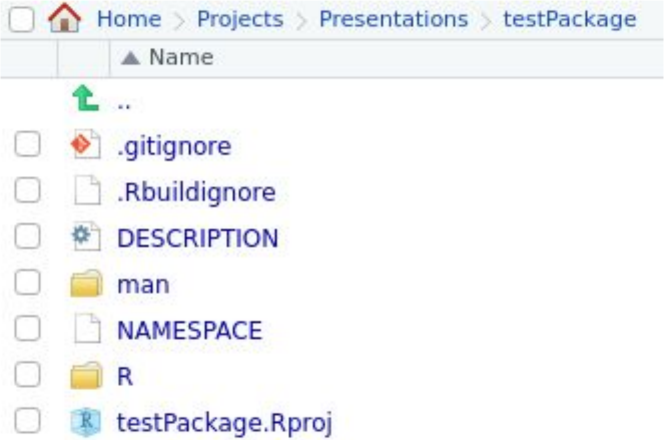
Creating Data Packages in R

Emanuel Rodriguez

erodriguez@flowwest.com

What is an R package?

- A collection of functions, documentation, tests and data
- Traditionally *functions* have been the purpose for package distribution
    - Write documentation to explain the usage
    - Write tests to validate them
    - Bundle data to showcase them



□ Home > Projects > Presentations > testPackage

▲ Name

..

□ .gitignore

□ .Rbuildignore

□ DESCRIPTION

□ man

□ NAMESPACE

□ R

□ testPackage.Rproj

What is a data package?

- This time the data is the purpose for distribution
  - Write functions, documentation and tests to support the data
  - Write documentation to explain usage, detail the source and add additional metadata *(roxygen)*
  - Write tests to qa/qc *(testthat)*
  - Write functions to facilitate additional processing, and visualizations

Why create data packages?

- Create reproducible workflows for data distribution
    - Capture and version control the entire process of going from raw to clean
    - Facilitate team workflow
- Bundle documentation with the data
    - Using roxygen and a well defined documentation workflow create a package that details sources, usage and purpose for data
- Package and release versioned data alongside other products
    - Develop data packages as part of larger deliverable or “ecosystem”
    - Package deliverable data
- Use a handful of additional packages to extend usage of data
    - Build websites with pkgdown
    - Create API’s with plumbr
    - Integrate into Shiny Apps

The Workflow

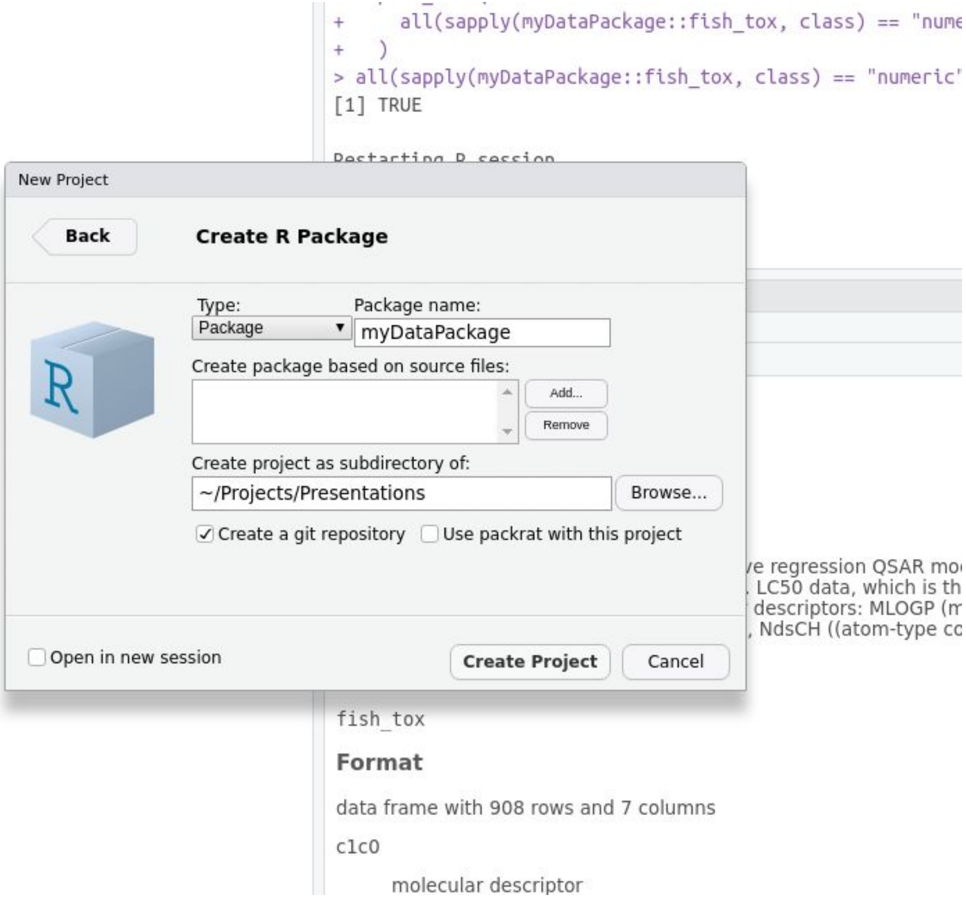
Use the R package framework to capture all aspects of going from raw to clean

