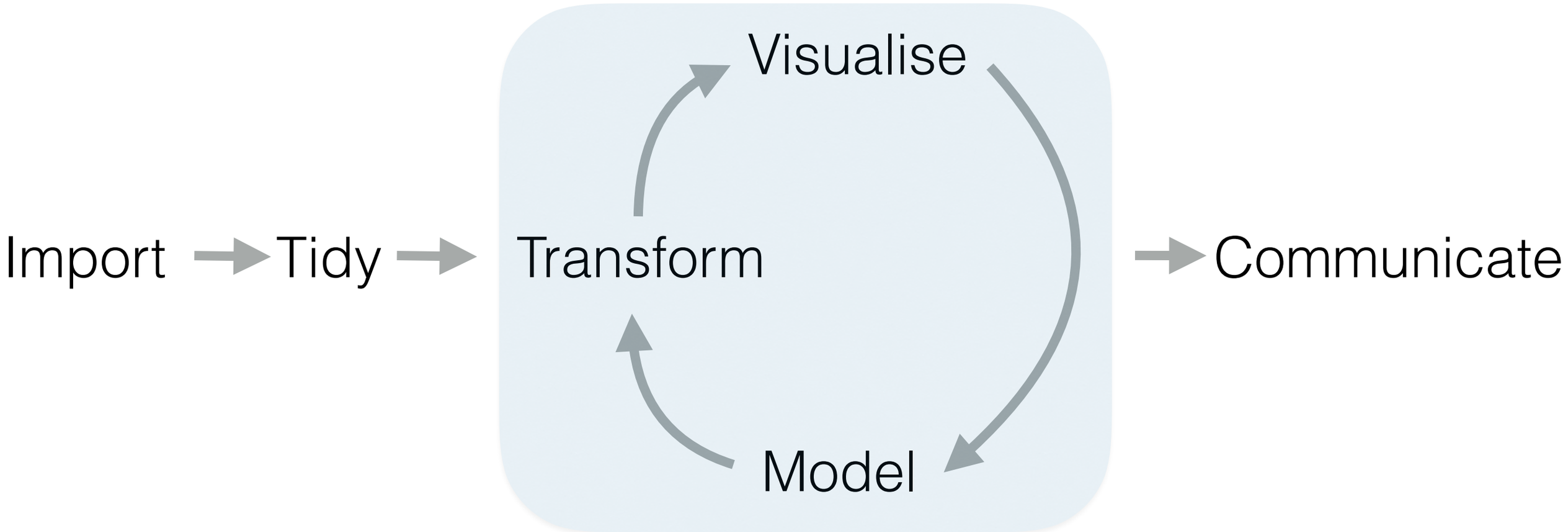
Excel to R
An introduction

Data analysis workflow



Doing this in Excel + Minitab/Graphpad etc

Why reproducible research?

- To demonstrate correctness / audit
- To allow others to follow methods
- To remember what happened…
  - Students graduate
  - Future you!

"An article about computational science in a scientific publication is not the scholarship itself it is merely advertising of the scholarship"

–From Claerbout and Karrenbach (1992)

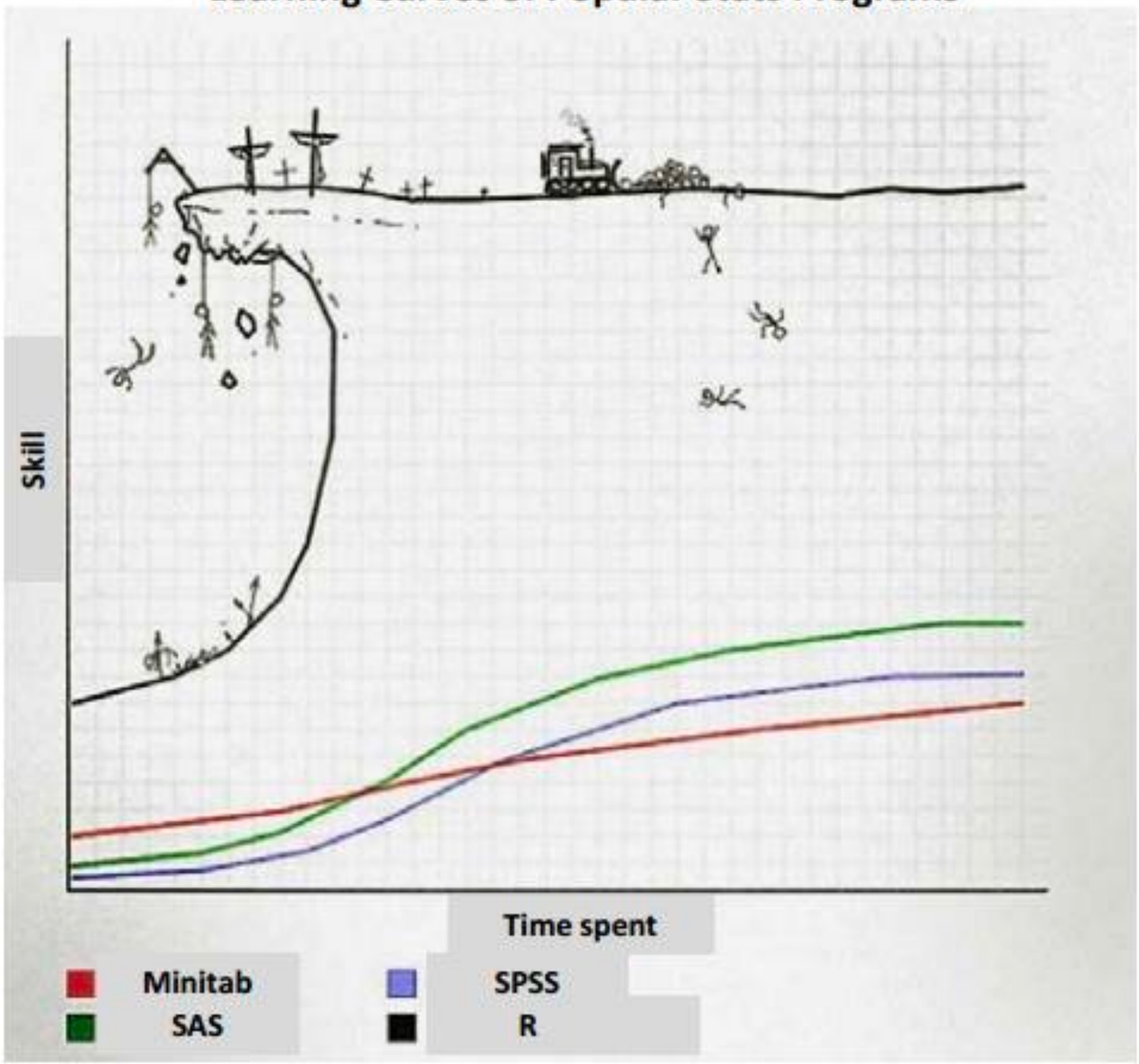
What is R?

- Programming language / system for data science
- Open source and **free**
- Home for lots of new methods
- Extended by > 10,000 packages…

“This huge variety of packages is one of the reasons that R is so successful: the chances are that someone has already solved a problem that you're working on, and you can benefit from their work by downloading their package.”

