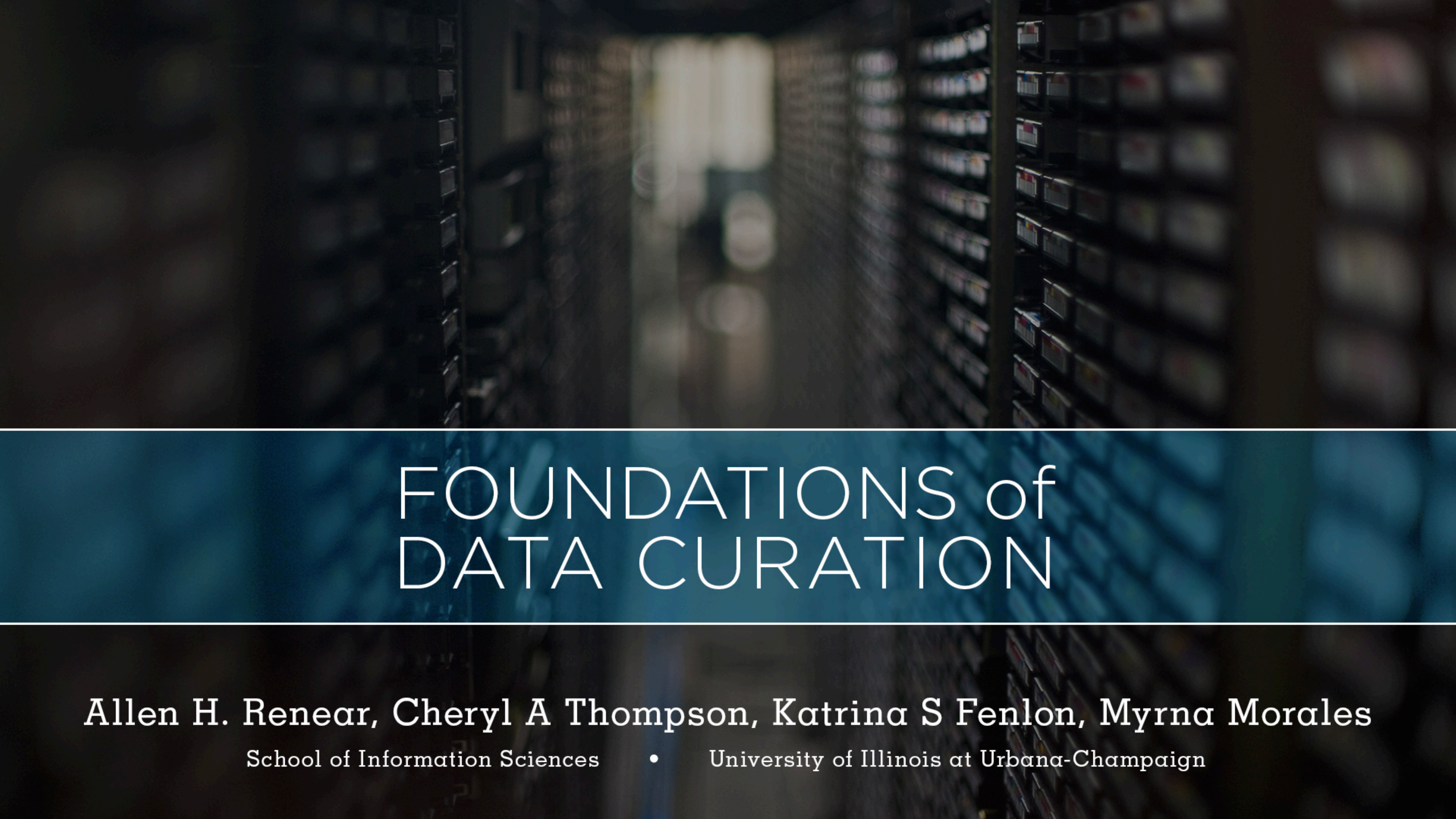


FOUNDATIONS of DATA CURATION

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales

School of Information Sciences • University of Illinois at Urbana-Champaign

Workflow and Provenance

(anything profound, and the cool slides, is from Bertram Ludäscher. Everything else is from Renear

Workflow and Provenance

V1. Workflow

V2: Provenance

V3: Workflow systems

V1. Workflow

What is [data] workflow?

