Teaching data analysis

Hadley Wickham

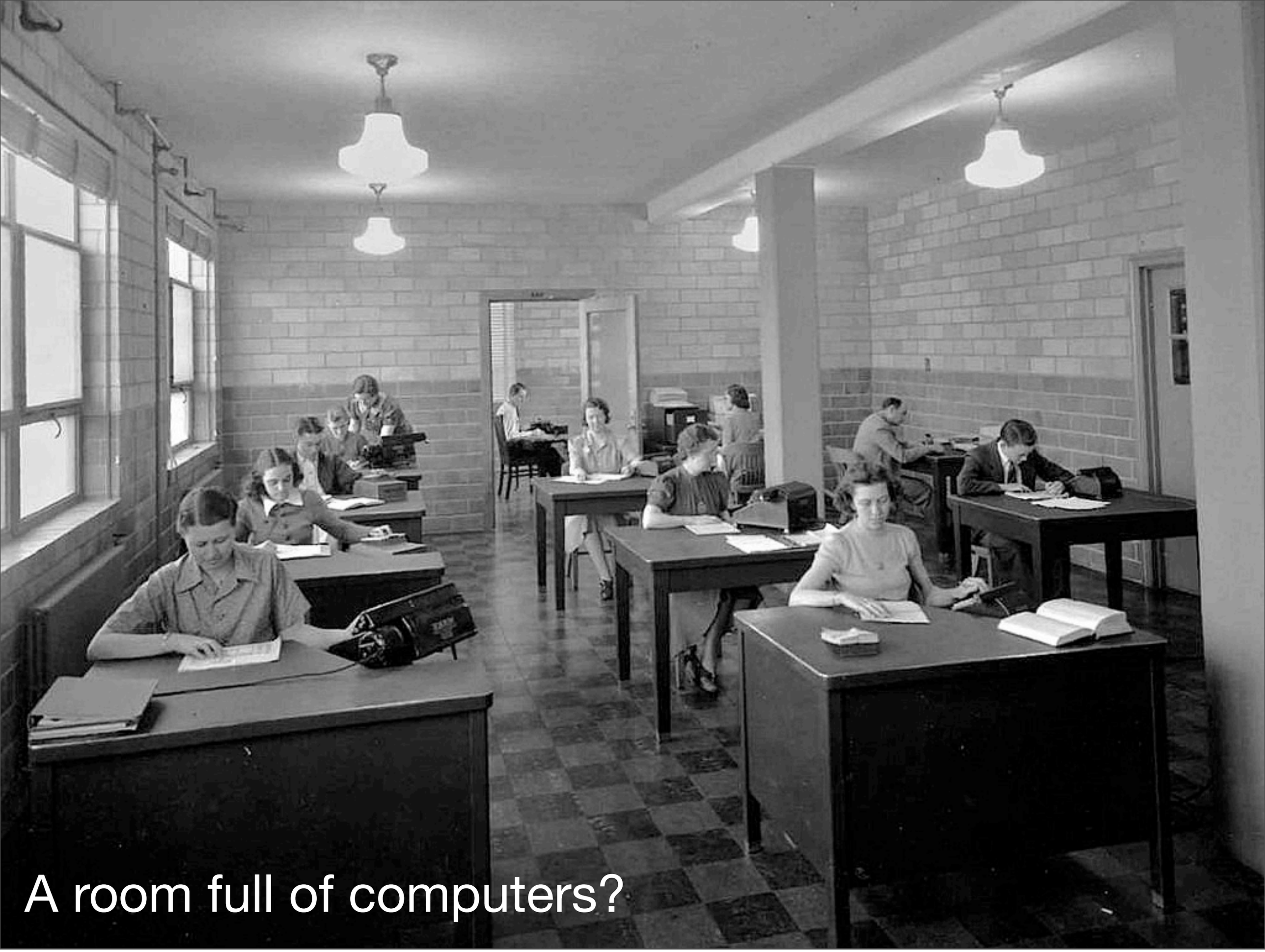
Assistant Professor / Dobelman Family Junior Chair

