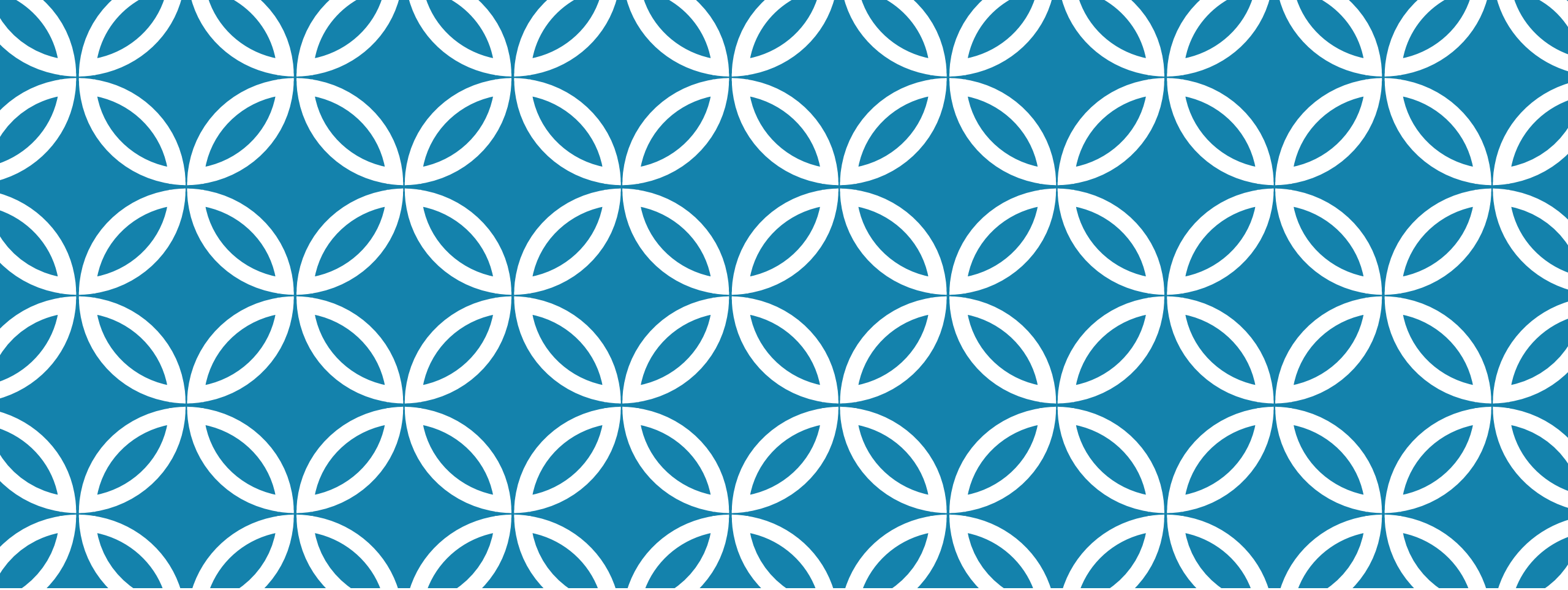


Reproducible Clinical Data Analysis with R and RStudio

Session 2

Getting Data

March 28, 2018

Patient Data

When is it OK to look at patient data?

