YesWorkflow: A User-Oriented, Language-Independent Tool for **Recovering Workflow** Information **from Scripts**

Timothy McPhillips[1], Tianhong Song[2], Tyler Kolisnik[3], Steve Aulenbach[4], Khalid Belhajjame[5], Kyle Bocinsky[6], Yang Cao[1], Fernando Chirigati[7], Saumen Dey[2], Juliana Freire[7], Deborah Huntzinger[11], Christopher Jones[8], David Koop[9], **Paolo Missier**[10], Mark Schildhauer[8], Christopher Schwalm[11], Yaxing Wei[12], James Cheney[13], Mark Bieda[3], **Bertram Ludäscher**[1, 14]

[1]Graduate School for Library and Information Science (GSLIS), University of Illinois at Urbana-Champaign (UIUC); [2]Dept. of Computer Science, University of California, Davis; [3]University of Calgary; [4]University Corporation for Atmospheric Research (UCAR) and U.S. Global Change Research Program (USGCRP); [5]Paris Dauphine University, LAMSADE; [6]Department of Anthropology, Washington State University, Pullman, WA; [7]New York University; [8]University of California, Santa Barbara; [9]University of Massachusetts, Dartmouth; [10]University of Newcastle, UK; [11]Northern Arizona University; [12]Oak Ridge National Laboratoy; [13]University of Edinburgh, Scotland; [14]National Center for Advanced Supercomputing Applications (NCSA), UIUC.

10th Intl. Digital Curation Conference (IDCC)
30 Euston Square, London, UK

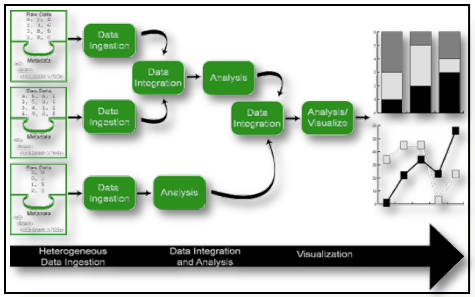
Overview

- *Enter* **Scientific Workflows** …
  - Kepler, Taverna, VisTrails, …
- … *Exeunt*
- *Enter* **Scripts** …
  - Python, R, Matlab, …

- *Enter* **YesWorkflow**
  - Complements **noWorkflow**
    - … ***runtime*** (= *retrospective*) provenance from scripts
  - Combines the best of both worlds:
    - Scripts + **Comments**

    => **Workflow Views** (***prospective*** provenance) from scripts

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**

TRIANA

- **Provenance**
  - wfs should capture **processing history**, **data lineage**

➔ traceable data- and wf-evolution

➔ **Reproducible Science**

ASKALON

Taverna

Science Example: Paleoclimate Reconstruction

- *Kohler & Bocinsky*: study rain-fed maize of **Ancestral Pueblo, 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 reconstructing climate at a given time and place.

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

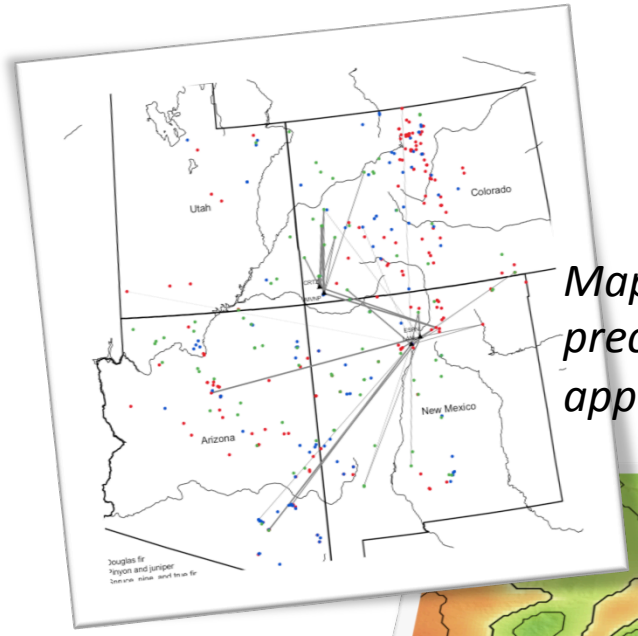
```
## Gene Ontology Statistics are Calculated Here.

 # 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)))
```

