Online CDS 101 *Introduction to Computational and Data Sciences*

Reproducible toolbox

Reproducible research toolkit

Reproducibility checklist

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

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? (e.g., how were parameter settings chosen?)

Long-term goals:

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

Meet the toolkit

Primary tool: *R*

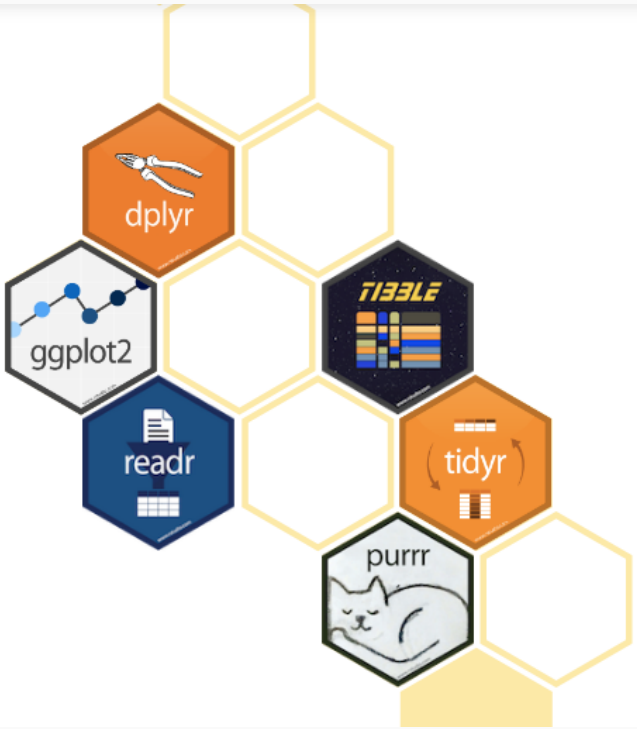
```
side_one <- 3
side_two <- 4
hypotenuse <- sqrt(side_one^2 + side_two^2)
result <- paste(
  "A triangle with sides of length",
  side_one,
  "and length",
  side_two,
  "has a hypotenuse of length",
  hypotenuse
)

print(result)
```

```
## [1] A triangle with sides of
## length 3 and length 4 has a
## hypotenuse of length 5
```

