

Theory & Practice of Data Cleaning

Workflow Automation and Provenance

Reproducibility Crisis (hitting the airwaves)

The Economist (October 2014) - "How Science Goes Wrong": A satirical article from The Economist's October 2014 issue, featuring a cover with the title "HOW SCIENCE GOES WRONG".

Retraction Watch: A screenshot of a website titled "Retraction Watch" showing an article about a chemistry paper retraction in the journal Science.

BBC Radio 4 - "Science has a 'reproducibility crisis'": A screenshot of a BBC Radio 4 website page discussing the reproducibility crisis, featuring a photo of several test tubes.

BBC Radio 4 - "When the Revolution Came for Amy Cuddy": A screenshot of a BBC Radio 4 website page featuring a photo of Amy Cuddy sitting on a stool.

- e.g. Science, Economist, Nature (via SciAM):
 - <https://www.scientificamerican.com/video/is-there-a-reproducibility-crisis-in-science/>
- BBC4 report :
 - Causes: culture that rewards novel, eye-catching results, i.e.,
 - inappropriate statistics
 - selective reporting
 - hyping



Computational Reproducibility

- Different sciences have different reproducibility crises ...
- Our focus: **Computational Reproducibility**
 - **Scientific workflows**,
 - ... and **scripts**: R, Matlab, Python, ..
- *How to facilitate reproducibility for computational and data scientists?*
 - **Workflow Automation** (workflow systems, scripts)
 - **Transparency**
 - **Provenance**

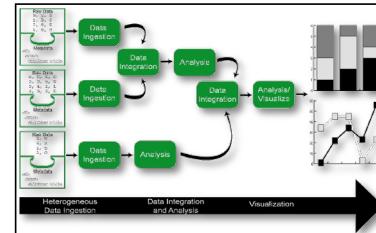
Reproducibility Studies

- **Successful** reproducibility study:
 - increases trust in prior study 😊
 - ... but no surprises 😕
- **Failed** reproducibility study :
 - decreases trust (or falsifies) prior study 😥
 - ... but surprising failure yields new info/knowledge 😊
- Learning from failures!
 - Not really a new, revolutionary idea..
 - What is a positive vs negative result anyways?
 - ... *fail early, fail often* ...

Scientific Workflows: ASAP

- **Automation**

- wfs to **automate** computational aspects of science



- **Scaling** (exploit and optimize *machine* cycles)

- wfs should make use of **parallel compute resources**
- wfs should be able handle **large data**



- **Abstraction, Evolution, Reuse** (*human* cycles)

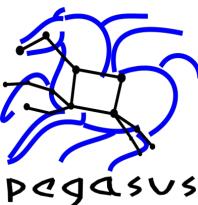
- wfs should be easy to **(re-)use, evolve, share**



- **Provenance**



- wfs should capture **processing history, data lineage**
- traceable data- and wf-evolution
- **Reproducible Science**



Trident
Workbench
Es war einmal ...

Essential Functions of a Scientific Workflow System

1. Automate programs and services scientists already use.
2. Schedule invocations of programs and services correctly and efficiently – in parallel where possible.
3. Manage dataflow to, from, and between programs and services.
4. Enable scientists (not just developers) to author or modify workflows easily.
5. Predict what a workflow will do when executed: **prospective provenance**.
6. Record what happened during workflow execution: **retrospective provenance**.
7. Reveal and query provenance – how workflow products were derived from inputs via programs and services.
8. Organize intermediate and final data products as desired by users.
9. Enable scientists to version, share and publish their workflows.
10. Empower scientists who wish to automate additional programs and services themselves.

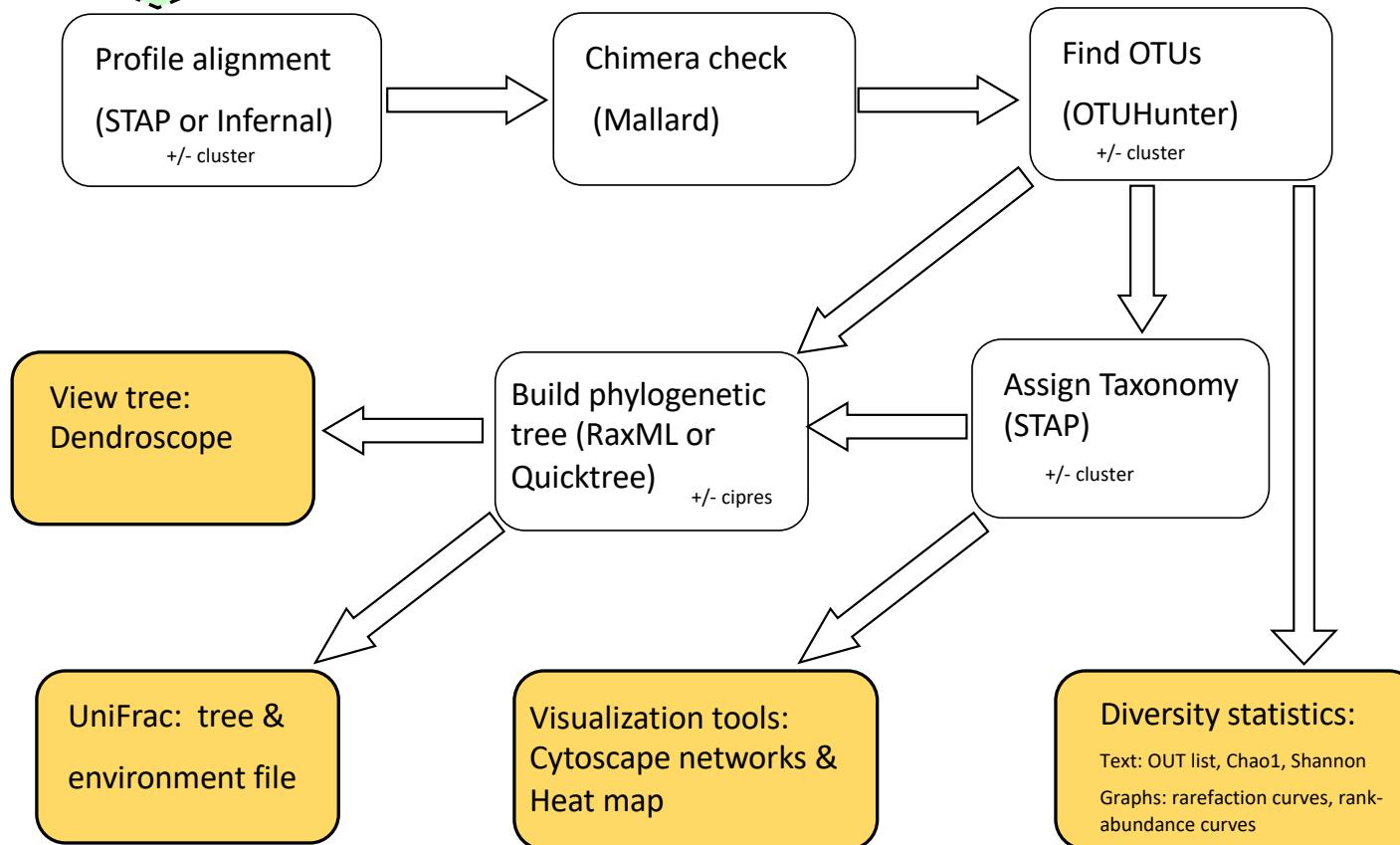
These functions (not just dataflow & actors) distinguish **scientific workflow automation** from general (scientific) software development.

