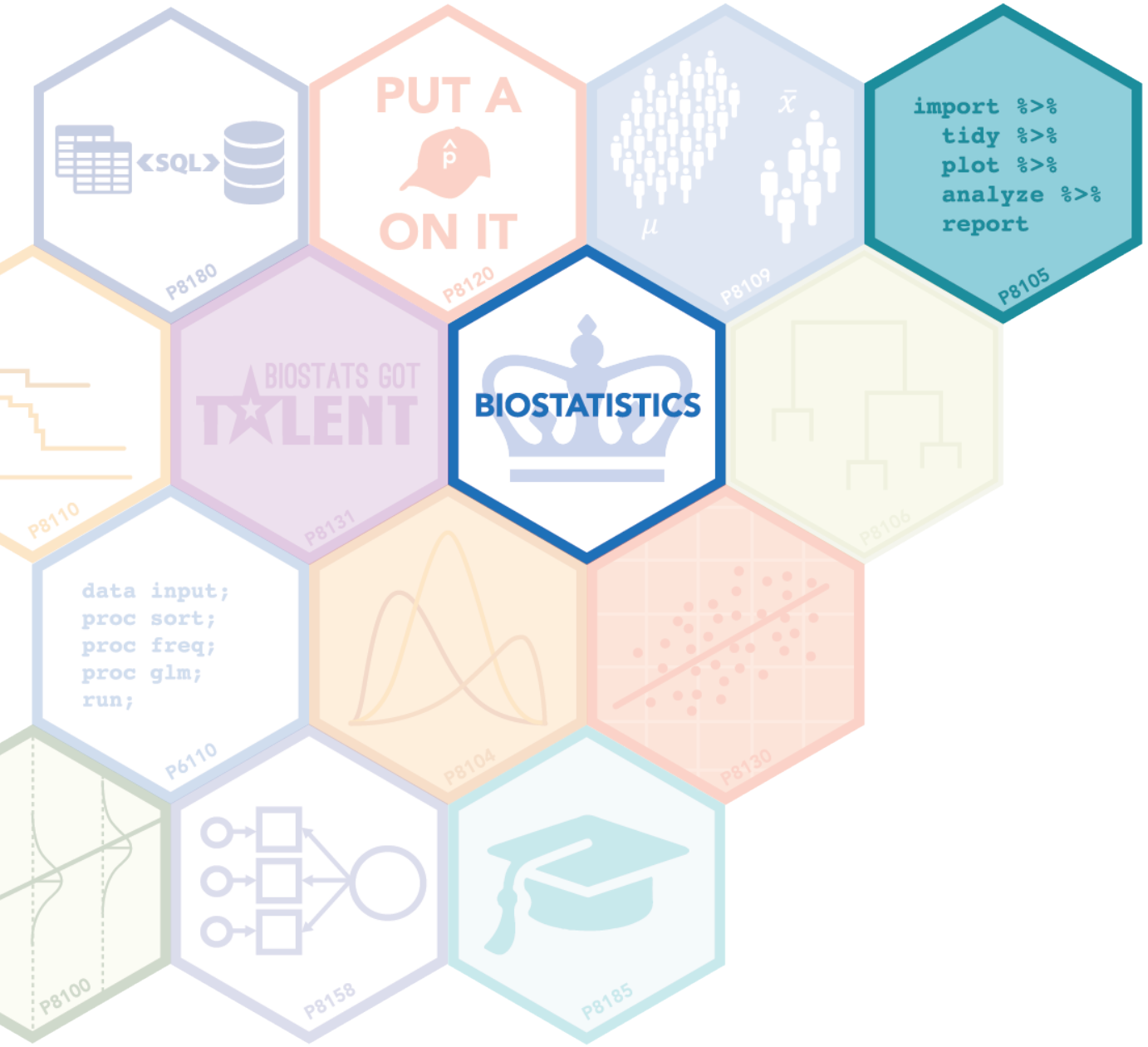



# DATA IMPORT AND MANIPULATION

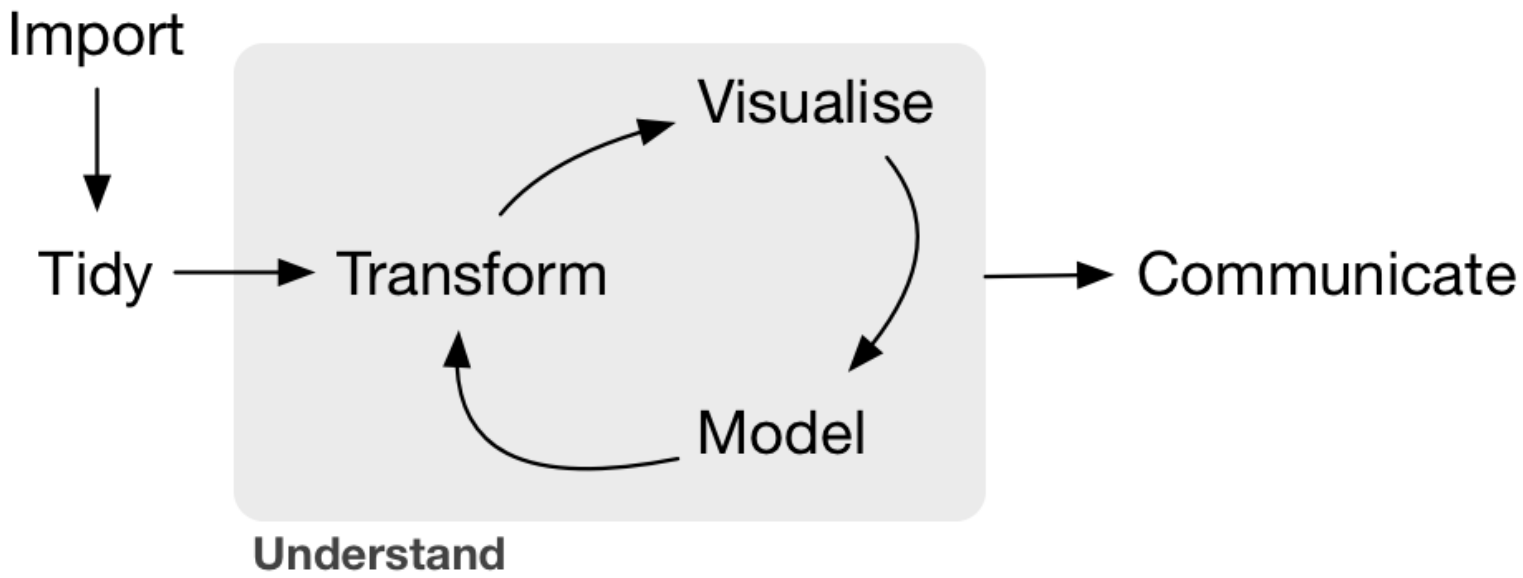
Jeff Goldsmith, PhD
Department of Biostatistics

# 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"

# Import

- "Import" is the first step to "wrangle"




