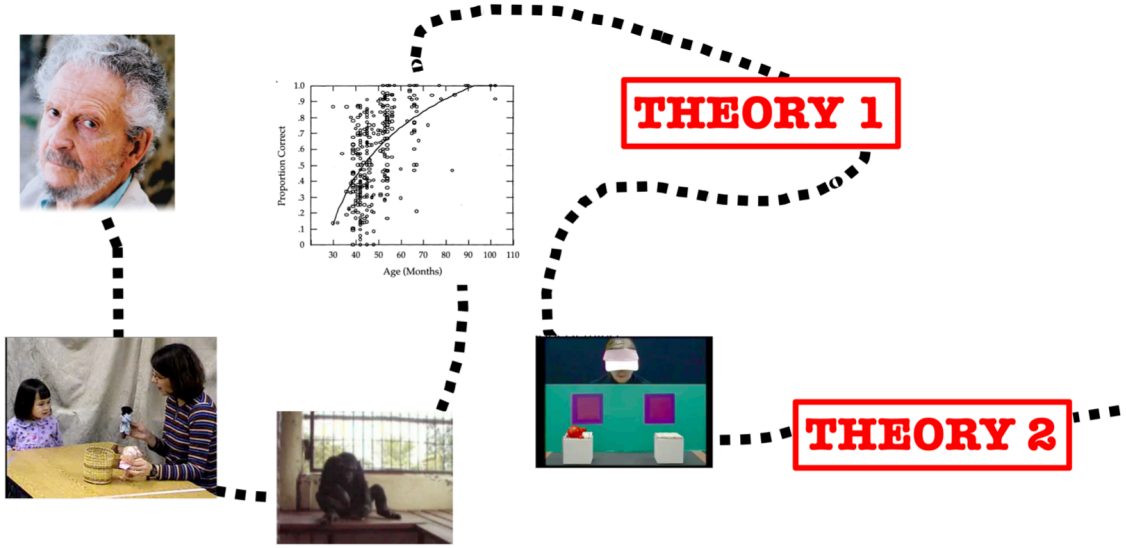
Working with Experimental Data

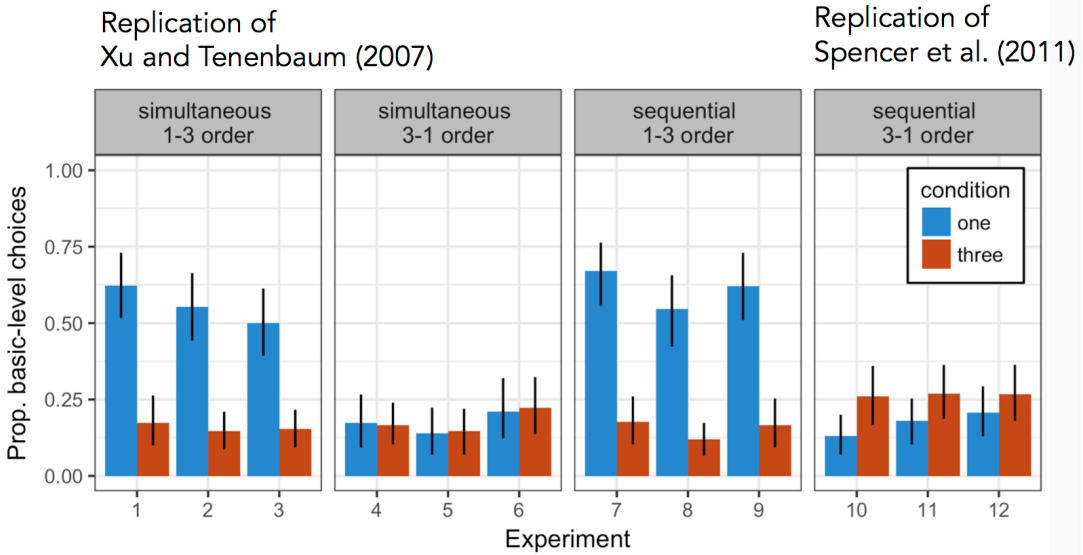
8 September 2021

Modern Research Methods

Artwork by @allison_horst

The Process of Cumulative Science





Overview of course

1) **The Process of Cumulative Science**

2) **The Single Experiment** – Experimental data, tools in R for working with data and plotting data, reproducibility

3) **Repeating an Experiment** – Intro to statistical concepts, replication of experiments