Why is workflow important?

Kinds of transformations.

What is data workflow?

Much of our work with data, especially in scientific applications,
consists in *transforming one data set into another*

In the abstract:
a software process embodying one or more algorithms
takes one dataset as input
and produces another as output

This process is what we mean (here) by workflow

These transformation scenarios range widely in kind and nature
and can be extremely complex.

They are fundamental to data science and a core focus of data curation

Data curation and data workflow

Data curation is concerned with transformations in *two* ways:

managing and documenting transformations involved in data analytics

performing transformation to realize data curation objectives.

(preservation, integration, format conversion, etc.)

Kinds of data transformations

Transformations where input and output datasets are identical in propositional content

transformation to a different data description language (or new version of a language)

transformation to a different serialization (or new version of a serialization)

Transformations where the input dataset mathematically contains the output dataset

transformation to a subset matching specific conditions
e.g. simple queries

transformation to a logically or mathematically entailed data of the same kind
e.g., summaries, statistics, visualizations

Transformations where the input dataset scientifically contains in the output dataset

transformation to scientifically entailed data of the same kind
here the resulting data set typically contains information different in kind
e.g., a data set about air pressure is transformed to a dataset about altitudes.

Workflow: Chaotic vs Organized vs Supported

There is always workflow.

even if it is an impenetrable, untrustworthy, chaotic mess

Organized workflows may be either

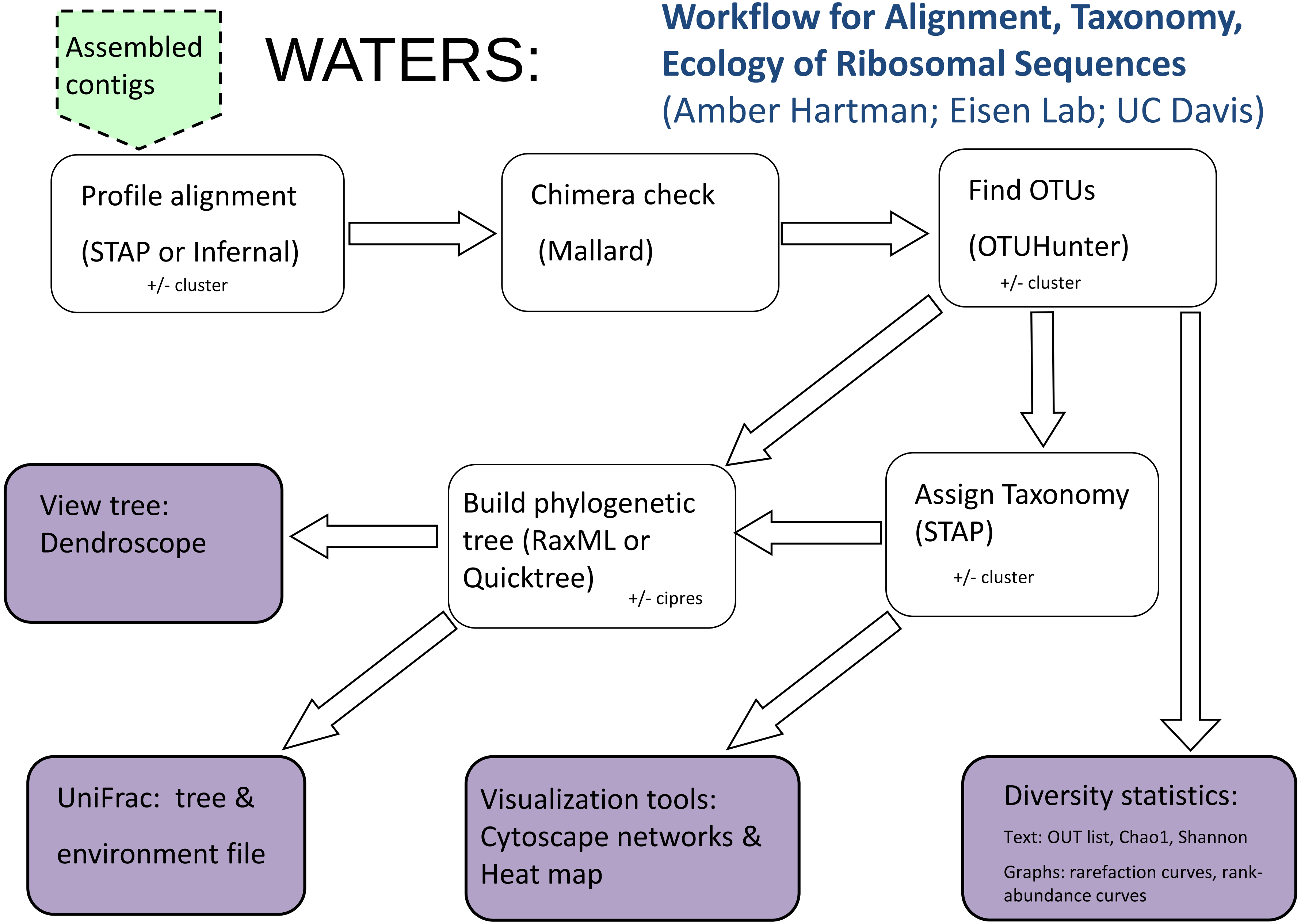
1) home grown (systems of documented scripts for instance)