Src: Tim McPhillips

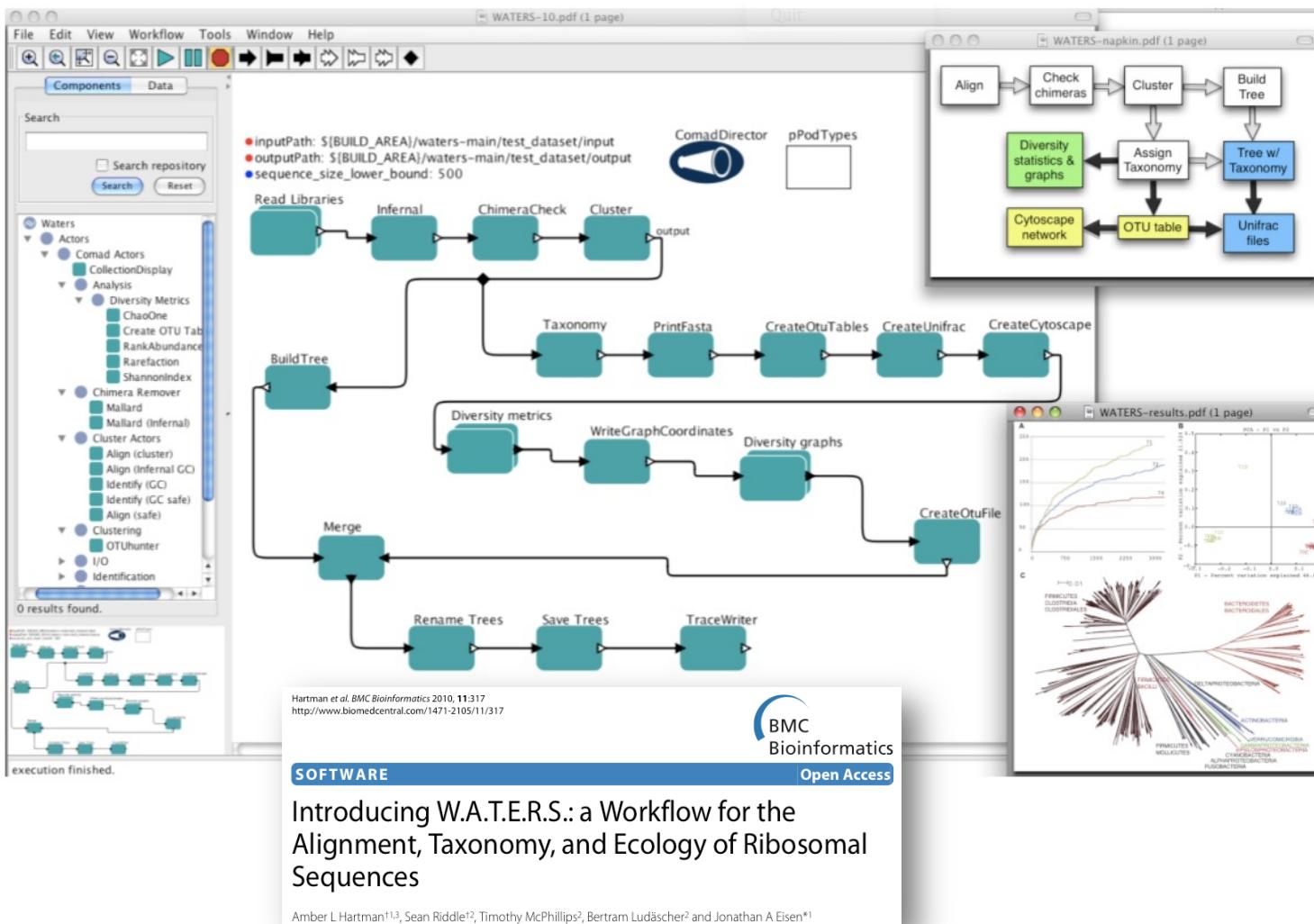
Assembled
contigs

WATERS:

Workflow for Alignment, Taxonomy,
Ecology of Ribosomal Sequences
(Amber Hartman; Eisen Lab; UC Davis)



