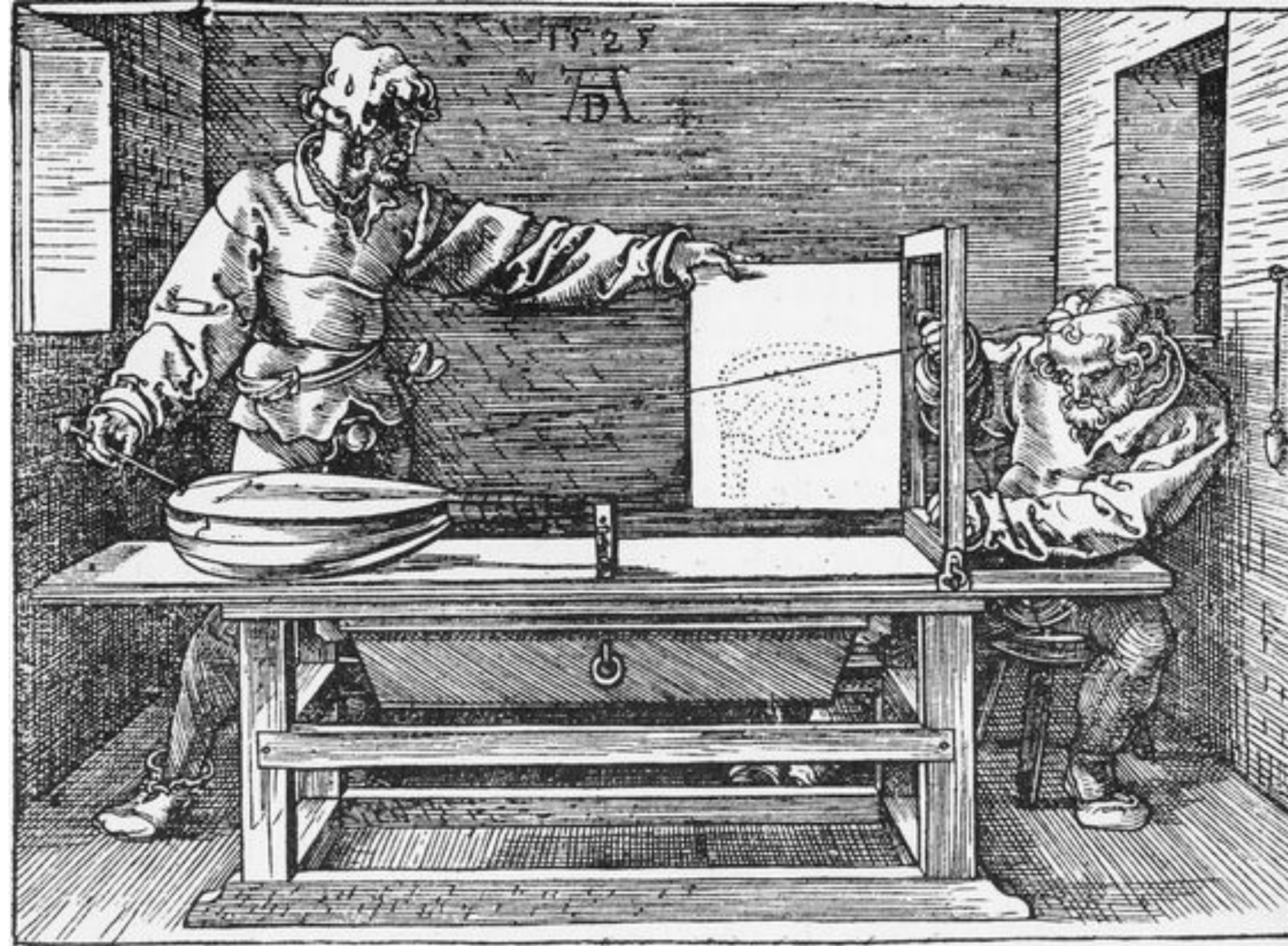


Master the Tidyverse



Garrett Grolmund

Data Scientist, Educator

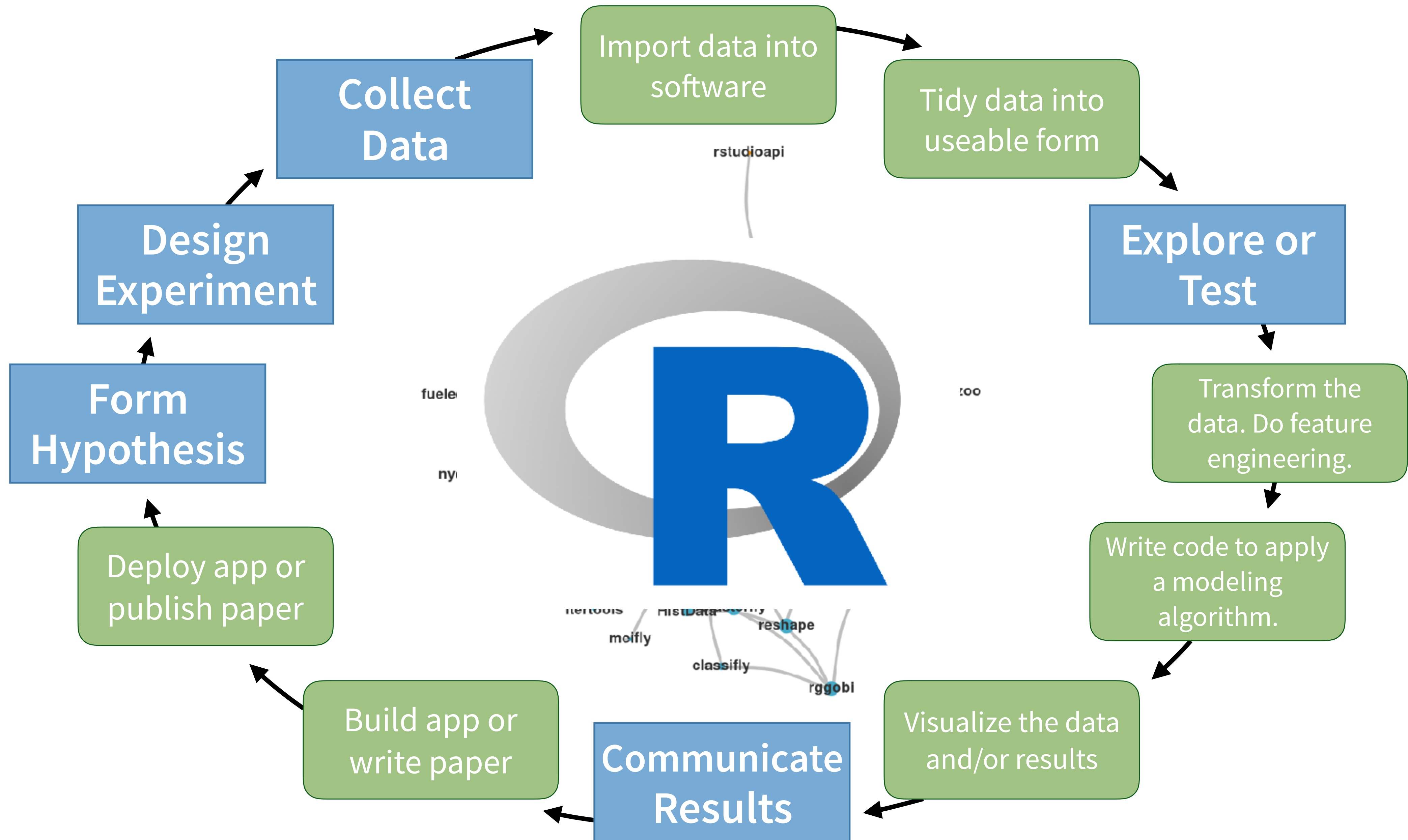
January 2017

RStudio

Your Turn

Re-introduce yourself to the people at your table. Then login to the classroom server at harpers-ferry.rstudio.com.

05:00

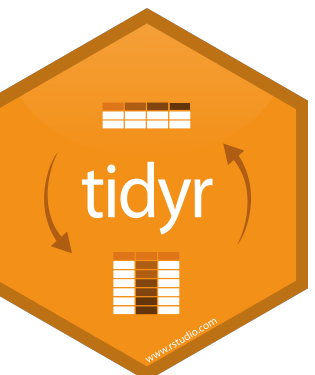


Tidy data

| country | year | cases | pop |
|-------------|------|-------|----------|
| Afghanistan | 1999 | 745 | 10137321 |
| Afghanistan | 2000 | 666 | 20125120 |
| Afghanistan | 2001 | 3153 | 17133432 |
| China | 2000 | 3153 | 1271372 |
| India | 1999 | 22258 | 1271372 |
| India | 2000 | 3766 | 12812363 |

