

Using Kepler Scientific Workflows to automate the modeling of Thresholds of Potential Concern (TPCs)



Mark Schildhauer

National Center for Ecological Analysis and Synthesis
University of California, Santa Barbara

<http://www.nceas.ucsb.edu/ecoinformatics>



Presentation at 4th KNP Science
Networking Meeting, Skukuza, KNP



March 14, 2006



Thresholds of Potential Concern at Kruger National Park

- Kruger TPC's are a set of ecological analyses--
 - based on long-term monitoring data
 - quantifying variability in ecologically relevant factors
 - for determining whether pre-defined conditions have been exceeded
 - so that management decisions can be made, and their empirical outcomes carefully documented

Exceedance of a TPC indicates an ecological condition within Kruger that is of serious concern.



How to implement TPCs?

KNP researchers would like to be able to easily run, review, and share the results of their TPC analyses

- Documents exist that describe the TPC criteria in detail
- Data are available to calculate the TPC's

however...

- Many TPC's have not yet been calculated
- For those that have been calculated, many cannot be re-calculated without extensive additional effort
 - Effort to collate and integrate updated information
 - Analyses re-derived each time (not saved as executable code)
- No easy way to review, revise and share TPC analyses



Advantages of Scientific Workflow approach

- Scientific Workflows are a new way of accomplishing ecological analysis, and may help address the challenges of running, reviewing, and sharing Kruger's TPC's
 - visually depict how the TPC works, as well as clarify how execution takes place
 - provide the ability to rapidly review and revise an analysis
 - provide direct access to the relevant data, via links to local or network storage
 - enable efficient execution and sharing of results, even for those with minimal quantitative skills



How to implement TPCs?

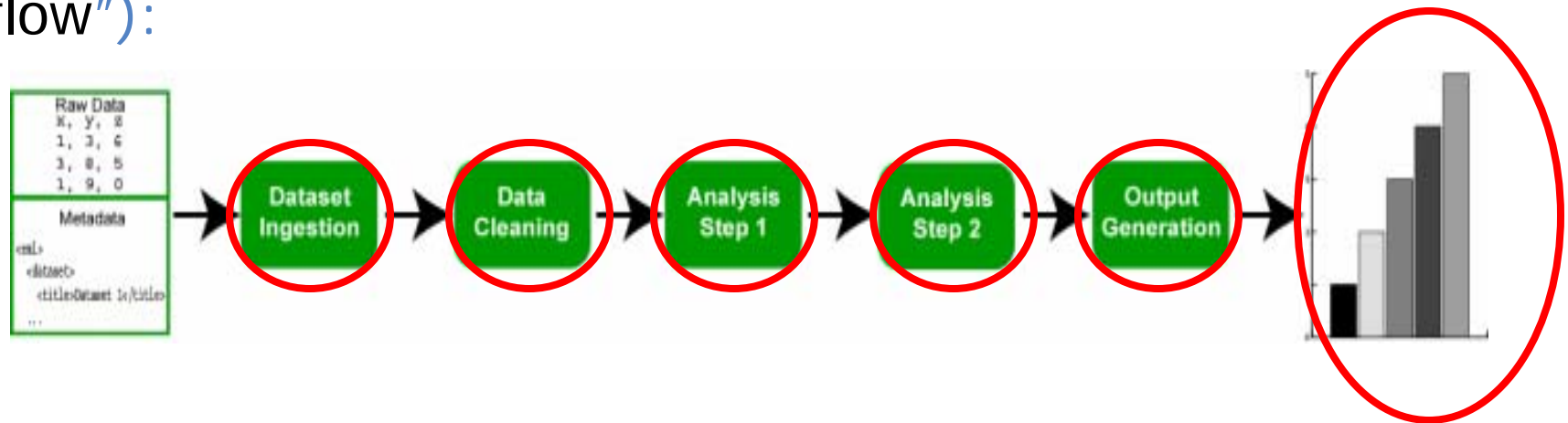
Judith Kruger & NCEAS staff are:

- prototyping the Buffalo TPC
- using a Scientific Workflow tool
- to investigate how it can assist Kruger researchers
- in developing, running, and sharing TPC analyses



Scientific Workflow approach

Think of ecological analysis and modeling as a sequence of “steps”– involving boxes (indicating **data** and **analytical processes**), which are joined by arrows (which indicate “flow”):

