Making Tidy Data

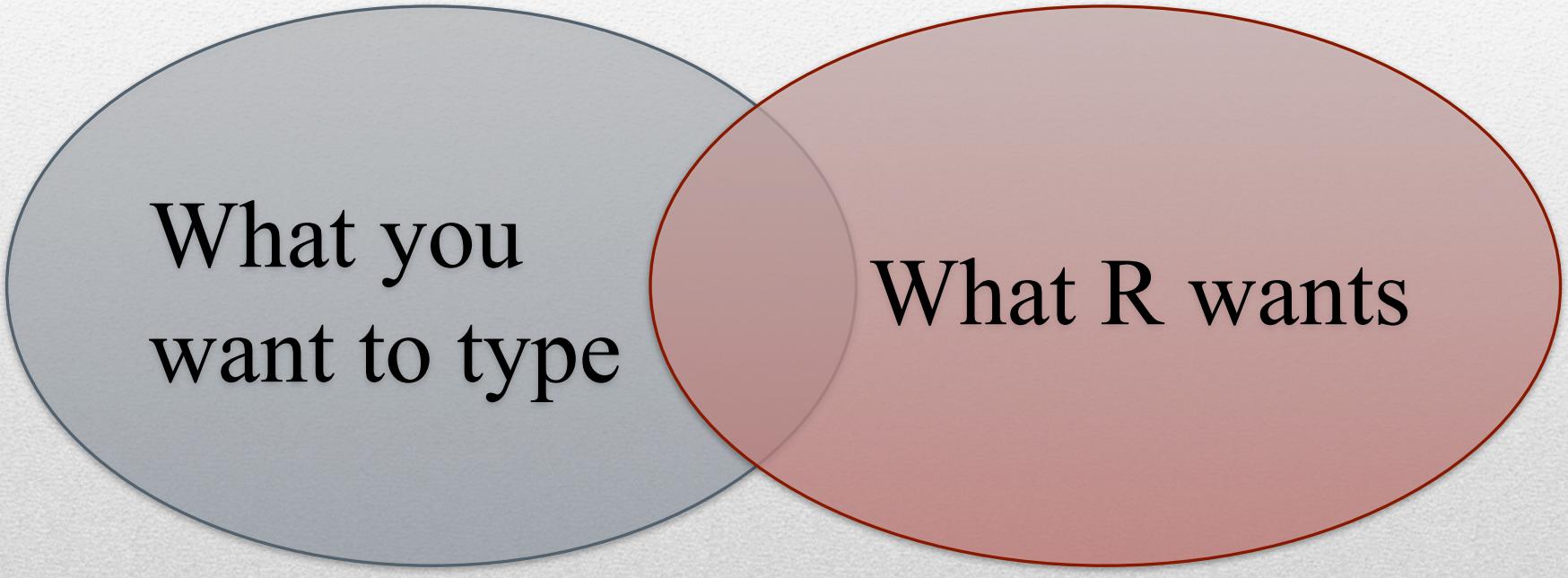
Elizabeth Wickes, @elliewix
Data Curation Specialist
Research Data Service, University of Illinois

- Download and install R
  - https://cran.mtu.edu/
- Download Rstudio
  - https://www.rstudio.com/

