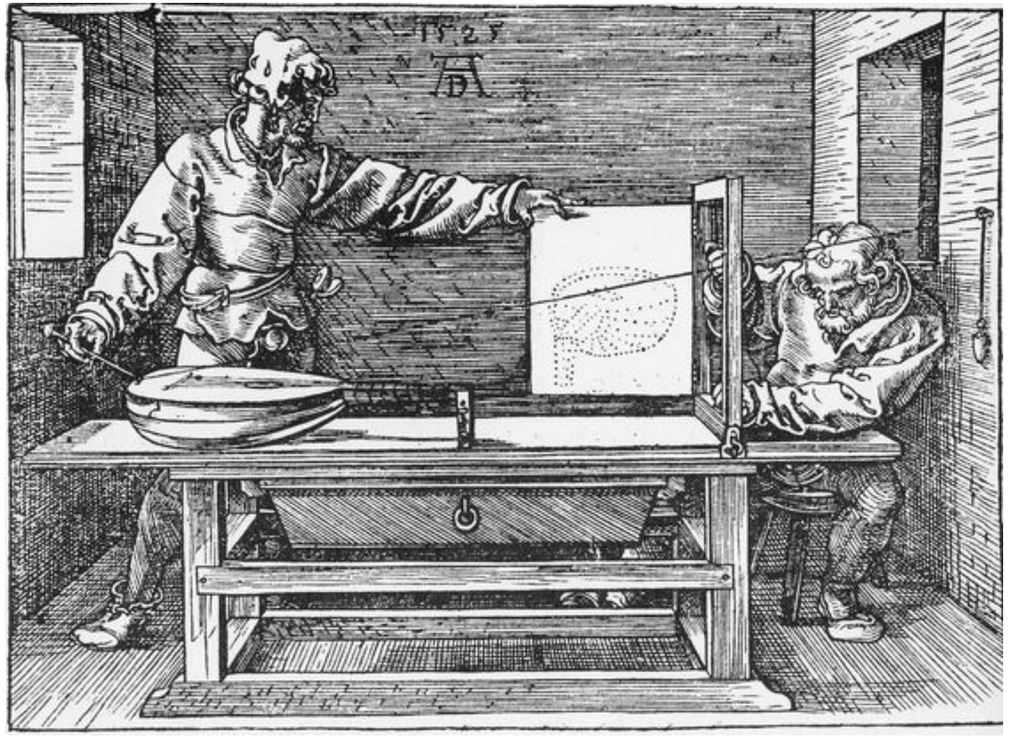
Master the Tidyverse



Garrett Grolemund

Data Scientist, Educator

July 2018

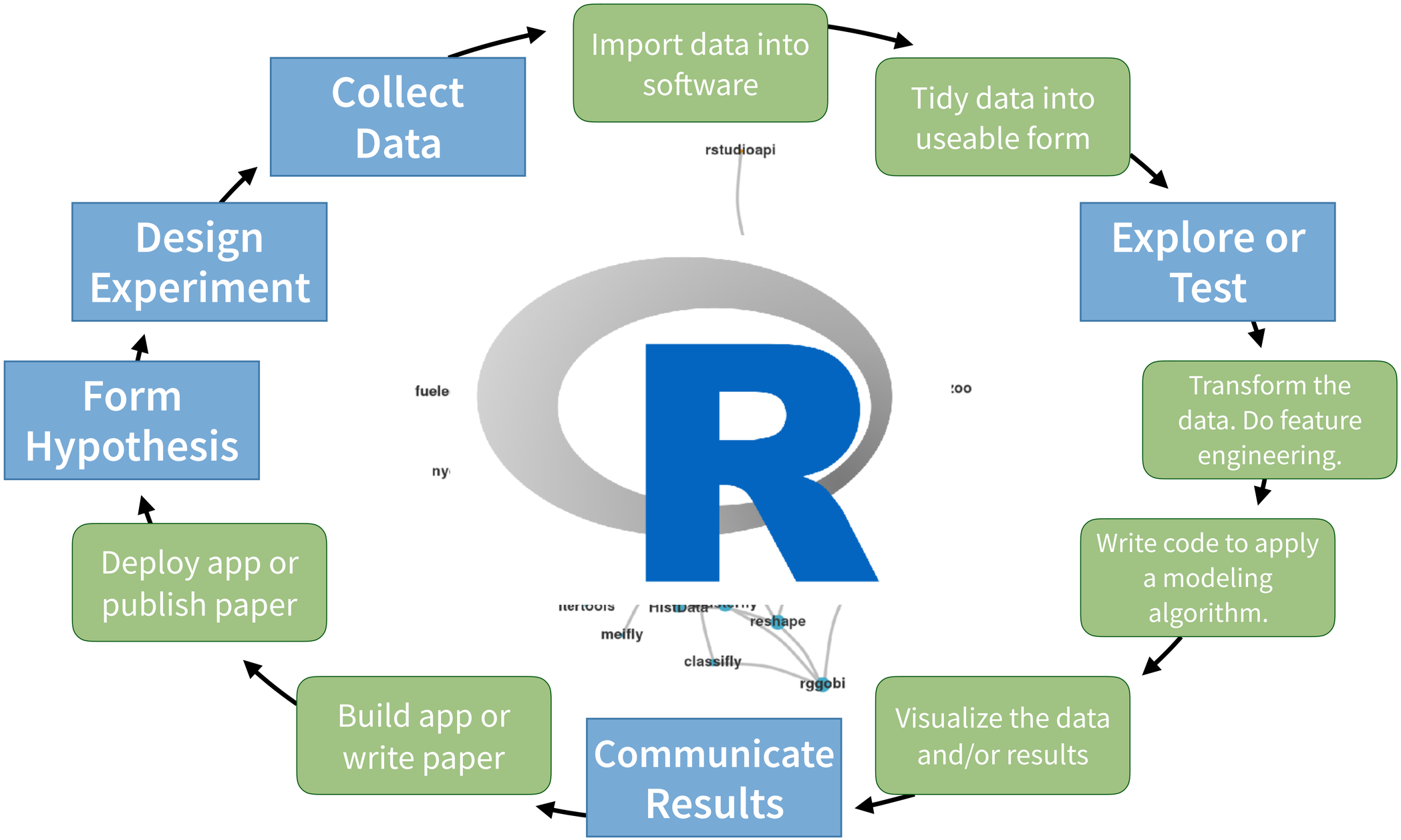
RStudio

Slides CC BY-SA RStudio

Your Turn

Re-introduce yourself to the people at your table. Then login to your rstudio.cloud project.

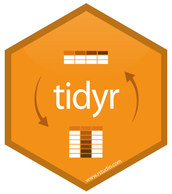
Restart your R Session:
Session > Restart R


Collect Data
Import data into software
Tidy data into useable form
Explore or Test
Transform the data. Do feature engineering.
Write code to apply a modeling algorithm.
Visualize the data and/or results
Communicate Results
Build app or write paper
Deploy app or publish paper
Form Hypothesis
Design Experiment
rstudioapi
fuele
ny
:oo
reshape
meifly
classifly
rggobi

Tidy data

A data set is **tidy** iff:

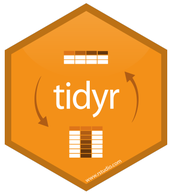
1. Each **variable** is in its own **column**
2. Each **case** is in its own **row**
3. Each **value** is in its own **cell**



Tidy tools

country	year	cases	pop
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

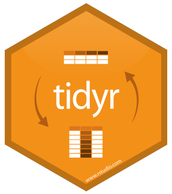
```
filter(df, year == 2000)
```



Tidy tools

country	year	cases	pop
Afghanistan	2000	2666	20595360
Brazil	2000	80488	174504898
China	2000	213766	1280428583