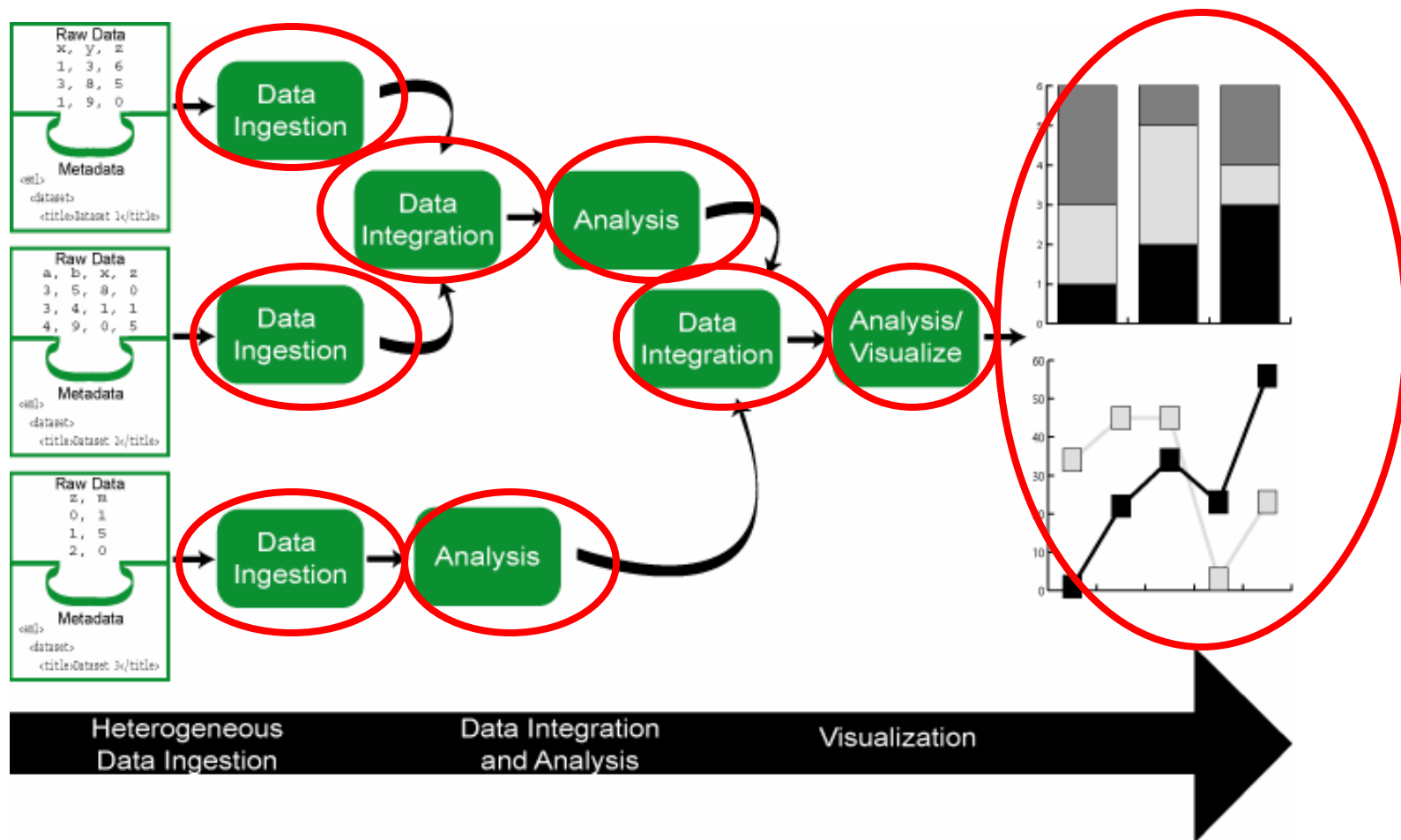


Resembles traditional “flow chart” approach to documenting analyses

But modern Scientific Workflow applications are very different, because you can *execute* these workflows

Scientific Workflow approach

Complex analyses and models can be constructed and executed using scientific workflow tools:





Scientific Workflows and TPCs

Kepler is a Scientific Workflow tool that we are developing as part of the “Science Environment for Ecological Knowledge” (SEEK) research project.

