



USGS-R: A community to support and expand R scientific computing capacity

E. Read, A. Appling, L. Carr, L. DeCicco, J. Read, J. Walker, and L. Winslow
U.S. Geological Survey, Office of Water Information

Background

“Waste water effluent loads estimated for two streams”

“National flood forecasting network now active”

“New regional lake study published”

“Streamflow reanalysis dataset released”

← Data/information/tool divide →

Local to regional-scale
descriptive science

Broad-scale
prediction

Challenges

- Collaborative interdisciplinary work
 - Exchange between hydrologists and limnologists
 - Engage software developers as equals
- Science in the era of web-enabled research
 - Web collaborations
 - Web tools
 - Data web services
- Shared tools improve efficiency
 - Open data -> open tools
 - Reusable solutions

Research workflows

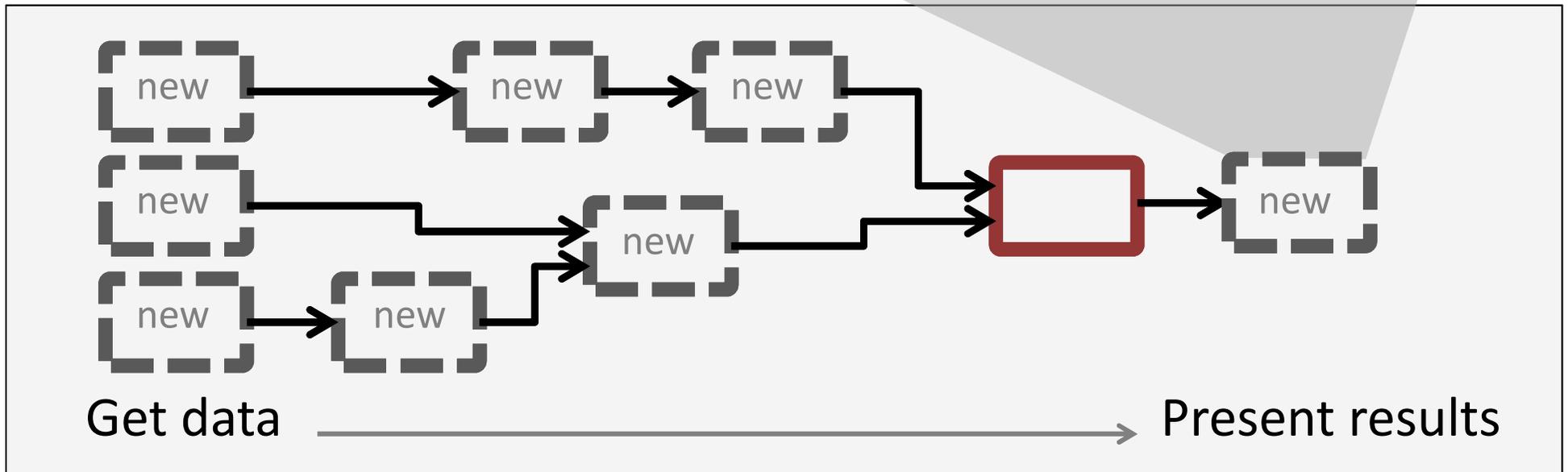
- The Status Quo:
 - Confusing code
 - Poor software development practices
 - Not repeatable
 - “One-offs”

```
source('C:/users/data/data_new/scripts_4_data/mungeDataClimate_FINAL_v4.R')
answer <- data_fix(data, arguments = c('do_math', 4, 45*2.3))

# *** THIS USED TO WORK!!!! NOW IT DOESN'T HELP
# enough <- frac>=maxMissing
# dataToday <- any(dailyDis$Year==2014)

#If there's not enough data between 1980-today, table mean and current discharge values will
meanDis <- mean(dailyDis[,dataColI])
itodayDis <- dailyDis$Year== (as.POSIXlt(endDate)$year+1900)
todayDis <- dailyDis[itodayDis,]
todayDis <-todayDis[,dataColI]
source('dis_points.R')
points <- dis_points()

lg_lim <- c(0.003, 47500)
tcks <- c(1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4)
minor_tcks <- minor_ticks()
fig_w = '650'
fig_h = '550'
r_mar = '100'
ect_1 <- newXMLNode("rect",parent=g_id, attrs = c(id="box1", x=l_mar, y=t_mar,
width=main_dim, height=main_dim,
style="fill: rgb(100%,100%,100%);fill-opacity: 1; stroke: none;"))
inset_dim = '210'
main_dim = '500'
x_bump = 20 # pixel bump to shift map
y_crt = c(as.numeric(t_mar)+as.numeric(main_dim), as.numeric(t_mar))
x_crt = c(as.numeric(l_mar), as.numeric(l_mar)+as.numeric(main_dim))
```



Research workflows

- Analytical building blocks:
 - Tested/versioned
 - Documented
 - Modular
 - Dependable

dataRetrieval

Linux: build passing

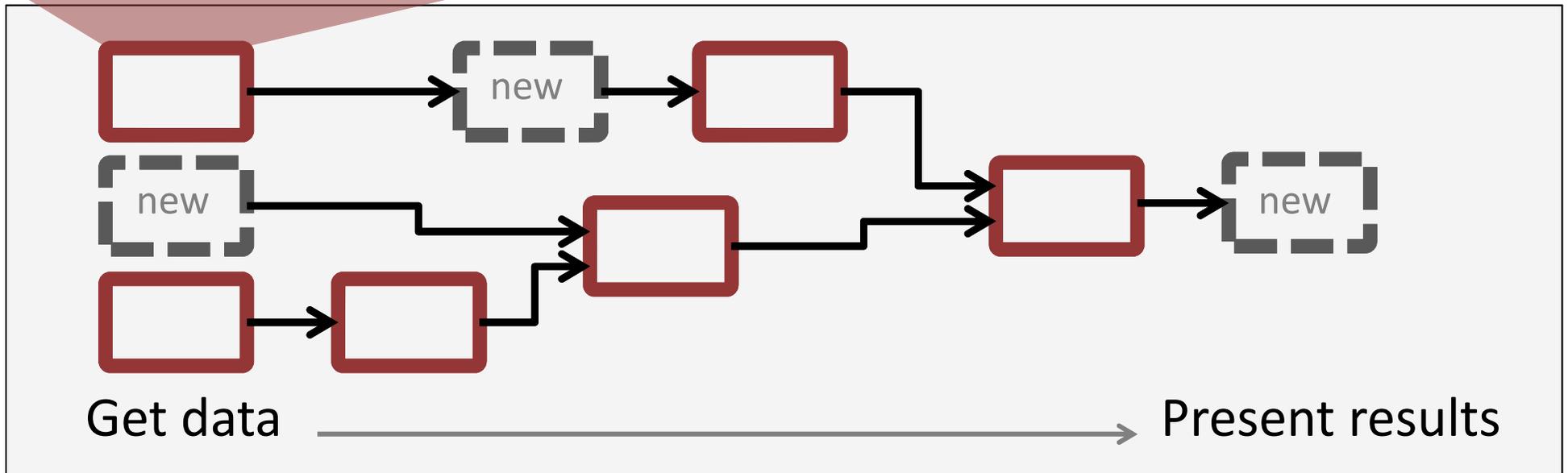
Windows: Build passing

Retrieval functions for USGS and EPA hydrologic and water quality data.
[dataRetrieval Issues page](#)

