http://courses.had.co.nz

Data science in R

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

July 2012

http://courses.had.co.nz

Data science in R

Hadley Wickham

Data scientist in resdience

Metamarkets

July 2012

HELLO

my name is

Hadley

Your turn

Introduce yourself to your neighbour. What do you want to get out of this course? What sort of data are you working with?

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

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

R

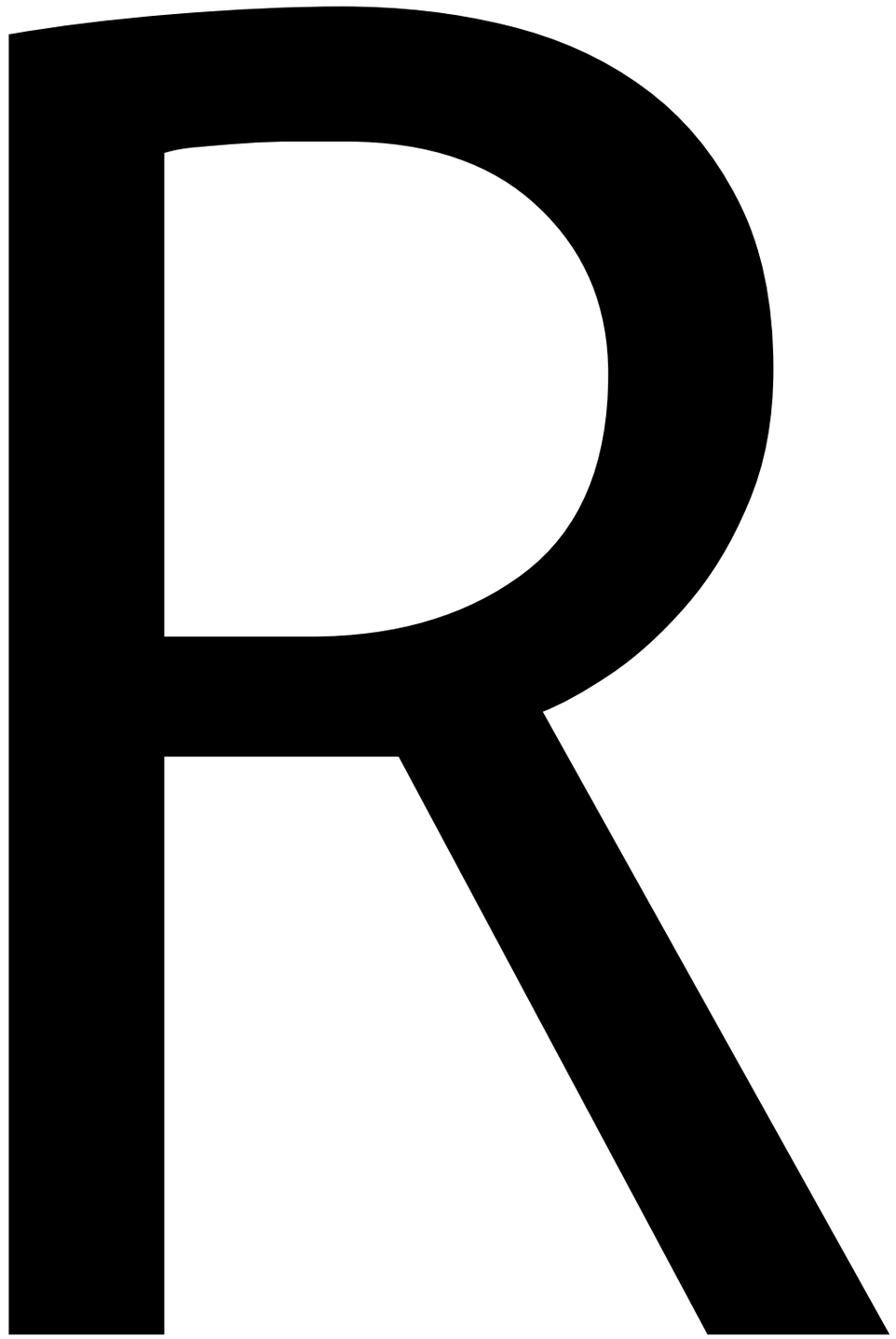
R

+

