Information visualizatio
- Data provenance
- …

o summarize…

orkflows are used for

- **documentation** of the analysis
- **visual representation** of analytical steps
- ability to work **across multiple software and platforms**
- **distributed data-parallel** scheduling and execution
- **reproducibility** of a given project with little effort
- **reuse** of part or all of a workflow in a different project

eathe!

https://www.youtube.com/watch?v=Qs-KaOKsxDU

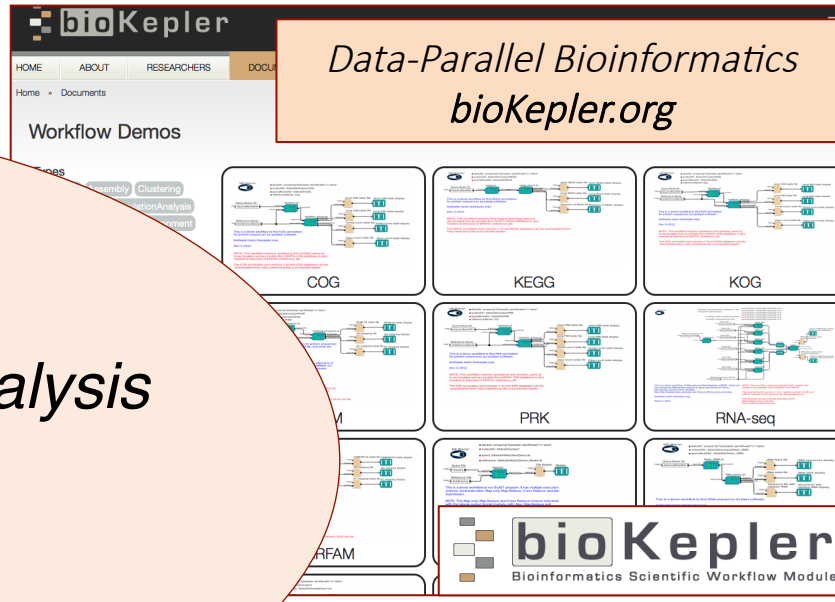
#HPCmatte

Computational Data Science Workflows

- Programmable, Reusable and Reproducible Scalability -

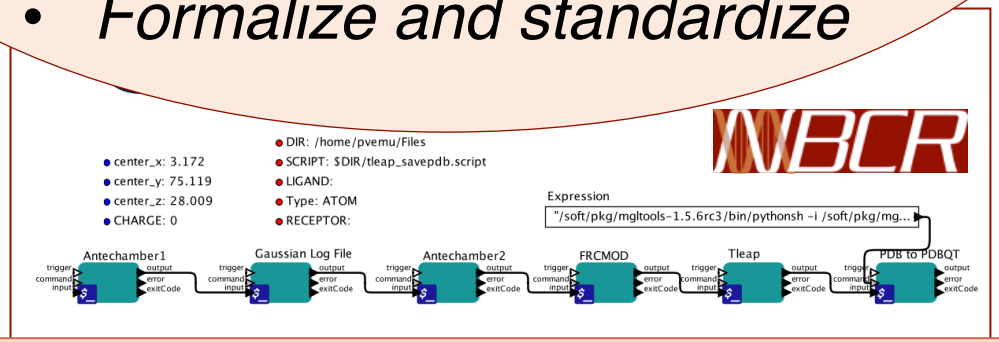
eal-Time Hazards Management
wifire.ucsd.edu

Data-Parallel Bioinformatics
bioKepler.org



WorDS Center

- *Access and query data*
- *Scale computational analysis*
- *Increase reuse and reproducibility*
- *Save time, energy and money*
- *Formalize and standardize*