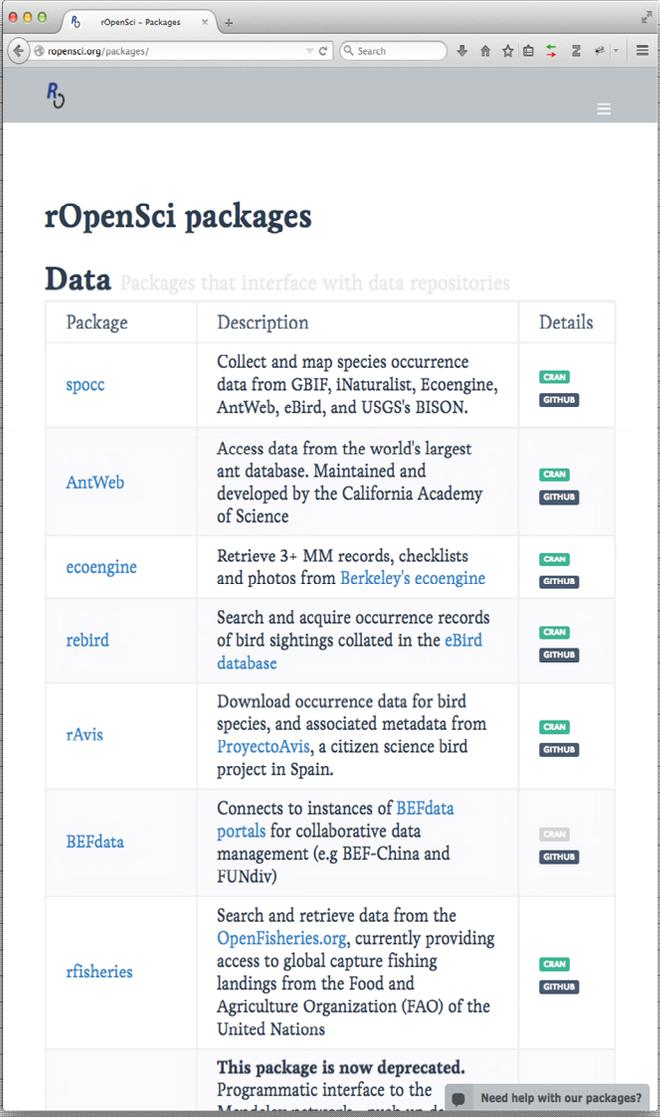
Function Overview

Web service retrieval functions:

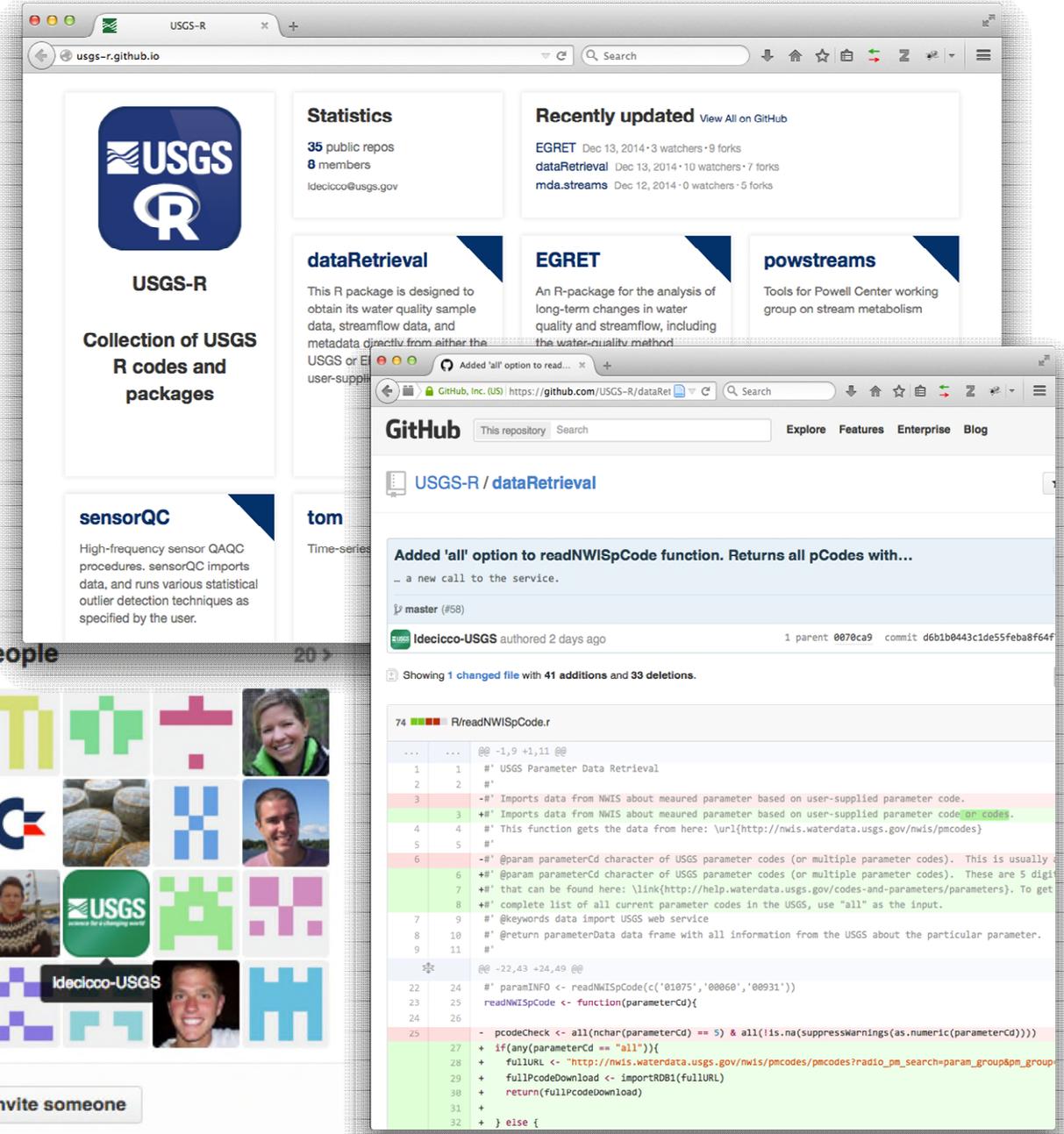
Function	Inputs	Description
readNWISdata	..., service	NWIS data using user-specified queries
readNWISdv	Common 3, parameterCd, statCd	NWIS daily data with Common query
readNWISqw	Common 3, parameterCd, expanded	NWIS water quality data with Common query



How do we make the leap?



Package	Description	Details
spocc	Collect and map species occurrence data from GBIF, iNaturalist, Ecoengine, AntWeb, eBird, and USGS's BISON.	 
AntWeb	Access data from the world's largest ant database. Maintained and developed by the California Academy of Science	 
ecoengine	Retrieve 3+ MM records, checklists and photos from Berkeley's ecoengine	 
rebird	Search and acquire occurrence records of bird sightings collated in the eBird database	 
rAvis	Download occurrence data for bird species, and associated metadata from ProyectoAvis, a citizen science bird project in Spain.	 
BEFdata	Connects to instances of BEFdata portals for collaborative data management (e.g BEF-China and FUNdiv)	 
rfisheries	Search and retrieve data from the OpenFisheries.org, currently providing access to global capture fishing landings from the Food and Agriculture Organization (FAO) of the United Nations	 
This package is now deprecated. Programmatic interface to the Modeller network, web and...		 



USGS-R
Collection of USGS R codes and packages

Statistics
35 public repos
8 members
ldecicco@usgs.gov

Recently updated View All on GitHub
EGRET Dec 13, 2014 · 3 watchers · 9 forks
dataRetrieval Dec 13, 2014 · 10 watchers · 7 forks
mda.streams Dec 12, 2014 · 0 watchers · 5 forks

dataRetrieval
This R package is designed to obtain its water quality sample data, streamflow data, and metadata directly from either the USGS or E...

EGRET
An R-package for the analysis of long-term changes in water quality and streamflow, including the water quality method...

powstreams
Tools for Powell Center working group on stream metabolism

sensorQC
High-frequency sensor QA/QC procedures. sensorQC imports data, and runs various statistical outlier detection techniques as specified by the user.

tom
Time-series

People 20 →

Invite someone

GitHub This repository Search Explore Features Enterprise Blog

USGS-R / dataRetrieval

Added 'all' option to readNWISpCode function. Returns all pCodes with...
... a new call to the service.

master (#58)

ldecicco-USGS authored 2 days ago 1 parent 0078ca9 commit d6b1b0443c1de55fba8f64f

Showing 1 changed file with 41 additions and 33 deletions.

```
74 R/readNWISpCode.R
... @@ -1,9 +1,11 @@
1 1 #' USGS Parameter Data Retrieval
2 2 #'
3 -#' Imports data from NWIS about measured parameter based on user-supplied parameter code.
4 +#' Imports data from NWIS about measured parameter based on user-supplied parameter code or codes.
5 4 #' This function gets the data from here: \url{http://nwis.waterdata.usgs.gov/nwis/pmcodes}
5 5 #'
6 -#' @param parameterCd character of USGS parameter codes (or multiple parameter codes). This is usually
7 +#' @param parameterCd character of USGS parameter codes (or multiple parameter codes). These are 5 digi
8 +#' that can be found here: \link{http://help.waterdata.usgs.gov/codes-and-parameters/parameters}. To get
9 +#' complete list of all current parameter codes in the USGS, use "all" as the input.
7 9 #' @keywords data import USGS web service
8 10 #' @return parameterData data frame with all information from the USGS about the particular parameter.
9 11 #'
@@ -22,43 +24,49 @@
22 24 #' paramINFO <- readNWISpCode(c('01875', '00060', '00931'))
23 25 readNWISpCode <- function(parameterCd){
24 26
25 - pcodeCheck <- all(nchar(parameterCd) == 5) & all(!is.na(suppressWarnings(as.numeric(parameterCd))))
27 + if(any(parameterCd == "all")){
28 + fullURL <- "http://nwis.waterdata.usgs.gov/nwis/pmcodes/pmcodes?radio_pm_search=param_group&pm_group=
29 + fullPcodeDownload <- importRDB1(fullURL)
30 + return(fullPcodeDownload)
31 + } else {
32 + }
```



How do we make the leap?