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 |

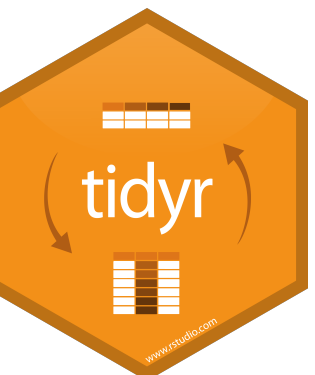
```
filter(df, year == 2000)
```



Tidy tools

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

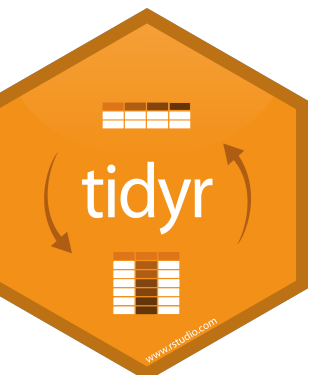
```
filter(df, year == 2000)  
select(df, -year)
```



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)
```



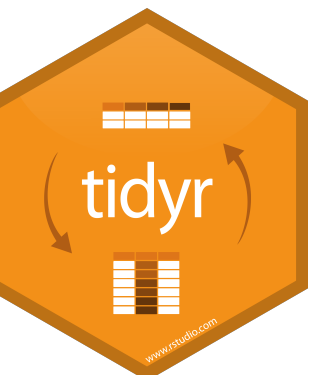
Tidy tools

| 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 |

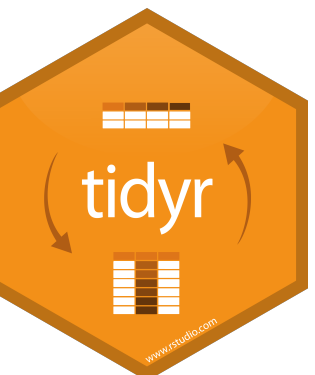
`df %>%`

`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
```