Executable WATERS Workflow in Kepler



Example Bioinformatics Workflow: *Motif-Catcher*

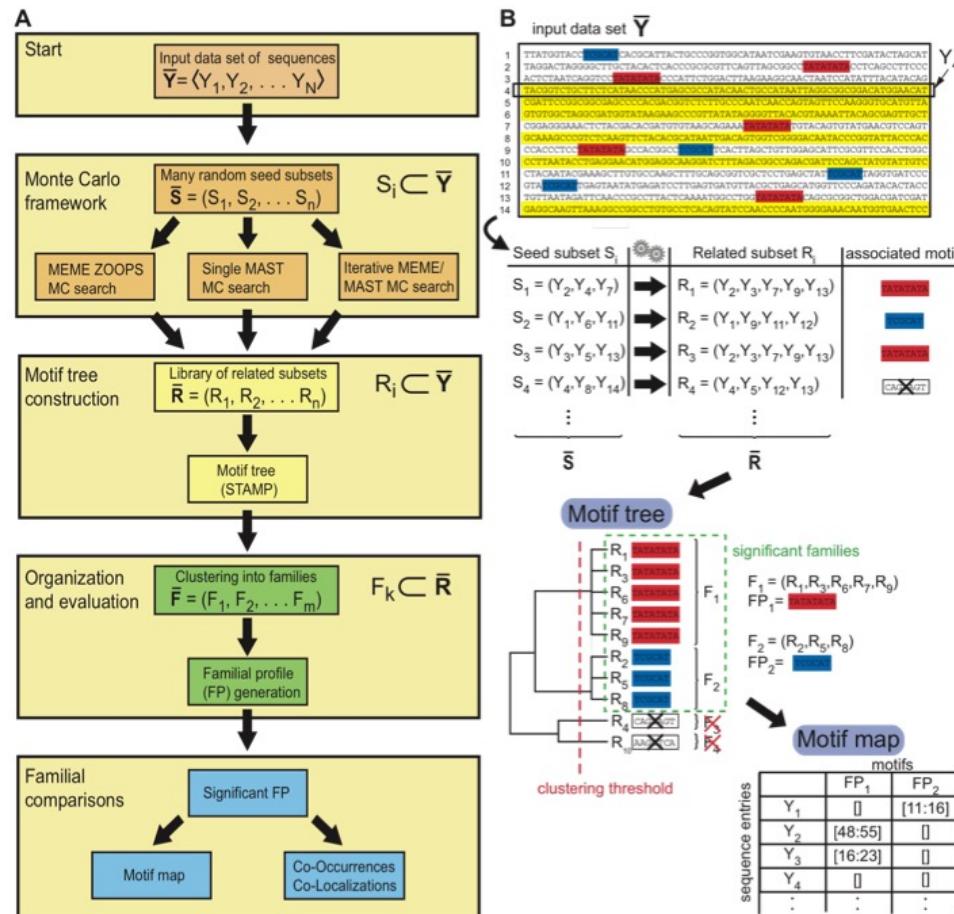
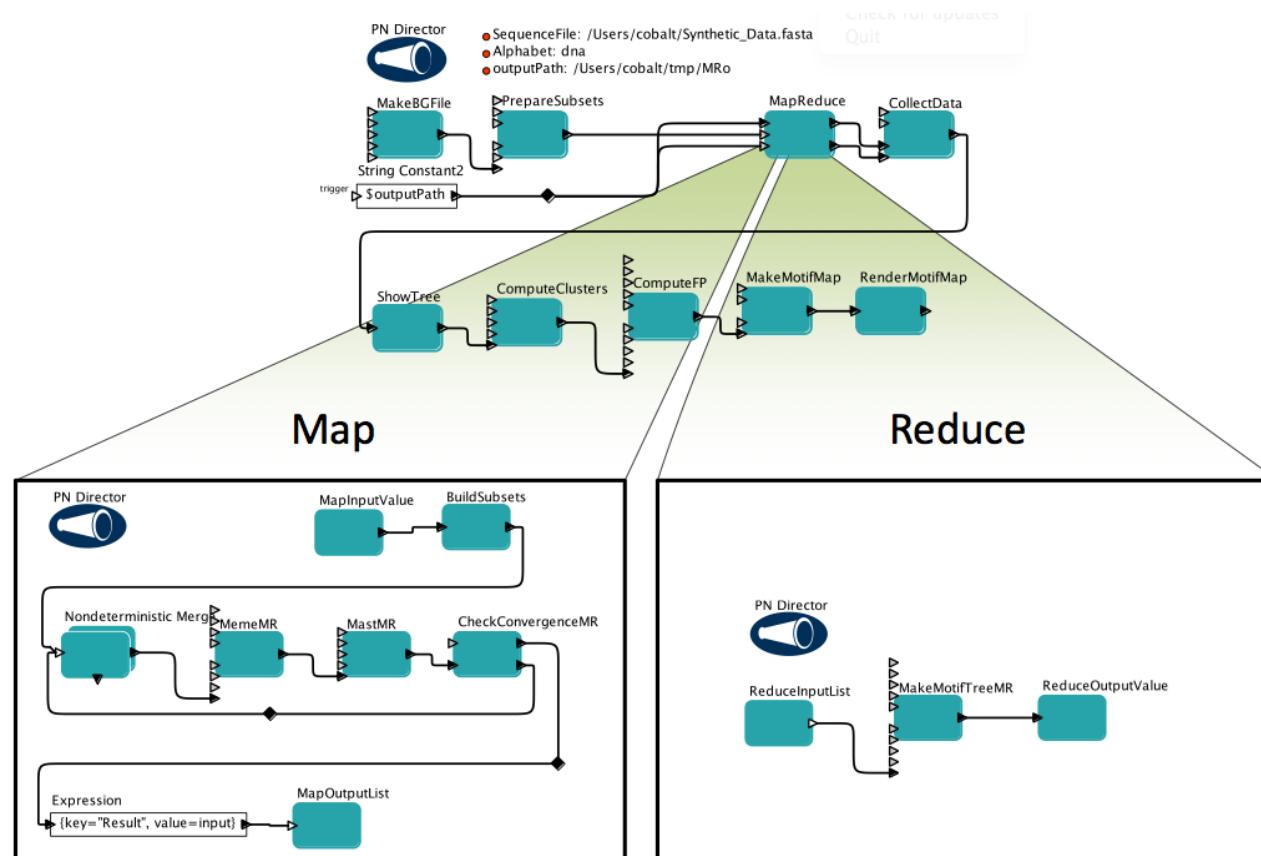


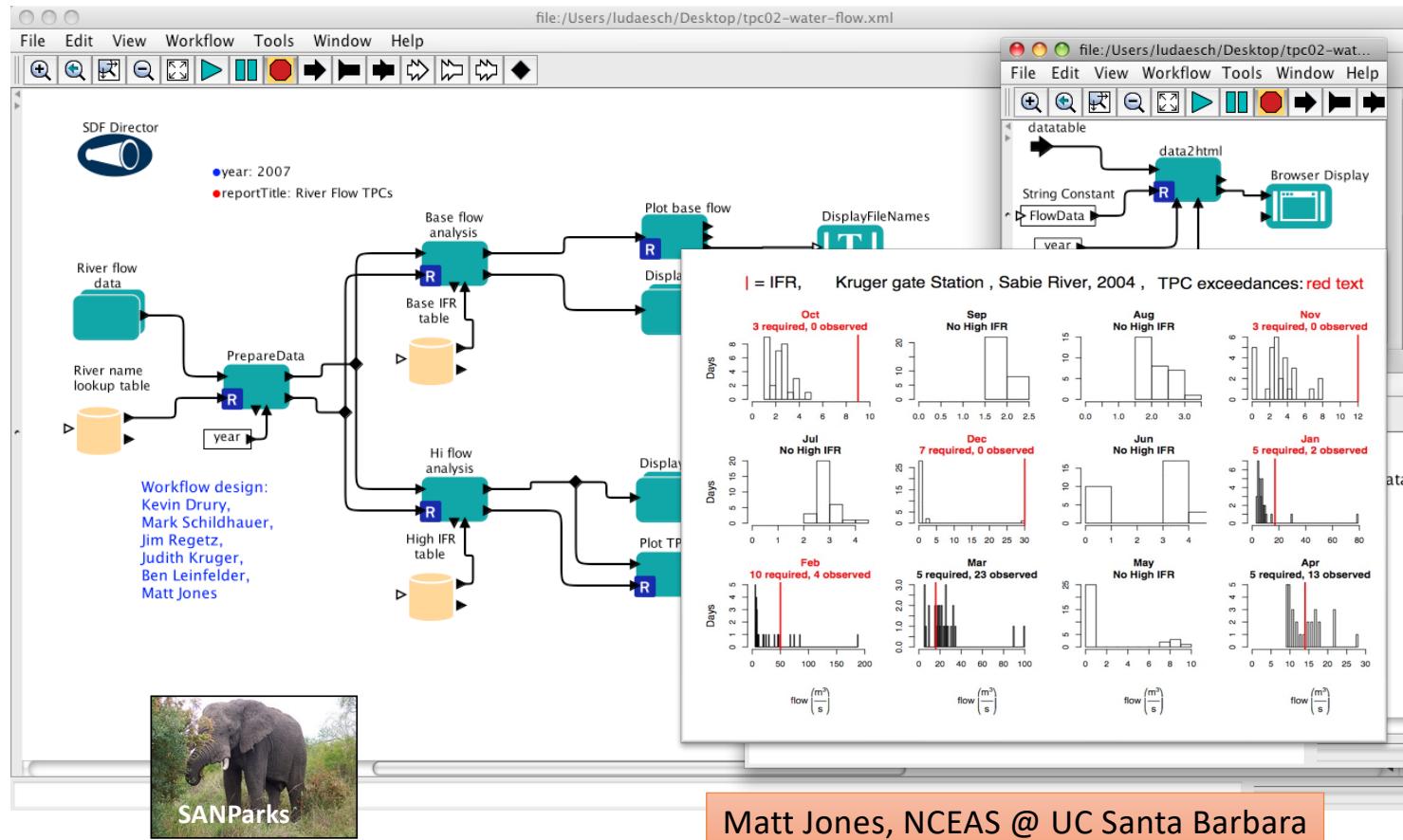
Figure 1: Concept of Monte-Carlo based detection and interpretation of motifs.
A) Abstract description of MotifCatcher process. B) Examples illustrating the process with sample data.