- Promote and teach best practices
- Catalog of open-source tools
- Develop a community of learners and teachers

rOpenSci packages

Data Packages that interface with data repositories

Package	Description	Details
spocc	Collect and map species occurrence data from GBIF, iNaturalist, Ecoengine, AntWeb, and USGS's BISON.	CRAN GitHub
AntWeb	Access data from the world's largest ant database. Maintained and developed by the California Academy of Science	CRAN GitHub
ecoengine	Retrieve 3+ MM records, checklists and photos from Ecoengine	CRAN GitHub
rebird	Search and retrieve data from the eBird database	CRAN GitHub
rAvis	Download occurrence data for bird species, and associated metadata from Project Avis, a global bird project	CRAN GitHub
BEFdata	Connects to instances of BEFdata portals for collaborative data management (e.g. BEF-Chicago and FUNdiv)	CRAN GitHub
rfisheries	Search and retrieve data from the OpenFisheries.org, currently providing access to global capture fishing landings from the Food and Agriculture Organization (FAO) of the United Nations	CRAN GitHub

This package is now deprecated. Programmatic interface to the Mendelian network, such as...

Need help with our packages?

USGS-R

Statistics: 35 public repos, 8 members, ldeicoco@usgs.gov

Recently updated: EGRET, dataRetrieval, mda.streams

powstreams: Tools for Powell Center working group on stream metabolism

Collection of USGS R codes and packages

USGS-R / dataRetrieval

Added all of the following functions: Returns all pCodes with...

ldeicoco-USGS authored 2 days ago

Showing 1 changed file with 41 additions and 33 deletions.

```
74 R/readNWISpCode.R
...
1 1 #' USGS Parameter Data Retrieval
2 2 #'
3 -#' Imports data from NWIS about measured parameter based on user-supplied parameter code.
4 3 +#' Imports data from NWIS about measured parameter based on user-supplied parameter code.
5 4 #' This function gets the data from here: \url{http://nwis.waterdata.usgs.gov/nwis/pmcodes}
6 5 #'
7 -#' @param parameterCd character of USGS parameter codes (or multiple parameter codes). This is usually
8 6 +#' @param parameterCd character of USGS parameter codes (or multiple parameter codes). These are 5 digit
9 7 +#' that can be found here: \link{http://help.waterdata.usgs.gov/codes-and-parameters/parameters}. To get
10 8 +#' complete list of all current parameter codes in the USGS, use "all" as the input.
11 9 #' @keywords data import USGS web service
12 10 #' @return parameterData data frame with all information from the USGS about the particular parameter.
13 11 #'
14 12
15 13
16 14
17 15
18 16
19 17
20 18
21 19
22 20
23 21
24 22
25 23 - pcodeCheck <- all(nchar(parameterCd) == 5) & all(!is.na(suppressWarnings(as.numeric(parameterCd))))
26 24 + if(any(parameterCd == "all")){
27 25 + fullURL <- "http://nwis.waterdata.usgs.gov/nwis/pmcodes/pmcodes?radio_pm_search=param_group&pn_group="
28 26 + fullPcodeDownload <- importRDB1(fullURL)
29 27 + return(fullPcodeDownload)
30 28 + } else {
31 29
32 30
```

People

Invite someone

ldeicoco-USGS



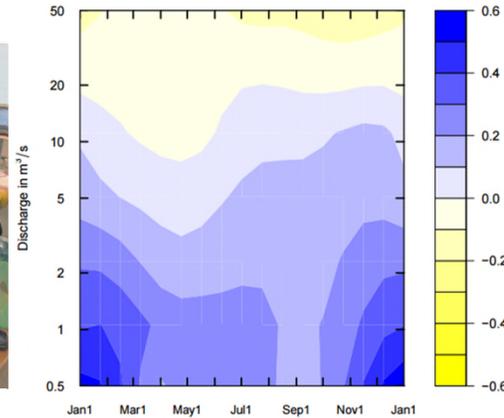
Promote and teach best practices

Outreach

Instruction

Advising

Collaboration



Promote and teach best practices

Instruction



Intro R Workshop Outline

22 January, 2016

- [Workshop Schedule](#)
- Day 1
 - [00 - Welcome (8:30 am - 9:00 am)]
 - 01 - Introduction (9:00 am - 10:00 am)
 - 02 - Get (10:15 am - 12:00 pm)
 - 03 - Clean (1:00 pm - 2:45 pm)
 - 04 - Explore (3:00 pm - 4:30 pm)
- Day 2
 - 05 - Analyze: Base (8:30 am - 9:30 am)
 - 06 - Analyze: Using Packages (9:30 am - 10:45 am)
 - 07 - Visualize: Base (11:00 pm - 12:00 pm)
 - 08 - Visualize: ggplot2 (1:00 pm - 2:30 pm)
 - 09 - Repeat (2:45 pm - 4:30 pm)
- Day 3
 - [10 - Practice (8:30 am - 10:30 am)]

This workshop is designed to provide training and tools to perform common data analysis workflow steps:
get -> clean -> explore -> analyze -> visualize -> repeat

The workshop and materials are hands-on and include examples and exercises. After completing the workshop, you won't be an R expert, but should have the foundation for getting started on your own data analysis work in R and will know where and how to get help.

We have borrowed from many sources for this material, most significantly from Jeffery W. Hollister's IntroR course. Thank you, Jeff, for openly sharing your materials. In addition, material has been drawn from [Software Carpentry](#), [Data Carpentry](#), and from R seminars given by numerous individuals at the USGS Center for Integrated Data Analytics and the USGS Wisconsin Water Science Center at the 2014-2015 R Lunch Bunch Data Crunch seminar series. We are grateful to all of these sources.

Each step of the workflow has a written component (blog post) describing the topic, a demonstration of example code, and hands-on exercises. The R code found in each blog post is also available as a stand-alone R script of the same name. The blog posts are intended to be used as a reference after the workshop.

Workshop Schedule

Day 1

[00 - Welcome (8:30 am - 9:00 am)]

Promote and teach best practices

Instruction



OWI-R x Emily

owi.usgs.gov/R/training.html

Product Review

★★★★★
Great job! It's a lot of material to cover with a large variety of people at different levels. I learned a great deal.

★★★★★
The patience of the instructors with students (including myself) given the variety of experience was greatly appreciated. I wish the course was longer as I would like to know more about R and exactly how to apply R to my current projects.

★★★★★
Great job. I think you reached a wide variety of students.

★★★★★
Great... a lot of folks in WSCs may be in more of a need to use "canned" scripts or methods (example Egret)

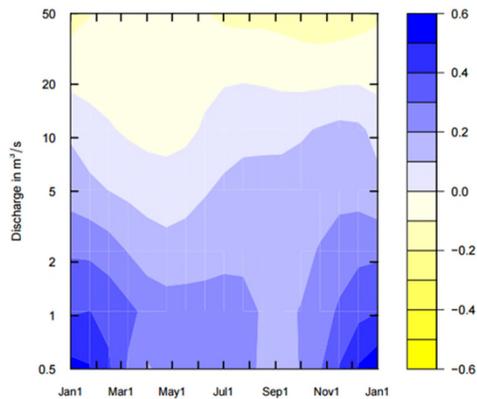
★★★★★
Instructors did a great job for the amount of material they had to convey and the diversity of the class members.