js

python

data scientists

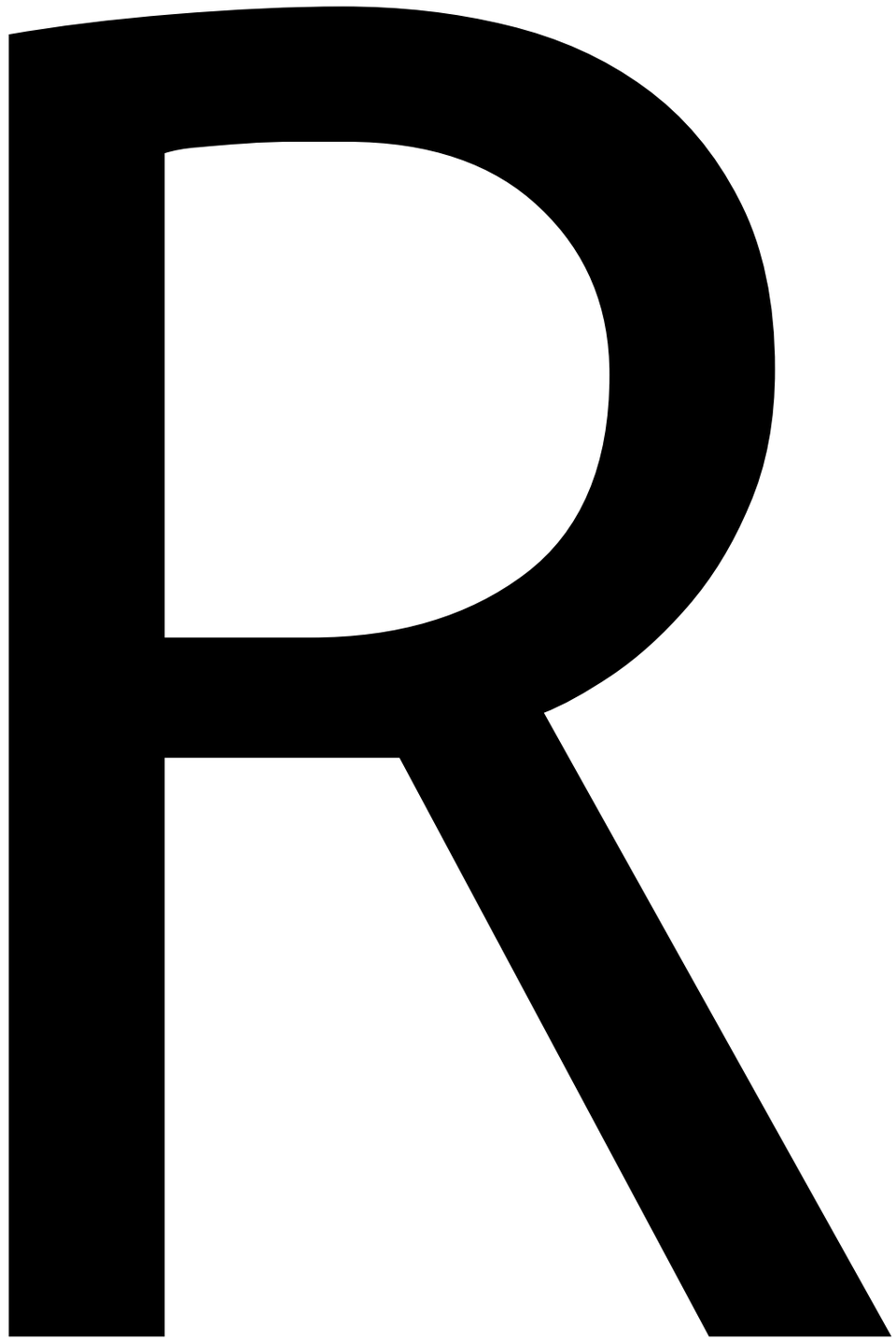


+
js
python

data scientists

+
sql
regex
xpath

data languages

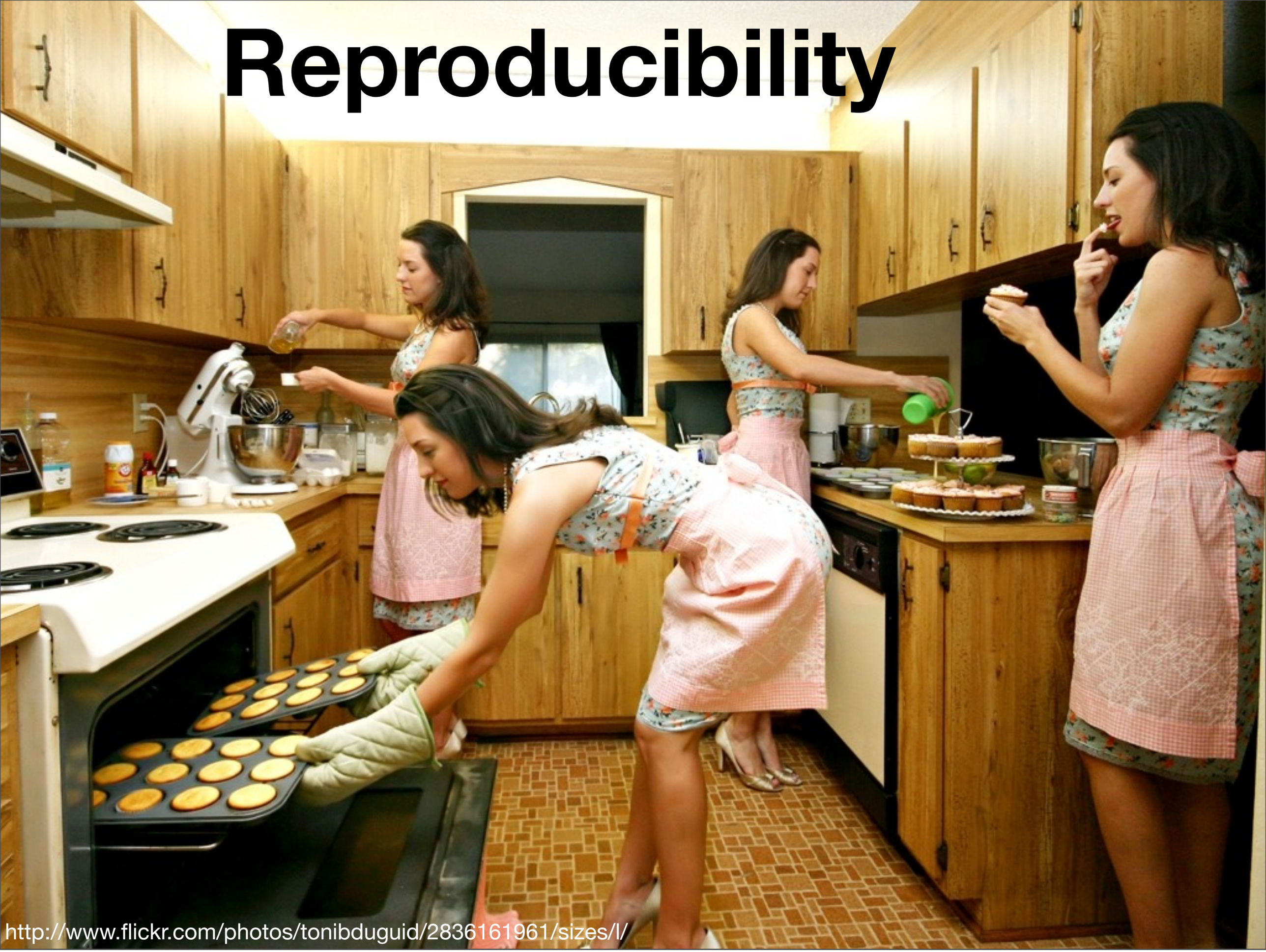


+
js
python
data scientists

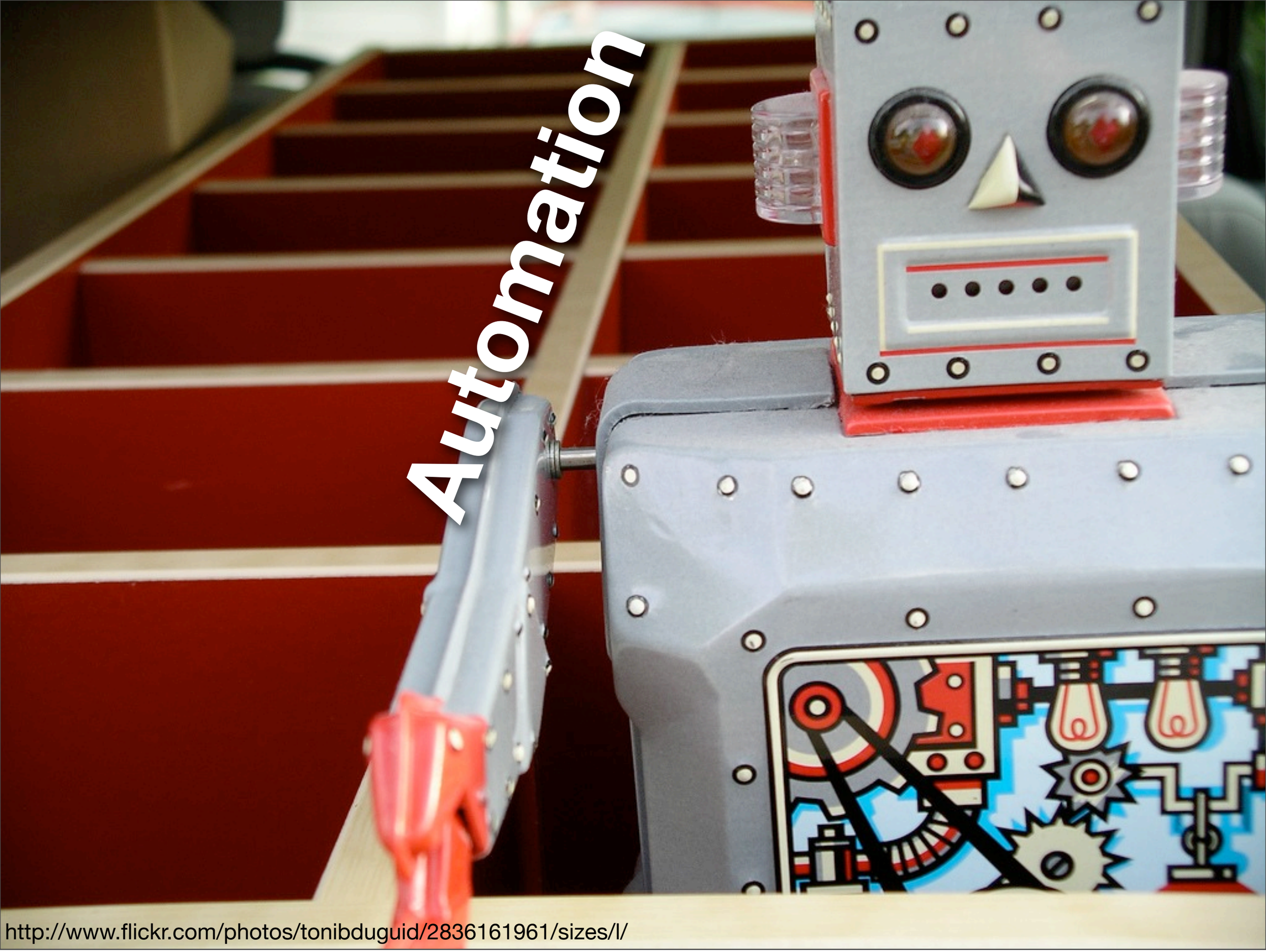
+
sql
regex
xpath
data languages

+
C/C++
fortran
scala
tool builders

Reproducibility



http://www.flickr.com/photos/tonibduguid/2836161961/sizes/l/



http://www.flickr.com/photos/tonibduguid/2836161961/sizes/l/

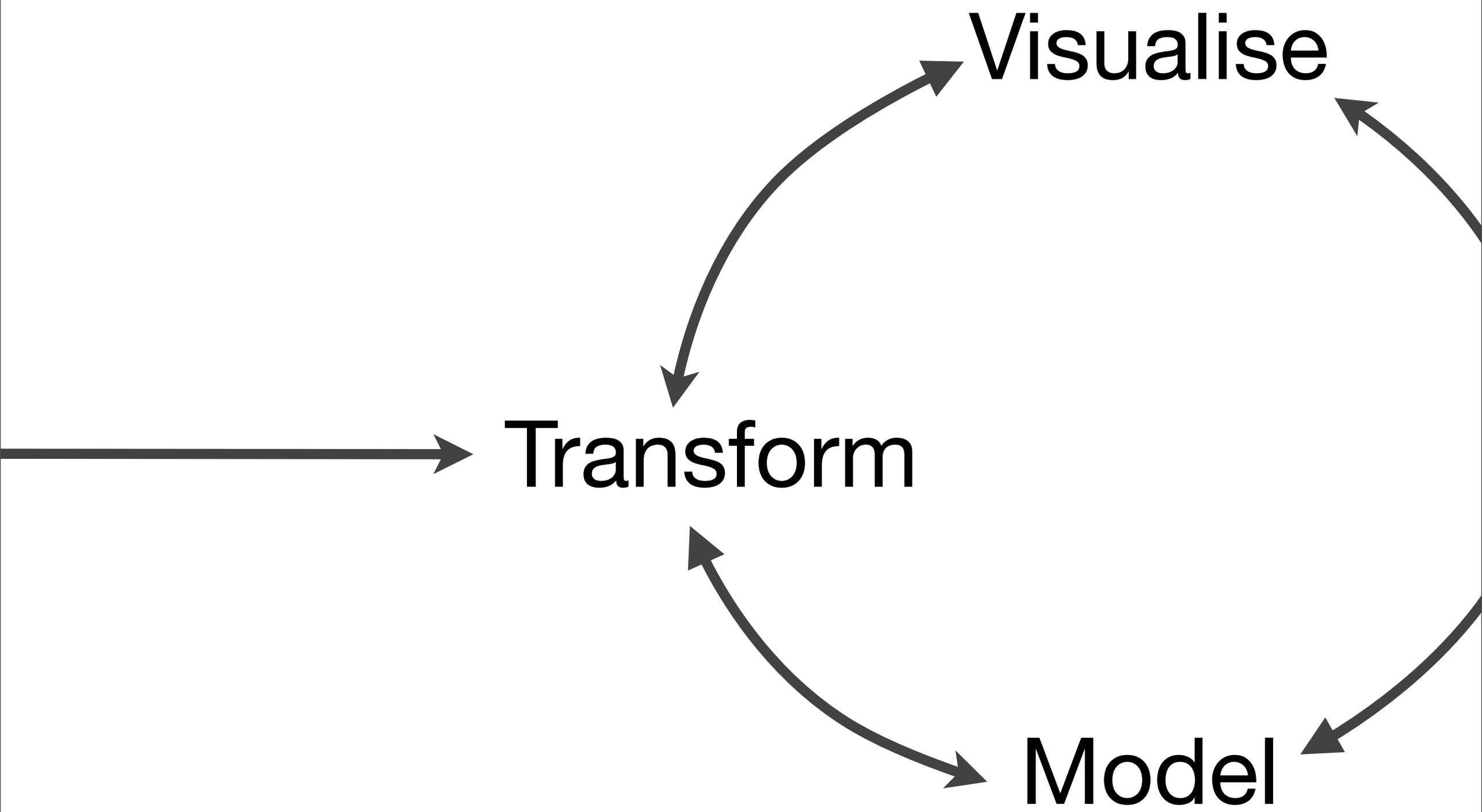
Just text

```