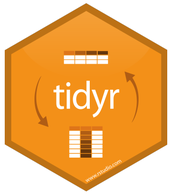
```
filter(df, year == 2000)
select(df, -year)
```



Tidy tools

country	cases	pop	rate
Afghanistan	2666	20595360	0.00013
Brazil	80488	174504898	0.00046
China	213766	1280428583	0.00017

```
filter(df, year == 2000)
select(df, -year)
mutate(df, rate = cases / pop)
```

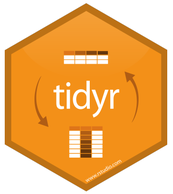


Tidy tools

country	cases	pop	rate
Afghanistan	2666	20595360	0.00013
Brazil	80488	174504898	0.00046
China	213766	1280428583	0.00017

avg
0.00025

```
filter(df, year == 2000)
select(df, -year)
mutate(df, rate = cases / pop)
summarise(df, avg = mean(rate))
```

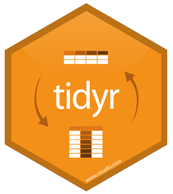


Tidy tools

country	cases	pop	rate
Afghanistan	2666	20595360	0.00013
Brazil	80488	174504898	0.00046
China	213766	1280428583	0.00017

avg
0.00025

```
df %>%
  filter(year == 2000) %>%
  select(-year) %>%
  mutate(rate = cases / pop) %>%
  summarise(avg = mean(rate))
```



Today

Functions for specific types of data.

strings factors dates times

Non-Tidy R

Lists

```
$city
[1] "New York" "New York" "London"
[4] "London"   "Beijing"  "Beijing"

$size
[1] "large" "small" "large" "small"
[5] "large" "small"

$amount
[1]  23  14  22  16 121 121

attr(,"row.names")
[1] 1 2 3 4 5 6
```

