dplyr and pipes

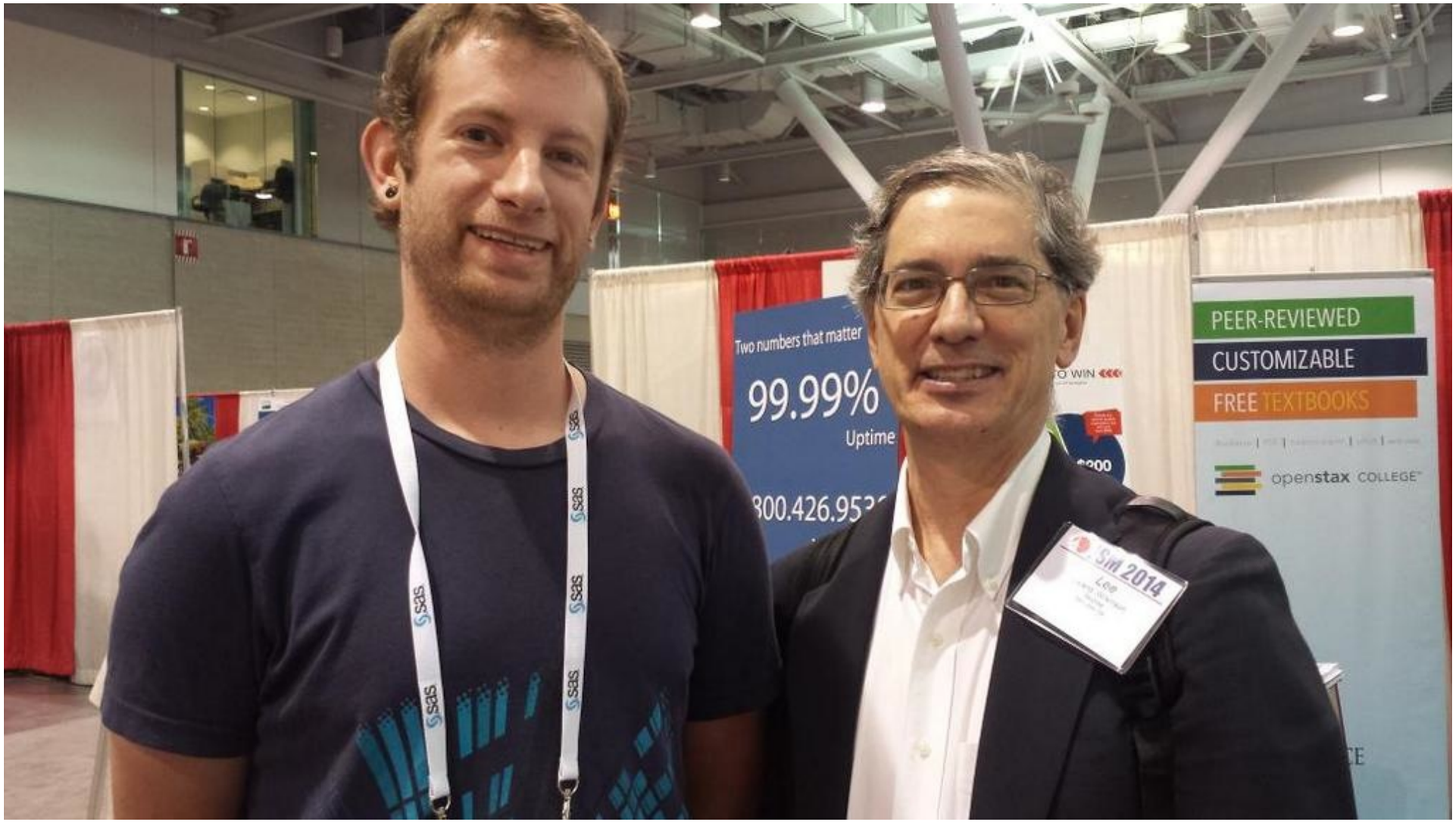
Sean Anderson
sean@seananderson.ca
@sean_anderson

