University of Pittsburgh

Tutorial
Workflow Systems: OCCAM

Big picture, workflows, examples, demo

Bruce Childers

Associate Dean for Strategic Initiatives
Professor of Computer Science
School of Computing & Information
childers@pitt.edu
http://www.cs.pitt.edu/~childers

Modeling the World's Systems, May 22, 2018

If a tree falls in a forest and no one is around to hear it, does it make a sound?

If you create a model and use it in an experiment, and nobody can access it (& repeat!), does it exist?

nature

Menu

NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

70% failed replication trial (50% of their own!); 45% due to data/code, 80% contributed

The New York Times

RESEARC

Estir

psyc

Open Scie

+ See all a

SCIENTIFIC AMERICAN

NOVANEXT

Why Should Scientific Results Be Reproducible?

Climate Science Can Be More Transparent, Researchers Say

omplicated by the large volume

open data

Virginia Gewin

Nature **529**, 117–119 (07 January 2016)

COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH

Home / Magazine Archive / March 2016 (Vol. 59, No. 3) / Repeatability in Computer Systems Research / Full Text

CONTRIBUTED ARTICLES

Repeatability in Computer Systems Research

The National Academies of SCIENCES ENGINEERING MEDICINE

BOARD ON BEHAVIORAL, COGNITIVE, AND SENSORY SCIENCES

Weak repeatability for 32.3% of 601 papers in computer systems research

PNAS Proceedings of the National Academy of Sciences of the United States of America

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund, Thomas E. Nichols, and Hans Knutsson

PLOS

Why Most Pub

se

John P. A. Ioannidis

Path to the ~~Emerald~~ *Model* City

I. Community and culture

- Galvanize, **incentivize**, & **educate**
- Review, mandates, **funding**, expectations
- Governance, policy & procedures
- Work with **existing practices to push boundaries**

II. Technology

- Ease effort to **accelerate research**
- **Leverage** for higher quality, new discovery
- **Path of least resistance** is “right one”
- Work with **existing methods to push boundaries**

Technology: Workflow Systems

A place to do experiments

- Computational experiments (model + data + compute)
- Setup your own, or use one that is available

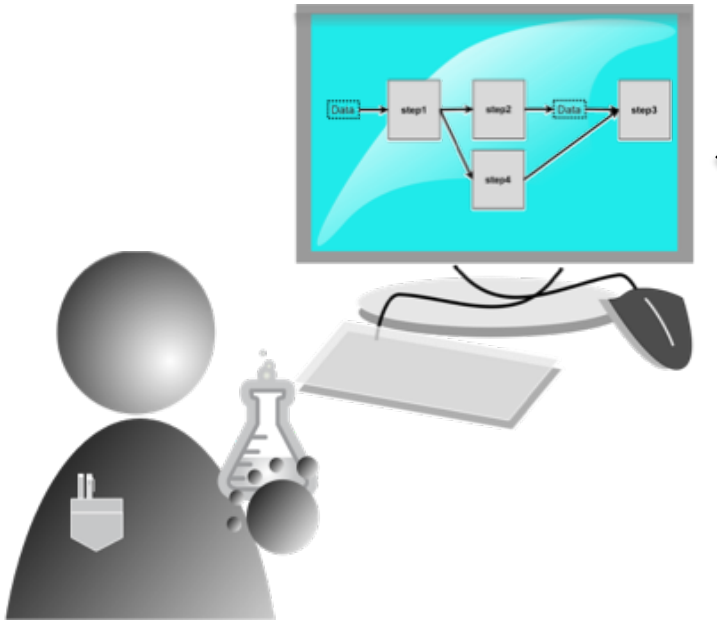
Accelerate your research

- Focus on research rather than the infrastructure
- Leverage shared data, models, & other resources
- Capture & document experiment (FAIR)

Workflow System

Workflow System

(1) Capture



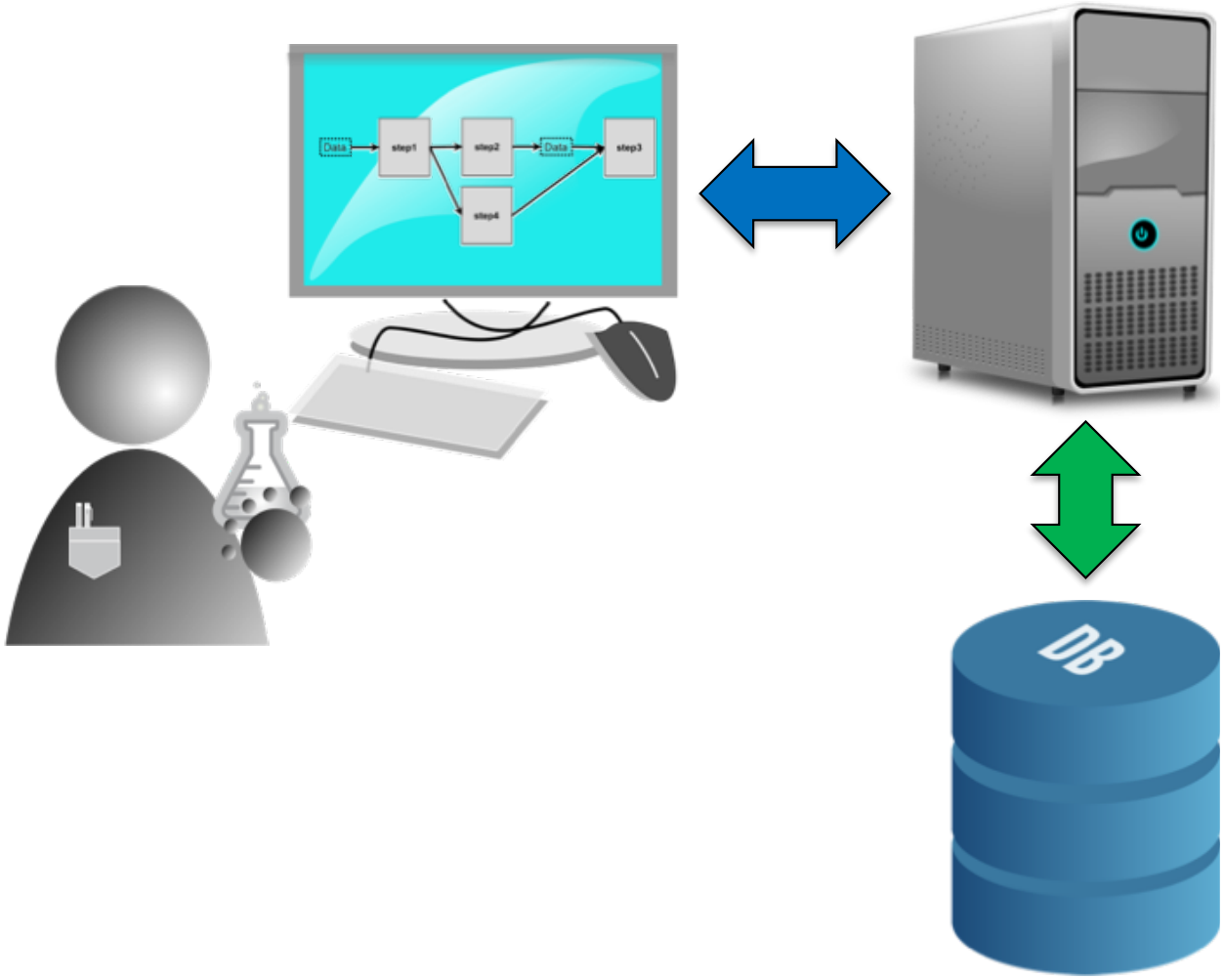
Front-end User Interaction

- Create & edit workflow
- View & manipulate results
- User workspace

Workflow System

(1) Capture

(2) Engine



Middle-End Management

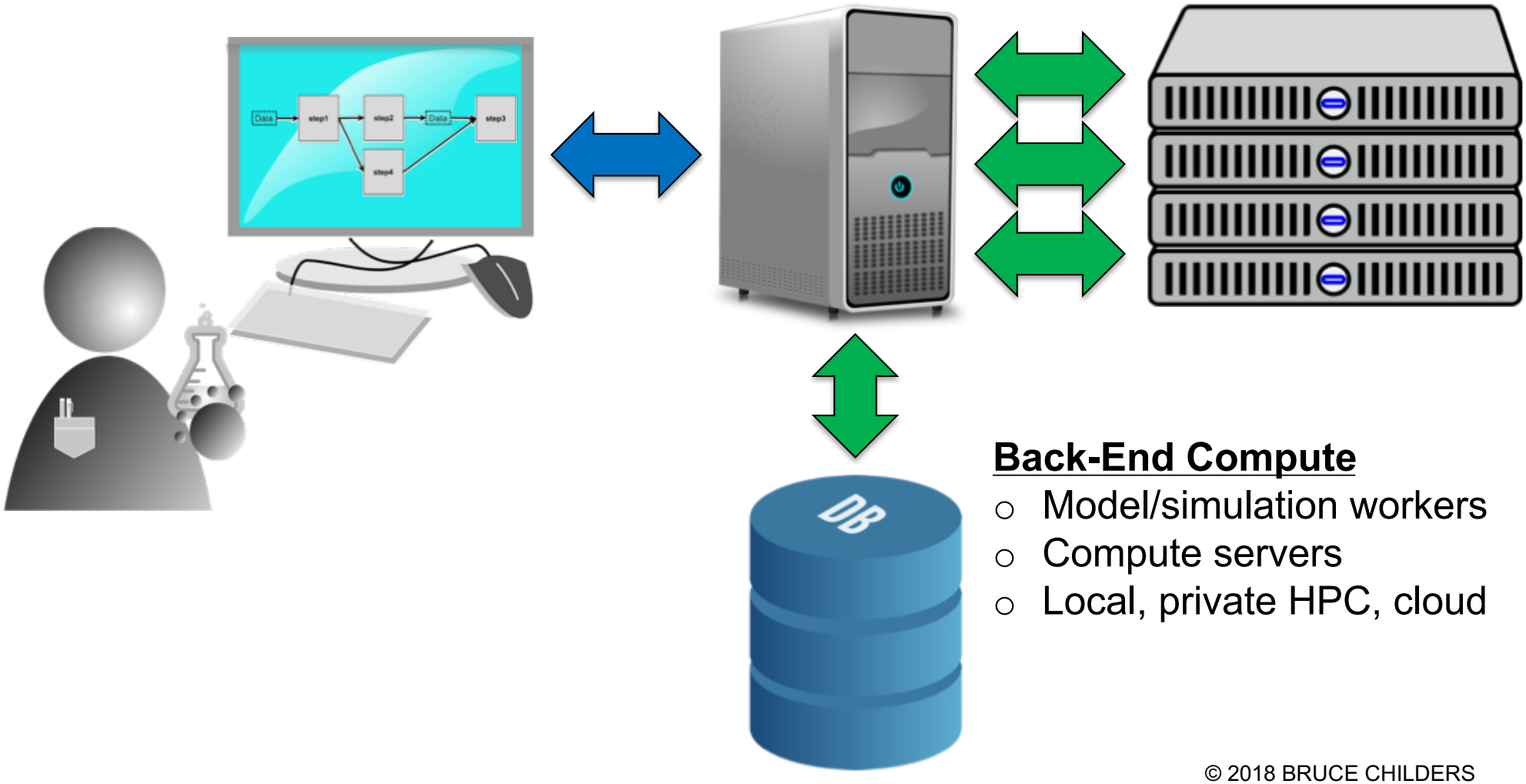
- User interface
- Generate & dispatch
- Repository (database)

