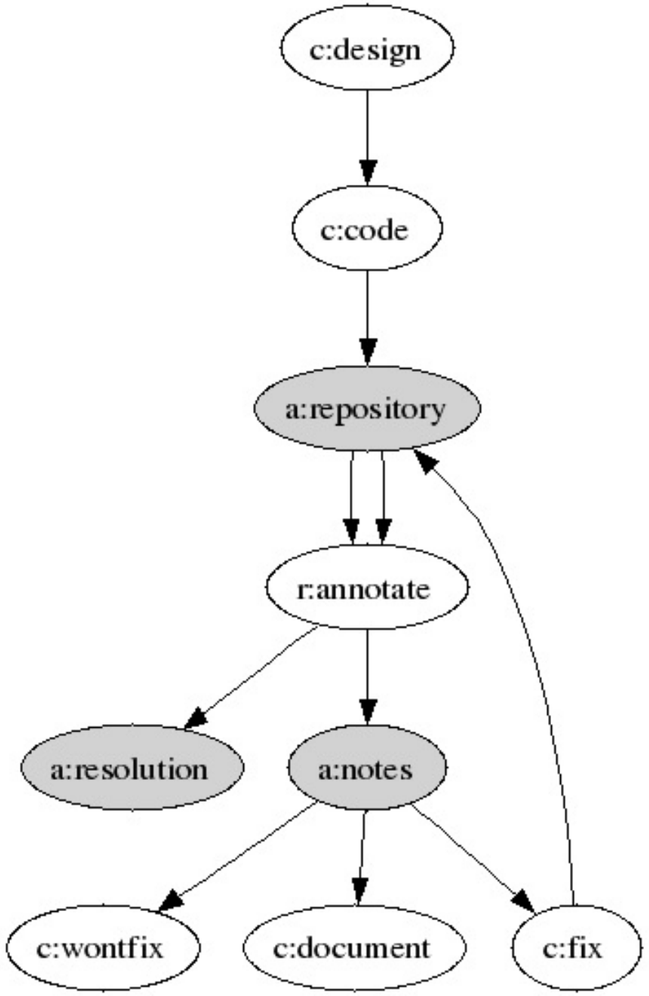
Data Analysis and Workflows

DataONE Community Engagement & Outreach Working Group

Lesson Topics

- Review of typical data analyses
- Reproducibility & provenance
- Workflows in general
- Informal workflows
- Formal workflows