R for Data Science

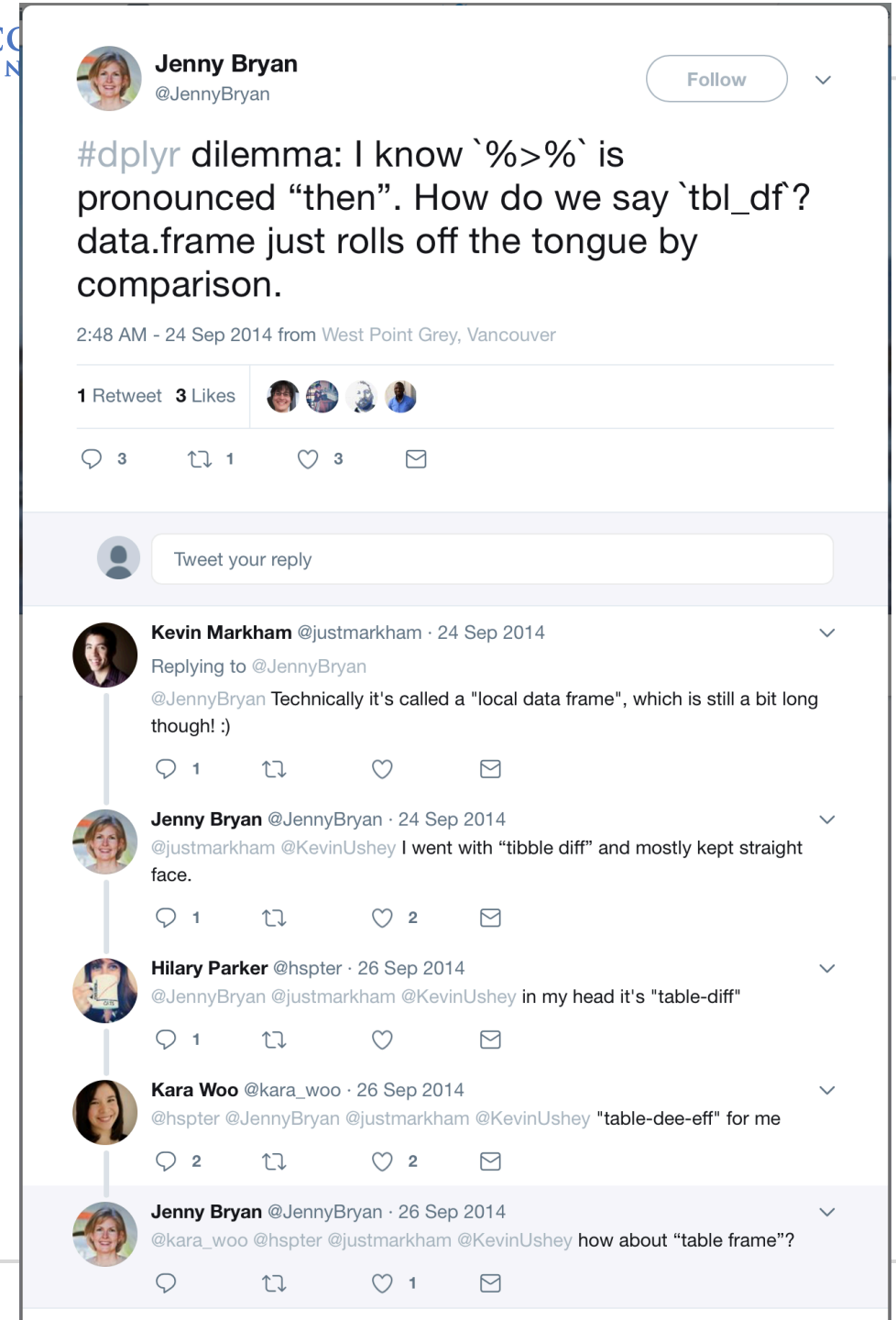
# Data tables