–Hadley Wickham, Chief Scientist RStudio

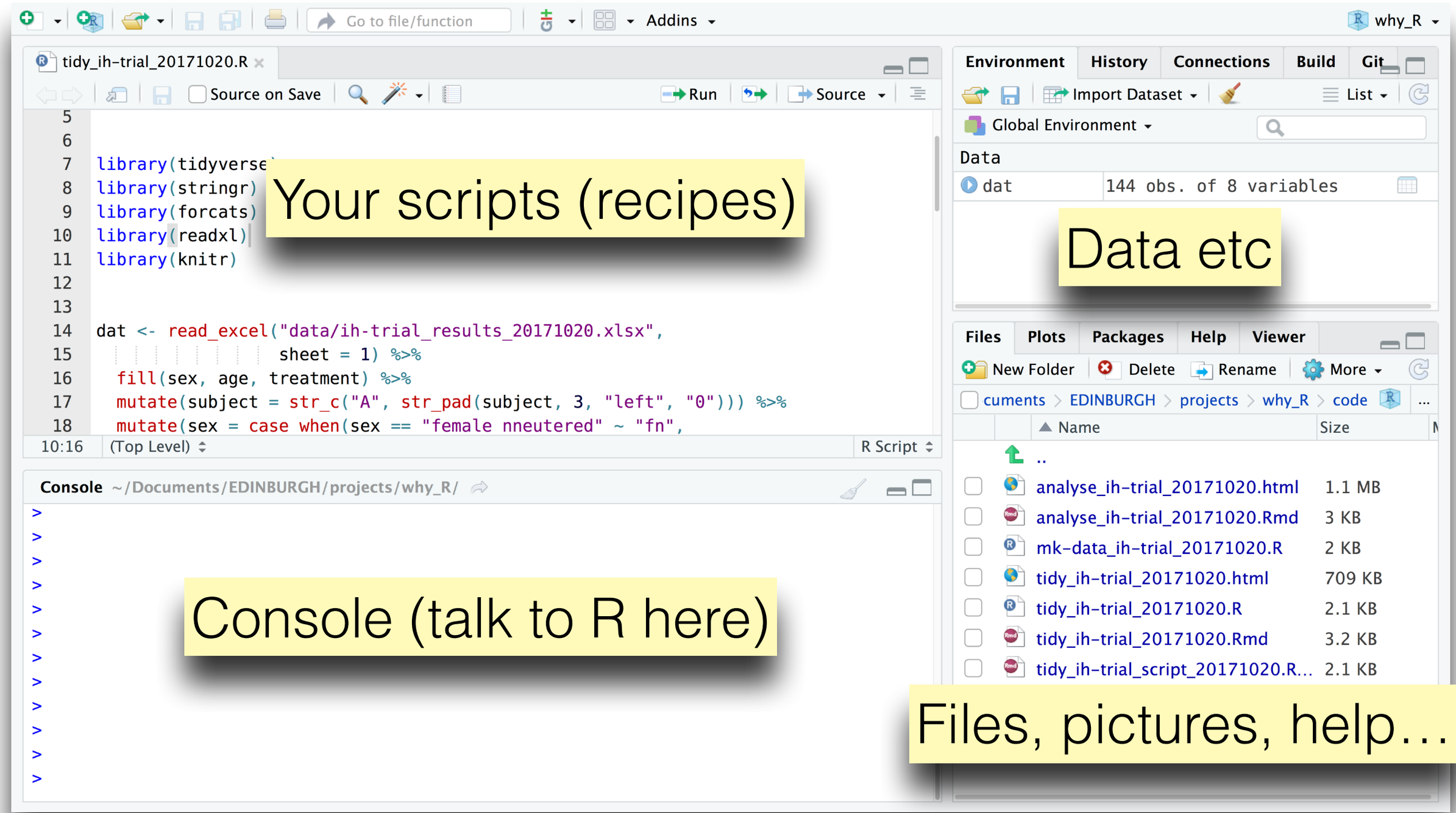
Learning Curves of Popular Stats Programs



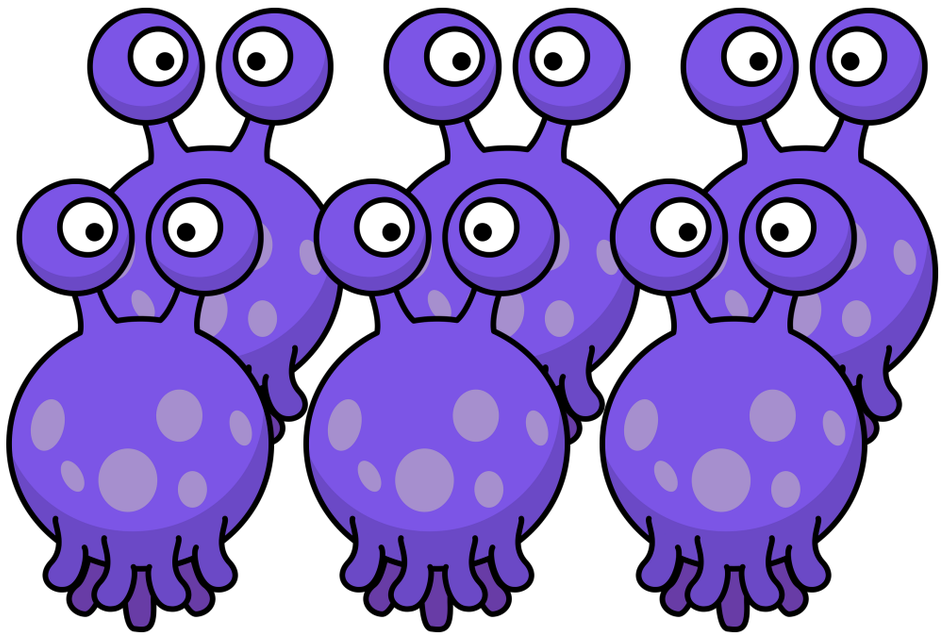
The Tidyverse

Import
Tidy
Manipulate/Transform
Visualise
Model
Report

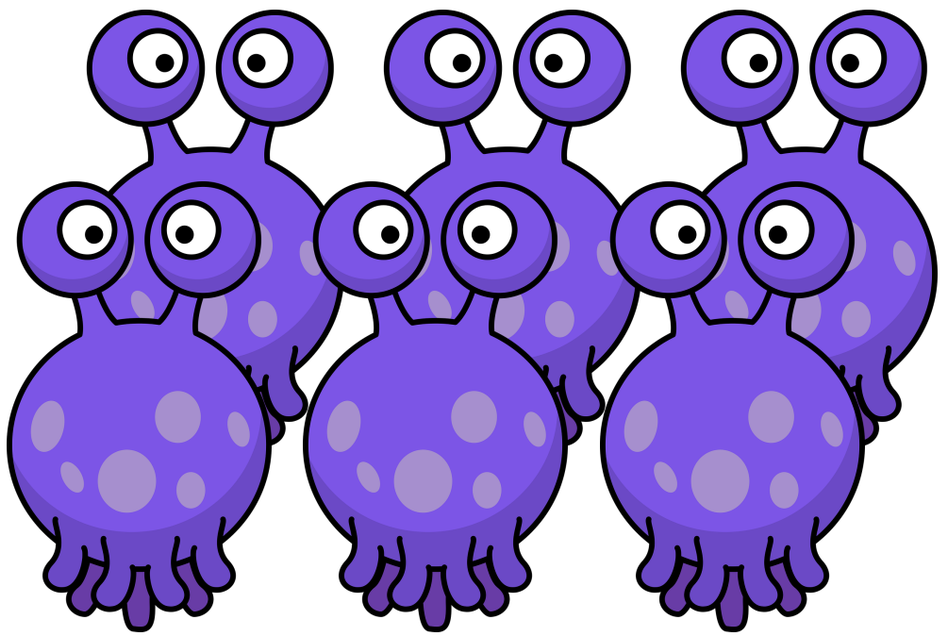
R and RStudio



Example - experiment



Treatment A



Treatment B

Glucose measured weekly (3 times) for 4 weeks

ih-trial_results_20171020

Search Sheet

Home | Insert | Page Layout | Formulas | Data | Review | View

Paste | Calibri (Body) | 11 | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | Editing