- Kepler is—
 - generically useful for scientific analysis and modeling (but our focus is specifically on ecological analysis and modeling)
 - developed through multi-institutional, multidisciplinary collaboration (including ecology, genomics, engineering, geology, and others)
 - freely available; runs on Windows, Mac, Linux





Data & Analyses in Kepler

Data—

- Kepler can be used to access data from the local hard drive, or over the network
- Kepler has close ties with the KNB data repositories described earlier, to provide powerful data discovery and access, based on metadata

Analyses—

Kepler has a rich set of built-in libraries for modeling, statistical analysis, etc.

Kepler workflows can include code written in “R”, MATLAB and other scientific programming languages, which can be executed from within the program



Kruger Buffalo TPC Case Study

A complex set of rules, specifying how to determine a “Wet Cycle”, based on measurements of monthly rainfall from a variety of stations around Kruger National Park

If, during a “Wet Cycle”, the zonal population growth of Buffaloes is below 5% for three consecutive years, and the total Buffalo population is $\leq 30,000$ animals, then the Buffalo TPC has been exceeded

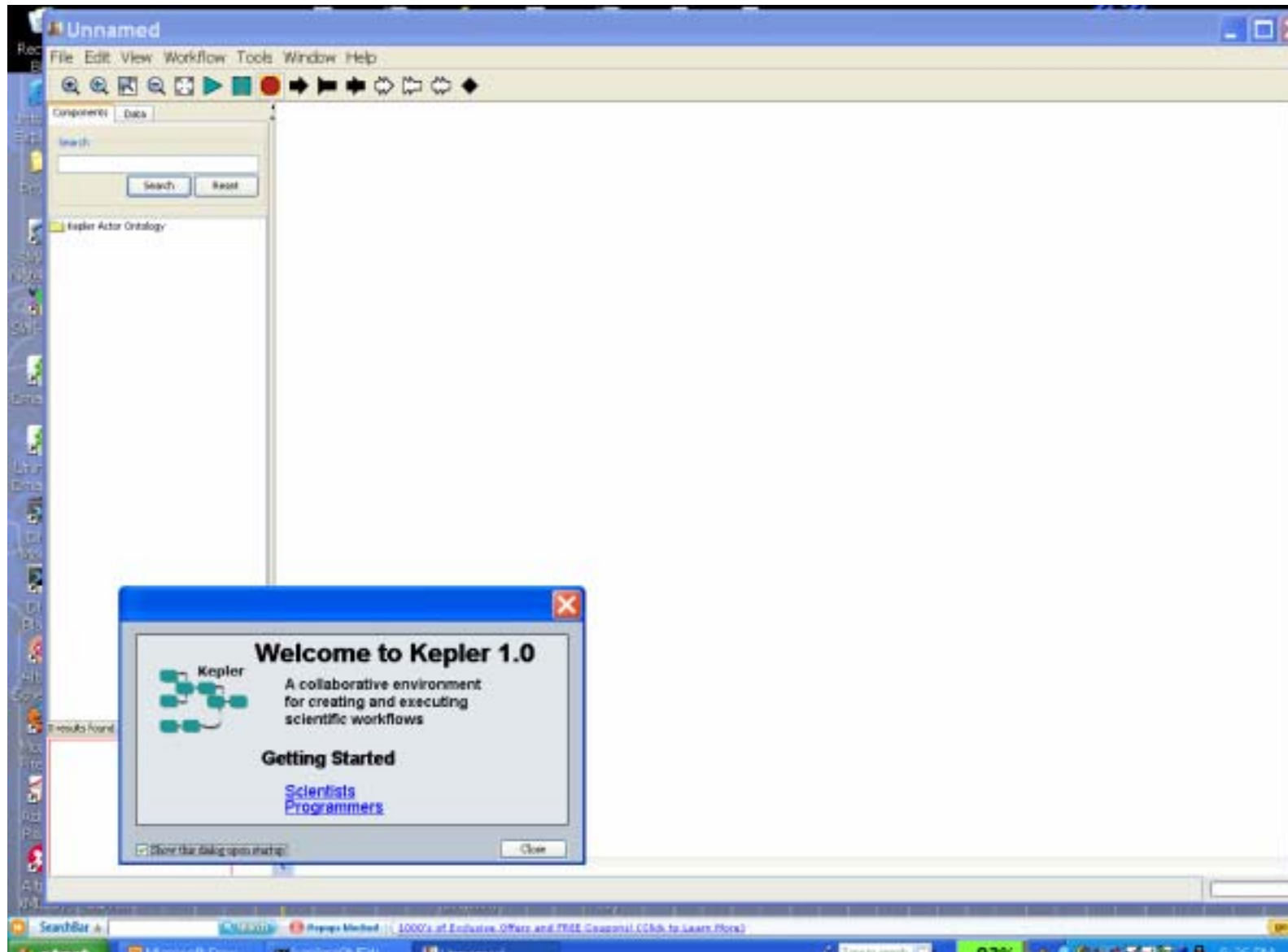
Task:

Use Kepler to develop data ingestion, analyses, and visualizations for the Buffalo TPC

Leverage analysis and graphics capabilities of “R”



Constructing the Buffalo TPC Workflow





Locate Data and Analysis functions

The screenshot shows the SDF Director software interface. A search bar at the top left contains the word "constant". Below it, a search results tree lists various components, with "String Constant" and "String Function" circled in red. A blue callout box points to the search bar with the text "Search for data and analyses here". Another blue callout box points to the search results tree with the text "Choices show up here". A third blue callout box points to a specific "Constant" entry in the results, which contains a file path: "C:/KrugerTPC/R-Stats/ExtractedData/Rain...". A blue callout box points to this entry with the text "point to your data". A fourth blue callout box points to the "String Constant" entry in the tree with the text "Drag choice across".

Search for data and analyses here

Choices show up here

"point" to your data

Drag choice across



Link up analysis to data

The screenshot shows a workflow editor window titled "file:/C:/KrugerTPC/R-Stats/rainfall1.xml". The interface includes a menu bar (File, Edit, View, Workflow, Tools, Window, Help), a toolbar with various icons, and a left-hand pane for components. The main workspace displays a workflow with two steps: a "Constant" step and an "R" step labeled "Rainfall-Wet cycle calculator". A red circle highlights the toolbar icons, with a blue callout box stating "Use arrows to link data to analysis". Another red circle highlights the "Constant" step, with a blue callout box stating "A simple two-step workflow linking a data set with a tool that calculates average rainfall". A third red circle highlights a text box at the bottom, with a blue callout box stating "It is very easy to add explanatory text to your workflow".

file:/C:/KrugerTPC/R-Stats/rainfall1.xml

File Edit View Workflow Tools Window Help

Components Data

SDF Director

Use arrows to link data to analysis

Constant
"C:/KrugerTPC/R-Stats/ExtractedData/Rain..."

Rainfall-Wet cycle calculator

R

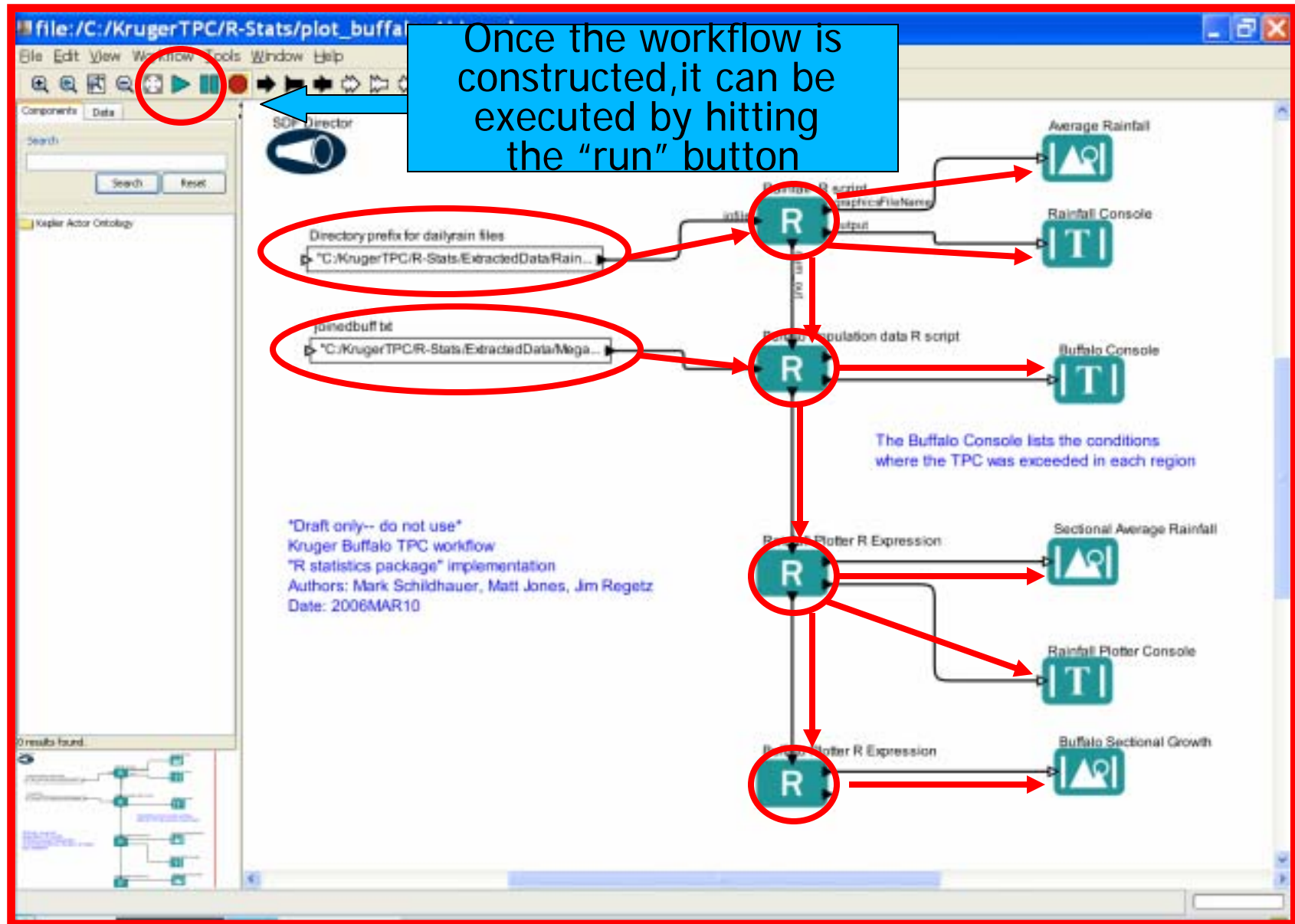
A simple two-step workflow linking a data set with a tool that calculates average rainfall