Patient Authorization not Needed

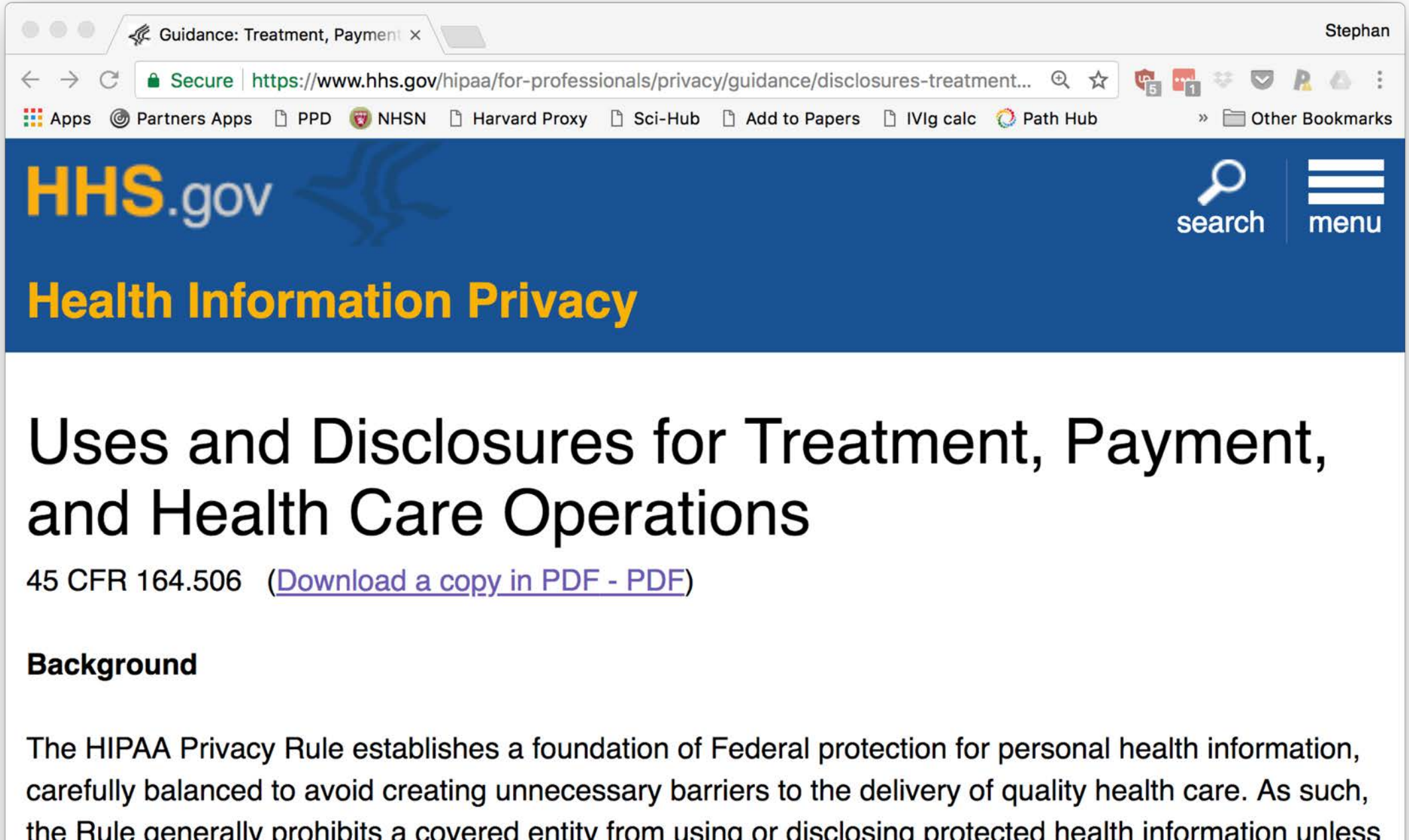
- Treatment
- Payment
- Operations

The Definition of "Healthcare Operations" Explicitly Includes Quality Improvement Activities

Health care operations means any of the following activities of the covered entity to the extent that the activities are related to covered functions:

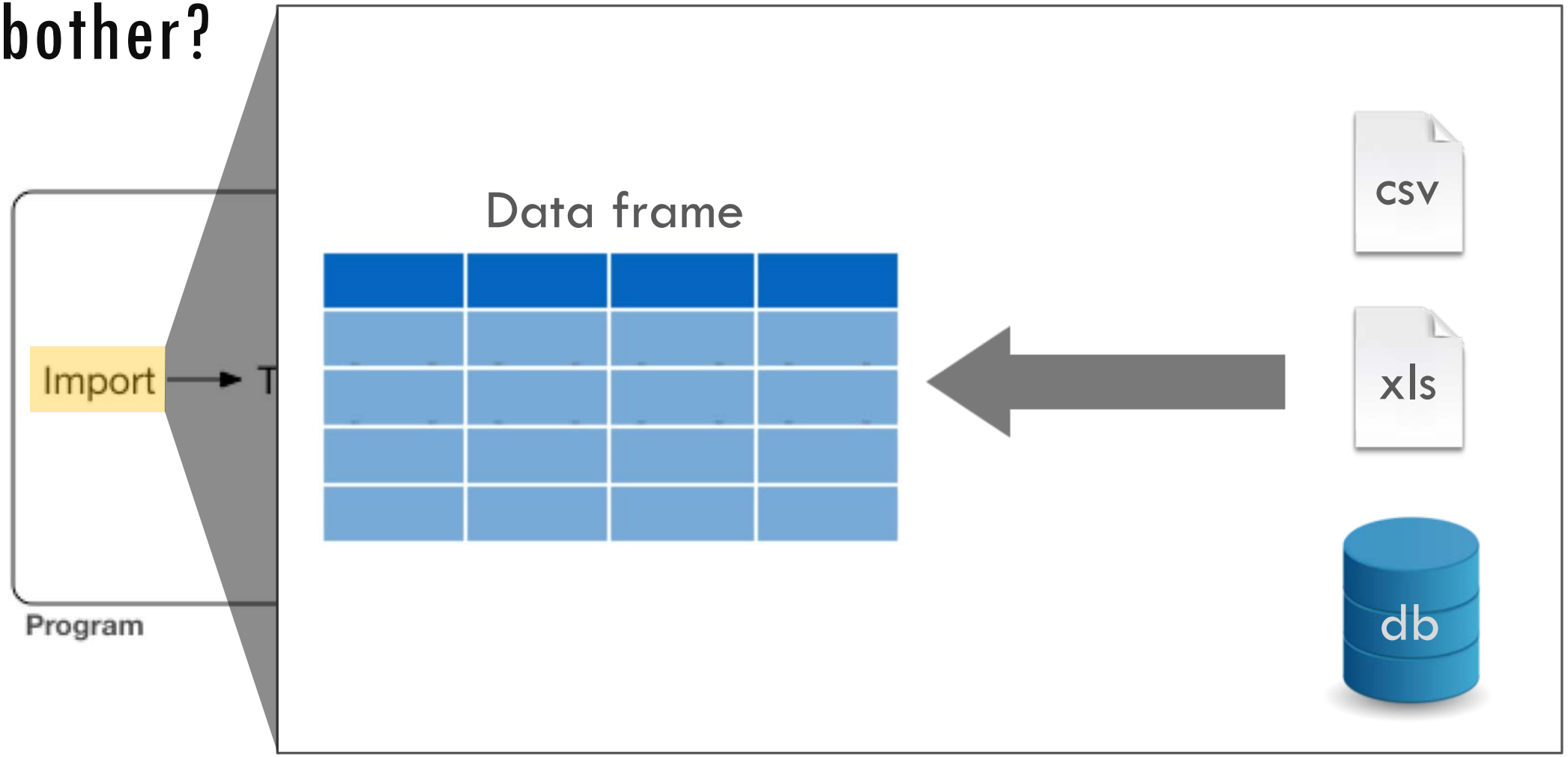
(1) Conducting quality assessment and improvement activities, including outcomes evaluation and development of clinical guidelines, provided that the obtaining of generalizable knowledge is not the primary purpose of any studies resulting from such activities; patient safety activities (as defined in 42 CFR 3.20); population-based activities relating to improving health or reducing health care costs, protocol development, case management and care coordination, contacting of health care providers and patients with information about treatment alternatives; and related functions that do not include treatment;