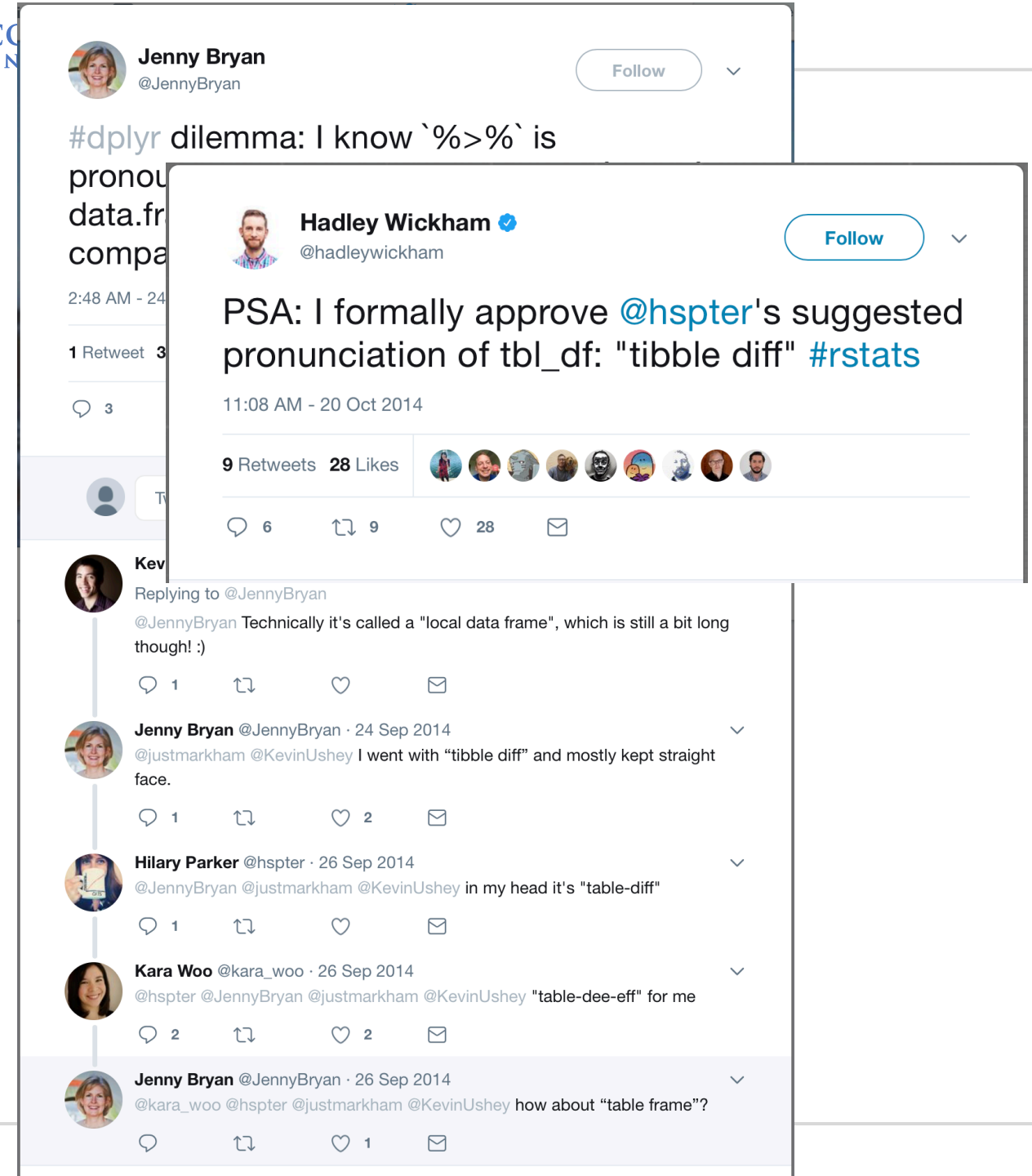
- Data often come in tables
  - Row = subject
  - Column = variable