A1 | subject

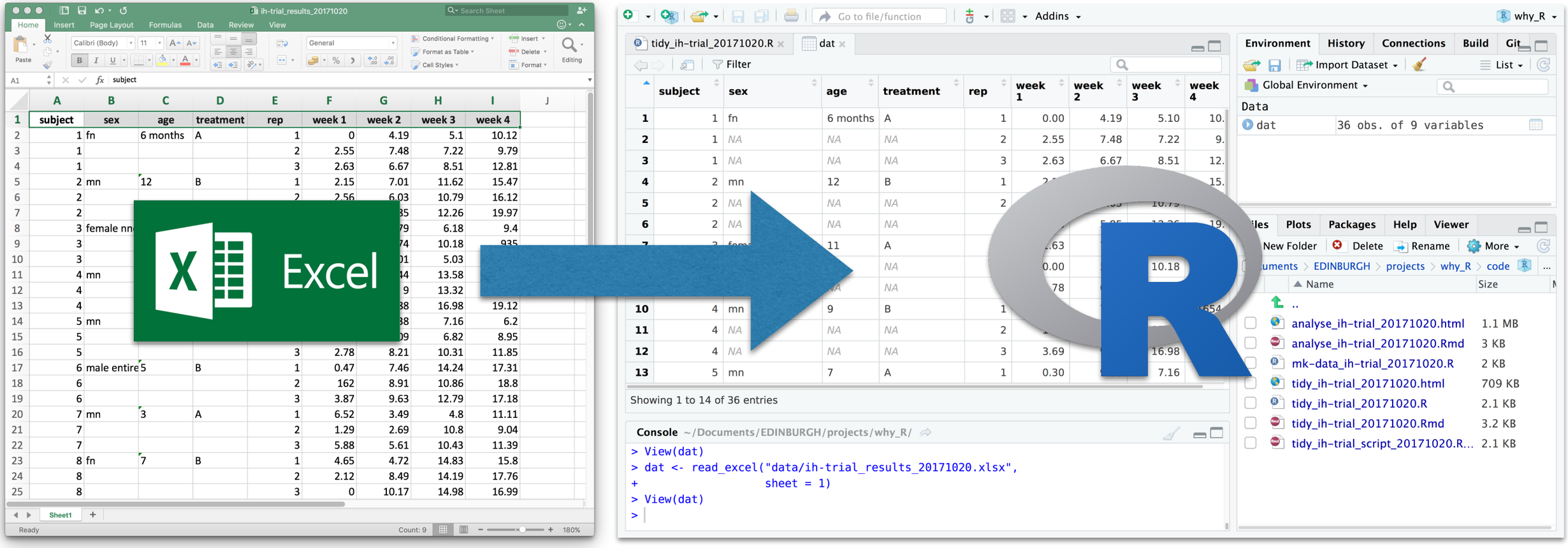
	A	B	C	D	E	F	G	H	I	J
1	subject	sex	age	treatment	rep	week 1	week 2	week 3	week 4	
2	1	fn	6 months	A	1	0	4.19	5.1	10.12	
3	1				2	2.55	7.48	7.22	9.79	
4	1				3	2.63	6.67	8.51	12.81	
5	2	mn	12	B	1	2.15	7.01	11.62	15.47	
6	2				2	2.56	6.03	10.79	16.12	
7	2				3	5.36	5.85	12.26	19.97	
8	3	female nne	11	A	1	1.63	7.79	6.18	9.4	
9	3				2	0	5.74	10.18	935	
10	3				3	5.78	6.01	5.03	13.86	
11	4	mn	9	B	1	3.45	7.44	13.58	1654	
12	4				2	1.36	9	13.32	19.09	
13	4				3	3.69	9.88	16.98	19.12	
14	5	mn	7	A	1	0.3	9.38	7.16	6.2	
15	5				2	4.08	9.09	6.82	8.95	
16	5				3	2.78	8.21	10.31	11.85	
17	6	male entire	5	B	1	0.47	7.46	14.24	17.31	
18	6				2	162	8.91	10.86	18.8	
19	6				3	3.87	9.63	12.79	17.18	
20	7	mn	3	A	1	6.52	3.49	4.8	11.11	
21	7				2	1.29	2.69	10.8	9.04	
22	7				3	5.88	5.61	10.43	11.39	
23	8	fn	7	B	1	4.65	4.72	14.83	15.8	
24	8				2	2.12	8.49	14.19	17.76	
25	8				3	0	10.17	14.98	16.99	

Sheet1 | +

Ready | Count: 9 | 180%

Import the data

```
dat <- read_excel("../data/ih-trial_results_20171020.xlsx",
                  sheet = 1)
```



Fill down missing data

```
dat <- dat %>%
  fill(sex, age, treatment)
```

	subject	sex	age	treatment	rep	week 1
1	1	fn	6 months	A	1	0.00
2	1	*NA*	*NA*	*NA*	2	2.55
3	1	*NA*	*NA*	*NA*	3	2.63
4	2	mn	12	B	1	2.15
5	2	*NA*	*NA*	*NA*	2	2.56
6	2	*NA*	*NA*	*NA*	3	5.36
7	3	female nneutered	11	A	1	1.63
8	3	*NA*	*NA*	*NA*	2	0.00
9	3	*NA*	*NA*	*NA*	3	5.78
10	4	mn	9	B	1	3.45
11	4	*NA*	*NA*	*NA*	2	1.36
12	4	*NA*	*NA*	*NA*	3	3.69

Fill down missing data

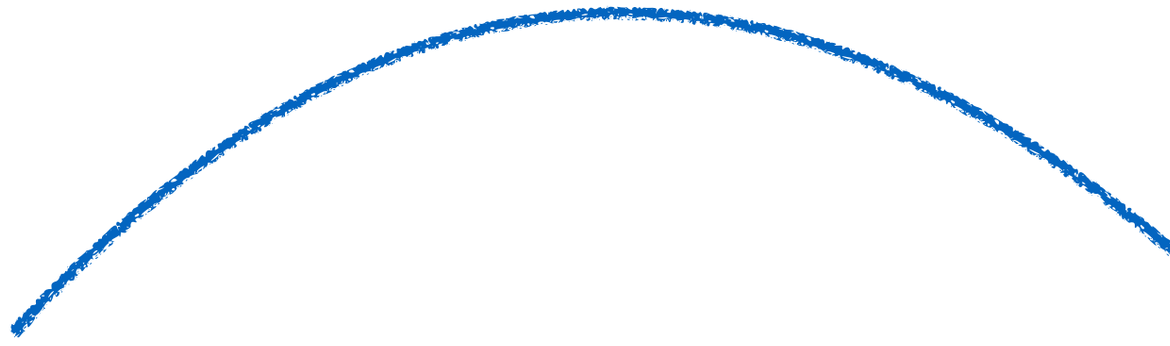
```
dat <- dat %>%
  fill(sex, age, treatment)
```

	subject	sex	age	treatment	rep	week 1
1	1	fn	6 months	A	1	0.00
2	1	*NA*	*NA*	*NA*	2	2.55
3	1	*NA*	*NA*	*NA*	3	2.63
4	2	mn	12	B	1	2.15
5	2	*NA*	*NA*	*NA*	2	2.56
6	2	*NA*	*NA*	*NA*	3	5.36
7	3	female nneutered	11	A	1	1.63
8	3	*NA*	*NA*	*NA*	2	0.00
9	3	*NA*	*NA*	*NA*	3	5.78
10	4	mn	9	B	1	3.45
11	4	*NA*	*NA*	*NA*	2	1.36
12	4	*NA*	*NA*	*NA*	3	3.69

	subject	sex	age	treatment	rep	week 1
1	1	fn	6 months	A	1	0.00
2	1	fn	6 months	A	2	2.55
3	1	fn	6 months	A	3	2.63
4	2	mn	12	B	1	2.15
5	2	mn	12	B	2	2.56
6	2	mn	12	B	3	5.36
7	3	female nneutered	11	A	1	1.63
8	3	female nneutered	11	A	2	0.00
9	3	female nneutered	11	A	3	5.78
10	4	mn	9	B	1	3.45
11	4	mn	9	B	2	1.36
12	4	mn	9	B	3	3.69

