Computational Skills for Biostatistics I: Lecture 5

Amy Willis, Biostatistics, UW

November 8, 2017

R packages

Why write an R package?

- The best way to contribute to science
- Allow others to build on your advances
- Promote yourself and your work
- Pass this class

$$\text{package} = \text{code} + \text{other stuff}$$

This class will focus on the other stuff

Install versus load

Brief refresher: `install.packages()` downloads the package from CRAN, and `library()` loads the library

- Loads the library? `attach()`-es it
  - Adds it to search path

Attaching

Recall from homework 1 that you can `attach()` data frames

```
attach(women)
height
```

```
##  [1] 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
```

```
height <- height*2.54
height
```

```
##  [1] 147.32 149.86 152.40 154.94 157.48 160.02 162.56 16
## [11] 172.72 175.26 177.80 180.34 182.88
```

```
women$height
```

```
##  [1] 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
```

Attaching

Loading a library places its functions in the search path, but the global environment will always be in the first position

```
ggplot  <- function(...) "Ha! Overwritten!"
library(ggplot2)
ggplot(iris, aes(x = Sepal.Length, y = Petal.Width))
```

```
## [1] "Ha! Overwritten!"
```

Attach

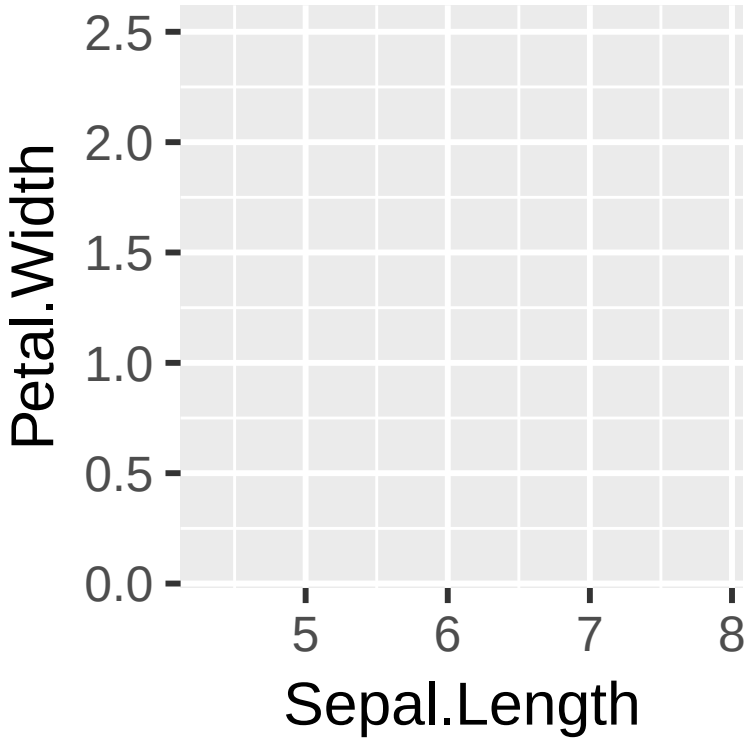
Moral of the story: Don't attach data frames, but it's fine to attach packages

If I ever used ‘require()‘ please forget it and use ‘library()‘ instead

Accessing functions in packages

The operator `x::y` means "the function `y` in the namespace of package `x`"

```
ggplot  <- function(...) "Ha! Overwritten!"
ggplot2::ggplot(iris, aes(x = Sepal.Length,
                          y = Petal.Width))
```



Namespaces

All packages have a `NAMESPACE` file: a collection of objects to be exported and imported

