Provenance, Workflows, and Reproducibility - *Cui Bono*?

Bertram Ludäscher

Director, Center for Informatics Research in Science & Scholarship (CIRSS)

School of Information Sciences

University of Illinois, Urbana-Champaign

`ludaesch@illinois.edu`

School of Information Sciences
The iSchool at Illinois

NCSA

What's up on Day 1? *(evolving)*

- On DataONE (Bill M)
- On Provenance (yours truly)
- Status/overview of provenance tools (Matt J)
- Goals, agenda check (Dave V, Amber, Kyle B)
- *Break*
- Introductions come first (sort of)
- *Lunch*
- Tool demos & tutorials (hands-on!)
- *… Day 2 …*

https://github.com/DataONEorg/provathon-2017

https://dataoneorg.slack.com #prov-a-thon

A quick *Tour de Provenance*

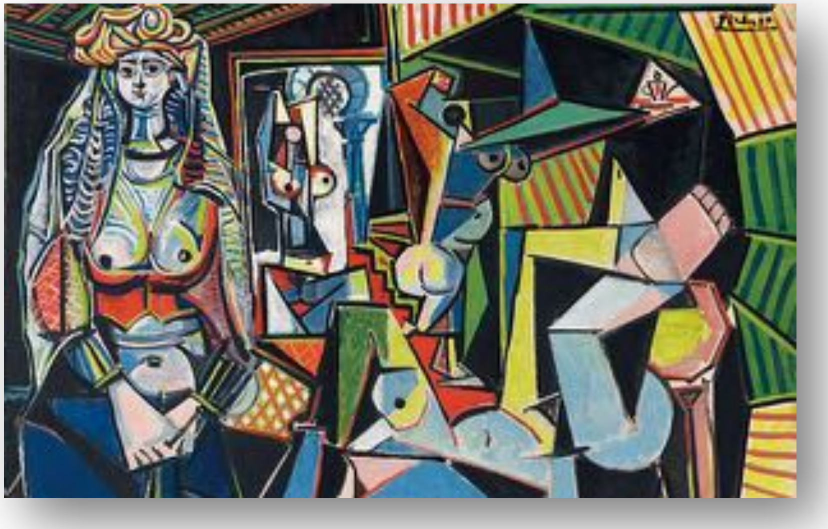
- **Provenance**
  - *... by whom, for whom, for what, how-to*
  - **Prospective** provenance (≈ scientific workflows)
  - **Retrospective** provenance (≈ runtime events, traces)

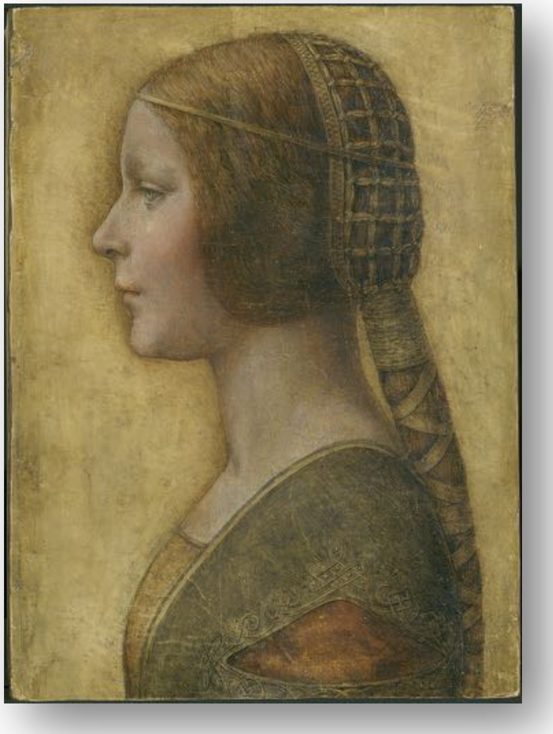
  ➔ **Hybrid** provenance ...
- **Reproducibility** & **Transparency**
  - *... of what and for what?*
- **... projects**: DataONE, Whole Tale, SKOPE, ...

- **... and tools:** recordR, YesWorkflow, WT
- My provenance:

CS/DB@{KIT,Freiburg} .. SDSC .. UC Davis ... {iSchool,NCSA,CS}@UIUC

Provenance: The Million $$$ Question ..

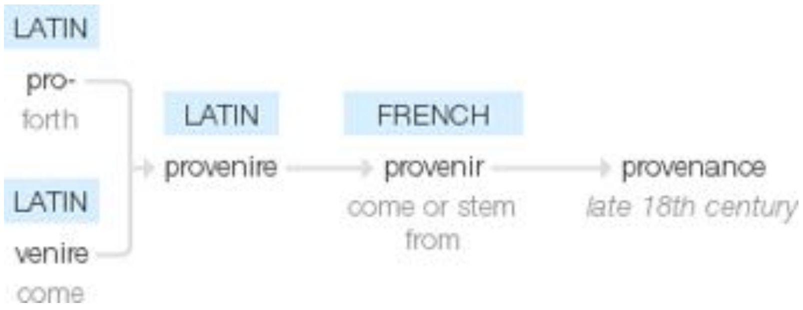




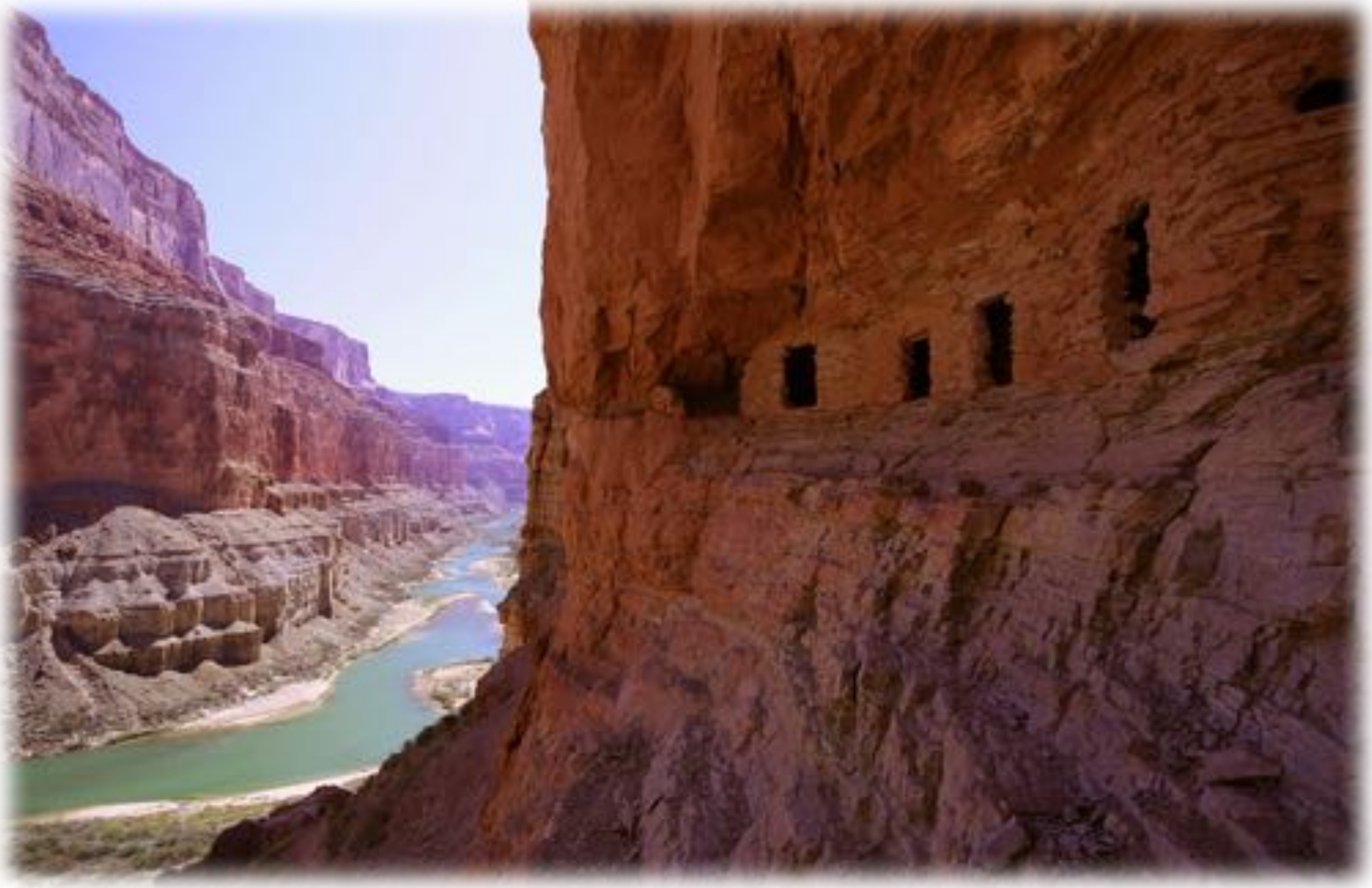
- One of these is has been sold for nearly $180 million.
- The other *could* be worth as much or more.
- Which is which?
- What is the difference?

Provenance defined …

- **Oxford English Dictionary**
  - The **place of origin** or earliest **known history** of something:
    - *an orange rug of Iranian **provenance***
  - The **beginning** of something's existence; its **origin**:
    - *they try to understand the whole universe, its **provenance** and fate*
  - A **record of ownership** of a work of art or an antique, used as a guide to **authenticity** or **quality**:
    - *the manuscript has a distinguished **provenance***
- **Similar but different: *Provenience***



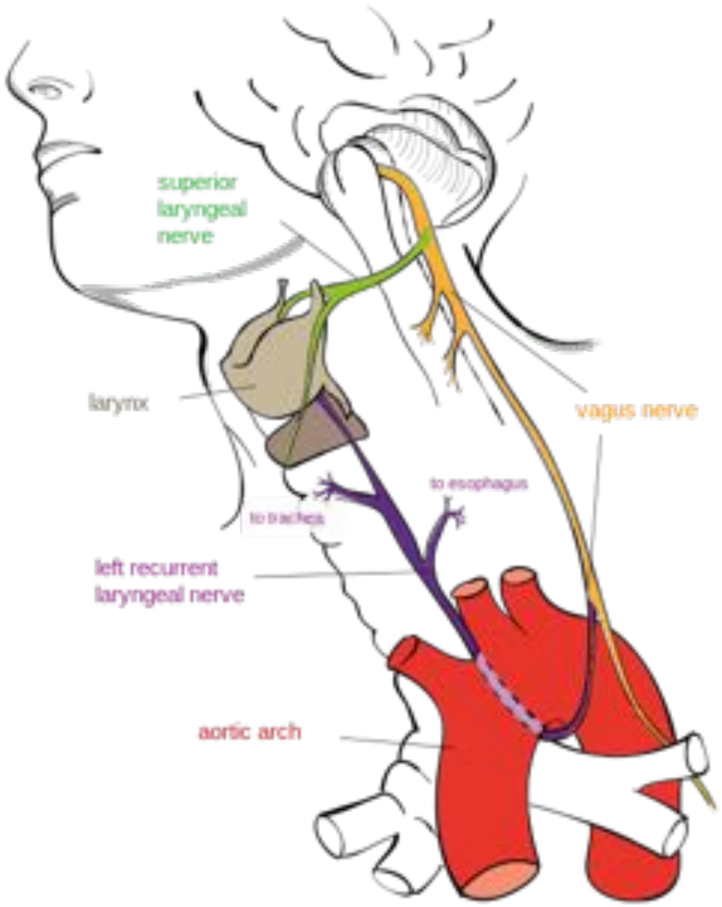
Provenance in the Science Sciences



- Can you “see provenance” in this image?
- Grand Canyon’s rock layers are a record of the early **geologic history** of North America. The ancestral puebloan granaries at Nankoweap Creek tell archaeologists about more recent **human history**. (By Drenaline, licensed under CC BY-SA 3.0)

Science Science: Biology & Natural History

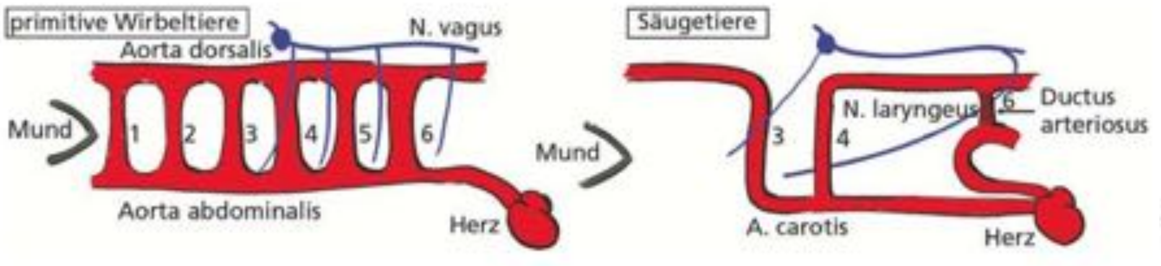
Provenance = Understanding what happened…



Author: Jkwchui (Based on drawing by Truth-seeker2004)

Zrzavý, Jan, David Storch, and Stanislav Mihulka. Evolution: Ein Lese-Lehrbuch. Springer-Verlag, 2009.

334 5 Adaptation

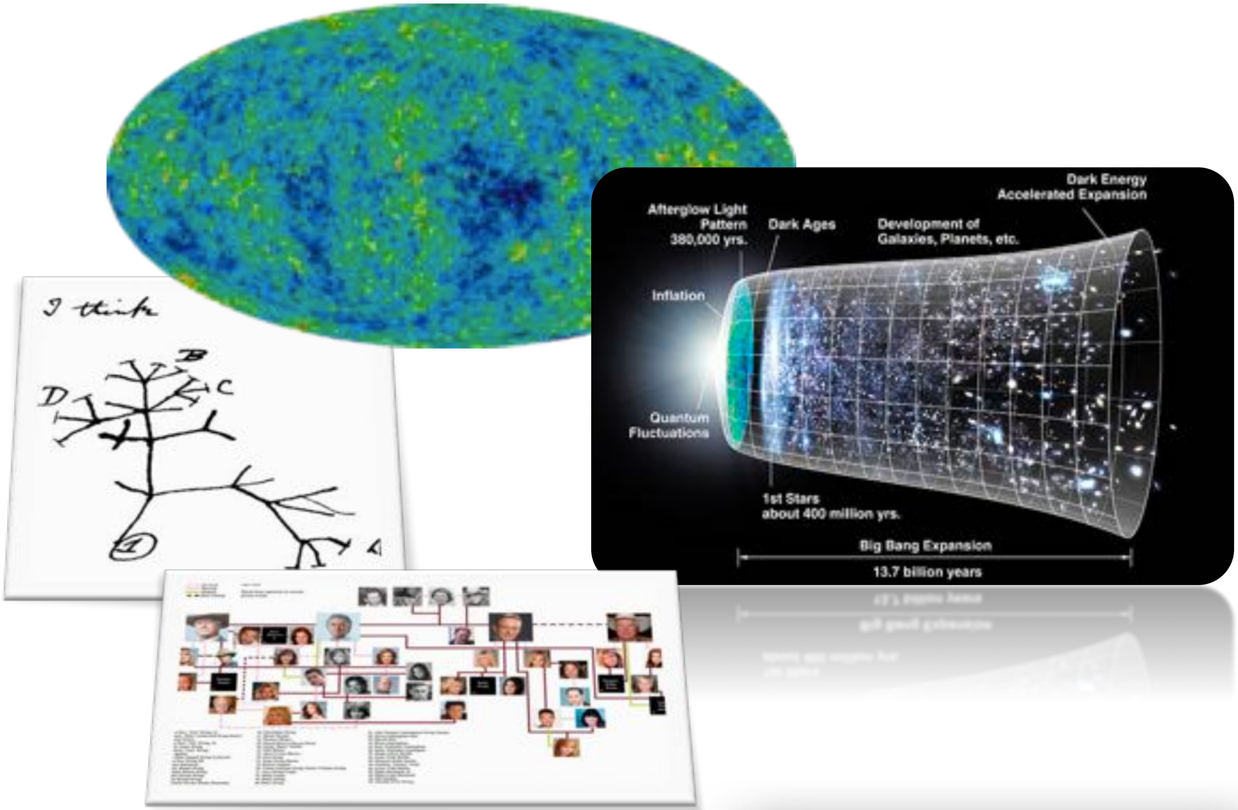


5.16 Evolution der Schleife des rückläufigen Kehlkopfnervs (Nervus laryngeus recurrens) der Wirbeltiere. Dieser Nerv stellt den vierten Ast des Nervus vagus dar. Bei ursprünglichen Wassertieren sandte der Vagusnerv seine Äste zu den Kiemenarterien, die die Bauch- und die Rückenaorta verbanden. Während der Phylogenese der Wirbeltiere haben sich allerdings die Kiemenbögen und mit ihnen auch die Kiemenarterien verändert und das Herz wurde nach kaudal verschoben. Aus der sechsten Arterie wurde bei den Säugetieren der Ductus arteriosus; der vierte Ast des Vagus, der heute den Kehlkopf (Larynx) innerviert, liegt stets *hinter* der ehemaligen sechsten Arterie, also hinter dem Ductus arteriosus. Daher führt dieser Nerv vom Gehirn aus nach hinten, windet sich unter dem Ductus hindurch und kehrt nach vorne zurück, um den Larynx zu innervieren.