Workflow System

(1) Capture

(2) Engine

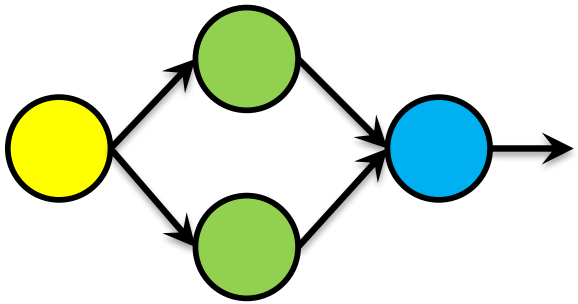
(3) Workers



Back-End Compute

- Model/simulation workers
- Compute servers
- Local, private HPC, cloud

Workflow



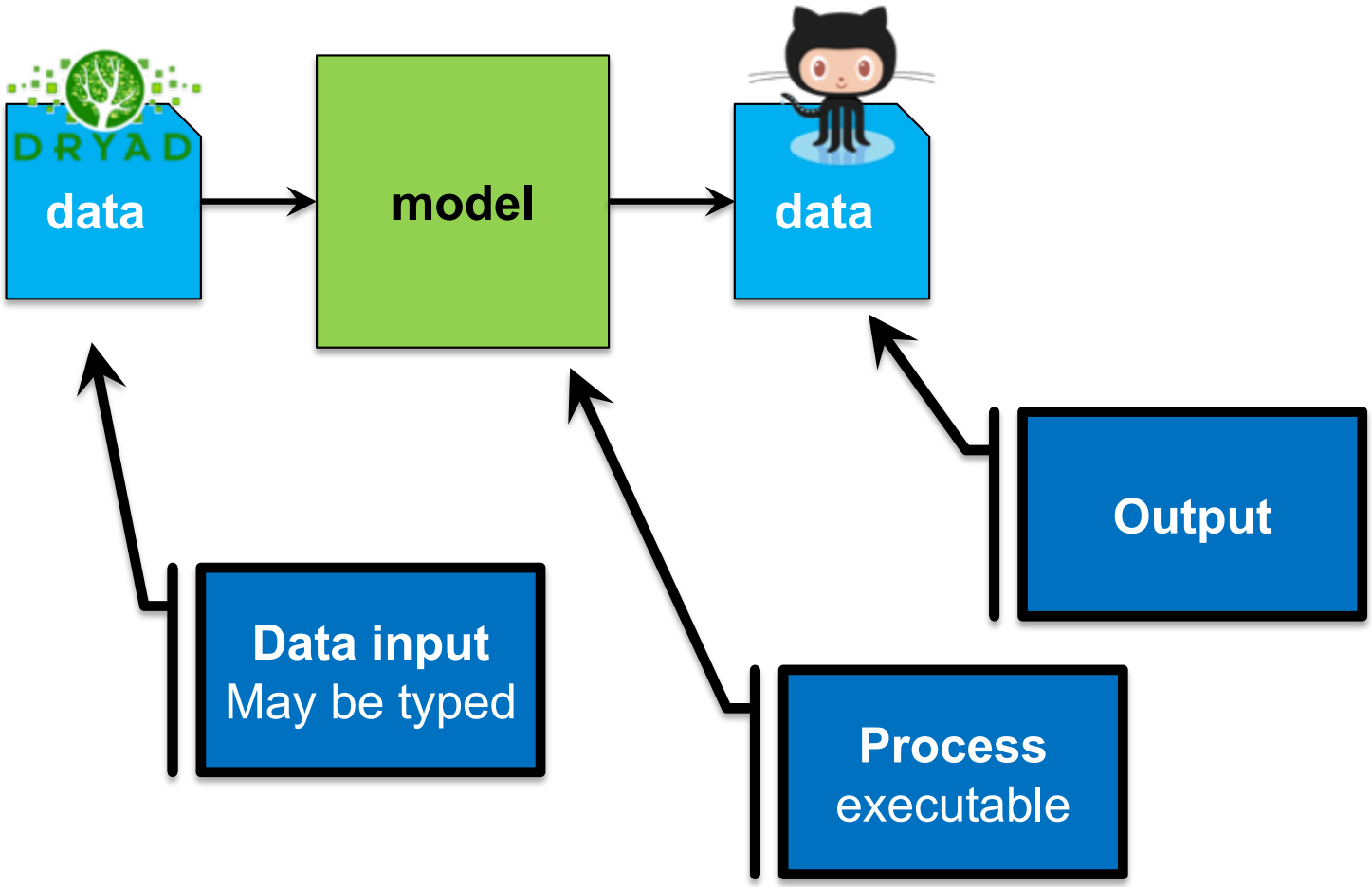
Sequence of computational steps

- ***Directed acyclic graph*** (DAG)
- **Nodes**: Process (executes), data (input/output)
- **Edges**: data flow of one step to the next
- **Operators***:* of the language (e.g., transformation)

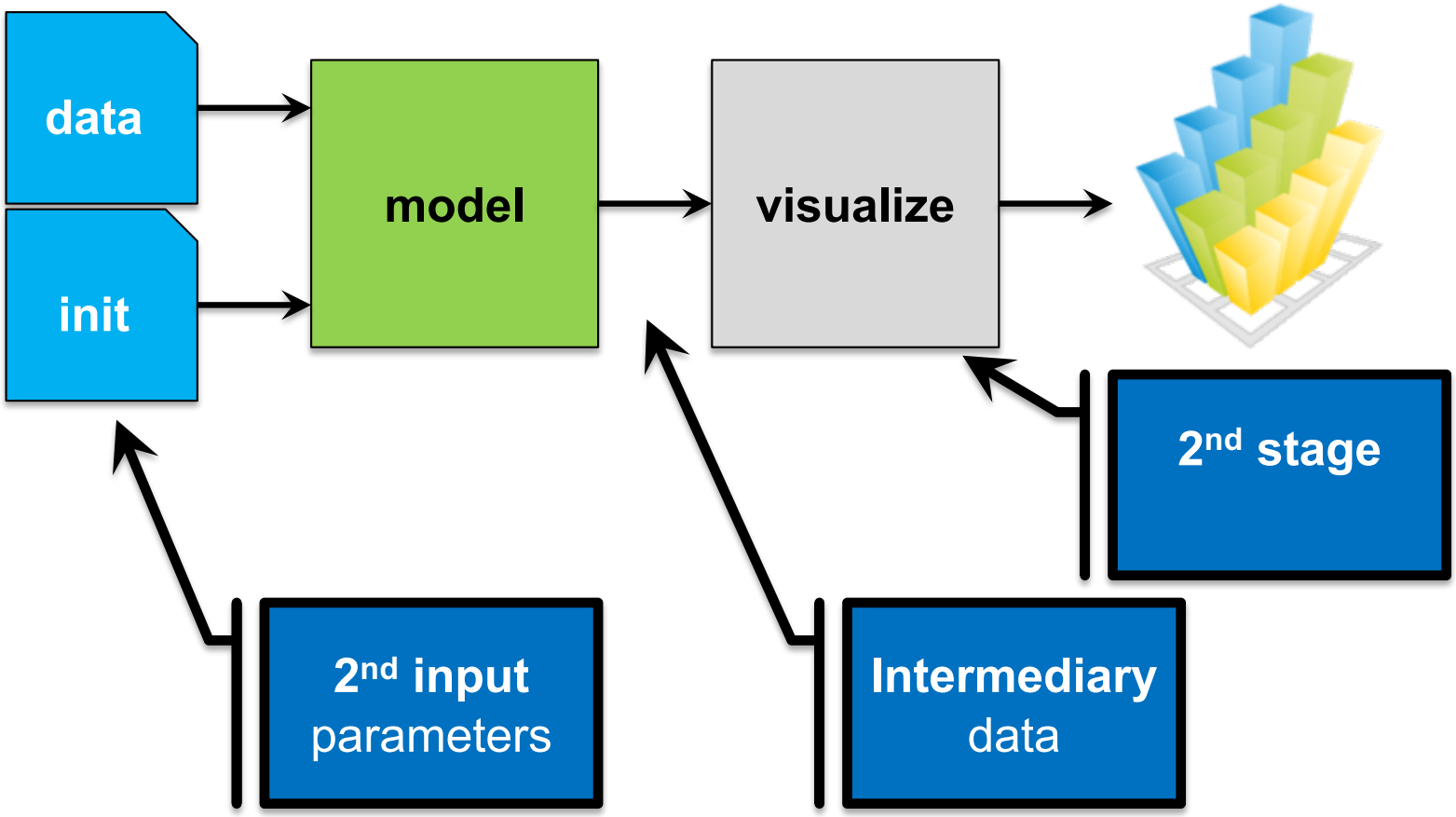
By itself, ***workflow is a specification***

- Realized: Program, script, *visual language*
- Represents experiment structure, data, steps