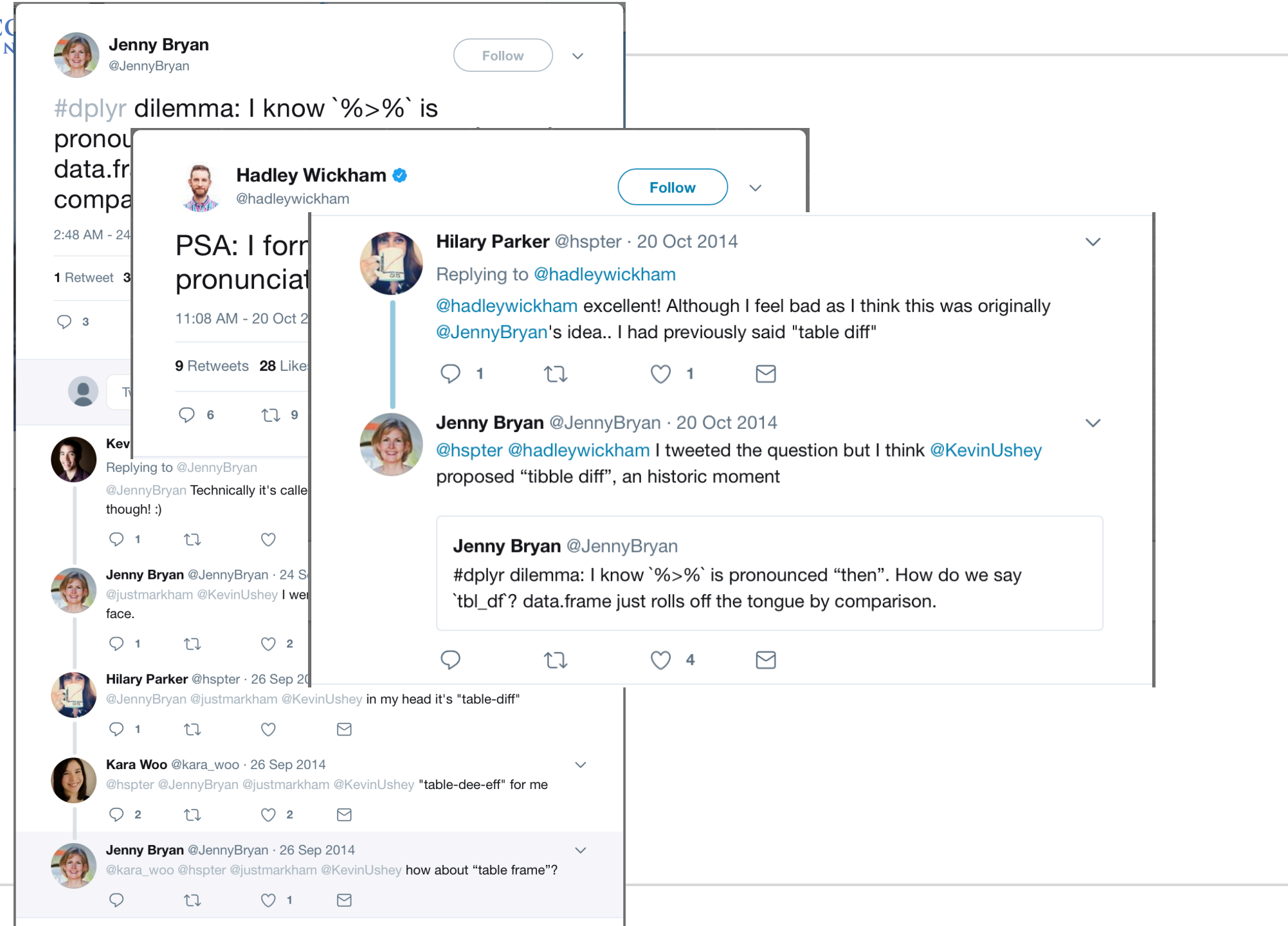
- The variables may be of different types
- In R, `data.frames` are designed to hold this kind of dataset
  - Looks like a matrix
  - Actually a very specific list