"Draft only-- do not use"
Kruger Buffalo TPC workflow
"R statistics package" implementation
Authors: Mark Schildhauer, Matt Jones, Jim Regetz
Date: 2006MAR10

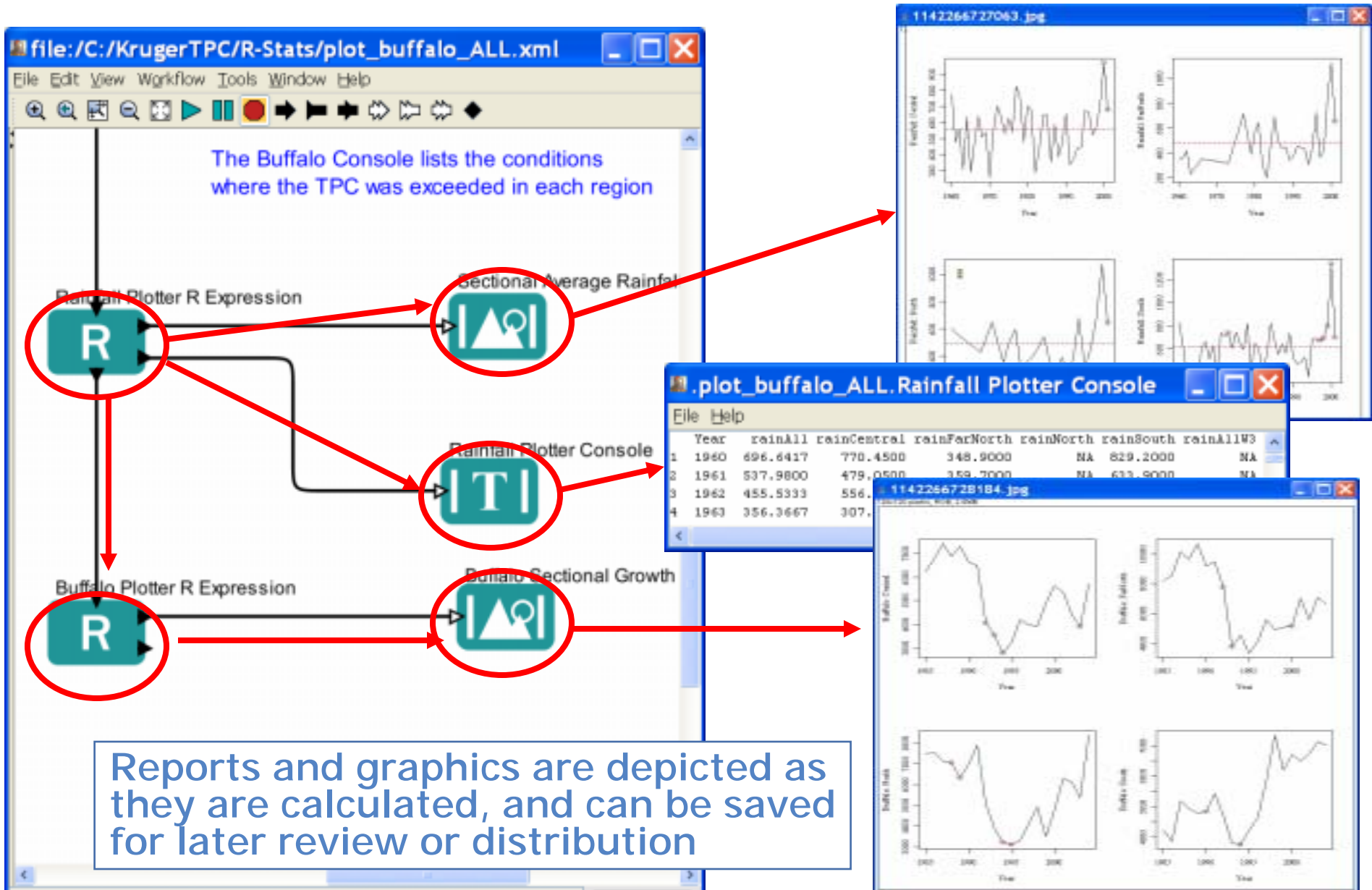
It is very easy to add explanatory text to your workflow