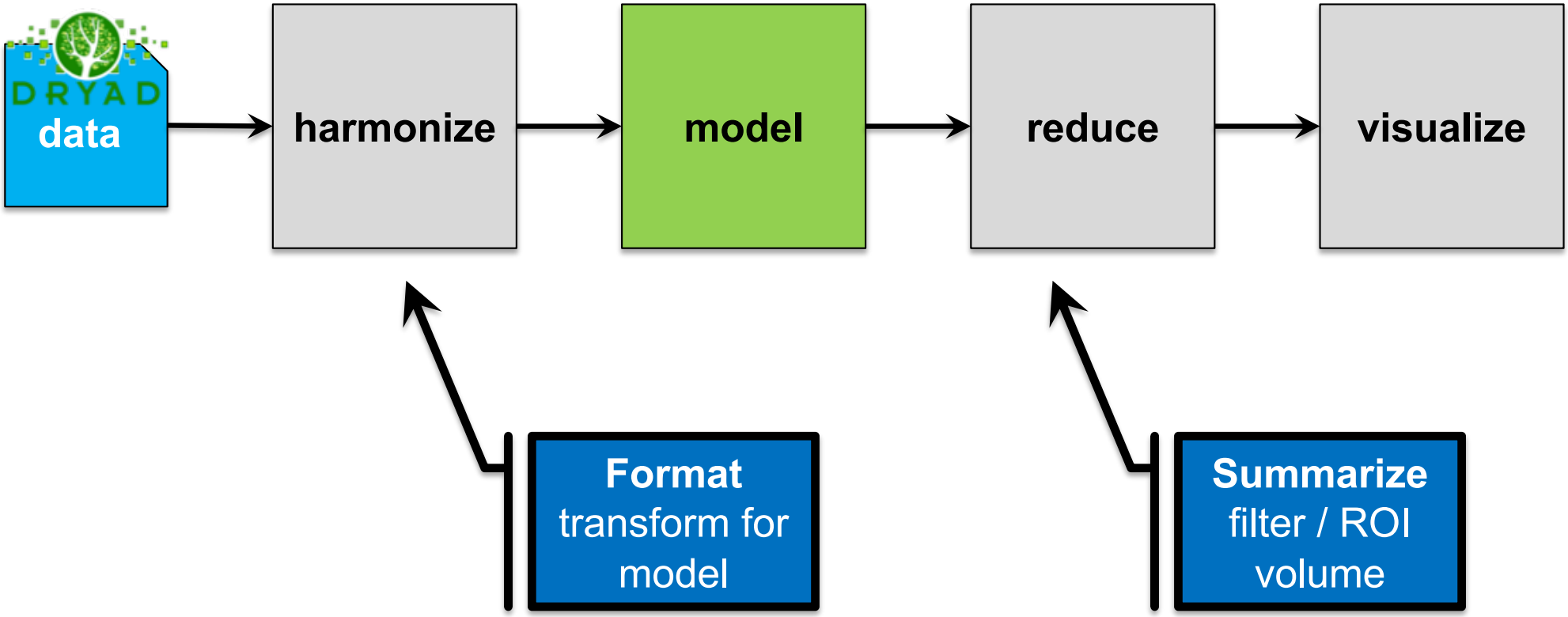
Workflow Pattern: Process



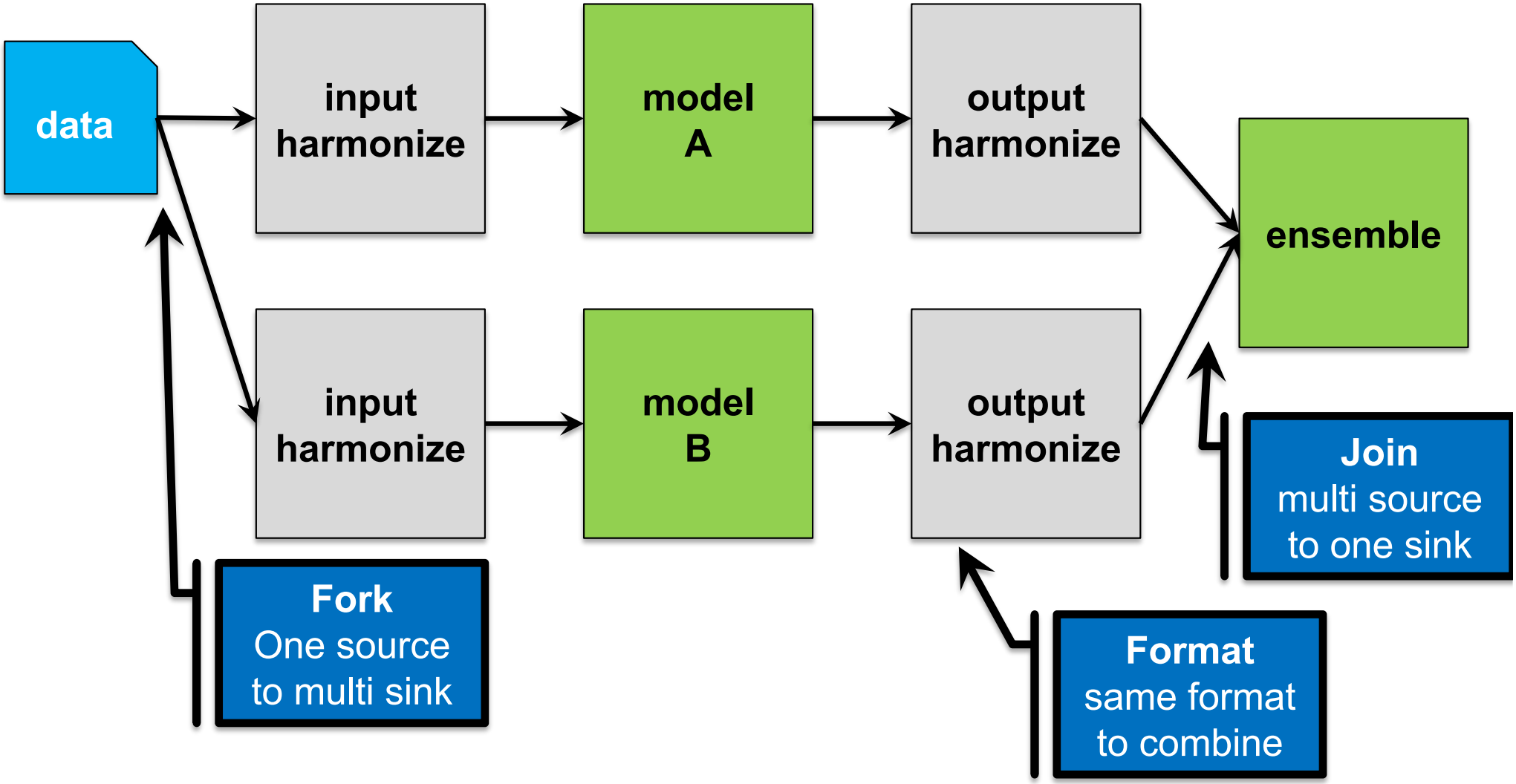
Workflow Pattern: Pipeline



Workflow Pattern: Pipeline



Workflow Pattern: Fork & Join



Capture: How to Specify

- Text

Specification
print.cwl

```
cwlVersion: cwl:draft-3
class: CommandLineTool
baseCommand: echo
inputs:
  - id: message
    type:string
    inputBinding:
      position: 1
outputs: []
```

Data Input
print.cwl

```
message: Hello World Modelers!
```

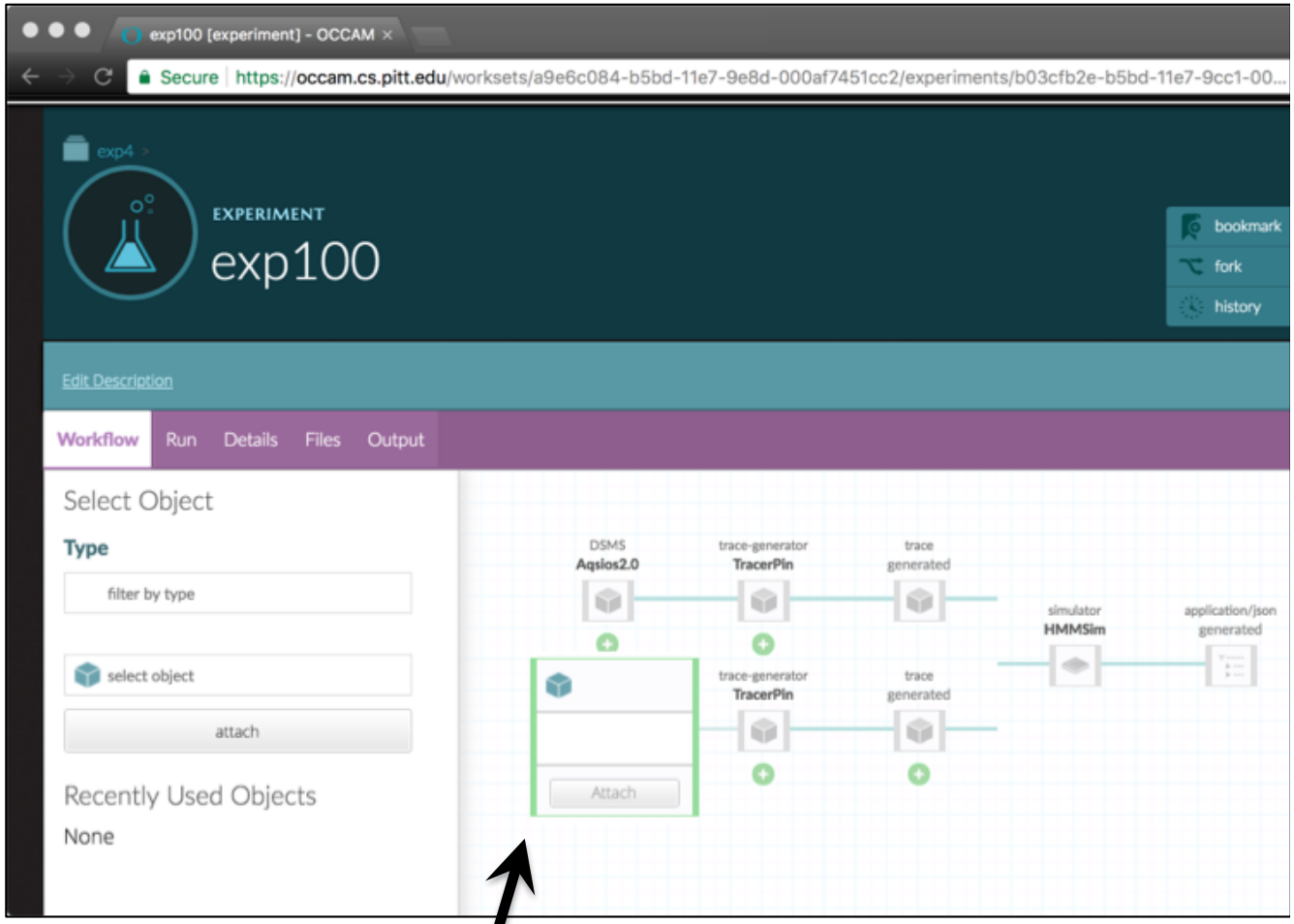
Execution
Output

```
$cwl-runner print.cwl mesg.yml
Hello World Modelers!
```

Credit: CWL 3.0 hello world example

Capture: How to Specify

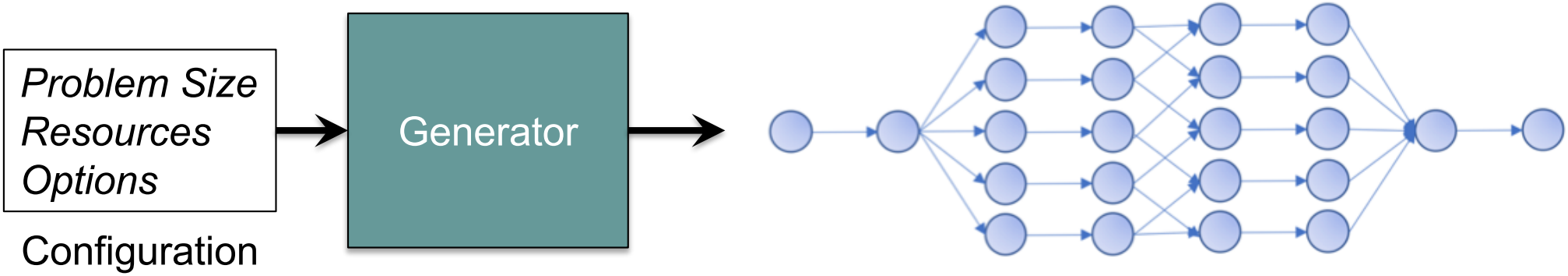
- Text
- Draw (visual)



Editor
Drawing Workflow

Capture: How to Specify

- Text
- Draw (visual)
- Generate



Example generators:

- ❑ Montage: Image stitching for astronomy
- ❑ CyberShake: Earthquake modeling for So. California
- ❑ LIGO: Analyze gravitational waveforms

Capture: How to Specify

- Text
- Draw (visual)
- Generate
- Program

```
input_file <- "data/data.csv"
output_file <- "data/results.csv"

# read input
input_data <- read.csv(input_file)
# get number of samples in data
sample_number <- nrow(input_data)
# generate results
results <- some_other_function(input_file, sample_number)
# write results
write.table(results, results_file)
```

Credit: SW Carpentry Programming with R

Workflow Execution

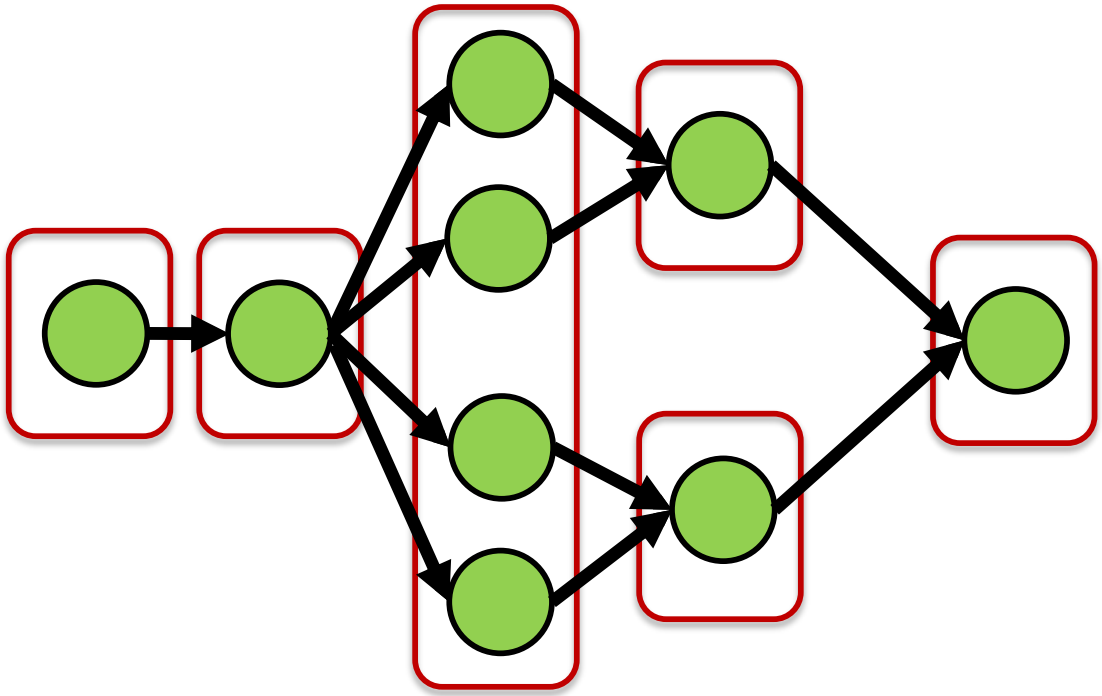
Need to run (i.e., “conduct experiment”)

