Why Should I Trust Your Data?

Barbara Lerner
Mount Holyoke College
February 14, 2019

Overview

- What is provenance?
- What is scientific data provenance?
- Provenance and R
  - RDataTracker
  - provSummarizeR
  - provDebugR
- Hands-on exercise

Skull of Rene Descartes

or is it?

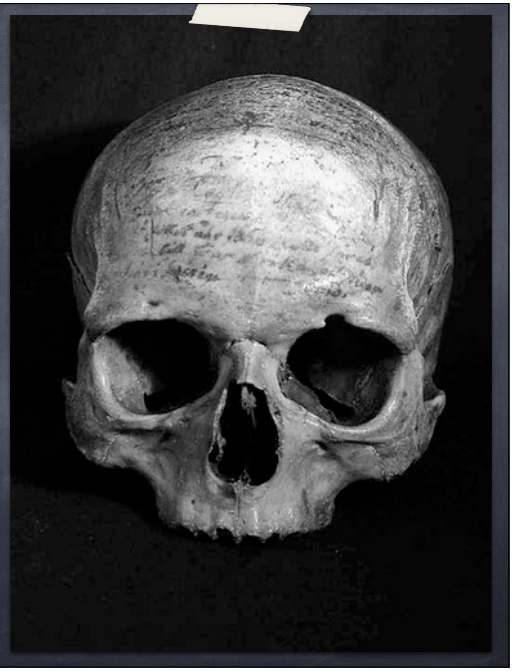
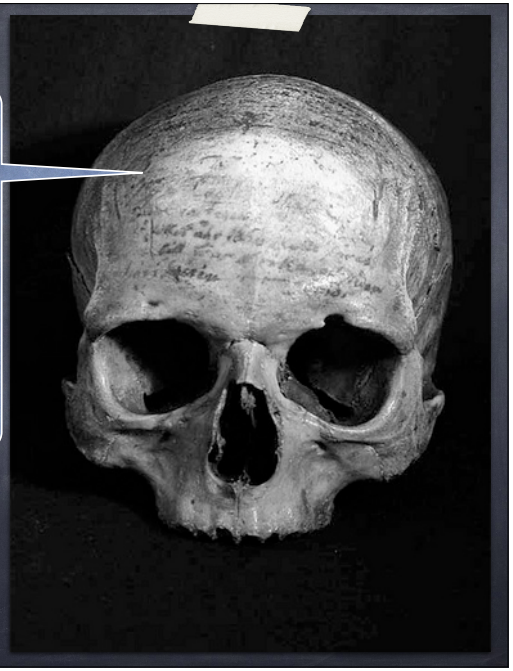


Photo from Russell Shorto, Descartes' Bones, Vintage Books, 2008.

This small skull once belonged to the great Cartesius,
The rest of his remains are hidden far away in the land of France,
But all around the circle of the globe his genius is praised,
And his spirit still rejoices in the sphere of heaven.

The skull of Descartes, taken by J. Fr. Planström, the year 1666, at the time when the body was being returned to France.



Provenance

- PROVENANCE: The history of ownership of a particular item. It allows the buyer to secure additional insight as to the origin of the item. From www.pbs.org

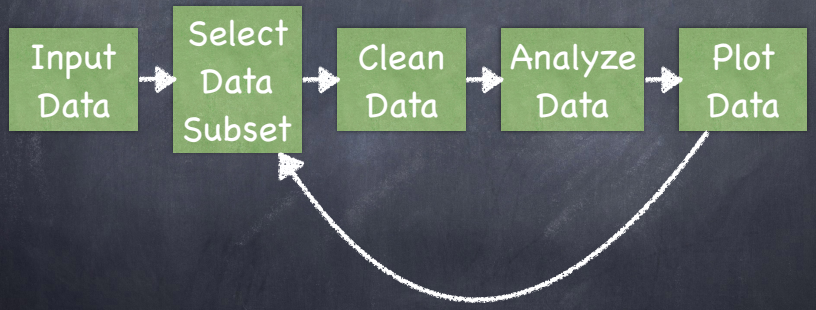
ClimateGate

Data Provenance

- DATA PROVENANCE: Information about the source and history of particular data items or sets, which is generally necessary to ensure their integrity, currency and reliability. From www.nature.com.

