Reproducible computational workflows with **signac**

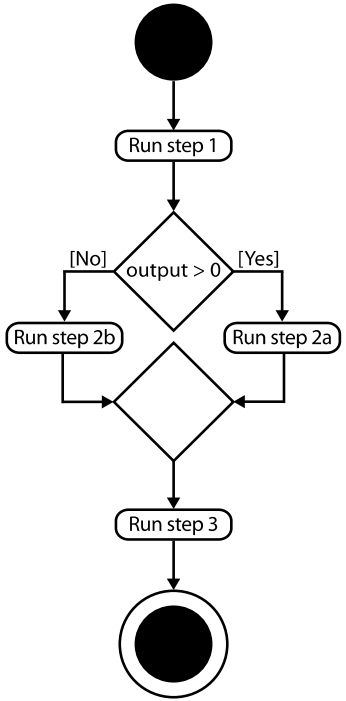
Bradley D. Dice, Carl S. Adorf, Vyas Ramasubramani, Sharon C. Glotzer

MICDE Symposium for the Center for Network and Storage Enabled Collaborative Computational Science

October 15, 2018, Ann Arbor, MI

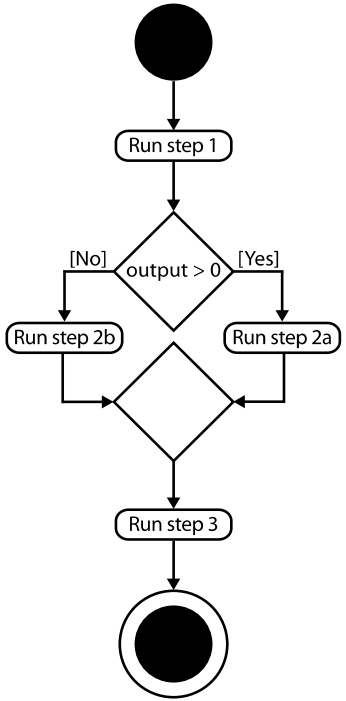
Planning a Parameter Study

```
concentration_A_0.25/
concentration_A_0.50/
concentration_A_0.75/
```



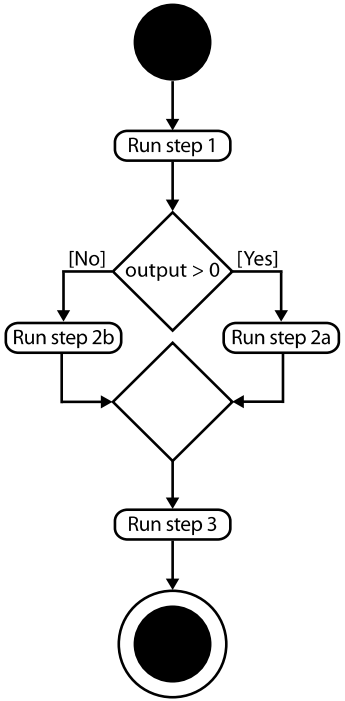
Planning a Parameter Study

```
concentration_A/0.25
concentration_A/0.50
concentration_A/0.75
```



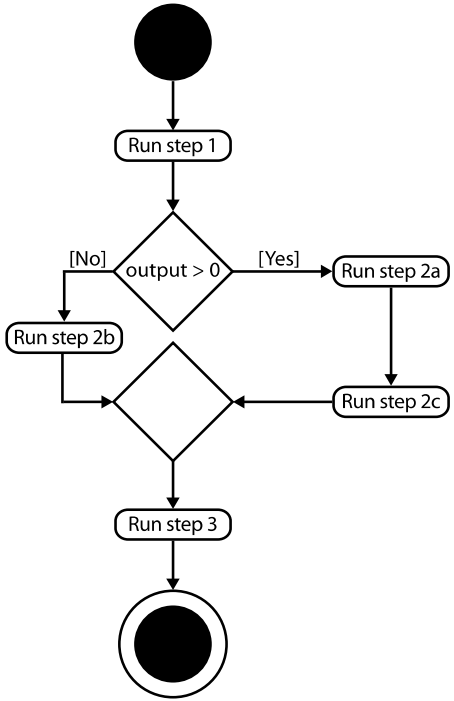
Planning a Parameter Study

```
conc_A/0.25
conc_A/0.50
conc_A/0.75
```