oject.org

WorDS.sdsc.ed

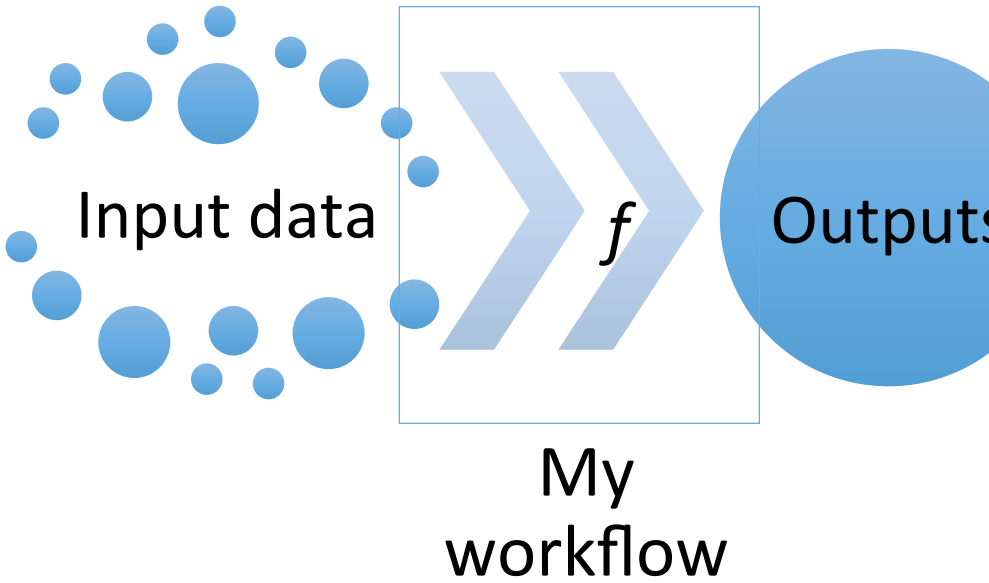
Scalable Automated Molecular Dynamics and Drug Discovery
nbcr.ucsd.edu

How Do I Start Conceptualizing a Computational Workflow?