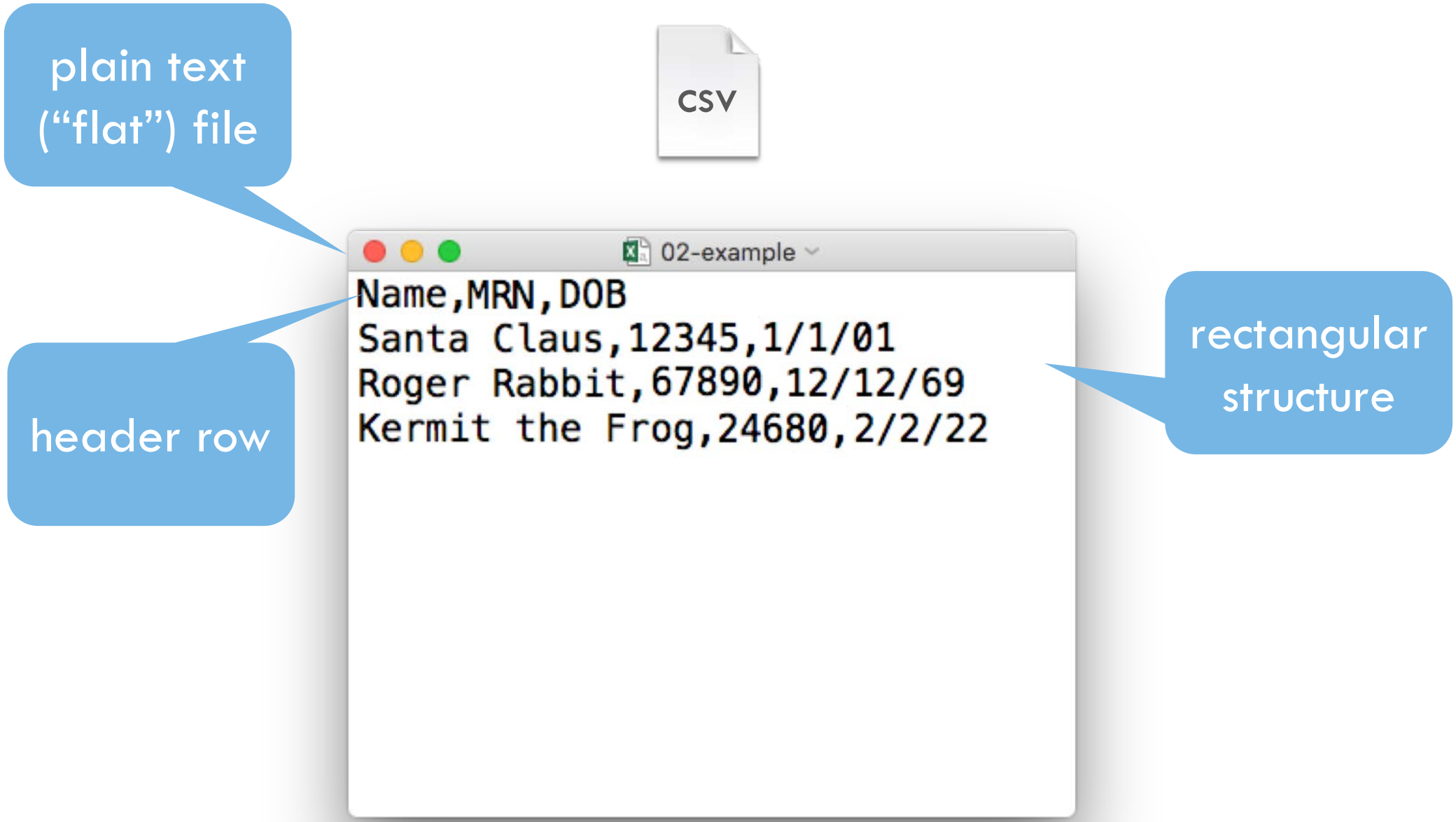
2017 45 CFR § 164.501


Guidance: Treatment, Payment
Stephan
Secure
https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-treatment...
Apps
Partners Apps
PPD
NHSN
Harvard Proxy
Sci-Hub
Add to Papers
IVIg calc
Path Hub
Other Bookmarks
HHS.gov
search
menu
Health Information Privacy
Uses and Disclosures for Treatment, Payment, and Health Care Operations
45 CFR 164.506 (Download a copy in PDF - PDF)
Background
The HIPAA Privacy Rule establishes a foundation of Federal protection for personal health information, carefully balanced to avoid creating unnecessary barriers to the delivery of quality health care. As such, the Rule generally prohibits a covered entity from using or disclosing protected health information unless

Importing Data

Why bother?

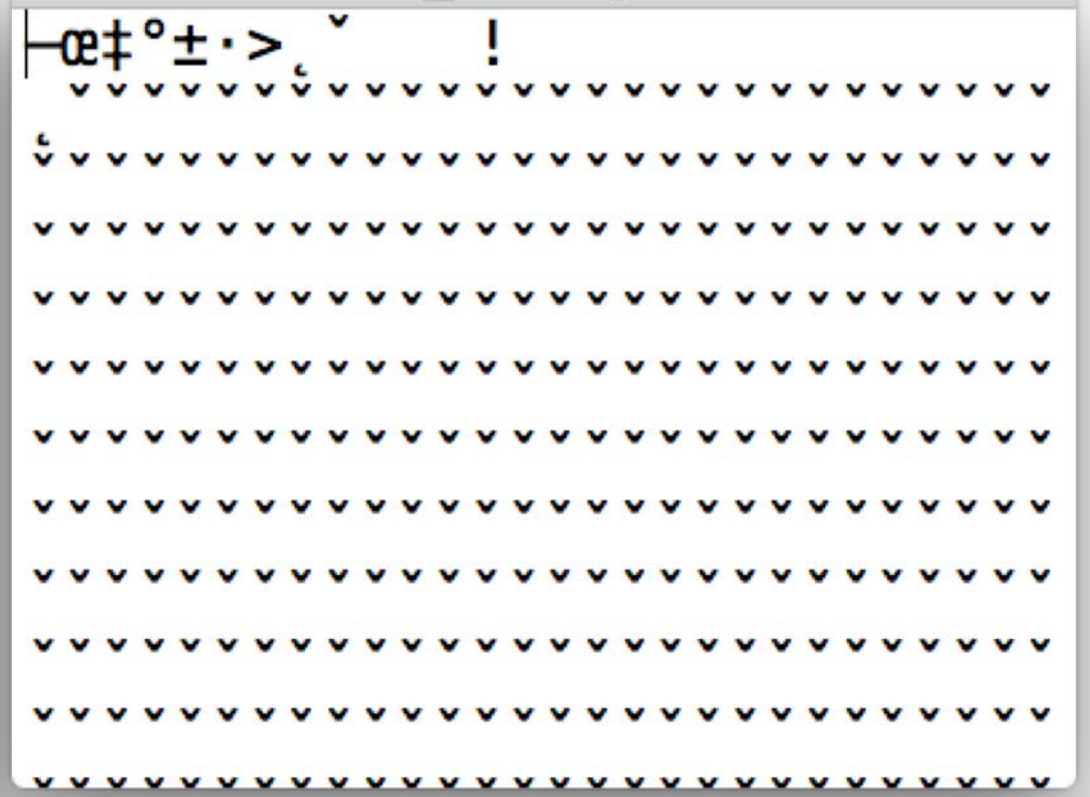



plain text
("flat") file
CSV
02-example
Name,MRN,DOB
Santa Claus,12345,1/1/01
Roger Rabbit,67890,12/12/69
Kermit the Frog,24680,2/2/22
header row
rectangular
structure

02-example

```
Name,MRN,DOB
Santa Claus,12345,1/1/01
Roger Rabbit,67890,12/12/69
Kermit the Frog,24680,2/2/22
```

02-example



Import flat rectangular data files (.csv, .tsv, ...)