- Workflow (few steps to 100Ks of steps)
- Actual input and configuration values
- Resources (e.g., machines)

Workflow engine handles the “magic”

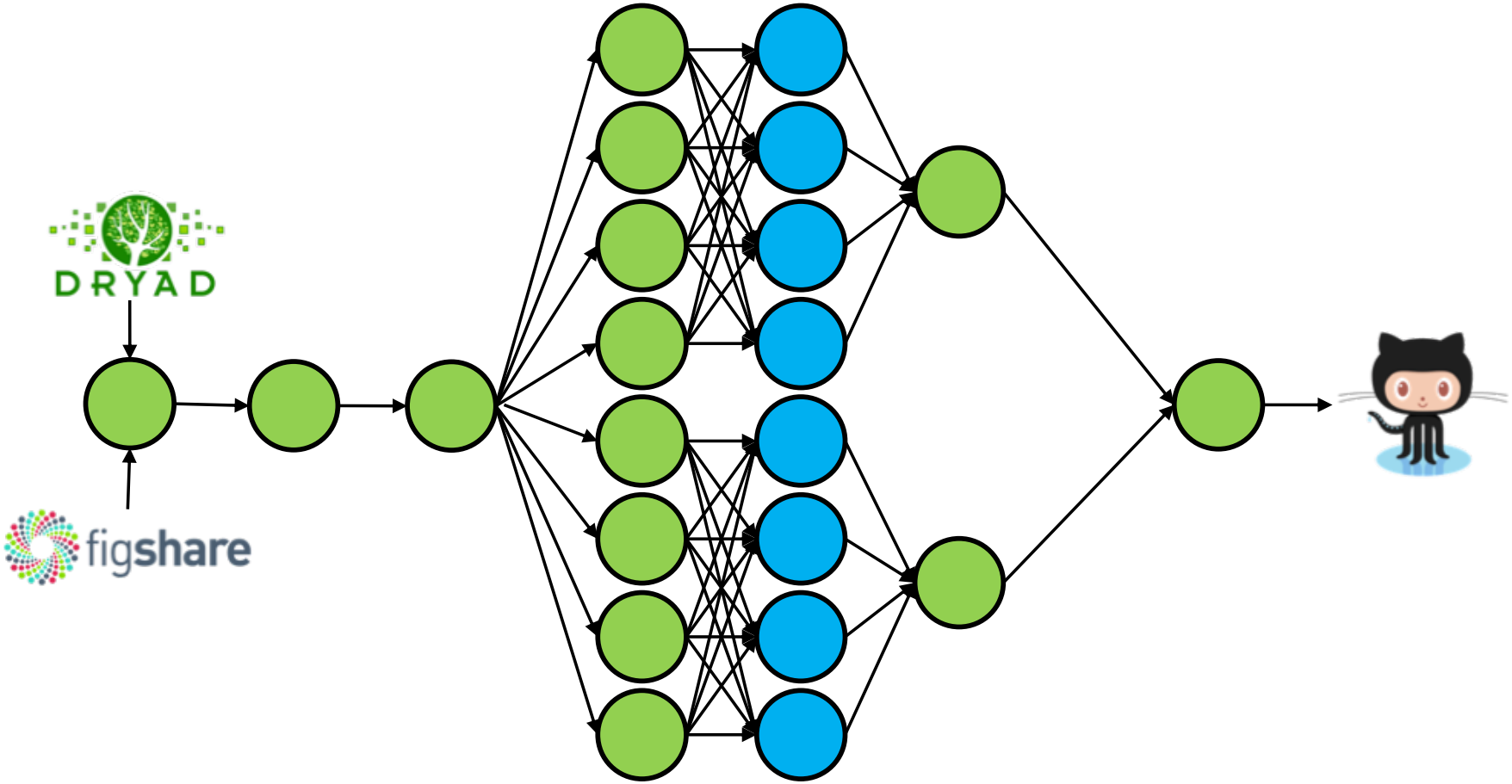
- Compute scheduling: dispatch processes to resources
- Data orchestration: input, intermediary, result
- Management: Monitoring, logging, error recovery
- Metadata for experiment, workflow preserved

Workflow Execution (Simple)

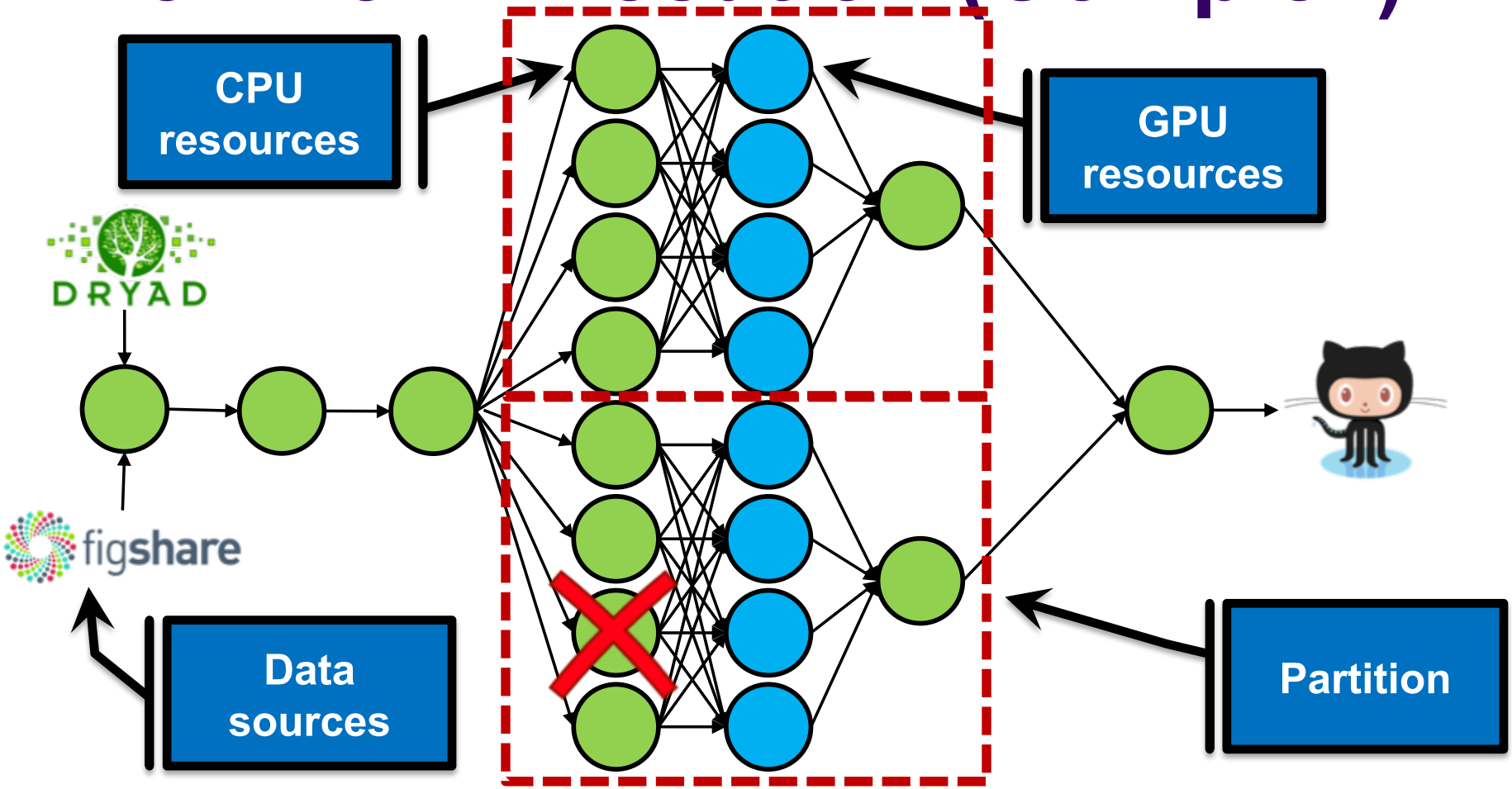


Simple scheduling on non-distributed resources
Order imposed by workflow
Inputs and resource available to execute step

Workflow Execution (Complex)



Workflow Execution (Complex)



Data sources and sinks
Grouping computational steps & allocate resources
Failure and recovery

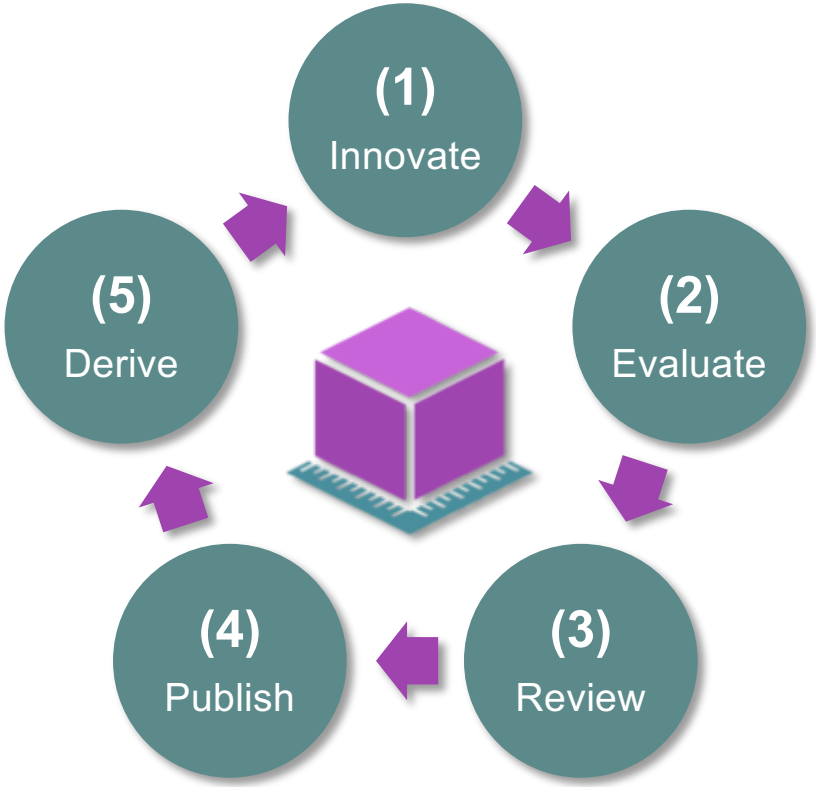
Some Factors to Consider

- Science domain
- Interactivity (vs. generated)
- Scalability
- Resources (e.g., laptop vs. distributed cloud)
- Collaboration (e.g., workspaces, identifiers)
- Metadata & preservation

System	Application	Notable features
Galaxy	Genomics	Commons, HPC, repeatability
Pegasus	General science	Data-centric, scalability, community
Kepler	Astronomy, others	Derived embedded systems modeling
Traverna	Biomedicine, others	Workbench, provenance
VisTrails	General science	Provenance, visualization, exploration
CK	Machine learning	Packaging, comparison, auto-tuning
ReproZip	Data science	Record & replay for repeatability
Pachyderm	General science	Repeatability, provenance, commercial
CodeOcean	General science	Publication, editing & run, commercial