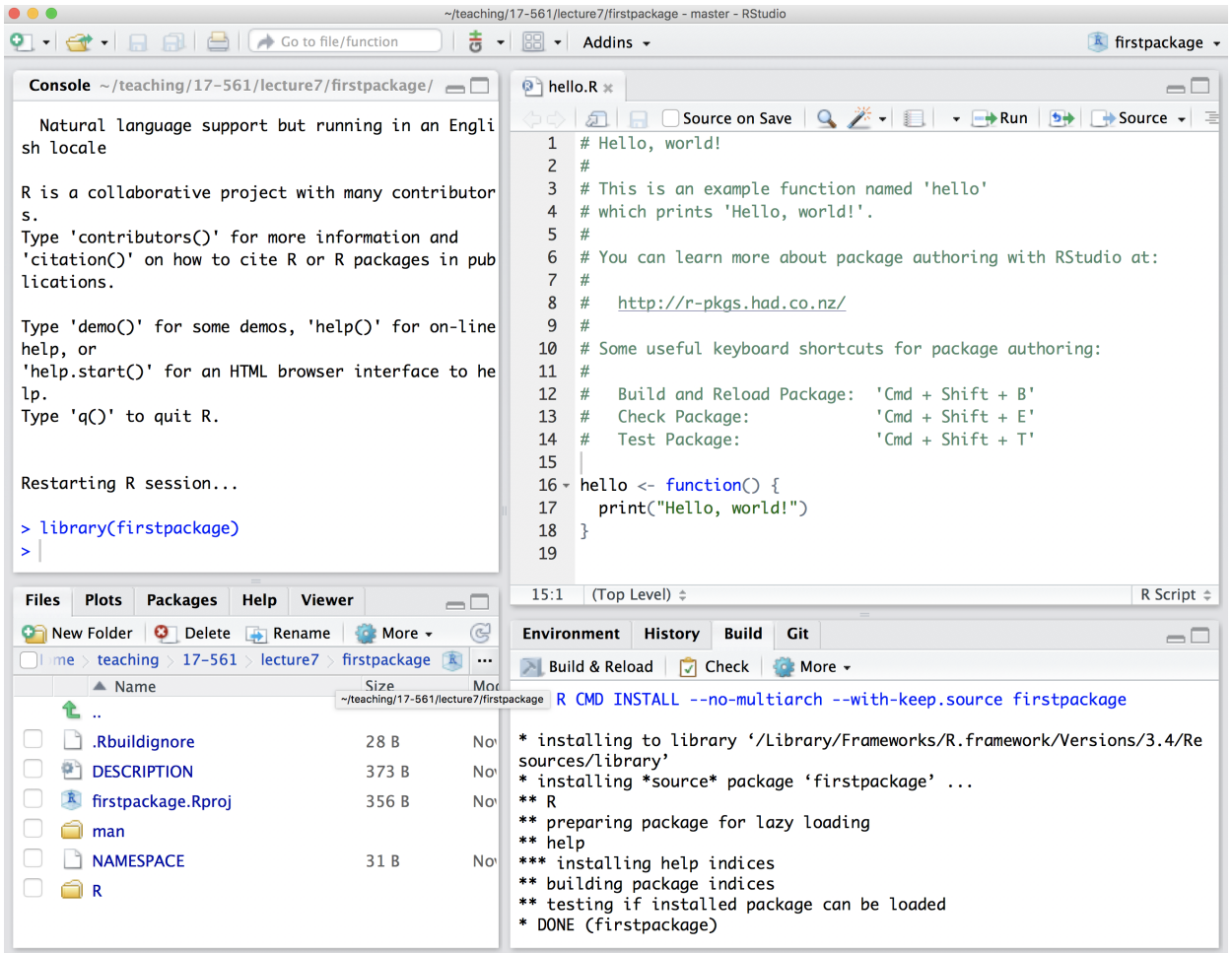
- To avoid overwriting users' variables
- To avoid ambiguity in function calls
- To ensure the package has everything it needs to run
- To encourage modular code

Your `NAMESPACE` is auto generated using *tags*; never directly modify your `NAMESPACE` file

Starting an R package

Ideally, create packages from scratch as soon as you begin on a project

- RStudio->File->New project->New Directory->R Package



Starting an R package: DESCRIPTION

Short-term: Keeps track of imports (dependencies)

Long-term: Help others find your package

hello.R | DESCRIPTION | NAMESPACE | hello.Rd

```
1  Package: firstpackage
2  Type: Package
3  Title: What the Package Does (Title Case)
4  Version: 0.1.0
5  Author: Who wrote it
6  Maintainer: The package maintainer <yourself@somewhere.net>
7  Description: More about what it does (maybe more than one line)
8      Use four spaces when indenting paragraphs within the Description.
9  License: What license is it under?
10 Encoding: UTF-8
11 LazyData: true
12 Imports:
13     dplyr,
14     ggplot
15 Suggests:
16     breakaway
17
```