★★★★★
The instructors were excellent and have extensive R background. I was impressed at how well they taught the material and gave a broad understanding of the potential tools available.

★★★★
I really appreciate the patience and knowledge you guys have. I understand that it is difficult to come into a room of such varied experience and teach at a level that is broadly accepted and useful. I really look forward to possibly communicating about projects in the future. Thanks so much, you guys did a great job!

Promote and teach best practices

Advising



USGS Water Use in the United States

water.usgs.gov/watuse/

USGS science for a changing world

USGS Home Contact USGS Search USGS

Water Use in the United States

Home Publications Data About Us Contact Water Use Data and Research Internal

Water Use in the United States

The U.S. Geological Survey's National Water-Use Information Program is responsible for compiling and disseminating the nation's water-use data. The USGS works in cooperation with local, State, and Federal environmental agencies to collect water-use information. USGS compiles these data to produce water-use information aggregated at the county, state, and national levels. Every five years, data at the county level are compiled into a national water-use data system and state-level data are published in a [national circular](#). Over the history of these circulars, the [water-use categories have had some changes](#). The [USGS Water-Use Data and Research program](#) seeks to develop improved water-use data through agreements with State water-resources agencies.

Estimated Use of Water in the United States in 2010 is available (published November 2014). [Report](#) | [Data download](#) | [Fact sheet](#)

Work on the 2015 report began in calendar year 2016.

Water Use Overviews

Total Water Use

Total water use: Estimated total water use for all categories and sources by State.

[More](#)

Surface Water and Groundwater Use

Surface Water and Groundwater Use: Water-use estimates for groundwater and surface water by State.

[More](#)

Trends in Water Use

Trends: How water use is changing over time, starting with the initial USGS estimates for 1950.

[More](#)

Categories of water use

Public Supply

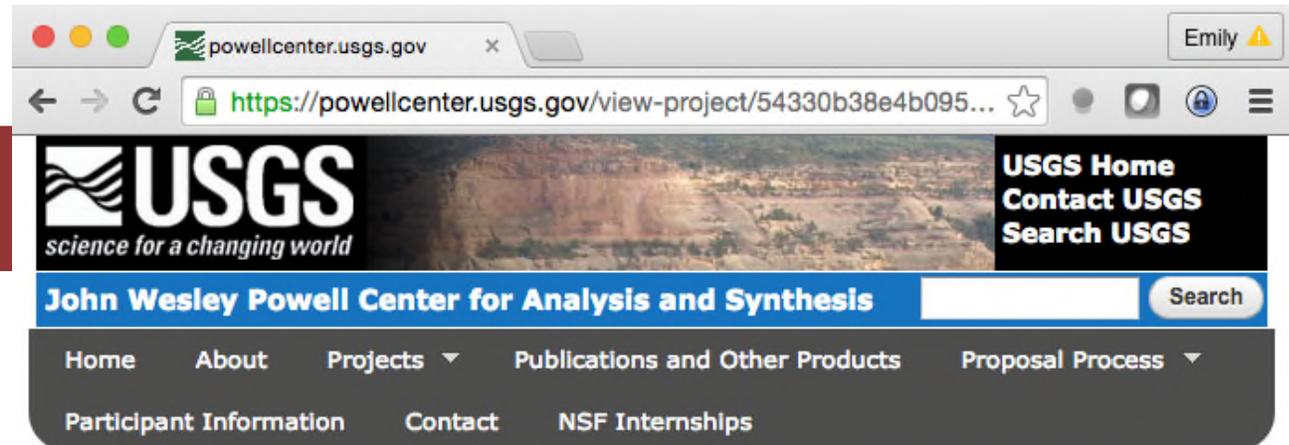
Domestic

Irrigation

Thermoelectric Power

Promote and teach best practices

Collaboration



Powell Center Working Group Project Information

Continental-scale overview of stream primary productivity, its links to water quality, and consequences for aquatic carbon biogeochemistry

Principal Investigator(s):

Edward Stets (*USGS Branch of Regional Research, Central Region*)

Emily Stanley (*University of Wisconsin-Madison*)

Jordan S Read (*Center for Integrated Data Analytics (CIDA)*)

Robert Hall (*University of Wyoming*)

Participant(s):

Charles B Yackulic (*Grand Canyon Monitoring and Research Field Station, SBSC*)

David L Lorenz (*Minnesota Water Science Center*)

Judson W Harvey (*Branch of Regional Research, Eastern Region*)

Natalie Griffiths (*Oak Ridge National Laboratory*)

Alison Appling (*University of Wisconsin*)

Emily Bernhardt (*Duke University*)

Jim Heffernan (*Duke University*)

Maite Arroita (*University of the Basque Country*)

Catalog of tools

rOpenSci packages

Data Packages that interface with data repositories

Package	Description	Details
spocc	Collect and map species occurrence data from GBIF, iNaturalist, Ecoengine, AntWeb, eBird, and USGS's BISON.	CRAN GITHUB
AntWeb	Access data from the world's largest ant database. Maintained and developed by the California Academy of Science	CRAN GITHUB
ecoengine	Retrieve 3+ MM records, checklists and photos from Berkeley's ecoengine	CRAN GITHUB
rebird	Search and acquire occurrence records of bird sightings collated in the eBird database	CRAN GITHUB

USGS-R

USGS R codes and packages

Statistics
35 public repos
8 members
ldecicco@usgs.gov

Recently updated View All on GitHub
EGRET Dec 13, 2014 · 3 watchers · 9 forks
dataRetrieval Dec 13, 2014 · 10 watchers · 7 forks
mda.streams Dec 12, 2014 · 0 watchers · 5 forks

dataRetrieval
This R package is designed to obtain its water quality sample data, streamflow data, and metadata directly from either the USGS or EPA, as well as user-supplied text files.

EGRET
An R-package for the analysis of long-term changes in water quality and streamflow, including the water-quality method Weighted Regressions on Time, Discharge, and Season (WRTDS)

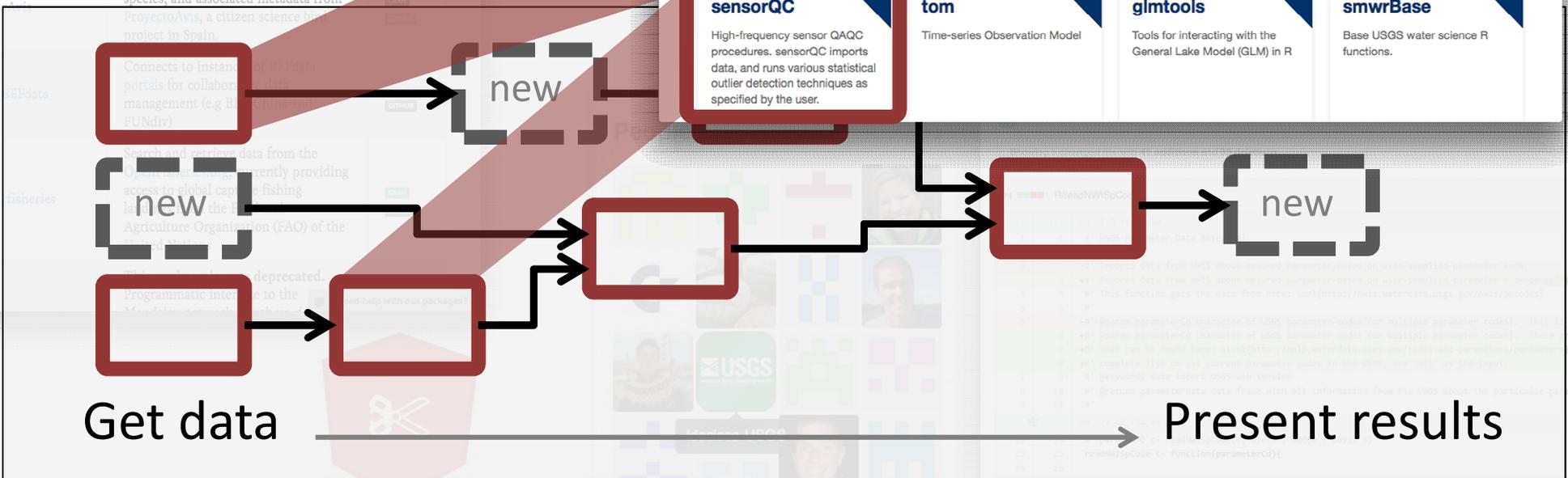
powstreams
Tools for Powell Center working group on stream metabolism

sensorQC
High-frequency sensor QA/QC procedures. sensorQC imports data, and runs various statistical outlier detection techniques as specified by the user.

tom
Time-series Observation Model

glmtools
Tools for interacting with the General Lake Model (GLM) in R

smwrBase
Base USGS water science R functions.