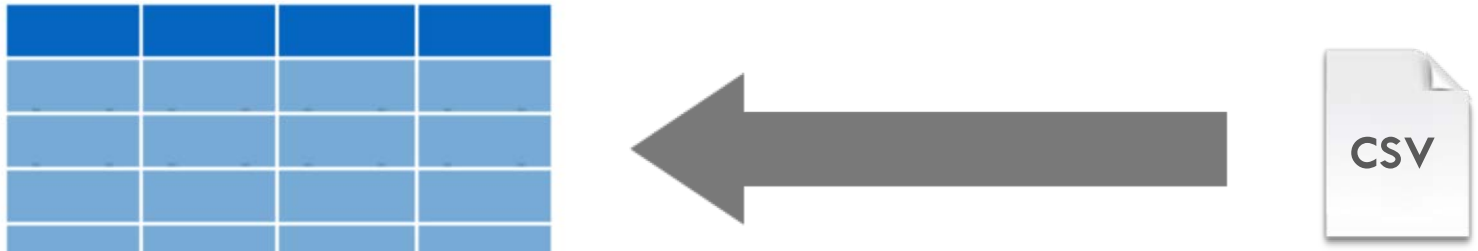
```
library(readr)
```

Alternatively,

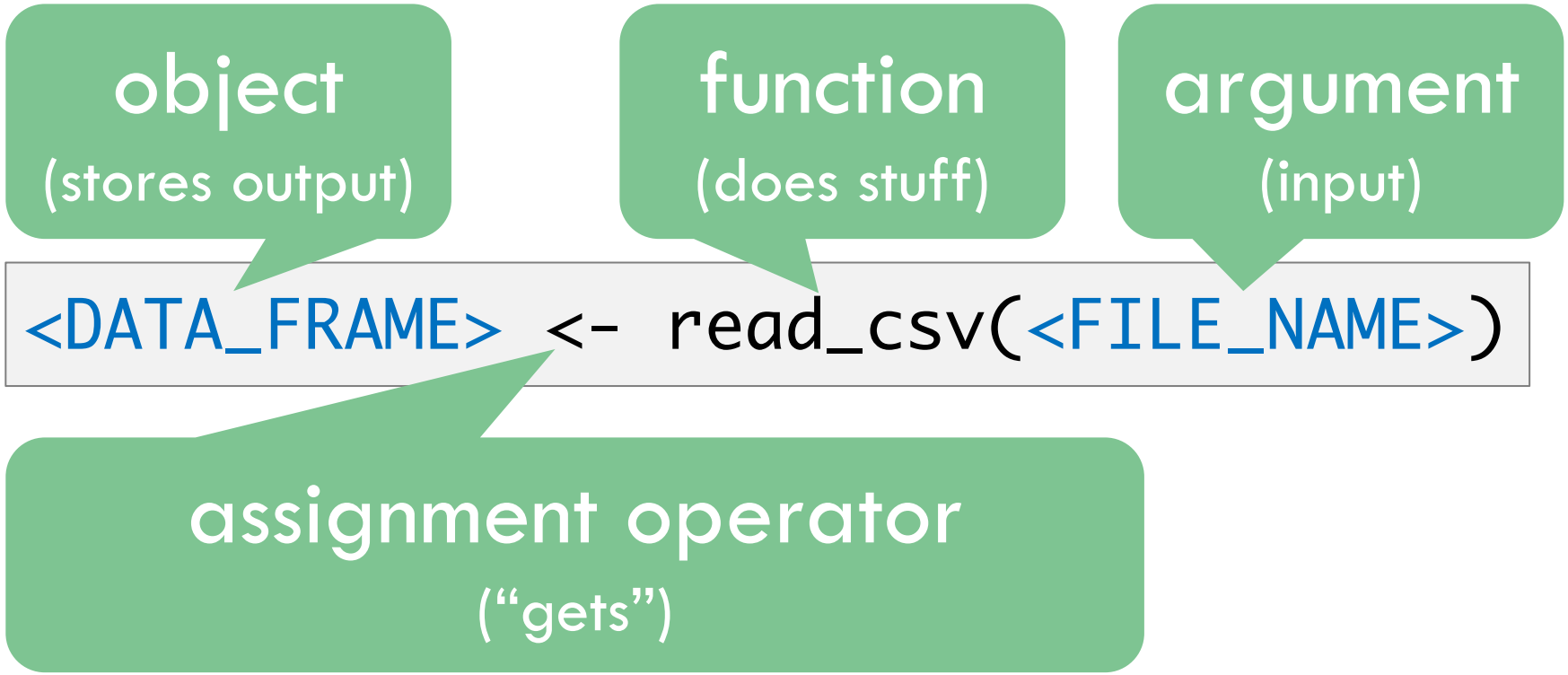
```
library(tidyverse)
```

read_csv()

```
<DATA_FRAME> <- read_csv(<FILE_NAME>)
```



Aside: R Syntax Basics

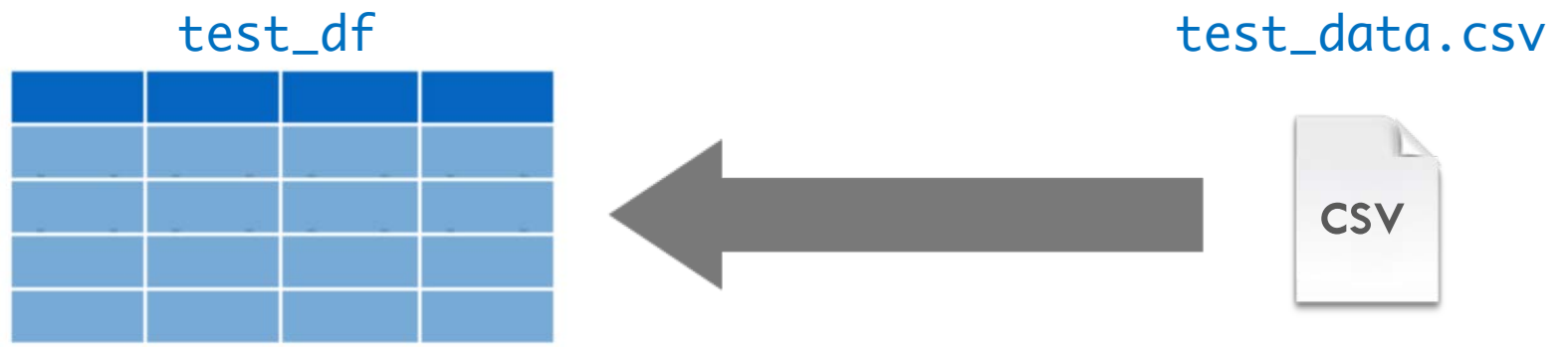


read_csv()

data frame to read data into

name of CSV file

```
test_df <- read_csv("test_data.csv")
```



readr

Your Turn #1

Download the files I sent to you by e-mail:

1. `02-getting-data.Rmd`
2. `esr_data.csv`

Open `02-getting-data.Rmd` in RStudio. Work through the section "Your Turn #1".

Then complete the sections "Patient Data" and "Importing Data" on Handout 2

Import Excel files
(.xls, .xlsx)

```
library(readxl)
```

read_excel()

data frame to read data into

name of Excel file

```
<DATA_FRAME> <- read_excel(<FILE_NAME>)
```

How to Get Help When You Get Stuck

