

DATA IMPORT

Jeff Goldsmith, PhD
Department of Biostatistics

Data wrangling

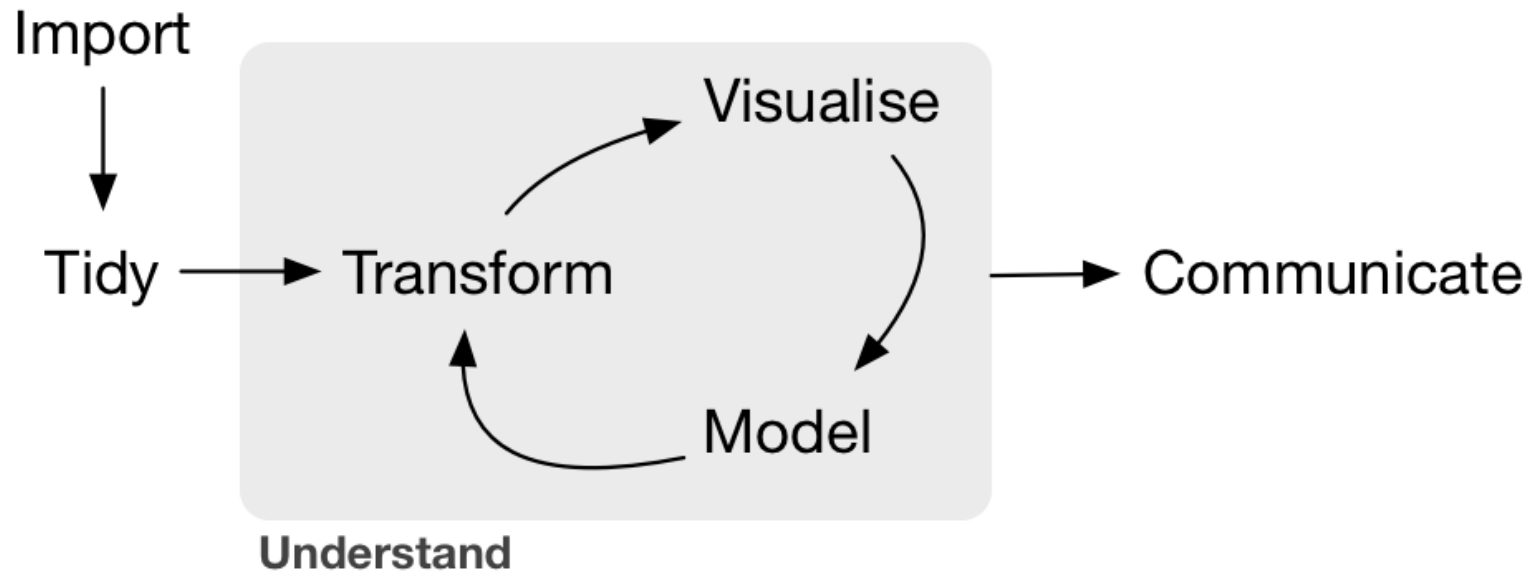
- Data don't magically appear in your R session
- They're rarely even in the form you need
- The process of taking data in whatever form they exist and transforming them to the form you need is “wrangling”

You're going to have to wrangle

- Call it what you want – there really isn't a way around the need to load, organize, and transform data
- If you expect someone to do this for you, that person will also do the rest of your job

Import

- “Import” is the first step to “wrangle”



R for Data Science

Data tables

- Data often come in tables
 - Row = subject
 - Column = variable
- The variables may be of different types
- In R, `data.frames` are designed to hold this kind of dataset
 - Looks like a matrix
 - Actually a very specific list

Tibbles

... formerly tbl_df ...



Jenny Bryan

@JennyBryan

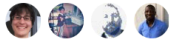
Follow



#dplyr dilemma: I know `%>%` is pronounced “then”. How do we say `tbl_df`? data.frame just rolls off the tongue by comparison.