Meet the toolkit

Utility belt: *tidyverse packages*



R packages for data science

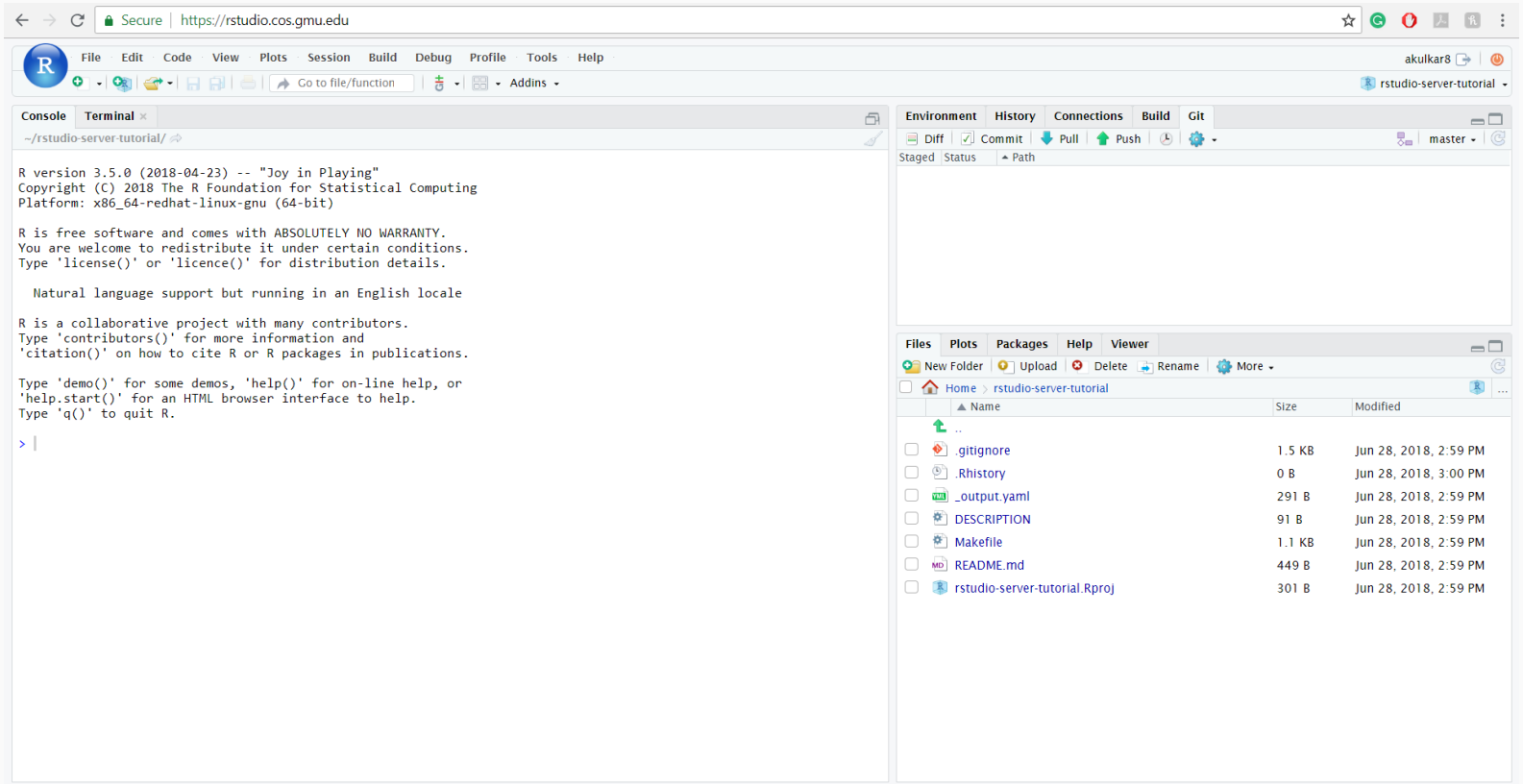
The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