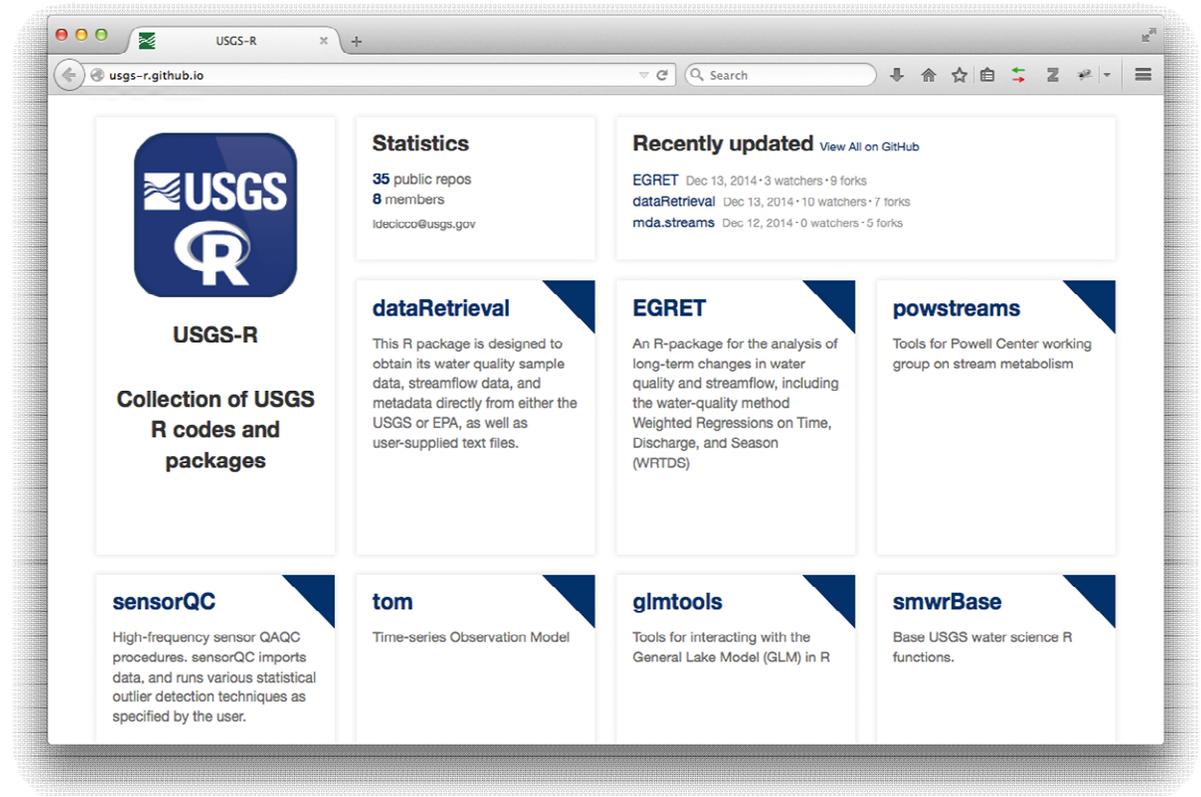
Invite someone

```

27 + if(any(parameterCd == "all")){
28 +   fullURL <- "http://mwis.waterdata.usgs.gov/mwis/pmcodes/pmcodes?radio_pm_search=param_group&pm_group=
29 +   fullPcodeDownload <- importRDBI(fullURL)
30 +   return(fullPcodeDownload)
31 + } else {
32 + }
  
```

Catalog of tools

	packagename	d1_count
1	dataRetrieval	1376
2	smwrGraphs	1223
3	smwrBase	1209
4	smwrQW	1054
5	smwrStats	1002
6	smwrData	934
7	WQReview	478
8	rloadest	392
9	glmtools	386
10	EGRET	345
11	sbtools	328
12	GLMr	321
13	EflowStats	294
14	unitted	249
15	geoknife	240



Catalog of tools

rOpenSci packages

Data Packages that interface with data repositories

Package	Description	Details
spocc	Collect and map species occurrence data from GBIF, iNaturalist, Ecoengine, AntWeb, eBird, and USGS's BISON.	
AntWeb	Access data from the world's largest ant database. Maintained and developed by the California Academy of Science	
ecoengine	Retrieve 3+ MM records, checklists and photos from Berkeley's ecoengine	
rebird	Search and acquire occurrence records of bird sightings collated in the eBird database	
rAvis	Download occurrence data for bird species, and associated metadata from ProyectoAvis, a citizen science bird project in Spain.	
BEFdata	Connects to instances of BEFdata portals for collaborative data management (e.g BEF-China and FUNdiv)	
rfisheries	Search and retrieve data from the OpenFisheries.org, currently providing access to global capture fishing landings from the Food and Agriculture Organization (FAO) of the United Nations	

USGS-R

USGS

Collection of USGS R codes and packages

Statistics

- 35 public repos
- 8 members
- ldecicco@usgs.gov

Recently updated

- EGRET Dec 13, 2014 · 3 watchers · 9 forks
- dataRetrieval Dec 13, 2014 · 10 watchers · 7 forks
- mda.streams Dec 12, 2014 · 0 watchers · 5 forks

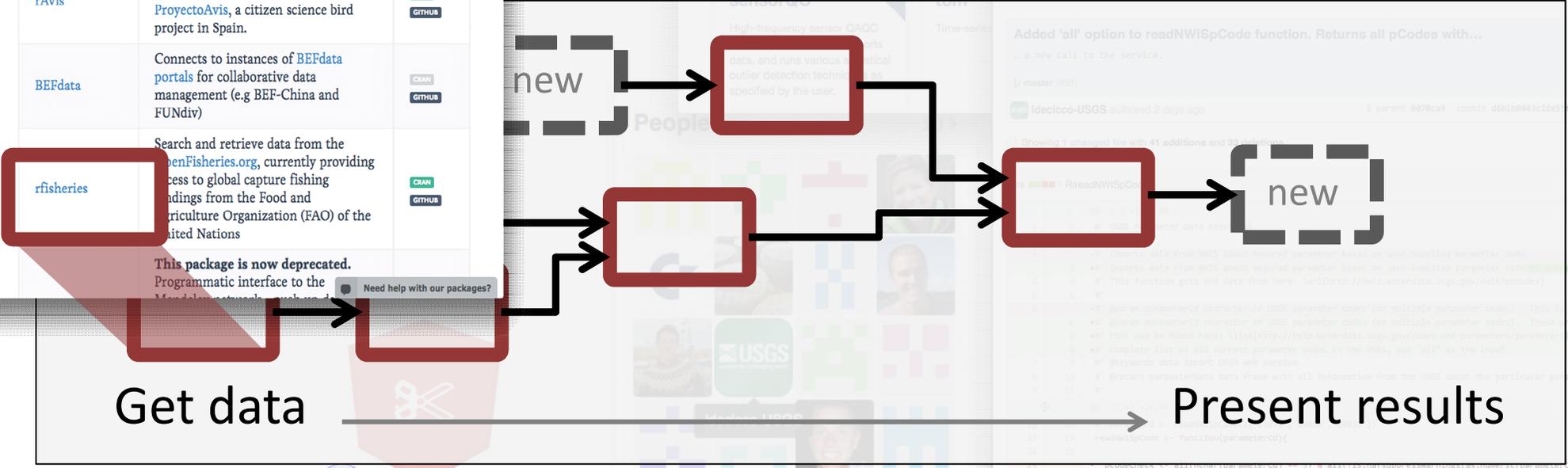
dataRetrieval

EGRET

powstreams

Added 'all' option to readNWISpCode function. Returns all pCodes with...