4) **Aggregating Many Experiments** – Meta-analysis

Chapter 2 Working with data

2.1 What are data?

The first important point about data is that data *are* – meaning that the word “data” is plural (though some people disagree with me on this). You might also wonder how to pronounce “data” – I say “day-tah”, but I know many people who say “dah-tah”, and I have been able to remain friends with them in spite of this. Now, if I heard them say “the data is” then that would be a bigger issue…

Data

1. **Variable** – unique measurement or quantity
   - e.g., temperature, mood, attendance, # of books owned, reaction time, color
2. **Observation** – Smallest unit you have data about
   - e.g., person, trial in an experiment, city, school, unit of time
3. **Value** – Quantity/quality associated with a particular variable and observation

Variable – Jan. high temp. in PGH
Observation – day
Value – 38 degrees

Variable – Jan. weather in PGH
Observation – sensor ID
Value – rainy, snowy, clear, other

Variable – age of students in MRM
Observation – student
Value – 19.3 years

Variable – Native Pittsburgher
Observation – person
Value – yes or no

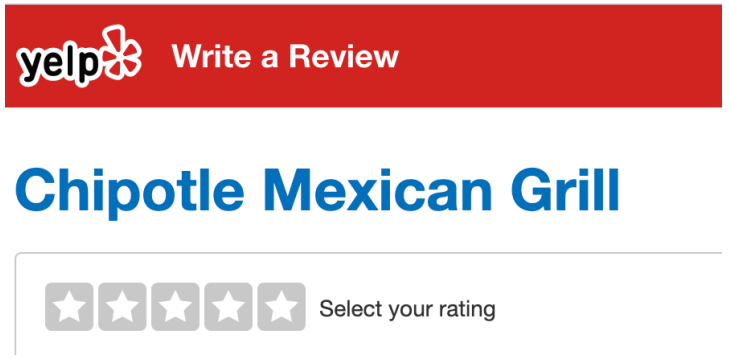
Types of variables

Discrete – measurement can only take one of a set of values

- Days of the week, dog breeds, # of children, # of Twitter followers, # of Yelp stars
- No "middle ground"

Continuous – measurement that is real number, and could take one of any range of values

- Most quantitative variables (reaction time, judgement on slider scale)
- Limited by precision of instrument



CONTINUOUS

measured data, can have ∞ values within possible range.

I AM 3.1" TALL

I WEIGH 34.16 grams

DISCRETE

OBSERVATIONS CAN ONLY EXIST AT LIMITED VALUES, OFTEN COUNTS.

I HAVE 8 LEGS and 4 SPOTS!

@allison_horst

Artwork by @allison_horst

Types of variables

- **Qualitative** – describe quality (no intrinsic ordering)
- **Quantitative** – describe quantity
  - **Binary** – 1 or 0 (or, TRUE or FALSE)
  - **Integers** – whole numbers
  - **Real numbers** – have fractional/decimal part

Variable – Jan. high temp. in PGH
Observation – day
Value – 38 degrees

Variable – Jan. weather in PGH
Observation – sensor ID
Value – rainy, snowy, clear, other

Variable – age of students in MRM
Observation – student
Value – 19.3 years

Variable – Native Pittsburgher
Observation – person
Value – yes or no

An example: class attendance

How many students showed up to class today? (*binary*)

Student	Attendance
Kyla	TRUE
Kara	FALSE
Zara	TRUE

How many times did each student comment in class today? (*integer*)

Student	Attendance
Kyla	1
Kara	NA
Zara	3

What was the most common type of attention of students in class today? (*qualitative*)

Student	Attendance
Kyla	“quiet”
Kara	NA
Zara	“inquisitive”

How many seconds early was each student today? (*real number*)

Student	Attendance
Kyla	120.457
Kara	NA
Zara	125.332

(1)

Variable – How much food Rhoda eats
Observation – ?
Value – ? [quantitative, binary]

(2)

Variable – How much food Rhoda eats
Observation – ?
Value – ? [quantitative, integer]

(3)