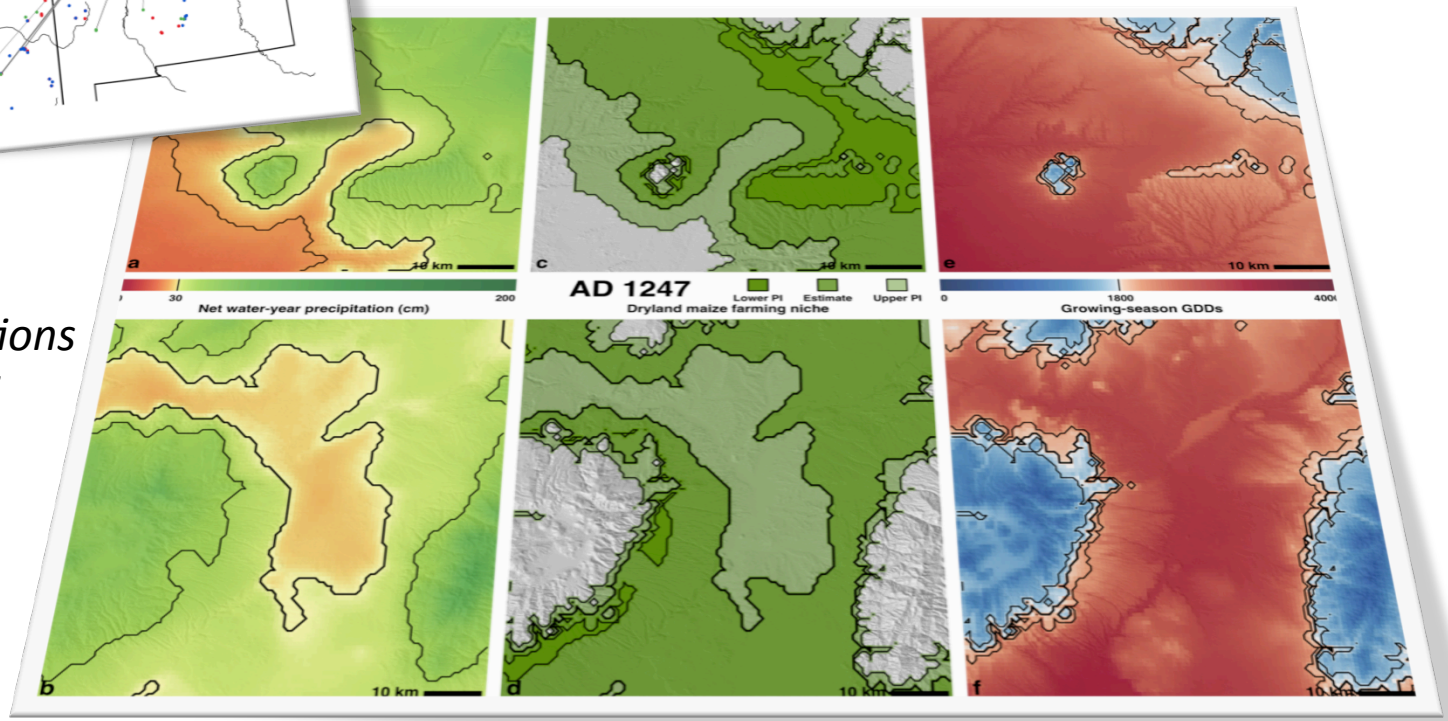
… implemented as an R Script …

… Paleoclimate Reconstruction …



Map showing the "selected" trees for reconstructing precipitation at four sites in our CAR regression approach (Correlation-Adjusted corRelation).