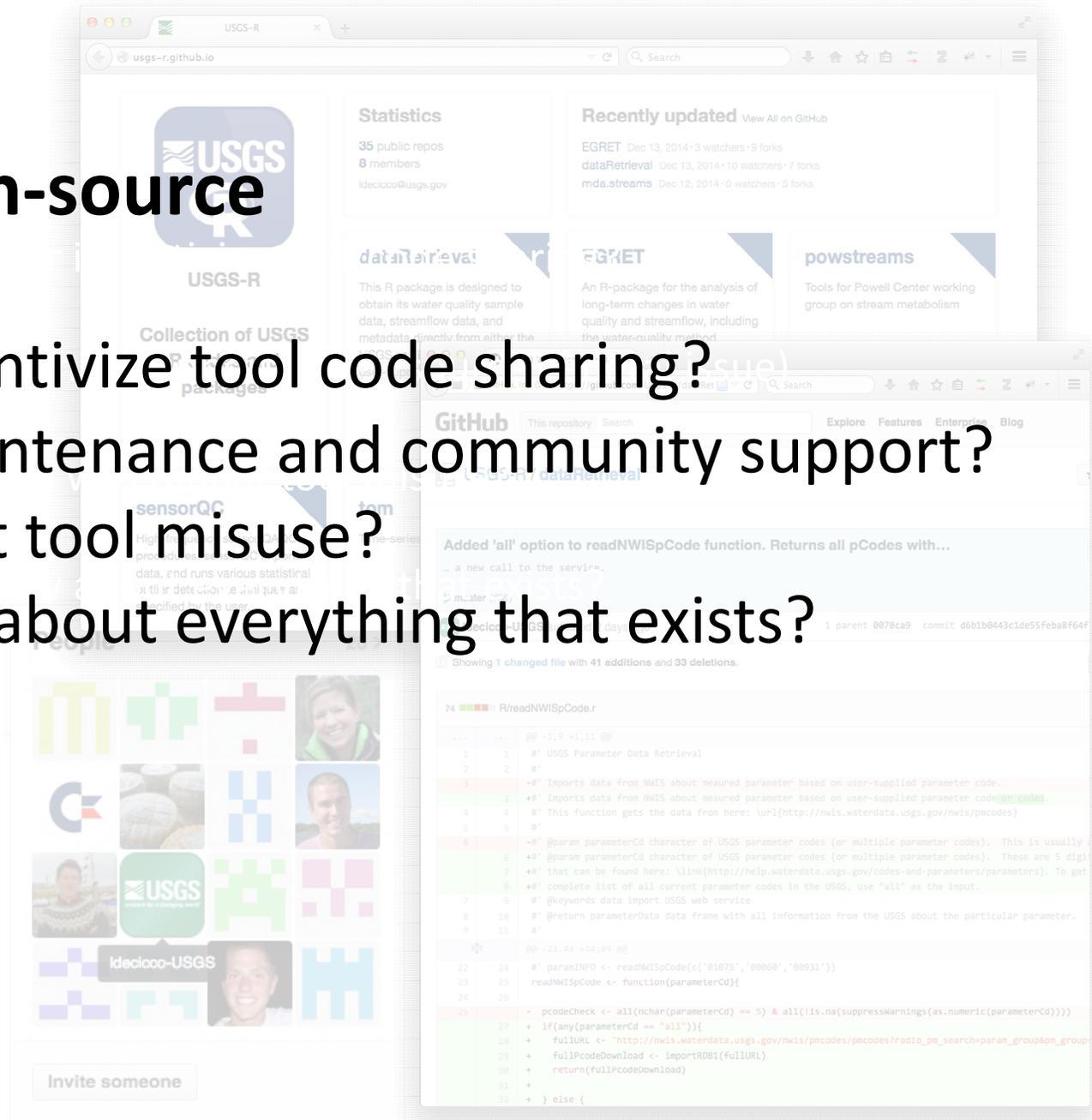
```
1/ master (406)
ldecicco-USGS authored 2 days ago
Showing 1 changed file with 41 additions and 35 deletions
R/readNWISpCode.R
+ readNWISpCode <- function(parameterCd, fullURL, fullPcodeDownload) {
+   data <- readNWISpCodeFromNWIS(fullURL, parameterCd)
+   if (fullPcodeDownload) {
+     fullPcodeDownload <- importRDB1(fullURL)
+     return(fullPcodeDownload)
+   } else {
+     return(data)
+   }
+ }
```



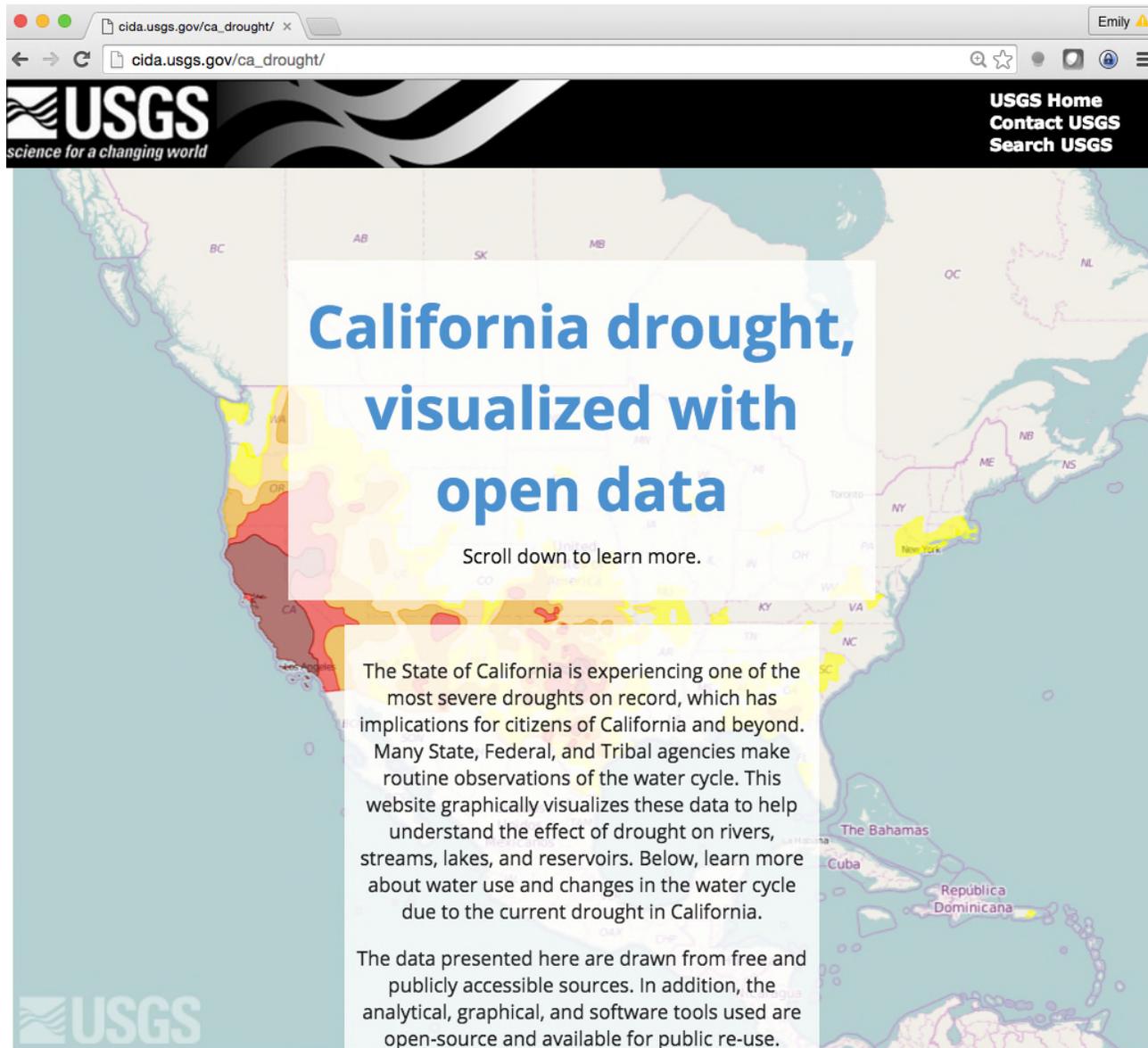
Catalog of tools

Challenges to open-source

- Can we better incentivize tool code sharing?
- How to ensure maintenance and community support?
- How do we combat tool misuse?
- How can we know about everything that exists?