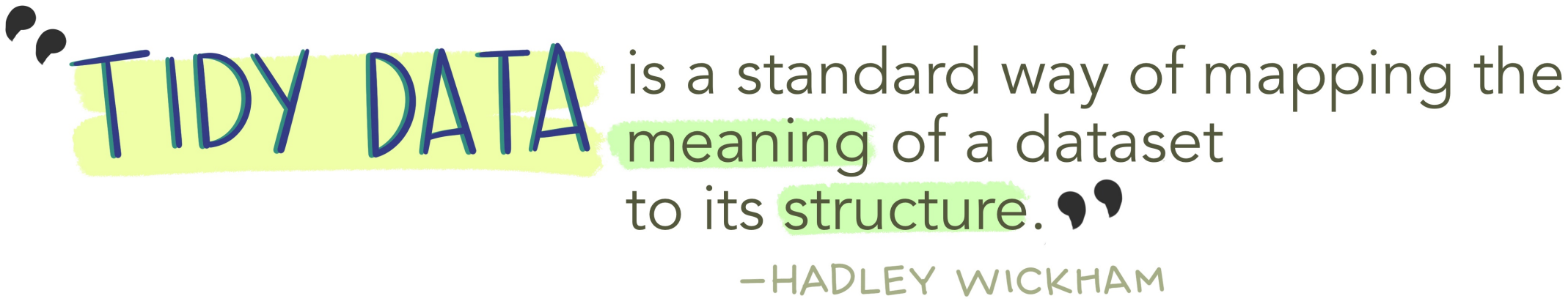
Variable – How much food Rhoda eats
Observation – ?
Value – ? [quantitative, real]

(4)

Variable – How much food Rhoda eats
Observation – ?
Value – ? [qualitative]

Structuring data

- Most data has this structure (variable, observation, value)
- Lots of ways we could take this and put it into a spreadsheet
- In this class, we're going to structure our data in a particular way, called **tidy data**

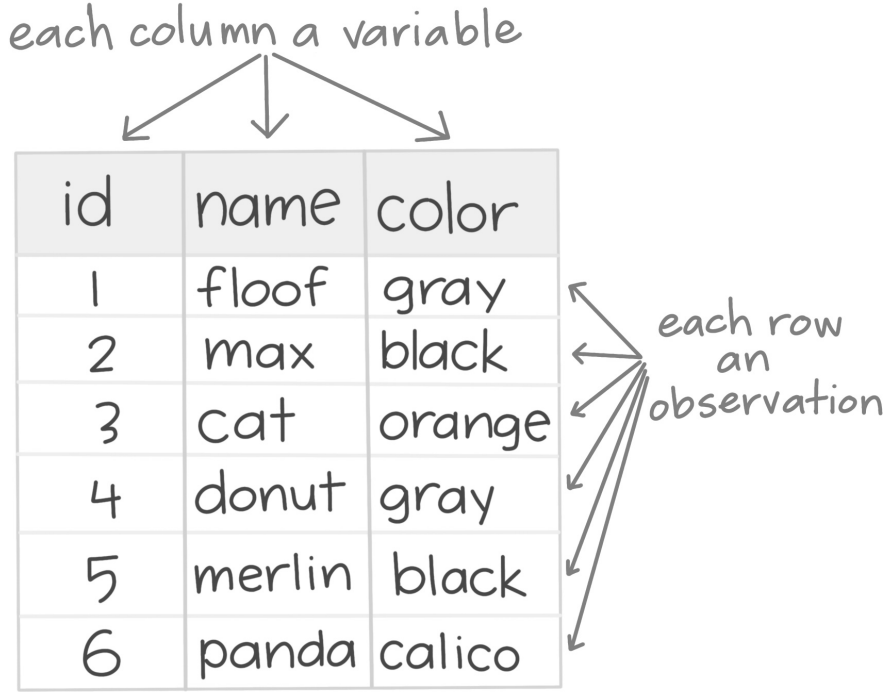


"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

—HADLEY WICKHAM

In tidy data:

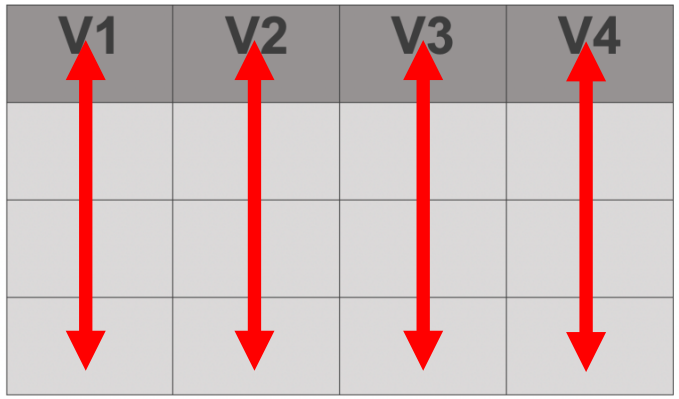
- each variable forms a column
- each observation forms a row
- each cell is a single measurement