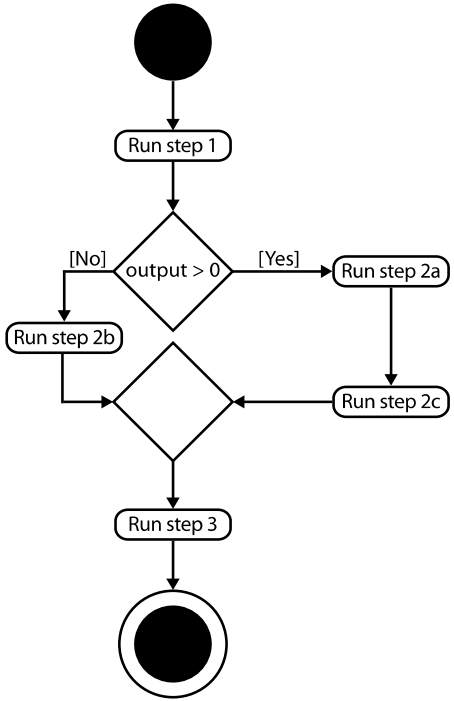
Planning a Parameter Study

```
conc_A/0.25/temp_08
conc_A/0.25/temp_1
conc_A/0.50
conc_A/0.75
```



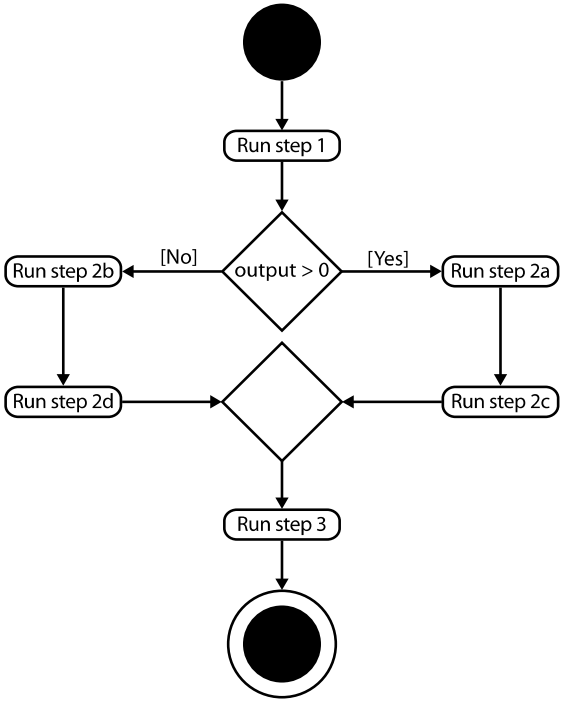
Planning a Parameter Study

```
temp_08/conc_A/0.25
temp_08/conc_A/0.50
temp_08/conc_A/0.75
temp_1/conc_A/0.25
```



Planning a Parameter Study

```
temp_08/conc_A/25/conc_B/05
temp_08/conc_A/50
temp_08/conc_A/75
temp_1/conc_A/25
```