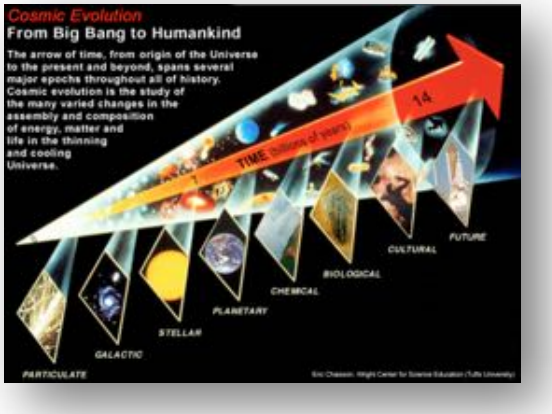
5.17 Suboptimale evolutionäre Konstruktionslösung:

Provenance in Science- Sciences Palooza

- What are those?
- *Cosmology*
- *Geology, Stratigraphy*
- *Phylogeny*
  - the *Tree of Life*
- *Genealogy*
  - your family: literally
- *Academic Pedigree*
  - *“Doktorvater” (oder Doktor-Mutter)*
- *Etymology*
- *Chain of custody*
  - of art(ifacts)
- Yes: all about **origins** and **history** …



Thibaut
Pfaff
Möbius
Gudermann
Dirksen
Gauß
Dedekind
Weierstraß
Jacobi
Gerling
Kovalevskaya
Gordan
Plücker
Noether
Euler
Lagrange
Laplace
Fourier
Poisson
Dirichlet
Lipschitz
Klein
Lindemann
Furtwängler
Hilbert


Cosmic Evolution
From Big Bang to Humankind
The arrow of time, from origin of the Universe to the present and beyond, spans several major epochs throughout all of history. Cosmic evolution is the study of the many varied changes in the assembly and composition of energy, matter and life in the thinning and cooling Universe.
TIME (billions of years)
1
14
PARTICULATE
GALACTIC
STELLAR
PLANETARY
CHEMICAL
BIOLOGICAL
CULTURAL
FUTURE

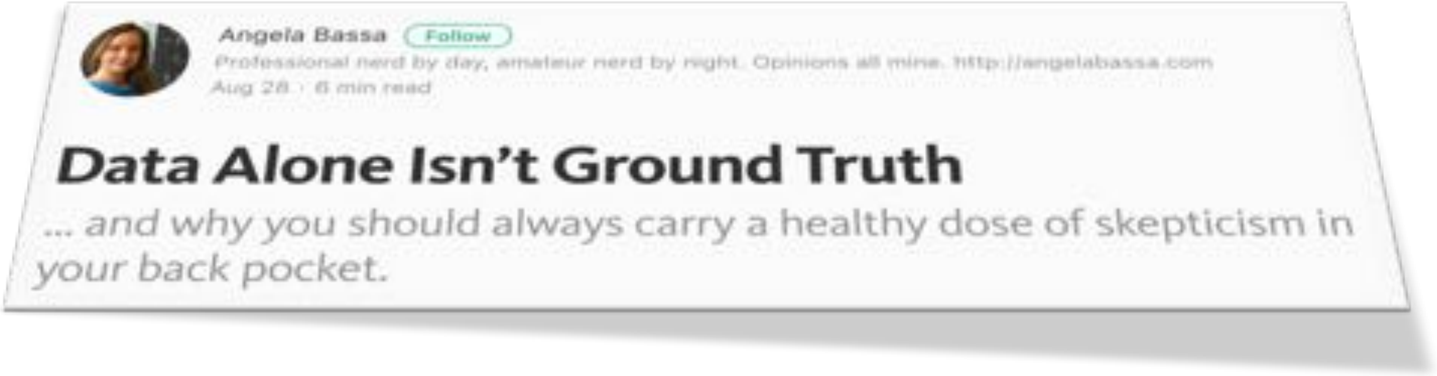
All Science is …

- … physics or stamp collecting
- … ~~physics or stamp collecting~~ **provenance**!

- **Provenance as …**
  - ***evidence*** (in science … computational science…)
  - ***explanation*** (in science … computational science ..)
- As we discuss, work with tools, ask …
  - … is this for provenance ***recording***?
  - … is this for (method, …) ***explaining***?
  - … ***by*** whom?
  - … ***for*** whom?

Why provenance?

- … to *document* what happened
- ... to *understand* what happened
- ... to *explain* what happened
- ... to *anticipate* what (might/will) happen?


Angela Bassa

Professional nerd by day, amateur nerd by night. Opinions all mine. http://angelabassa.com

Aug 28 · 6 min read

Data Alone Isn’t Ground Truth

... and why you should always carry a healthy dose of skepticism in your back pocket.

… never forget …

- *"The government are very keen on amassing statistics. They collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases."*

SOME ECONOMIC FACTORS
IN MODERN LIFE

BY

SIR JOSIAH STAMP, G.B.E.
Fellow of the British Academy

LONDON
P. S. KING & SON, LTD.
ORCHARD HOUSE, WESTMINSTER
1922

Computational Provenance defined …

- **Origin** and **processing history** of an **artifact**
  - usually: **data (products),** figures, ...
  - sometimes: **workflow** (and script) evolution …

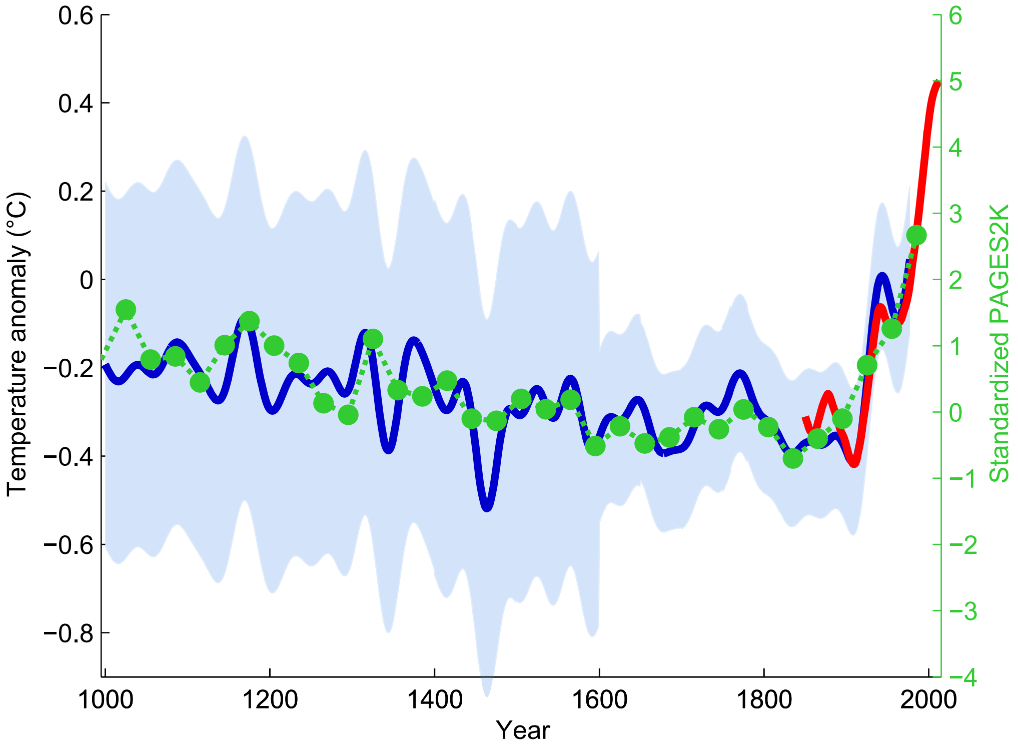
- Different sub-communities:
  - Provenance in **(scientific) workflows** …
  - Provenance in **databases** …
  - ... and of course there is more:
    - ... programming languages, systems/security, …
    - … information science, archival science, diplomatics
    - ... **science science**!

Using **Provenance** for Transparency, Reproducibility

- What ***input data*** went into this study?

- What ***methods*** were used?
- … with what ***parameter*** settings, ***calibrations***, …?

- *Can we* ***trust*** *the data and methods?*



- **Provenance** (*lineage*): track **origin** and **processing history** of data ➔ trust, data quality ~ audit trail for attribution, credit

- **Discovery** of data, methodologies, experiments

Climate Change: Whodunnit?

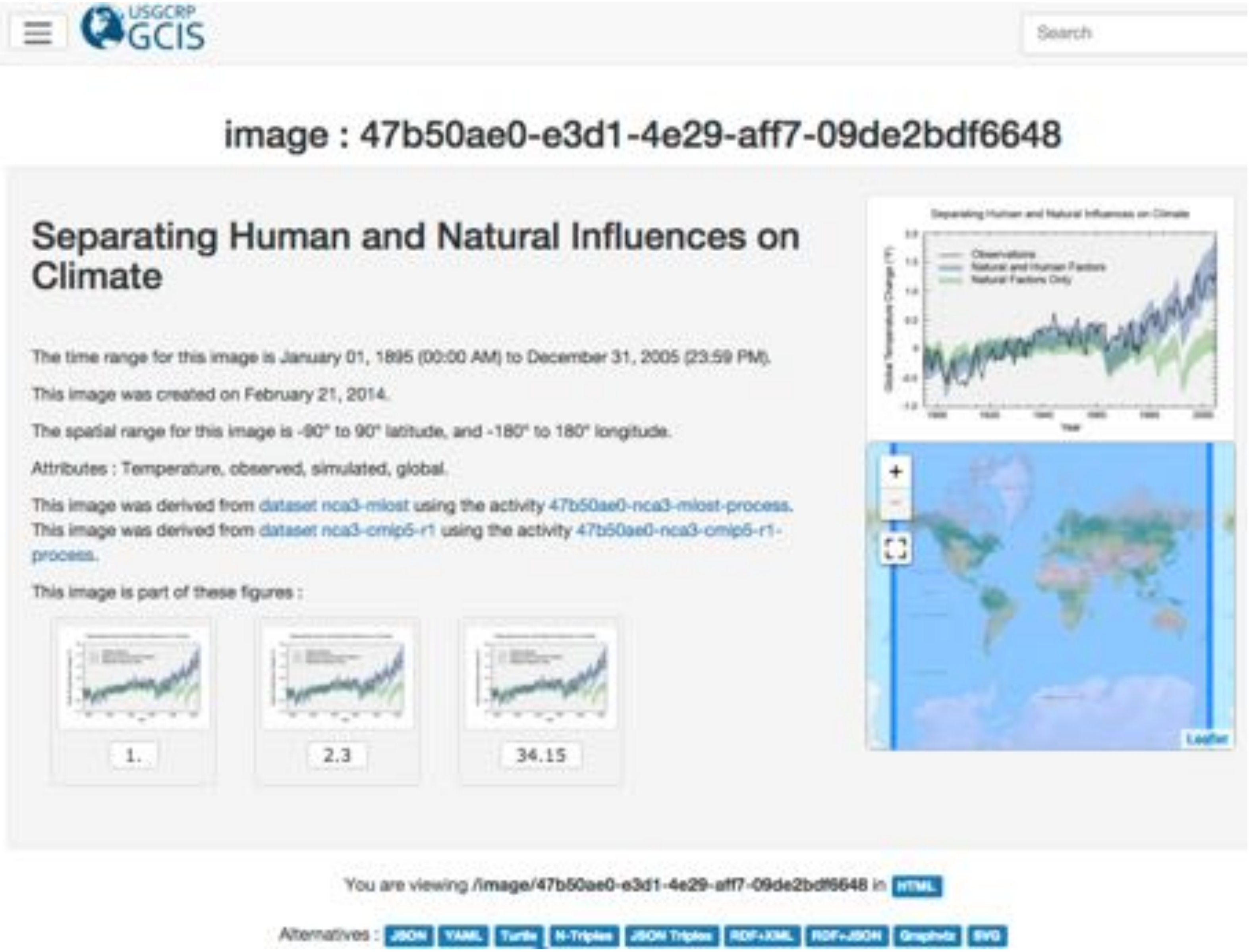


image : 47b50ae0-e3d1-4e29-aff7-09de2bdf6648

Separating Human and Natural Influences on Climate

The time range for this image is January 01, 1895 (00:00 AM) to December 31, 2005 (23:59 PM).

This image was created on February 21, 2014.

The spatial range for this image is -90° to 90° latitude, and -180° to 180° longitude.

Attributes : Temperature, observed, simulated, global.

This image was derived from dataset nca3-mlost using the activity 47b50ae0-nca3-mlost-process.
This image was derived from dataset nca3-cmip5-r1 using the activity 47b50ae0-nca3-cmip5-r1-process.

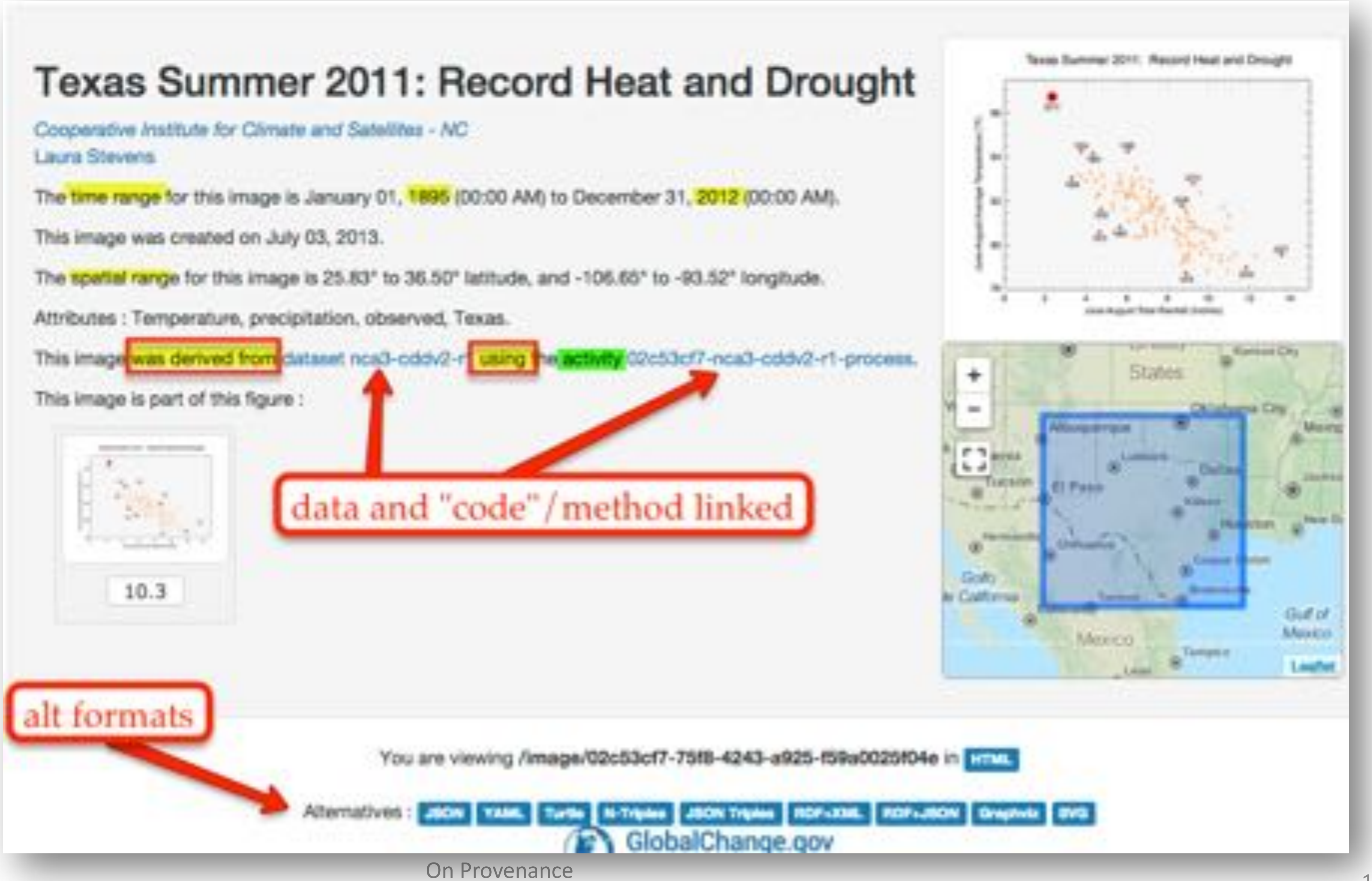
This image is part of these figures :

1.
2.3
34.15
Separating Human and Natural Influences on Climate
Observations
Natural and Human Factors
Natural Factors Only
Global Temperature Change (°F)
Year
+
−
Leaflet

You are viewing /image/47b50ae0-e3d1-4e29-aff7-09de2bdf6648 in HTML

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

Tracing the sources (data, code)



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, 1895 (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-r1 using the activity 02c53cf7-nca3-cddv2-r1-process.

This image is part of this figure :

10.3

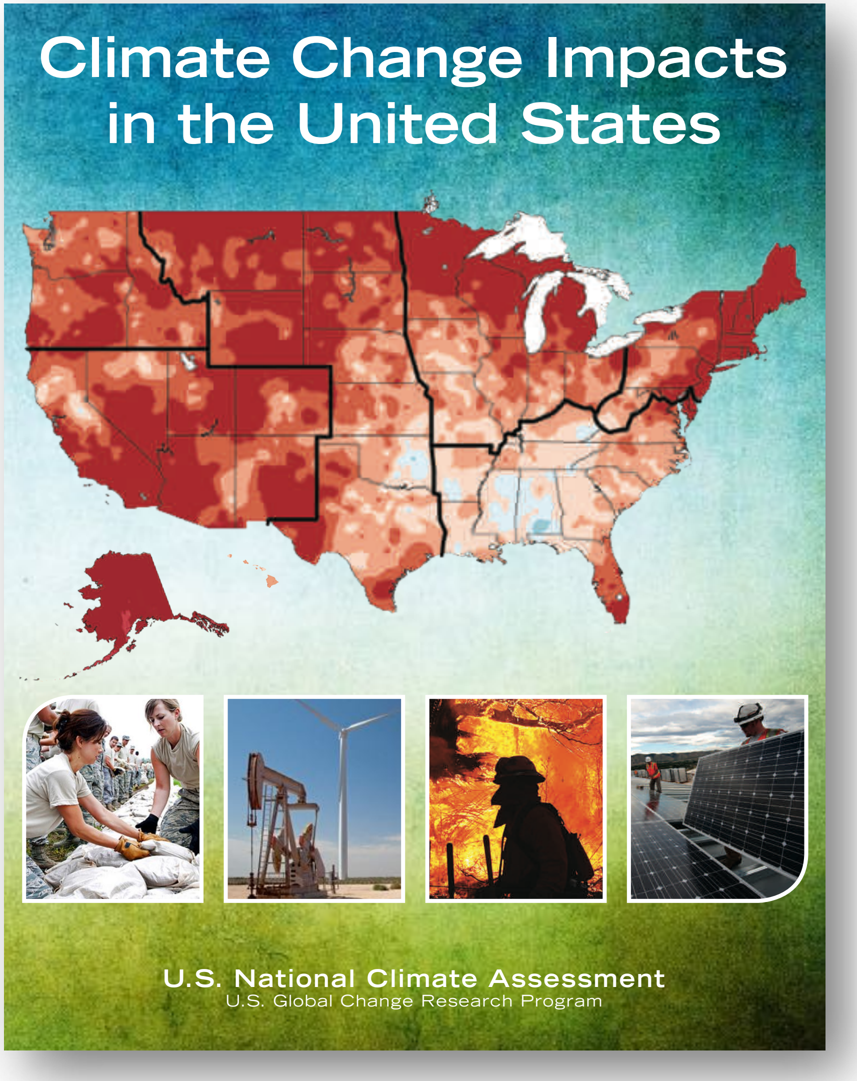
data and "code"/method linked

alt formats

You are viewing /image/02c53cf7-75f8-4243-a925-f59a0025f04e in HTML

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

Provenance today: Important but *hard*



Climate Change Impacts in the United States

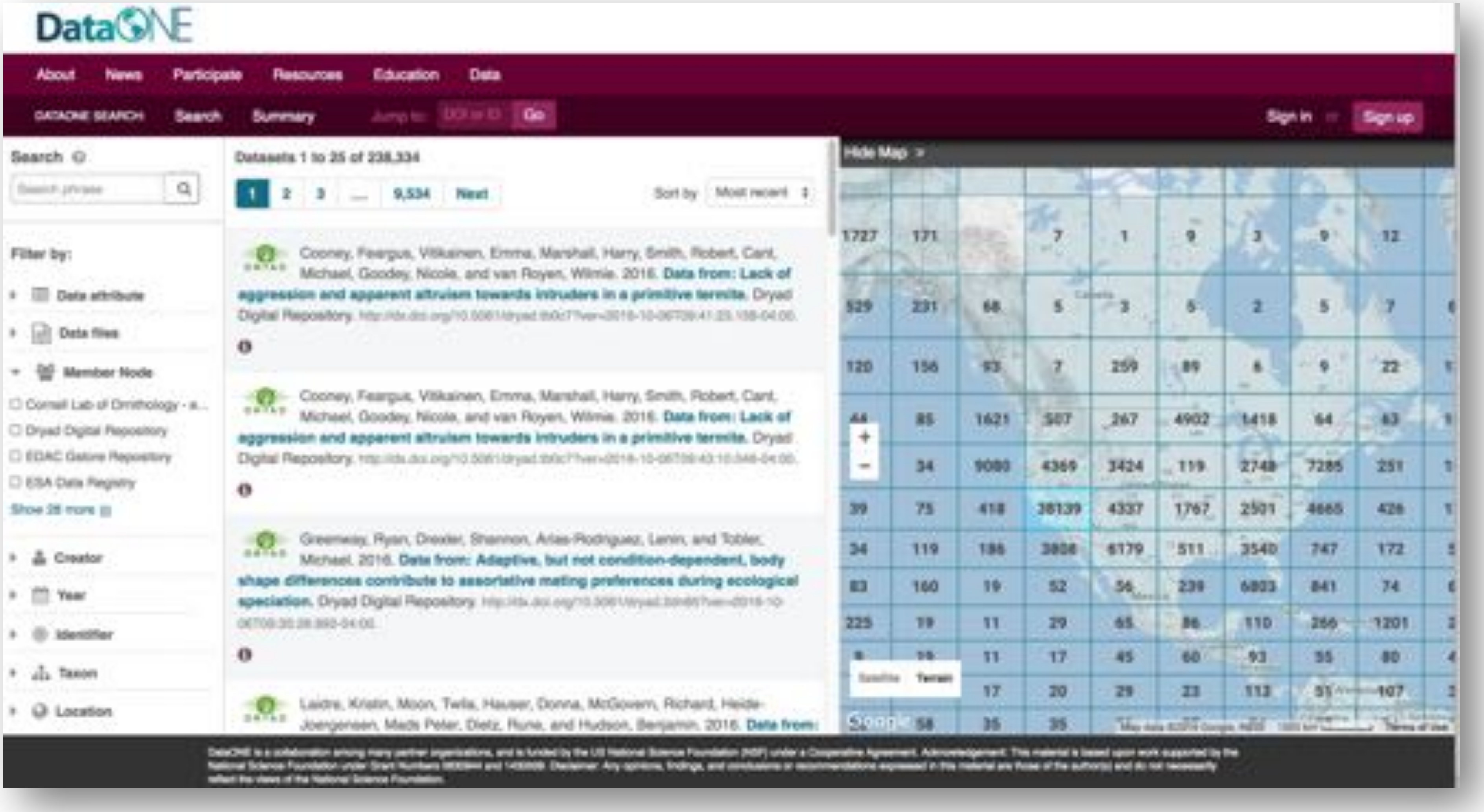
U.S. National Climate Assessment
U.S. Global Change Research Program