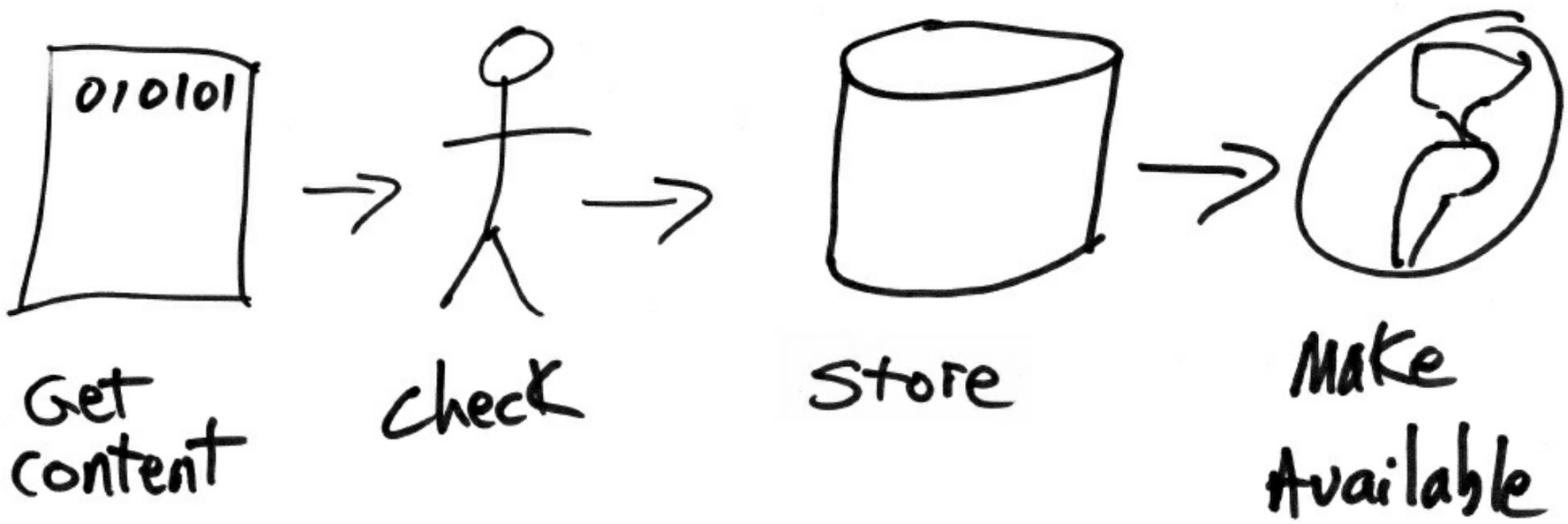


Code Review Process

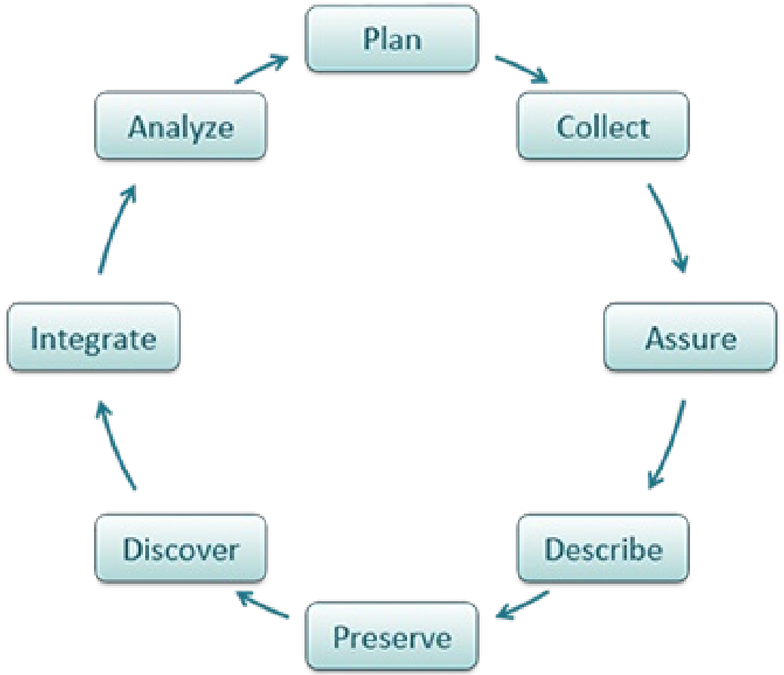
Learning Objectives

After completing this lesson, the participant will be able to:

- Understand a subset of typical analyses used
- Define a workflow
- Understand the concepts informal and formal workflows
- Discuss the benefits of workflows