Create a sensible subject ID

```
dat <- dat %>%
  mutate(subject = str_c("A", str_pad(subject, 3, "left", "0")))
```



	subject	sex	age
1	1	fn	6 months
2	1	fn	6 months
3	1	fn	6 months
4	2	mn	12
5	2	mn	12
6	2	mn	12

	subject	sex	age
1	A001	fn	6 months
2	A001	fn	6 months
3	A001	fn	6 months
4	A002	mn	12
5	A002	mn	12
6	A002	mn	12

Examine the sex/status column

	subject	sex	age	treatment	rep
1	A001	fn	6 months	A	1
2	A001	fn	6 months	A	2
3	A001	fn	6 months	A	3
4	A002	mn	12	B	1
5	A002	mn	12	B	2
6	A002	mn	12	B	3
7	A003	female nneutered	11	A	1
8	A003	female nneutered	11	A	2
9	A003	female nneutered	11	A	3
10	A004	mn	9	B	1
11	A004	mn	9	B	2
12	A004	mn	9	B	3
13	A005	mn	7	A	1
14	A005	mn	7	A	2
15	A005	mn	7	A	3
16	A006	male entire	5	B	1
17	A006	male entire	5	B	2
18	A006	male entire	5	B	3
19	A007	mn	3	A	1
20	A007	mn	3	A	2
21	A007	mn	3	A	3

etc…

```
dat %>%
  group_by(sex) %>%
  tally()
```

sex	n
female nneutered	3
fn	9
male entire	6
mn	15
MN	3

Correct sex/status typos (etc)

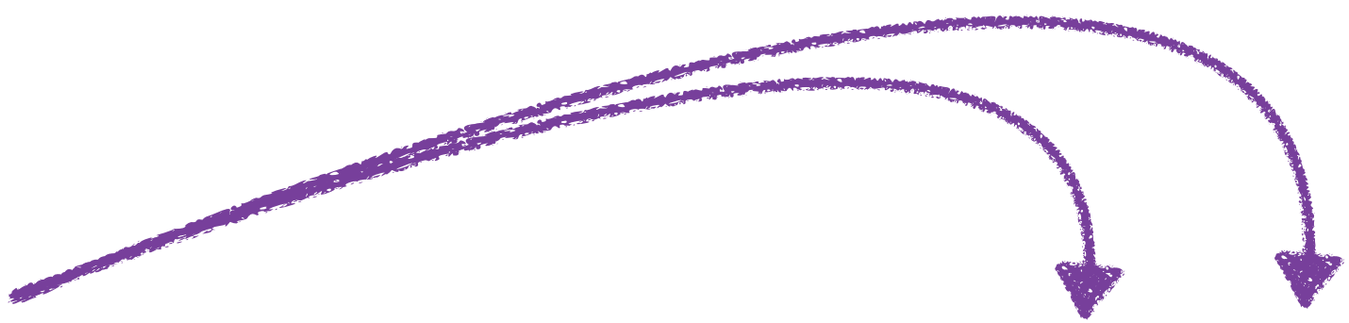
```
dat <- dat %>%
  mutate(sex = case_when(sex == "female nneutered" ~ "fn",
                         sex == "male entire" ~ "me",
                         sex == "MN" ~ "mn",
                         TRUE ~ sex))
```

	subject	sex	age	treatment
1	A001	fn	6 months	A
2	A001	fn	6 months	A
3	A001	fn	6 months	A
4	A002	mn	12	B
5	A002	mn	12	B
6	A002	mn	12	B
7	A003	female nneutered	11	A
8	A003	female nneutered	11	A
9	A003	female nneutered	11	A
10	A004	mn	9	B
11	A004	mn	9	B
12	A004	mn	9	B
13	A005	mn	7	A
14	A005	mn	7	A
15	A005	mn	7	A
16	A006	male entire	5	B
17	A006	male entire	5	B
18	A006	male entire	5	B

	subject	sex	age	treatment
1	A001	fn	6 months	A
2	A001	fn	6 months	A
3	A001	fn	6 months	A
4	A002	mn	12	B
5	A002	mn	12	B
6	A002	mn	12	B
7	A003	fn	11	A
8	A003	fn	11	A
9	A003	fn	11	A
10	A004	mn	9	B
11	A004	mn	9	B
12	A004	mn	9	B
13	A005	mn	7	A
14	A005	mn	7	A
15	A005	mn	7	A
16	A006	me	5	B
17	A006	me	5	B
18	A006	me	5	B

Separate sex/status into two columns

```
dat <- dat %>%
  separate(sex, c("sex", "neuter_status"), sep = 1)
```

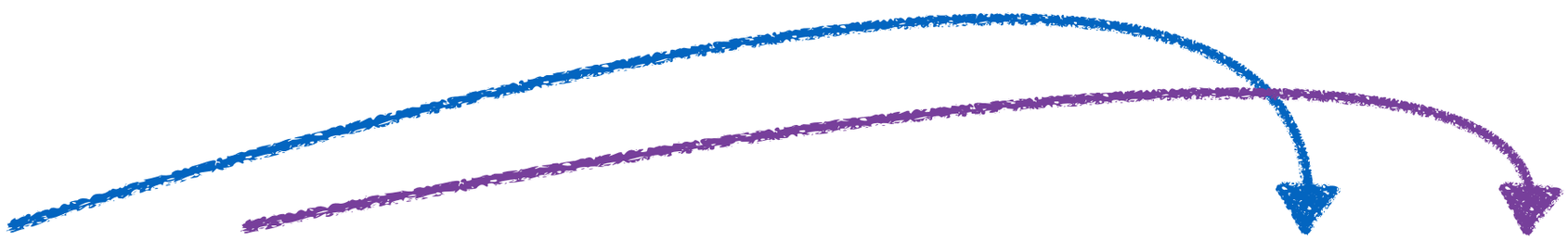


	subject	sex	age
1	A001	fn	6 months
2	A001	fn	6 months
3	A001	fn	6 months
4	A002	mn	12
5	A002	mn	12
6	A002	mn	12
7	A003	fn	11
8	A003	fn	11
9	A003	fn	11

	subject	sex	neuter_status	age
1	A001	f	n	6 months
2	A001	f	n	6 months
3	A001	f	n	6 months
4	A002	m	n	12
5	A002	m	n	12
6	A002	m	n	12
7	A003	f	n	11
8	A003	f	n	11
9	A003	f	n	11

Expand sex and status labels

```
dat <- dat %>%
  mutate(sex = fct_recode(sex,
                          male = "m",
                          female = "f"),
         neuter_status = fct_recode(neuter_status,
                                    neutered = "n",
                                    entire = "e"))
```