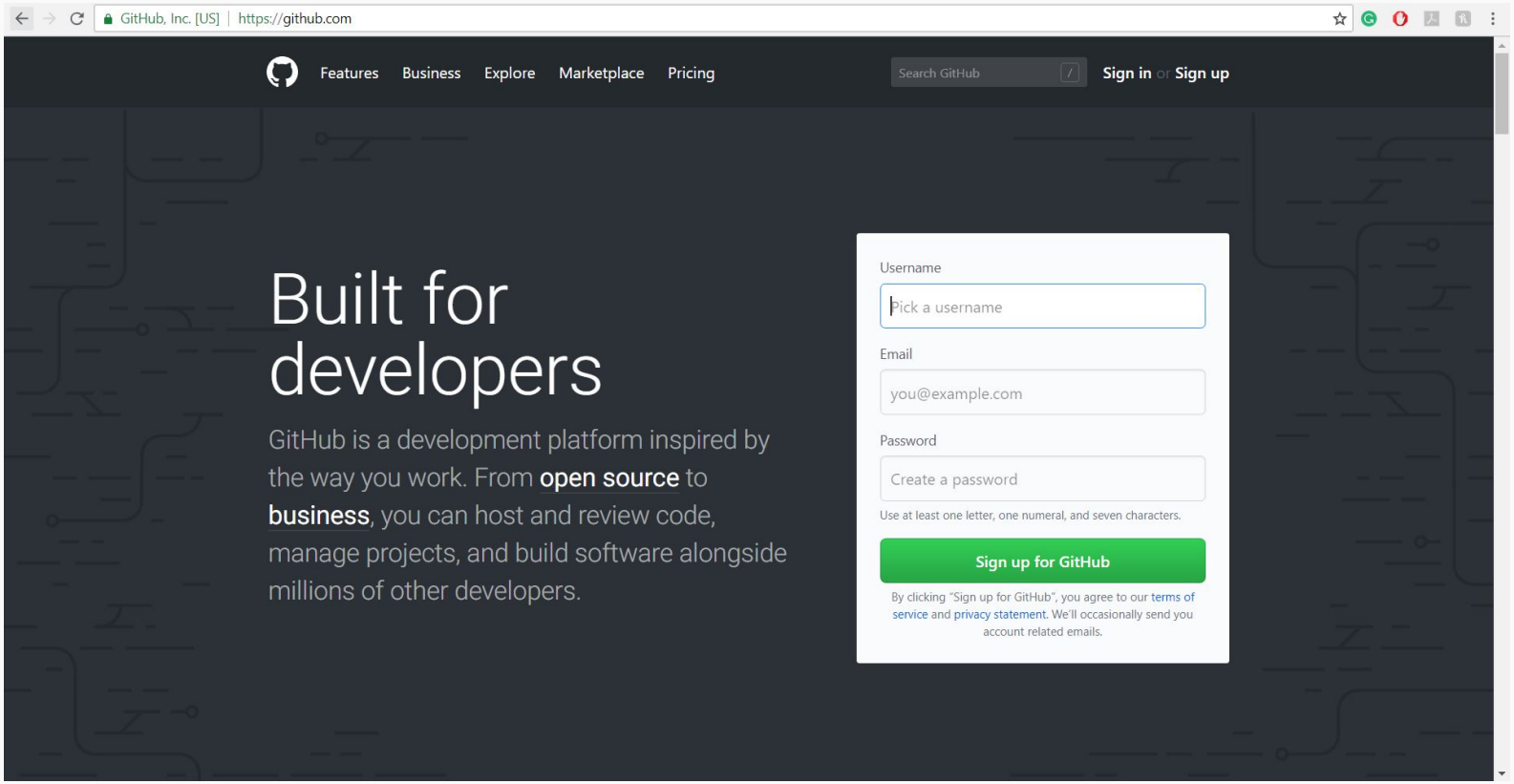
Meet the toolkit

Workshop: *RStudio Server*



Meet the toolkit

Storage room: *GitHub*



Some R history

- The first stable version of R, v1.0.0, was released on February 29, 2000.
- R itself is an implementation of the S programming language, which was designed at Bell Laboratories in the mid-1970s.
- *Base R* was built for statisticians and for doing data analysis, but not necessarily for modern Data Science
- It's age and legacy brings along old implementations of data structures and abbreviated function (commands) names

Source: David Smith, *Over 16 years of R project history*, Revolutions blog, last updated on March 4, 2016, accessed September 20, 2017, http://blog.revolutionanalytics.com/2016/03/16-years-of-r-history.html

Modernizing R with `tidyverse`

Over the last 3 years, chief scientist at RStudio, Hadley Wickham, has brought R into the modern era with the `tidyverse`.

> *The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying philosophy and common APIs.*
>
> — Front page of the Tidyverse website

In practice, this meant reducing everything to a small, core set of commands that all behave in a similar way.

R essentials

A short list (for now):

R essentials

A short list (for now):