Reconstructions for AD 1247



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

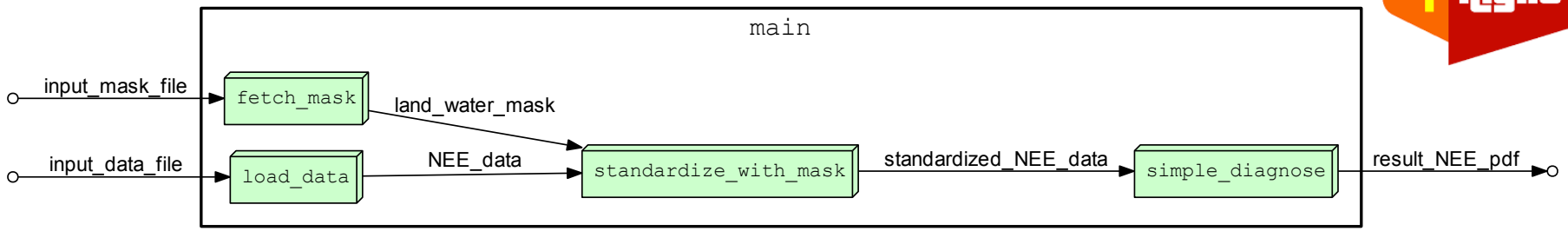
YesWorkflow = Scripts + **Comments**

- Scripts can be hard to digest, communicate
- Idea:
  - Add structured comments (cf. JavaDoc)

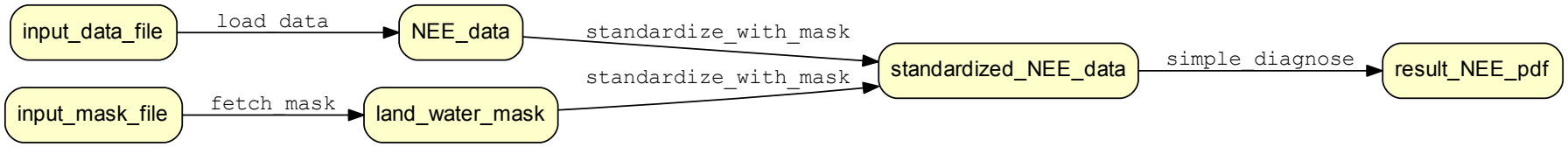
  => reveal **workflow structure** and **dataflow**

  => obtain some scientific workflow benefits
- … ASAP …

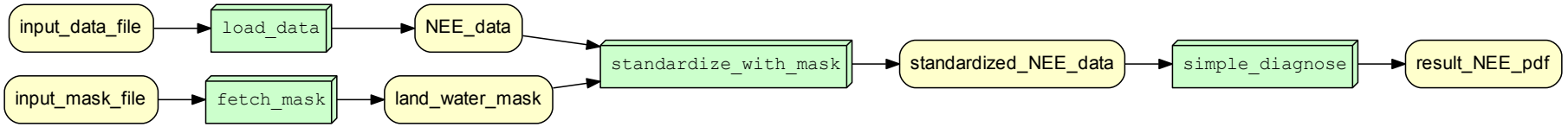
Get 3 views for the price of 1!