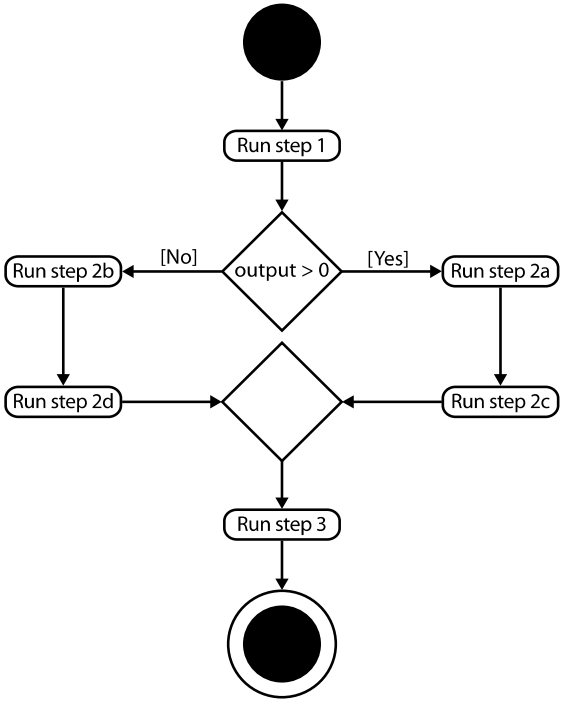


Planning a Parameter Study

```
#!/bin/bash
#SBATCH -J myproject
#SBATCH -N 16
#SBATCH -A ${MYACCOUNT}
#SBATCH -p ${QUEUE}
#SBATCH --ntasks-per-node 8
#SBATCH -t 12:00:00

cd ${WORKING_DIRECTORY}

mpirun -n 16 ./myscript.sh
```



The **signac** Framework

A lightweight, application-agnostic <u>software framework</u> that unobtrusively helps users manage and scale file-based workflows, facilitating data reuse, sharing, and reproducibility.

Python 2/3 | open source BSD-3 | install with pip or conda

www.signac.io

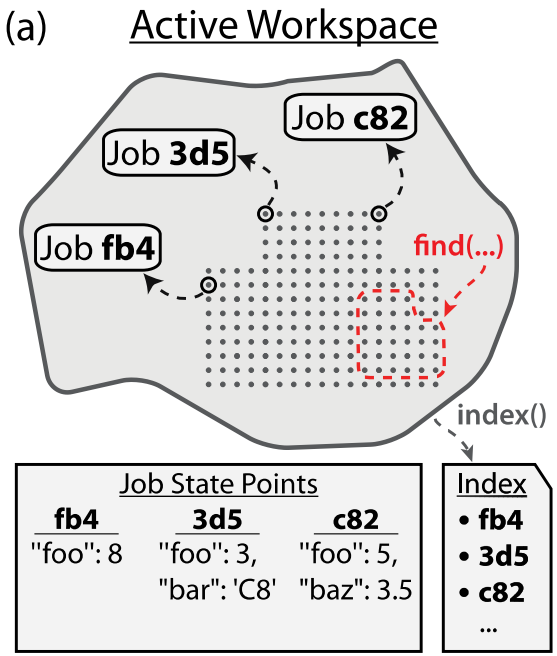
- Simple data and workflow management with the signac framework, JCMS 146, 220-229, (2018)

Topic Overview

1. Introduction
2. Projectile Demo
3. Recent Development and Future Goals

Data (**signac**) and Workflow (**signac-flow**) Management

signac

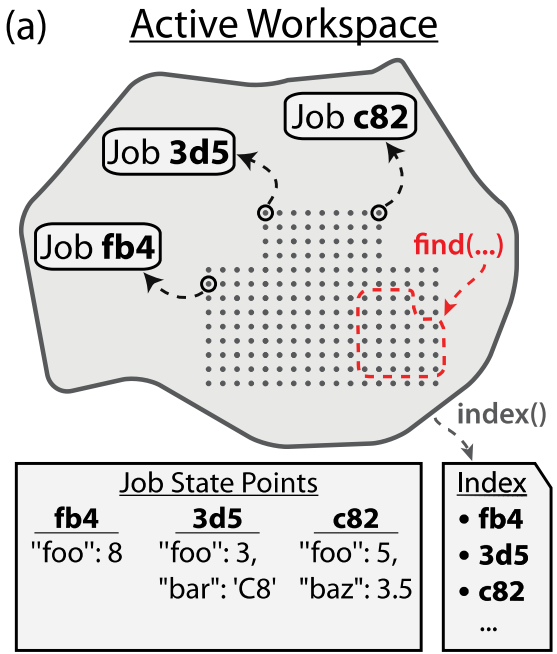


- signac: A Python framework for data and workflow management. In Proceedings of the 17th Python in Science Conference, *2018*.