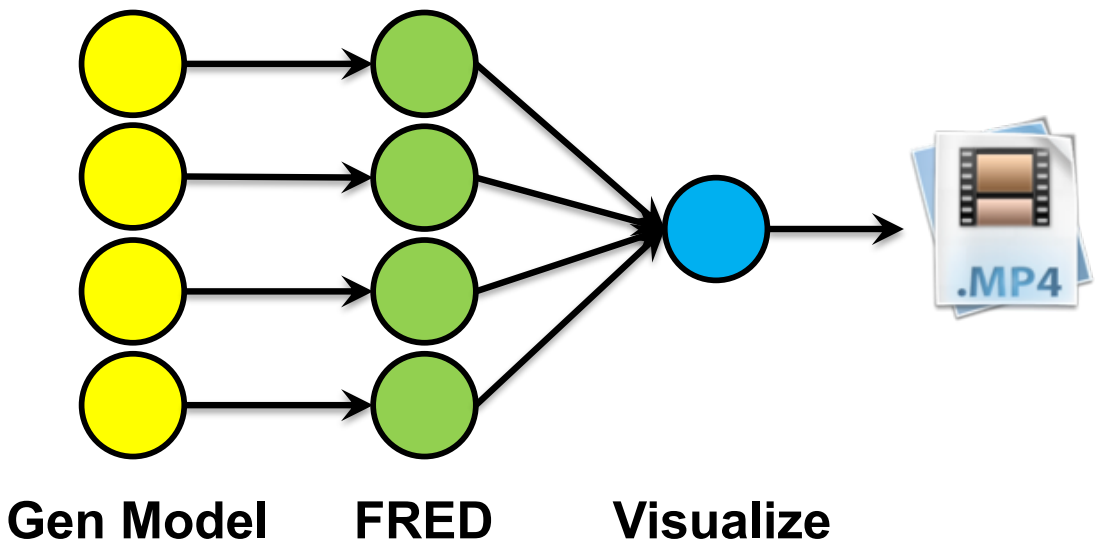
OCCAM: *Experiment* System

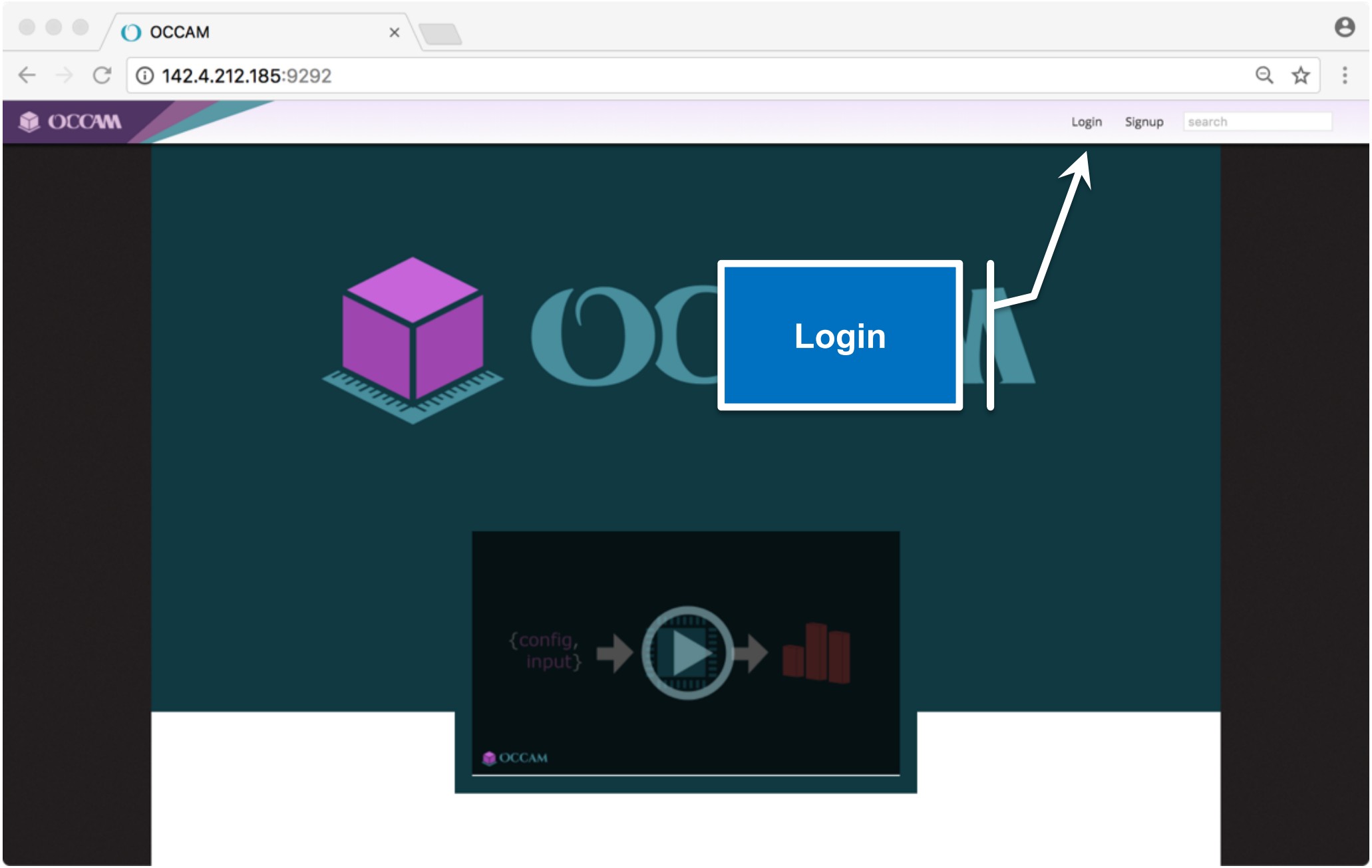
- Workflows
- Research life cycle
- Contributory & collaborate
- Preservation, provenance
- FAIR metadata

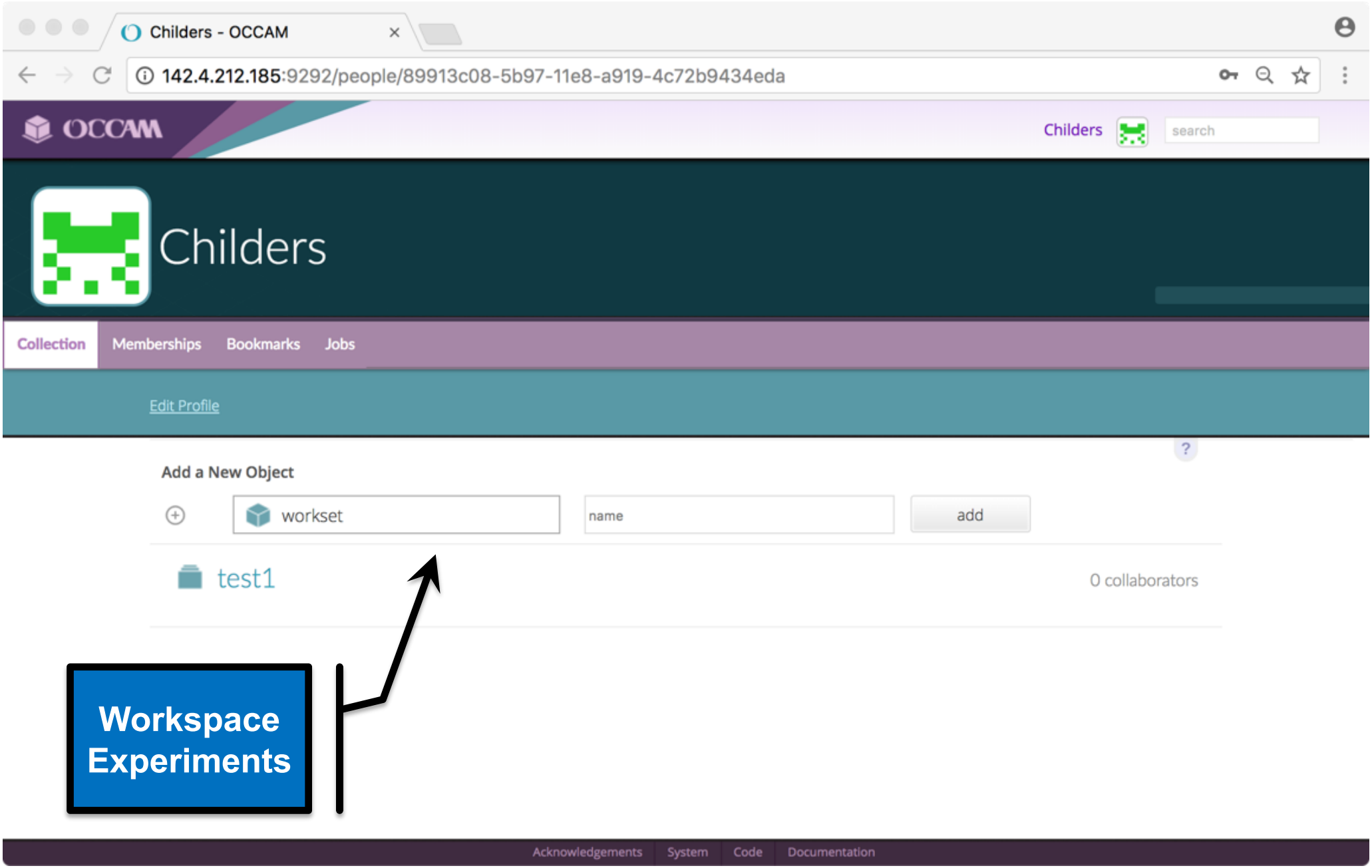


Demo! #1

- Start of research lifecycle
  - Creating, running & sharing an experiment
  - Influenza model with FRED
  - Visualize results







Workspace Experiments

test1 [workset] - OCCAM

Not Secure | 142.4.212.185:9292/94f48412-5bdb-11e8-a324-4c72b9434eda/19c227b8e4f5ea7b1791be661389833774e5...