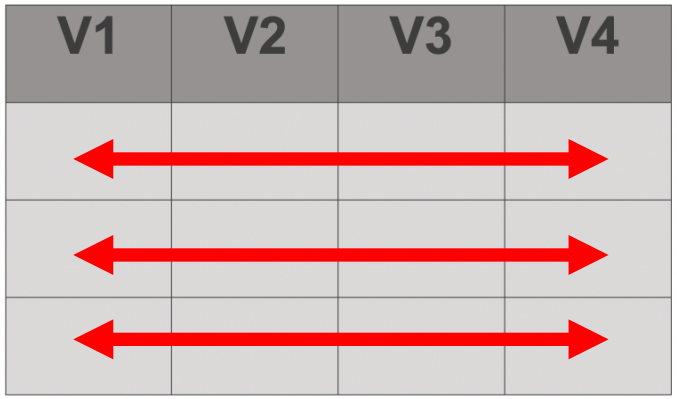
Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

Tidy data

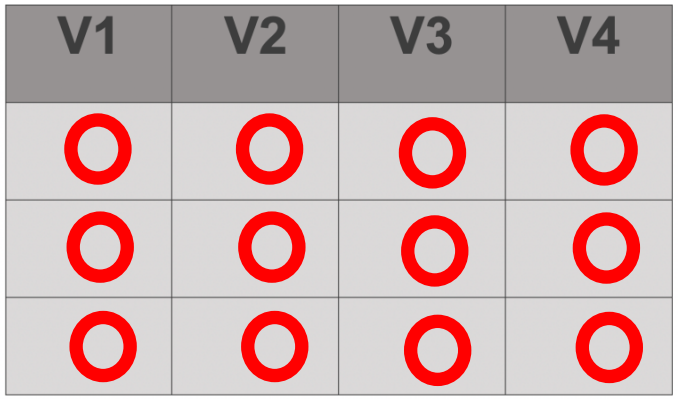
Each **variable** is its own column



Each **observation** is its own row



Each **value** is its own cell



* Allows R to use vectorized observations

Tidy data

Variable – Jan. high temp in PGH
Observation – day
Value – 38 degrees

Pittsburgh January 2020 temperature data

Variable

Observation

Value

Date	Low Temp.	High Temp.
1/1/2020	28	38
1/2/2020	40	48
1/3/2020	49	51

Tidy data

Variable – age of students in MRM
Observation – student
Value – 19.3 years

MRM student age data

Student	Age	Year
Sam	19.3	2
Zara	20	3
Caitlin	20.2	3

Sketch the tidy data…

Variable – How much food Rhoda eats
Observation – ?
Value – ? [quantitative, binary]

Variable – How much food Rhoda eats
Observation – ?
Value – ? [quantitative, integer]

Variable – How much food Rhoda eats
Observation – ?
Value –? [quantitative, real]

Variable – How much food Rhoda eats
Observation – ?
Value – ? [qualitative]

	A	AA	AB	AC	AD	AE	AF	AG	AH
1	Estimated HIV Prevalence% - (Ages 15-49)	2004	2005	2006	2007	2008	2009	2010	2011
2	Abkhazia								
3	Afghanistan						0.06	0.06	0.06
4	Akrotiri and Dhekelia								
5	Albania								
6	Algeria	0.1	0.1	0.1	0.1	0.1			
7	American Samoa								
8	Andorra								
9	Angola	1.9	1.9	1.9	1.9	2	2.1	2.1	2.1
10	Anguilla								
11	Antigua and Barbuda								
12	Argentina	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4
13	Armenia	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2
14	Aruba								
15	Australia	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2
16	Austria	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4
17	Azerbaijan	0.06	0.06	0.06	0.1	0.1	0.1	0.1	0.1
18	Bahamas	3	3	3	3.1	3.1	2.9	2.8	2.8