Your Turn #2

Suppose you want to import data from a file that's similar to a CSV but uses pipes (|) instead of commas (,) as the delimiter between data fields.

Discuss with your group mates how you could figure out how to do this.

The Basic Approach to Getting Un-Stuck

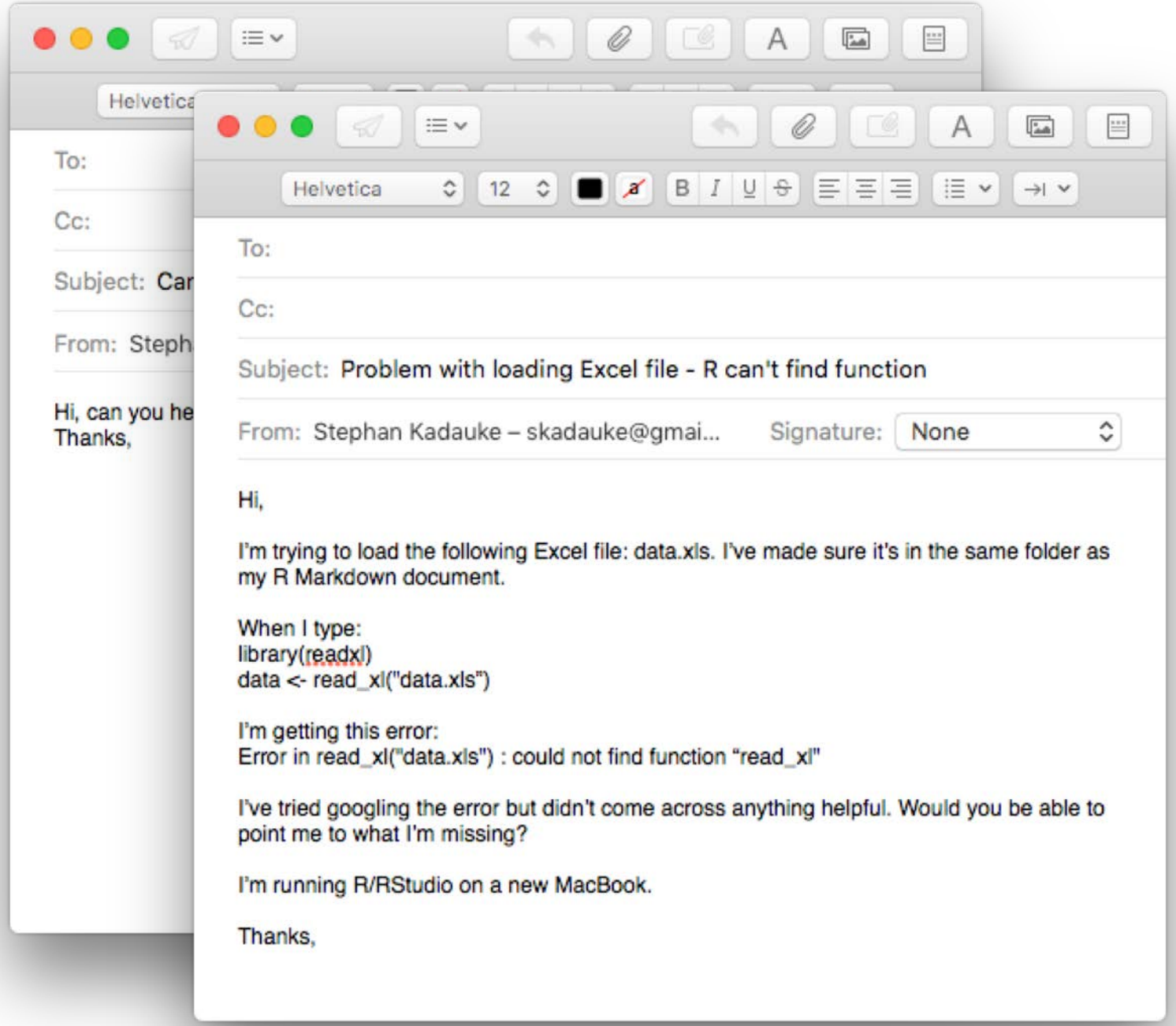
Spend a limited amount of time actively looking for the solution

Ask for help

Scenario 1: Unexpected Result or Error Message

- Check "spelling"
  - Missing quotes, commas, parentheses
  - Misspelled function names
- Read the error message and try to make sense of it
- Google the error message (cut + paste)
- Ask for help using the following formula:

> **I tried to accomplish** (goal) **so I tried this code**: …
> **I expected this**: …
> **But instead I got this result/error**: …
> **I'm running R/RStudio on** (operating system).