Stats Beerz September 2014
Simon Fraser University

CC BY NC Many slides from workshop by Hadley Wickham, June 2014

Workshop notes and files:

http://tinyurl.com/dplyr-1



https://twitter.com/knitesh/status/496670944963080192

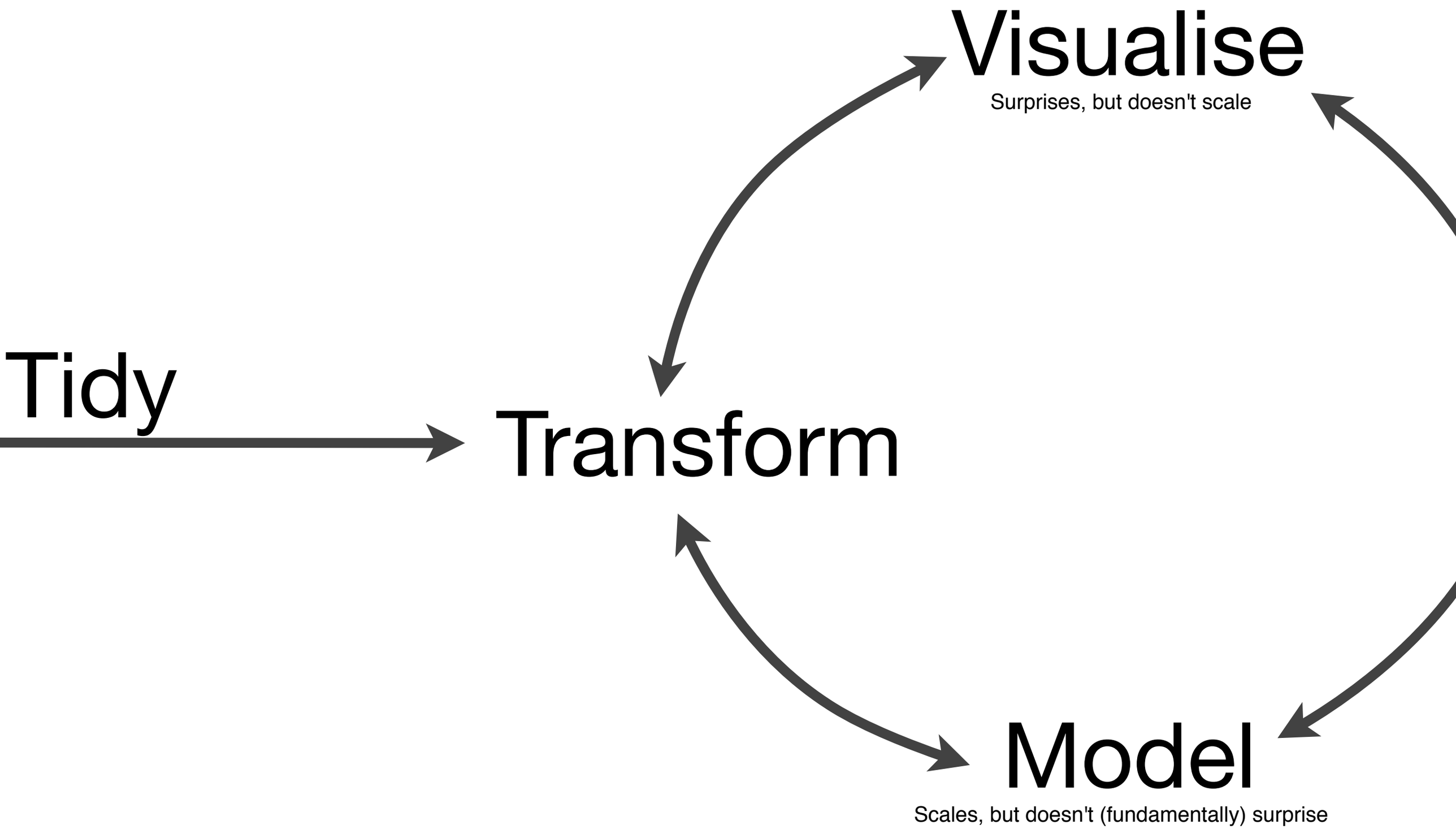
Data analysis is the process by which data becomes understanding, knowledge and insight

Data analysis is the process by which data becomes understanding, knowledge and insight