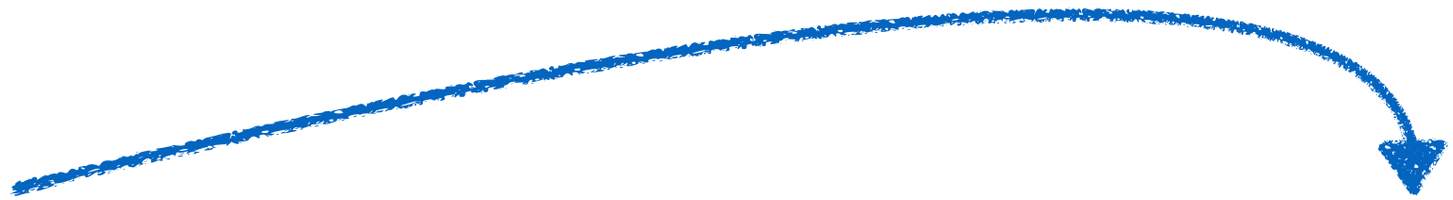


	subject	sex	neuter_status
1	A001	f	n
2	A001	f	n
3	A001	f	n
4	A002	m	n
5	A002	m	n
6	A002	m	n
7	A003	f	n

	subject	sex	neuter_status
1	A001	female	neutered
2	A001	female	neutered
3	A001	female	neutered
4	A002	male	neutered
5	A002	male	neutered
6	A002	male	neutered
7	A003	female	neutered

Convert ages in ‘months’ to ‘years’

```
dat <- dat %>%
  mutate(age = case_when(
                    str_detect(age, "month") ~ parse_number(age) / 12,
                    TRUE ~ parse_number(age)))
```



	subject	sex	neuter_status	age
1	A001	female	neutered	6 months
2	A001	female	neutered	6 months
3	A001	female	neutered	6 months
4	A002	male	neutered	12
5	A002	male	neutered	12
6	A002	male	neutered	12
7	A003	female	neutered	11
8	A003	female	neutered	11
9	A003	female	neutered	11

	subject	sex	neuter_status	age
1	A001	female	neutered	0.5
2	A001	female	neutered	0.5
3	A001	female	neutered	0.5
4	A002	male	neutered	12.0
5	A002	male	neutered	12.0
6	A002	male	neutered	12.0
7	A003	female	neutered	11.0
8	A003	female	neutered	11.0
9	A003	female	neutered	11.0

Gather the 4 results columns

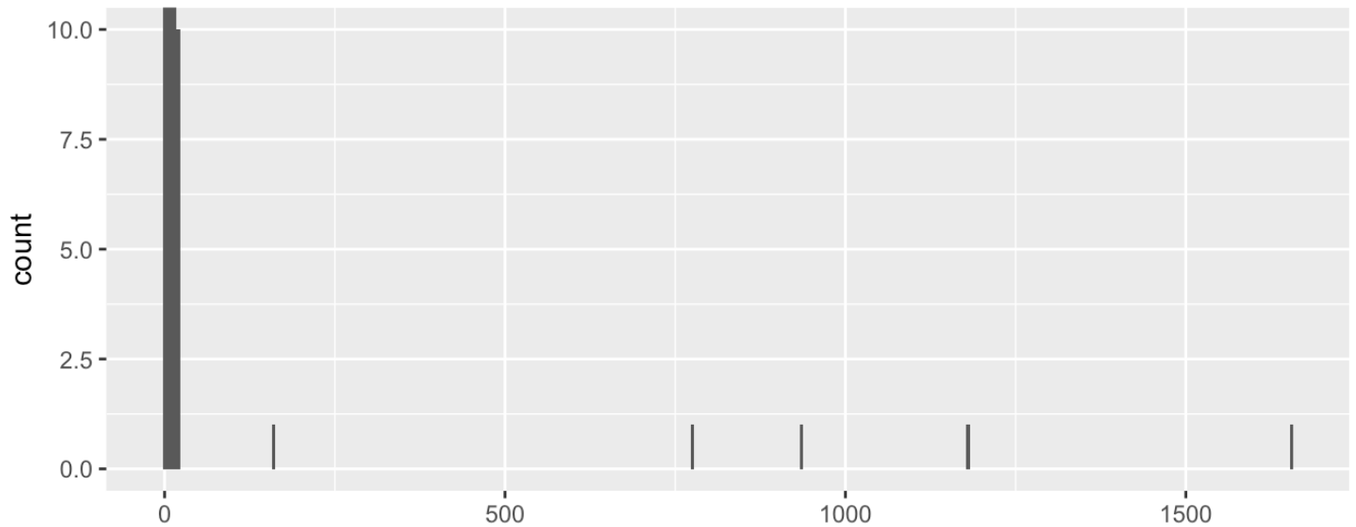
```
dat <- dat %>%
  gather("week", "glucose", `week 1`:`week 4`) %>%
  mutate(week = parse_number(week))
```

	subject	rep	week 1	week 2	week 3	week 4
1	A001	1	0.00	4.19	5.10	10.12
2	A001	2	2.55	7.48	7.22	9.79
3	A001	3	2.63	6.67	8.51	12.81

subject	rep	week	glucose
A001	1	1	0.00
A001	1	2	4.19
A001	1	3	5.10
A001	1	4	10.12
A001	2	1	2.55
A001	2	2	7.48
A001	2	3	7.22
A001	2	4	9.79
A001	3	1	2.63
A001	3	2	6.67
A001	3	3	8.51
A001	3	4	12.81

Explore the biochem results

```
ggplot(dat, aes(x = glucose)) +
  geom_histogram(binwidth = 5) +
  coord_cartesian(ylim = c(0, 10))
```




```
dat %>%
  filter(glucose > 50) %>%
  select(subject, week, rep, glucose)
```

subject	week	rep	glucose
A006	1	2	162
A012	2	2	776
A012	3	1	1178
A003	4	2	935
A004	4	1	1654

Correct age data typos

```
dat <- dat %>%
  mutate(glucose = case_when(subject == "A006" &
                               week == 1 &
                               rep == 2 ~          1.62,

                             subject == "A012" &
                               week == 2 &
                               rep == 2 ~          7.76,

                             subject == "A012" &
                               week == 3 &
                               rep == 1 ~          11.78,

                             subject == "A003" &
                               week == 4 &
                               rep == 2 ~          9.35,

                             subject == "A004" &
                               week == 4 &
                               rep == 1 ~          16.54,

                             TRUE ~                glucose))
```