"This report is the result of a **three-year** analytical effort by a team of **over 300 experts**, overseen by a broadly constituted Federal Advisory Committee of **60 members**. It was developed from information and analyses gathered in over 70 workshops and listening sessions held across the country."

➔many research projects, groups conduct R&D on provenance methods, tools, …

Example: DataONE

A scientific data federation: DataONE
Data Observation Network for Earth



Datasets 1 to 25 of 238,334

1 2 3 ... 9,534 Next

□ Cornell Lab of Ornithology - a...
□ Dryad Digital Repository
□ EDAC Gstore Repository
□ ESA Data Registry
Show 28 more

Cooney, Feargus, Vitikainen, Emma, Marshall, Harry, Smith, Robert, Cant, Michael, Goodey, Nicola, and van Royen, Wilmie. 2016. Data from: Lack of aggression and apparent altruism towards intruders in a primitive termite. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.tb0c7?ver=2016-10-06T09:41:23.158-04:00.

Cooney, Feargus, Vitikainen, Emma, Marshall, Harry, Smith, Robert, Cant, Michael, Goodey, Nicola, and van Royen, Wilmie. 2016. Data from: Lack of aggression and apparent altruism towards intruders in a primitive termite. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.tb0c7?ver=2016-10-06T09:43:10.346-04:00.

Greenway, Ryan, Drexler, Shannon, Arias-Rodriguez, Lenin, and Tobler, Michael. 2016. Data from: Adaptive, but not condition-dependent, body shape differences contribute to assortative mating preferences during ecological speciation. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.3dn8t?ver=2016-10-06T09:30:28.992-04:00.

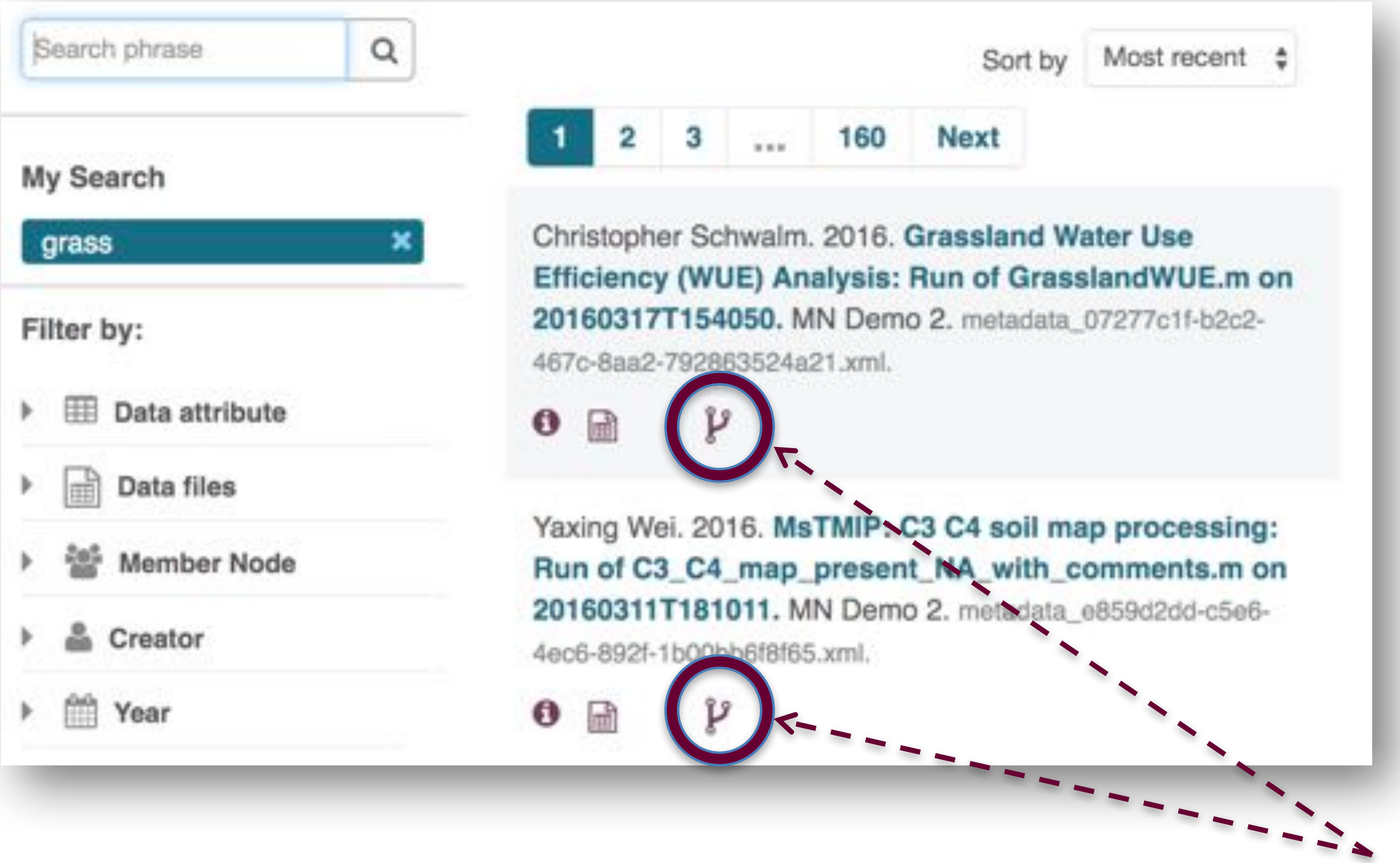
Laidre, Kristin, Moon, Twila, Hauser, Donna, McGovern, Richard, Heide-Joergensen, Mads Peter, Dietz, Rune, and Hudson, Benjamin. 2016. Data from:

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

DataONE Search

search.dataone.org

Provenance in Action: Benefits & Impact

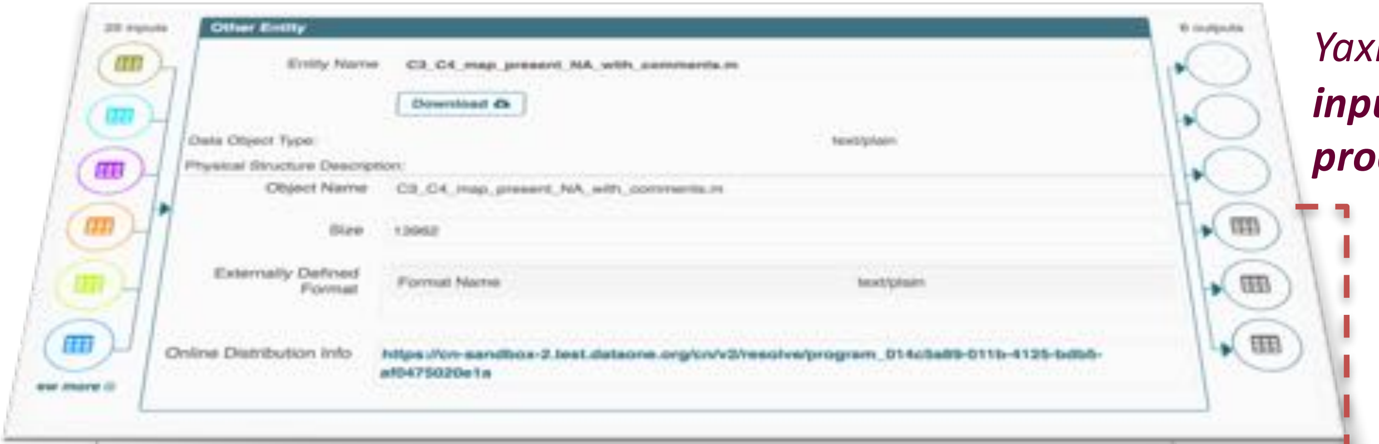


Christopher Schwalm. 2016. **Grassland Water Use Efficiency (WUE) Analysis: Run of GrasslandWUE.m on 20160317T154050.** MN Demo 2. metadata_07277c1f-b2c2-467c-8aa2-792863524a21.xml.

Yaxing Wei. 2016. **MsTMIP: C3 C4 soil map processing: Run of C3_C4_map_present_NA_with_comments.m on 20160311T181011.** MN Demo 2. metadata_e859d2dd-c5e6-4ec6-892f-1b00bb6f8f65.xml.

A DataONE search (here: "grass") yields different packages with provenance

DataONE: Support for Provenance



Other Entity

Entity Name: C3_C4_map_present_NA_with_comments.m

Download

Data Object Type: text/plain

Physical Structure Description:

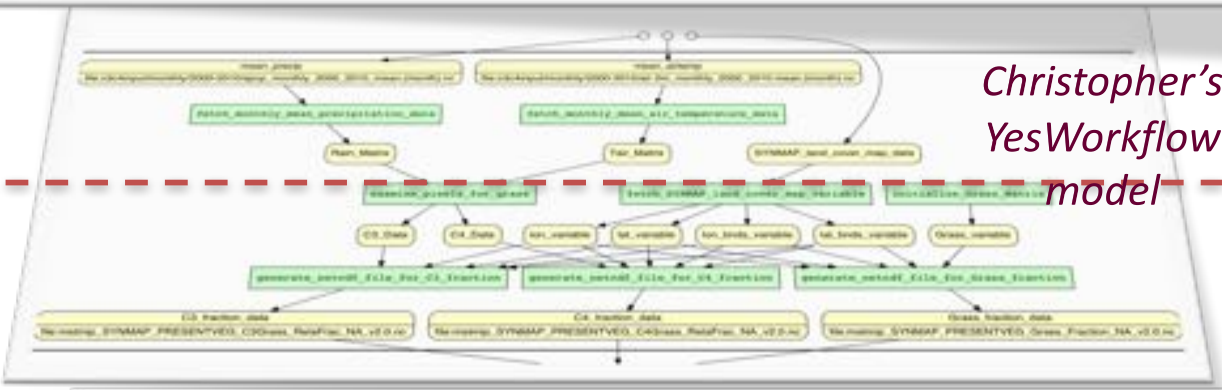
Object Name: C3_C4_map_present_NA_with_comments.m

Size: 13962

Externally Defined Format: Format Name text/plain

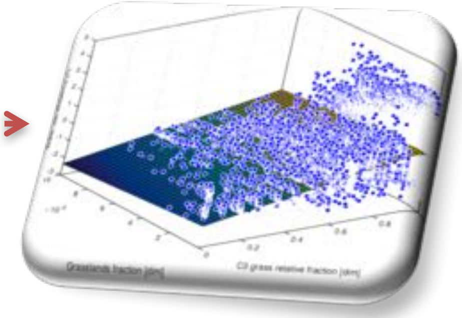
Online Distribution Info: https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program_014c5a89-011b-4125-bdb5-af0475020e1a

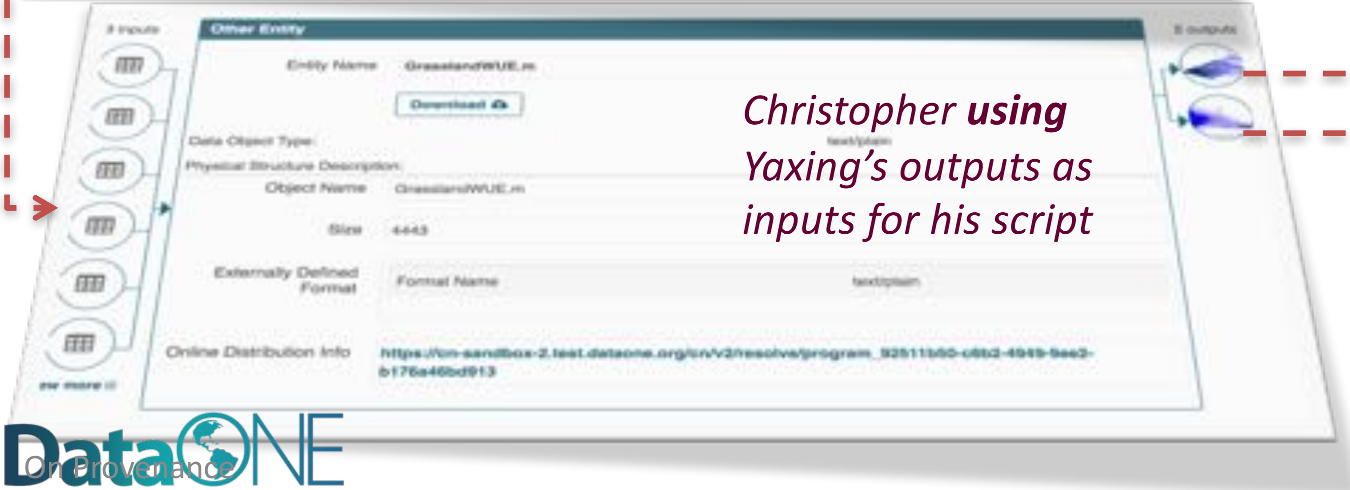
*Yaxing's **script** with*
inputs** & **output
products


mean_precip
mean_airtemp
fetch_monthly_mean_precipitation_data
fetch_monthly_mean_air_temperature_data
Rain_Matrix
Tair_Matrix
SYNMAP_land_cover_map_data
examine_pixels_for_grass
fetch_SYNMAP_land_cover_map_variable
initialize_Grass_Matrix
C3_Data
C4_Data
lon_variable
lat_variable
lon_bnds_variable
lat_bnds_variable
Grass_variable
generate_netcdf_file_for_C3_fraction
generate_netcdf_file_for_C4_fraction
generate_netcdf_file_for_Grass_fraction
C3_fraction_data
C4_fraction_data
Grass_fraction_data

Christopher's YesWorkflow model

Christopher's results can be traced back all the way to Yaxing's input





Other Entity

Entity Name: GrasslandWUE.m

Download

Data Object Type: text/plain

Physical Structure Description:

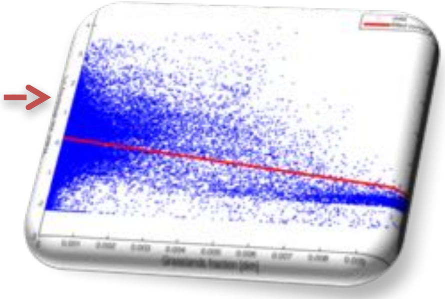
Object Name: GrasslandWUE.m

Size: 4443

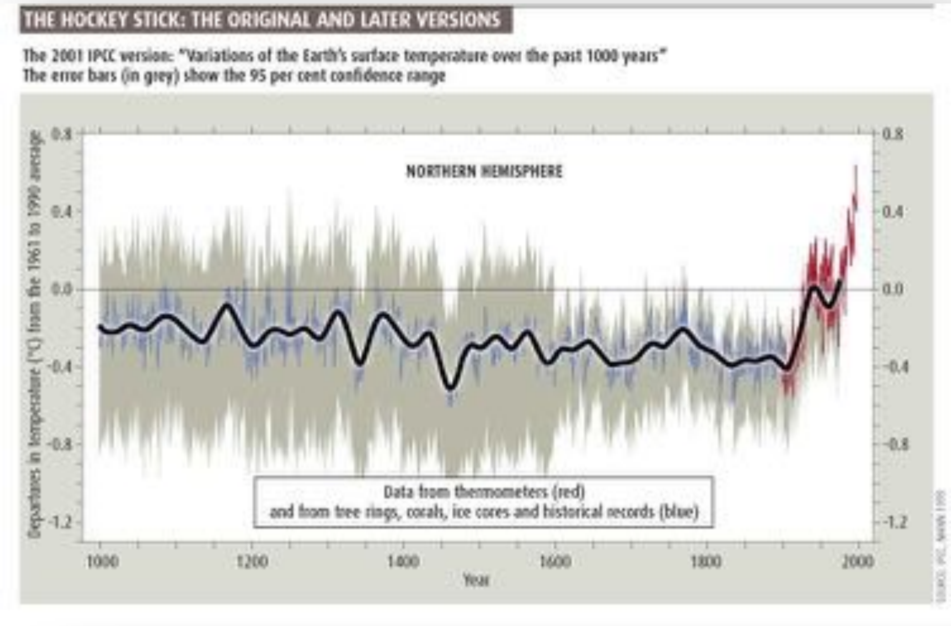
Externally Defined Format: Format Name text/plain

Online Distribution Info: https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program_92511b50-c6b2-4949-9ee3-b176a46bd913