Data (**signac**) and Workflow (**signac-flow**) Management

signac

*The workspace consists of **jobs**, data containers associated with distinct metadata mappings (called **state points**).*



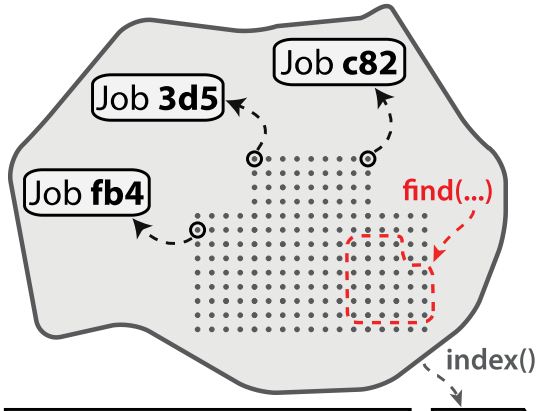
• signac: A Python framework for data and workflow management. In Proceedings of the 17th Python in Science Conference, *2018*.

Data (**signac**) and Workflow (**signac-flow**) Management

*The workspace consists of **jobs**, data containers associated with distinct metadata mappings (called **state points**).*

signac

(a) Active Workspace



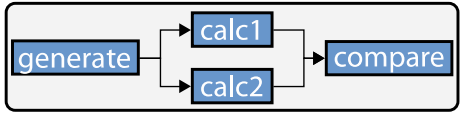
Job State Points

fb4	3d5	c82
"foo": 8	"foo": 3, "bar": 'C8'	"foo": 5, "baz": 3.5

Index
- **fb4**
- **3d5**
- **c82**
- ...

signac-flow

(b) Project Workflow



(c) Status Tracking

operation(job)	Status
generate(fb4)	✓
calc1(fb4)	✓
calc2(fb4)	✓
compare(fb4)	→
generate(3d5)	✓
calc1(3d5)	✓
calc2(3d5)	→
compare(3d5)	⊘
...	...

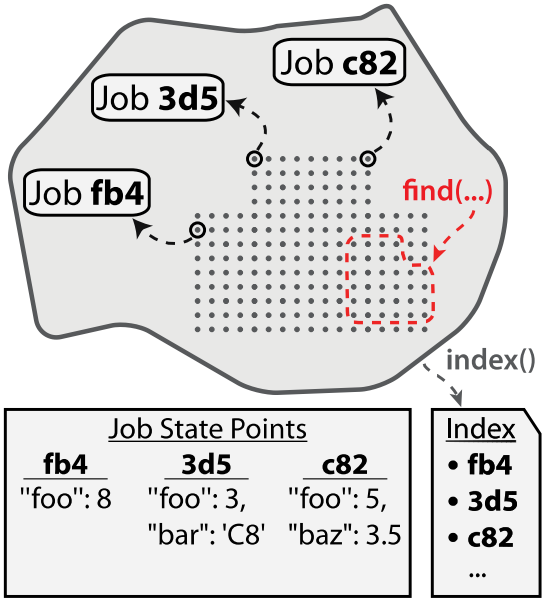
• signac: A Python framework for data and workflow management. In Proceedings of the 17th Python in Science Conference, *2018*.