2) supported by specialized workflow management systems (Kepler, Taverna, YesWorkflow, etc.)

Practical workflow support systems are usually more or less language independent, allowing the use of R, Python, XSLT, MATLAB, etc.

WATERS: Workflow for Alignment, Taxonomy, Ecology of Ribosomal Sequences

(Amber Hartman; Eisen Lab; UC Davis)



Assembled contigs

Profile alignment

(STAP or Infernal)

+/- cluster

Chimera check

(Mallard)

Find OTUs

(OTUHunter)

+/- cluster

Assign Taxonomy (STAP)

+/- cluster

Build phylogenetic tree (RaxML or Quicktree)

+/- cipres

View tree: Dendroscope

UniFrac: tree & environment file

Visualization tools: Cytoscape networks & Heat map

Diversity statistics:

Text: OUT list, Chao1, Shannon

Graphs: rarefaction curves, rank-abundance curves

Example Bioinformatics Workflow:

Motif-Catcher

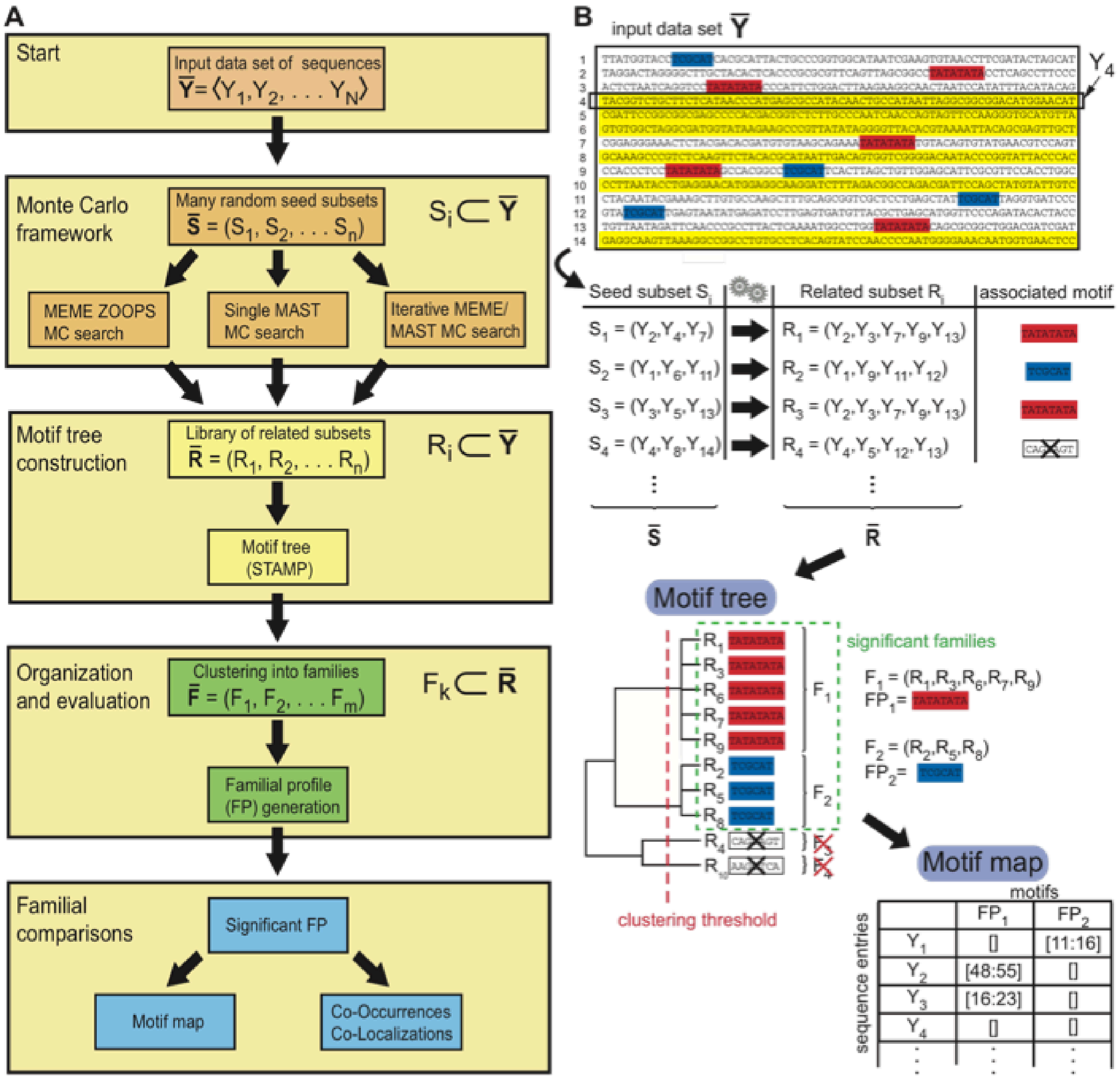


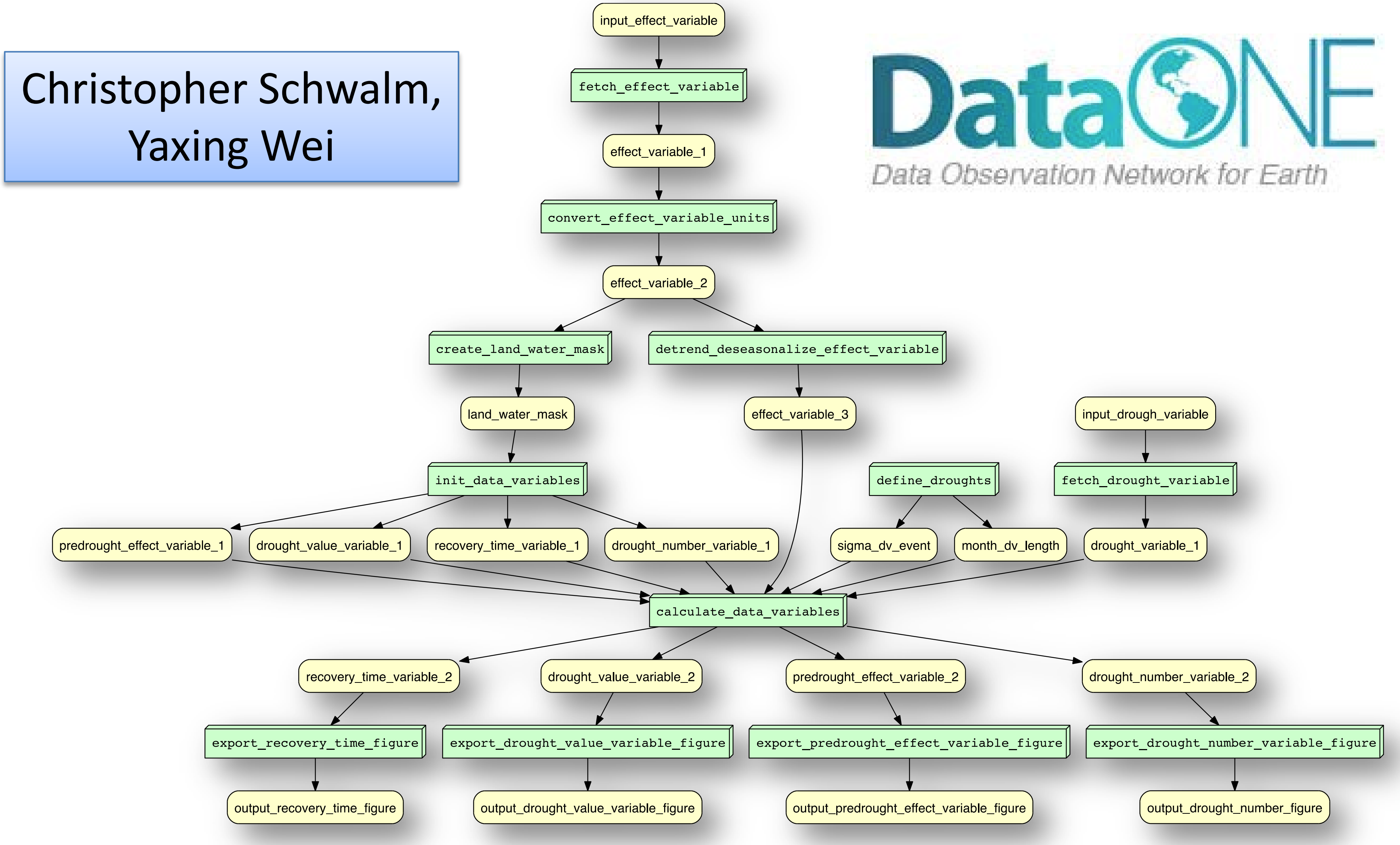
Figure 1: Concept of Monte-Carlo based detection and interpretation of motifs. A) Abstract description of MotifCatcher process. B) Examples illustrating the process with sample data.