- 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)
```

R essentials

A short list (for now):

- 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

A short list (for now):

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

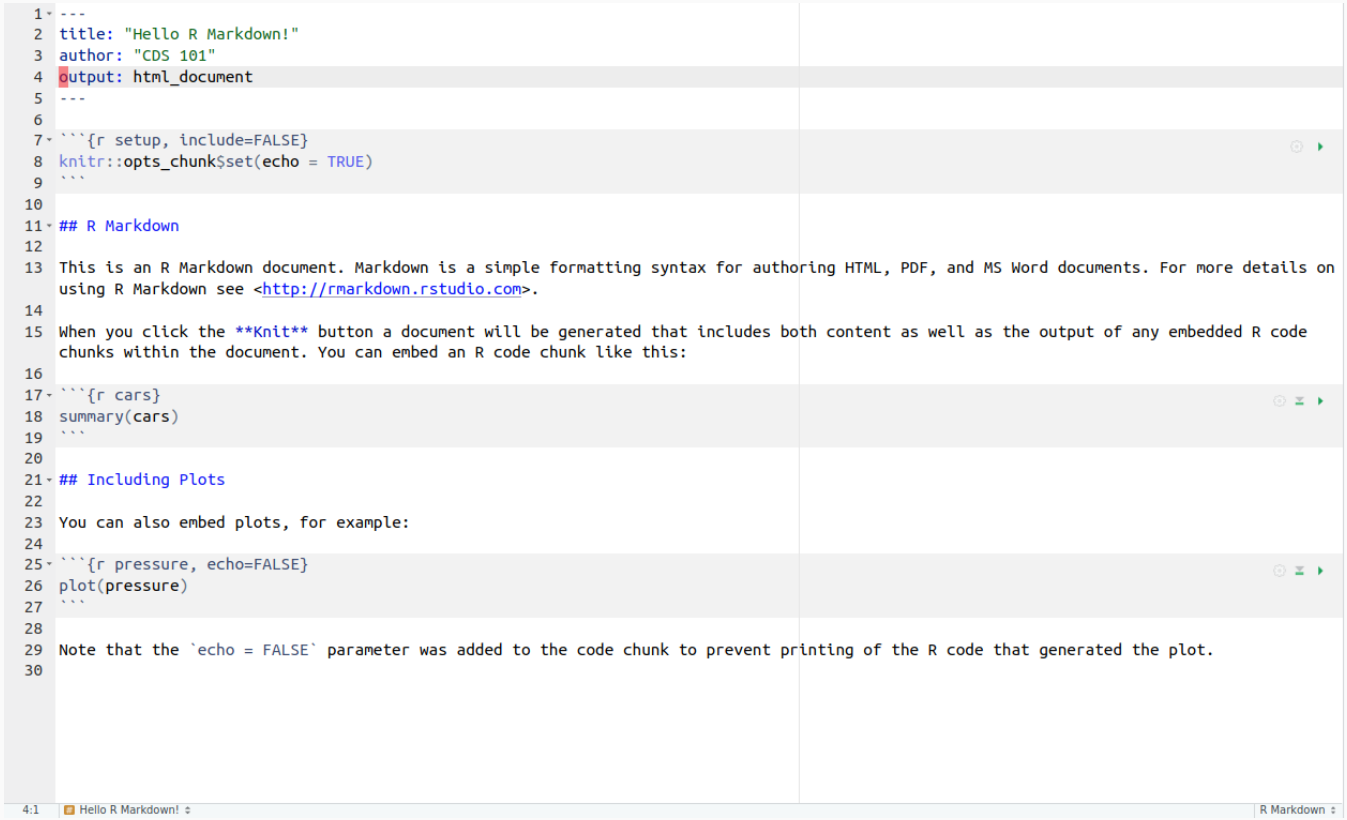
```
do_this(to_this)
do_that(to_this, to_that, with_those)
```

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

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

- Packages bring in additional functions for us to use!

R Markdown



```
---
title: "Hello R Markdown!"
author: "CDS 101"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