*Christopher **using** Yaxing's outputs as inputs for his script*

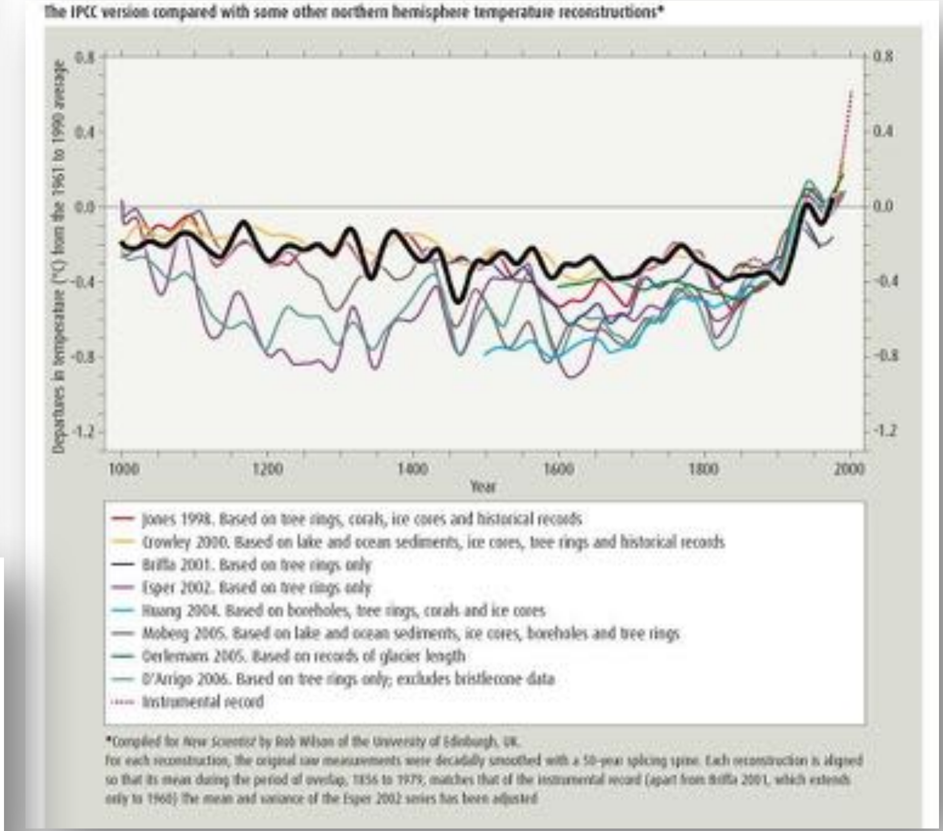


REWIND: From Provenance to Reproducible Science …



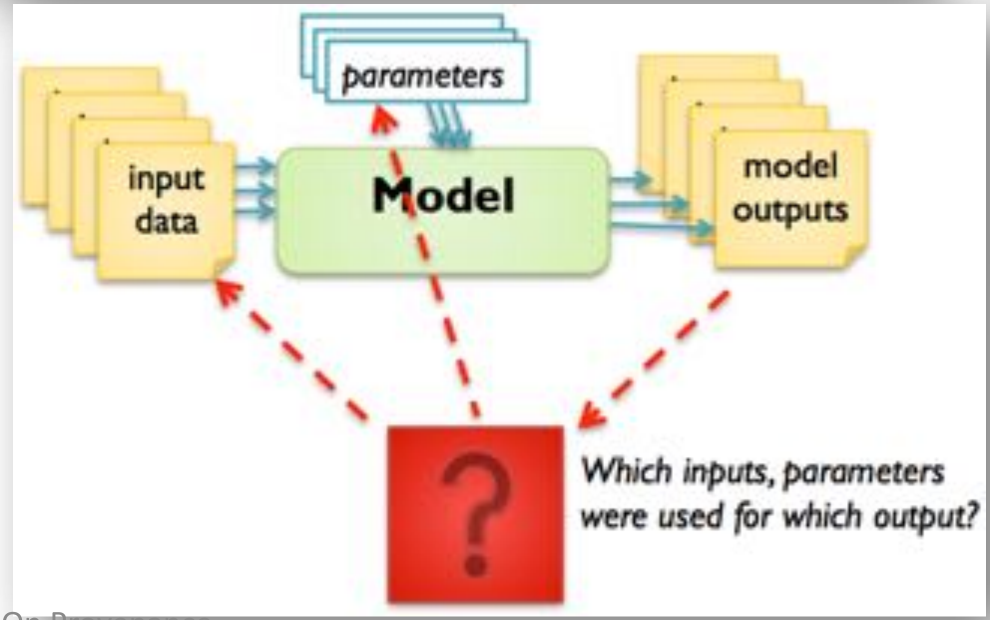
THE HOCKEY STICK: THE ORIGINAL AND LATER VERSIONS

The 2001 IPCC version: "Variations of the Earth's surface temperature over the past 1000 years"
The error bars (in grey) show the 95 per cent confidence range



The IPCC version compared with some other northern hemisphere temperature reconstructions*

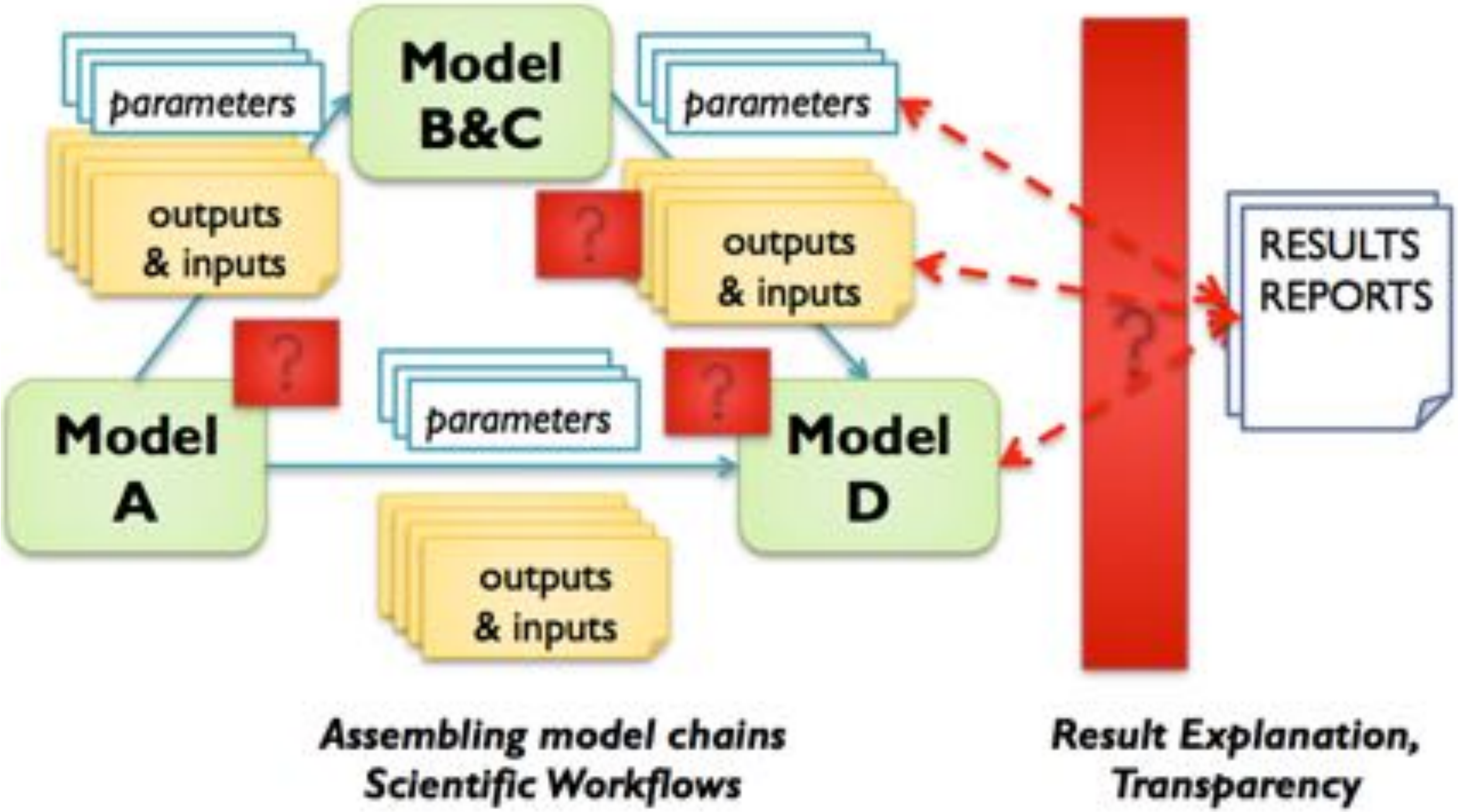
*Compiled for *New Scientist* by Rob Wilson of the University of Edinburgh, UK.
For each reconstruction, the original raw measurements were decadally smoothed with a 50-year splicing spine. Each reconstruction is aligned so that its mean during the period of overlap, 1856 to 1979, matches that of the instrumental record (apart from Briffa 2001, which extends only to 1960) The mean and variance of the Esper 2002 series has been adjusted



Which inputs, parameters were used for which output?

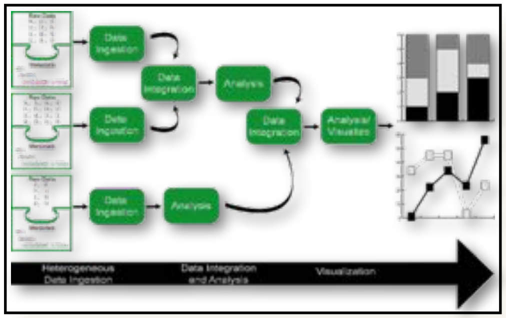
Capturing **provenance** is crucial for transparency, interpretation, debugging, …
=> *repeatable experiments,*
=> *reproducible science*
=> *need workflow-system agnostic model*

... via scientific workflows (… and scripts)



Tour Stop: **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**

Kepler

swift

- **Abstraction, Evolution, Reuse** *(**human** cycles)*
  - wfs should be easy to **(re-)use, evolve, share**

Taverna

ASKALON

VisTrails

TRIANA

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

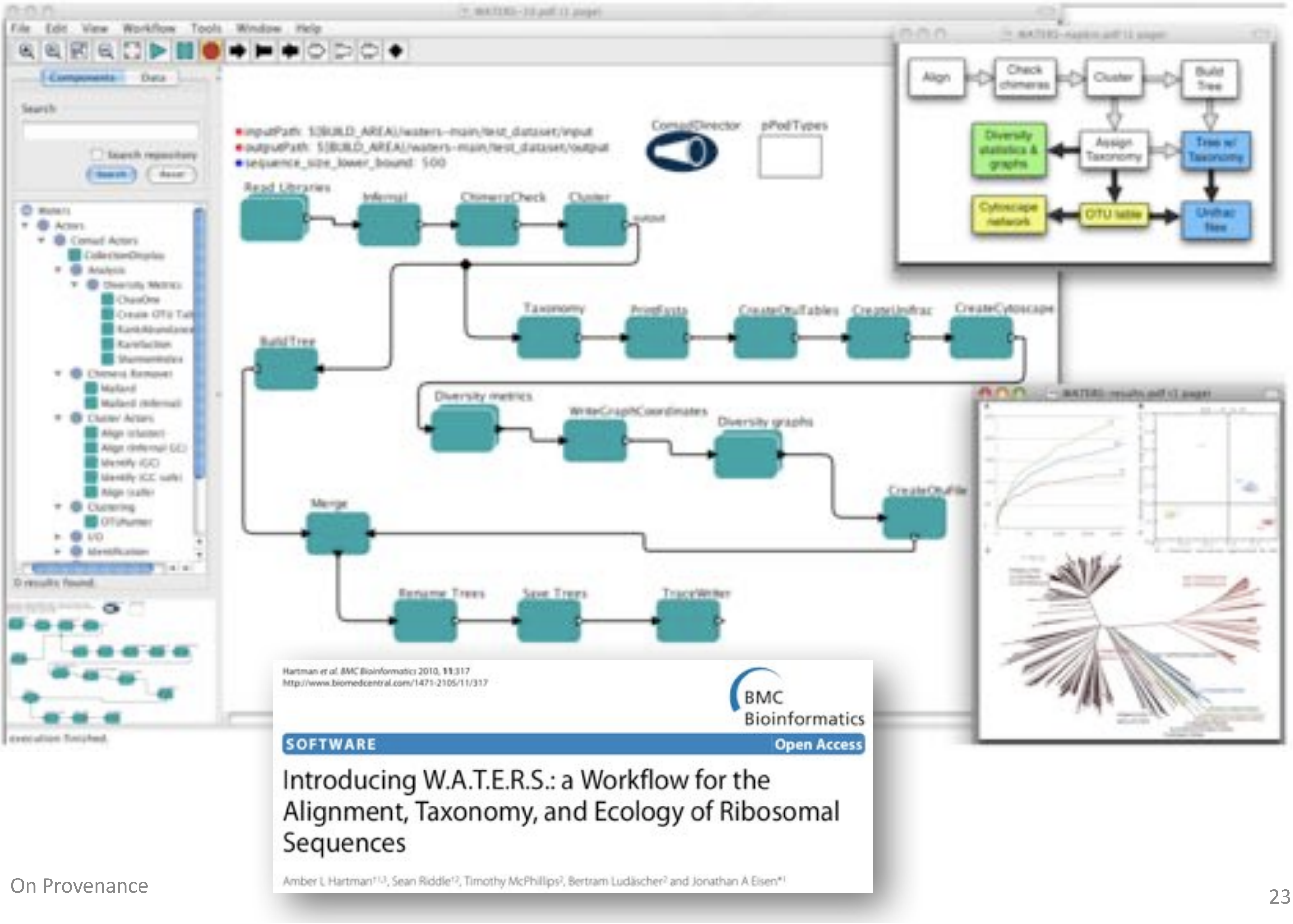
WINGS

pegasus

Trident Workbench

Es war einmal …

Executable WATERS Workflow in Kepler



- inputPath: $BUILD_AREA)/waters-main/test_dataset/input
- outputPath: $BUILD_AREA)/waters-main/test_dataset/output
- sequence_size_lower_bound: 500

Hartman *et al. BMC Bioinformatics* 2010, **11**:317
http://www.biomedcentral.com/1471-2105/11/317

BMC Bioinformatics

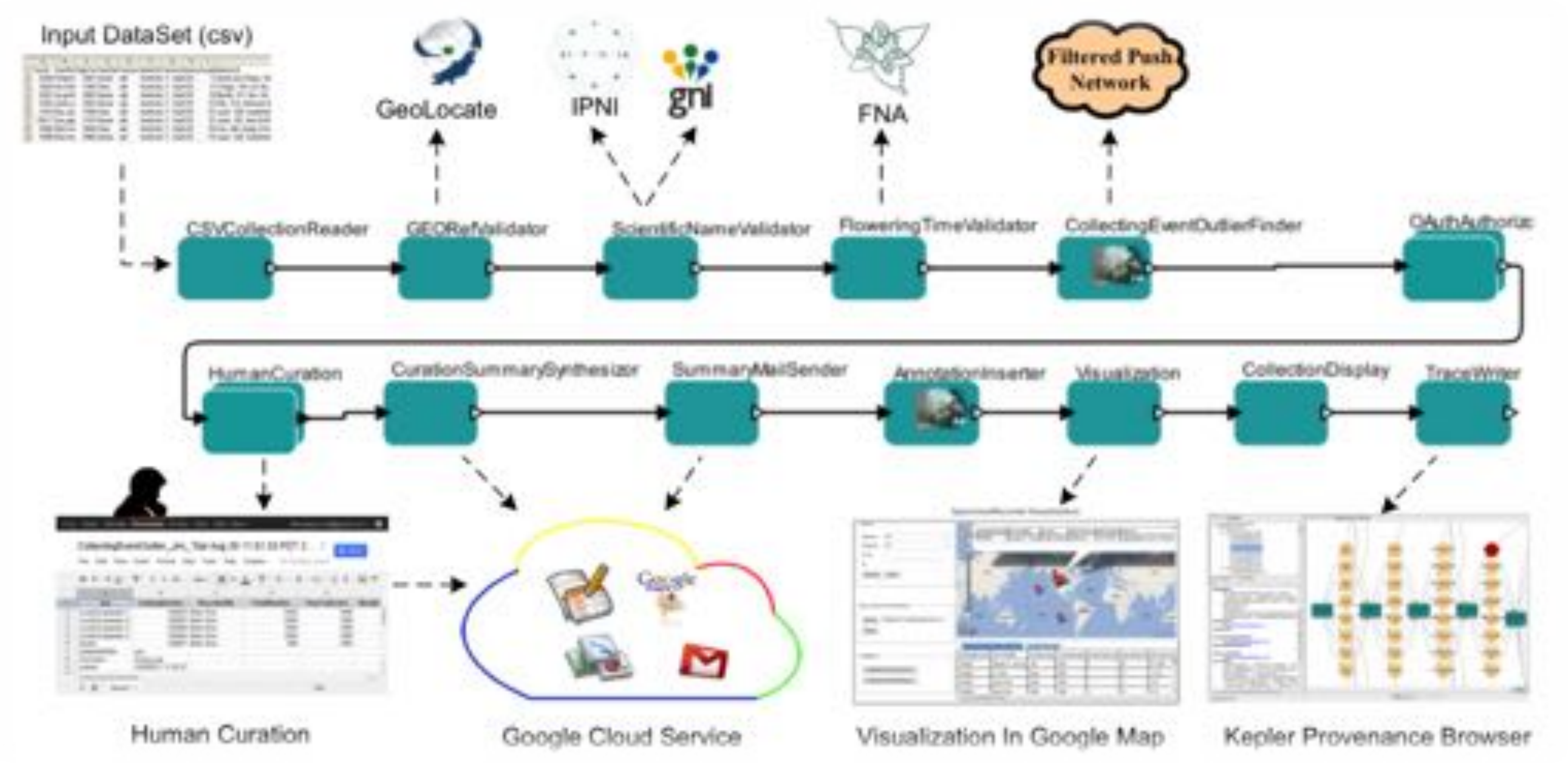
SOFTWARE Open Access

Introducing W.A.T.E.R.S.: a Workflow for the Alignment, Taxonomy, and Ecology of Ribosomal Sequences

Amber L Hartman[†1,3], Sean Riddle[†1,2], Timothy McPhillips[2], Bertram Ludäscher[2] and Jonathan A Eisen[*1]

Data Curation Workflows

(Filtered-Push … Kepler … Kurator projects)