2:48 AM - 24 Sep 2014 from West Point Grey, Vancouver

1 Retweet 3 Likes



3 1 3



Tweet your reply



Kevin Markham @justmarkham · 24 Sep 2014



Replying to @JennyBryan

@JennyBryan Technically it's called a "local data frame", which is still a bit long though! :)

1



Jenny Bryan @JennyBryan · 24 Sep 2014



@justmarkham @KevinUshey I went with “tibble diff” and mostly kept straight face.

1 2



Hilary Parker @hspter · 26 Sep 2014



@JennyBryan @justmarkham @KevinUshey in my head it's "table-diff"

1



Kara Woo @kara_woo · 26 Sep 2014



@hspter @JennyBryan @justmarkham @KevinUshey "table-dee-eff" for me

2 2



Jenny Bryan @JennyBryan · 26 Sep 2014



@kara_woo @hspter @justmarkham @KevinUshey how about “table frame”?

1



Jenny Bryan
@JennyBryan

Follow



#dplyr dilemma: I know `%>%` is
pronounced
data.frame
comparison

2:48 AM - 24 Sep 2014

1 Retweet 3



3



Kevin Ushey

Replying to @JennyBryan

@JennyBryan Technically it's called a "local data frame", which is still a bit long though! :)



1



Jenny Bryan @JennyBryan · 24 Sep 2014

@justmarkham @KevinUshey I went with "tibble diff" and mostly kept straight face.



1



2



Hilary Parker @hspter · 26 Sep 2014

@JennyBryan @justmarkham @KevinUshey in my head it's "table-diff"



1



Kara Woo @kara_woo · 26 Sep 2014

@hspter @JennyBryan @justmarkham @KevinUshey "table-dee-eff" for me



2



2



Jenny Bryan @JennyBryan · 26 Sep 2014

@kara_woo @hspter @justmarkham @KevinUshey how about "table frame"?



1



Hadley Wickham ✓
@hadleywickham

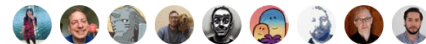
Follow



PSA: I formally approve @hspter's suggested
pronunciation of tbl_df: "tibble diff" #rstats

11:08 AM - 20 Oct 2014

9 Retweets 28 Likes



6



9



28





Jenny Bryan

@JennyBryan

Follow



#dplyr dilemma: I know `%>%` is

pronounced

data.frame

comparison

2:48 AM - 24

1 Retweet 3

3



Hadley Wickham

@hadleywickham

Follow



PSA: I form
pronounced

11:08 AM - 20 Oct 2

9 Retweets 28 Likes

6

9



Hilary Parker @hspter · 20 Oct 2014

Replying to @hadleywickham

@hadleywickham excellent! Although I feel bad as I think this was originally @JennyBryan's idea.. I had previously said "table diff"

1



1



Jenny Bryan @JennyBryan · 20 Oct 2014

@hspter @hadleywickham I tweeted the question but I think @KevinUshey proposed "tibble diff", an historic moment

Jenny Bryan @JennyBryan

#dplyr dilemma: I know `%>%` is pronounced "then". How do we say `tbl_df`? data.frame just rolls off the tongue by comparison.



4



Kev

Replying to @JennyBryan

@JennyBryan Technically it's called "tibble" though! :)

1



Jenny Bryan @JennyBryan · 24 Sep 2014

@justmarkham @KevinUshey I was a bit of a face.

1



2



Hilary Parker @hspter · 26 Sep 2014

@JennyBryan @justmarkham @KevinUshey in my head it's "table-diff"

1



Kara Woo @kara_woo · 26 Sep 2014

@hspter @JennyBryan @justmarkham @KevinUshey "table-dee-eff" for me

2



2



Jenny Bryan @JennyBryan · 26 Sep 2014

@kara_woo @hspter @justmarkham @KevinUshey how about "table frame"?