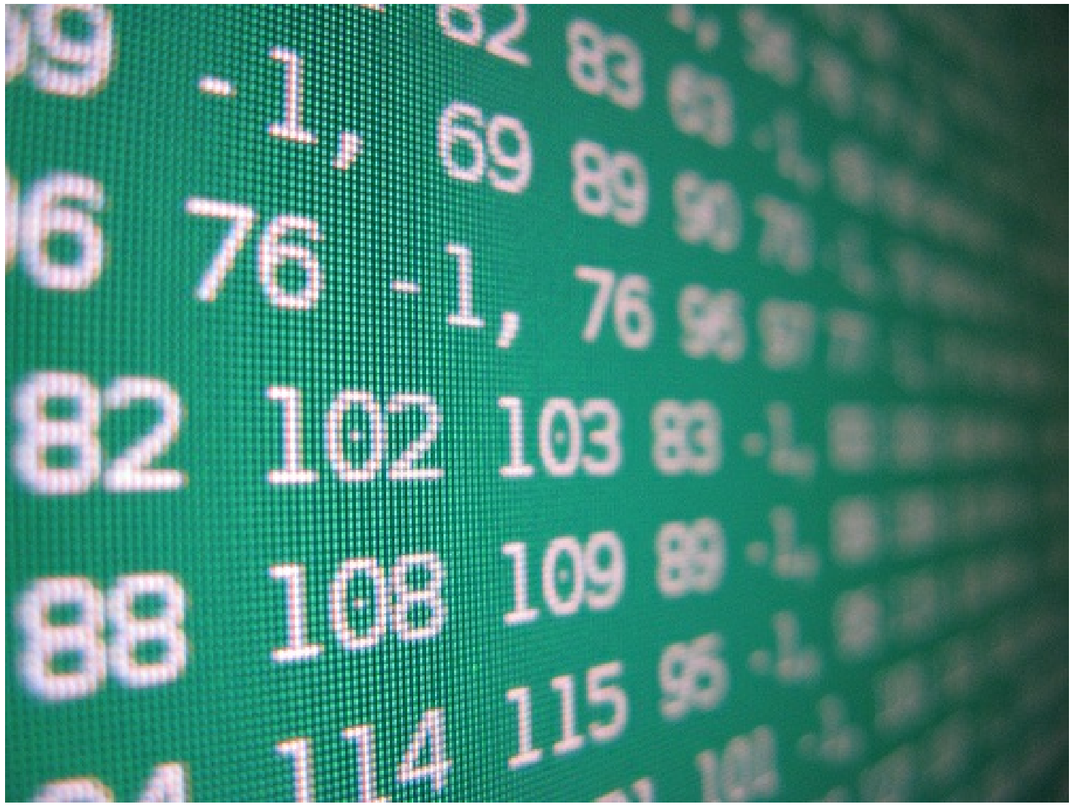
The Data Life Cycle



Data Analyses

Processes:

- Conducted via personal computer, grid, cloud computing
- Statistics, model runs, parameter estimations, graphs/plots, etc.



DataONE

Types of Analyses

Processing: subsetting, merging, manipulating

- Reduction: important for high-resolution datasets
- Transformation: unit conversions, linear and nonlinear algorithms

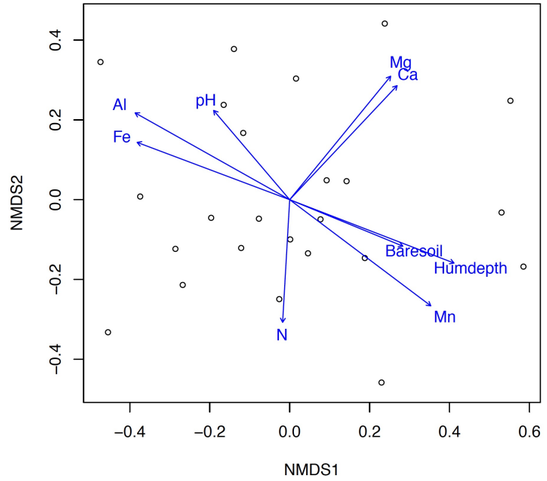
Types of Analyses

- Graphical analyses
  - Visual exploration of data: search for patterns
  - Quality assurance: outlier detection

Statistical Analyses

Conventional Statistics

- Experimental data
- Examples: ANOVA, MANOVA, linear and nonlinear regression
- Rely on assumptions: random sampling, random & normally distributed error, independent error terms, homogeneous variance



From Oksanen (2011) Multivariate Analysis of Ecological Communities in R: vegan tutorial

Descriptive Statistics

- Observational or descriptive data
- Examples: diversity indices, cluster analysis, quadrant variance, distance methods, principal component analysis, correspondence analysis

Types of Analyses

- Statistical analyses (continued)
  - Temporal analyses: time series
  - Spatial analyses: for spatial autocorrelation
  - Nonparametric approaches useful when conventional assumptions violated or underlying distribution unknown
  - Other mis. analyses: risk assessment, generalized linear models, mixed models, etc.
- Analyses of very large datasets
  - Data mining and discovery
  - Online data processing

After Data Analysis

- Re-analysis of outputs
- Final visualizations: charts, graphs, simulations, etc

Science is iterative: The process that results in the final product can be complex

Reproducibility

- Reproducibility at core of scientific method
- Complex process = more difficult to reproduce

- Good documentation required for reproducibility

  - Metadata: data about data
  - Process metadata: data about process used to create, manipulate, and analyze data

DataONE

Ensuring Reproducibility: Documenting the Process

- Process metadata: Information about process (analysis, data organization, graphing) used to get to data outputs
- Related concept: data provenance
  - Origins of data
  - Good provenance = able to follow data throughout entire life cycle
  - Allows for
  - Replication & reproducibility
  - Analysis for potential defects, errors in logic, statistical errors
  - Evaluation of hypotheses

Workflows: The Basics

- Formalization of process metadata
- Precise description of scientific procedure
- Conceptualized series of data ingestion, transformation, and analytical steps
- Three components
  - Inputs: information or material required
  - Outputs: information or material produced & potentially used as input in other steps
  - Transformation rules/algorithms (e.g. analyses)
- Two types:
  - Informal
  - Formal/Executable

Informal Workflows

- **Inputs or outputs** include data, metadata, or visualizations
- **Analytical processes** include operations that change or manipulate data in some way
- **Decisions** specify conditions that determine the next step in the process
- **Predefined processes** or subroutines specify a fixed multi-step process

Informal Workflows

- **Workflow diagrams:** Simple linear flow chart
  - Conceptualizing analysis as a sequence of steps
  - arrows indicate flow

Informal Workflows

- **Flow Charts:** simplest form of workflow

Informal Workflows

- **Flow charts:** simplest form of workflow
  - Transformation Rules

??? These steps are known in workflows as “transformation rules”. Transformation rules describe what is done to/with the data to obtain the relevant outputs for publication.