HARVARD UNIVERSITY HERBARIA

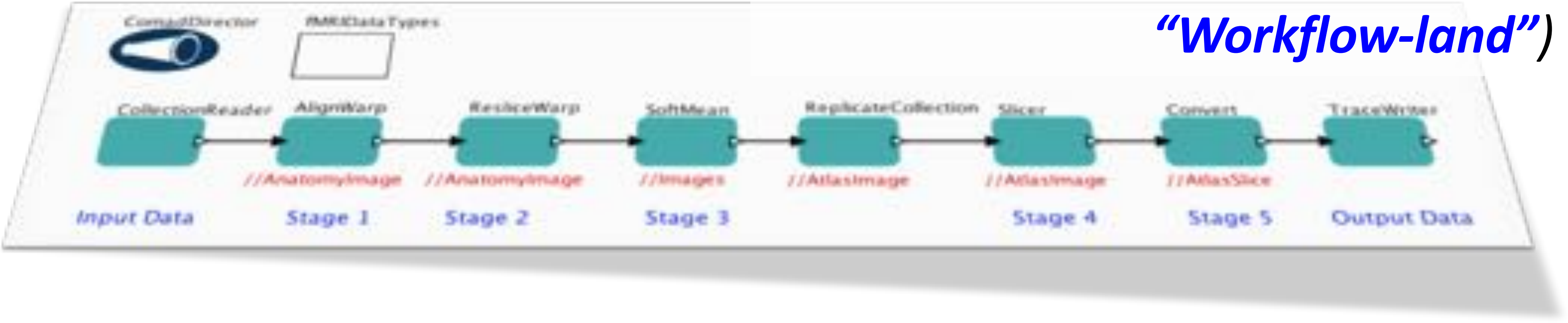
MUSEUM OF COMPARATIVE ZOOLOGY
HARVARD UNIVERSITY

UCDAVIS genome center

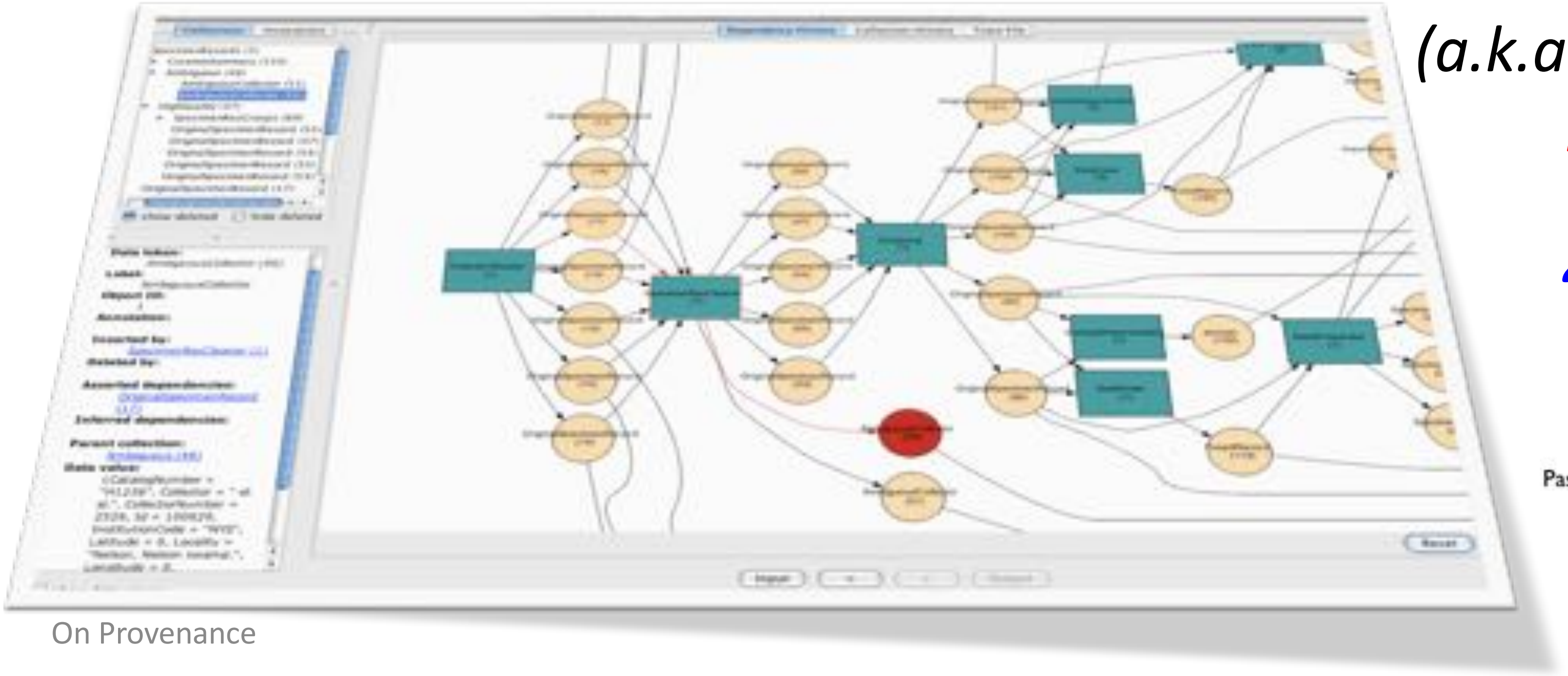
I 1867 ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Provenance ⇔ Workflows

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

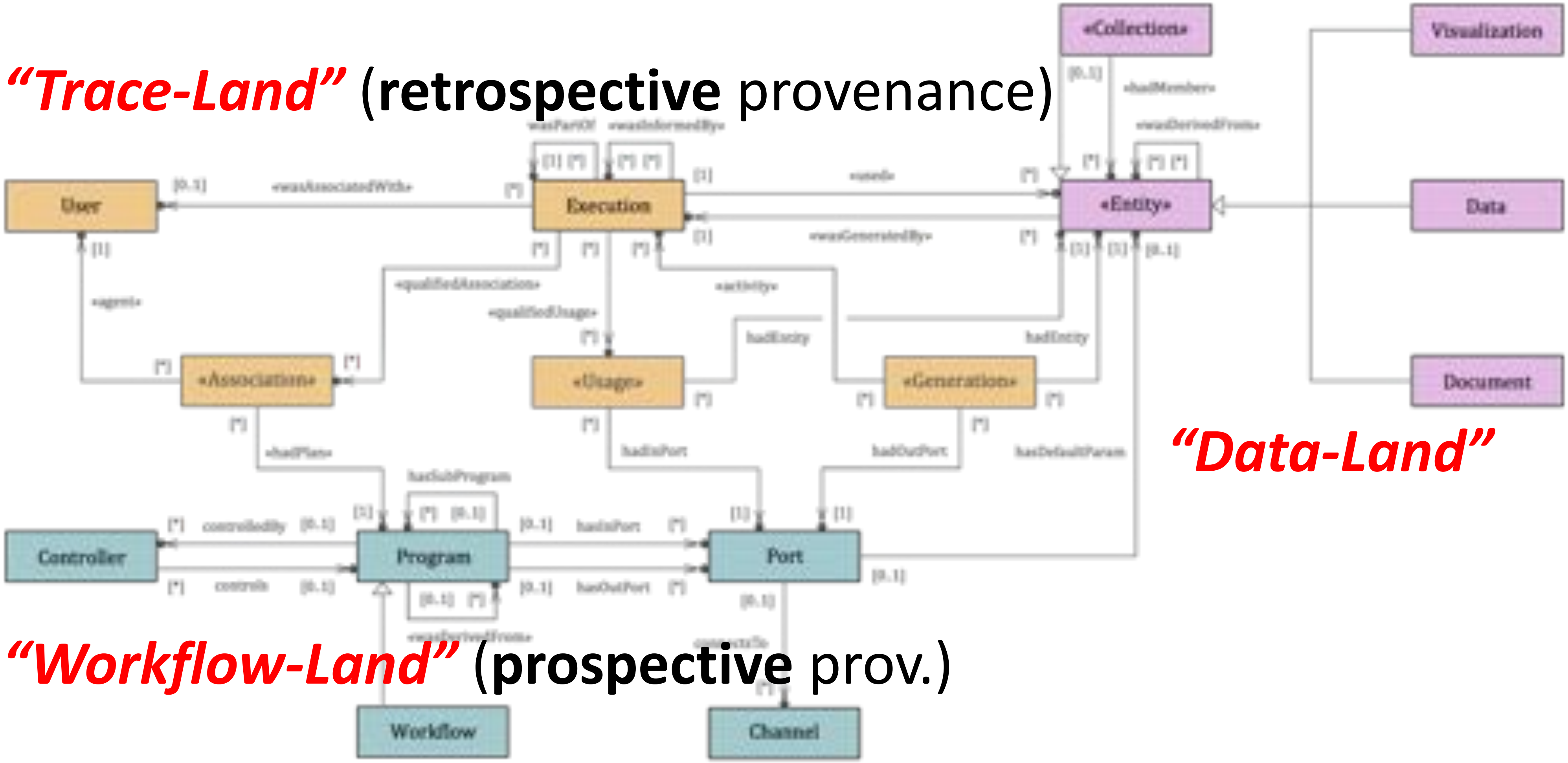


Runtime Provenance
(a.k.a. ***traces, logs,***
retrospective
provenance,
"Trace-land"*)*



ProvONE: PROV for scientific workflows

(Transfer station to any of several other "standard extensions")



"Trace-Land" (**retrospective** provenance)

"Data-Land"

"Workflow-Land" (**prospective** prov.)

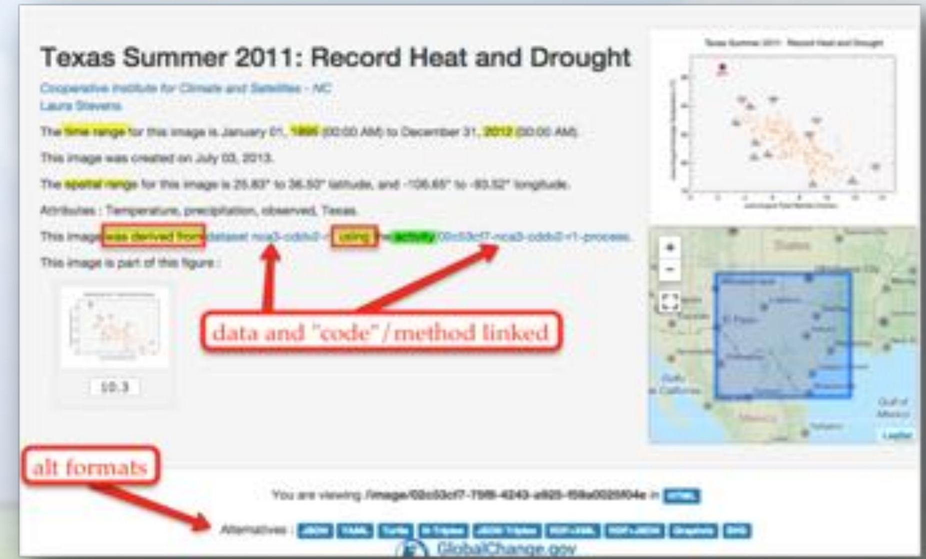
Also: OPM-W (G & G et al), others

…

But … **how to prime the provenance pump??**

Must support "**Provenance for Self**" !

Provenance for Self?!



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, 1895 (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 :

10.3

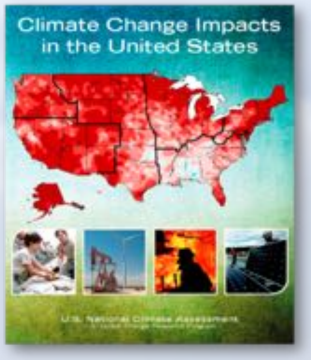
data and "code"/method linked

alt formats

You are viewing /image/02c53cf7-75f8-4243-a925-f59a0025f04e in HTML

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

Texas Summer 2011 - Record Heat and Drought
Dallas
El Paso
Gulf of California
Mexico
Gulf of Mexico
Leaflet
GlobalChange.gov



Climate Change Impacts in the United States
U.S. National Climate Assessment

Provenance for Others

- ✔ 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.**

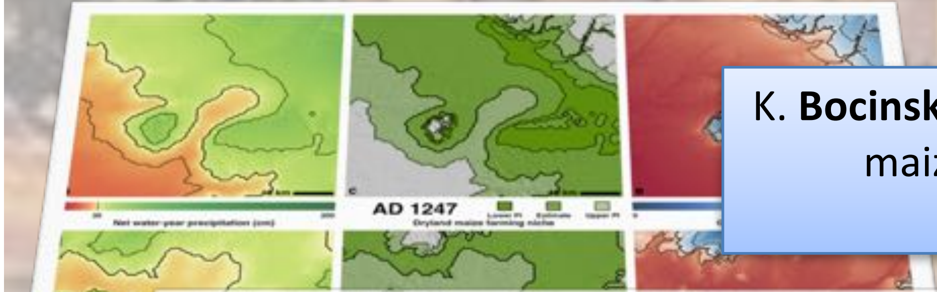
From Workflows & Provenance to **Provenance for Script-based Workflows …**

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

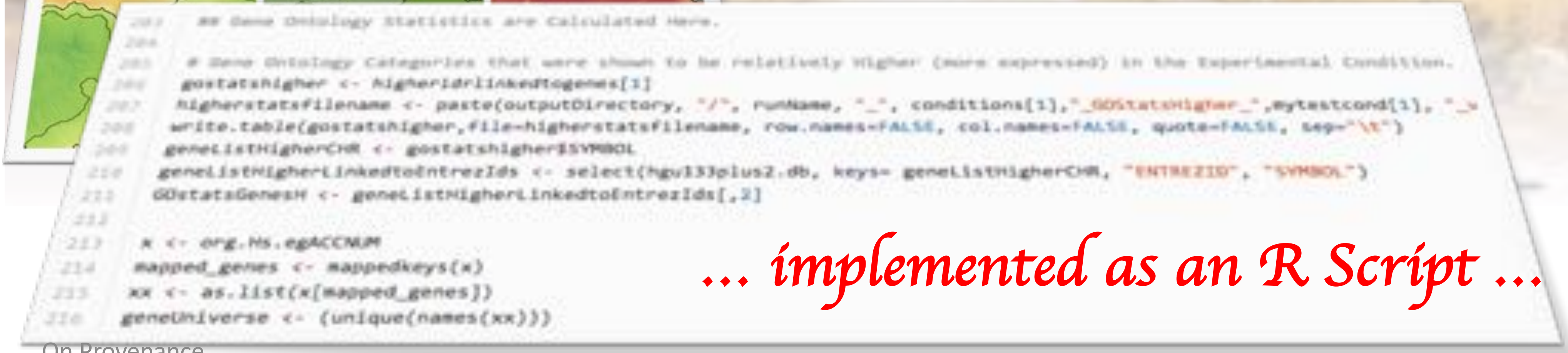
SKOPE: Synthesized Knowledge Of Past Environments

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.


AD 1247
Net water-year precipitation (cm)
Dryland maize farming niche

K. **Bocinsky**, T. **Kohler**, A 2000-year reconstruction of the rain-fed maize agricultural niche in the US Southwest. ***Nature Communications***. doi:10.1038/ncomms6618



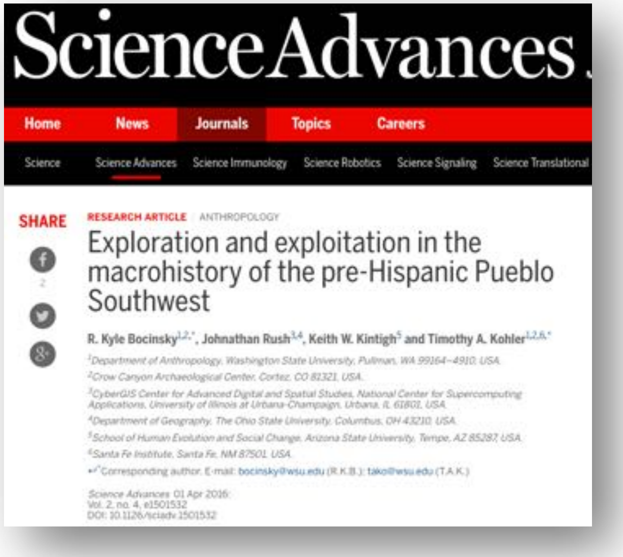
```
 ## Gene Ontology Statistics are Calculated Here.

  # Gene Ontology Categories that were shown to be relatively Higher (more expressed) in the Experimental Condition.
 gostatshigher <- higheridrlinkedtogenes[1]
 higherstatsfilename <- paste(outputDirectory, "/", runName, "_", conditions[1],"_GOStatsHigher_",mytestcond[1], "_v
 write.table(gostatshigher,file=higherstatsfilename, row.names=FALSE, col.names=FALSE, quote=FALSE, sep="\t")
 geneListHigherCHR <- gostatshigher$SYMBOL
 geneListHigherLinkedtoEntrezIds <- select(hgu133plus2.db, keys= geneListHigherCHR, "ENTREZID", "SYMBOL")
 GOstatsGenesH <- geneListHigherLinkedtoEntrezIds[,2]

 x <- org.Hs.egACCNUM
 mapped_genes <- mappedkeys(x)
 xx <- as.list(x[mapped_genes])
 geneUniverse <- (unique(names(xx)))
```

... implemented as an R Script ...

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?



RESEARCH ARTICLE | ANTHROPOLOGY

Exploration and exploitation in the macrohistory of the pre-Hispanic Pueblo Southwest

R. Kyle Bocinsky[1,2,*], Johnathan Rush[3,4], Keith W. Kintigh[5] and Timothy A. Kohler[1,2,6,*]

[1] Department of Anthropology, Washington State University, Pullman, WA 99164–4910, USA.

[2] Crow Canyon Archaeological Center, Cortez, CO 81321, USA.

[3] CyberGIS Center for Advanced Digital and Spatial Studies, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

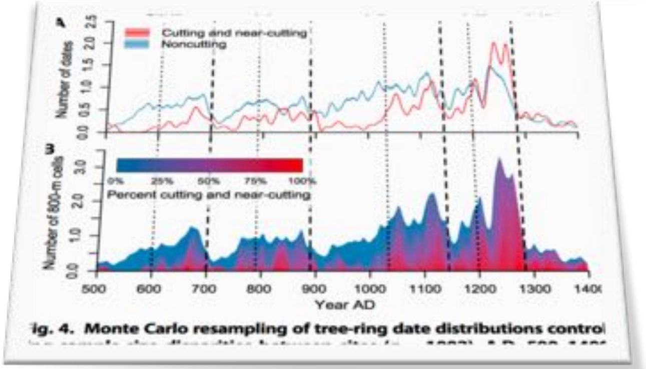
[4] Department of Geography, The Ohio State University, Columbus, OH 43210, USA.

[5] School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287, USA.

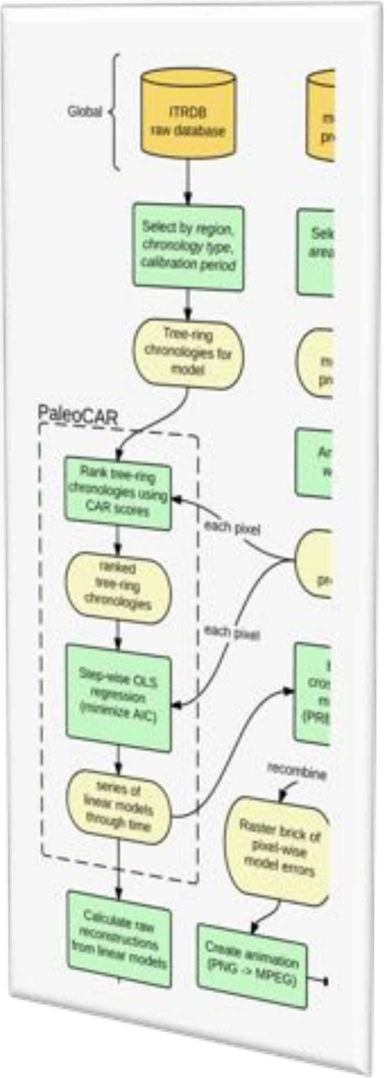
[6] Santa Fe Institute, Santa Fe, NM 87501, USA.