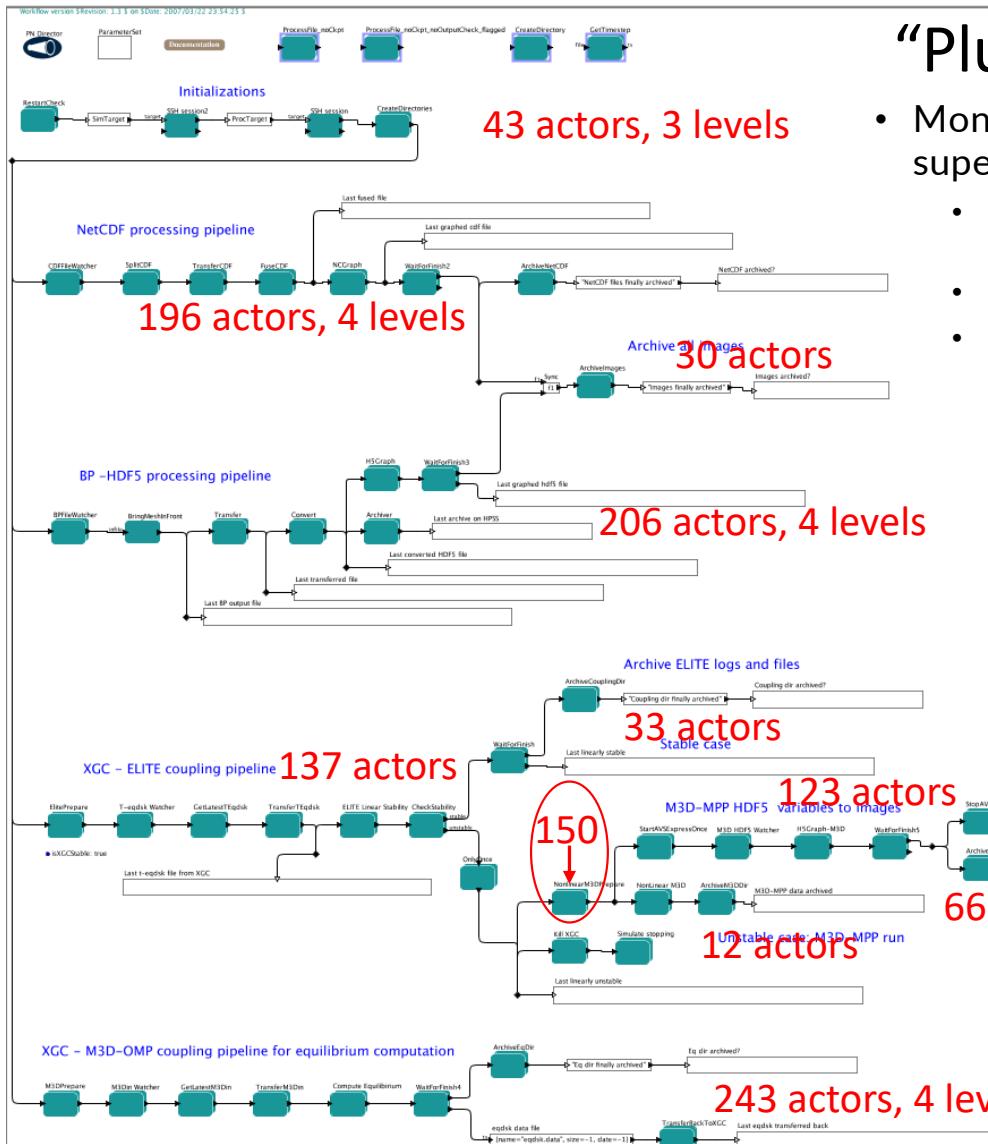
Motif-Catcher workflow, implemented in Kepler



S Köhler et al. Improved Motif Detection in Large Sequence Sets with Random Sampling in a Kepler workflow, ICCS-WS, 2012

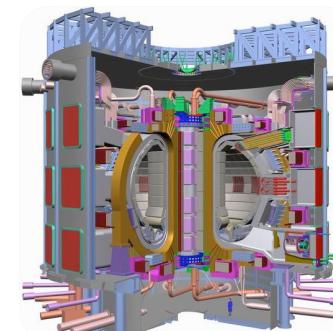
Kepler Workflows & Decision Making (Kruger Natl. Park, South Africa)