# Load data and create smaller subsets
tb <- read.csv("tb.csv")
tb2008 <- subset(tb, year == 2008)

# Choropleth map ------------------------------------------------------------
borders <- read.csv("world-borders.csv")
choro <- merge(tb2008, borders, by = "iso2")
choro <- choro[order(choro$order), ]

qplot(long, lat, data = choro, fill = cut_number(rate, 5), geom = "polygon", group =
group) + scale_fill_brewer("Rate", pal = "Blues")

# Bubble maps ---------------------------------------------------------------
centres <- read.csv("world-centres.csv")
bubble <- merge(centres, tb2008, by = "iso2")

world_coord <- coord_map(xlim = c(-180, 180), ylim = c(-50, 70))

# This is basically what a choropleth is showing us
qplot(long, lat, data = bubble, size = area, colour = rate) +
  scale_area(to = c(2, 25), legend = FALSE) +
  world_coord

# More traditional options
qplot(long, lat, data = bubble, size = rate) + world_coord
qplot(long, lat, data = bubble, size = log10(pop), colour = rate) +
  world_coord

# Even better if we add world boundaries
ggplot(bubble, aes(long, lat)) +
  geom_polygon(data = borders, aes(group = group)) +
  geom_point(aes(colour = rate)) +
  coord_map()
ggsave("world-4.png", width = 8, height = 6, dpi = 128)

# Works better if we tweak aesthetics
ggplot(bubble, aes(long, lat)) +
  geom_polygon(data = borders, aes(group = group), colour = "grey70",
```

http://www.flickr.com/photos/altemark/337248947/sizes/l/

Outline

1. Introduction to the R language and ecosystem
2. Visualisation
3. Transformation
4. Modelling
5. Case study


Visualise
Transform
Model

```
install.packages(c("ggplot2", "plyr", "stringr"))
```