*Corresponding author. E-mail: bocinsky@wsu.edu (R.K.B.); tako@wsu.edu (T.A.K.)

Science Advances 01 Apr 2016:
Vol. 2, no. 4, e1501532
DOI: 10.1126/sciadv.1501532

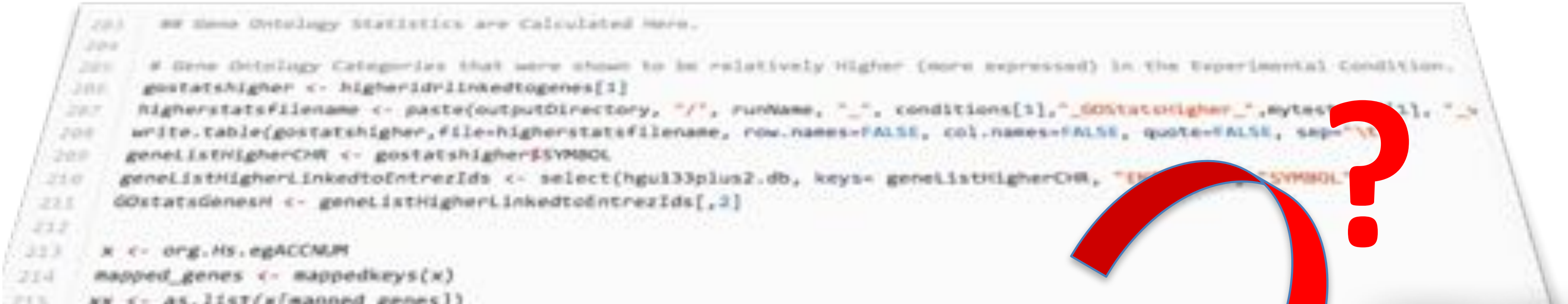


ig. 4. Monte Carlo resampling of tree-ring date distributions contro


Global
ITRDB raw database
Select by region, chronology type, calibration period
Tree-ring chronologies for model
PaleoCAR
Rank tree-ring chronologies using CAR scores
each pixel
ranked tree-ring chronologies
each pixel
Step-wise OLS regression (minimize AIC)
series of linear models through time
recombine
Raster brick of pixel-wise model errors
Calculate raw reconstructions from linear models
Create animation (PNG -> MPEG)

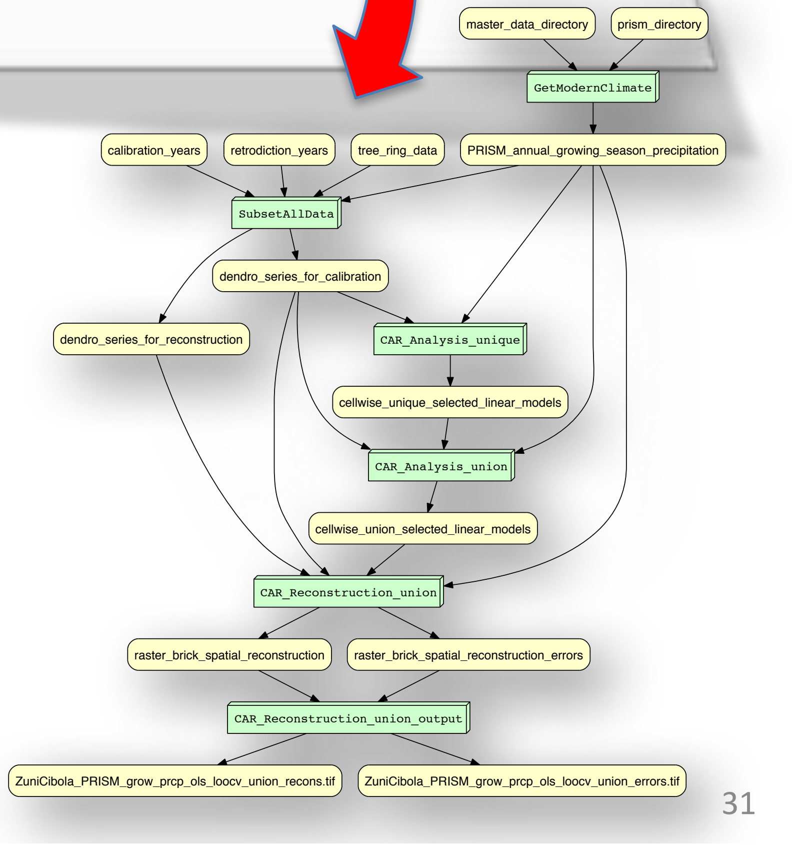
YesWorkflow:

Yes, scripts are workflows, too!



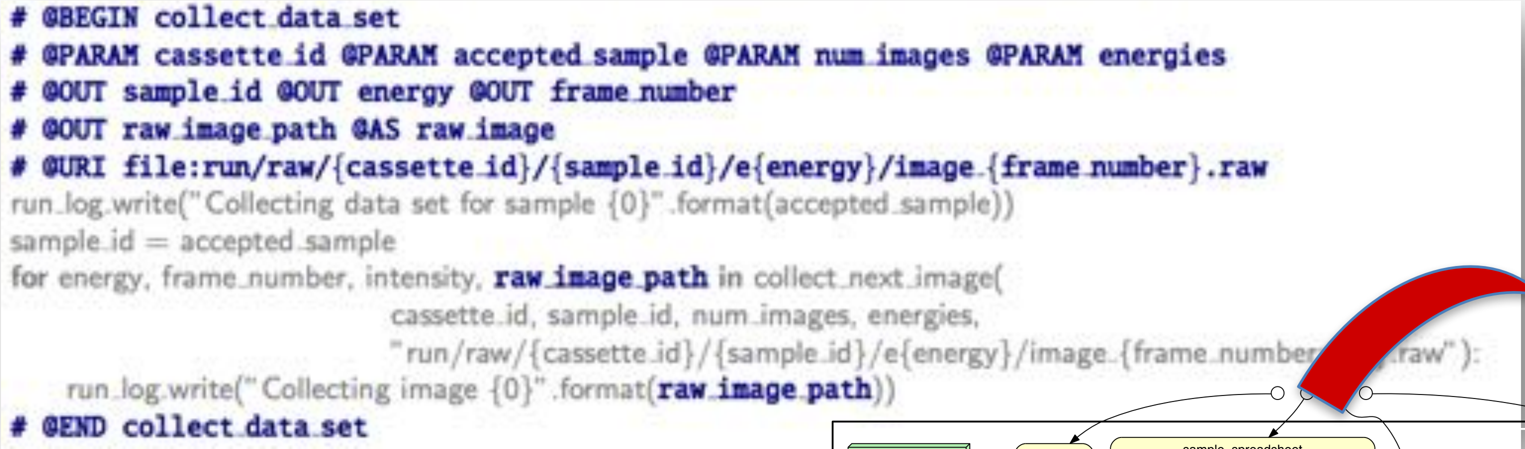
```
 ## Gene Ontology Statistics are Calculated Here.

  # Gene Ontology Categories that were shown to be relatively Higher (more expressed) in the Experimental Condition.
  gostatshigher <- higheridrlinkedtogenes[1]
  higherstatsFilename <- paste(outputDirectory, "/", runName, "_", conditions[1],"_GOStatsHigher_",mytest    1], "_>
  write.table(gostatshigher,file=higherstatsFilename, row.names=FALSE, col.names=FALSE, quote=FALSE, sep="\t
  geneListHigherCHR <- gostatshigher$SYMBOL
  geneListHigherLinkedtoEntrezIds <- select(hgu133plus2.db, keys= geneListHigherCHR, "EN          "SYMBOL"
  GOstatsGenesH <- geneListHigherLinkedtoEntrezIds[,2]

  x <- org.Hs.egACCNUM
  mapped_genes <- mappedkeys(x)
  xx <- as.list(x[mapped_genes])
  geneUniverse <- (unique(names(xx)))
```

- **Script** *vs* Workflows/**ASAP**:
  - **A**utomation: *****
  - **S**caling: **
  - **A**bstraction: *
  - **P**rovenance: **



master_data_directory
prism_directory
GetModernClimate
calibration_years
retrodiction_years
tree_ring_data
PRISM_annual_growing_season_precipitation
SubsetAllData
dendro_series_for_calibration
dendro_series_for_reconstruction
CAR_Analysis_unique
cellwise_unique_selected_linear_models
CAR_Analysis_union
cellwise_union_selected_linear_models
CAR_Reconstruction_union
raster_brick_spatial_reconstruction
raster_brick_spatial_reconstruction_errors
CAR_Reconstruction_union_output
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_recons.tif
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_errors.tif

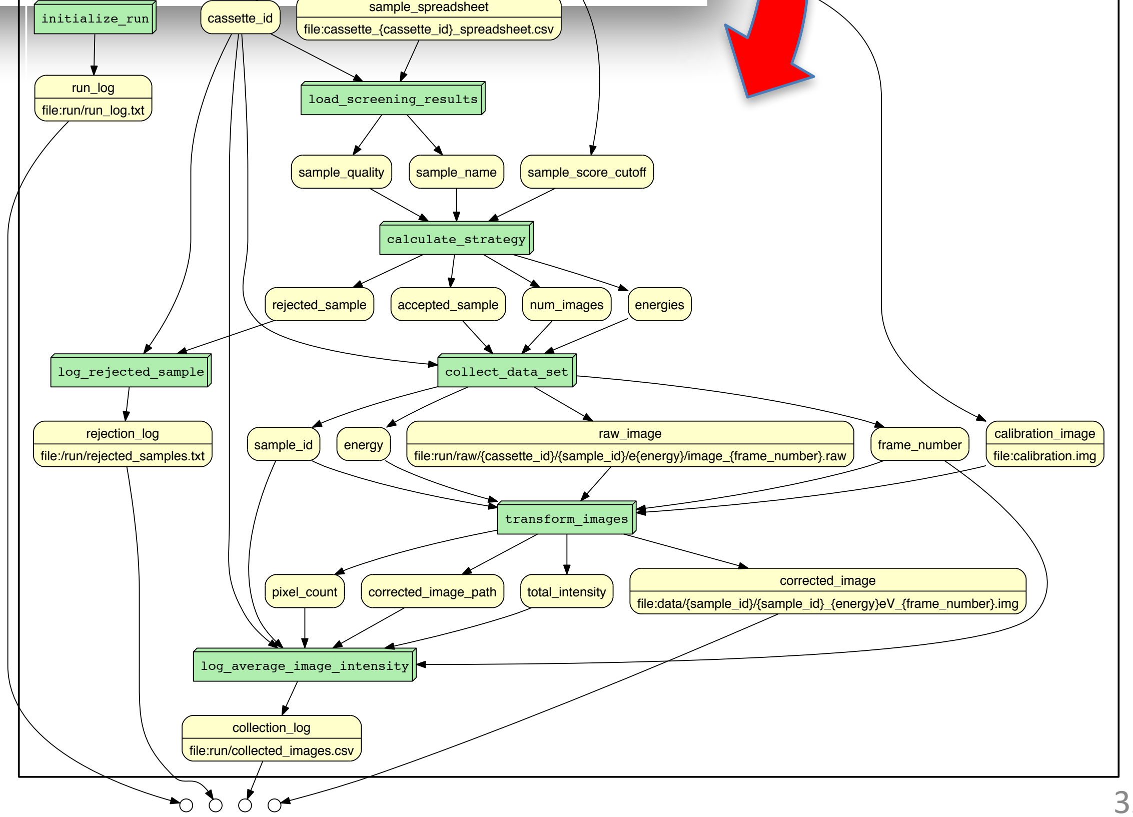
YesWorkflow: Prospective & Retrospective Provenance … (almost) for free!



```
# @BEGIN collect_data_set
# @PARAM cassette_id @PARAM accepted_sample @PARAM num_images @PARAM energies
# @OUT sample_id @OUT energy @OUT frame_number
# @OUT raw_image_path @AS raw_image
# @URI file:run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number}.raw
run_log.write("Collecting data set for sample {0}".format(accepted_sample))
sample_id = accepted_sample
for energy, frame_number, intensity, raw_image_path in collect_next_image(
                        cassette_id, sample_id, num_images, energies,
                        "run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number}.raw"):
    run_log.write("Collecting image {0}".format(raw_image_path))
# @END collect_data_set
```

YW!

- YW annotations in the script (R, Python, Matlab) are used to **recreate the workflow view** from the script …

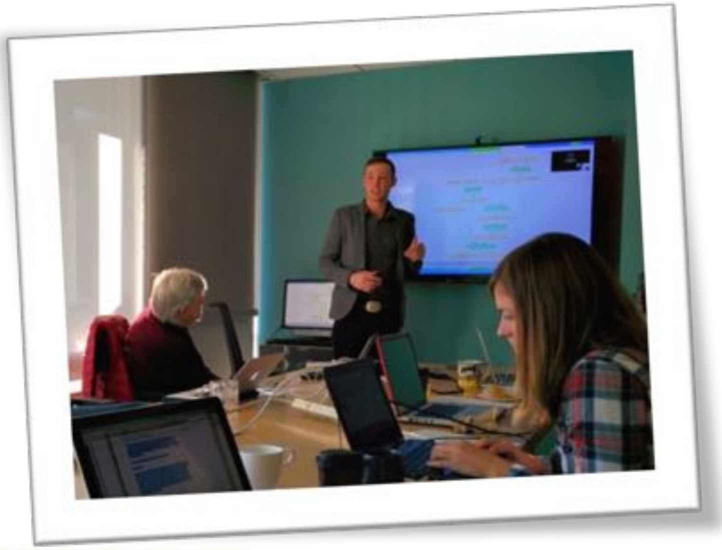


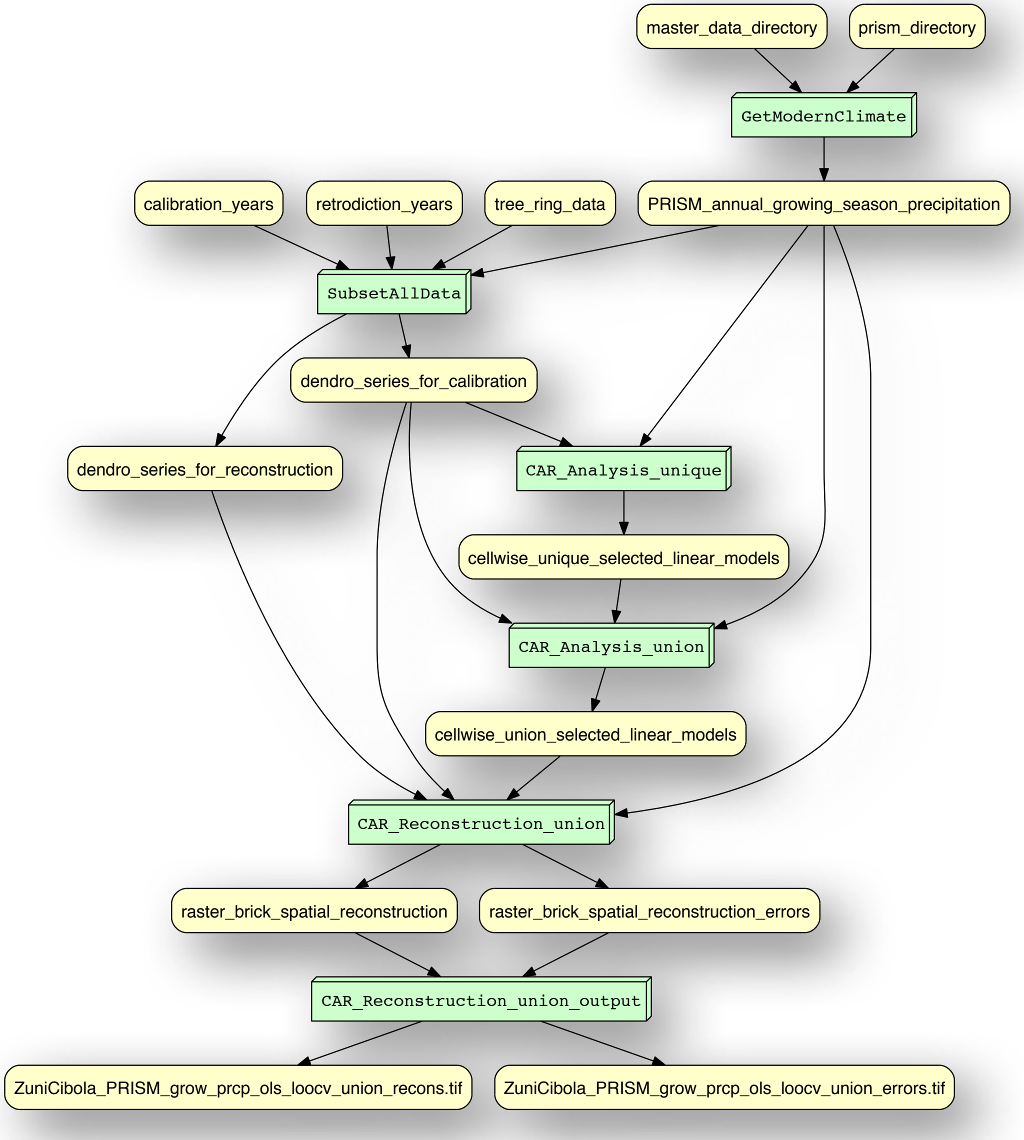
Paleoclimate Reconstruction (openSKOPE.org)

- … explained using **YesWorkflow!**

Kyle B., (computational) archaeologist:

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





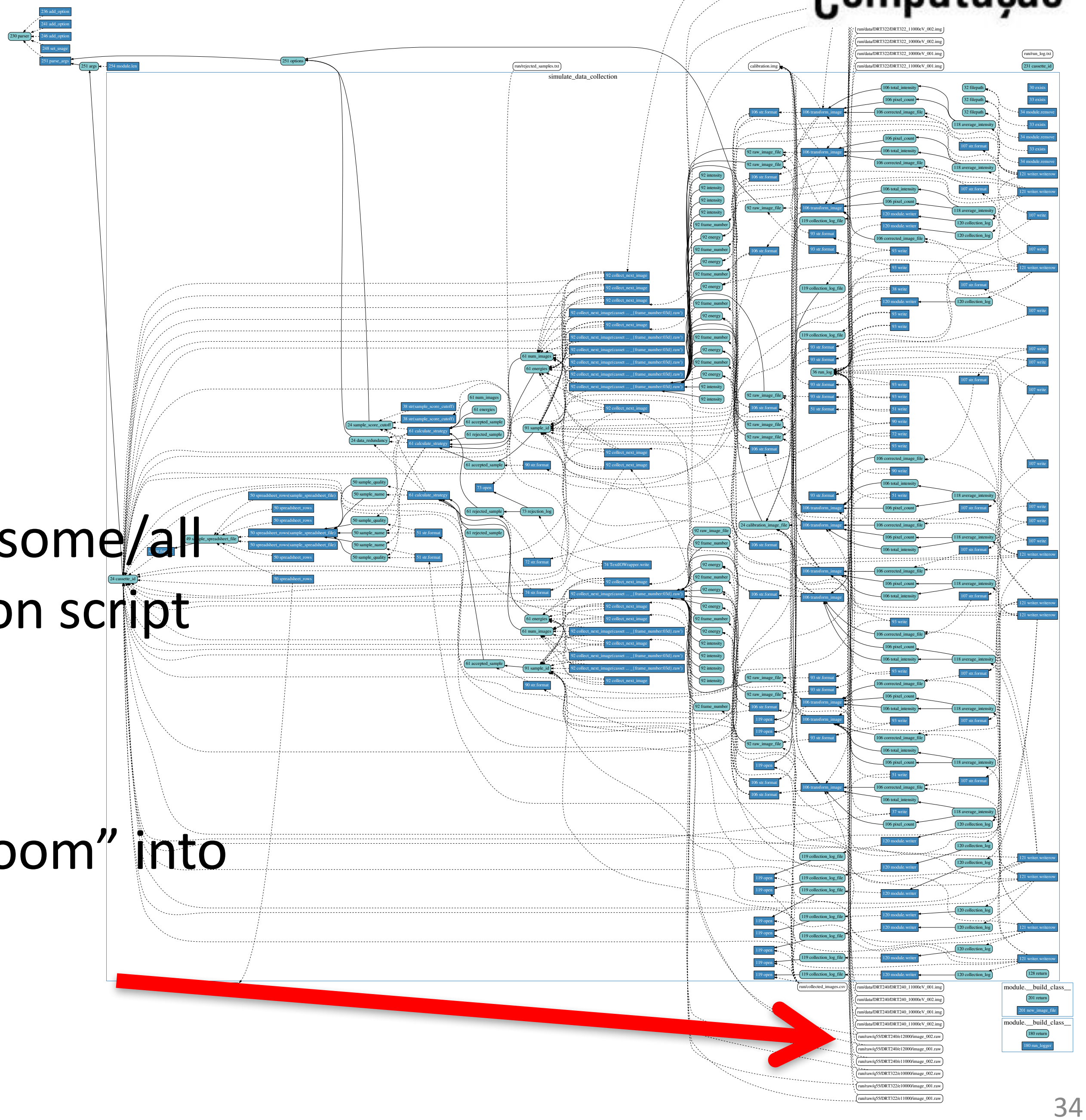
master_data_directory
prism_directory
GetModernClimate
calibration_years
retrodiction_years
tree_ring_data
PRISM_annual_growing_season_precipitation
SubsetAllData
dendro_series_for_calibration
dendro_series_for_reconstruction
CAR_Analysis_unique
cellwise_unique_selected_linear_models
CAR_Analysis_union
cellwise_union_selected_linear_models
CAR_Reconstruction_union
raster_brick_spatial_reconstruction
raster_brick_spatial_reconstruction_errors
CAR_Reconstruction_union_output
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_recons.tif
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_errors.tif

SKOPE + **Kurator**

+ **DataONE**
Data Observation Network for Earth

=> YesWorkflow.org

noWorkflow: not only Workflow!

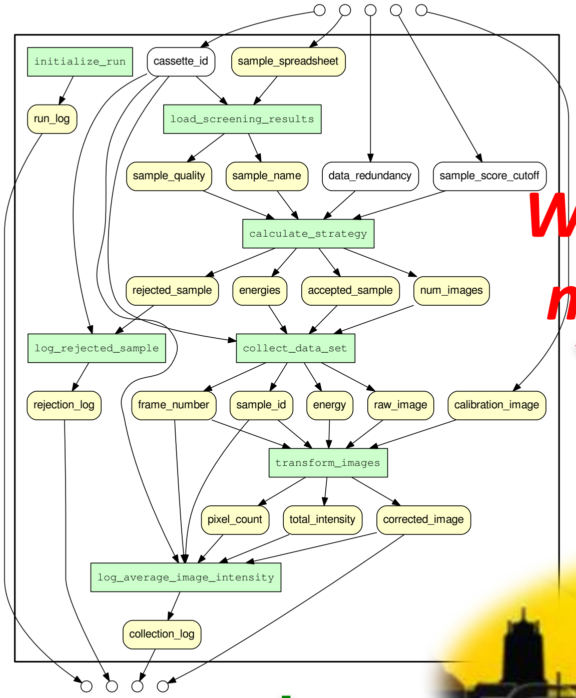


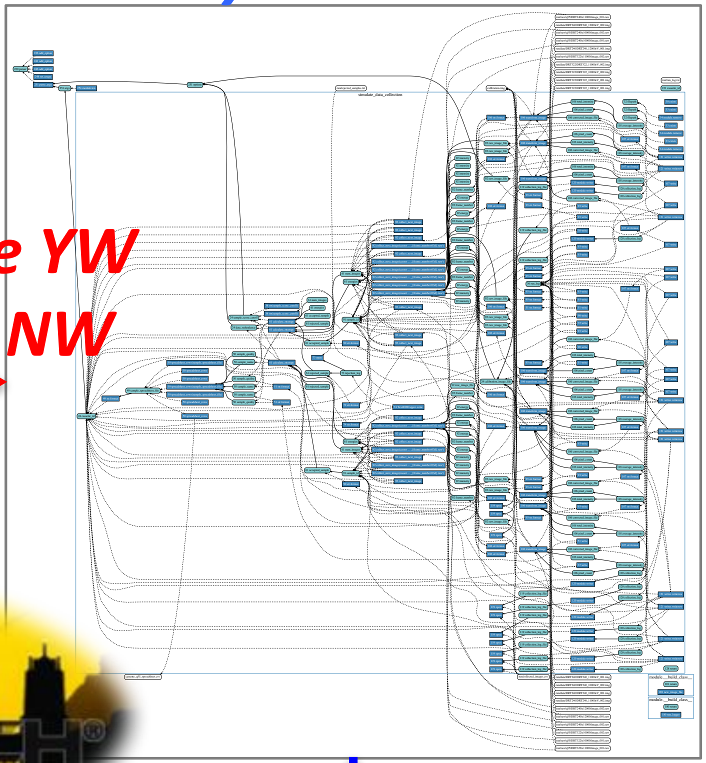
- **Transparently** capture some/all provenance from Python script **runs**.

- Use filter **queries** to “zoom” into relevant parts ..

YesWorkflow: ***Conceptual*** *workflow model*

noWorkflow: ***Python*** *trace model*

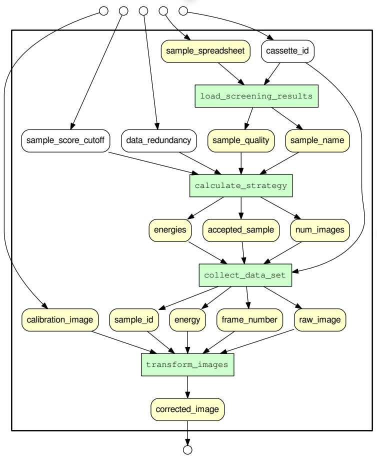




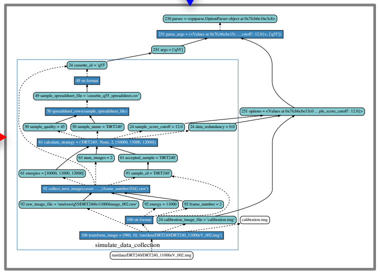
Would like to use YW model to query NW data!

lineage query

lineage query



But how do we bridge this gap???

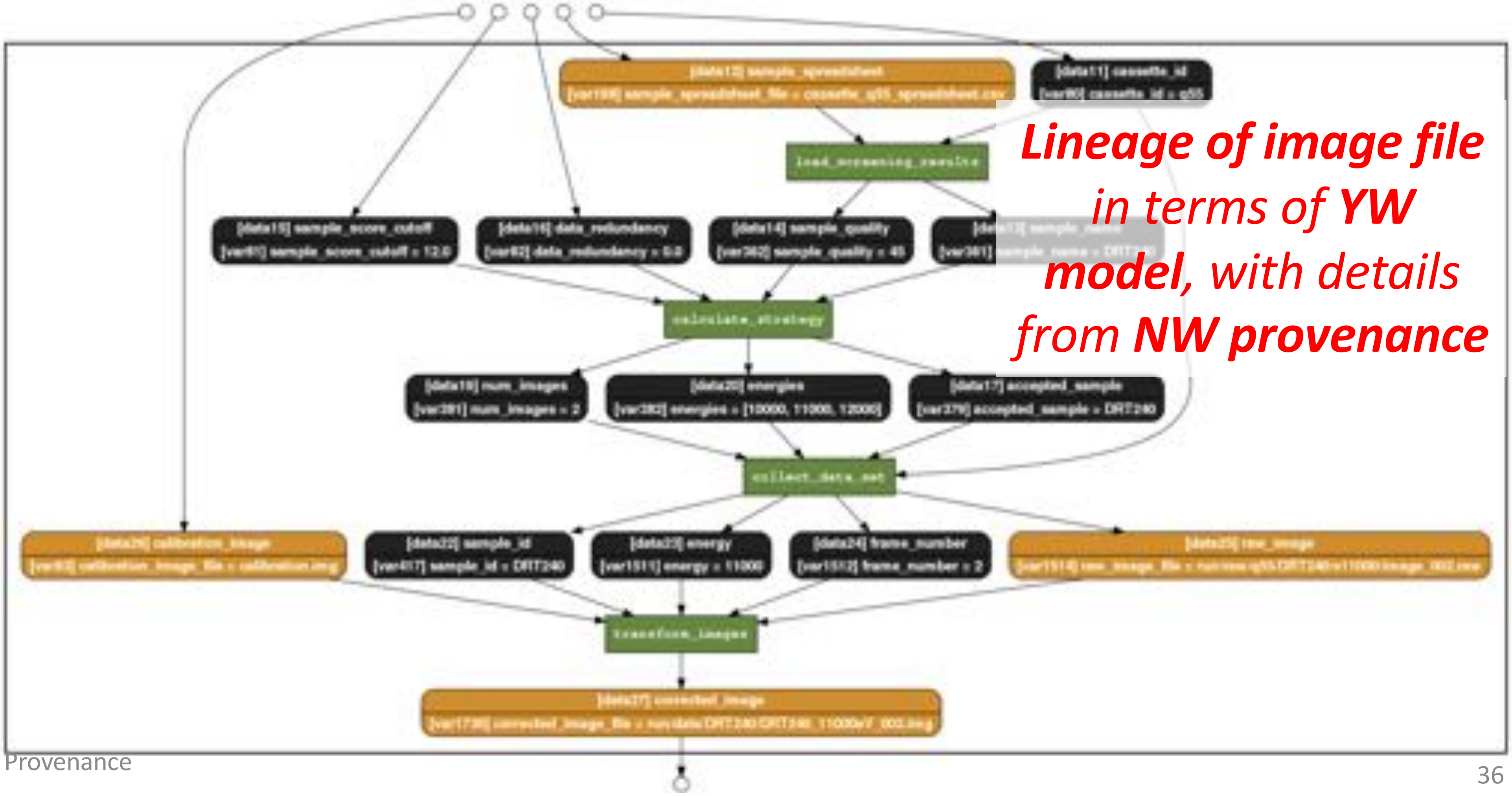


Habemus Pons!

We’ve got the Bridge!

The bridge is the journey..

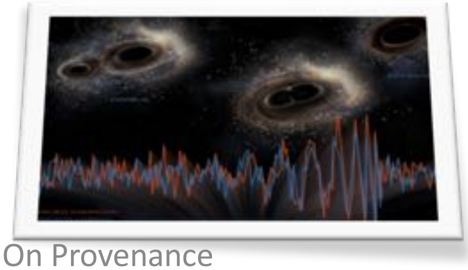
(The journey is the destination)

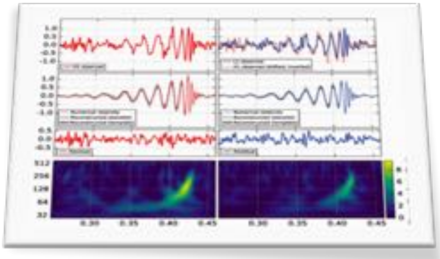

[data12] sample_spreadsheet
[var198] sample_spreadsheet_file = cassette_q55_spreadsheet.csv
[data11] cassette_id
[var80] cassette_id = q55
load_screening_results
[data15] sample_score_cutoff
[var81] sample_score_cutoff = 12.0
[data16] data_redundancy
[var82] data_redundancy = 0.0
[data14] sample_quality
[var382] sample_quality = 45
[data13] sample_name
[var381] sample_name = DRT240
calculate_strategy
[data19] num_images
[var381] num_images = 2
[data20] energies
[var382] energies = [10000, 11000, 12000]
[data17] accepted_sample
[var379] accepted_sample = DRT240
collect_data_set
[data26] calibration_image
[var83] calibration_image_file = calibration.img
[data22] sample_id
[var417] sample_id = DRT240
[data23] energy
[var1511] energy = 11000
[data24] frame_number
[var1512] frame_number = 2
[data25] raw_image
[var1514] raw_image_file = run/raw/q55/DRT240/e11000/image_002.raw
transform_images
[data27] corrected_image
[var1730] corrected_image_file = run/data/DRT240/DRT240_11000eV_002.img

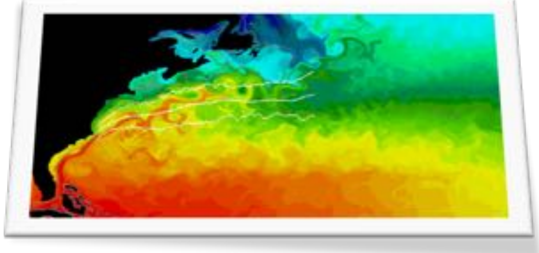
Lineage of image file *in terms of* ***YW model****, with details from* ***NW provenance***

YW-IDCC’17 Demo Use Cases

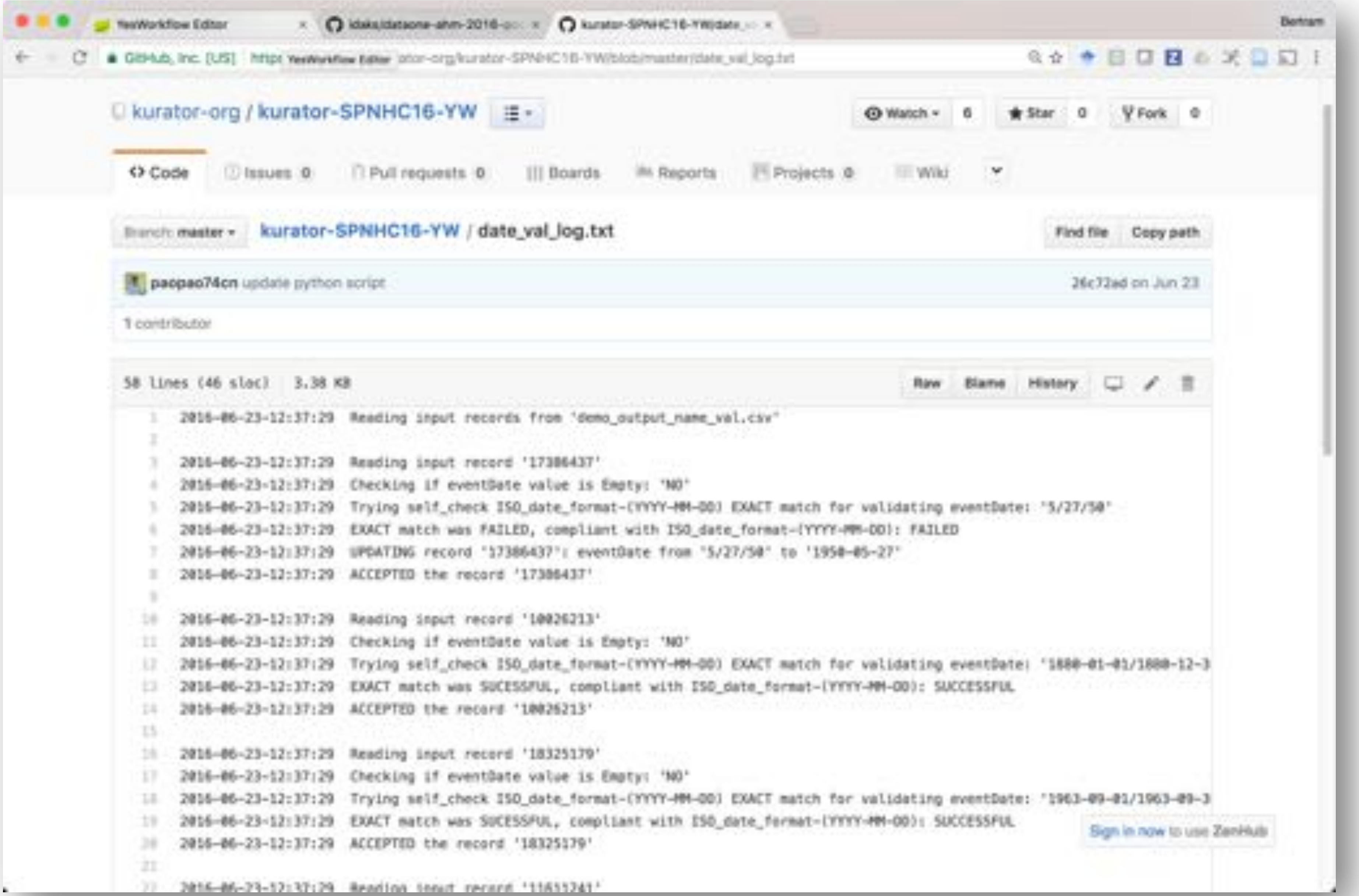
Domain	Use case	Programming language	Provenance methods
Climate science	C3C4	MATLAB	YW + MATLAB RunManager
Astrophysics	LIGO	Python	YW + NW (code-level)
Protein crystal samples	Simulate data collection	Python	YW + NW (code-level)
Biodiversity data curation	kurator-SPNHC	Python	YW-recon + YW-logging
Social network analysis	Twitter	Python	YW + NW (file-level)
Oceanography	OHIBC Howe Sound (multi-run multi-script)	R	YW + R RunManager







Finer-grained Provenance: User Log Files!