“Plumbing” workflow

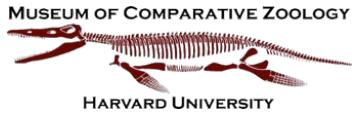
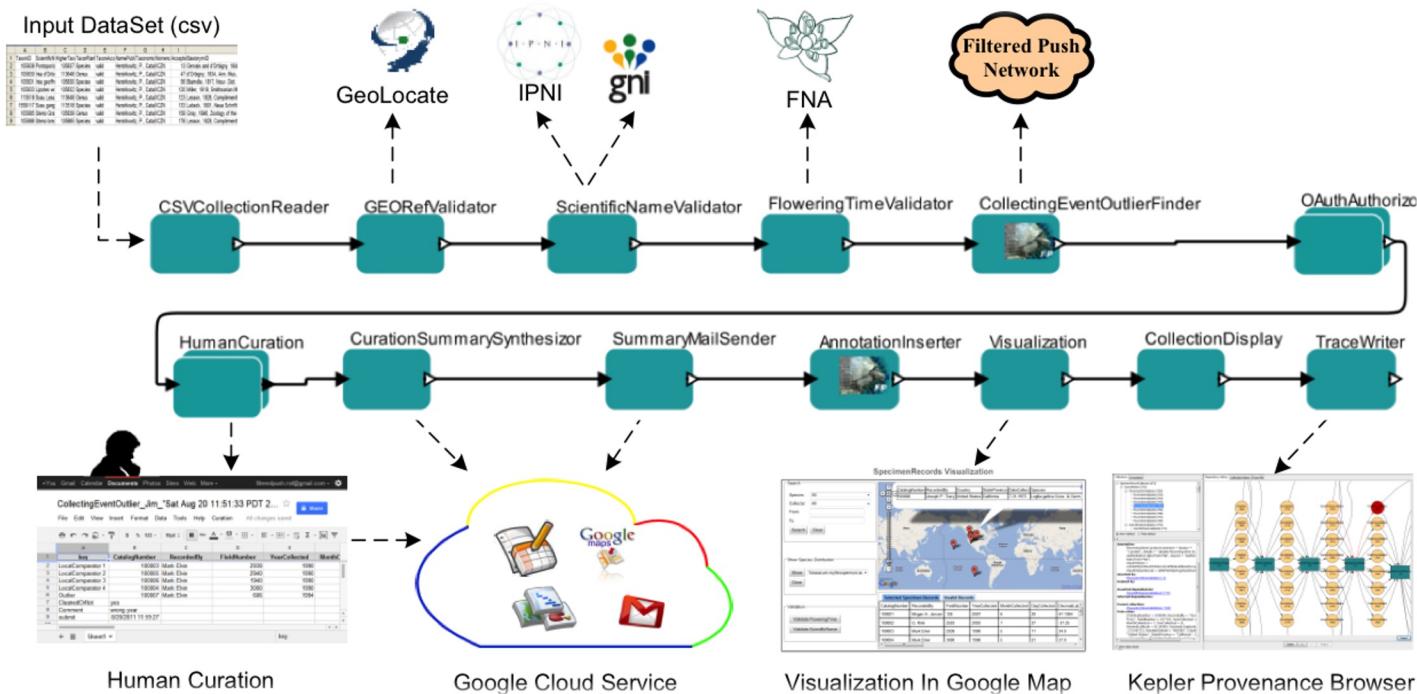
- Monitor and control supercomputer simulations
- 50+ composite actors (subworkflows)
- 4 levels of hierarchy
- 1000+ atomic (Java) actors



Norbert Podhorszki
ORNL (then: UC Davis)

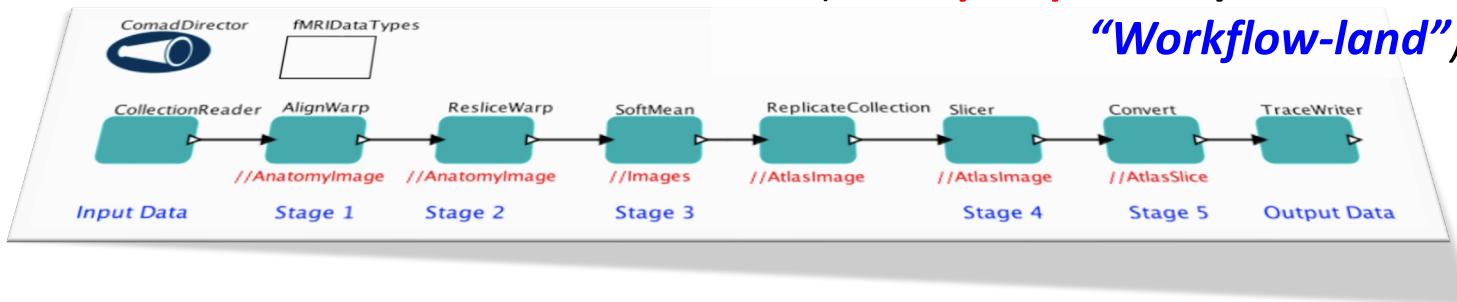
Data Curation Workflows

(Filtered-Push ... Kepler ... Kurator projects)

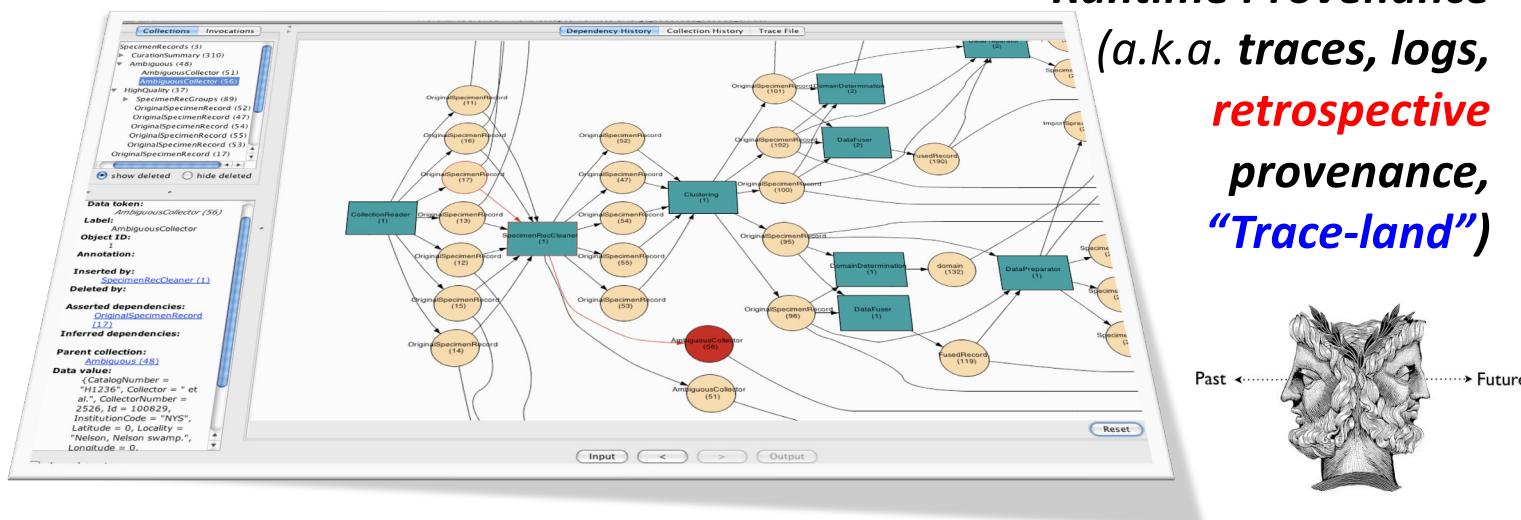


Provenance \leftrightarrow Workflows

Workflow Modeling & Design
*(a.k.a. **prospective** provenance)*
"Workflow-land"



Runtime Provenance
*(a.k.a. **traces**, **logs**,
retrospective provenance,
"Trace-land")*