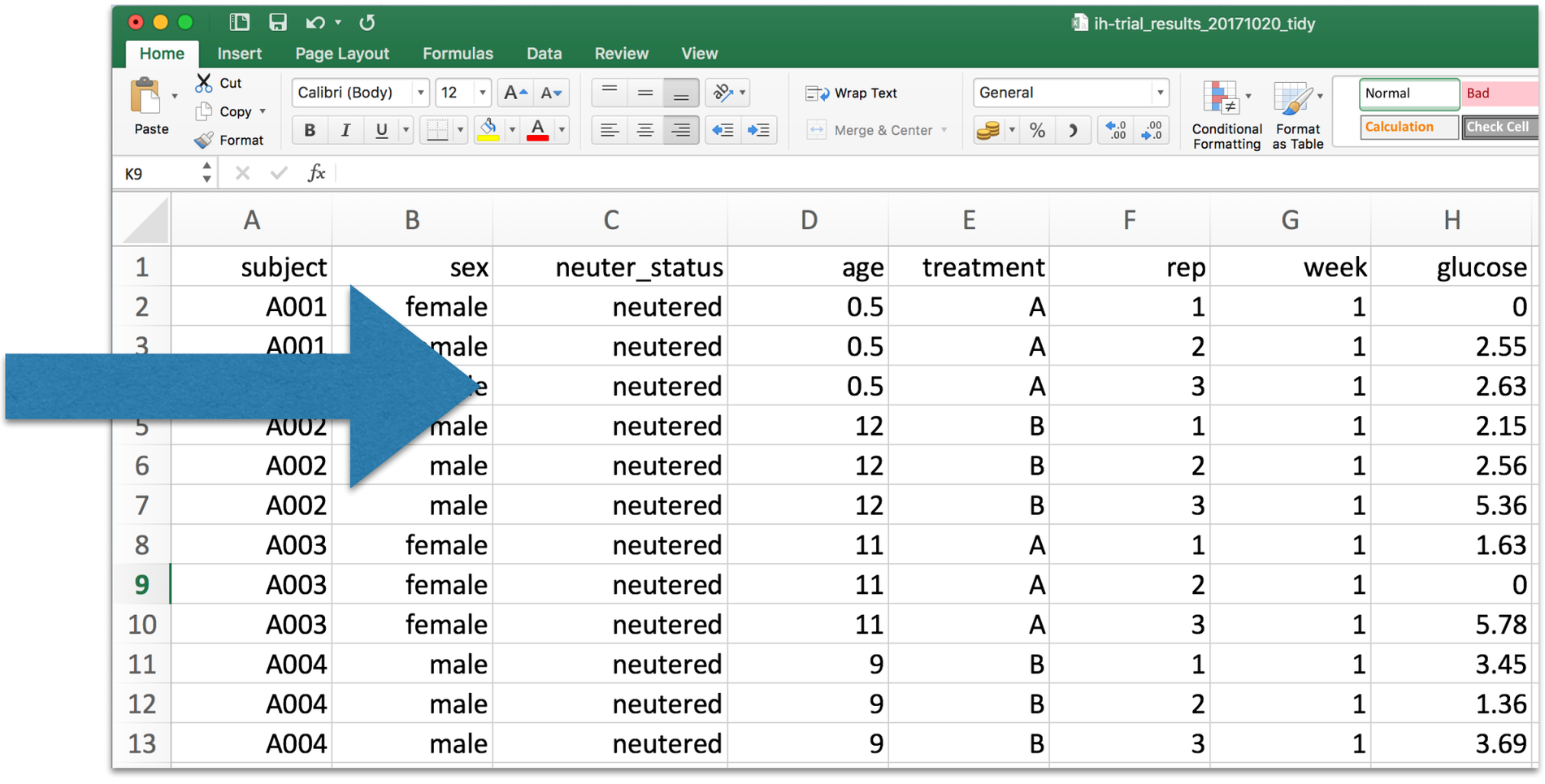
subject	rep	week	glucose
A006	1	1	0.47
A006	2	1	162.00
A006	3	1	3.87

subject	rep	week	glucose
A006	1	1	0.47
A006	2	1	1.62
A006	3	1	3.87

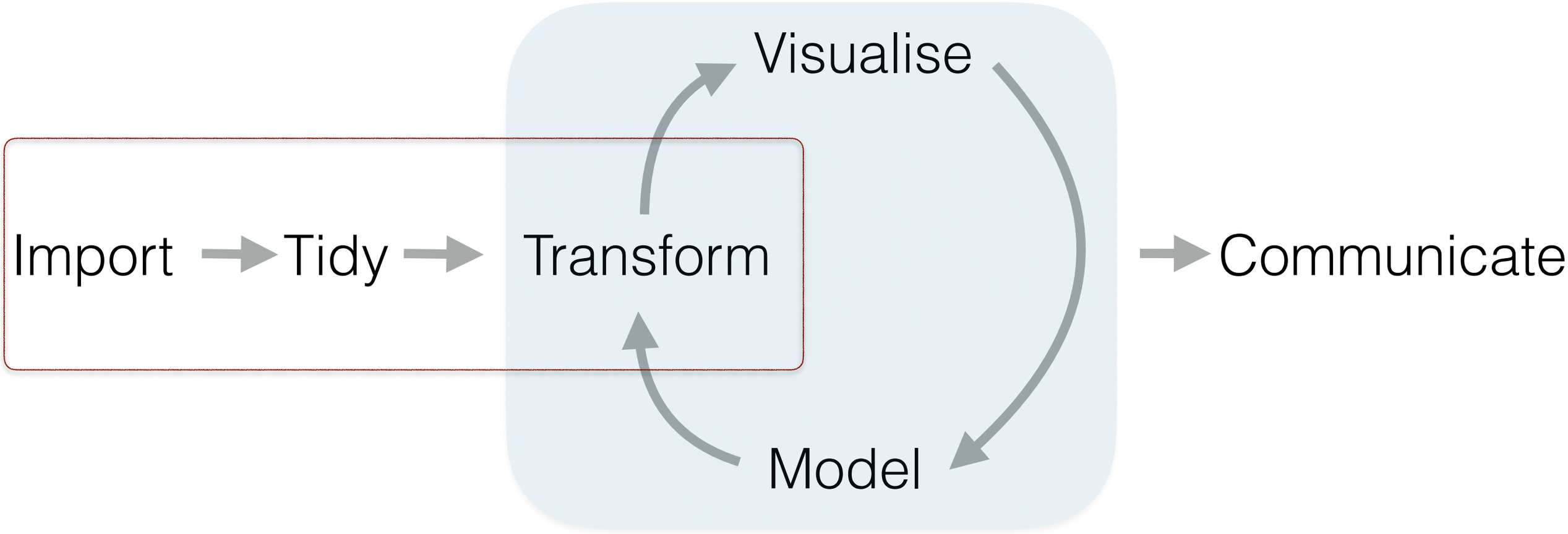
etc…

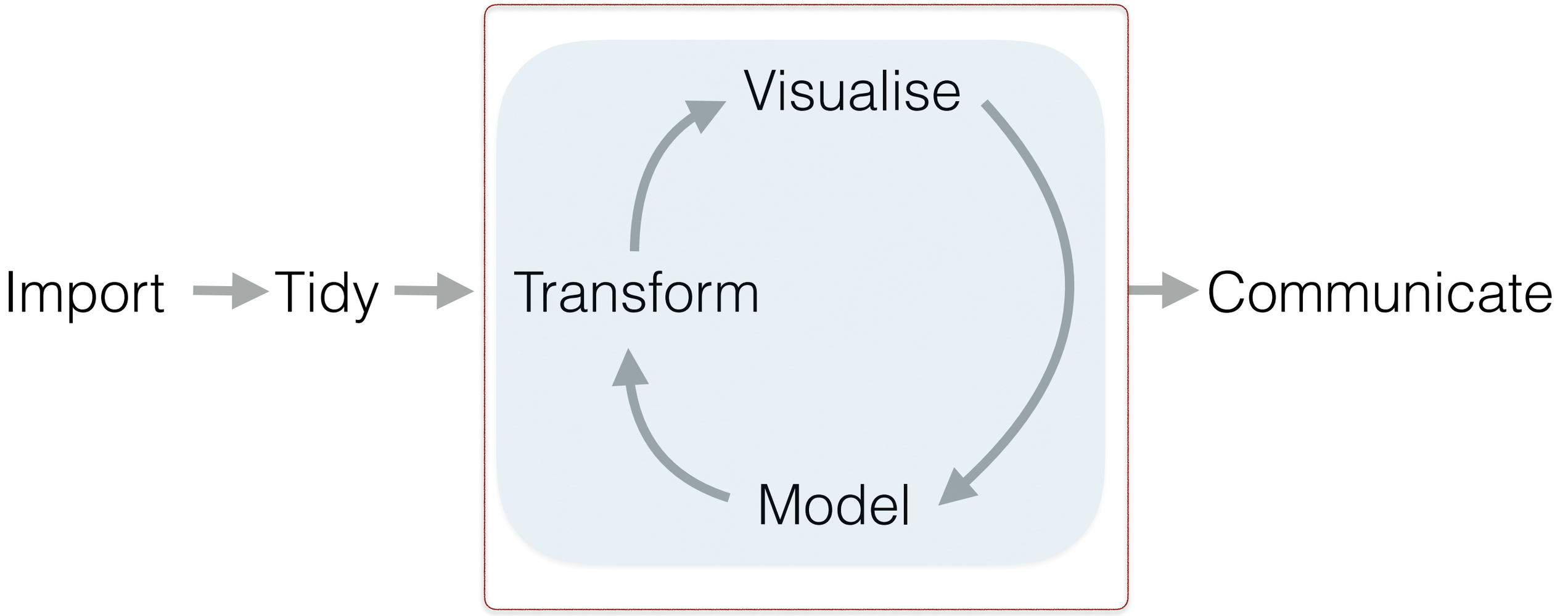
Save tidied data

```
write_csv(dat, "../data/ih-trial_results_20171020_tidy.csv")
```