Process view



Data view



Combined view

User Comments: YW Annotations

```
188  ## @begin GO_Analysis
189  # @in hgCutoff @as GO_stats_p_value_cutoff
190  # @in higheridrlinkedtogenes @as DEG_list_higher_in_test_condition
191  # @in loweridrlinkedtogenes @as DEG_list_lower_in_test_condition
192  # @out gostatshigher @as GO_stats_gene_list_higher_in_test_condition
193  # @out BP_SummH_File @as GO_stats_BP_higher_in_test_condition
194  # @out CC_SummH_File @as GO_stats_CC_higher_in_test_condition
195  # @out MF_SummH_File @as GO_stats_MF_higher_in_test_condition
196  # @out gostatslower @as GO_stats_gene_list_lower_in_test_condition
197  # @out BP_SummL_File @as GO_stats_BP_lower_in_test_condition
198  # @out CC_SummL_File @as GO_stats_CC_lower_in_test_condition
199  # @out MF_SummL_File @as GO_stats_MF_lower_in_test_condition
200
201  ############################ Begin GOStats Block ##########################
202
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  gostatshigher <- higheridrlinkedtogenes[1]
207  higherstatsfilename <- paste(outputDirectory, "/", runName, "_", conditions[1],"_GOStatsHigher_",mytestcond[1], "_vs_",baseline,".
208  write.table(gostatshigher,file=higherstatsfilename, row.names=FALSE, col.names=FALSE, quote=FALSE, sep="\t")
209  geneListHigherCHR <- gostatshigher$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)))
```

@begin GO_Analysis

@in hgCutoff

@in …

@out BP_Summl_file

@out …

@end GO_Analysis

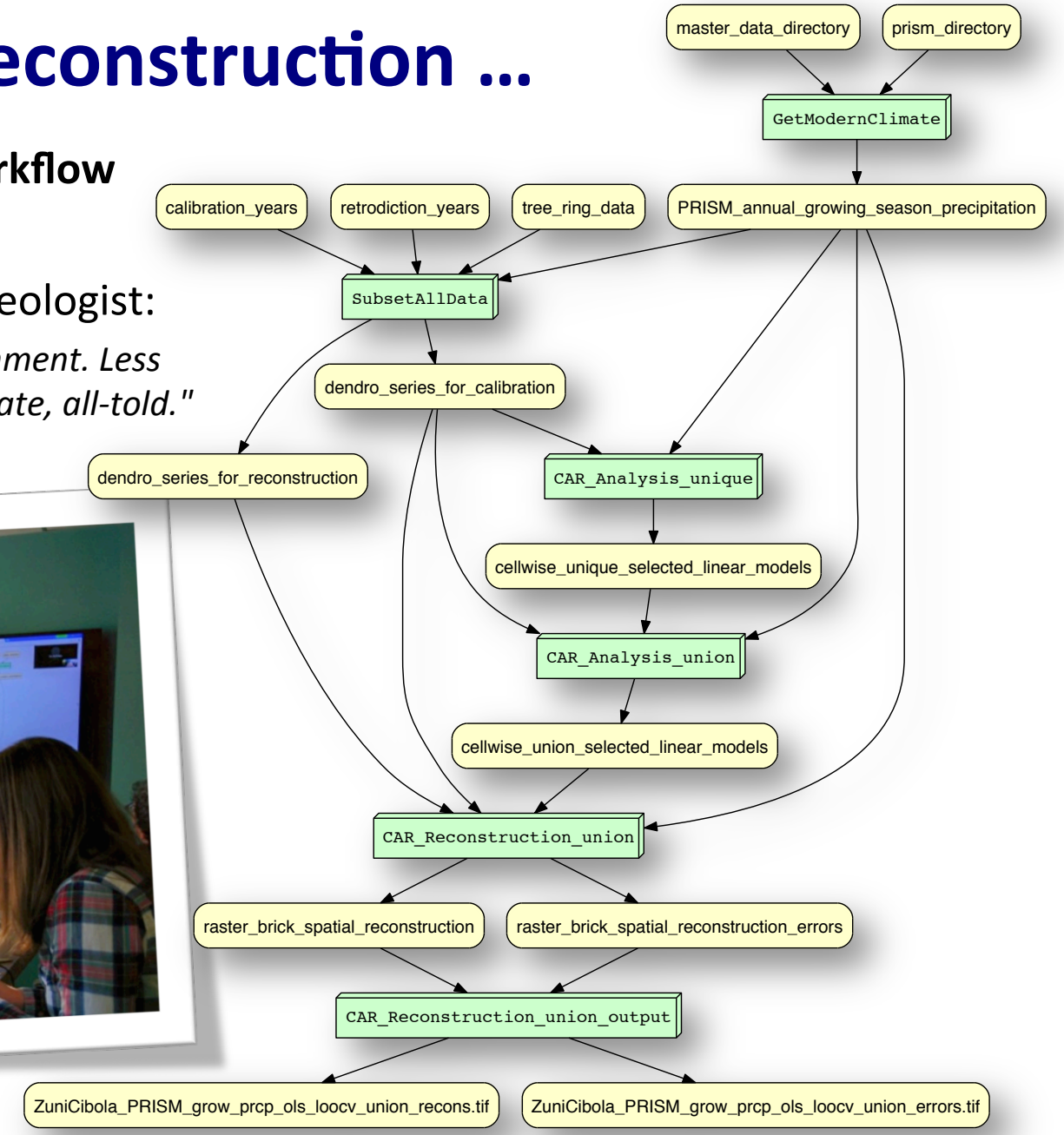
. . .