Provenance Standards vs Tools

- Do we need more standards to sort this out?
 - ... or do we already have too many “standards”?
- How should we **think** about provenance?
 - ... in workflows and databases?
- What can we **do** with provenance?
 - ... in workflows and databases?
- Tools to create, share, **use** provenance
 - ... not just for “provenance for others”
 - ... need more “**provenance for self**”

A simple (simplistic?) World View: *Workflow-Land* \leftrightarrow *Trace-Land*

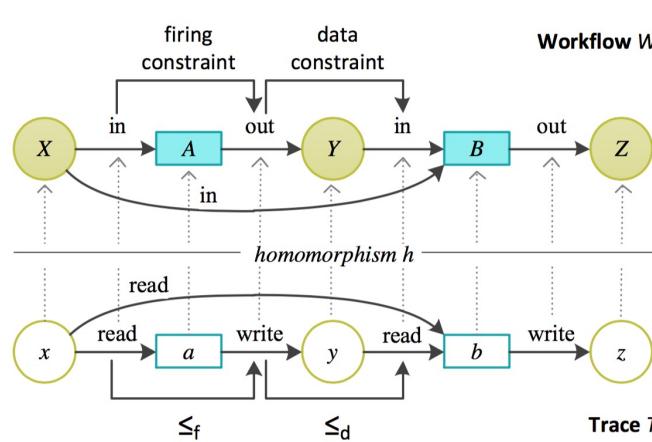


Figure 6: A homomorphism h from trace T to workflow W guarantees structural validity. Workflow-level constraints induce temporal constraints \leq_f and \leq_d on traces [DKBL12].

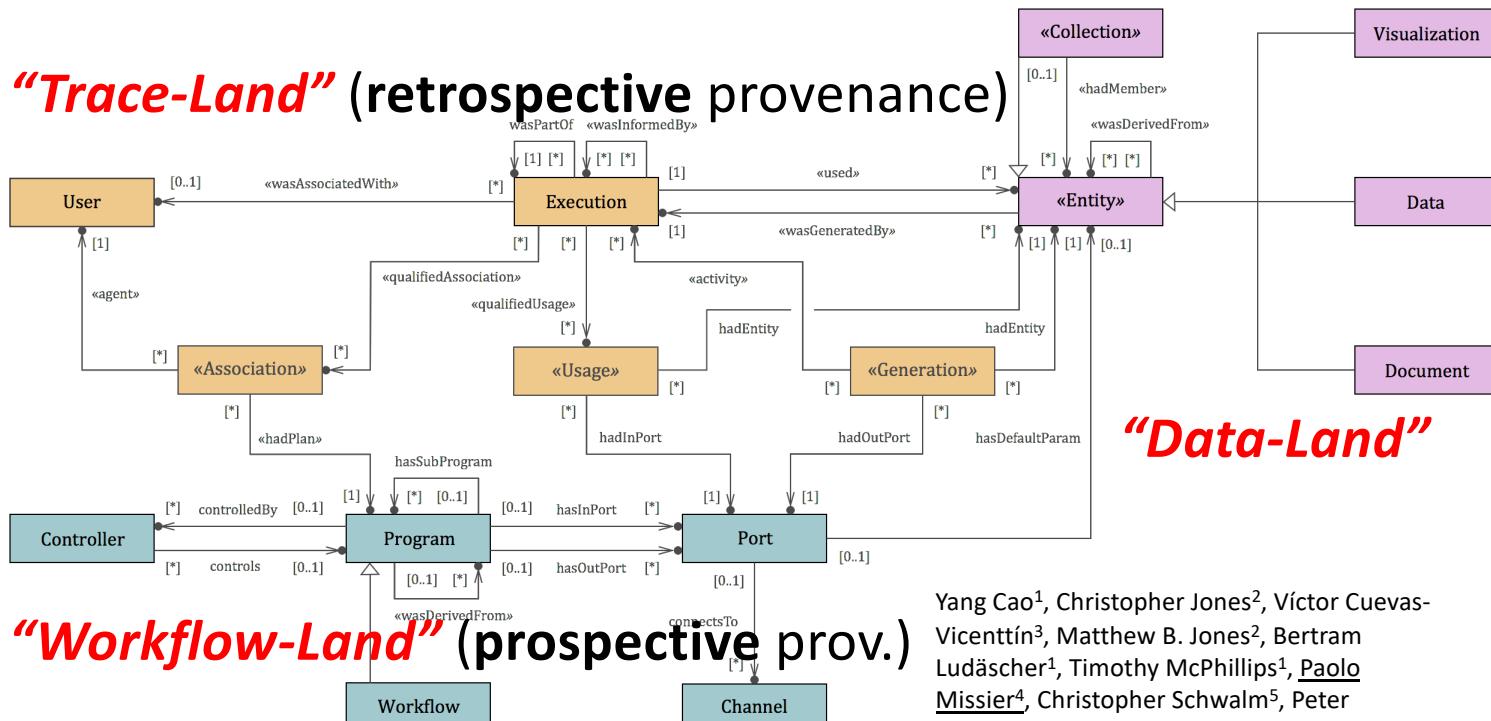
- Not a “standard” – but helps (at least me...) think about **workflows** (**prospective** provenance) and **traces** (**retrospective** provenance).

ProvONE: PROV for scientific workflows

(Transfer station to any of several other “standard extensions”)



“Trace-Land” (retrospective provenance)



“Data-Land”

Yang Cao¹, Christopher Jones², Víctor Cuevas-Vicentín³, Matthew B. Jones², Bertram Ludäscher¹, Timothy McPhillips¹, Paolo Missier⁴, Christopher Schwalm⁵, Peter Slaughter², Dave Vieglais⁶, Lauren Walker², Yaxing Wei⁷

¹University of Illinois, Urbana-Champaign, ²National Center for Ecological Analysis and Synthesis, UCSB, ³Universidad Popular Autónoma del Estado de Puebla, Mexico, ⁴School of Computing Science, Newcastle University, UK, ⁵Woods Hole Research Center, Falmouth, MA, ⁶University of Kansas, Lawrence, ⁷Environmental Sciences Division, Oak Ridge National Lab, TN

Also: A. Marinho, L. Murta, C.Werner, V.Braganholo, S. Serra da Cruz, E.Ogasawara, M. Mattoso. “ProvManager: A Provenance Management System for Scientific Workflows.” Concurrency and Computation: Practice and Experience 24, no. 13 (2012): 1513–1530
...

But ... how to prime the provenance pump?? Must support “Provenance for Self” !

The diagram shows a flow from a scientist working at a desk (labeled 'Provenance for Self?!') to a public climate report (labeled 'Provenance for Others').

Provenance for Self?!

Provenance for Others

Texas Summer 2011: Record Heat and Drought

Cooperative Institute for Climate and Satellites - NC
Laura Stevens

The time range for this image is January 01, 1995 (00:00 AM) to December 31, 2012 (00:00 AM). This image was created on July 03, 2013.

The spatial range for this image is 25.83° to 36.50° latitude, and -106.65° to -93.52° longitude.