	A	B	C	D	E	F	G	H
1	subject	sex	neuter_status	age	treatment	rep	week	glucose
2	A001	female	neutered	0.5	A	1	1	0
3	A001	male	neutered	0.5	A	2	1	2.55
4		e	neutered	0.5	A	3	1	2.63
5	A002	male	neutered	12	B	1	1	2.15
6	A002	male	neutered	12	B	2	1	2.56
7	A002	male	neutered	12	B	3	1	5.36
8	A003	female	neutered	11	A	1	1	1.63
9	A003	female	neutered	11	A	2	1	0
10	A003	female	neutered	11	A	3	1	5.78
11	A004	male	neutered	9	B	1	1	3.45
12	A004	male	neutered	9	B	2	1	1.36
13	A004	male	neutered	9	B	3	1	3.69


Visualise
Import
Tidy
Transform
Communicate
Model


Import
Tidy
Transform
Visualise
Model
Communicate

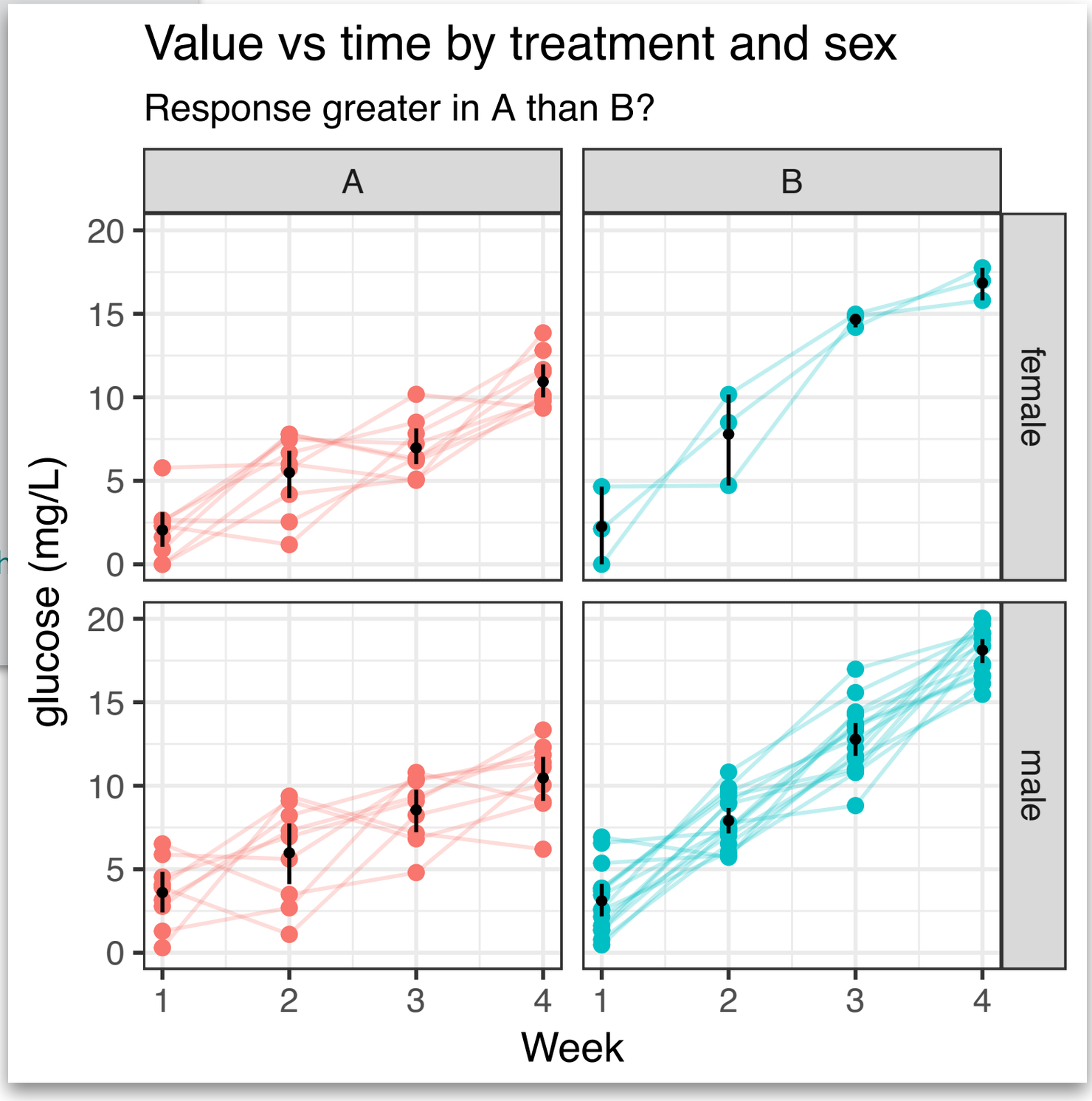
Summarise data

```
dat %>%
  select(subject, treatment, age) %>%
  distinct() %>%
  group_by(treatment) %>%
  summarise(n = sum(!is.na(age)),
            mean = mean(age),
            median = median(age),
            sd = sd(age),
            min = min(age),
            max = max(age)) %>%
  ungroup()
```

treatment	n	mean	median	sd	min	max
A	72	5.08	4.5	3.42	0.5	11
B	72	8.00	8.0	2.60	5.0	12