Marc Facciotti *et al.*
UC Davis Genome Center

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

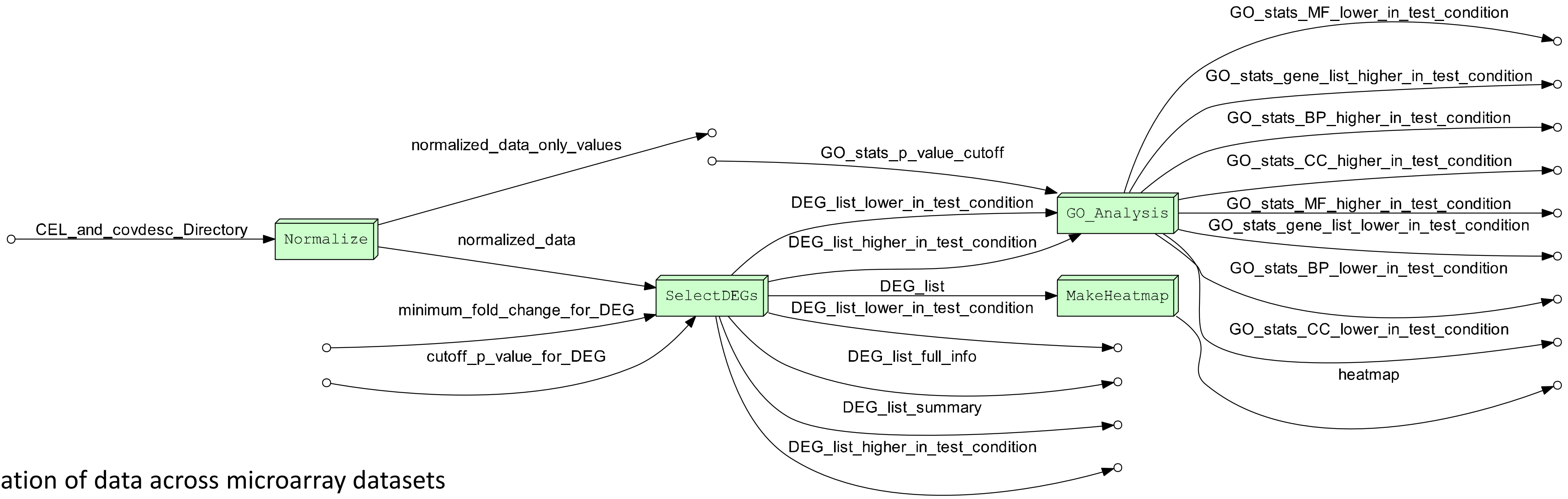
Christopher Schwalm,
Yaxing Wei

DataONE
Data Observation Network for Earth



input_effect_variable
fetch_effect_variable
effect_variable_1
convert_effect_variable_units
effect_variable_2
create_land_water_mask
detrend_deseasonalize_effect_variable
land_water_mask
effect_variable_3
input_drough_variable
init_data_variables
define_droughts
fetch_drought_variable
predrought_effect_variable_1
drought_value_variable_1
recovery_time_variable_1
drought_number_variable_1
sigma_dv_event
month_dv_length
drought_variable_1
calculate_data_variables
recovery_time_variable_2
drought_value_variable_2
predrought_effect_variable_2
drought_number_variable_2
export_recovery_time_figure
export_drought_value_variable_figure
export_predrought_effect_variable_figure
export_drought_number_variable_figure
output_recovery_time_figure
output_drought_value_variable_figure
output_predrought_effect_variable_figure
output_drought_number_figure

Gene Expression Microarray Data Analysis


CEL_and_covdesc_Directory
Normalize
normalized_data_only_values
normalized_data
minimum_fold_change_for_DEG
cutoff_p_value_for_DEG
SelectDEGs
GO_stats_p_value_cutoff
DEG_list_lower_in_test_condition
DEG_list_higher_in_test_condition
DEG_list
DEG_list_lower_in_test_condition
DEG_list_full_info
DEG_list_summary
DEG_list_higher_in_test_condition
GO_Analysis
MakeHeatmap
GO_stats_MF_lower_in_test_condition
GO_stats_gene_list_higher_in_test_condition
GO_stats_BP_higher_in_test_condition
GO_stats_CC_higher_in_test_condition
GO_stats_MF_higher_in_test_condition
GO_stats_gene_list_lower_in_test_condition
GO_stats_BP_lower_in_test_condition
GO_stats_CC_lower_in_test_condition
heatmap

- [**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

Why is workflow important

Thoughtfully designed organized workflows support:

Efficiency

Reliability

Modifiability

Reuse

Reproducibility

FOUNDATIONS OF DATA CURATION (IS531)

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales
School of Information Sciences
University of Illinois at Urbana-Champaign

Includes material adapted from work by Carole Palmer, Melissa Cragin, David Dubin, Karen Wickett, Bertram Ludaescher, Ruth Duerr and Simone Sacchi.

Comments and corrections to: renear@illinois.edu.