Attributes : Temperature, precipitation, observed, Texas.

This image was derived from dataset nca3-cddv2-r using the activity 02c53cf7-nca3-cddv2-r1-process.

This image is part of this figure :

data and "code"/method linked

alt formats

You are viewing /image/02c53cf7-75fb-4243-a925-f59a0025f04e in HTML.

Alternatives : JSON YAML Turtle N-Triples JSON Triples RDF-XML RDF-JSON Graphviz SVG

GlobalChange.gov

Climate Change Impacts in the United States

U.S. National Climate Assessment

The diagram illustrates the flow of provenance information from a specific scientific finding (Texas Summer 2011 drought) to a broader, publicly accessible report on climate change impacts.

- ✓ Provenance *capture* (Matlab, R, Python, ... scientific workflow systems)
Uploading, *sharing, linking* provenance through various provenance tools
- ✗ Tools for scientists to *exploit* (= *capture, share, link*) provenance for their own day-to-day work.
- ➔ Prime the provenance pump and **increase provenance generation**
- ➔ Scientists **accelerate their work via new, active uses of provenance.**

From Workflows & Provenance to Provenance for Script-based Workflows ...

- What workflow tools are (most) scientists using?
 - Workflow systems
 - ... vs scripts (Python, R, MATLAB, ...)
- What provenance tools are there?
 - Workflow system support
 - Tools for “workflow” scripts!?

Yes, scripts are (can be) workflows too!

Reproducible academic publications

This section contains academic papers that have been published in the peer-reviewed literature or pre-print sites such as the ArXiv that include one or more notebooks that enable (even if only partially) readers to reproduce the results of the publication. If you include a publication here, please link to the journal article as well as providing the notebook link (and any other relevant resources associated with the paper).

- Automatic segmentation of odor maps in the mouse olfactory bulb using regularized non-negative matrix factorization, by J. Soeller et al. (NeuroImage 2014, Open Access). The notebook allows to reproduce most figures from the paper and provides a deeper look at the data. The full code repository is also available.
- Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss, by A. Gross et al. (Nature Genetics 2014). The full collection of notebooks to replicate the results.
- powerlaw: a Python package for analysis of heavy-tailed distributions, by J. Alstott et al.. Notebook of examples in manuscript, ArXiv link and project repository.
- Collaborative cloud-enabled tools allow rapid, reproducible biological insights, by B. Ragan-Kelley et al.. The main notebook, the full collection of related notebooks and the companion site with the Amazon AMI information for reproducing the full paper.
- A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data, by C.T. Brown et al.. Full notebook, ArXiv link and project repository.
- The kinematics of the Local Group in a cosmological context by J.E. Forero-Romero et al.. The Full notebook and also all the data in a github repo.
- Warming Ocean Threatens Sea Life, an article in Scientific American backed by a notebook, `main.ipynb`. By Roberto de Almeida from Mariana Forde, via a notebook, `main.ipynb`: `display(i)`

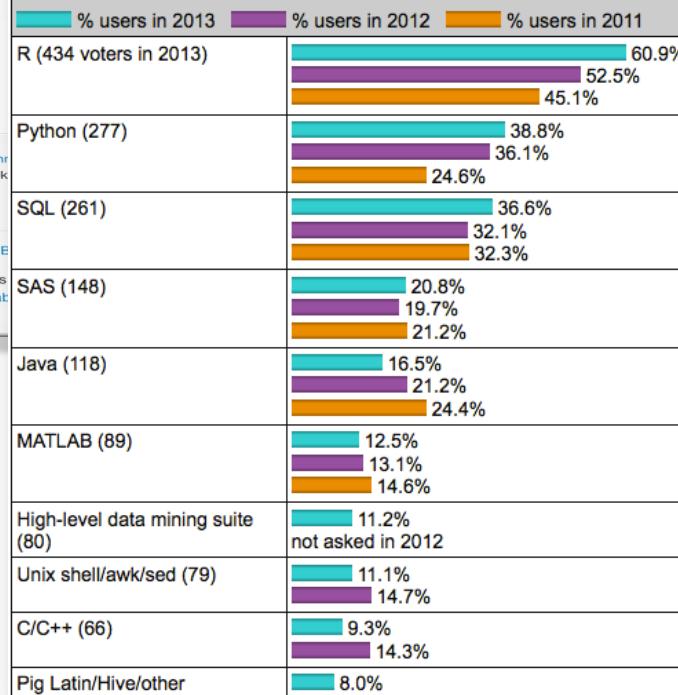


```
In [3]: from IPython.display import SVG  
SVG(filename='python-logo.svg')  
  
Out[3]:
```

The Python logo is a stylized yellow and blue icon resembling a snake.

- Data-driven journalism**
- AtomPy: An Open Atomic Data Curation Environment for Applications, by C. Mendoza, J. Boswell, D. Ajok
 - The Need for Openness in Data Journalism, by E. St. Louis County Segregation Analysis , analysis Area Is Even More Segregated Than You Probably Think, by Singer-Vine.

What programming/statistics languages you used for an analytics / data mining / data science work in 2013? [713 votes total]



SKOPE: Synthesized Knowledge Of Past Environments

Bocinsky, Kohler *et al.* study rain-fed maize of Anasazi

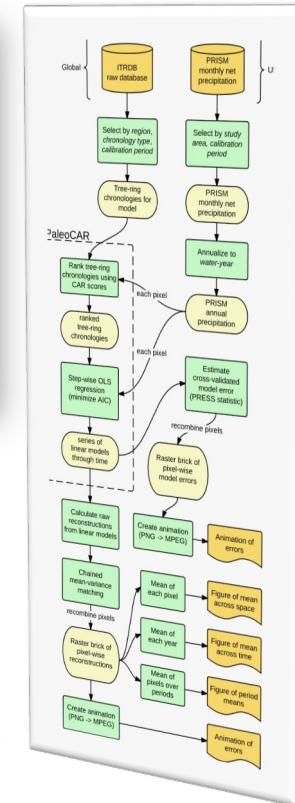
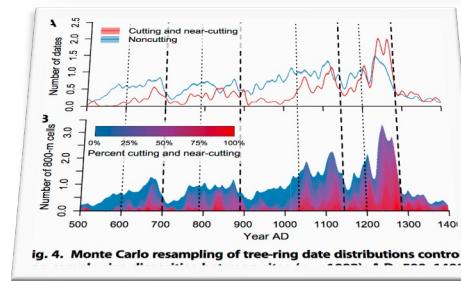
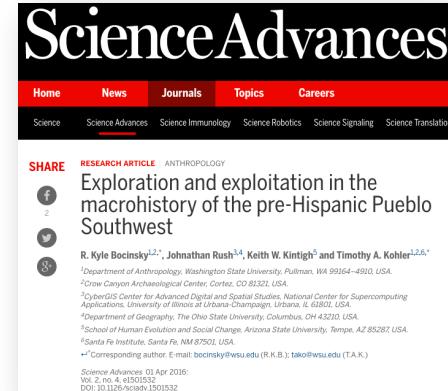
- Four Corners; AD 600–1500. Climate change influenced Mesa Verde Migrations; late 13th century AD. Uses network of tree-ring chronologies to reconstruct a spatio-temporal climate field at a fairly high resolution (~800 m) from AD 1–2000. Algorithm estimates joint information in tree-rings and a climate signal to identify “best” tree-ring chronologies for climate reconstructing.