Web-enabled analysis



The image shows a screenshot of a web browser displaying a USGS page. The browser's address bar shows the URL `cida.usgs.gov/ca_drought/`. The page features the USGS logo and navigation links. A map of the United States is shown with a color-coded overlay indicating drought severity, with California highlighted in red and orange. A large text box is overlaid on the map with the following content:

California drought, visualized with open data

Scroll down to learn more.

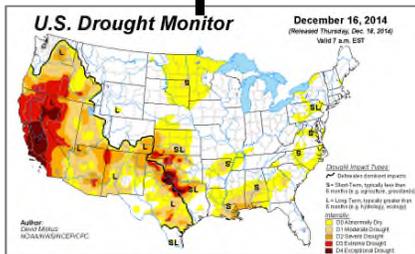
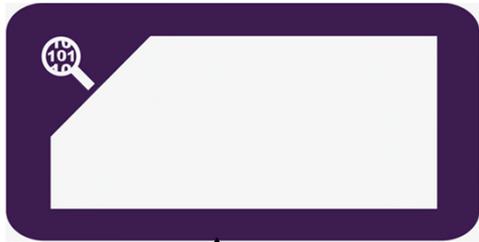
The State of California is experiencing one of the most severe droughts on record, which has implications for citizens of California and beyond. Many State, Federal, and Tribal agencies make routine observations of the water cycle. This website graphically visualizes these data to help understand the effect of drought on rivers, streams, lakes, and reservoirs. Below, learn more about water use and changes in the water cycle due to the current drought in California.

The data presented here are drawn from free and publicly accessible sources. In addition, the analytical, graphical, and software tools used are open-source and available for public re-use.

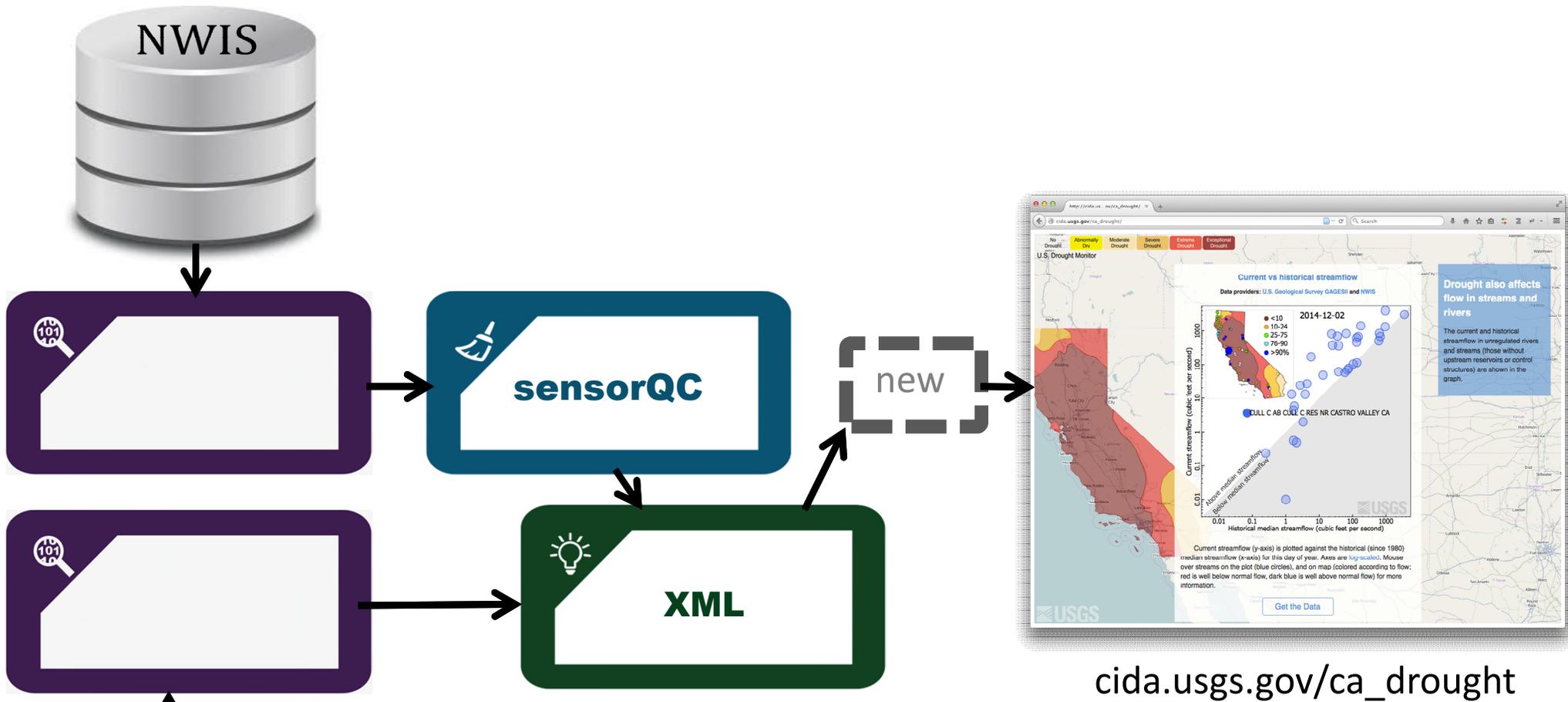
Web-enabled analysis



“dataRetrieval is designed to obtain its water quality sample data, streamflow data, and metadata directly from either the USGS or EPA, as well as user-supplied text files.”