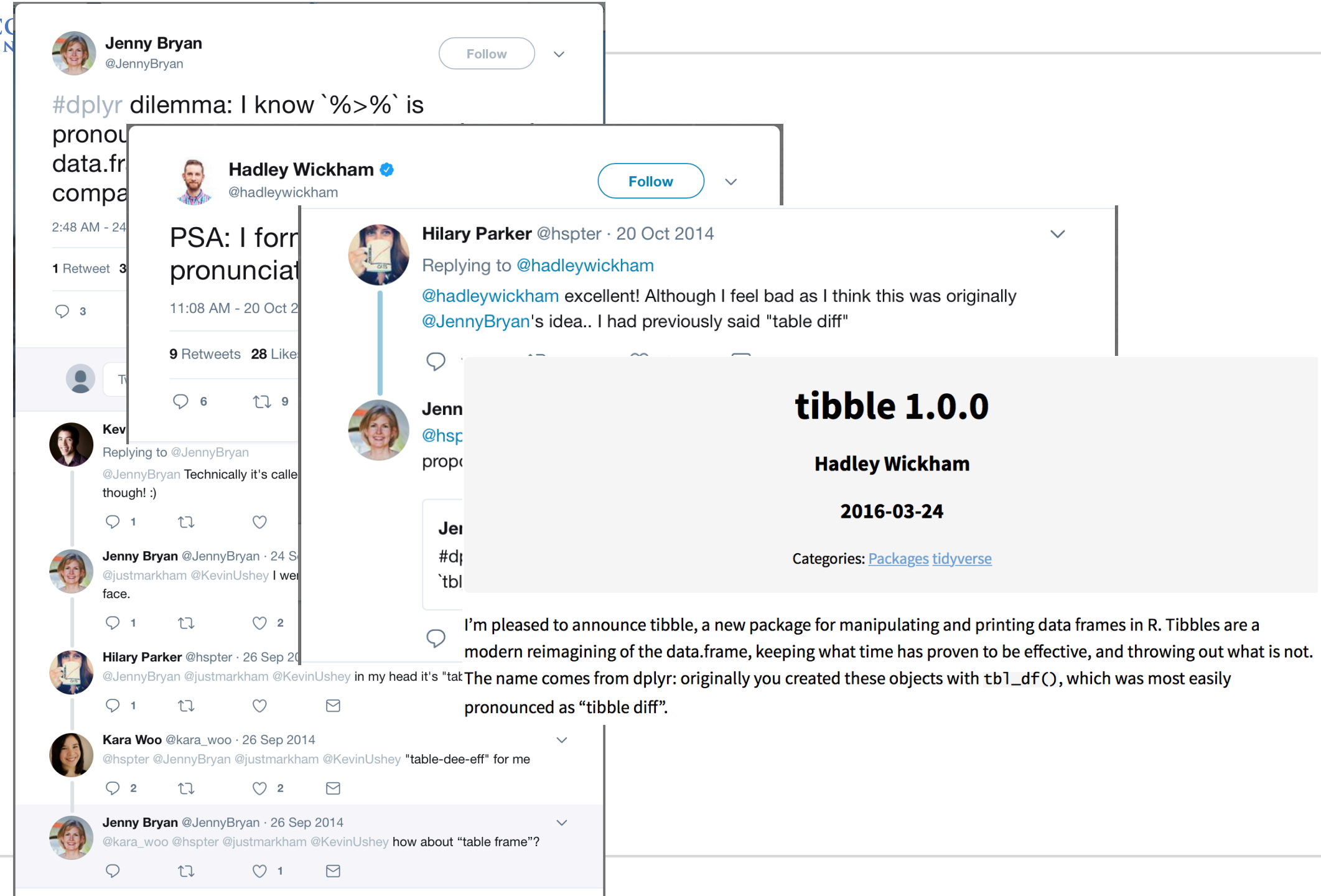
# Tibbles

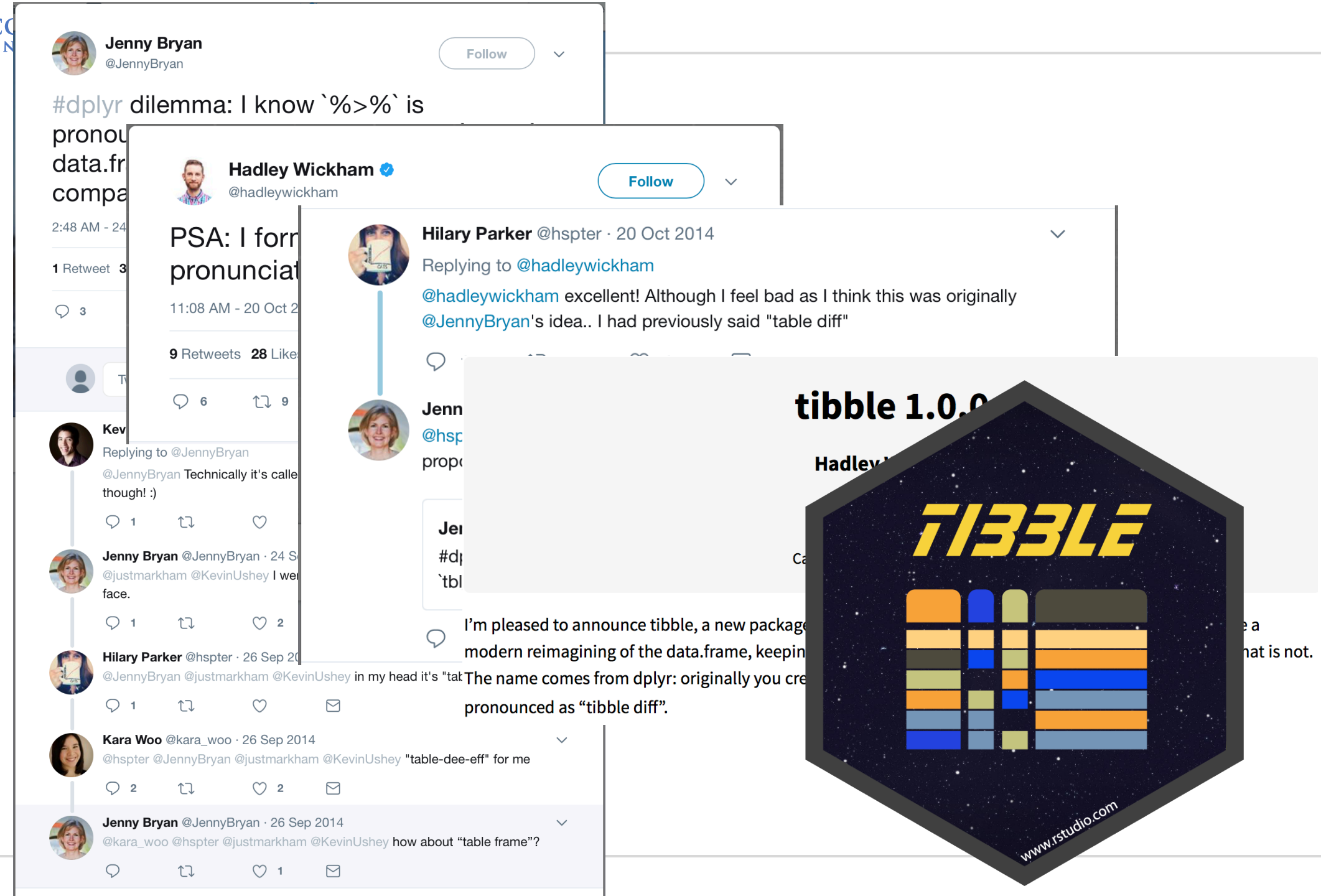
… formerly tbl_df …


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
pronou
data.fr
compa
2:48 AM - 24
1 Retweet 3
3
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
Kev
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
pronou
data.fr
compa
2:48 AM - 24
1 Retweet 3
3
Hadley Wickham
@hadleywickham
Follow
PSA: I forr
pronunciat
11:08 AM - 20 Oct 2
9 Retweets 28 Like
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 calle
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 "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
pronou
data.fr
compa
2:48 AM - 24
1 Retweet 3
3
Hadley Wickham
@hadleywickham
Follow
PSA: I forr
pronunciat
11:08 AM - 20 Oct 2
9 Retweets 28 Like
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
@hsp
propo
Jer
#dp
`tbl
Kev
Replying to @JennyBryan
@JennyBryan Technically it's calle
though! :)
1
Jenny Bryan @JennyBryan · 24 S
@justmarkham @KevinUshey I wer
face.
1
2
Hilary Parker @hspter · 26 Sep 20
@JennyBryan @justmarkham @KevinUshey in my head it's "tab
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
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”.