kurator-org / kurator-SPNHC16-YW

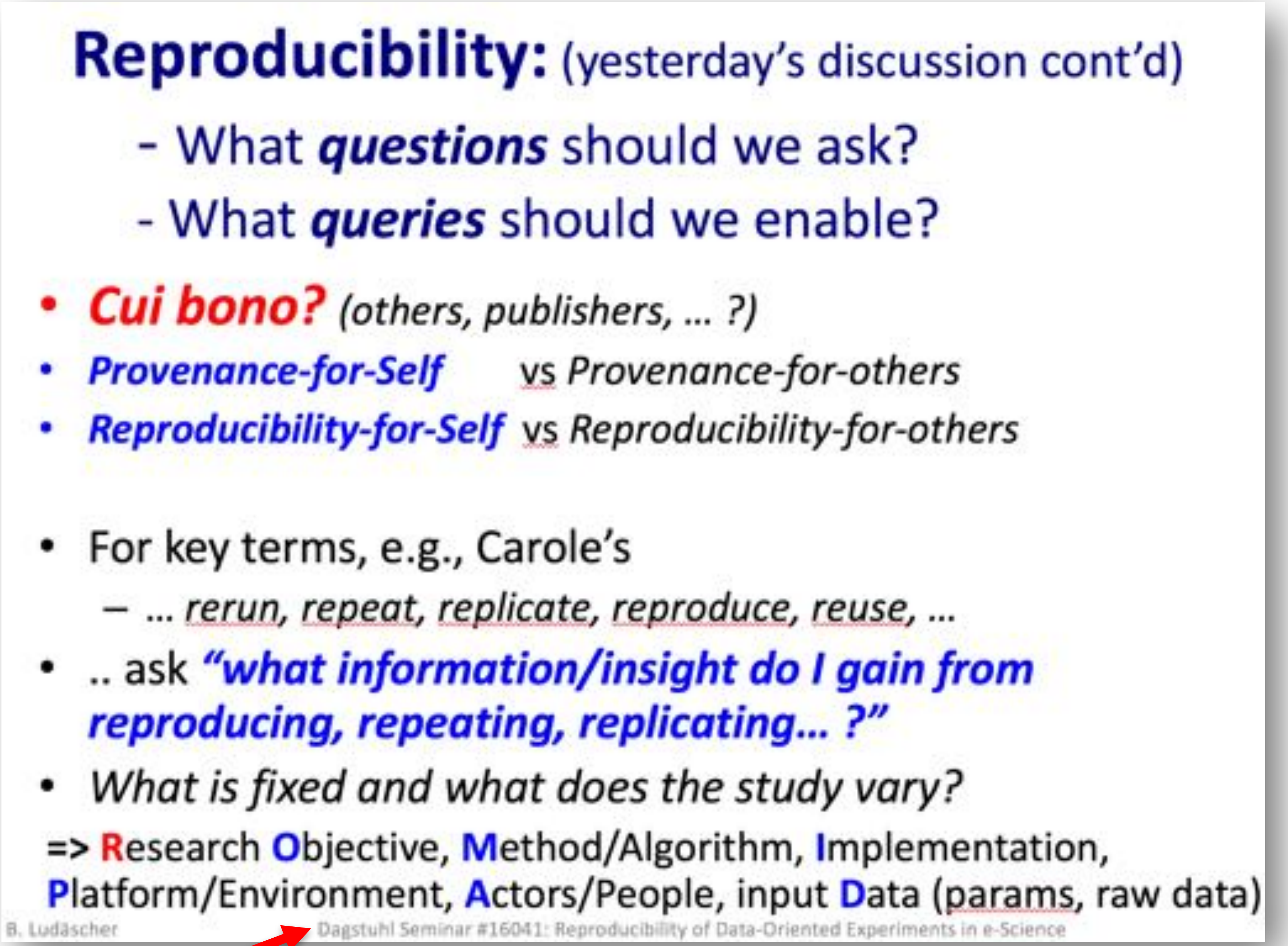
Branch: master ▾ kurator-SPNHC16-YW / date_val_log.txt

paopao74cn update python script 26c72ad on Jun 23

1 contributor

58 lines (46 sloc) 3.38 KB

```
2016-06-23-12:37:29  Reading input records from 'demo_output_name_val.csv'

2016-06-23-12:37:29  Reading input record '17386437'
2016-06-23-12:37:29  Checking if eventDate value is Empty: 'NO'
2016-06-23-12:37:29  Trying self_check ISO_date_format-(YYYY-MM-DD) EXACT match for validating eventDate: '5/27/50'
2016-06-23-12:37:29  EXACT match was FAILED, compliant with ISO_date_format-(YYYY-MM-DD): FAILED
2016-06-23-12:37:29  UPDATING record '17386437': eventDate from '5/27/50' to '1950-05-27'
2016-06-23-12:37:29  ACCEPTED the record '17386437'

2016-06-23-12:37:29  Reading input record '10026213'
2016-06-23-12:37:29  Checking if eventDate value is Empty: 'NO'
2016-06-23-12:37:29  Trying self_check ISO_date_format-(YYYY-MM-DD) EXACT match for validating eventDate: '1880-01-01/1880-12-3
2016-06-23-12:37:29  EXACT match was SUCESSFUL, compliant with ISO_date_format-(YYYY-MM-DD): SUCCESSFUL
2016-06-23-12:37:29  ACCEPTED the record '10026213'

2016-06-23-12:37:29  Reading input record '18325179'
2016-06-23-12:37:29  Checking if eventDate value is Empty: 'NO'
2016-06-23-12:37:29  Trying self_check ISO_date_format-(YYYY-MM-DD) EXACT match for validating eventDate: '1963-09-01/1963-09-3
2016-06-23-12:37:29  EXACT match was SUCESSFUL, compliant with ISO_date_format-(YYYY-MM-DD): SUCCESSFUL
2016-06-23-12:37:29  ACCEPTED the record '18325179'

2016-06-23-12:37:29  Reading input record '11611241'
```



Reproducibility: (yesterday's discussion cont'd)

- What ***questions*** should we ask?
- What ***queries*** should we enable?

- ***Cui bono?*** *(others, publishers, … ?)*
- ***Provenance-for-Self*** vs *Provenance-for-others*
- ***Reproducibility-for-Self*** vs *Reproducibility-for-others*

- For key terms, e.g., Carole's
  - … *rerun, repeat, replicate, reproduce, reuse*, …
- .. ask ***"what information/insight do I gain from reproducing, repeating, replicating… ?"***
- *What is fixed and what does the study vary?*

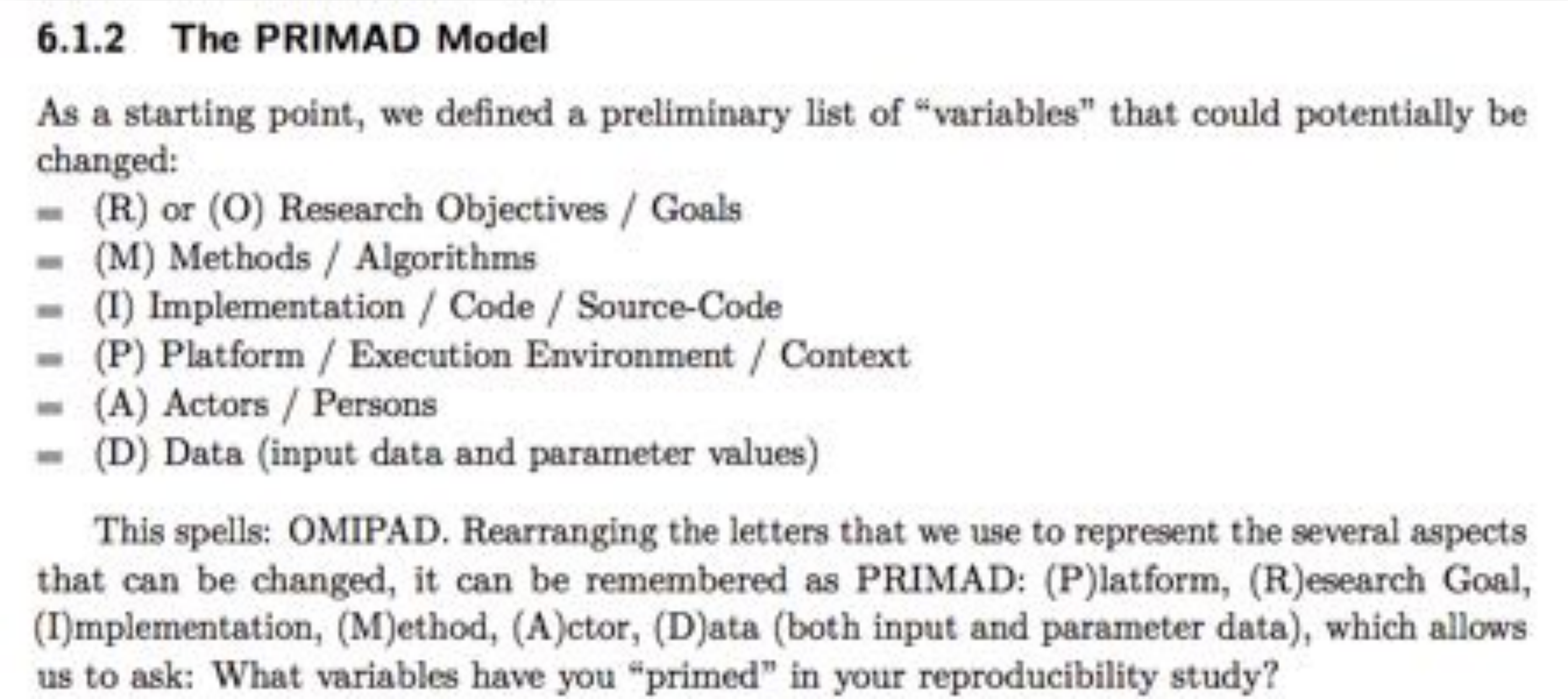
=> **R**esearch **O**bjective, **M**ethod/Algorithm, **I**mplementation, **P**latform/Environment, **A**ctors/People, input **D**ata (params, raw data)

B. Ludäscher Dagstuhl Seminar #16041: Reproducibility of Data-Oriented Experiments in e-Science

Reproducibility Crisis *(reprised)*

- **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 …*

PRIMAD *(what have you "primed"?)*



6.1.2 The PRIMAD Model

As a starting point, we defined a preliminary list of "variables" that could potentially be changed:

- (R) or (O) Research Objectives / Goals
- (M) Methods / Algorithms
- (I) Implementation / Code / Source-Code
- (P) Platform / Execution Environment / Context
- (A) Actors / Persons
- (D) Data (input data and parameter values)

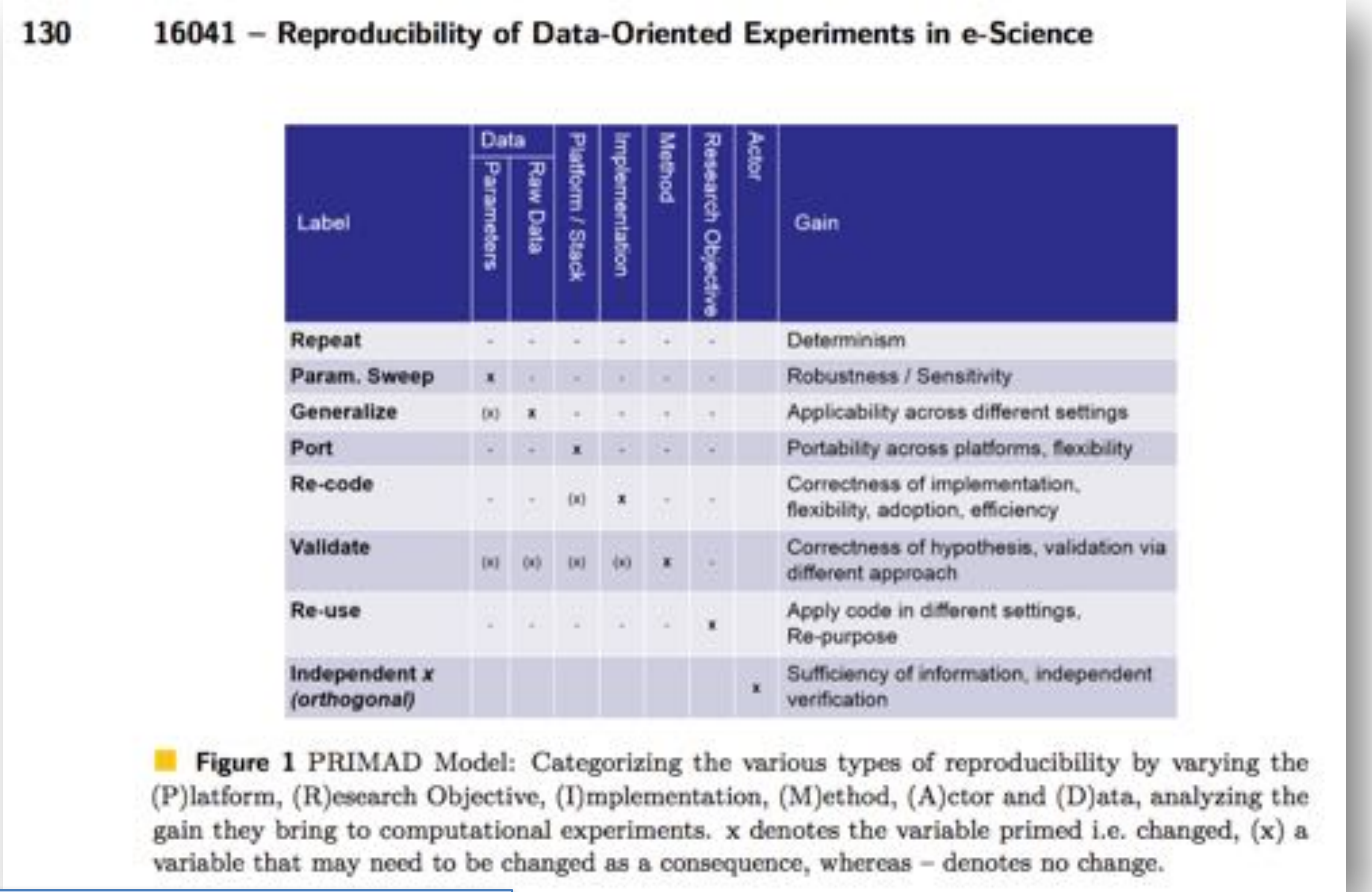
This spells: OMIPAD. Rearranging the letters that we use to represent the several aspects that can be changed, it can be remembered as PRIMAD: (P)latform, (R)esearch Goal, (I)mplementation, (M)ethod, (A)ctor, (D)ata (both input and parameter data), which allows us to ask: What variables have you "primed" in your reproducibility study?

Dagstuhl Seminar #16041 Report

Outputs = Exec(M,I,P,D) | RO, A

- M = parsimony/bootstrap/..
- I = package XYZ
- P = MacOS ..
- D = (Params, Files)

PRIMAD *(what have you "primed"?)*



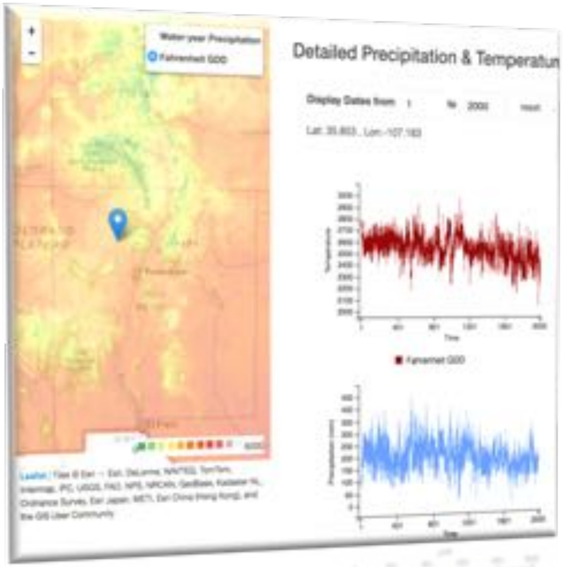
130 16041 – Reproducibility of Data-Oriented Experiments in e-Science

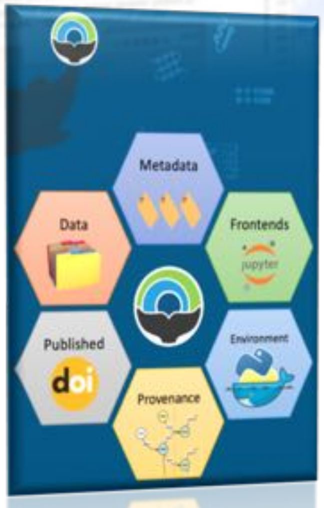
Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	x	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	x	-	-	-	-		Applicability across different settings
Port	-	-	x	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	x	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	x	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	x		Apply code in different settings, Re-purpose
Independent *x* *(orthogonal)*							x	Sufficiency of information, independent verification

Figure 1 PRIMAD Model: Categorizing the various types of reproducibility by varying the (P)latform, (R)esearch Objective, (I)mplementation, (M)ethod, (A)ctor and (D)ata, analyzing the gain they bring to computational experiments. x denotes the variable primed i.e. changed, (x) a variable that may need to be changed as a consequence, whereas – denotes no change.

Dagstuhl Seminar #16041 Report

Some related Projects ..

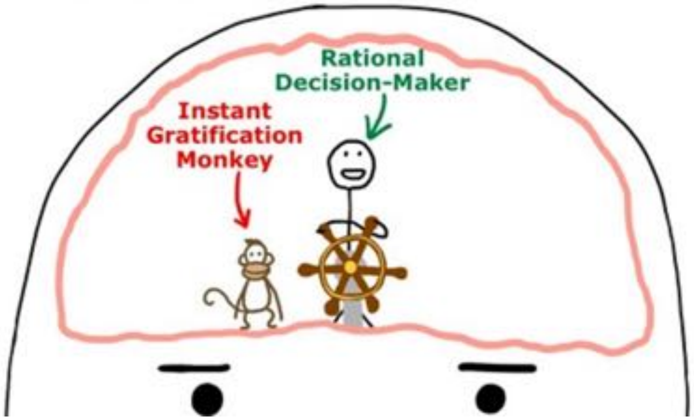

Water year Precipitation
Fahrenheit GDD
Detailed Precipitation & Temperature
Display Dates from 1 to 2000 result
Lat: 35.863 , Lon: -107.183
Fahrenheit GDD



- **DataONE:** …
  ➔ **Posters** by Linh Hoang & Hui Lyu, Xiaoliang Jiang
- **SKOPE**: system and tools to discover, access, analyze, visualize **paleoenvironmental data**
  – unprecedented ability to explore provenance (detailed, comprehensible record of computational derivation of results)
  – for researchers, tinkerers, and modelers
  ➔ **WT /SKOPE poster** by Pratik Shrivastava
- **Whole Tale:**
  – **leverage & contribute to existing CI** to support the whole tale (“living paper”), from workflow run to scholarly publication
  – integrate tools & CI (DataONE, Globus, iRODS, NDS, ...) to **simplify use and promote best practices.**
  – **driven by science WGs** (Archaeology/SKOPE, materials science, astro, bio ..)

Preliminary Conclusions

- **Goal**: allow researchers (+*tinkers*, +*modelers*) to tell the whole tale of a science study, transparently and reproducibly.
- **Provenance** …
  - … is key to transparency, reproducibilty, comprehensibility
  - ... comes in many (hybrid) forms (*workflow graphs*, *log files*, *trace events*, ...)
  - … is metadata (=> “*a love note to the future*”)
  - … should be *actionable today* (feed both, your IGM & RDM)
- **Provenance-for-Self** …
  - ... asks: *how does provenance help me get my work done today?*
  - … is what provenance technologists and tool builders could/should do more of!



Inside the mind of a master procrastinator (TED Talk by Tim Urban)

Truth or Consequences … (ice breaker)

- Which of these are true/false?
- As a high-school student I …
  - ... worked at a nuclear power plant
  - ... worked at a historic Roman bath
  - ... migrated code from FORTRAN to Pascal

- Later in life I …
  - … toured the backwaters from Cochin to Alleppey
  - … toured the Isle of Skye on motorbike
  - ... toured Sri Lanka on motorbike
  - ... became a national master in chess
  - ... discovered game provenance, invented provenance games