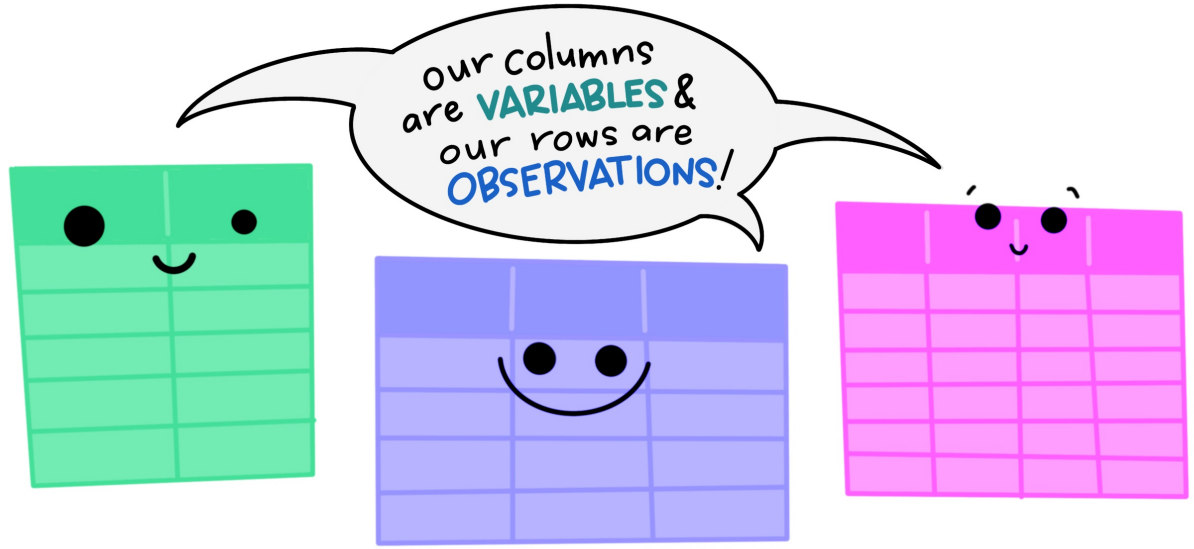
Source: Gapminder, Estimated HIV prevalence among 15-49 year olds

Airplanes on Hand in the AAF, By Major Type: Jul 1939 to Aug 1945

End of Month	Total	Very Heavy Bombers	Heavy Bombers	Medium Bombers	Light Bombers	Fighters	Recon-naissance	Transports	Trainers	Communi-cations
1939										
Jul	2,402	-	16	400	276	494	356	118	735	7
Aug	2,440	-	18	414	276	492	359	129	745	7
[Germany invades Poland, 1 Sep 1939]										
Sep	2,473	-	22	428	278	489	359	136	754	7
Oct	2,507	-	27	446	277	490	365	137	758	7
Nov	2,536	-	32	458	275	498	375	136	755	7
Dec	2,546	-	39	464	274	492	378	131	761	7
1940										
Jan	2,588	-	45	466	271	464	409	128	798	7
Feb	2,658	-	49	470	271	458	415	128	860	7
Mar	2,709	-	54	468	267	453	415	125	920	7
Apr	2,806	-	54	468	263	451	416	125	1,022	7
May	2,906	-	54	470	259	459	410	124	1,123	7
Jun	2,966	-	54	478	166	477	414	127	1,243	7
[France surrenders to Germany, 25 Jun 1940] [Battle of Britain begins, 10 July 1940]										
Jul	3,102	-	56	483	161	500	410	128	1,357	7
Aug	3,295	-	65	485	158	539	407	128	1,506	7