Paleoclimate Reconstruction …

- … explained using **YesWorkflow**

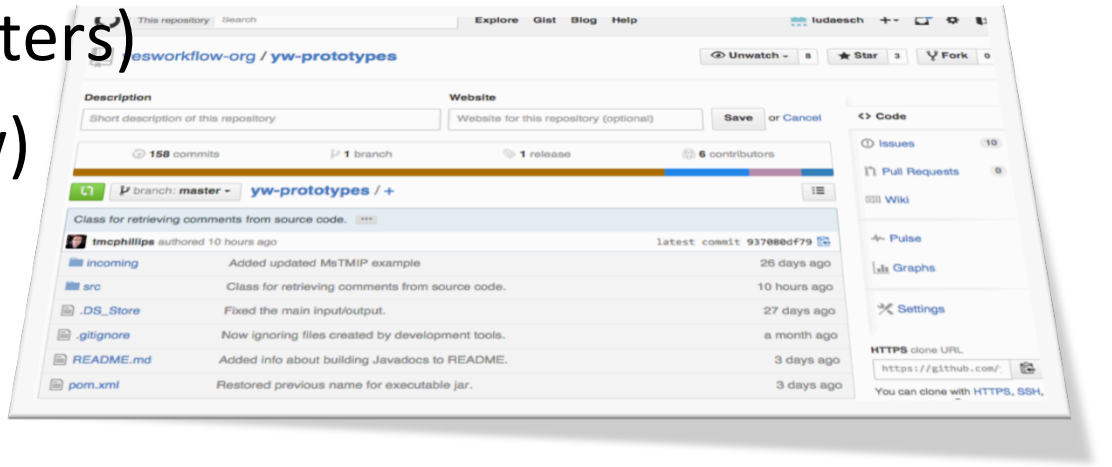
Kyle B., (computational) archeologist:

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



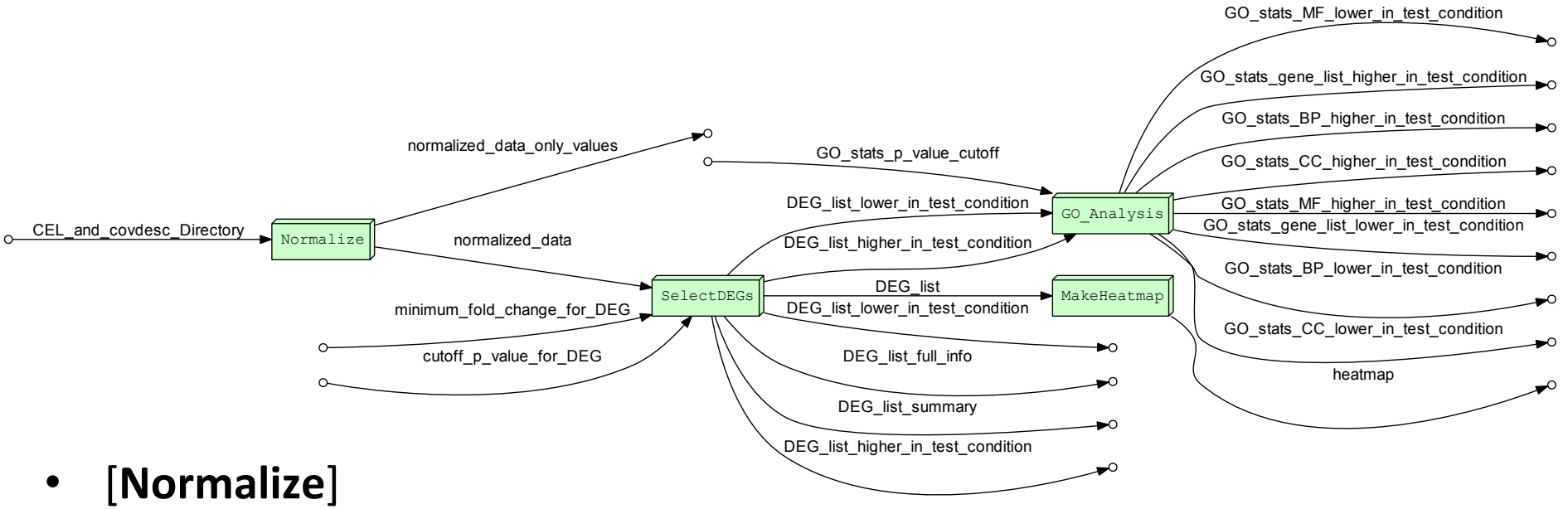
YesWorkflow Architecture: KISS!

- **YW-Extract**
  - … structured comments
- **YW-Model**
  - Program Block, Workflow
  - Port (data, parameters)
  - Channels (dataflow)



- **YW-Graph**
  - … using GraphViz/DOT files
- **YW-Query**, **YW-Validate**, **YW-CLI**