Informal Workflows

- **Flow charts:** simplest form of workflow
    - Inputs and Outputs

??? Now we focus on the actual data. The Inputs & outputs of this workflow are shown here in red. The first inputs are the raw temperature & salinity data. These are imported into R.The output of this process is the data in R format. That data in R format then become the input for the quality control and data cleaning step. The output of this step is “clean” temperature and salinity data, which is then the input for the analysis step. The output of the analysis step is the summary statistics, such as mean and standard deviation by month. These are subsequently the inputs for the visualization step.

Informal Workflows

- **Workflow diagrams:** adding decision points

Informal Workflows

- **Workflow diagrams:** a simple example

Informal Workflows

- **Workflow diagrams:** a complex example

Informal Workflows

Commented scripts: best practices

- Well-documented code is easier to review, share, enables repeated analysis
- Add high-level information at the top
  - Project description, author, date
  - Script dependencies, inputs, and outputs
  - Describes parameters and their origins
- Notice and organize sections
  - What happens in the section and why
  - Describe dependencies, inputs, and outputs
- Construct “end-to-end” script if possible
  - A complete narrative
  - Runs without intervention from start to finish

Formal/Executable Workflows

- Analytical pipeline
- Each step can be implemented in different software systems
- Each step & its parameters/requirements formally recorded
- Allows reuse of both individual steps and overall workflow

Formal/Executable Workflows

Benefits

- Single access point for multiple analyses across software packages
- Keeps track of analysis and provenance: enables reproducibility
  - Each step & its parameters/requirements formally recorded
- Workflow can be stored
- Allows sharing and reuse of individual steps or overall workflow
  - Automate repetitive tasks
  - Use across different disciplines and groups
  - Can run analyses more quickly since not starting from scratch

Formal/Executable Workflows

Example: Kepler Software

- Open-source, free, cross-platform
- Drag-and-drop interface for workflow construction
- Steps (analyses, manipulations etc) in workflow represented by “actor”
- Actors connect from a workflow
- Possible applications
  - Theoretical models or observational analyses
  - Hierarchical modeling
  - Can have nested workflows
  - Can access data from web-based sources (e.g. databases)
- Downloads and more information at kepler-project.org

Formal/Executable Workflows

Example: Kepler Software

Formal/Executable Workflows

Example: Kepler Software

Formal/Executable Workflows

Example: Kepler Software

Formal/Executable Workflows

Example: VisTrails

- Open source
- Workflow and provenance management support
- Geared toward exploratory computational tasks
  - Can manage evolving SWF
  - Maintains detailed history about steps and data
- www.vistrails.org

Workflows in General

- Science is becoming more computationally intensive
- Sharing workflows benefits science
  - Scientific workflow systems make documenting workflows easier
- Minimally: document your analysis via informal workflows
- Emerging workflow applications (formal/executable workflows) will
  - Link software for executable end-to-end analysis
  - Provide detailed info about data & analysis
  - Facilitate re-use & refinement of complex, multi-step analyses
  - Enable efficient swapping of alternative models & algorithms
  - Help automate tedious tasks

Best Practices for Data Analysis

- Scientists should document workflows used to create results
  - Data provenance
  - Analyses and parameters used
  - Connections between analyses via inputs and outputs
- Documentation can be informal (e.g. flowcharts, commented scripts) or formal (e.g. Kepler, VisTrails)

Summary

- Modern science is computer-intensive
  - Heterogeneous data, analyses, software
- Reproducibility is important
- Workflows = process metadata
- Use of informal or formal workflows for documenting process metadata ensures reproducibility, repeatability, validation

Resources for Data Analysis & Workflows

1. W. Michener and J. Brunt, Eds. Ecological Data: Design, Management and Processing. (Blackwell, New York, 2000).

The full slide deck may be downloaded from: http://www.dataone.org/education-modules

Suggested citation: DataONE Education Module: Analysis and Workflows. DataONE. Retrieved October 26 2016. From http://www.dataone.org/sites/all/documents/L9_Analysis Workflows.pptx