Helvetica
To:
Cc:
Subject: Car
From: Steph
Hi, can you he
Thanks,
Helvetica
12
To:
Cc:
Subject: Problem with loading Excel file - R can't find function
From: Stephan Kadauke – skadauke@gmai...
Signature: None
Hi,
I'm trying to load the following Excel file: data.xls. I've made sure it's in the same folder as my R Markdown document.
When I type:
library(readxl)
data <- read_xl("data.xls")
I'm getting this error:
Error in read_xl("data.xls") : could not find function "read_xl"
I've tried googling the error but didn't come across anything helpful. Would you be able to point me to what I'm missing?
I'm running R/RStudio on a new MacBook.
Thanks,

Scenario 2: You Know What You Want to Do But Don't Know How to Do it in R

- If you think you know what function to use:

```
?function
```

- RStudio Cheat Sheets
- Google
  - Extremely powerful
  - But there's an art to this
  - Less useful for very basic questions
- Ask for help
- Do a tutorial

Cheatsheets – RStudio

Stephan

Secure | https://www.rstudio.com/resources/cheatsheets/

Apps | Partners Apps | PPD | NHSN | Harvard Proxy | Sci-Hub | Add to Papers | Other Bookmarks

RStudio

Products | Resources | Pricing | About Us | Blogs

RStudio Cheat Sheets

The cheat sheets below make it easy to learn about and use some of our favorite packages. From time to time, we will add new cheat sheets to the gallery. If you'd like us to drop you an email when we do, let us know by clicking the button to the right.

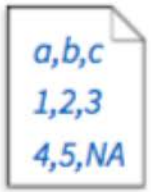
SUBSCRIBE TO CHEAT SHEET UPDATES HERE

Deep Learning with Keras Cheat Sheet

Keras is a high-level neural networks API developed with a focus on enabling fast experimentation. Keras supports both convolution based networks and recurrent networks (as

Read Tabular Data - These functions share the common arguments:

read_*(file, col_names = TRUE, col_types = NULL, locale = default_locale(), na = c("", "NA"), quoted_na = TRUE, comment = "", trim_ws = TRUE, skip = 0, n_max = Inf, guess_max = min(1000, n_max), progress = interactive())



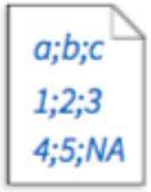
A	B	C
1	2	3
4	5	NA

Comma Delimited Files

read_csv("file.csv")

To make file.csv run:

write_file(x = "a,b,c\n1,2,3\n4,5,NA", path = "file.csv")

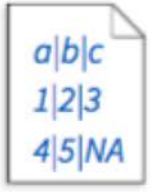


A	B	C
1	2	3
4	5	NA

Semi-colon Delimited Files

read_csv2("file2.csv")

write_file(x = "a;b;c\n1;2;3\n4;5;NA", path = "file2.csv")

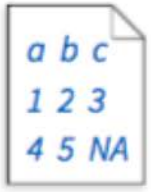


A	B	C
1	2	3
4	5	NA

Files with Any Delimiter

read_delim("file.txt", delim = "|")

write_file(x = "a|b|c\n1|2|3\n4|5|NA", path = "file.txt")



A	B	C
1	2	3
4	5	NA

Fixed Width Files

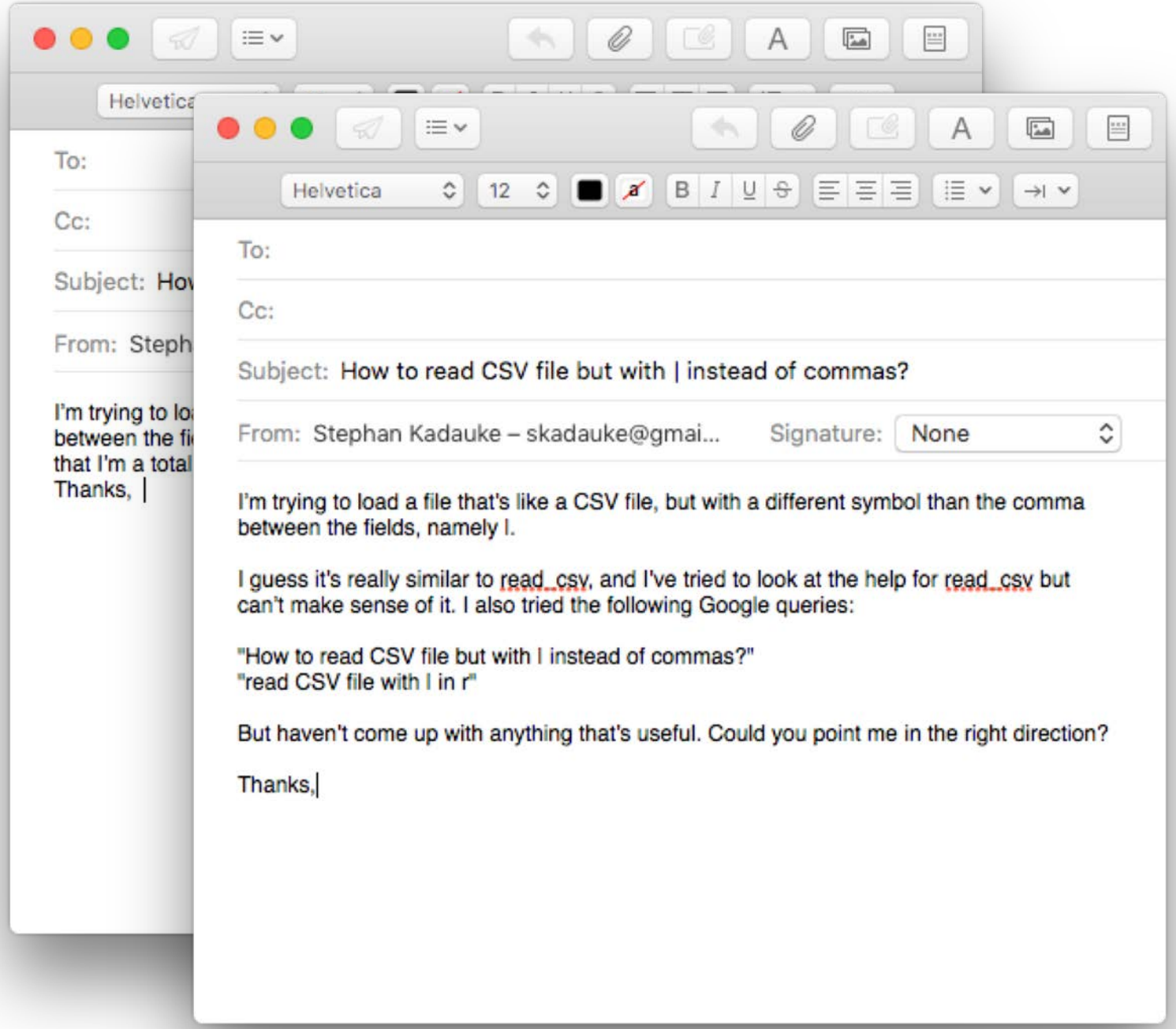
read_fwf("file.fwf", col_positions = c(1, 3, 5))