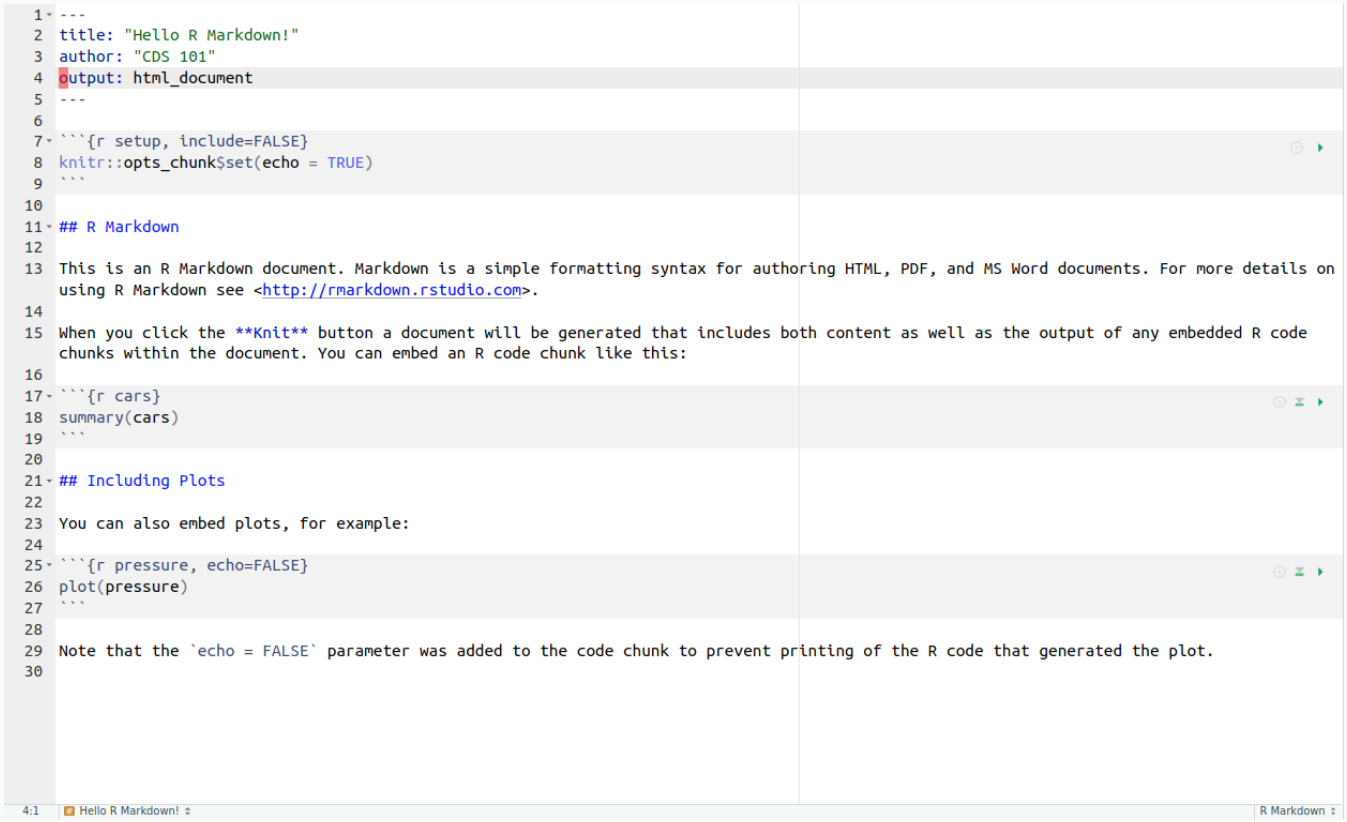
You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

- Fully reproducible reports – each time you knit the analysis is ran from the beginning
- Simple markdown syntax for text
- Code goes in chunks, defined by three back-ticks, narrative goes outside of chunks

R Markdown



```
---
title: "Hello R Markdown!"
author: "CDS 101"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

If it works for you, then it works for me!

How will we use R Markdown?

- You will submit all your homework assignments and the final group project as an R Markdown document
- You will be shown examples in the form of "real-time" demos inside an R Markdown document. While watching, you should create your own document and follow along!
- For assignments, you'll always have a template R Markdown document to start with

Credits

License

Acknowledgments

Content adapted from the Meet the toolkit slides by Mine Çetinkaya-Rundel and made available under the CC BY 4.0 license.