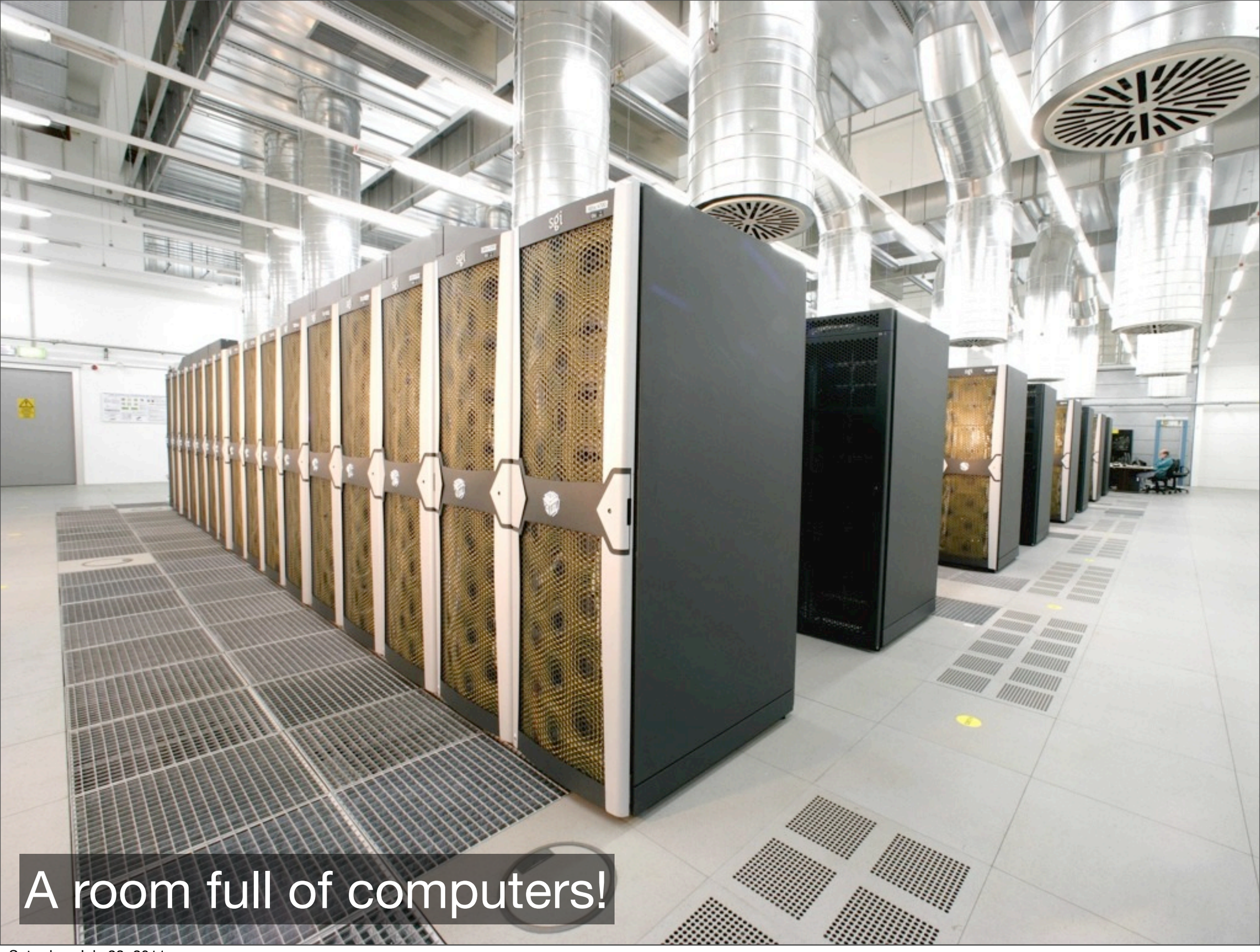
Department of Statistics / Rice University

July 2011

1. Why is programming an important part of data analysis?
2. How can we make programming engaging and accessible?
3. What types of practice and feedback help students learn?