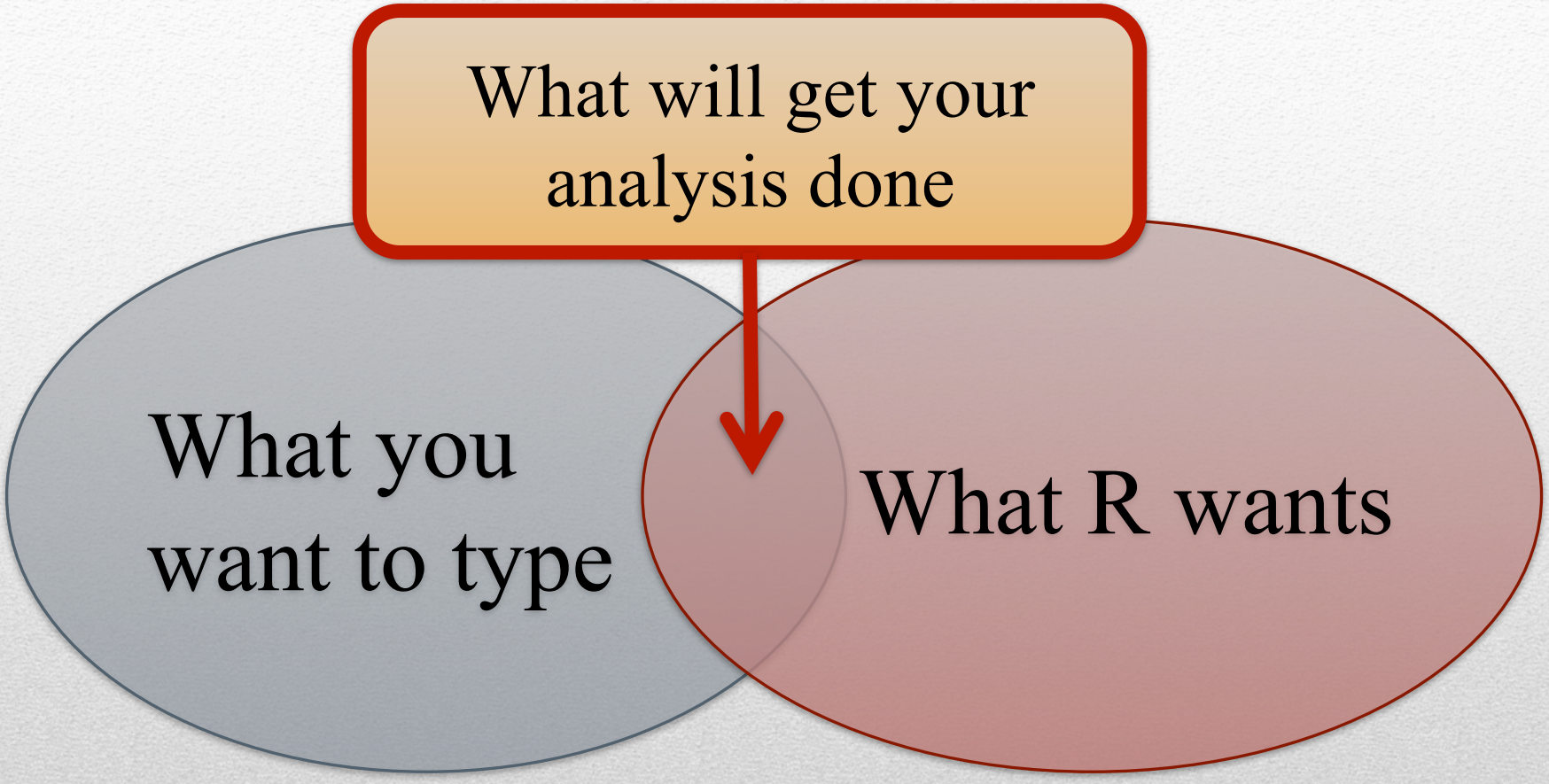
While I'm yakking

Why do we care?

- Because R does
- Many of the headaches about R are with just trying to reshape the data into something that the functions are expecting.
- But if you can start with this design from the beginning, you'll save yourself some headaches.



Humans versus R



Humans versus R

- https://ndownloader.figshare.com/files/2252083
- Open this file in your spreadsheet program of choice.
- We've got several tables per sheet, formatted for readability

Playing with a spreadsheet

Try reformatting it

- Pair up or work with 1-2 other people
- Play around with the spreadsheet
- Think about how you can reorganize this into a single sheet
  - Determine the new column headers
  - Make a new tab and start moving the data over
- Green stickies up when done

- What did you discover about the spreadsheet?

Discussion

- What was your strategy for reorganization?
- What were the column headers that you selected?

Discussion

- How would adding the 2014 data impact your design choices?