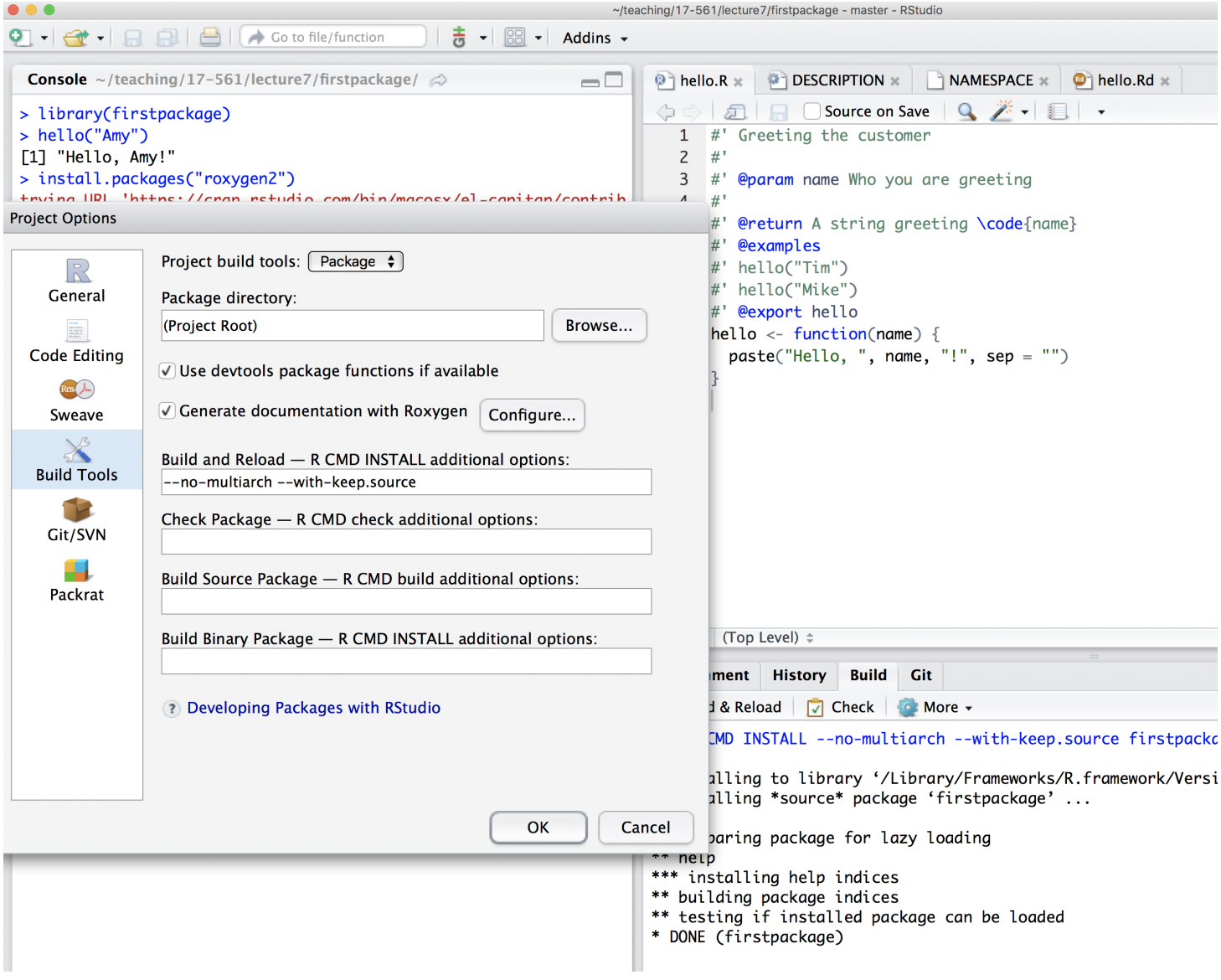
Writing R packages: documentation

`roxygen` documentation makes everything incredibly easy

```
install.packages("devtools")
devtools::install_github("hadley/devtools")
install.packages("roxygen2")
```