*A Proposed Workflow (using the **usethis** package)*

1. *Start at the data-raw/ folder, place all raw data here*
2. *Start scripting the wrangling process, place these scripts in data-raw*
3. *Once complete create a binary of the data in the data/ folder and document*
4. *Create test suite for the data*
5. *Build, (Publish?) and repeat*

Package Creation

- Follow Package naming conventions
- Enable git repository



```
+     all(sapply(myDataPackage::fish_tox, class) == "nume
+   )
> all(sapply(myDataPackage::fish_tox, class) == "numeric"
[1] TRUE

Restarting R session
```

Type: Package ▾

Package name: myDataPackage

Create package based on source files:

Create project as subdirectory of: ~/Projects/Presentations

☑ Create a git repository ☐ Use packrat with this project

☐ Open in new session

ve regression QSAR mo
. LC50 data, which is th
descriptors: MLOGP (m
, NdsCH ((atom-type co

fish_tox

Format

data frame with 908 rows and 7 columns

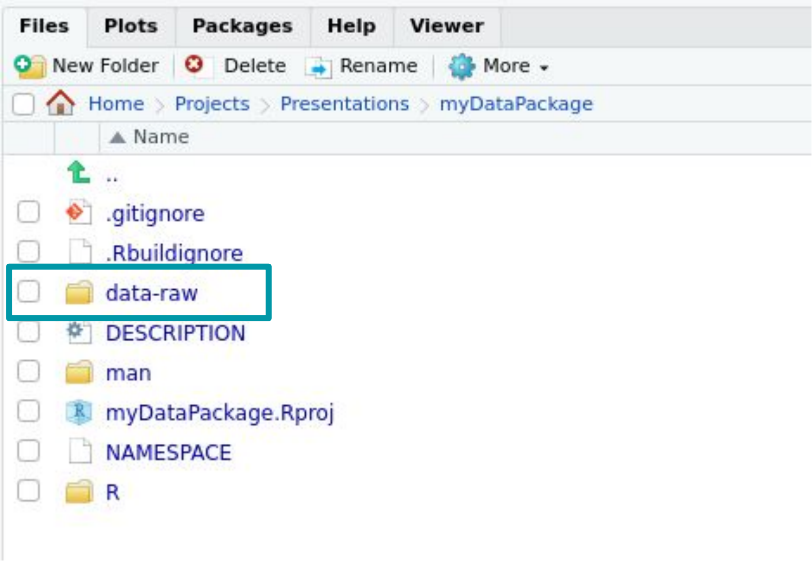
c1c0

molecular descriptor

Package Creation

```
usethis::use_data_raw()
```

```
✓ Setting active project to
'/home/emanuel/Projects/Presentations/myDataPackage'
✓ Creating 'data-raw/'
✓ Adding '^data-raw$' to '.Rbuildignore'
✓ Writing 'data-raw/DATASET.R'
● Modify 'data-raw/DATASET.R'
● Finish the data preparation script in
'data-raw/DATASET.R'
● Use `usethis::use_data()` to add prepared data to
package
```



Home › Projects › Presentations › myDataPackage

..
.gitignore
.Rbuildignore
data-raw
DESCRIPTION
man
myDataPackage.Rproj
NAMESPACE
R

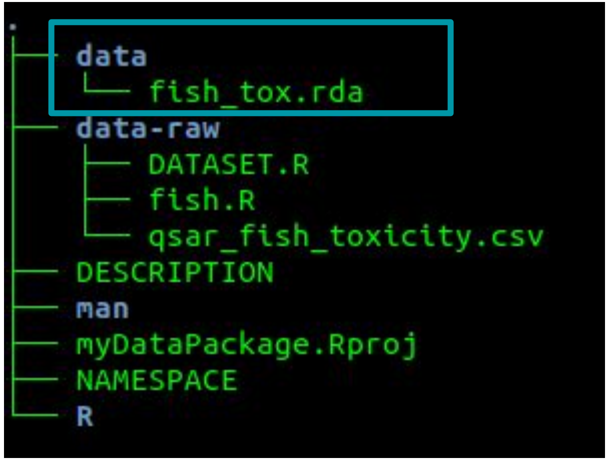
Wrangling and Packaging

```
# data-raw/fish.R
library(tidyverse)
col_names <- c("c1c0", "sm1",
               "gat", "ndsch",
               "ndssc", "mlo",
               "resp")

fish_tox <-
read_delim("data-raw/qsar_fish_toxicity.csv",
               col_names = col_names,
               delim = ";")


usethis::use_data(fish_tox, overwrite = TRUE)
```

```
✓ Setting active project to
'/home/emanuel/Projects/Presentations/myDataPackage'
✓ Creating 'data/'
✓ Saving 'fish_tox' to 'data/fish_tox.rda'
```



```
.
├── data
│   └── fish_tox.rda
├── data-raw
│   ├── DATASET.R
│   ├── fish.R
│   └── qsar_fish_toxicity.csv
├── DESCRIPTION
├── man
├── myDataPackage.Rproj
├── NAMESPACE
└── R
```