Prototype Buffalo TPC Workflow



(Close-up of) Final Part of TPC



The Buffalo Console lists the conditions where the TPC was exceeded in each region

Sectional Average Rainfall

Rainfall Plotter Console

Buffalo Plotter R Expression

Binary Sectional Growth

1142266727063.jpg

1142266728184.jpg

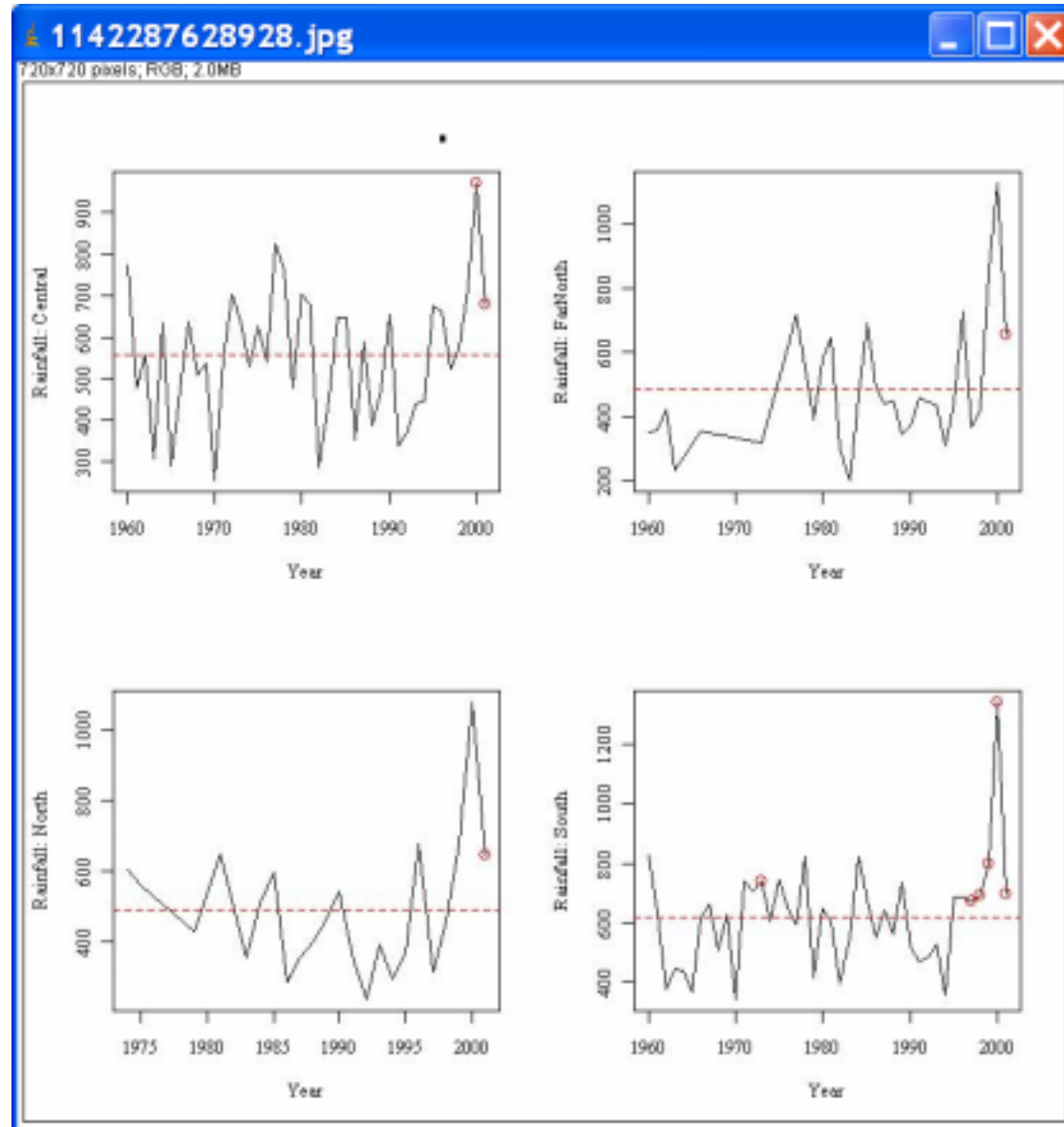
Year rainAll rainCentral rainFarNorth rainNorth rainSouth rainAllW3