Data Analysis Workflow

Data Analysis Workflow



Data Provenance

- Who collected the data?
- Where was the data collected?
- What instruments were used?
- What was the weather like?

Basic Metadata

- How has the data been manipulated?

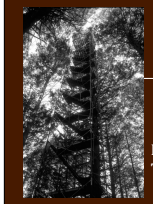
Scientific Processing Metadata

Scientific Processing Metadata

- What software / Web services were used to compute the data?
- Which version of my script did I use?
- What was the input to the software?
- Where did that input come from, etc.?
- How are errors in sensor readings or computations identified and handled?

Harvard Forest Hydrological Data Network

Meteorological Station



Eddy Flux Tower

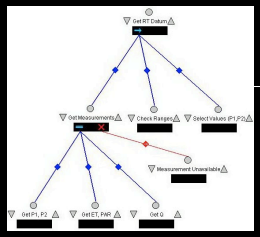
Wireless Network

Stream Sensors

Adapted from Boose et. al., 2007. Ensuring reliable datasets for environmental models and forecasts.

Data Processing:

Near-real-time quality control, modeling, and process metadata capturing



Internet

Consumers

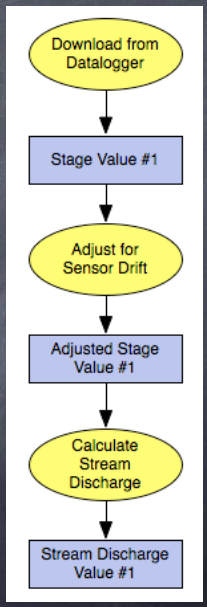
Image: Larisa Proulx

Calculating Stream Discharge

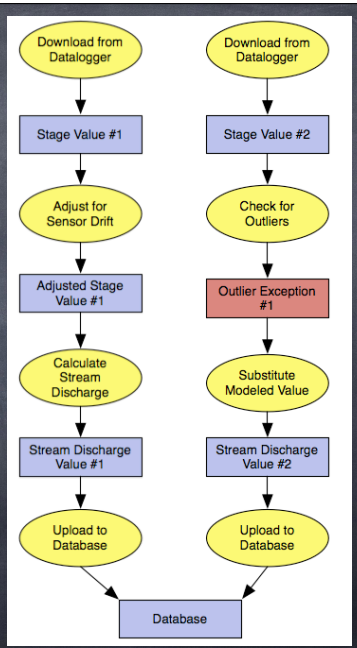
1. Download 15-minute stage values from datalogger at stream gauge.
2. Check for outliers and discard them.
3. Adjust values to account for sensor drift.
4. Calculate stream discharge values.
5. Upload stream discharge values to a public database.

```
Date,ID,Depth,NumWidth,I/Pflow,M2inTransect/quadrat,Bldr/bedrock%,Boulder%,Cobble%,Gravel%,Sand%,Silt%,Leafpack%,WoodyDebris%,Clay/
hardpan,Artificial,UnscouredForestFloor%,ScouredForestFloor%,EmbeddednessOfSubstrate,Pools%,Cascade%,Run%,Riffle%,Moss%,Roots%,DebrisPi
les%,Damp/drySand%,Damp/
dryRocks%,RiparianZoneL,RiparianZoneR,BankQual,LandUsesL,LandUsesR,AdjacentHabitatL,AdjacentHabitatR,Sinuosity,Gradient,FlowChar,Distan
ceFromPond,Elevation
2002-10-08,BPA253,11,2.6,1,1,0,0,0,0,13,0,87,0,0,0,0,f,2,70,0,0,0,0,5,25,0,3,3,1,0,0,2,2,0,1,6,2,324
2002-10-07,BPA356,8,2.6,1,1,0,0,0,92,0,0,8,0,0,0,0,0,2,100,0,0,0,0,0,0,0,0,3,3,1,0,0,2,2,0,1,6,3,327
```