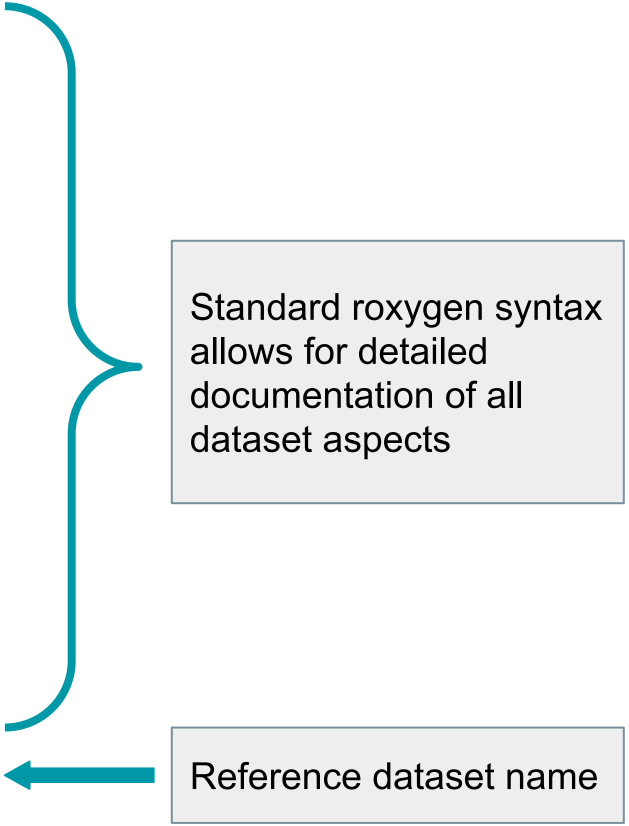
Document Data

- Use roxygen2 to document datasets in detail
- Create and use a data.R file in R the directory
- Build documentation
  - `devtools::document(roclets = c('rd', 'collate', 'namespace')) or`
  - **`Ctrl+Shift+B`** `(build package)`