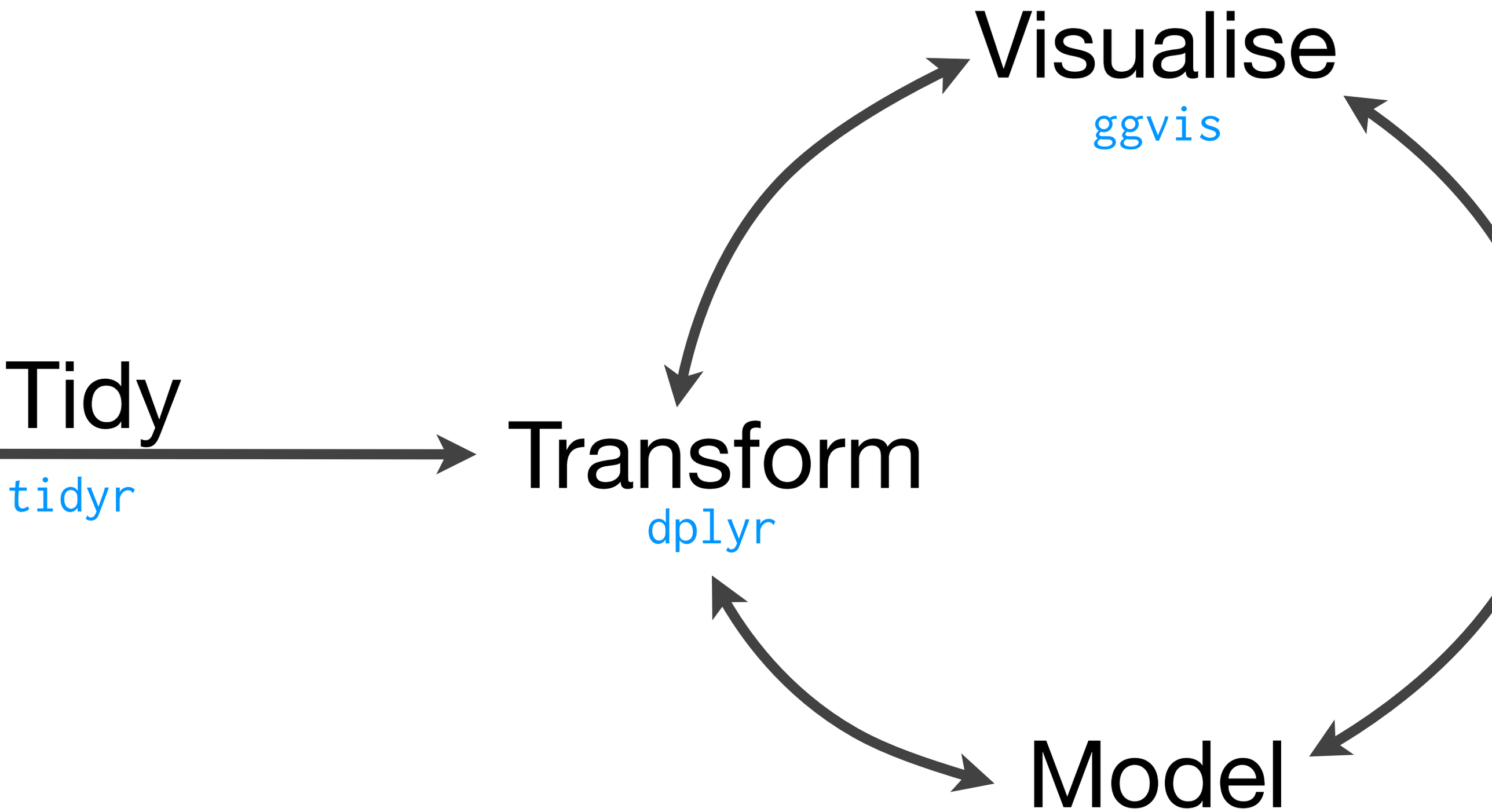
Slide by Hadley Wickham

Data analysis is the process by which data becomes understanding, knowledge and insight