Data (**signac**) and Workflow (**signac-flow**) Management

*The workspace consists of **jobs**, data containers associated with distinct metadata mappings (called **state points**).*

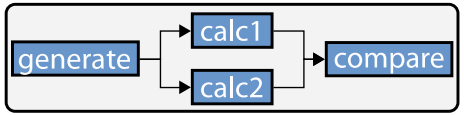
signac

(a) Active Workspace



signac-flow

(b) Project Workflow



(c) Status Tracking

operation(job)	Status
generate(fb4)	✓
calc1(fb4)	✓
calc2(fb4)	✓
compare(fb4)	→
generate(3d5)	✓
calc1(3d5)	✓
calc2(3d5)	→
compare(3d5)	⊘
...	...

*The **workflow** consists of **operations,** linked through condition functions. This forms a directed graph.*

• signac: A Python framework for data and workflow management. In Proceedings of the 17th Python in Science Conference, *2018*.

The **signac** Team

Carl S. Adorf
Lead developer &
co-maintainer,
signac and signac-flow

Vyas Ramasubramani
Developer &
co-maintainer,
signac and signac-flow

Bradley Dice
Lead developer &
maintainer,
signac-dashboard

Top 6 Features Released In Last 12 Months

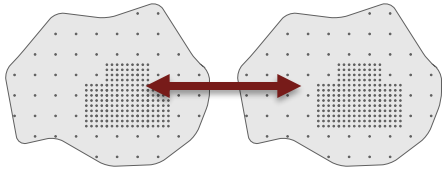
Advanced Searching and Aggregation

```
find({"T.$gt": 298})
  .groupby("P")
```

Schema Detection

```
'T': int([298, 300, 302])
'P': float([0.1, 1.0])
```

Synchronization



One-script Projects

PY

Implement workflows in ***< 10*** *lines.*

Improved Container Support

docker

Templating

jinja2

- container by DPIcons from the Noun Project
- Relational Schema by Becris from the Noun Project
- search data by SBTS from the Noun Project
- Workflow by ProSymbols from the Noun Project

News, Documentation & Chat Room Support

http://www.signac.io http://signac-docs.readthedocs.io https://gitter.im/signac/Lobby

signac - data management framework

Overview Installation Documentation Support News

signac - simple data management

The signac framework aids in the management of large and heterogeneous data spaces.

Get started

The signac framework supports **researchers** in managing project-related data with a well-defined *indexable* storage layout for data and metadata. This streamlines post-processing and analysis and makes data collectively *accessible*.

Signac is **open-source**[1] and **freely** available for **Python** versions 2.7.x and 3.4+.

What you can do with signac

- **Manage** files and document-oriented **data** on the local file system
- **Search** and access **data** within Python or on the command line
- **Develop computational workflows** and data pipelines that are highly scalable
- **Integrate external databases**, such as MongoDB, to manage metadata and share data

To **get started**, follow the installation instructions or have a look at the minimal examples first.

1. The signac source code is licensed under the BSD 3-Clause License. ↩

Signac is a product of The Glotzer Group at the University of Michigan. Development and deployment supported by MICCoM, as part of the Computational Materials Sciences Program funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, under Subcontract No. 6F-30844. Project conceptualization and implementation supported by the National Science Foundation.

signac-docs

Welcome to the signac framework documentation!

The **signac** framework supports researchers in managing project-related data with a well-defined indexable storage layout for data and metadata. This streamlines post-processing and analysis and makes data collectively accessible.

This is the overall **framework documentation**. It provides a comprehensive overview on what you can do with packages that are part of the **signac** framework. If you are new to **signac**, the best place to start is to read the Introduction and the Tutorial.