Year	rainAll	rainCentral	rainFarNorth	rainNorth	rainSouth	rainAllW3
1960	696.6417	770.4500	348.9000	NA	829.2000	NA
1961	537.9800	479.0500	359.7000	NA	633.9000	NA
1962	455.5333	556.0000				
1963	356.3667	307.0000				

Reports and graphics are depicted as they are calculated, and can be saved for later review or distribution

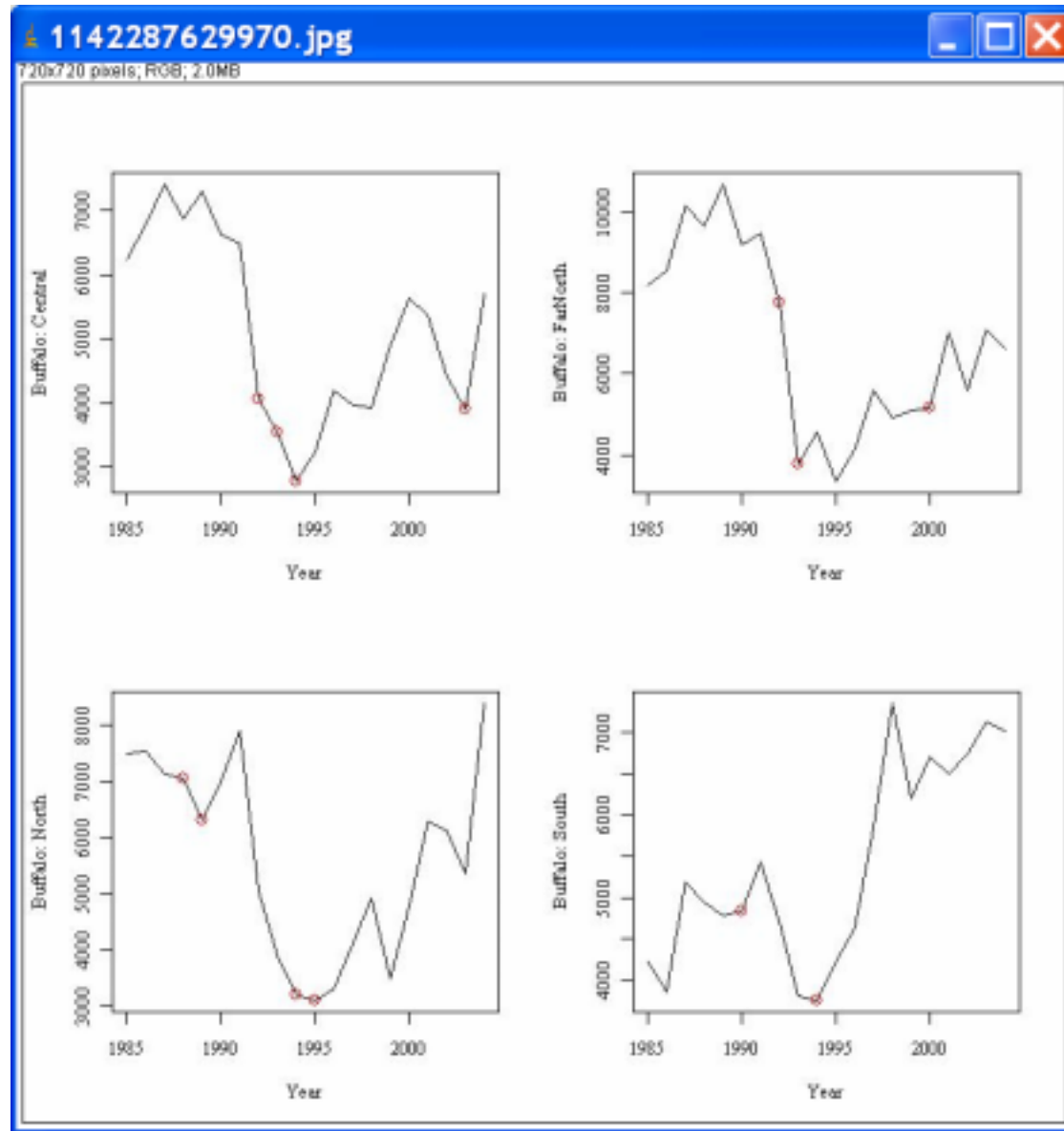


Sectional Rainfall Averages





Buffalo Sectional Population Growth





Buffalo Growth TPC Output Data Set

```
.plot_buffalo_ALL.Buffalo Console
File Help
> TPC_South <- subset(dataTPC2, D3South==FALSE & rainSouthW3==TRUE)
> TPC_South
  Year  rainAll  rainCentral  rainFarNorth  rainNorth  rainSouth  rainAllW3
38 1997  483.3923    521.0033    366.7399    312.7450    674.0054    FALSE
39 1998  562.8390    574.0500    418.6000    456.4750    692.0317    FALSE
40 1999  773.6544    717.2683    850.4295    676.9667    800.3486    FALSE
41 2000 1135.2477    968.1333    1126.2750  1077.8571  1340.7571    FALSE
42 2001  668.0937    678.5333    653.0938    646.3750    698.2000     TRUE
  rainCentralW3  rainFarNorthW3  rainNorthW3  rainSouthW3  buffAll  Central
38          FALSE          FALSE          FALSE          TRUE    19477    3958
39          FALSE          FALSE          FALSE          TRUE    21095    3913
40          FALSE          FALSE          FALSE          TRUE    19644    4880
41           TRUE          FALSE          FALSE          TRUE    22264    5631
42           TRUE           TRUE          TRUE          TRUE    25155    5364
  FarNorth  North  South  buffSub30k  GrAll  D3All  GrCentral  D3Central
38    5577  4084  5842      TRUE 16.768585 FALSE  -5.491882    FALSE
39    4912  4909  7351      TRUE  8.307234 FALSE  -1.136938    FALSE