1



1





Jenny Bryan
@JennyBryan

Follow

#dplyr dilemma: I know `%>%` is pronounced data.frame compatible

2:48 AM - 24 Sep 2014

1 Retweet 3

3



Hadley Wickham ✓
@hadleywickham

Follow

PSA: I form
pronounced

11:08 AM - 20 Oct 2014

9 Retweets 28 Likes

6 9



Hilary Parker @hspter · 20 Oct 2014

Replying to @hadleywickham

@hadleywickham excellent! Although I feel bad as I think this was originally @JennyBryan's idea.. I had previously said "table diff"



Jenny Bryan @hspter · 24 Sep 2014

Jenny Bryan

#dplyr
tbl

tibble 1.0.0

Hadley Wickham

2016-03-24

Categories: [Packages](#) [tidyverse](#)

I'm pleased to announce tibble, a new package for manipulating and printing data frames in R. Tibbles are a modern reimagining of the data.frame, keeping what time has proven to be effective, and throwing out what is not. The name comes from dplyr: originally you created these objects with `tbl_df()`, which was most easily pronounced as "tibble diff".



Kevin Ushey @KevinUshey · 24 Sep 2014
Replying to @JennyBryan
@JennyBryan Technically it's called "table" though! :)

1 2



Jenny Bryan @JennyBryan · 24 Sep 2014
@justmarkham @KevinUshey I went with "table" face.

1 2



Hilary Parker @hspter · 26 Sep 2014
@JennyBryan @justmarkham @KevinUshey in my head it's "table"

1 2



Kara Woo @kara_woo · 26 Sep 2014

@hspter @JennyBryan @justmarkham @KevinUshey "table-dee-eff" for me

2 2



Jenny Bryan @JennyBryan · 26 Sep 2014

@kara_woo @hspter @justmarkham @KevinUshey how about "table frame"?

1



Jenny Bryan
@JennyBryan

Follow

#dplyr dilemma: I know `%>%` is pronounced data.frame, but I don't want to compare

2:48 AM - 24

1 Retweet 3

3



Hadley Wickham ✓
@hadleywickham

Follow

PSA: I form
pronounced

11:08 AM - 20 Oct 2

9 Retweets 28 Likes

6

9



Hilary Parker @hspter · 20 Oct 2014

Replying to @hadleywickham

@hadleywickham excellent! Although I feel bad as I think this was originally @JennyBryan's idea.. I had previously said "table diff"



Jenn
@hspter
propo

Jen

#d

`tbl

I'm pleased to announce tibble, a new package
modern reimagining of the data.frame, keeping
The name comes from dplyr: originally you cre
pronounced as "tibble diff".



Kev

Replying to @JennyBryan

@JennyBryan Technically it's called
though! :)

1

↺

♥



Jenny Bryan @JennyBryan · 24 S
@justmarkham @KevinUshey I we
face.

1

↺

♥ 2



Hilary Parker @hspter · 26 Sep 20
@JennyBryan @justmarkham @KevinUshey in my head it's "tak

1

↺

♥

✉



Kara Woo @kara_woo · 26 Sep 2014

@hspter @JennyBryan @justmarkham @KevinUshey "table-dee-eff" for me

2

↺

♥ 2

✉



Jenny Bryan @JennyBryan · 26 Sep 2014

@kara_woo @hspter @justmarkham @KevinUshey how about "table frame"?

1

↺

♥ 1

✉

tibble 1.0.0

Hadley

Ca



www.rstudio.com

is a
that is not.



Jenny Bryan
@JennyBryan

Follow

#dplyr dilemma: I know `%>%` is pronounced data.frame, but I don't want to use the word data.frame in my package name.

2:48 AM - 24 Oct 2014

1 Retweet 3 Likes

3



Hadley Wickham ✓
@hadleywickham

Follow

PSA: I form pronounced

11:08 AM - 20 Oct 2014

9 Retweets 28 Likes

6 9



Hilary Parker @hspter · 20 Oct 2014

Replying to @hadleywickham

@hadleywickham excellent! Although I feel bad as I think this was originally @JennyBryan's idea.. I had previously said "table diff"



Jenny Bryan
@hspter
proposed



Kevin Ushey @KevinUshey · 24 Sep 2014
Replying to @JennyBryan
@JennyBryan Technically it's called data.frame though! :)

1 1



Jenny Bryan @JennyBryan · 24 Sep 2014
@justmarkham @KevinUshey I was thinking of using the word data.frame.