Document

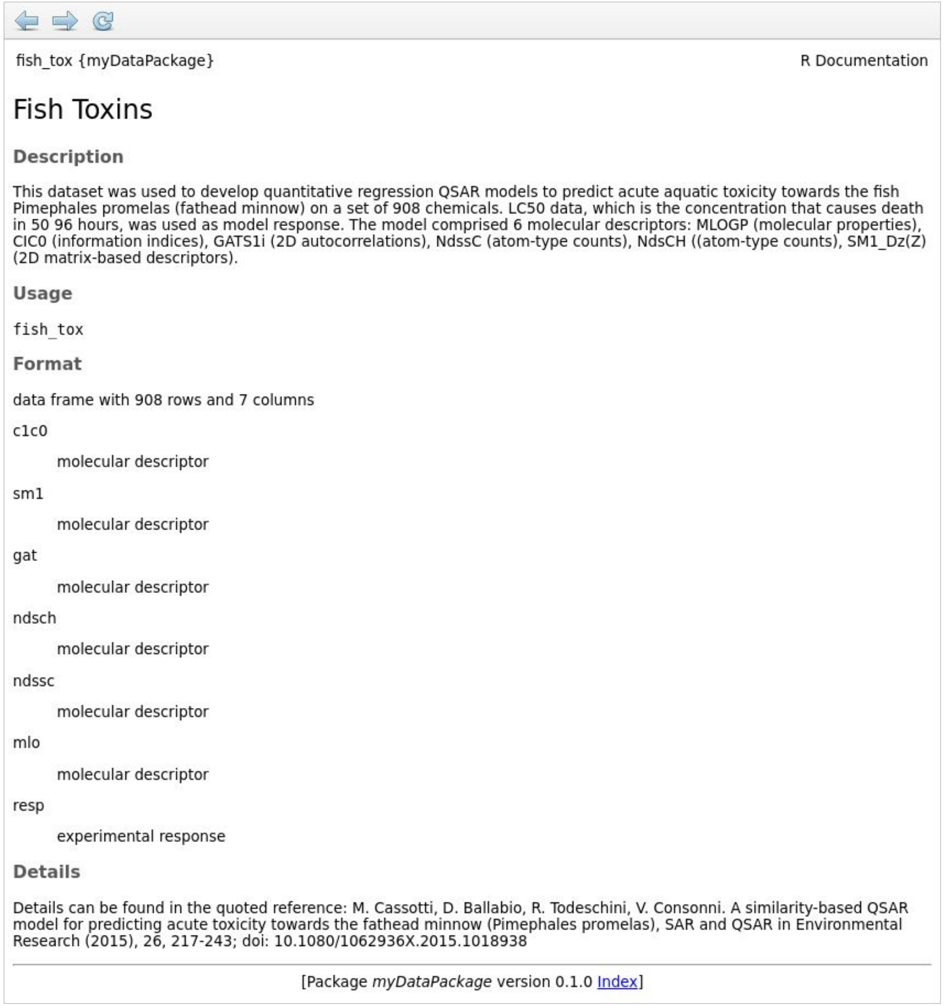
```
R/data.R


#' @title Fish Toxins
#' @description This dataset was used to develop quantitative regression
QSAR
#' models to predict acute aquatic toxicity towards the fish Pimephales
promelas
#' (fathead minnow) on a set of 908 chemicals. LC50 data, which is the
#' concentration that causes death in 50% of test fish over a test duration
of
#' 96 hours, was used as model response. The model comprised 6 molecular
#' descriptors: MLOGP (molecular properties), CIC0 (information indices),
#' GATS1i (2D autocorrelations), NdssC (atom-type counts),
#' NdsCH ((atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors).
#' @format data frame with 908 rows and 7 columns
#' \describe{
#' \item{c1c0}{molecular descriptor}
#' \item{sm1}{molecular descriptor}
#' \item{gat}{molecular descriptor}
#' \item{ndsch}{molecular descriptor}
#' \item{ndssc}{molecular descriptor}
#' \item{mlo}{molecular descriptor}
#' \item{resp}{experimental response}
#' }
#' @details
#' Details can be found in the quoted reference: M. Cassotti, D. Ballabio,
R. Todeschini,
#' V. Consonni. A similarity-based QSAR model for predicting acute toxicity
towards
#' the fathead minnow (Pimephales promelas), SAR and QSAR in Environmental
#' Research (2015), 26, 217-243; doi: 10.1080/1062936X.2015.1018938
"fish_tox"
```



Standard roxygen syntax allows for detailed documentation of all dataset aspects

Reference dataset name



fish_tox {myDataPackage} R Documentation

Fish Toxins

Description

This dataset was used to develop quantitative regression QSAR models to predict acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) on a set of 908 chemicals. LC50 data, which is the concentration that causes death in 50 96 hours, was used as model response. The model comprised 6 molecular descriptors: MLOGP (molecular properties), CIC0 (information indices), GATS1i (2D autocorrelations), NdssC (atom-type counts), NdsCH ((atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors).

Usage

```
fish_tox
```

Format

data frame with 908 rows and 7 columns

c1c0

molecular descriptor

sm1

molecular descriptor

gat

molecular descriptor

ndsch

molecular descriptor

ndssc

molecular descriptor

mlo

molecular descriptor

resp

experimental response

Details

Details can be found in the quoted reference: M. Cassotti, D. Ballabio, R. Todeschini, V. Consonni. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas), SAR and QSAR in Environmental Research (2015), 26, 217-243; doi: 10.1080/1062936X.2015.1018938

[Package *myDataPackage* version 0.1.0 Index]

Build a test suite for your data

- Use testthat to build a suite of test for data
- Make running test part of the workflow for data wrangling

Test Data

Write test as guided by the testthat package, set up using **usethis::use_testthat()**

```
# test/testthat/test-fish-data.R

context("Fish data")
library(myDataPackage)

test_that("Columns are of the correct type", {
  expect_true(
     all(sapply(myDataPackage::fish_tox, class) ==
"numeric")
  )
})
```

```
Loading myDataPackage
Testing myDataPackage
✓ |  OK F W S | Context
✓ |   1       | Fish data

══ Results ═══════════════════════════════════════════
OK:       1
Failed:   0
Warnings: 0
Skipped:  0
```

Package and release versioned data

- Data is clean, has been documented and been put through a test suite
- Distribute as need; https://github.com/ERGZ/myDataPackage
- Build a website for it and deploy on Github pages

Use case

https://github.com/FlowWest/cvpiaHabitat

More info

Usethis: https://usethis.r-lib.org/

Testthat: https://testthat.r-lib.org/

roxygen2: http://r-pkgs.had.co.nz/man.html

Devtools: https://devtools.r-lib.org/