Writing R packages: documentation

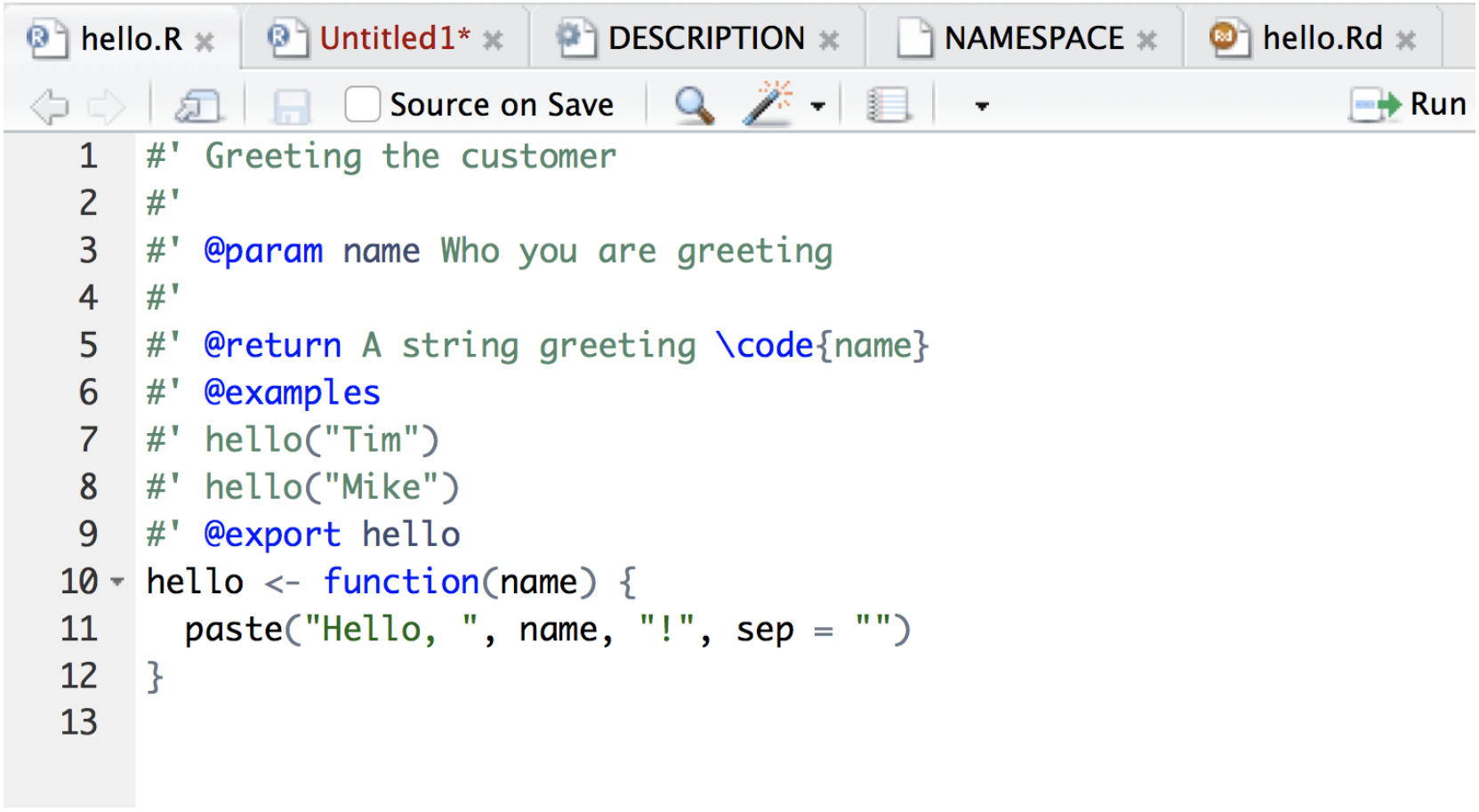
Under the `Build` tab, under Build tools, check `Generate documentation with Roxygen`