Contents

- Introduction
  - Overview
  - Packages
- Installation
  - Install with conda
  - Install with pip
  - Installation from Source
- Quickstart
- Tutorial
  - Basics
  - Workflows
  - Job scripts and cluster submission
- Examples
  - Ideal Gas
  - MD with HOOMD-blue
- Recipes
  - How to migrate (change) the data space schema.
  - How to integrate signac-flow with MATLAB or other software without Python interface
  - How to implement MPI-parallelized operations
  - How to handle containers
- Tips and Tricks
  - Do not replicate job metadata in file names.
  - Do not hard-code job ids in your scripts.
  - How to design a good schema.
- Support
- Contribute
- Changelog
  - signac
  - signac-flow
- License
- How to cite signac

Indices and tables

- Index
- Module Index
- Search Page

signac/Lobby

```
usage: project.py [-h] [-d] {status,next,run,script,submit} ...
project.py: error: unrecognized arguments: -q=regular
```

Stephen Thomas @amburan Sep 11 11:52
did'nt realize you can supply additional parameters directly in the command line without waiting to add in that line in the code. Nice

Carl Simon Adorf @csadorf Sep 11 11:52
Make sure to separate the "flow arguments" and the "scheduler arguments" with `--`.
The two dashes mean "everything after this goes directly to the scheduler."
So in your case `python src/project.py submit -o equil --nn 1 -w 10--bundle 10 --serial -- -q=regular`.
`python src/project.py submit -o equil --nn 1 -w 10 --bundle 10 --serial -- -q=regular`

Matt Thompson @mattwthompson Sep 11 11:56
'-q' even thought slurm syntax is -q? (I think so, anyway)

Carl Simon Adorf @csadorf Sep 11 11:56
Oh, sry, I just guessed that.
You might need to specify a partition: http://www.nersc.gov/users/computational-systems/edison/running-jobs/batch-jobs/

Matt Thompson @mattwthompson Sep 11 11:59
Just found that too, looks like `python src/project.py submit -- --partition=regular` is working

Carl Simon Adorf @csadorf Sep 11 11:59
awesome 😊

Matt Thompson @mattwthompson Sep 11 12:00
It looks like things are workable right now
Thanks for your help! Will come back if I run into big problems or can help you guys out too.

Carl Simon Adorf @csadorf Sep 11 12:00
Cool, thank you! 😊

Click here to type a chat message. Supports GitHub flavoured markdown.

PEOPLE

ADD DETAILS

ACTIVITY

csadorf pushed to glotzer/signac
Merge branch 'develop' into sync (compare)

csadorf pushed to glotzer/signac
2 commits

csadorf pushed to glotzer/signac
Add Job.merge() method. (compare)

csadorf pushed to glotzer/signac
Add Job.merge() method. (compare)

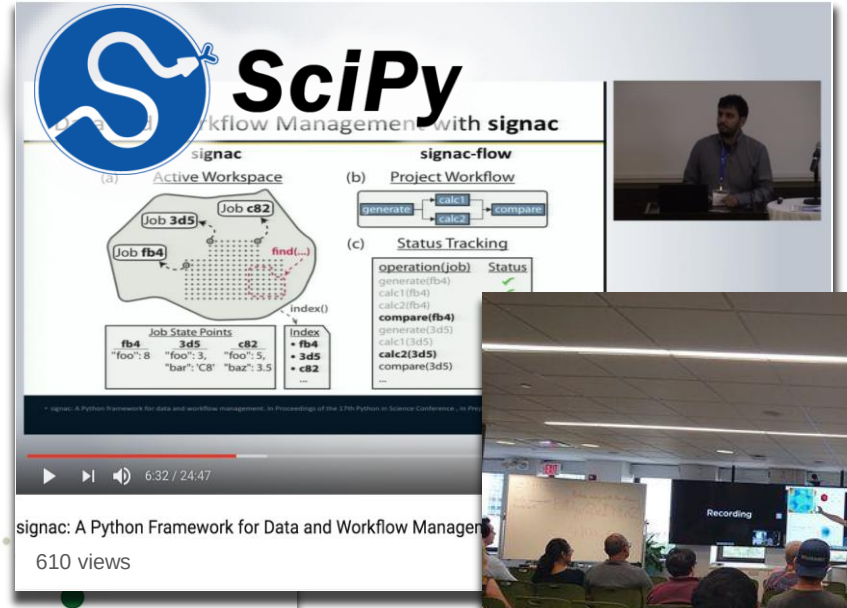
csadorf pushed to glotzer/signac
Use relative paths for sync logging output. (compare)