Slide by Hadley Wickham



Slide by Hadley Wickham



Slide by Hadley Wickham

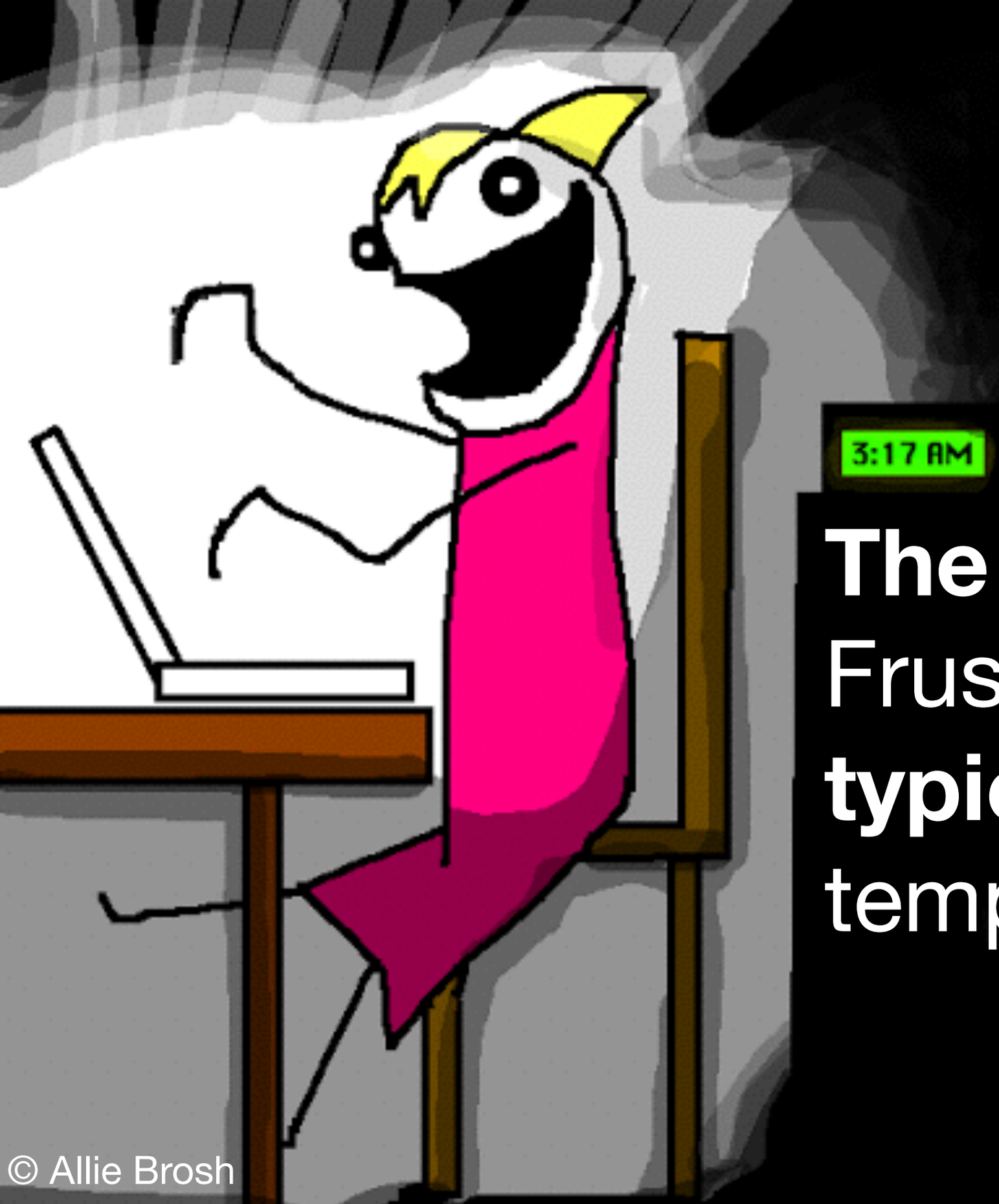
1. Mammals
2. Verbs
3. Pipes

The bad news:
It’s going to be frustrating

Slide by Hadley Wickham

© Allie Brosh

http://hyperboleandahalf.blogspot.com/2010/09/four-levels-of-social-entrapment.html



The good news:
Frustration is **typical** and temporary

Slide by Hadley Wickham

© Allie Brosh

http://hyperboleandahalf.blogspot.com/2010/06/this-is-why-ill-never-be-adult.html

Rstudio projects

- Isolate code and results from different projects. Restart where you left off.
- Double-click `dplyr-intro.Rproj` file to open. (One R file for each section)
- (If you don't use RStudio, just change working directories)

Slide by Hadley Wickham

PanTHERIA mammals data

Jones, K.E., *et al*. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90:2648. <http://esapubs.org/archive/ecol/E090/184/>

```
library(dplyr)
source("1-data.R")

mammals
glimpse(mammals)
```

dplyr verbs

select

```
df <- data.frame(
  color = c("blue", "black", "blue", "blue", "black"),
  value = 1:5)
```

Slide by Hadley Wickham

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

color
blue
black
blue
blue
black

```
select(df, color)
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

value
1
2
3
4
5

```
select(df, -color)
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

Read `?select`

What are 3 different ways you could select columns from `mammals`?