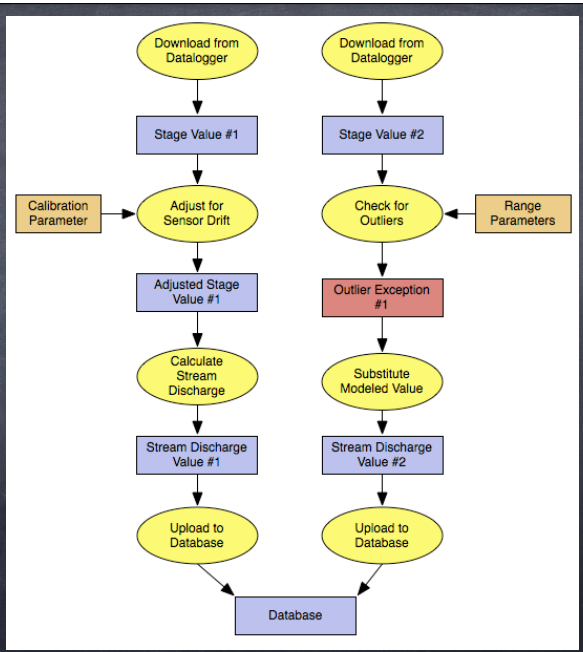
Provenance for Calculating Stream Discharge



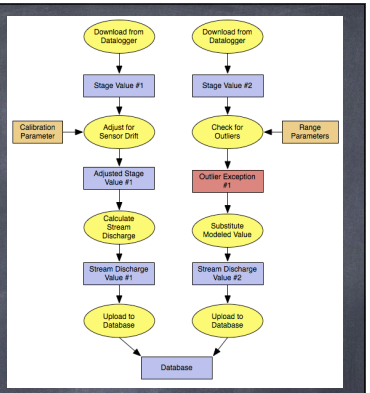
Provenance to Account for Errors



Provenance with Configuration Information

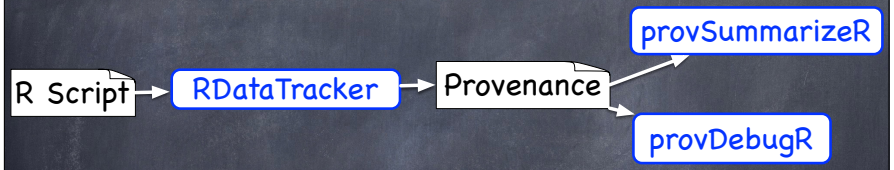


What does the user want to know???

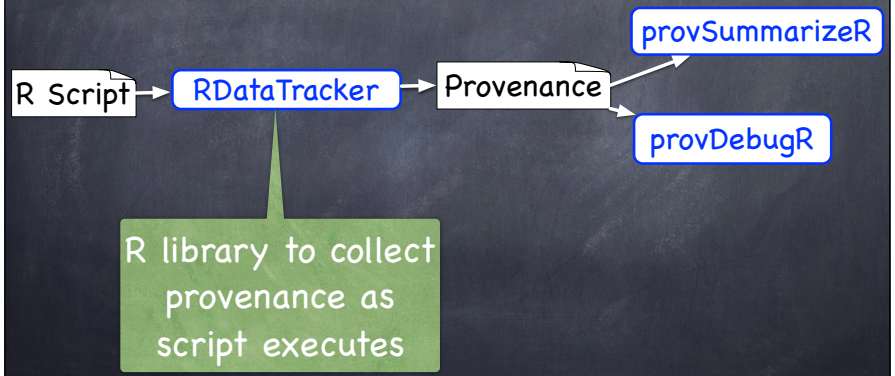


- What data / process / computation led to this output?
- Why did 2 experiments produce different results? Did the data, the process, or the computations change?
- If the data / process / computation changes, what part of the process should we re-run?