Provenance Support for Reproducible Science Example: Paleoclimate Reconstruction

Science paper (OA) uses:

- open source code:
 - R, PaleoCAR, ...
- Is that all we need?
- What was the “workflow”?
- Is there prospective and/or retrospective provenance?



Final Tour Stop: YesWorkflow: Yes, scripts are workflows, too!

```
203 ## Gene Ontology Statistics are Calculated Here.  
204  
205 # Gene Ontology Categories that were shown to be relatively Higher (more expressed) in the Experimental Condition.  
206 gosstatshigher <- higheridrlinkedtogenes[1]  
207 higherstatsfilename <- paste(outputDirectory, "/", runName, "_", conditions[1], "_GOSTatsHigher_", mytest, "[1], "_v  
208 write.table(gosstatshigher, file=higherstatsfilename, row.names=FALSE, col.names=FALSE, quote=FALSE, sep="\t"  
209 geneListHigherCHR <- gosstatshigher$SYMBOL  
210 geneListHigherLinkedtoEntrezIds <- select(hgu133plus2.db, keys= geneListHigherCHR, "ENTREZID", "SYMBOL")  
211 GOSTatsGenesh <- geneListHigherLinkedtoEntrezIds[,2]  
212  
213 x <- org.Hs.egACCNUM  
214 mapped_genes <- mappedkeys(x)  
215 xx <- as.list(x[mapped_genes])  
216 geneUniverse <- (unique(names(xx)))
```

?

- **Script vs Workflows/ASAP:**
 - **Automation:** ****
 - **Scaling:** **
 - **Abstraction:** *
 - **Provenance:** **

YW annotations: Model your Workflow!

```
1 # @BEGIN collect_data_set
2 # @PARAM cassette_id @PARAM accepted_sample @PARAM num_images @PARAM energies
3 # @OUT sample_id @OUT energy @OUT frame_number
4 # @OUT raw_image_path @AS raw_image
5 # ... @URI file:run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number}.raw
6 run_log.write("Collecting data set for sample {}".format(accepted_sample))
7 sample_id = accepted.sample
8 for energy, frame_number, intensity, raw_image_path in collect.next_image(
9         cassette_id, sample_id, num_images, energies,
10        "run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number:03d}.raw"):
11    run_log.write("Collecting image {}".format(raw_image_path))
12 # @END collect_data_set
13
14 # @BEGIN transform_images
15 # @PARAM sample_id @PARAM energy @PARAM frame_number
16 # @IN raw_image_path @AS raw_image
17 # @IN calibration_image @URI file:calibration.img
18 # @OUT corrected_image @URI file:run/data/{sample_id}/{sample_id}-{energy}eV.{frame_number}.img
19 # @OUT corrected_image_path @OUT total_intensity @OUT pixel_count
20     corrected.image_path = "run/data/{}/{}/{}eV_{}{:03d}.img".format(sample_id, energy, frame_number)
21     (total_intensity, pixel_count) = transform.image(raw_image_path, corrected.image_path, "calibration.img")
22     run_log.write("Wrote transformed image {}".format(corrected.image_path))
23 # @END transform_images
```

mark the code block

... and data inputs/outputs

Figure 1. YW-annotated fragment of a Python script for data collection from protein crystal samples. YW-annotations `@BEGIN` and `@END` delimit code blocks; `@IN` and `@OUT` tags model relevant input and output data elements of a block; `@PARAM` identifies a block's parameters. `@URI` templates for raw images (line 5) and corrected images (line 18) link conceptual-level data elements such as `raw_image` with runtime resources (data files and their file paths). Executable script code is greyed out to emphasize YW-annotations. A program variable (`raw_image_path`) is highlighted in the code (lines 8, 11, 21); aliases (lines 4, 16) are used to link such program-level objects to the scientist's concepts (here: `raw_image`). Full example available from [MBL15].

Paleoclimate Reconstruction (EnviRecon.org) ...

Kyle B.,
(computational)
archaeologist:

"It took me about 20 minutes to comment. Less than an hour to learn and YW-annotate, all-told."

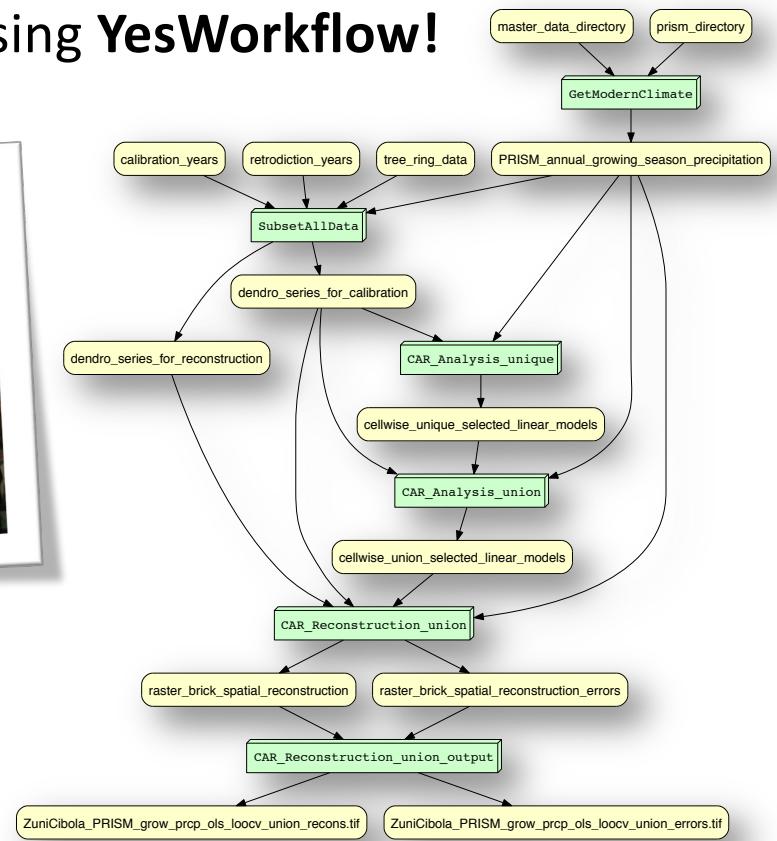


... explained using **YesWorkflow!**

SKOPE + **Kurator**

+ **DataONE**
Data Observation Network for Earth

=> **YesWorkflow.org**



Get 3 views for the price of 1!