Gene Expression Microarray Data Analysis



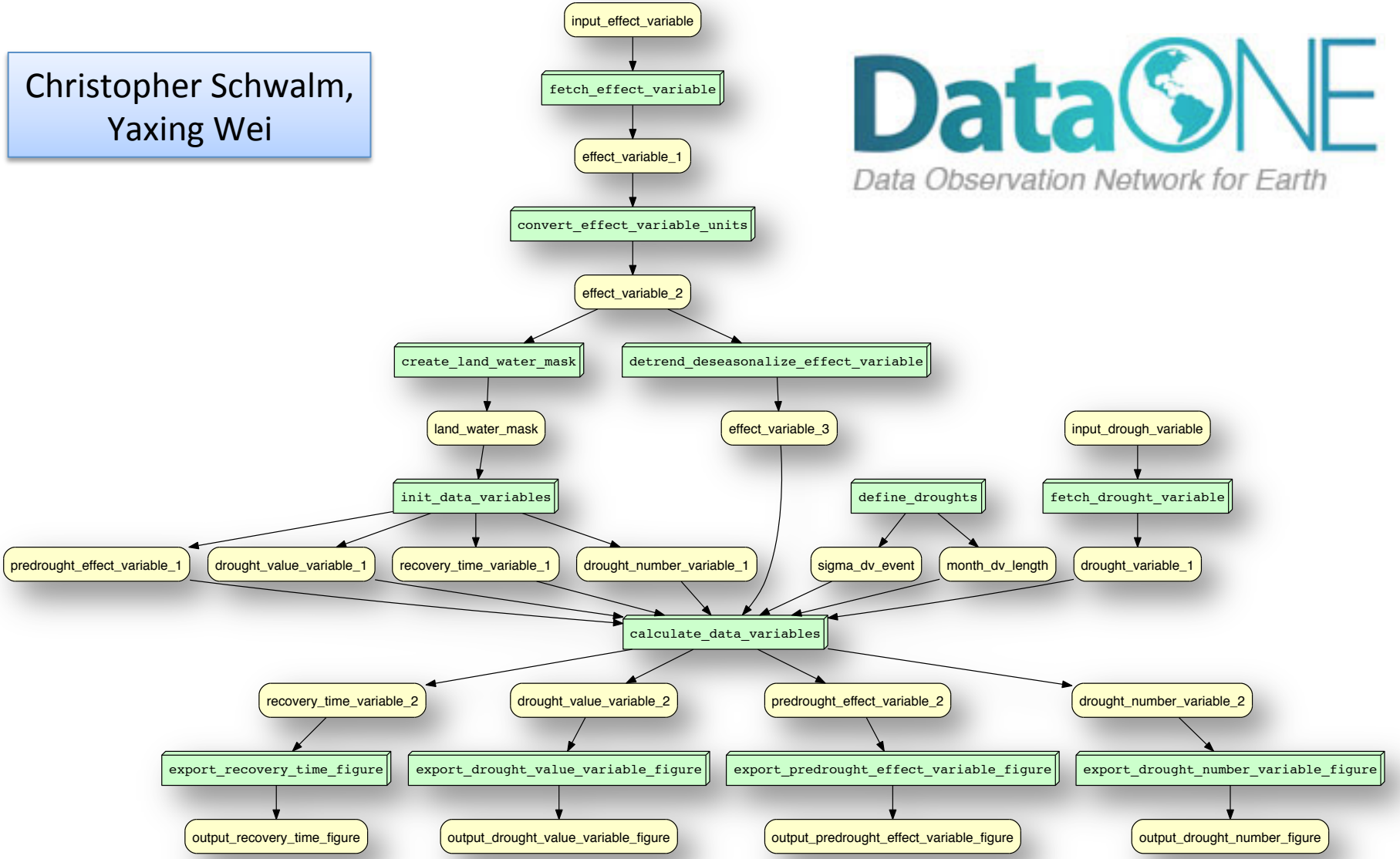
- [**Normalize**]
  - Normalization of data across microarray datasets
- [**SelectDEGs**]
  - Selection of differentially expressed genes between conditions
- [**GO Analysis**]
  - determination of gene ontology statistics for the resulting datasets
- [**MakeHeatmap**]
  - creation of a heatmap of the differentially expressed genes.

Tyler Kolisnik, Mark Bieda

Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP)

Christopher Schwalm,
Yaxing Wei

DataONE
Data Observation Network for Earth



Summary: Scientific Workflows

Scientific Workflows

- [+] **A**utomation
- [+] **S**calability
- [+] **A**bstraction
- [+] **P**rovenance
- …
- [+/0] Easy to use
  - [0] learning a new paradigm
- [-] Teaching resources
- [-] Special expertise needed for deep changes

  e.g. new Java actors, shims, …

Summary: Scripts + YesWorkflow

Scripts: [+] Automation, [0] Scalability, [-] Abstraction, [0/-] Provenance

Now: Scripts + YesWorkflow Annotations

- **[+] Abstraction**
  - explain your methods to mere mortals

  => encourage (re-)use
- **[+] Provenance:**
  - **noWorkflow** (***retrospective*** provenance)
  - **YesWorkflow** (***prospective*** provenance)
- [+] Language independent (R, Matlab, Python, …)
- [+] **Empower tool makers** (script programmers): give them …
  - … some **immediate** benefits (**workflow views**)
  - … some **medium** term improvements (**provenance integration**)
  - … some **long term** benefits: think about your methods differently

  => dataflow programming => [+] Scalability

YesWorkflow: Acknowledgements

- With support from NSF awards DBI-1356751 (*Kurator*), ACI-0830944 (*DataONE*), SMA-1439603 (*SKOPE*).

- Thanks to the IDCC organizers for choice of venues!