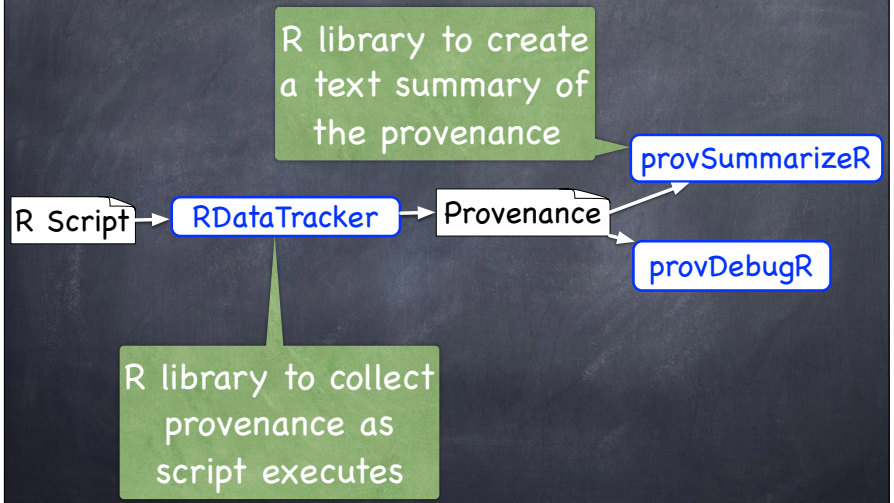
Provenance and R Scripts



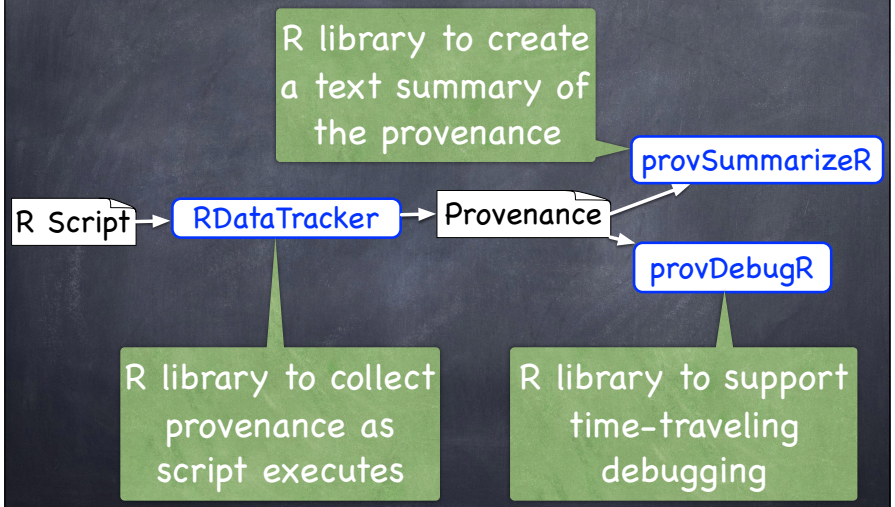
Provenance and R Scripts



Provenance and R Scripts



Provenance and R Scripts



Uses of Provenance

- Script debugging
- Experiment validation
  - Was an experiment run correctly?
  - What settings did the instruments have?
  - Was a piece of software / web service used correctly?
- Scientific analysis
  - Are some intermediate values strong indicators of a particular final result?
- Publication
  - Publish the process, not just the computed data

Challenges

- Not drowning in data
- Identifying the questions the user would like to be able to answer
- Presenting answers to questions in an understandable way

Take Away Message

- How do you know what you know?
  - Where are the potential errors introduced?
    - Instrument problems
    - Bad sampling techniques
    - Inappropriate statistical analysis
    - Errors in programming

Take Away Message

- How do you know what you know?
  - Where are the potential errors introduced?
    - Instrument problems
    - Bad sampling techniques
    - Inappropriate statistical analysis
    - Errors in programming

Having provenance doesn't make these problems go a way, but it can help you understand how you got there!

Related work

- Capturing provenance data
  - Kepler, Taverna, noWorkflow, rctrack, recordr, CXXR, and many more
- Data management
  - Chimera, Burrito
- Querying and visualization
  - Vistrails, Provenance Explorer, Zoom, ProvDB

Acknowledgements

Work done jointly with
- Emery Boose and Aaron Ellison, Harvard Forest
- Margo Seltzer, University of British Columbia
- Elizabeth Fong, Mount Holyoke College
- Joe Wonsil, Carthage College
- Orenna Brand, Columbia University
- Many REU students

Funded by
- NSF
- Mount Holyoke Center for the Environment

Let's try it!