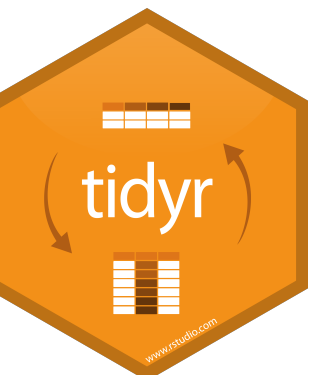
Models

```
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 | <table><tr><th>year</th><th>.resid</th></tr><tr><td>1952</td><td>-5.3071154</td></tr><tr><td>1957</td><td>-3.6144580</td></tr><tr><td>1962</td><td>-2.0158007</td></tr><tr><td>1967</td><td>-0.5411434</td></tr><tr><td>1972</td><td>1.8815140</td></tr><tr><td>1977</td><td>4.8731713</td></tr><tr><td>1982</td><td>6.7348287</td></tr><tr><td>1987</td><td>8.5694860</td></tr><tr><td>1992</td><td>7.3891434</td></tr><tr><td>1997</td><td>-3.1031993</td></tr><tr><td>2002</td><td>-9.3285420</td></tr><tr><td>2007</td><td>-5.5378846</td></tr></table> | 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 | <div><p>Call: lm(formula = lifeExp ~ year, data = .)</p><p>Coefficients: (Intercept) year -65.49586 0.06067</p></div> |
| 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 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Lesotho | 0.08 | <table><tr><th>year</th><th>.resid</th></tr><tr><td>1952</td><td>-5.2410256</td></tr><tr><td>1957</td><td>-2.8098543</td></tr><tr><td>1962</td><td>-0.5876830</td></tr><tr><td>1967</td><td>-0.3205117</td></tr><tr><td>1972</td><td>0.4766597</td></tr><tr><td>1977</td><td>2.4398310</td></tr><tr><td>1982</td><td>4.8320023</td></tr><tr><td>1987</td><td>6.4561737</td></tr><tr><td>1992</td><td>8.4833450</td></tr><tr><td>1997</td><td>3.8785163</td></tr><tr><td>2002</td><td>-7.5643124</td></tr><tr><td>2007</td><td>-10.0431410</td></tr></table> | 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 | <div><p>Call: lm(formula = lifeExp ~ year, data = .)</p><p>Coefficients: (Intercept) year -139.16529 0.09557</p></div> |
| 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 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Day 2

**ReIntroduction and
Data Types**

8:30 - 10:15

Morning Break

10:15 - 10:30

Iteration

10:30 - 12:30

Lunch

12:30 - 2:00

Modeling

2:00 - 3:15

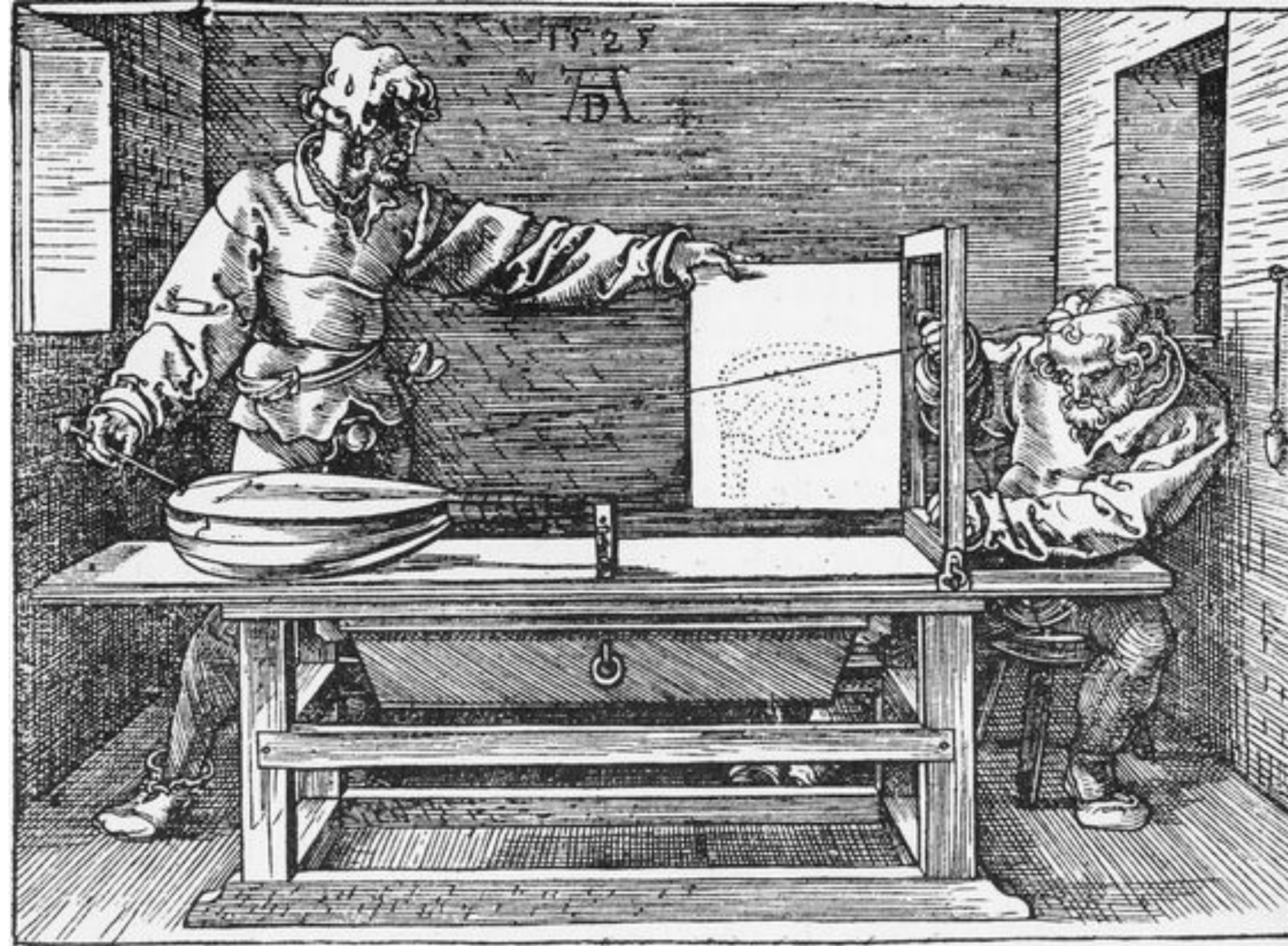
Afternoon Break

3:15 - 3:30

List Columns

3:30 - 5:00

Master the Tidyverse



Garrett Grolmund

Data Scientist, Educator

January 2017

RStudio