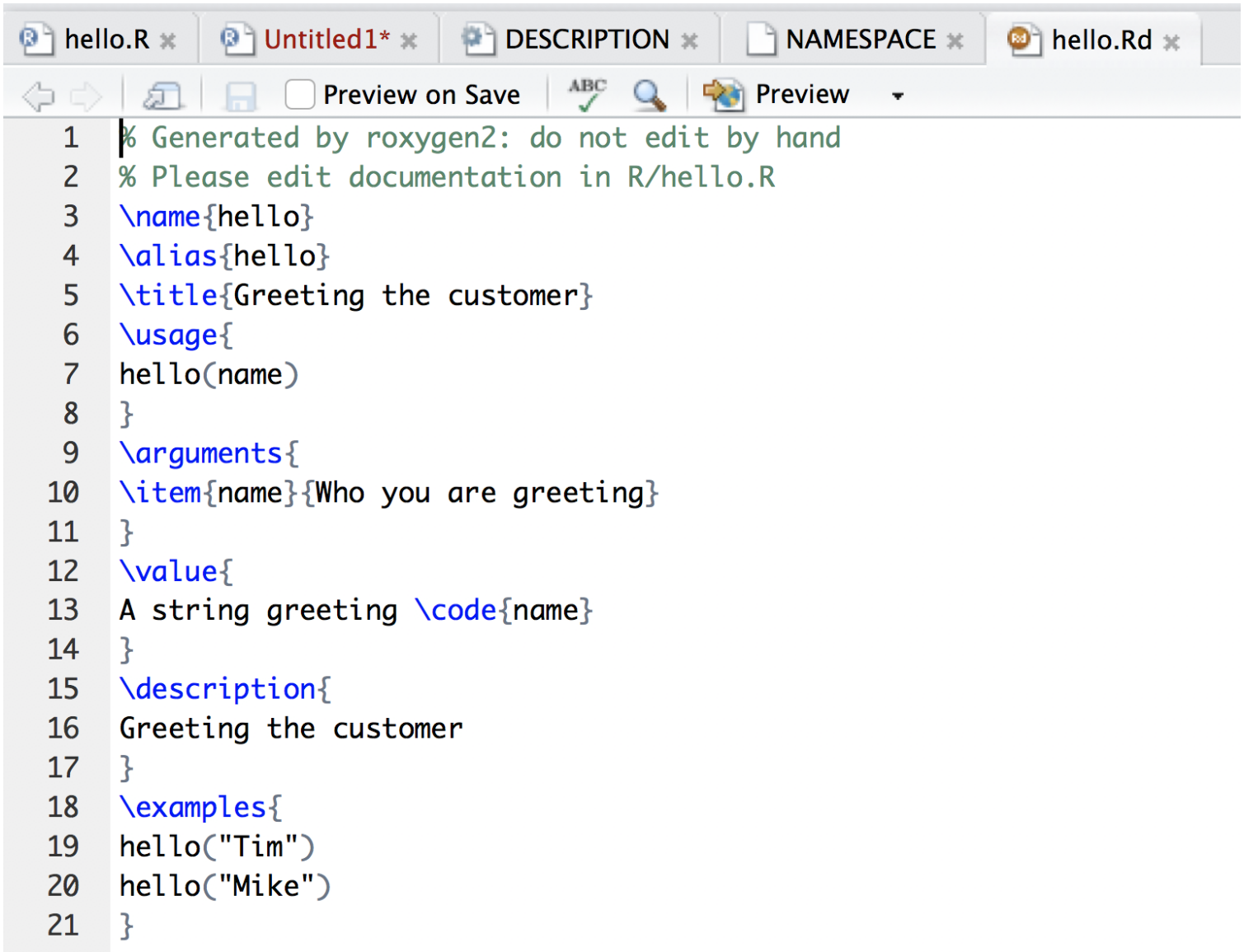
Writing R packages: documentation

1. Delete the default `NAMESPACE` file and the `man` folder
2. Write `roxygen` comments in your `.R` file
   - `roxygen` comments start with `#'`
3. `cmd/ctrl + shift + b` builds the package
4. `cmd/ctrl + shift + d` autogenerates documentation based on your roxygen comments