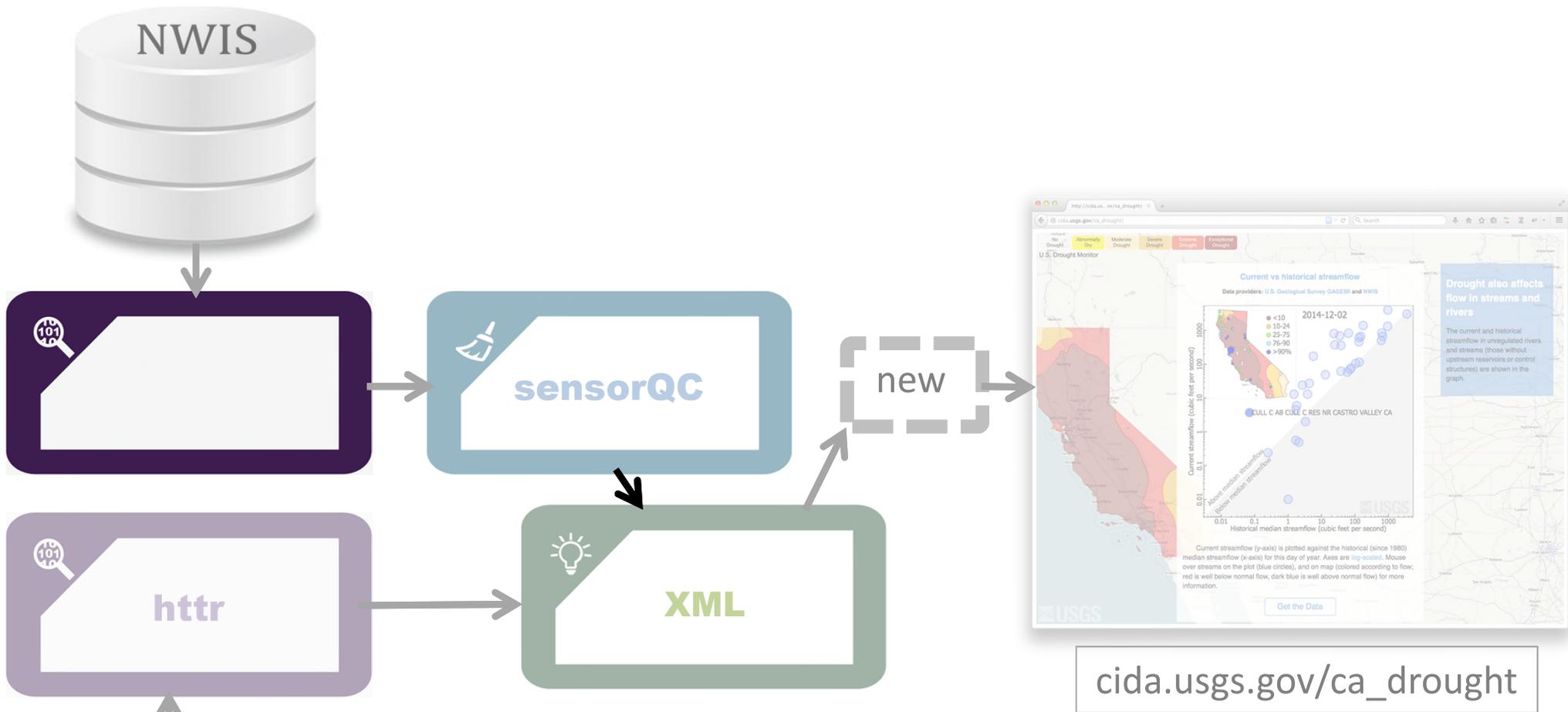


Web-enabled analysis



cida.usgs.gov/ca_drought

Community-based 'help' platform



Community-based 'help' platform



Issues · USGS-R/dataRetri x

GitHub, Inc. [US] https://github.com/USGS-R/dataRetrieval/issues

This repository Search Pull requests Issues Gist

USGS-R / dataRetrieval Unwatch

Code Issues 16 Pull requests 0 Wiki Pulse Graphs

Filters is:issue is:open Labels Milestones

16 Open 66 Closed Author Labels

- WQP HEAD request enhancement #202 opened 20 days ago by ldecicco-USGS
- Getting more parameters than I specified #200 opened 21 days ago by UserFeb2016
- update parameterCdFile? #193 opened on Mar 10 by jread-usgs
- user-friendly chunking of service calls? question #185 opened on Mar 3 by jread-usgs
- Support a retry argument for service calls? question #184 opened on Mar 3 by jread-usgs
- readNWISdata question question #180 opened on Feb 25 by ldecicco-USGS
- Allow more times zones when retrieving data from webservices please. question #152 opened on Dec 3, 2015 by cjhoard
- Error reading WaterML 2 data #131 opened on Jul 11, 2015 by jirikadlec2

Community-based 'help' platform



readWQPdata - fewer results returned than expected

Closed jjwill2 opened this issue 18 days ago · 4 comments

jjwill2 commented 18 days ago

I used the code below to try to query all nutrient chemistry data from lakes in OR. There was a warning message: "In importWQP(urlCall, FALSE, tz = tz) : 27571 sample results were expected, 1611 were returned". Is there any way to address or troubleshoot these warnings? Doing the same query through the WQ portal web interface yielded 27,374 results. What is happening here?

```
OR_lake_data <-readWQPdata(  
  
  # specify states  
  statecode = "US:41",  
  
  # specifies water samples  
  ActivityMediaName = "water",  
  
  # specifies lakes  
  siteType = "Lake, Reservoir, Impoundment",  
  
  # specifies parameter category  
  characteristicType = "Nutrient")
```

Community-based 'help' platform



Allow more times zones when retrieving data from webservices please. #152

Open cjhoard opened this issue on Dec 3, 2015 · 2 comments

cjhoard commented on Dec 3, 2015

I have an enhancement request, when retrieving data from waterservices using the readNWISuv command. All of the timeseries data collected in MI is stored as EST, so when we retrieve through dataRetrieval and use "America/New_York" (the only eastern timezone allowed) as the timezone all data retrieved during daylight savings time is 1 hour off. This poses a problem when we pull a year of data : some of the data is correct but some of the data is an hour off, so we have to do some data manipulation to adjust all that. If EST were an allowed timezone it would save us some time from having to reformat the times when we retrieve the data. Anyhow thank you for considering this enhancement request.

for example:

```
test<-readNWISuv('415318085243401',startDate="2015-06-01", endDate="2015-08-31",parameterCd = '
head(test,1)
```

agency_cd	site_no	dateTime	tz_cd	X_72019_00011	X_72019_00011_cd
1	USGS 415318085243401	2015-06-01	01:00:00	America/New_York	-0.26 P
2	USGS 415318085243401	2015-06-01	01:15:00	America/New_York	-0.27 P

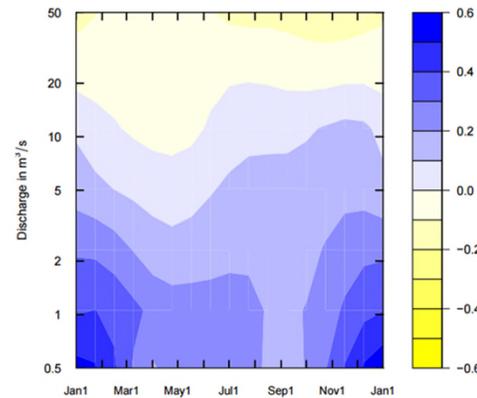
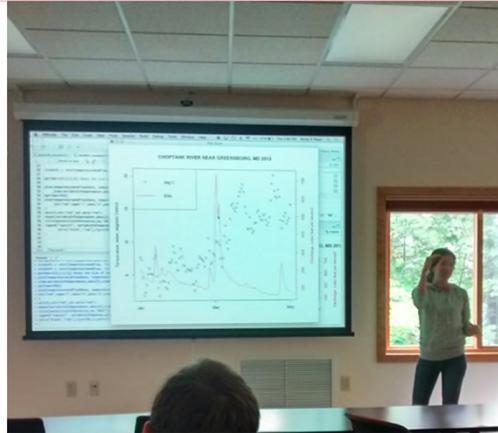
Community of learners and teachers

Outreach

Instruction

Advising

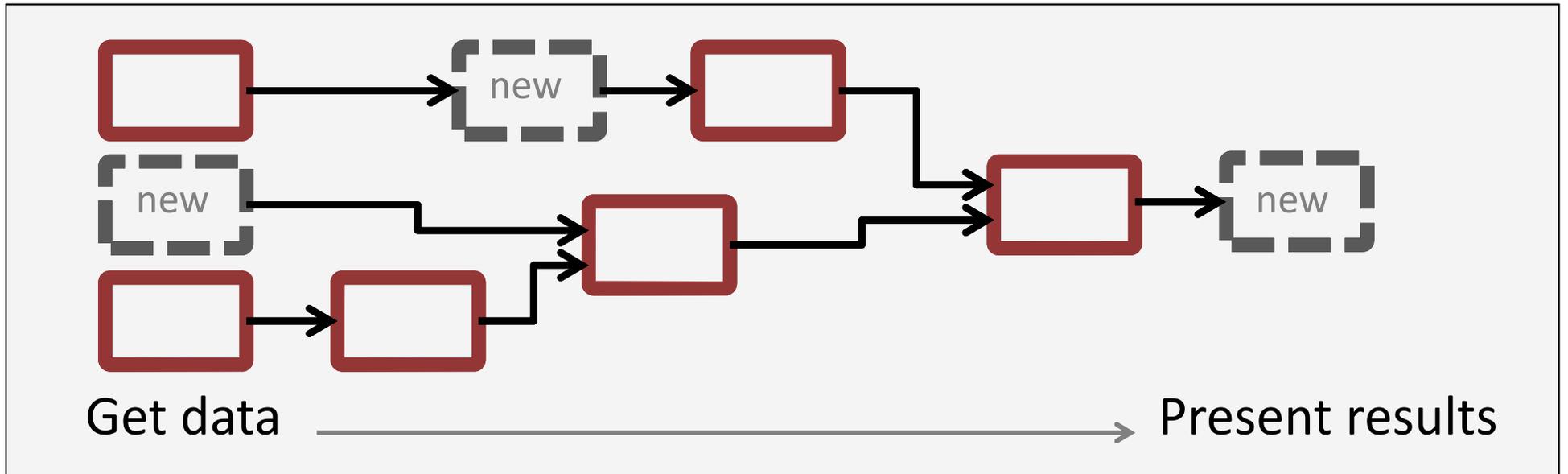
Collaboration



Local to regional-scale
descriptive science

← Data/information/tool divide →

Broad-scale
prediction



Outreach

Instruction

Advising

Collaboration

In summary

- Building blocks for analysis can expand scope and increase efficiency of science efforts
- A learning community and clear access to tools is important

eread@usgs.gov