filter

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

color	value
blue	1
blue	3
blue	4

```
filter(df, color == "blue")
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

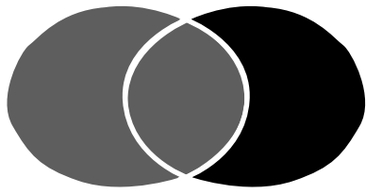
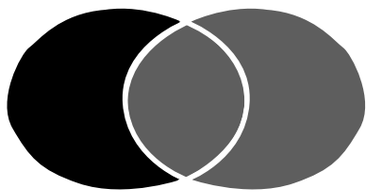
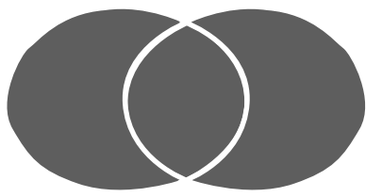
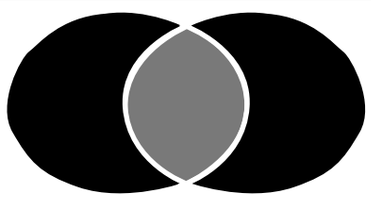
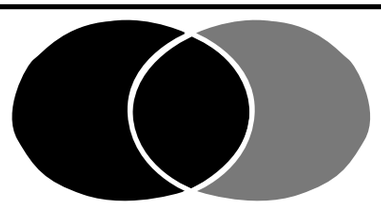
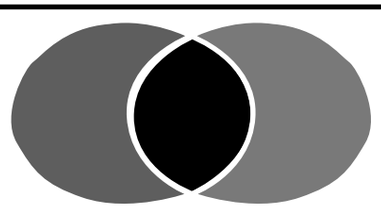
→

color	value
blue	1
blue	4

```
filter(df, value %in% c(1, 4))
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

	a
	b
	a	b
	a & b
	a & !b
	xor(a, b)

```
x > 1
x >= 1
x < 1
x <= 1
x != 1
x == 1
x %in% ("a", "b")
```

Slide by Hadley Wickham

1. Which mammals are heavier than 10,000 kg (1e7 g)?
2. Which mammals from the order Carnivora weigh less than 200g?

arrange

df

color	value
4	1
1	2
5	3
3	4
2	5

→

color	value
1	2
2	5
3	4
4	1
5	3

```
arrange(df, color)
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

df

color	value
4	1
1	2
5	3
3	4
2	5

→

color	value
5	3
4	1
3	4
2	5
1	2

```
arrange(df, desc(color))
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

1. Arrange mammals from small to large body mass.
2. Arrange mammals from large to small body mass.
3. Arrange mammals from small to large body mass within taxonomic order.

mutate

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

color	value	double
blue	1	2
black	2	4
blue	3	6
blue	4	8
black	5	10

```
mutate(df, double = 2 * value)
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

color	value	double	quadruple
blue	1	2	4
black	2	4	8
blue	3	6	12
blue	4	8	16
black	5	10	20

```
mutate(df, double = 2 * value,
    quadruple = 2 * double)
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

1. Make a new column with mass per body length.
2. Try making one column in g/mm and another in kg/mm.

summarise

(with group_by)

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

total
15

```
summarise(df, total = sum(value))
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

color	total
blue	8
black	7

```
by_color <- group_by(df, color)
summarise(by_color, total = sum(value))
```

Slide by Hadley Wickham

© 2014 RStudio, Inc. All rights reserved.

Summary functions

- `min(x), median(x), max(x), quantile(x, p)`
- `n(), n_distinct(), sum(x), mean(x)`
- `sum(x > 10), mean(x > 10)`
- `sd(x), var(x), iqr(x), mad(x)`

Slide by Hadley Wickham

1. What is the mean adult body mass?
2. What is the mean adult body mass by taxonomic order?

Data pipelines

$$\% > \%$$

```
f(data, …)
```

```
data %>% f(…)
```

try pronouncing %>% “then”

https://twitter.com/romain_francois/status/511791664570970112

https://twitter.com/hadleywickham/status/494875052467167232

https://twitter.com/yokkuns/status/505679441381433344

```
coffee %>%
grind(as="espresso") %>%
brew %>%
increase_productivity(by="drinking")
```

https://twitter.com/stefanbache/status/511911497346584576

```
increase_productivity(
  brew(
    grind(coffee,
          as="espresso")
    ),
  by="drinking
)
```

```
x <- c(1, 2, 3)
x %>% max
x %>% plot
x %>% sum(4)
```

1. Pipe mammals to the arrange function to sort by body mass.
2. Using pipes: Find the species with the highest mass to length ratio, sort by decreasing body-mass ratio, and select this new column.
3. Using pipes: What taxonomic orders have a median litter size greater than 3?

More information

https://github.com/hadley/dplyr

http://cran.r-project.org/web/packages/dplyr/vignettes/introduction.html

https://github.com/smbache/magrittr

http://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html