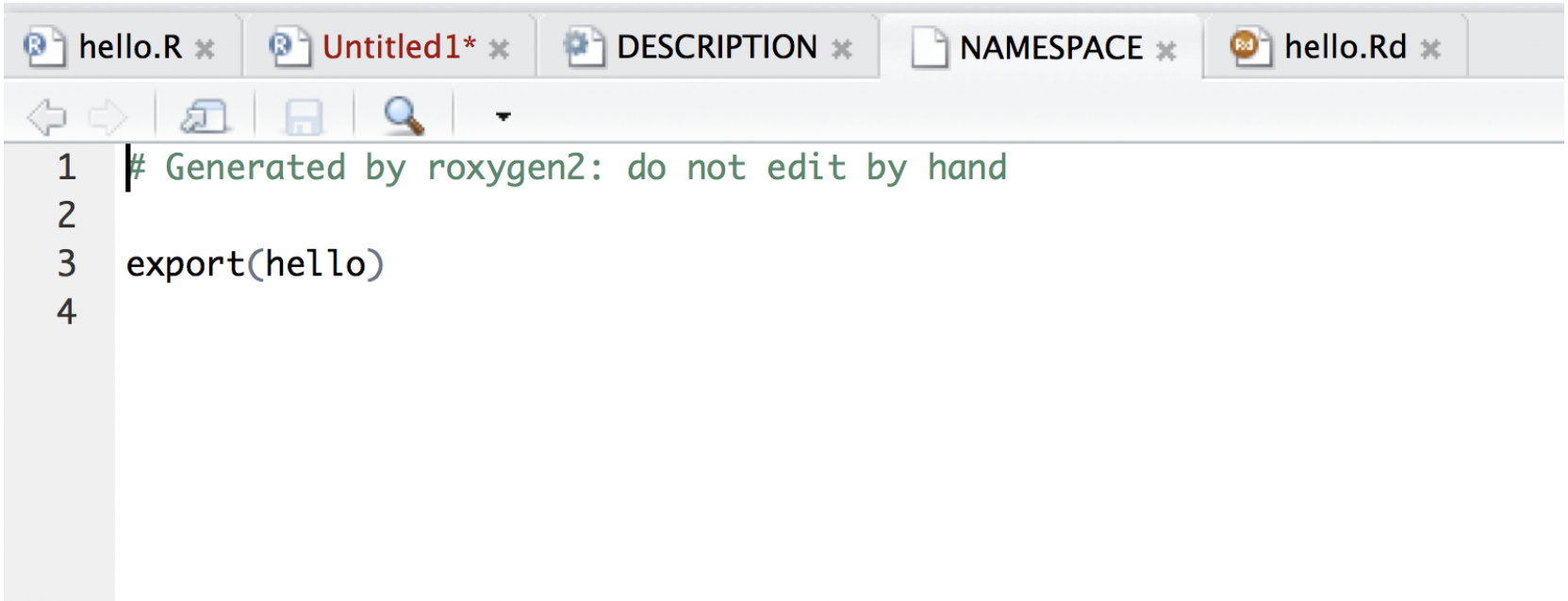
Writing R packages: documentation



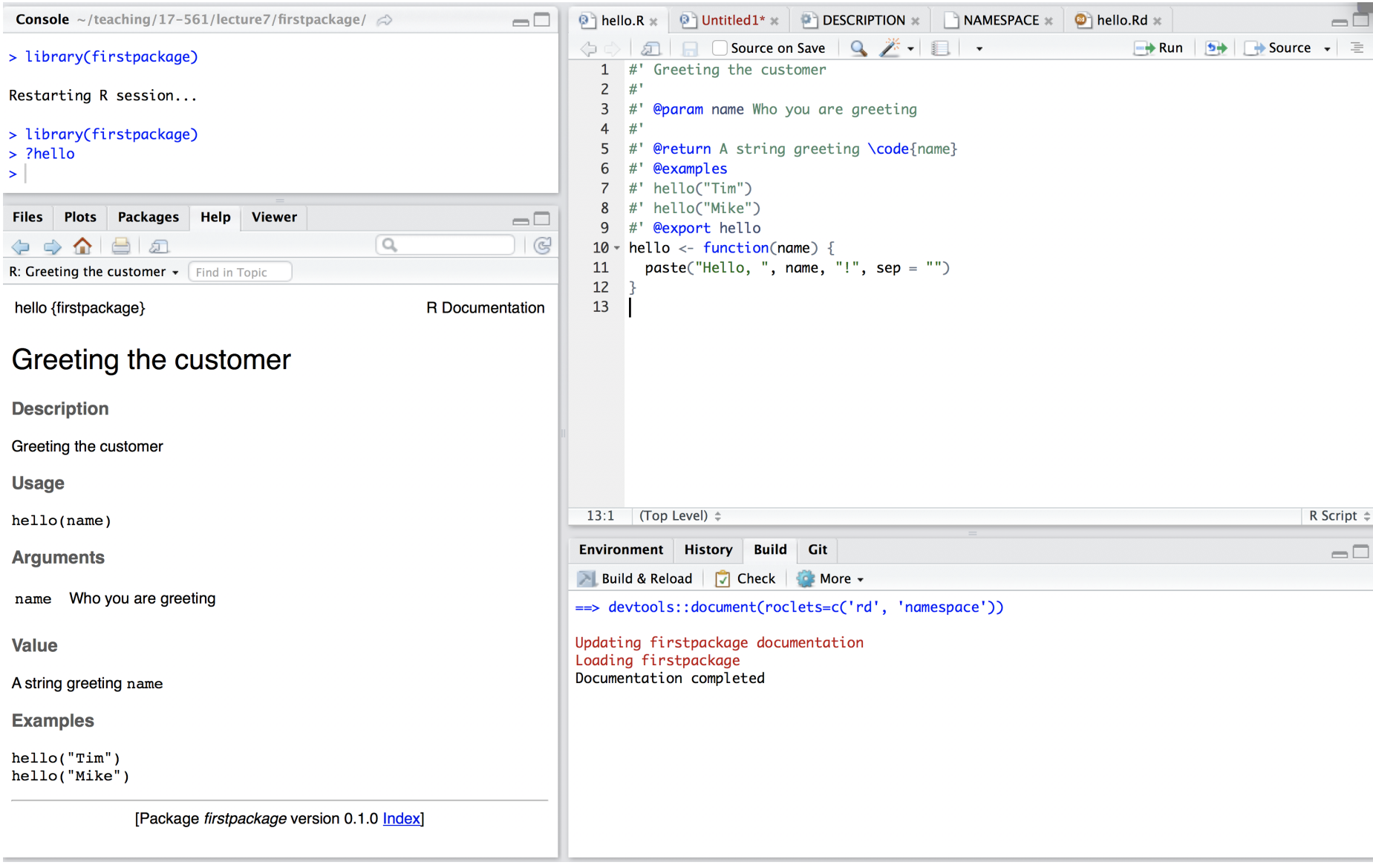
Writing R packages: documentation



Writing R packages: documentation



Writing R packages: documentation



Writing R packages: tags

```
#' \item{seest}{ The standard error in the diversity estimate.  } \item{full}{
#' The chosen nonlinear model for frequency ratios.  } \item{ci}{ An asymmetric
#' 95\% confidence interval for diversity.  }
#' @note \samp{breakaway} presents an estimator of species richness that is
#' well-suited to the high-diversity/microbial setting. However, many microbial
#' datasets display more diversity than the Kemp-type models can permit. In
#' this case, the log-transformed WLRM diversity estimator of Rocchetti et. al.
#' (2011) is returned. The authors' experience suggests that some datasets that
#' require the log-transformed WLRM contain ``false'' diversity, that is,
#' diversity attributable to sequencing errors (via an inflated singleton
#' count). The authors encourage judicious use of diversity estimators when the
#' dataset may contain these errors, and recommend the use of
#' \code{\link{breakaway_nof1}} as an exploratory tool in this case.
#' @author Amy Willis
#' @seealso \code{\link{breakaway_nof1}}; \code{\link{apples}}
#' @references Willis, A. and Bunge, J. (2015). Estimating diversity via
#' frequency ratios. \emph{Biometrics}, \bold{71}(4), 1042--1049.
#'
#' Rocchetti, I., Bunge, J. and Bohning, D. (2011). Population size estimation
#' based upon ratios of recapture probabilities. \emph{Annals of Applied