WORKSET

test1

Childers

bookmark

fork

Contents | Details | Files | Output | History | Access

Authors

username | add

Collaborators

username | add

Directory

- fred1
- fred2

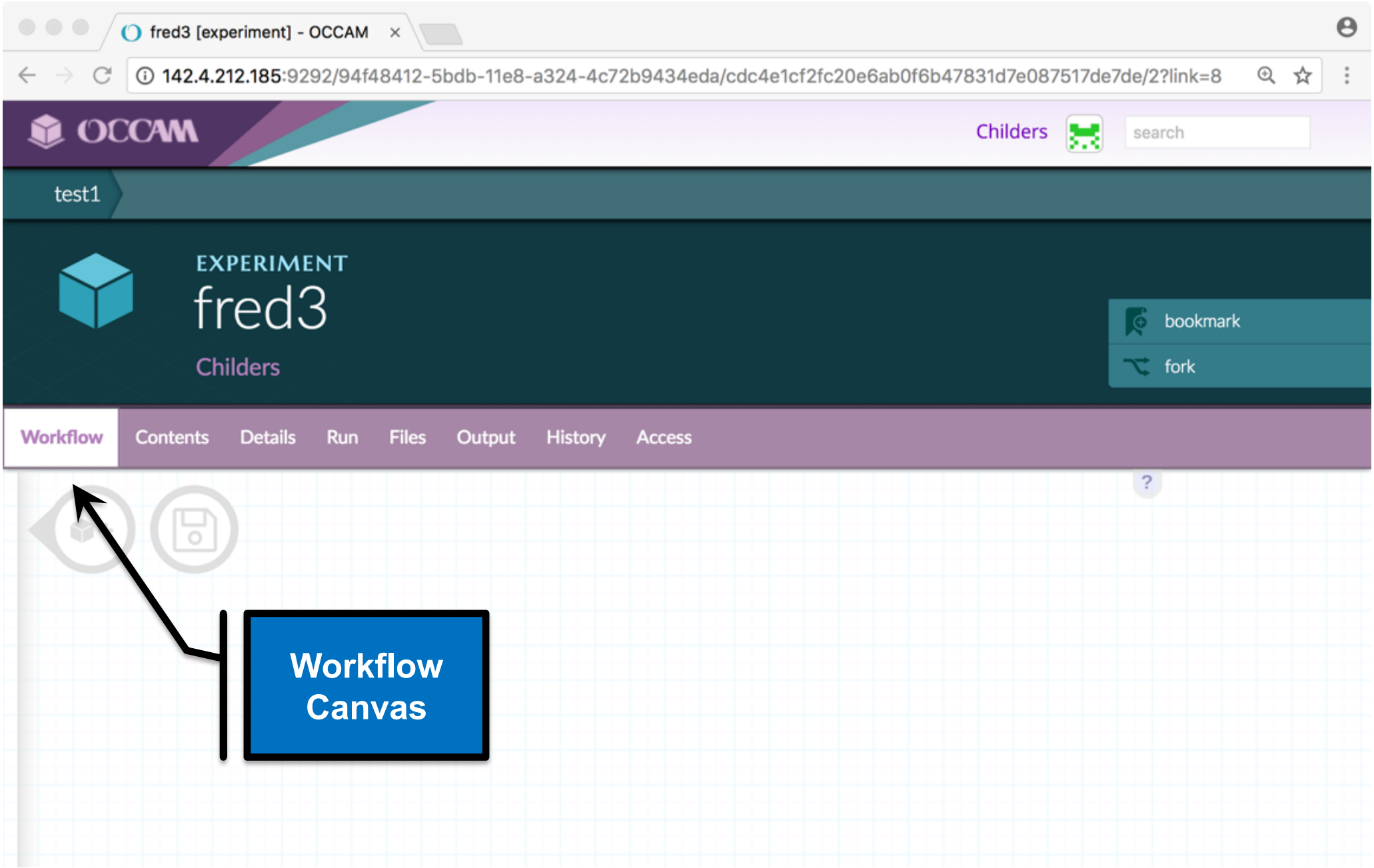
New Experiment

Add a New Object

experiment | Fred3 | add

Prior Experiments

Acknowledgements | System | Code | Documentation



fred3 [experiment] - OCCAM

Not Secure 142.4.212.185:9292/94f48412-5bdb-11e8-a324-4c72b9434eda/cdc4e1cf2fc20e6ab0f6b47831d7e087517de...

Childers

fork

Workflow | Contents | Details | Run | Files | Output | History | Access

Select Object

Type

simulator

Name

FRED

FRED simulation
Simulation

simulator
FRED

FRED result
Output

Workflow Templates

Recently Used Objects

None

FRED simulation
Simulation

simulator
FRED

FRED result
Output

Library

Instantiated FRED (1 of 4)

fred3 [experiment] - OCCAM

Not Secure 142.4.212.185:9292/94f48412-5bdb-11e8-a324-4c72b9434eda/cdc4e1cf2fc20e6ab0f6b47831d7e087517de7de/...

Childers

fork

Workflow | Contents | Details | Run | Files | Output | History | Access

FRED simulation
Simulation

simulator
FRED

FRED result
Output

FRED simulation
Simulation

simulator
FRED

FRED result
Output

FRED simulation
Simulation

simulator
FRED

FRED result
Output

FRED simulation
Simulation

simulator
FRED

FRED result
Output

Instantiated FREDs (4 of 4)

fred3 [experiment] - OCCAM

Not Secure 142.4.212.185:9292/94f48412-5bdb-11e8-a324-4c72b9434eda/cdc4e1cf2fc20e6ab0f6b47831d7e087517de7de/...

Childers

fork

Workflow Contents Details Run Files Output History Access

Select Object

Type

simulation

Name

vaccine

simulation vaccine

FRED simulation Simulation Params

Workflow Templates

Recently Used Objects

None

FRED simulation Simulation Params

FRED simulation Simulation

simulator FRED

FRED result Output

simulation vaccine

FRED simulation Simulation Params

FRED simulation Simulation

simulator FRED

FRED result Output

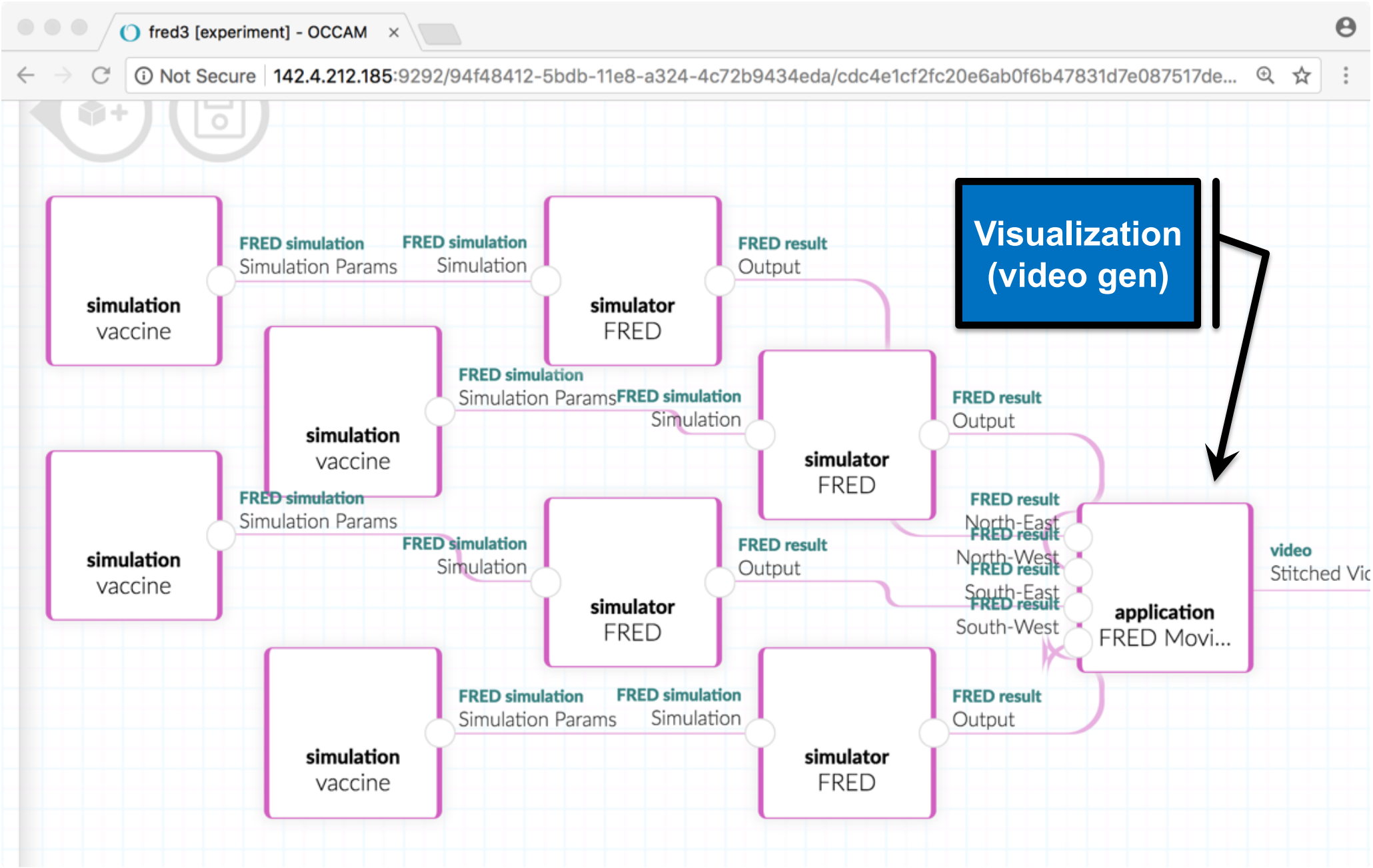
FRED simulation Simulation

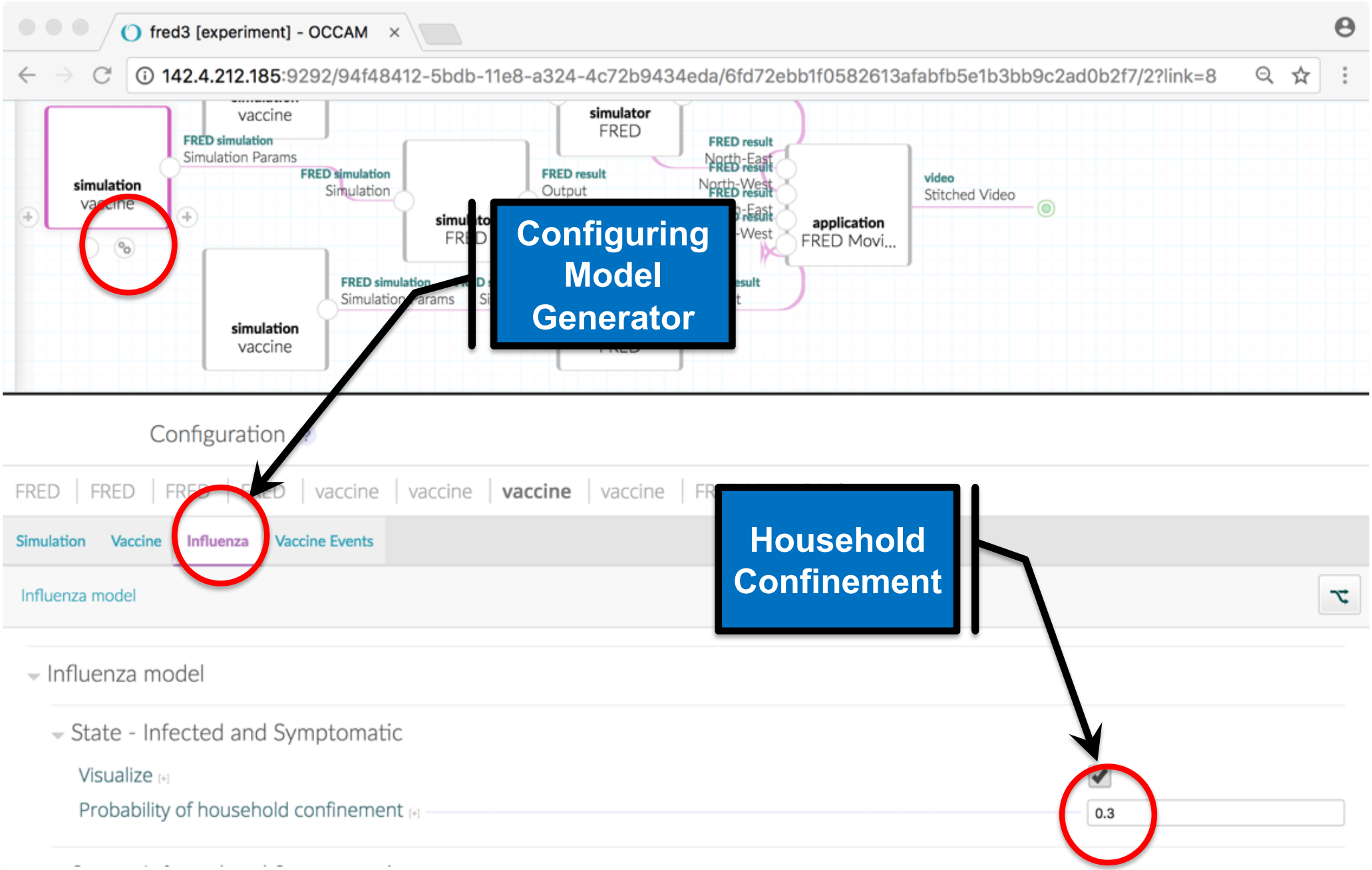
FRED result Output

Model Generator (config)

FRED result Output

simulator FRED





fred3 [experiment] - OCCAM

142.4.212.185:9292/94f48412-5bdb-11e8-a324-4c72b9434eda/6fd72ebb1f0582613afabfb5e1b3bb9c2ad0b2f7/2?link=8