1 2



Hilary Parker @hspter · 26 Sep 2014
@JennyBryan @justmarkham @KevinUshey in my head it's "table diff"

1 1



Kara Woo @kara_woo · 26 Sep 2014
@hspter @JennyBryan @justmarkham @KevinUshey "table-dee-eff" for me

2 2



Jenny Bryan @JennyBryan · 26 Sep 2014
@kara_woo @hspter @justmarkham @KevinUshey how about "table frame"?

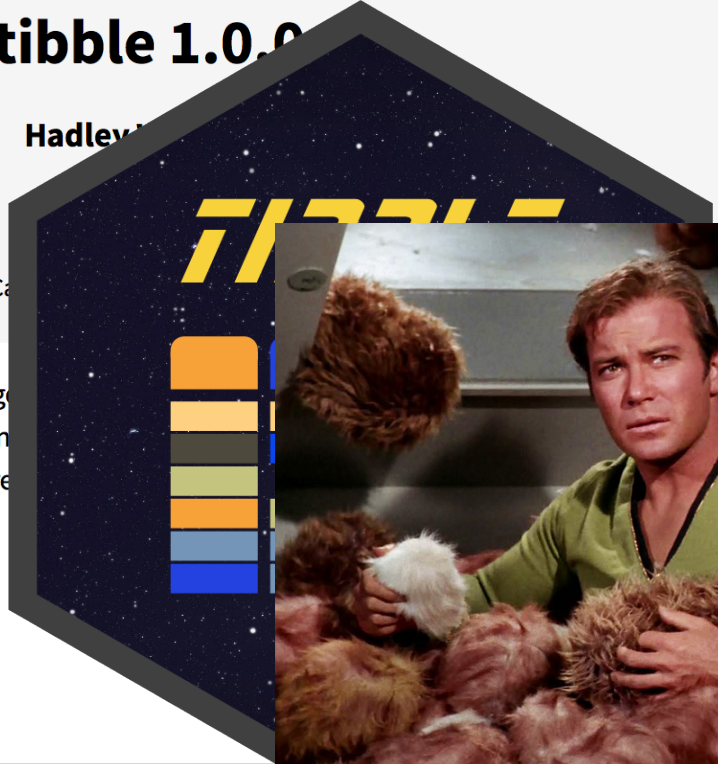
1

tibble 1.0.0

Hadley Wickham

Can

I'm pleased to announce tibble, a new package for modern reimagining of the data.frame, keeping the name as close to data.frame as possible. The name comes from dplyr: originally you create a data.frame, and then you use dplyr to manipulate it. The name comes from dplyr: originally you create a data.frame, and then you use dplyr to manipulate it. The name comes from dplyr: originally you create a data.frame, and then you use dplyr to manipulate it.



Why tibbles?

- `data.frames` have been around since R was introduced
- Some things change; base R is not one of those things
- Tibbles are data frames, just slightly different
 - They keep you from printing everything by accident
 - They make you type complete variable names

80/20 applies to data import

- Most data import is “easy”; the few hard cases will take up a lot of time
- You still have to learn to handle the easy cases
 - readr, haven, readxl
 - Parsing columns can be helpful
 - Watch out for inconsistencies in columns
 - Be sure you know what missing data looks like



“Raw” data

- You generally want the least-processed version of the data possible
- This gives you the ability to transform the data yourself
- This does not mean you are less likely to make mistakes in cleaning data than someone else
 - Your mistakes should be transparent
 - Fixing them shouldn't hurt your analysis pipeline
- Cleaning data is also how you really get to know it