#' Statistics}, \bold{5}.
#' @keywords diversity microbial models nonlinear
#' @examples
#'
#' breakaway(apples)
#' breakaway(apples, plot = FALSE, output = FALSE, answers = TRUE)
#'
#'
#' @export breakaway
breakaway <- function(my_data, output=TRUE, plot=FALSE, answers=FALSE, force=FALSE, useAll=

  ## read in data
  if( !(is.matrix(my_data) || is.data.frame(my_data))) {
```

Writing R packages: documentation

```
# Generated by roxygen2: do not edit by hand

export(betta)
export(betta_pic)
export(betta_points)
export(betta_random)
export(breakaway)
export(breakaway_clean)
export(breakaway_nof1)
export(build_frequency_count_tables)
export(chao1)
export(chao1_bc)
export(chao_bunge)
export(estimate_alpha_better)
export(frequency_count_or_proportion_or_column)
export(gini)
export(hill)
export(hill_better)
export(hill_pic)
export(inverse_simpson)
export(inverse_simpson_better)
export(make_frequency_count_table)
export(objective_bayes_geometric)
export(objective_bayes_mixedgeo)
export(objective_bayes_negbin)
export(objective_bayes_poisson)
export(plot_alpha)
export(proportions_instead)
export(resample_estimate)
export(rnbinomtable)
export(rztnbinomtable)
export(sample_size_estimate)
export(sample_size_figure)
export(shannon)
export(shannon_better)
export(simpson)
export(subsample_otu)
export(to_proportions)
export(wlrm_transformed)
export(wlrm_untransformed)
import(reshape2)
importFrom(MASS,mvrnorm)
importFrom(ggplot2,element_text)
importFrom(ggplot2,facet_wrap)
importFrom(ggplot2,geom_errorbar)
importFrom(plyr,is.discrete)
importFrom(stats,acf)
```