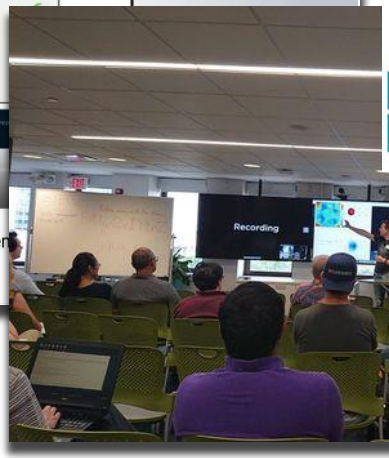
csadorf pushed to

Impact of signac

- In use by several research groups across the country, especially in materials science and molecular simulations
  - Princeton, Vanderbilt, Boise State, Air Force Research Laboratory, ...
- 10,000+ downloads of signac & signac-flow conda packages
- Actively used on projects with thousands of jobs and terabytes of data
- signac-flow has built-in support for Flux and XSEDE clusters, and adaptable templates for any SLURM/TORQUE scheduler

Community Building





PyData
Ann Arbor

17 AIChE
Annual Meeting, Minneapolis, MN

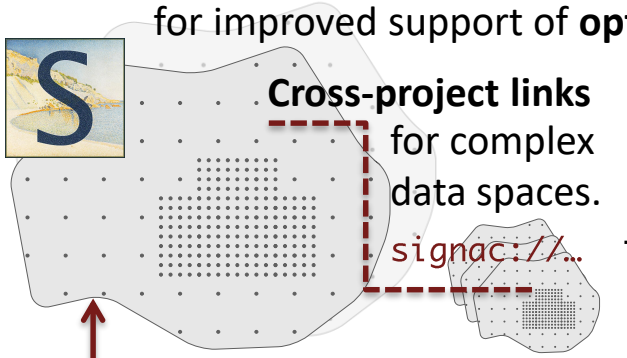
18 AIChE
Annual Meeting, Pittsburgh, PA

[**signac** Projectile Demo]

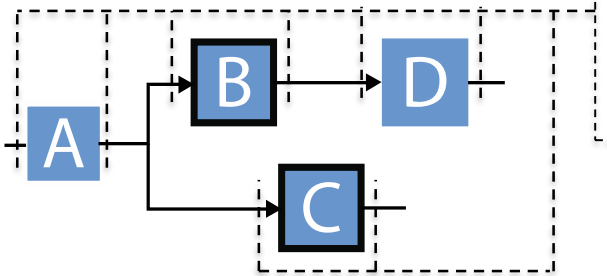
https://github.com/bdice/signac-micde-cnsccs-2018

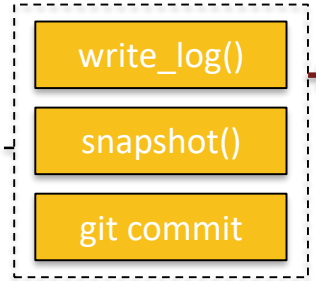
Latest Development

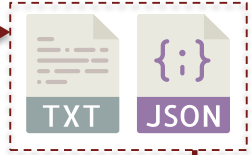
Better handling of **dynamically** growing data spaces for improved support of **optimization** workflows.



Cross-project links for complex data spaces.