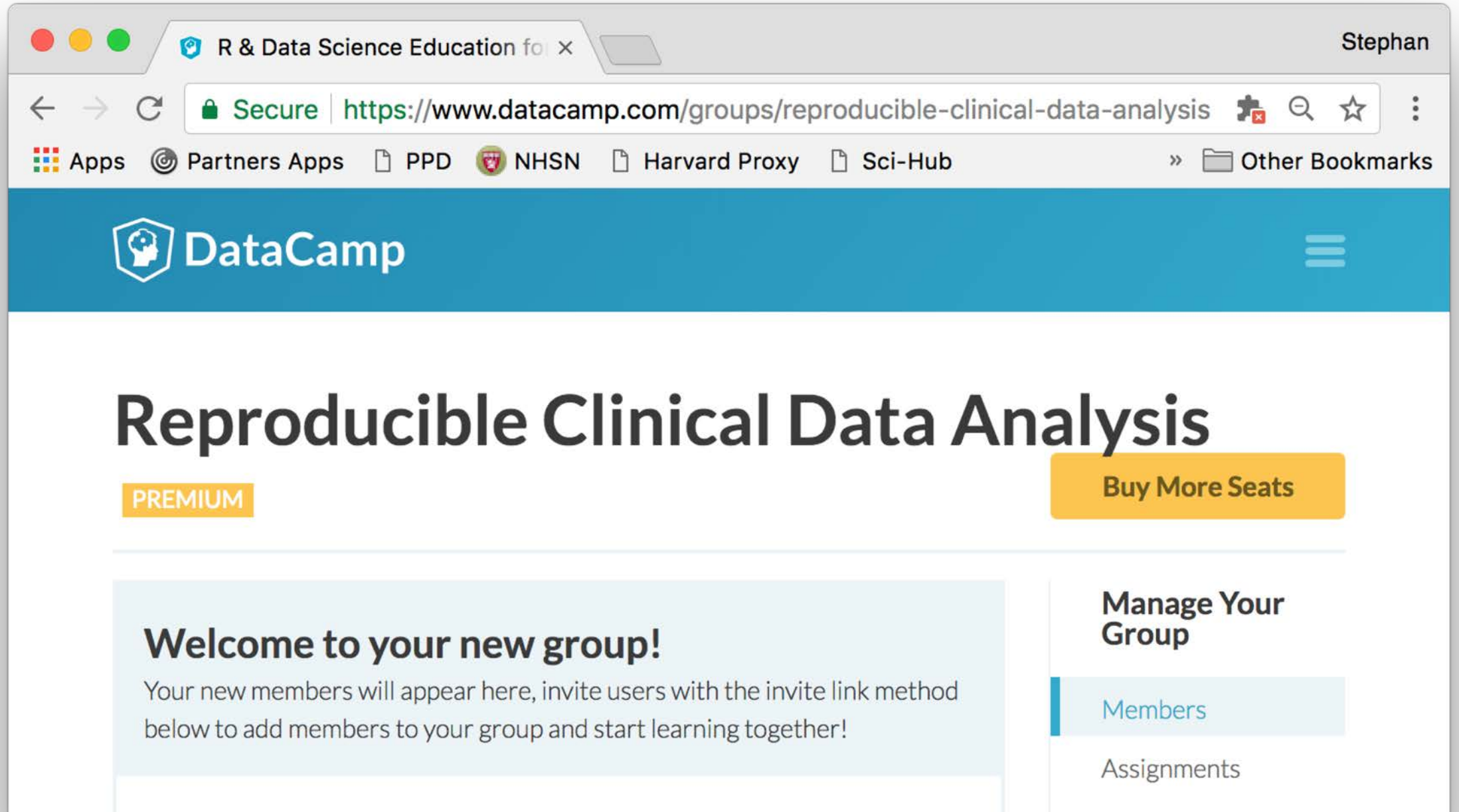
write_file(x = "a b c\n1 2 3\n4 5 NA", path = "file.fwf")

Tab Delimited Files

read_tsv("file.tsv") Also **read_table().**

write_file(x = "a\tb\tc\n1\t2\t3\n4\t5\tNA", path = "file.tsv")


Helvetica
To:
Cc:
Subject: How
From: Steph
I'm trying to lo
between the fi
that I'm a total
Thanks,
Helvetica
12
To:
Cc:
Subject: How to read CSV file but with | instead of commas?
From: Stephan Kadauke – skadauke@gmai...
Signature: None
I'm trying to load a file that's like a CSV file, but with a different symbol than the comma between the fields, namely |.
I guess it's really similar to read_csv, and I've tried to look at the help for read_csv but can't make sense of it. I also tried the following Google queries:
"How to read CSV file but with | instead of commas?"
"read CSV file with | in r"
But haven't come up with anything that's useful. Could you point me in the right direction?
Thanks,


R & Data Science Education fo
Stephan
Secure
https://www.datacamp.com/groups/reproducible-clinical-data-analysis
Apps
Partners Apps
PPD
NHSN
Harvard Proxy
Sci-Hub
Other Bookmarks
DataCamp
Reproducible Clinical Data Analysis
PREMIUM
Buy More Seats
Welcome to your new group!
Your new members will appear here, invite users with the invite link method below to add members to your group and start learning together!
Manage Your Group
Members
Assignments

Your Turn #3

Complete the section “Getting Help” on Handout 2.

