Meet the toolkit

Tidyverse & Co.

Filippo Chiarello, Ph.D.

Course toolkit

Doing data science

- Programming:
  - R
  - RStudio
  - tidyverse
  - R Markdown
- Version control and collaboration:
  - Git
  - GitHub

Learning goals

By the end of the module, you will be able to...

- gain insight from data
- gain insight from data, **reproducibly**
- gain insight from data, reproducibly, **using modern programming tools and techniques**
- gain insight from data, reproducibly **and collaboratively**, using modern programming tools and techniques
- gain insight from data, reproducibly and collaboratively, using modern programming tools and techniques, **to answer to key intelligence topics**

Reproducible data analysis

Reproducibility checklist

> What does it mean for a data analysis to be "reproducible"?

Near-term goals:

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear *why* it was done?

Long-term goals:

- Can the code be used for other data?
- Can you extend the code to do other things?

Toolkit for reproducibility

- Scriptability $\rightarrow$ R
- Literate programming (code, narrative, output in one place) $\rightarrow$ R Markdown
- Version control $\rightarrow$ Git / GitHub

R and RStudio

R and RStudio

- R is an open-source statistical **programming language**
- R is also an environment for statistical computing and graphics
- It's easily extensible with *packages*

- RStudio is a convenient interface for R called an **IDE** (integrated development environment), e.g. *"I write R code in the RStudio IDE"*
- RStudio is not a requirement for programming with R, but it's very commonly used by R programmers and data scientists

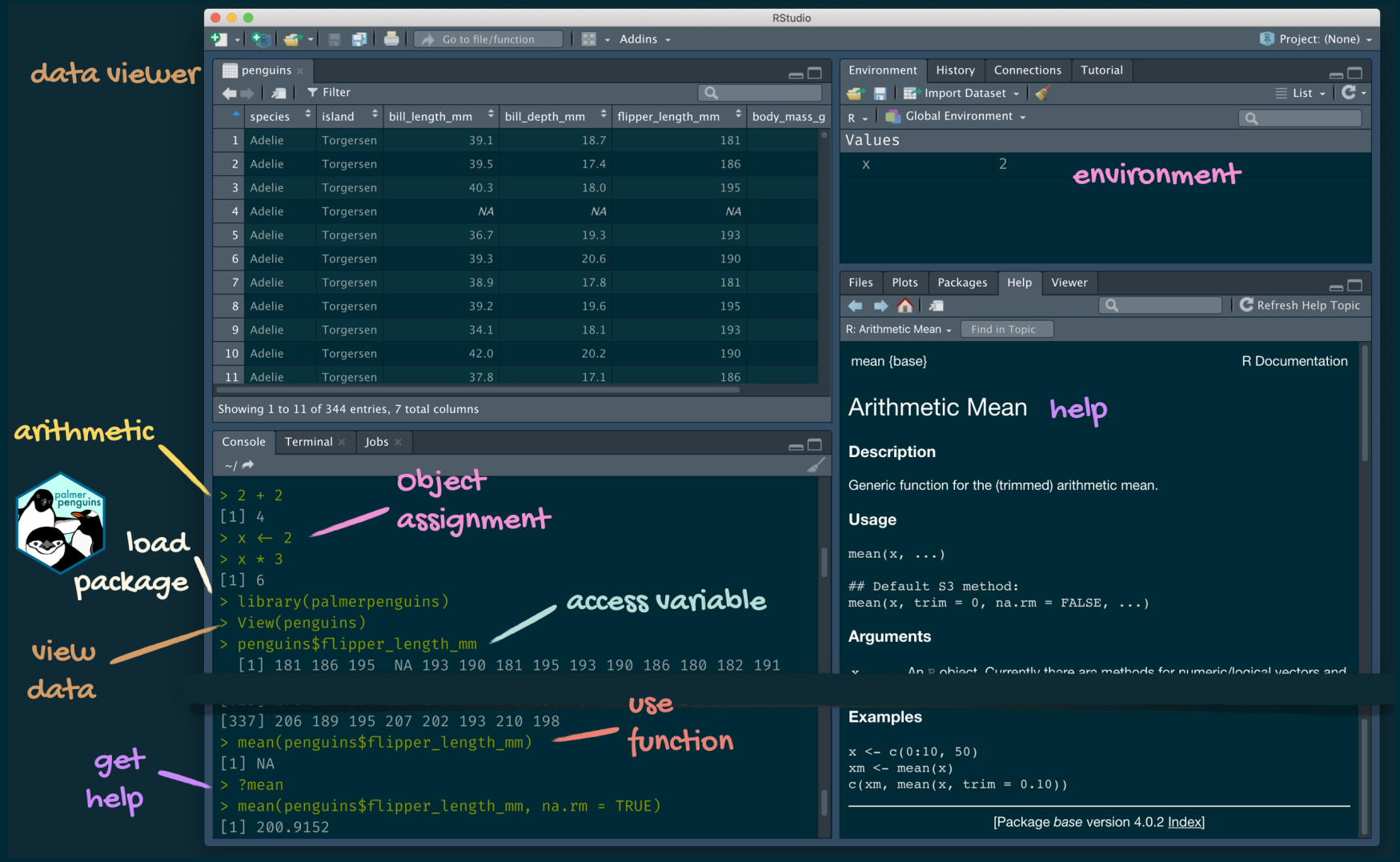
R packages

- **Packages** are the fundamental units of reproducible R code. They include reusable R functions, the documentation that describes how to use them, and sample data[1]
- As of September 2020, there are over 16,000 R packages available on **CRAN** (the Comprehensive R Archive Network)[2]
- We're going to work with a small (but important) subset of these!

[1] Wickham and Bryan, R Packages.

[2] CRAN contributed packages.

Tour: R and RStudio



A short list (for now) of R essentials

- Functions are (most often) verbs, followed by what they will be applied to in parentheses:

```
do_this(to_this)
do_that(to_this, to_that, with_those)
```

- Packages are installed with the `install.packages` function and loaded with the `library` function, once per session:

```
install.packages("package_name")
library(package_name)
```

R essentials (continued)

- Columns (variables) in data frames are accessed with `$`:

```
dataframe$var_name
```

- Object documentation can be accessed with `?`

```
?mean
```

tidyverse

tidyverse.org

- The **tidyverse** is an opinionated collection of R packages designed for data science
- All packages share an underlying philosophy and a common grammar