Building packages

Core "other stuff" for a package

- Code: written by you, goes into `R` folder (no subdirectories allowed)
- `DESCRIPTION`: written by you
- Documentation: autogenerated by `roxygen` comments, goes into `man` folder
- `NAMESPACE`: autogenerated by `roxygen` comments

documentation: vignettes

hello.R | breakaway.Rmd | DESCRIPTION | NAMESPACE | hello.Rd

Knit | Insert | Run

```
---
title: "Getting started with breakaway"
author: "Amy Willis"
date: "`r Sys.Date()`"
output: rmarkdown::html_vignette
vignette: >
  %\VignetteIndexEntry{getting-started}
  %\VignetteEngine{knitr::rmarkdown}
  %\VignetteEncoding{UTF-8}
---

`breakaway` is a package for species richness estimation and modelling. As the pac
grown and users have requested more functionality, it contains some basic generali
the statistical philosophy that underpins `breakaway` to the general alpha diversi
Because of the flexibility of the modelling strategies, most users of breakaway ar
microbial ecologists with very large OTU tables, however, nothing should exclude a
macroecologist from using the same tools. If you have a macroecology dataset and w
this package, I would love to hear from you so please feel free to contact me (ema
Github's Issues/Projects infrastructure).

## Vignette Info

This vignette will lead you through

- loading in your data and covariate information
- creating frequency tables
- using `breakaway` to estimate species richness
- using the objective bayes methods to estimate species richness
- estimating various alpha diversity indices and providing a very basic resample-b
method for ascribing variability to them
- using `betta` to model changes in alpha diversity with covariates

If there is something that you would like explained please feel free to request it

## Loading the package and data

Download the latest version of the package from github.

```{r}
### Run the first two lines at home! ####
# install.packages("devtools")
# devtools::install_github("adw96/breakaway")
library(breakaway)
data(toy_otu_table)
data(toy_metadata)
```
```

documentation: vignettes

277 lines (199 sloc) | 13.3 KB

Raw | Blame | History

title	author	date	output	vignette
Getting started with breakaway	Amy Willis	`r Sys.Date()`	rmarkdown::html_vignette	%\VignetteIndexEntry{getting-started} %\VignetteEngine{knitr::rmarkdown} %\VignetteEncoding{UTF-8}

`breakaway` is a package for species richness estimation and modelling. As the package has grown and users have requested more functionality, it contains some basic generalisations of the statistical philosophy that underpins `breakaway` to the general alpha diversity case. Because of the flexibility of the modelling strategies, most users of breakaway are microbial ecologists with very large OTU tables, however, nothing should exclude a macroecologist from using the same tools. If you have a macroecology dataset and want to use this package, I would love to hear from you so please feel free to contact me (email or via Github's Issues/Projects infrastructure).

Vignette Info

This vignette will lead you through

- loading in your data and covariate information
- creating frequency tables
- using `breakaway` to estimate species richness
- using the objective bayes methods to estimate species richness
- estimating various alpha diversity indices and providing a very basic resample-based method for ascribing variability to them
- using `betta` to model changes in alpha diversity with covariates

If there is something that you would like explained please feel free to request it!

Loading the package and data

Download the latest version of the package from github.

```
### Run the first two lines at home! ####
# install.packages("devtools")
# devtools::install_github("adw96/breakaway")
library(breakaway)
data(toy_otu_table)
data(toy_metadata)

## For historical reasons we going to call them:
otu_data <- toy_otu_table
meta_data <- toy_metadata
```

Other things to keep in mind

1. `testthat` is a H.W. package to write unit tests
2. Datasets take a special format
    - Don't release data in the same way as functions
3. `rm(list=ls(all=TRUE))` removes everything in the global environment
    - But not packages!
4. Consider making custom classes (S3 or S4)
5. `pkgdown` is a H.W. package that can autogenerate a website `build_site()`

Coming up

- New homework: your first R package
- Homework due Tuesday 21 November
  - Good luck with your 512 mid-term!
  - Move office hours from 3 p.m. Tuesday 14 November to 3 p.m. Friday 17 November?
- Next week: scripting + departmental computing with Brian Williamson