Simulation | Vaccine | Influenza | Vaccine Events

Vaccines become available

Vaccines become available

Start of event [+]	10
End of event [+]	10
Maximum number of events [+]	0
Percentage of candidates vaccinated [+]	0.2
Center latitude of circle of exposure [+]	0
Center longitude of circle of exposure [+]	0
Radius of exposure (km) [+]	0
Fips code of affected region (zero - disabled). [+]	0
Minimum age [+]	0
Maximum age [+]	999

Candidates Vaccinated

fred3 [experiment] - OCCAM

142.4.212.185:9292/94f48412-5bdb-11e8-a324-4c72b9434eda/6fd72ebb1f0582613afabfb5e1b3bb9c2ad0b2f7/2/run?lin...

fred3

Childers

bookmark

fork

Workflow | Contents | Details | Run | Files | Output | History | Access

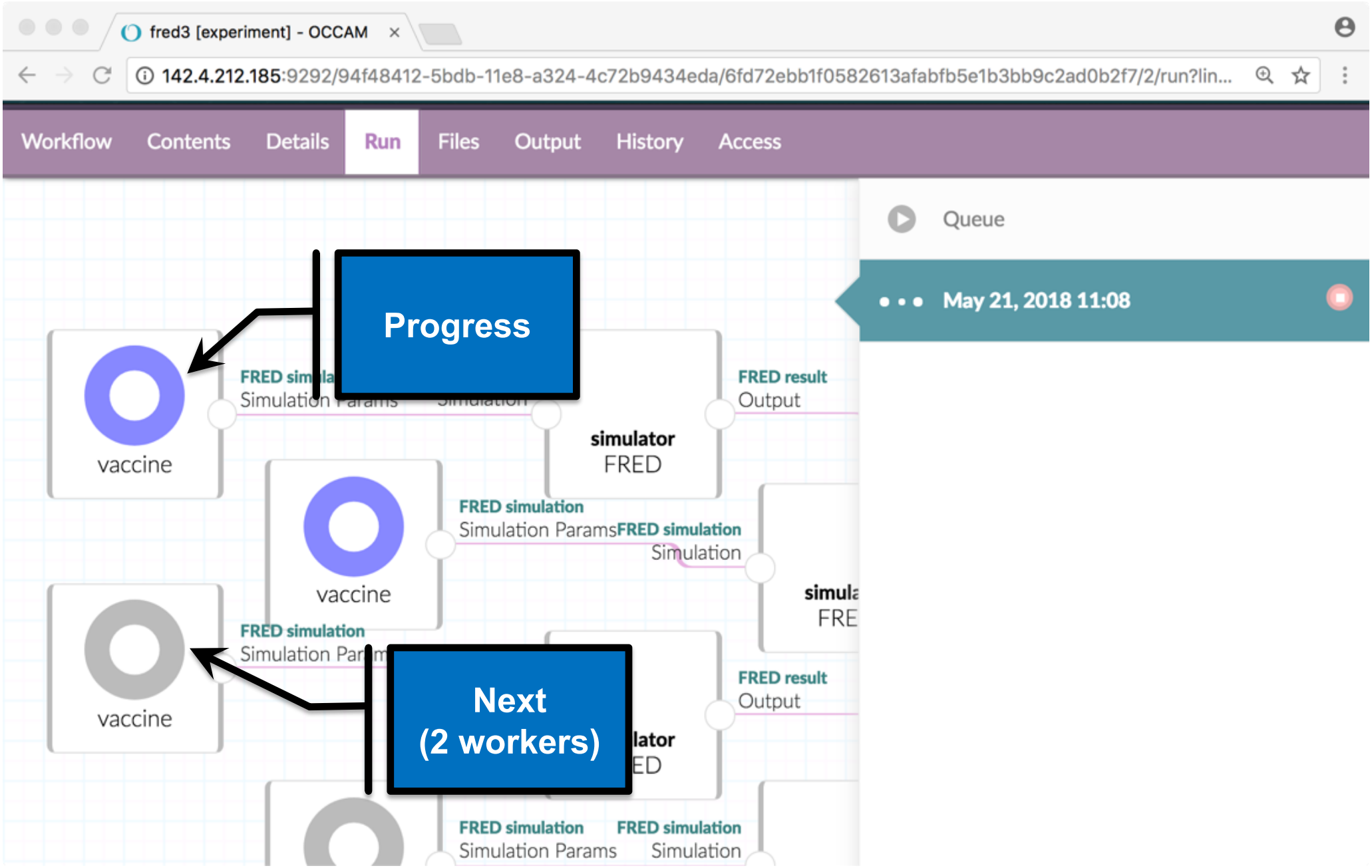
Queue

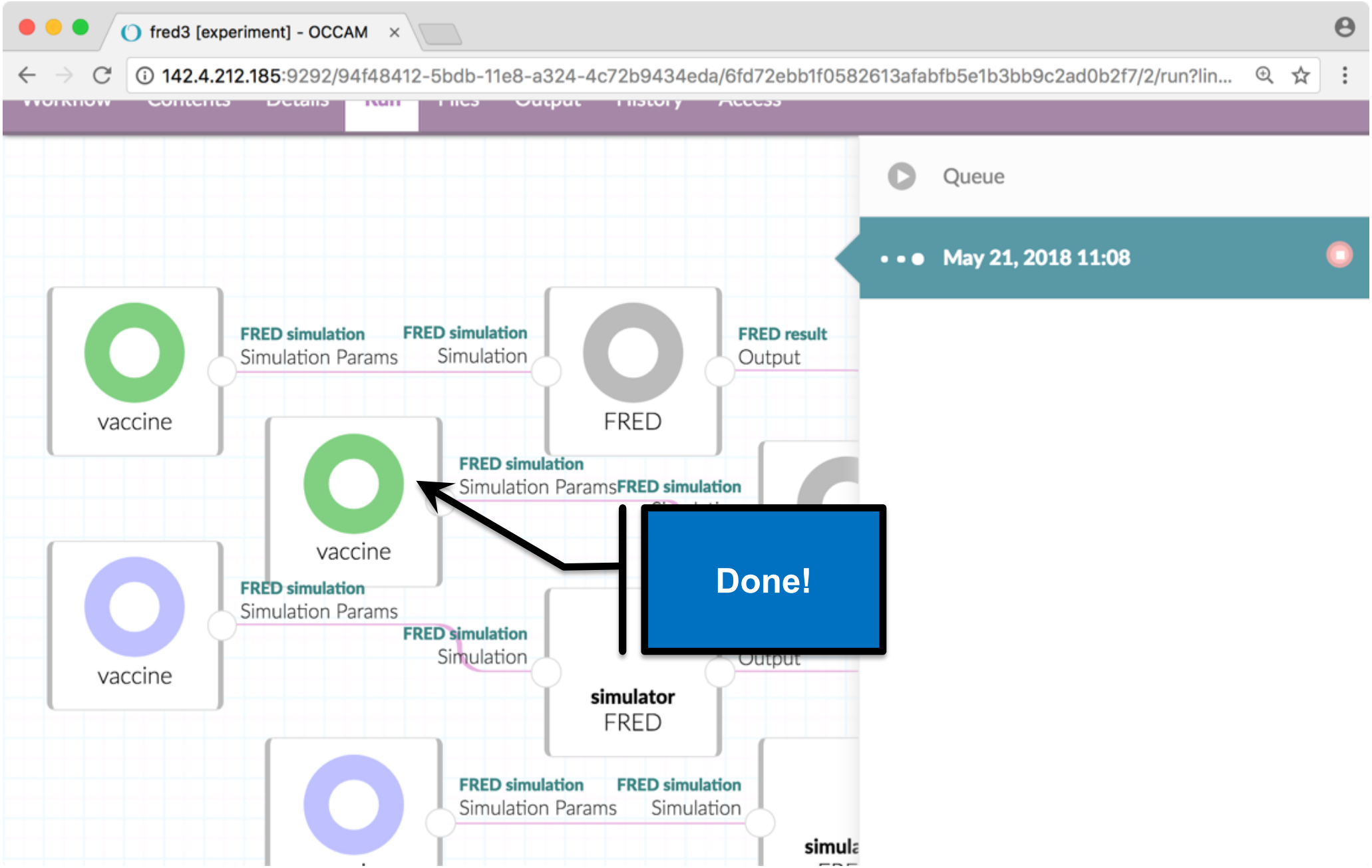
Run!