Start with the Workflow As a Blackbox

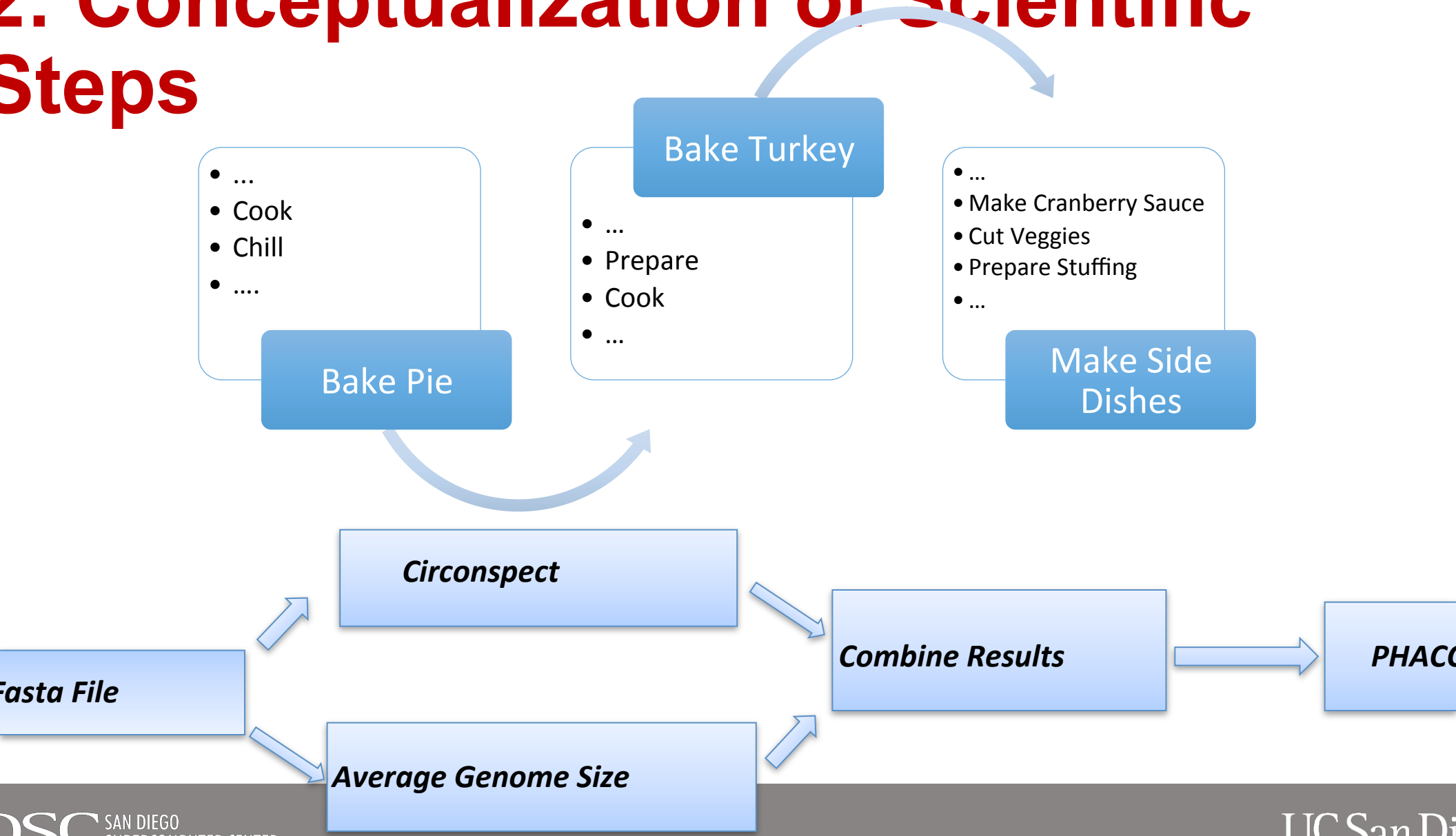
reat the whole workflow as a
lackbox

- What is the usecase/application?
  - What is the science question this workflow is solving?
- What is the input data?
- What are the expected outcomes?



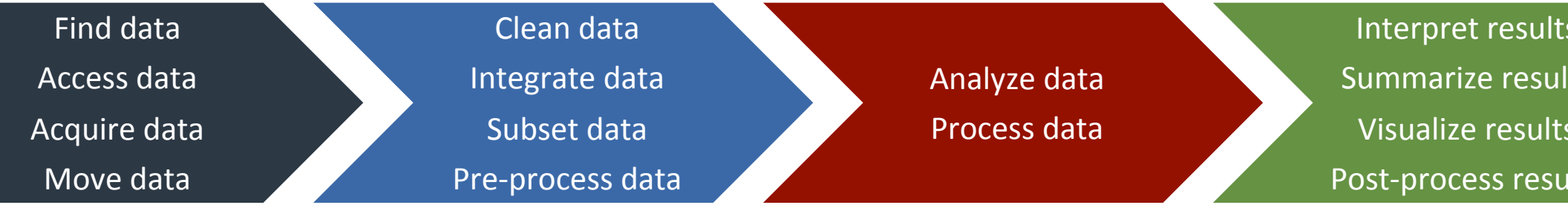
ive the workflow a title based on initial assessment!

2: Conceptualization of Scientific Steps



3: Treat Each Step Like a Workflow

- until you reach an atomic functional step -



SHOP ***PREPARE*** ***COOK*** ***STORE***

Some questions to ask:

- *Where and how do I get the data?*
- *What format is the data in?*
- *How do I integrate or subset datasets?*
- *How do I analyze the data and what is the analysis function?*
- *What are the parameters to customize each step?*
- *What are the computing needs to schedule and run each step?*
- *How do I make sure the results are useful for the next step or as scientific products?*

Searching for Data

- You need to know:
  - What you are looking for
    - Some data formats: FASTA, FASTQ, PHYLIP, PDB, PFAM, SAM, BAM,…
    - Meaning of data: how it scientifically/conceptually relates to the analysis you are implementing.
  - Where to find what you are looking for
    - Searchable data archives/repositories
    - A database system
    - Sensor networks or other scientific instruments

Moving Data

- **Data access technologies:**
  - ftp, scp, web services, …
- **Data querying technologies:**
  - rbdms, opendap, …

re-processing data

Clean: Reformat, QA/QC, duplicate removal, etc.