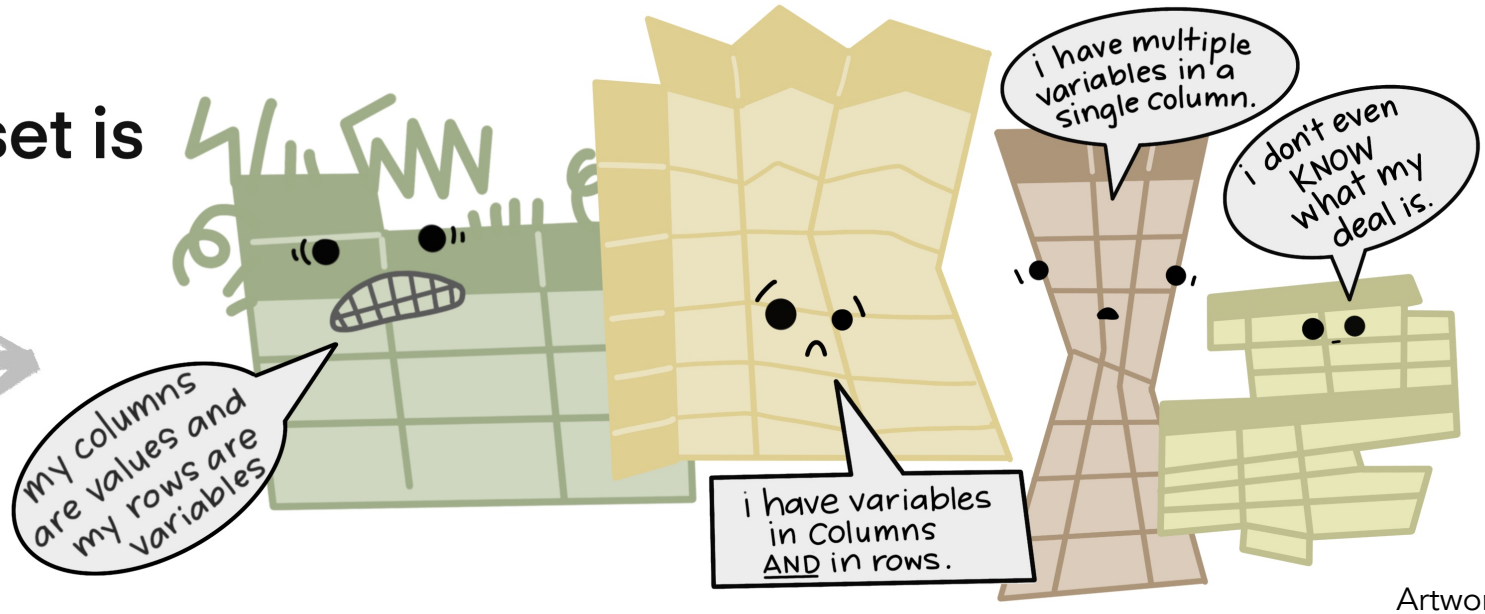
Source: Army Air Forces Statistical Digest, WW II

The standard structure of tidy data means that "tidy datasets are all alike..."



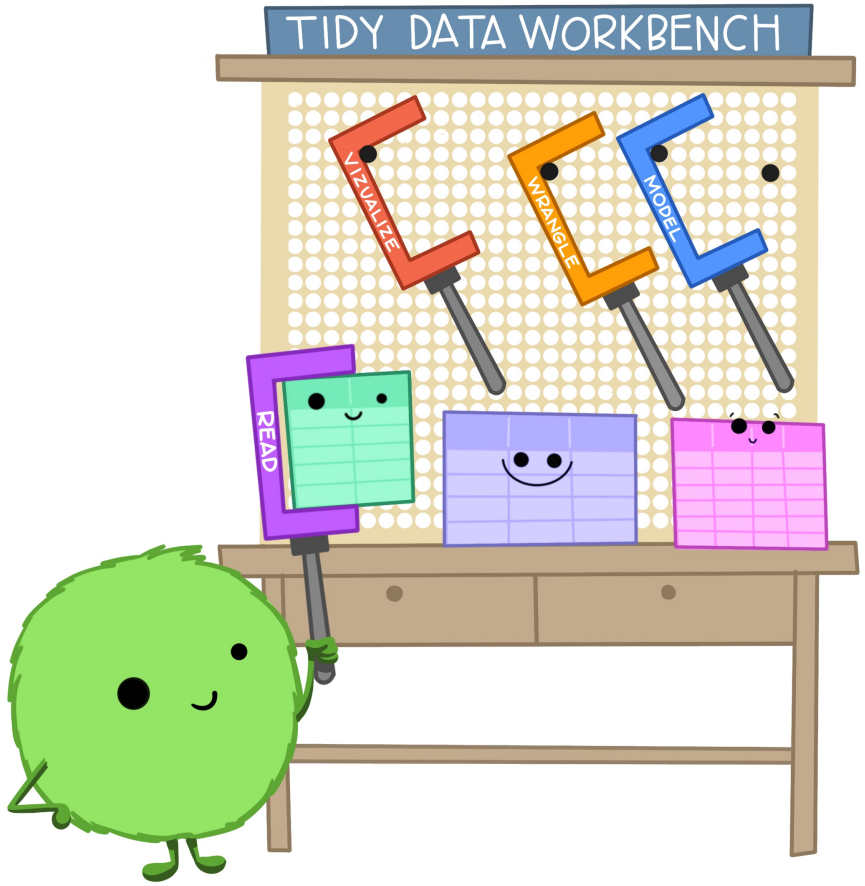
"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM

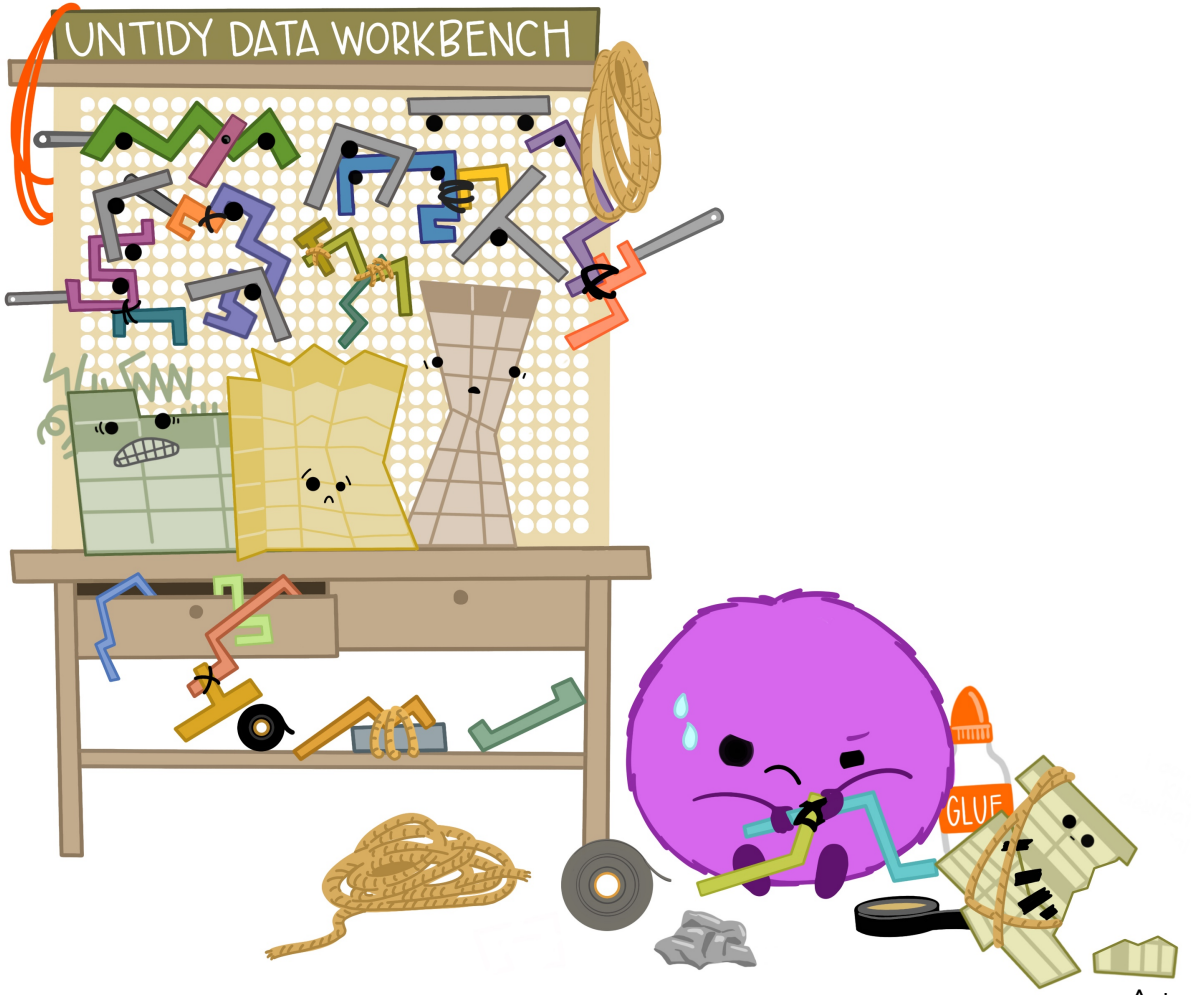


Artwork by @allison_horst

When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



Artwork by @allison_horst

Next Time: Lab

- Start to learn tools for working with tidy data
- Learn how to produce a "report" in R using Rmarkdown
- No reading, but short video (linked on website)