Jenny Bryan
@JennyBryan
Follow
#dplyr dilemma: I know `%>%` is
Hadley Wickham
@hadleywickham
Follow
PSA: I forr
pronunciat
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"
Replying to @JennyBryan
@JennyBryan Technically it's calle
though! :)
Jenny Bryan @JennyBryan · 24 S
@justmarkham @KevinUshey I wer
face.
Hilary Parker @hspter · 26 Sep 20
@JennyBryan @justmarkham @KevinUshey in my head it's "tab
Kara Woo @kara_woo · 26 Sep 2014
@hspter @JennyBryan @justmarkham @KevinUshey "table-dee-eff" for me
Jenny Bryan @JennyBryan · 26 Sep 2014
@kara_woo @hspter @justmarkham @KevinUshey how about “table frame”?
tibble 1.0.0
I’m pleased to announce tibble, a new package
modern reimagining of the data.frame, keepin
The name comes from dplyr: originally you cre
pronounced as “tibble diff”.
TIBBLE
www.rstudio.com

Jenny Bryan
@JennyBryan
Follow

#dplyr dilemma: I know `%>%` is pronou
data.fr
compa

2:48 AM - 24

1 Retweet 3

3

Hadley Wickham
@hadleywickham
Follow

PSA: I forr
pronunciat

11:08 AM - 20 Oct 2

9 Retweets 28 Like

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
@hsp
propo

Kev
Replying to @JennyBryan
@JennyBryan Technically it's calle
though! :)