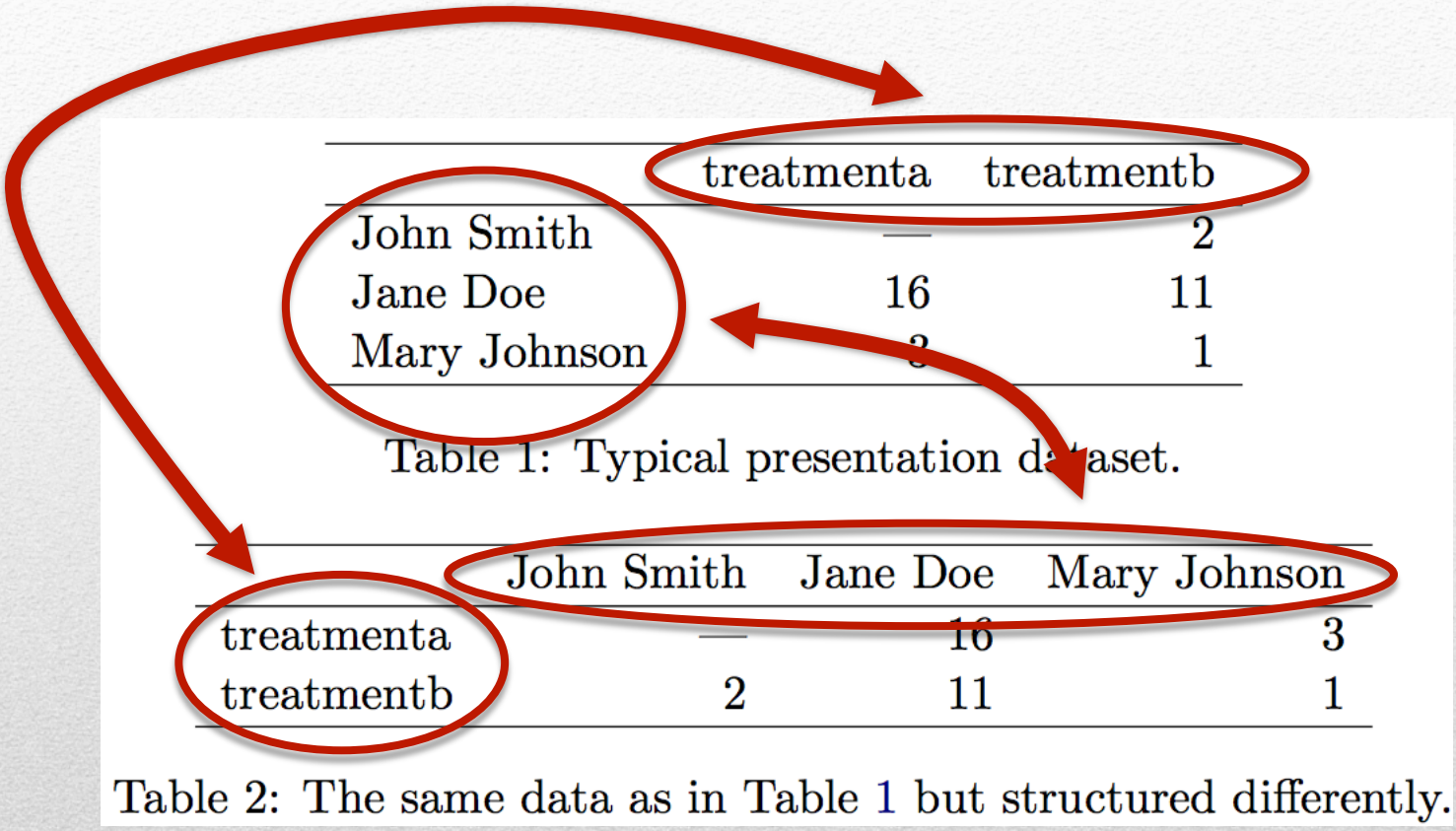
Discussion

All snark aside

- Setting up your data file to play nicely with what R expects will save you time and usually means that your format will flexible moving forward.
- We're going to explore what the tidy data format is in this workshop.

What is tidy data?

- Wickham, Hadley. (August, 2014). “Tidy Data.” *Journal of Statistical Software,* 10(59).
- Don’t describe data in terms of rows and columns.
- Data consists of:
  - Variables
  - Observations
  - Values (at the intersection of variables and observations)



	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table 2: The same data as in Table 1 but structured differently.

Semantic groups

Semantic groups

- What are the nouns that we see?
  - Humans (names – text codes)
  - Treatments (text codes)
  - Results (numerical)
- These are now our columns

The tidy* version

Each variable forms a column

Each observation forms a row.

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Each type of observational unit forms a table.

Why this feels like a database

- This is Codd's 3rd normal form (Codd 1990), but with the constraints framed in statistical language, and the focus put on a single dataset rather than the many connected datasets common in relational databases.

Melting data

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

But I just want to make a graph

- Open up Rstudio
- Live demo time
  - Load the data via Tools -> Import Dataset -> From Text File
  - Rstudio tour
  - Using plot(data) to get a quick impression
  - sapply(cleanedsurvey, summary) for quick descriptive statistics
  - install.packages("ggplot2") # see the quotes?
  - library(ggplot2) # see the lack of quotes?

But I just want to make a graph

- ggplot2
- External graphic package
- Basic syntax:
  - ggplot2(tell it what the data is…) + otheroptions() …
- These things layer together to make what you need
- Don’t try to memorize these options, just find references you like
- R syntax, particularly between packages, is ALL OVER THE PLACE. Don’t try, just copy/paste.

But I just want to make a graph

Make the ggplot object…

State the data.frame variable name

Use + to add graphic elements

- ggplot(data = cleanedsurvey) +
  geom_histogram(aes(x = Weight))

We want a histogram

"aesthetic mappings that describe how variables in the data are mapped to visual properties (aesthetics) of geoms."

Name of column where the data is that you want to plot

Your data frame goes here

Use + to add stuff

- ggplot(data = cleanedsurvey) +
  geom_histogram(aes(x = Weight))

Name of column where the data is that you want to plot

But I just want to make a graph

The joy of tidy

- Having a restricted number of columns keeps our own results and work tidier.
- For example, just trying to get the average weight for all… we can just use mean(cleanedsurvey$Weight).
- We don't have to reshape, melt, or combine things.
- We can also split the data apart in ways that make sense.
- A little more like database queries.

Acknowledgments

- Original data file from the spreadsheets lesson at Data Carpentry
  - http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- My R graphics reference was *R for Everyone*