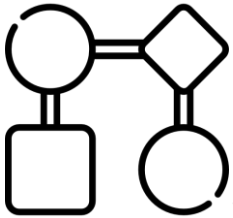




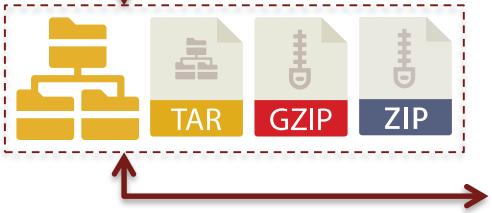


Execution hooks allow consistent tracking of all operations that manipulate the data space.

An **operations log** can be used to generate a graph of actions.



Import from and **export** to directories and archive files.



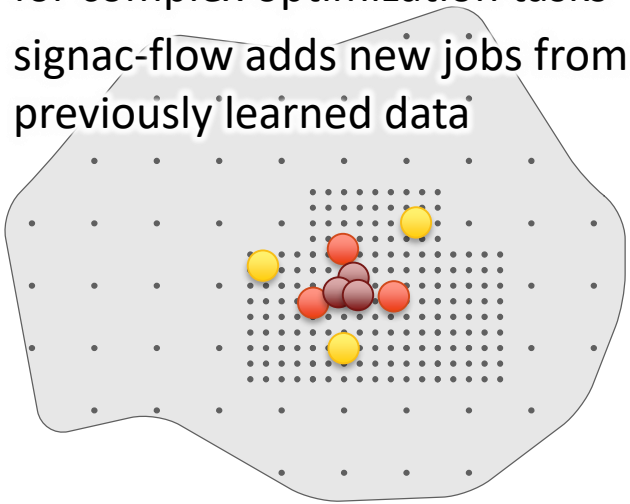
Simplifies **archival** in external collections.

Future goal:

Automated **Machine Learning** Workflows

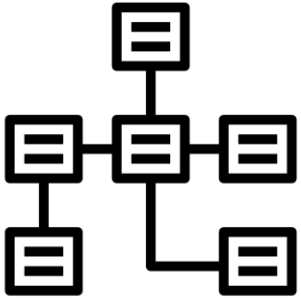
From **dynamic** workflows...

- signac data spaces are suited for complex optimization tasks
- signac-flow adds new jobs from previously learned data



... to distributed, iterative **exploration**

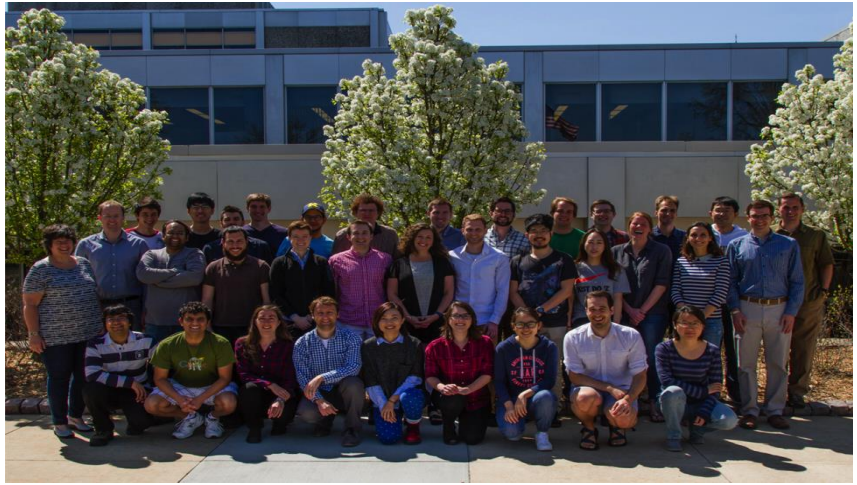
- prioritized exploration in high dimensional spaces
- distribution across systems



• Relational Schema by Becris from the Noun Project

Thank you!

signac is a product of The Glotzer Group at the University of Michigan. Development and deployment supported by MICCoM, as part of the Computational Materials Sciences Program funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, under Subcontract No. 6F-30844. Project conceptualization and implementation supported by the National Science Foundation, Award # DMR 1409620.



www.signac.io