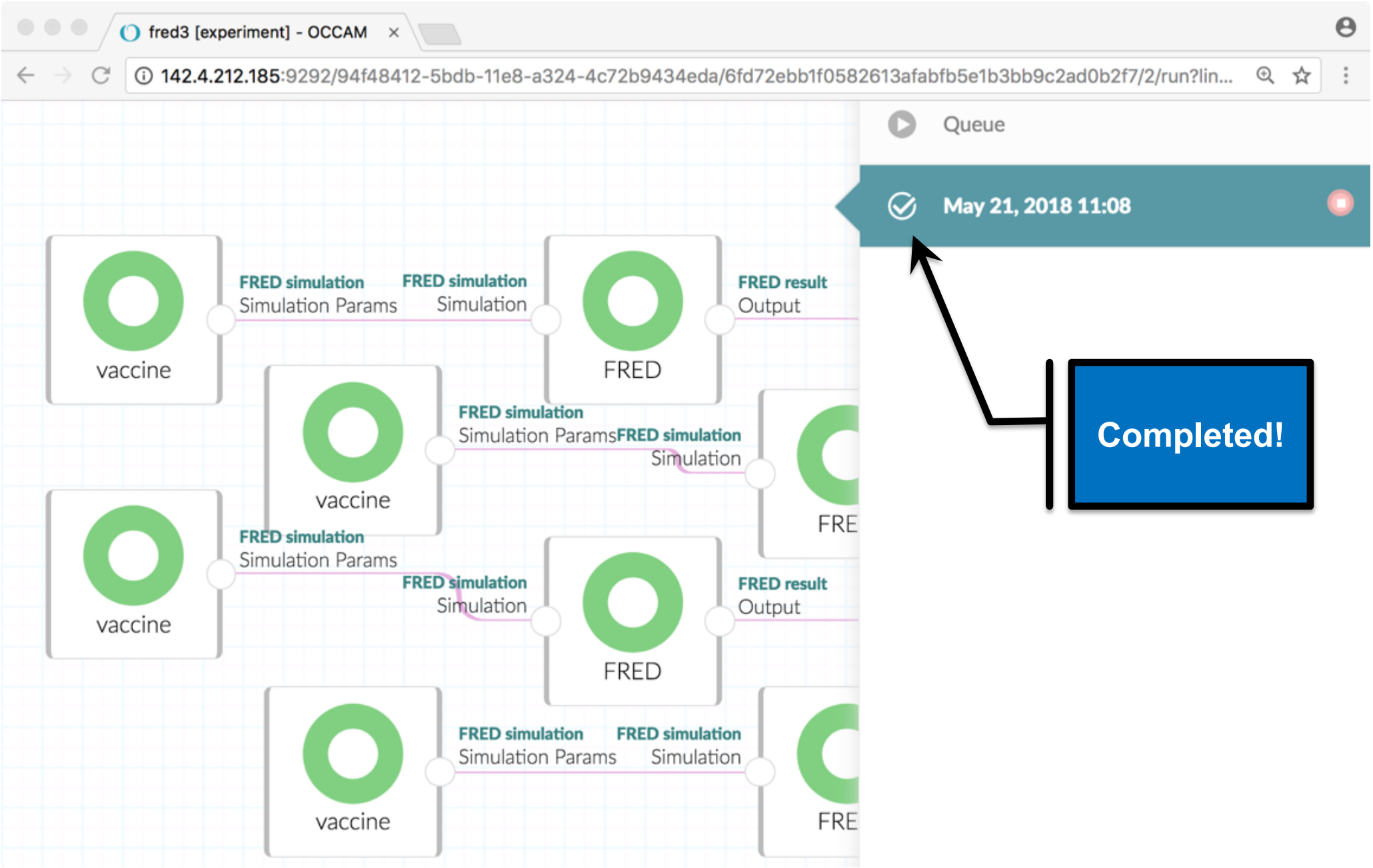
Dispatch To

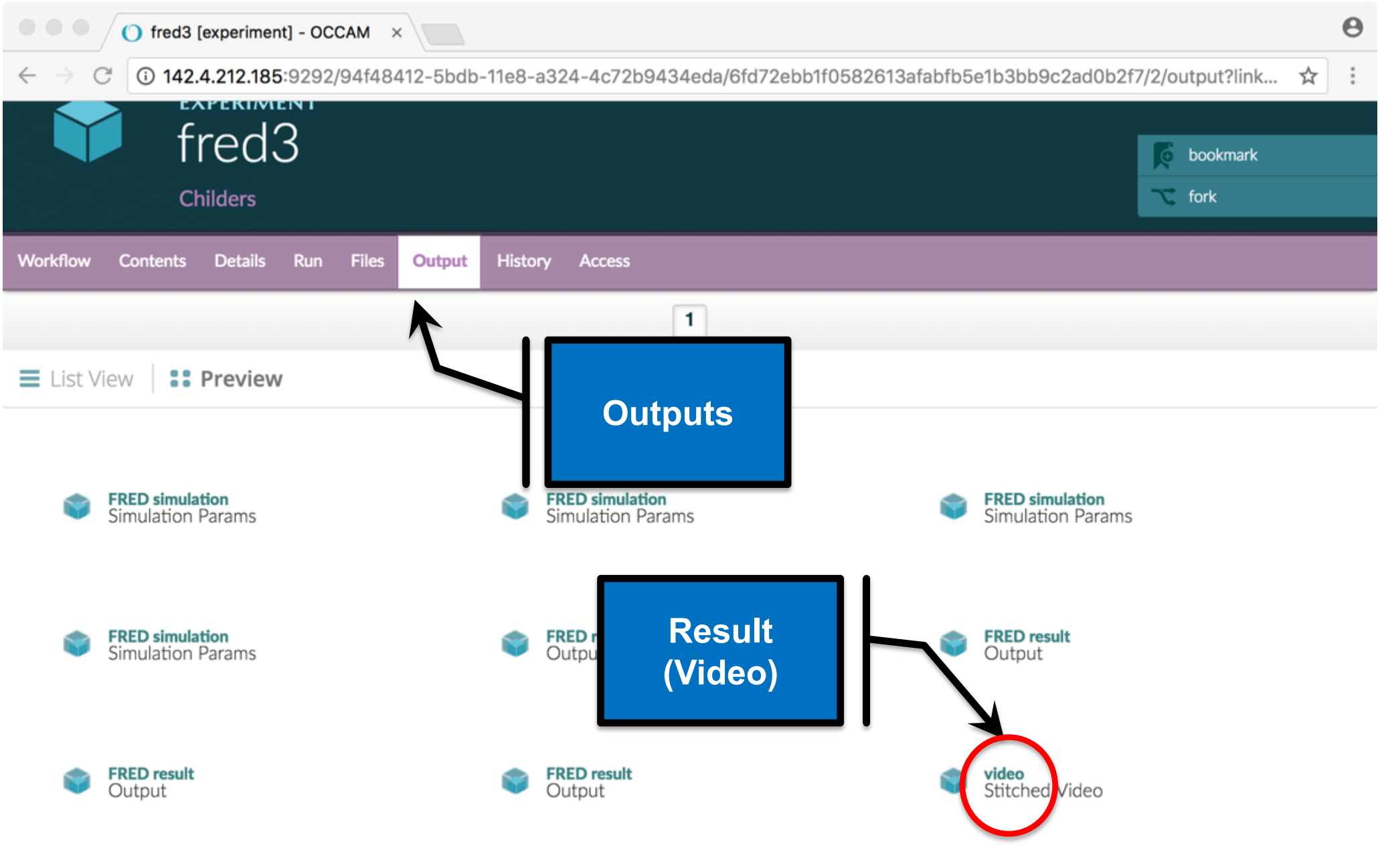
142.4.212.185

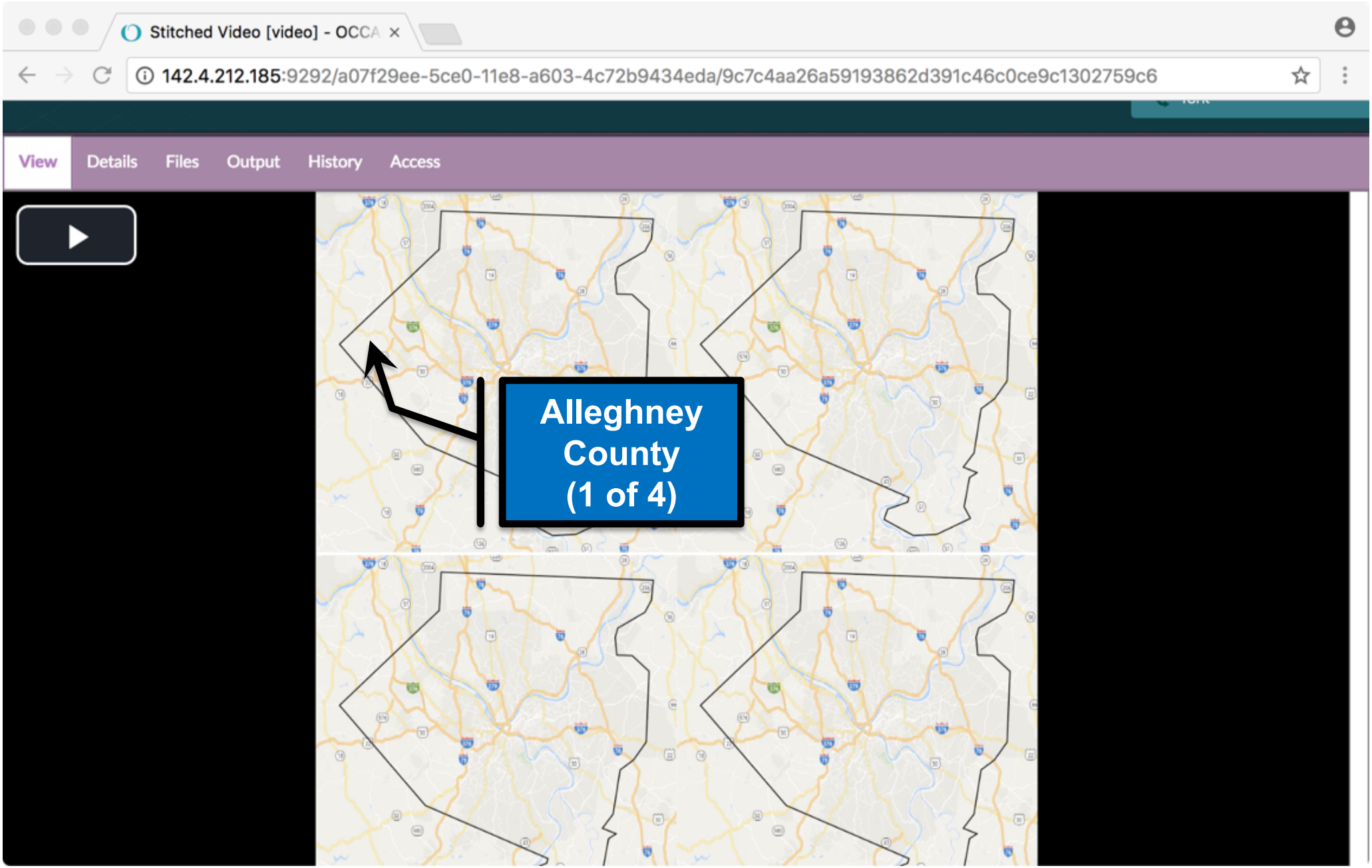
Run

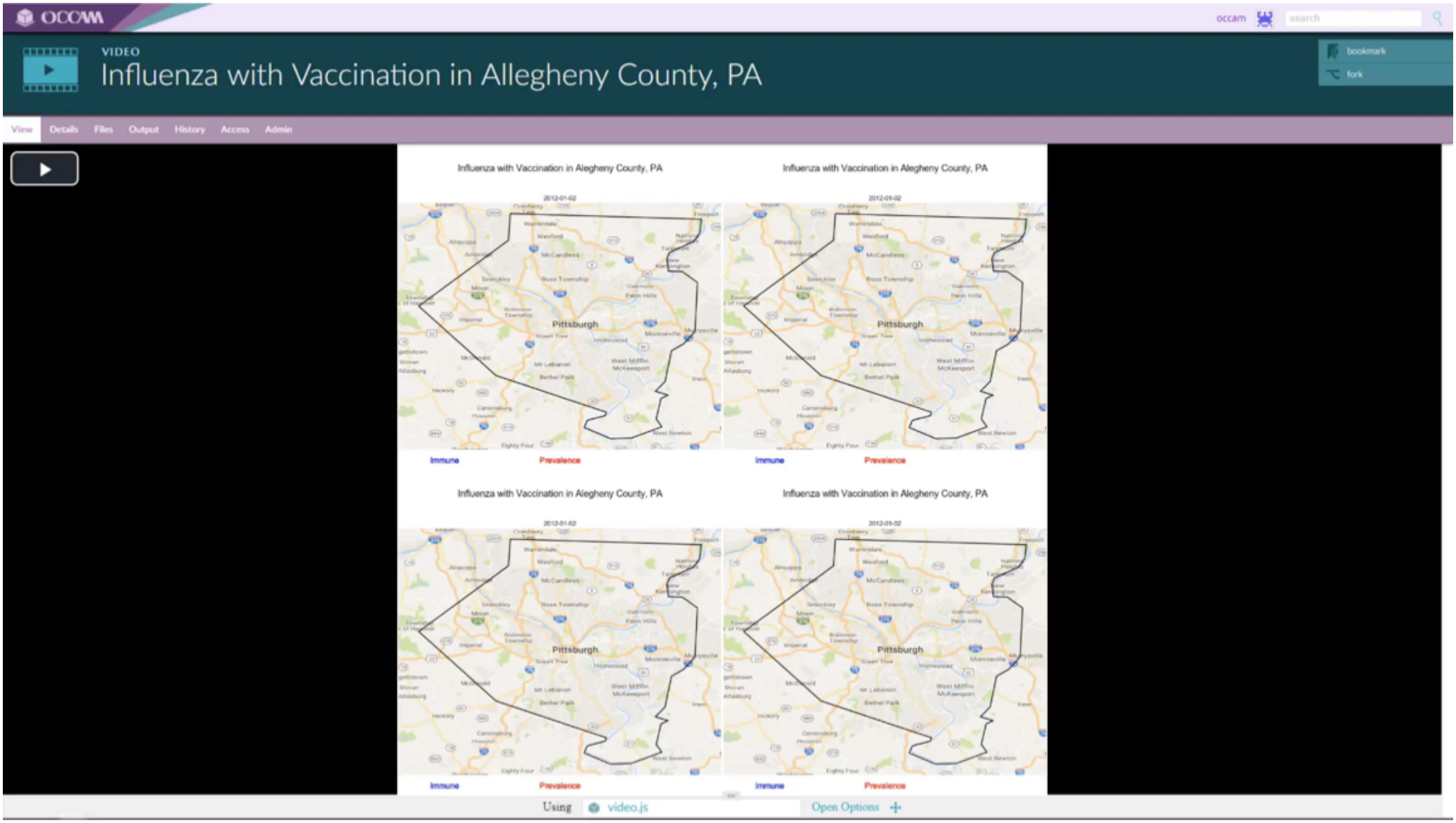












Summary

Workflow systems!

- Accelerate research progress
- Leverage models, data, experiments
- Collaboration for modeling

Part of solution to “reproducibility crisis”

Collaboration, sharing, metadata as well

Community and culture are equally important