Databases

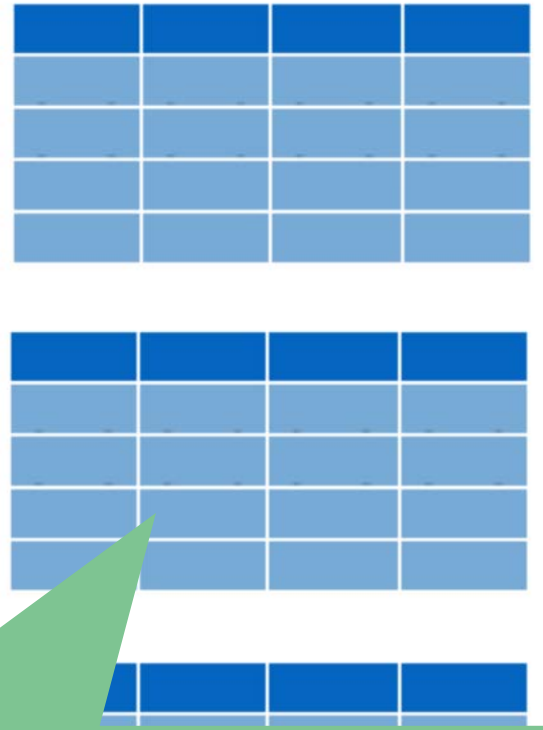
Database Server

PHSSQL2057
("Datamart")

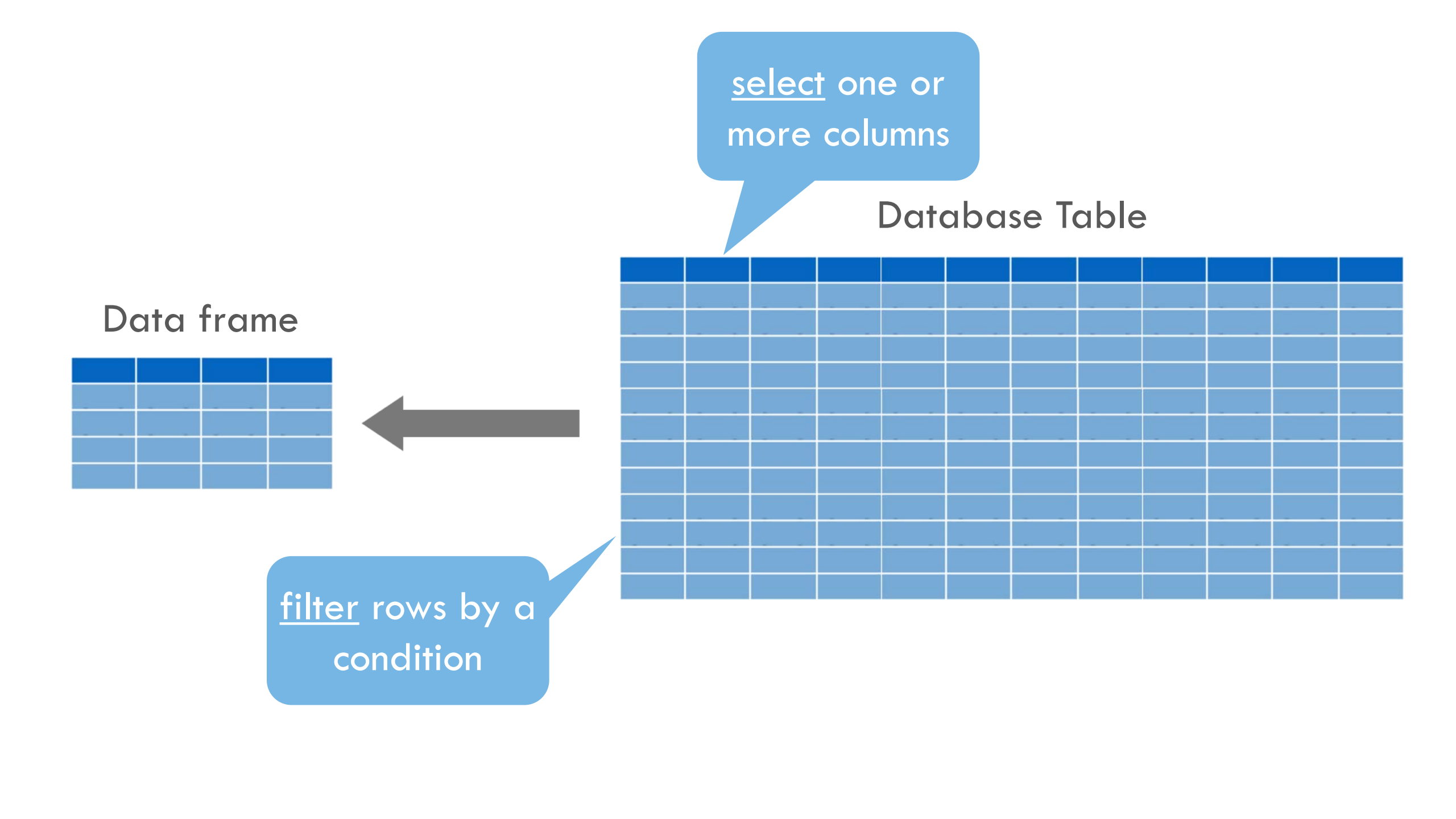
Database

MGHLABUTIL

Database Tables



MGHLABUTIL_LabResults


select one or more columns
Database Table
Data frame
filter rows by a condition

Databases: Best Practices

- Know and follow the rules of permissible use
- Let the database server do as much of the computation as possible
- While you develop with a database query, limit results

Your Turn #4

Work through the section "Your Turn #4" of `02-getting-data.Rmd`.

Then complete the section "Databases" on Handout 2