Visualise data

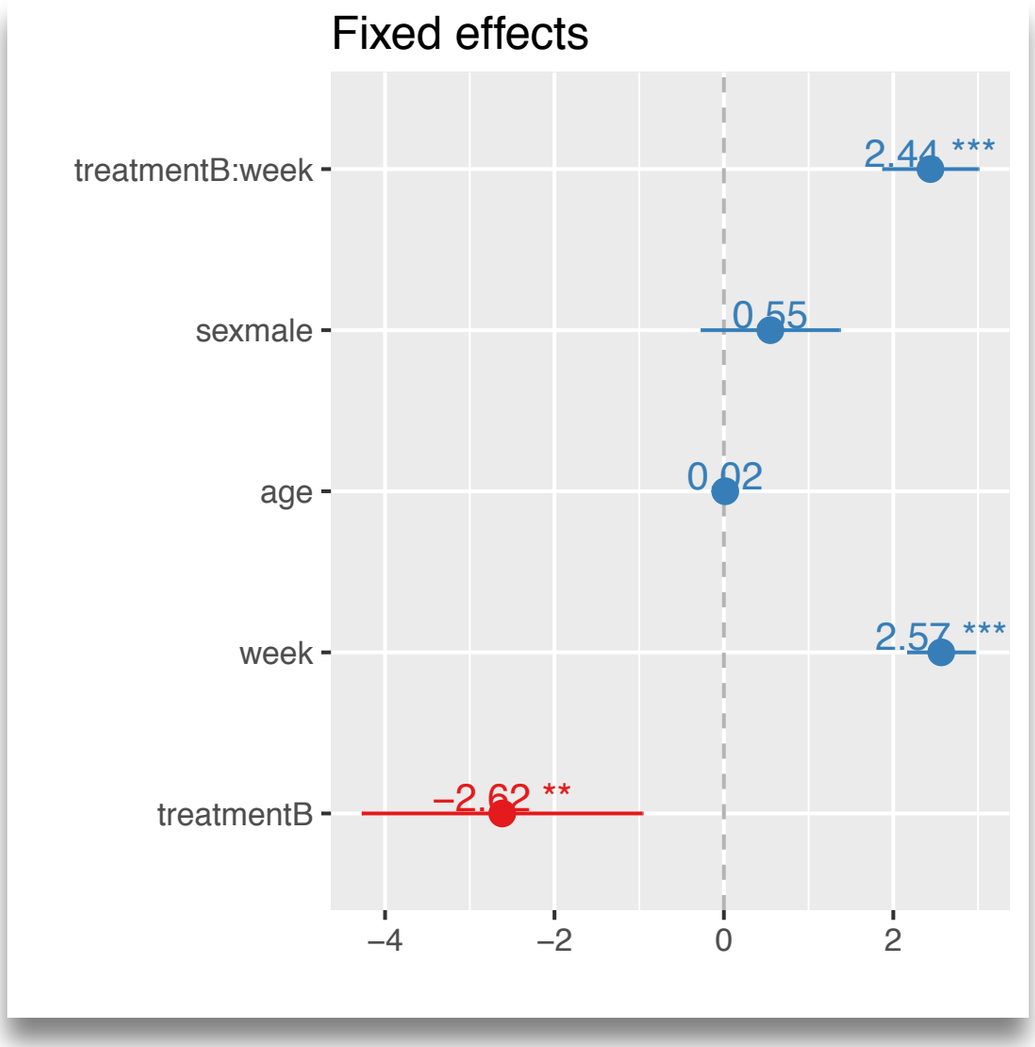
```
ggplot(dat) +
  aes(week, glucose,
      group = paste(subject, rep),
      colour = treatment) +
  geom_point() +
  geom_line(alpha = 0.25) +
  stat_summary(fun.data = "mean_cl_boot",
               geom = "pointrange",
               size = 0.5,
               fatten = 0.2,
               colour = "black",
               group = 1) +
  facet_grid(sex ~ treatment) +
  guides(colour = FALSE) +
  theme_bw() +
  labs(title = "Value vs time by treatment
       subtitle = "Response greater in A th
       x = "Week",
       y = "glucose (mg/L)")
```

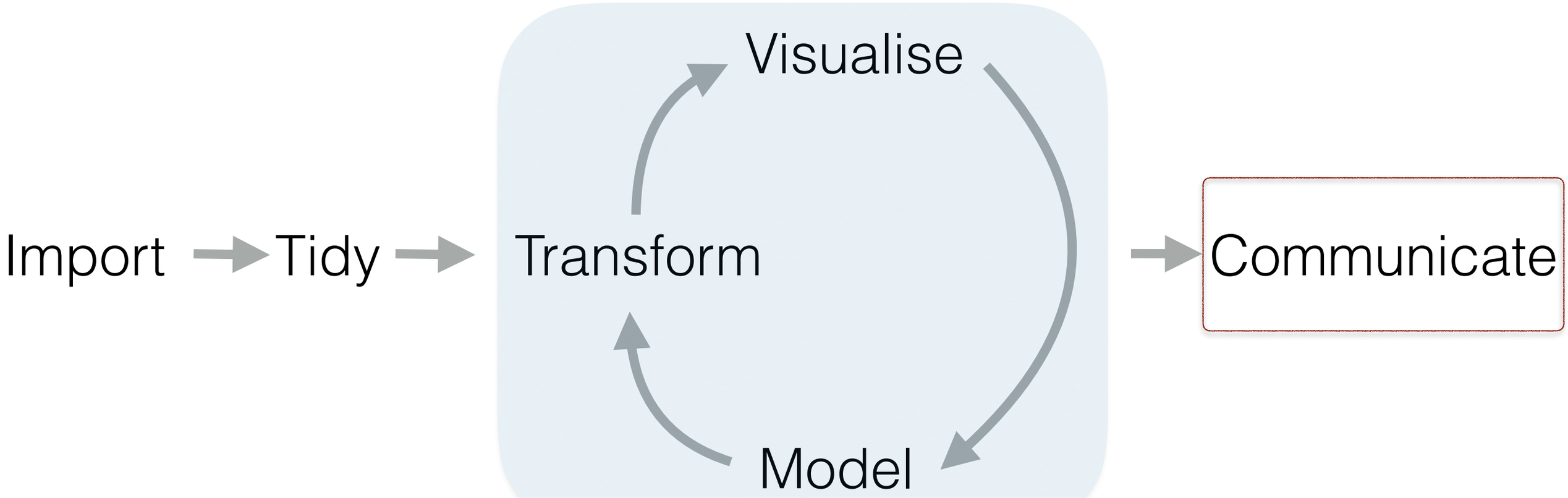


Do some ‘stats’

```
mod <- lmer(glucose ~ treatment * week + age + sex + (1 | subject), data = dat)
print(summary(mod)$coef, digits = 2)
sjPlot::sjp.lmer(mod, type = "fe")
```

```
Fixed effects:
                Estimate Std. Error t value
(Intercept)       -0.014      0.693   -0.02
treatmentB        -2.617      0.846   -3.09
week               2.566      0.205   12.53
age                0.016      0.061    0.26
sexmale            0.549      0.419    1.31
treatmentB:week    2.439      0.290    8.42
```




Import
Tidy
Visualise
Transform
Model
Communicate

+ RMarkdown

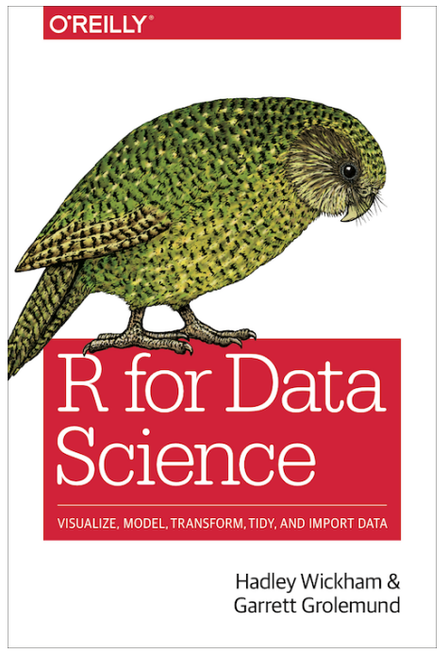
- Excel hides the logic/method/program
- R hides the data
- RMarkdown lets you see both data and method

Reproducible research?

- To demonstrate correctness / audit
- To allow others to follow methods
- To remember what happened…
  - Students graduate
  - Future you!

What next?

- Half day workshop here (the saga continues)
- R for Datascience - Book / Free eBook
- Datacamp
- Coursera
- & the ‘Andy Law method’



http://r4ds.had.co.nz

Data-methods-club

- How to get help (with R and other stuff)
- Half day workshop on R
- DMC Mailing list - just ask
- Website - resources

attendance