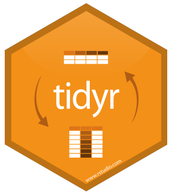
Models

```
Call:
lm(formula = lifeExp ~ year, data = gapminder)

Residuals:
   Min      1Q  Median      3Q     Max
-39.949  -9.651   1.697  10.335  22.158

Coefficients:
                Estimate Std. Error t value   Pr(>|t|)
(Intercept) -585.65219   32.31396  -18.12   <2e-16 ***
year           0.32590    0.01632   19.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.63 on 1702 degrees of freedom
Multiple R-squared:  0.1898,    Adjusted R-squared:  0.1893
F-statistic: 398.6 on 1 and 1702 DF,  p-value: < 2.2e-16
```



List Columns

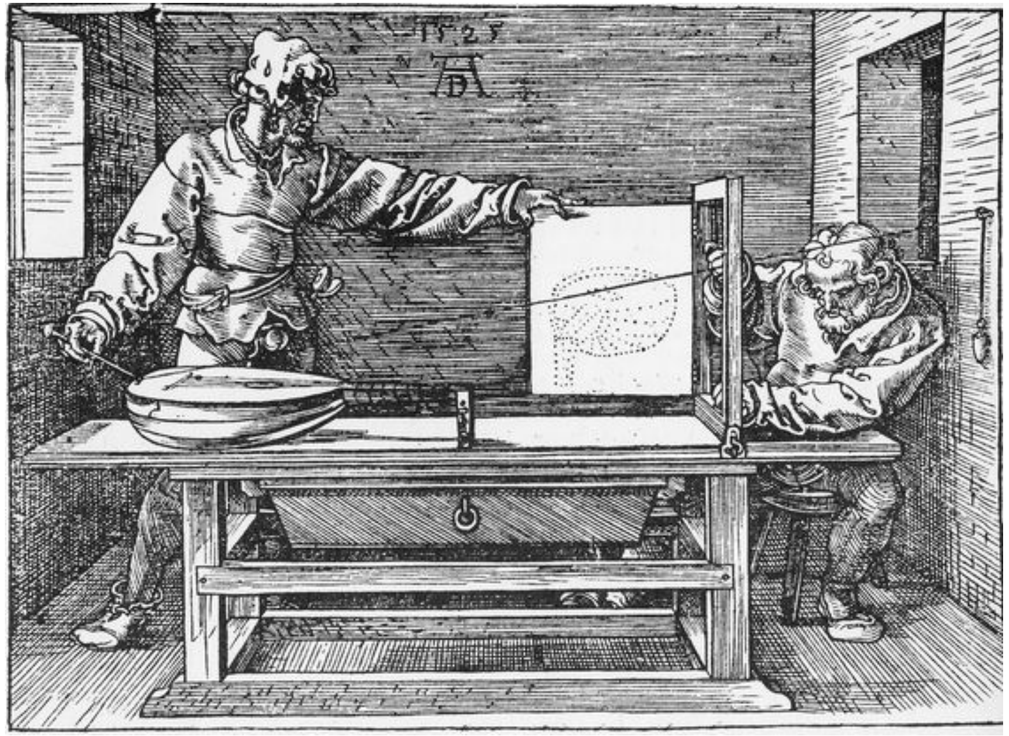
A table is …an organizational structure …that you can manipulate.

country	r.squared	data	model
Botswana	0.03	year .resid 1952 -5.3071154 1957 -3.6144580 1962 -2.0158007 1967 -0.5411434 1972 1.8815140 1977 4.8731713 1982 6.7348287 1987 8.5694860 1992 7.3891434 1997 -3.1031993 2002 -9.3285420 2007 -5.5378846	Call: lm(formula = lifeExp ~ year, data = .) Coefficients: (Intercept) year -65.49586 0.06067
Lesotho	0.08	year .resid 1952 -5.2410256 1957 -2.8098543 1962 -0.5876830 1967 -0.3205117 1972 0.4766597 1977 2.4398310 1982 4.8320023 1987 6.4561737 1992 8.4833450 1997 3.8785163 2002 -7.5643124 2007 -10.0431410	Call: lm(formula = lifeExp ~ year, data = .) Coefficients: (Intercept) year -139.16529 0.09557

Day 2

ReIntroduction and Data Types	9:00 - 10:45
Morning Break	10:45 - 11:00
Iteration	11:00 - 12:30
Lunch	12:30 - 1:30
Modeling	1:30 - 3:15
Afternoon Break	3:15 - 3:30
List Columns	3:30 - 5:00

Master the Tidyverse



Garrett Grolemund

Data Scientist, Educator

October 2017

RStudio

Slides CC BY-SA RStudio