Divide/chop/slice-n-dice: query, subset, filter, …

Transform: fasta->fastq; .fcs->.txt

Reduce: statistical operations, and

Integrate: combining data from previous steps

p 4: Start Building Each Step Including the Alternativ

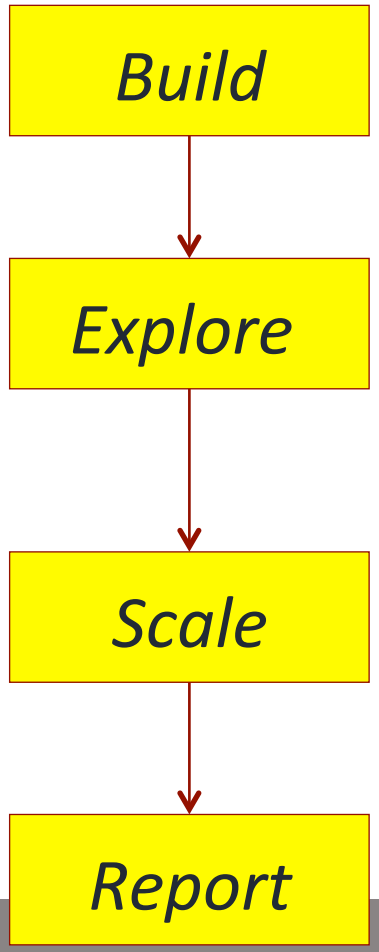
tegration of Many Tools to Serve a Purpose --

ernative tools

ultiple modes of scalability

pport for each step of the velopment and production ocess

ferent reporting needs for oloration and production ges



Need toolboxes with tools for:

- *data access,*
- *analysis,*
- ***scalable executio***
- *fault tolerance,*
- *provenance trackir*
- *reporting*
- *...*

d 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

Step 5: Save and Share Reports and Final Products with your Team

- Computational data science is in the middle bridging the gap between data and insight
  - → Computational Data Scientists tackle challenging questions and define the steps to achieve the results as a workflow
- Developers/computer scientists use their favorite tools to implemen the methods in the workflow
- The process is kept sharable, standardized, scalable and accountable

Why 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 Helps with Accountability and Reproducibility

- *Chronology of ownership, custody, or location of historical object* – Wikipedia
- Data and Process Provenance
  - Inputs, outputs, intermediate results
  - Workflow: actors, links, parameters, etc.

Why is Provenance Useful?

- Keep the association of results to processes
- Make it easier to validate/regenerate results and processes
- Enable comparison between different workflow versions
- Smart re-reruns
- Failure recovery

The Need for Provenance Interchange

- Applications and Tools:
  - Workflow systems
  - Databases
  - Programming Languages
  - File Systems
- Vocabularies and Ontologies:
  - Open Provenance Model (OPM)
  - Dublin Core

PROV

- Set of W3C recommendations
- Finalized September 2013
- Core concepts
  - Identifying objects, processes, etc.
- Conceptual data model & serializations:
  - OWL2 ontology
  - XML schema
  - Human-readable notation

Example Provenance Questions

- How long did workflow run *n* take?
- What parameter values were used in run *n*?
- What actors does workflow *m* have?
- How long does a given actor usually take to run?
- What workflows did a specific user run on a given date?

Sharing Process and Data Products

re are many reasons for sharing workflows

Collaborative scientific research

- Inform collaborators of your progress
- Formalize and standardize application development
- Increase reuse and repurposing of existing workflows
- Run collaborators' workflows on similar infrastructure

Reproducible publications

- Rerun/replay workflow executions that generated the published pro
- Validate outputs

Sharing Workflow Results

Version Control Repositories

- e.g., Git

Cloud File Hosting Services

- e.g., Dropbox, Google Drive, Microsoft OneDrive
- File locker
- Synchronization across devices
- Share files and directories

SeedMe

- Web platform to share results in near real-time
- Specialized viewers for plots, sequences, videos, etc.

Create an iPython notebook to share the execution history, input parameters and results

HAPPY ENDING?

re are no happy endings.
lings are the saddest part,
ust give me a happy middle
d a very happy start.

İlkay Altintas
altintas@sdsc.edu