1

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

1 2

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

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

Jer
#dp
`tbl

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

tibble 1.0.0

Hadley

# Why tibbles?

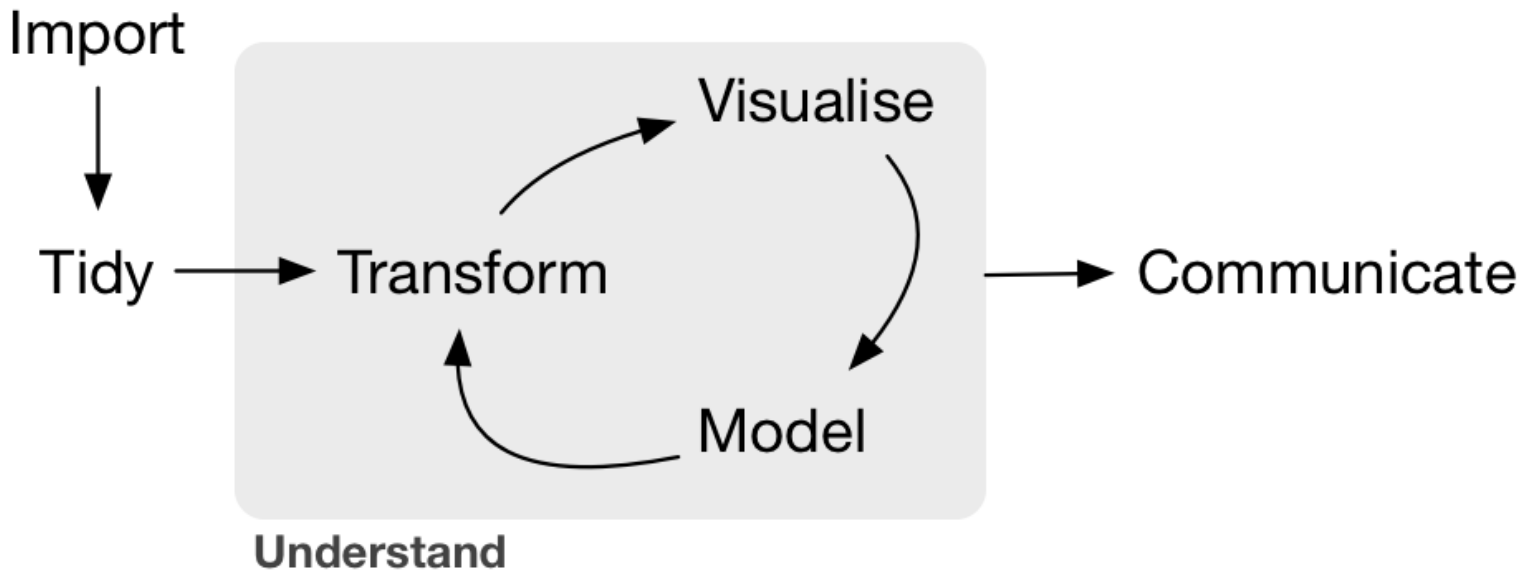
- `data.frames` have been around since R was introduced
- Some things change; base R is not one of those things
- Tibbles are data frames, just slightly different
  - They keep you from printing everything by accident
  - They make you type complete variable names

# Tools for data import

- The tools I use most for data import are `readr`, `haven`, `readxl`
  - Useful functions for importing from several sources
  - Produce tibbles
  - Fairly consistent interfaces

# Data manipulation

- Manipulate (aka transform, manage, clean) is the third step in wrangling




R for Data Science

# Major steps

- There are a few things you're going to do a lot of when you manipulate data:
  - Select relevant variables
  - Filter out unnecessary observations
  - Create new variables, or change existing ones
  - Arrange in an easy-to-digest format

# dplyr

- The `dplyr` package has specific functions that map to each of these major steps
  - `select` relevant variables
  - `filter` out unnecessary observations
  - `mutate` (sorry) new variables, or change existing ones
  - `arrange` in an easy-to-digest format

# dplyr

- The `dplyr` package has specific functions that map to each of these major steps
  - `select` relevant variables
  - `filter` out unnecessary observations
  - `mutate` (sorry) new variables, or change existing ones
  - `arrange` in an easy-to-digest format

# dplyr

- The modularity is intentional
  - Each function is designed to do one thing, and do it well
  - This is true of other functions as well (and there are several others)

- These functions share a structure: the first argument is always a data frame, and the returned objects is always a data frame
  - tibble comes in, tibble goes out, you can't explain that …

# Pipes

- Piping allows you to tie together a sequence actions
  - "New" to R (2014)
  - Comes from the `magrittr` package; loaded by everything in the tidyverse


# Pipes

- Sequence of actions to start my days
  - Wake up
  - Brush teeth
  - Do data science

- In “R”, I can nest these actions:

```
happy_jeff = do_ds(brush_teeth(wake_up(asleep_jeff)))
```

- Alternatively, I could name a bunch of intermediate objects

```
awake_jeff = wake_up(asleep_jeff)
clean_teeth_jeff = brush_teeth(awake_jeff)
happy_jeff = do_ds(clean_teeth_jeff)
```

# Pipes

- Using pipes is easier to read and understand, and avoids clutter

```
happy_jeff =
  wake_up(asleep_jeff) %>%
  brush_teeth() %>%
  do_ds()
```

- Read "%>%" as "and then"
- The result of one function gets passed as the first argument to the next one by default, although you can be more specific
- Works very well with "tibble goes in, tibble comes out" philosophy

- You will probably never fully appreciate how great piping is
  - You should be glad that that's true

# Time to code!!