40    5086  3478  6200      TRUE -6.878407 FALSE  24.712497    FALSE
```



Summary

- Kepler can be used effectively to calculate TPC's
 - Buffalo TPC is an end-to-end test case
 - starts with data ingestion
 - progresses thru multiple processing steps, including data integration
 - produces scientific visualizations, and relevant output data sets
 - Buffalo TPC workflow can easily be revised & extended
 - This test case produced generically useful Kruger code (rainfall trending and averaging routines and graphs)
 - Anyone with Kepler and "R" can now compute the Buffalo TPC



Future Directions

Kruger and NCEAS hope to continue developing Kruger TPC's using the Scientific Workflow approach

- directly link the TPC workflows with Kruger data sets that are stored on the network
- produce a set of generically useful scientific visualizations and algorithms that also operate across Kruger TPC's

These Scientific Workflows will provide a long-lasting and flexible way for Kruger researchers to accurately and efficiently report on the status of "Thresholds of Potential Concern" within Kruger National Park



Acknowledgements

This material is based upon work supported by:

The National Science Foundation under Grant Numbers 9980154, 9904777, 0131178, 9905838, 0129792, and 0225676.

The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.

The Andrew W. Mellon Foundation.

Primary Collaborators: NCEAS, University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research)

Kepler contributors: SEEK, Ptolemy II, SDM/SciDAC, GEON

Special thanks to Jim Regetz, Dan Higgins, Josh Madin, and Matt Jones for their assistance in developing the example workflows.