Why programming?


A room full of computers?

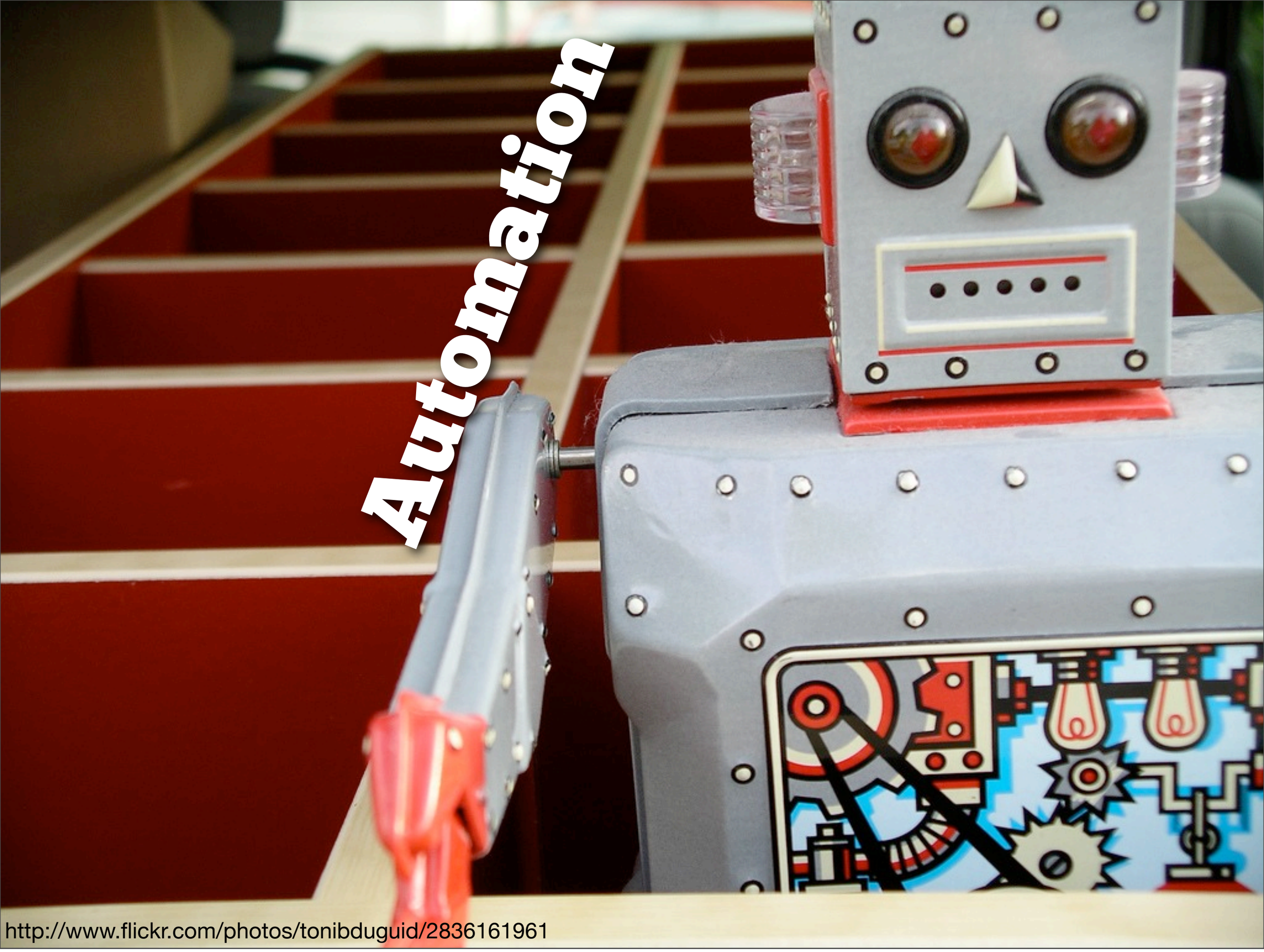

A room full of computers!

3 reasons

Reproducibility

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

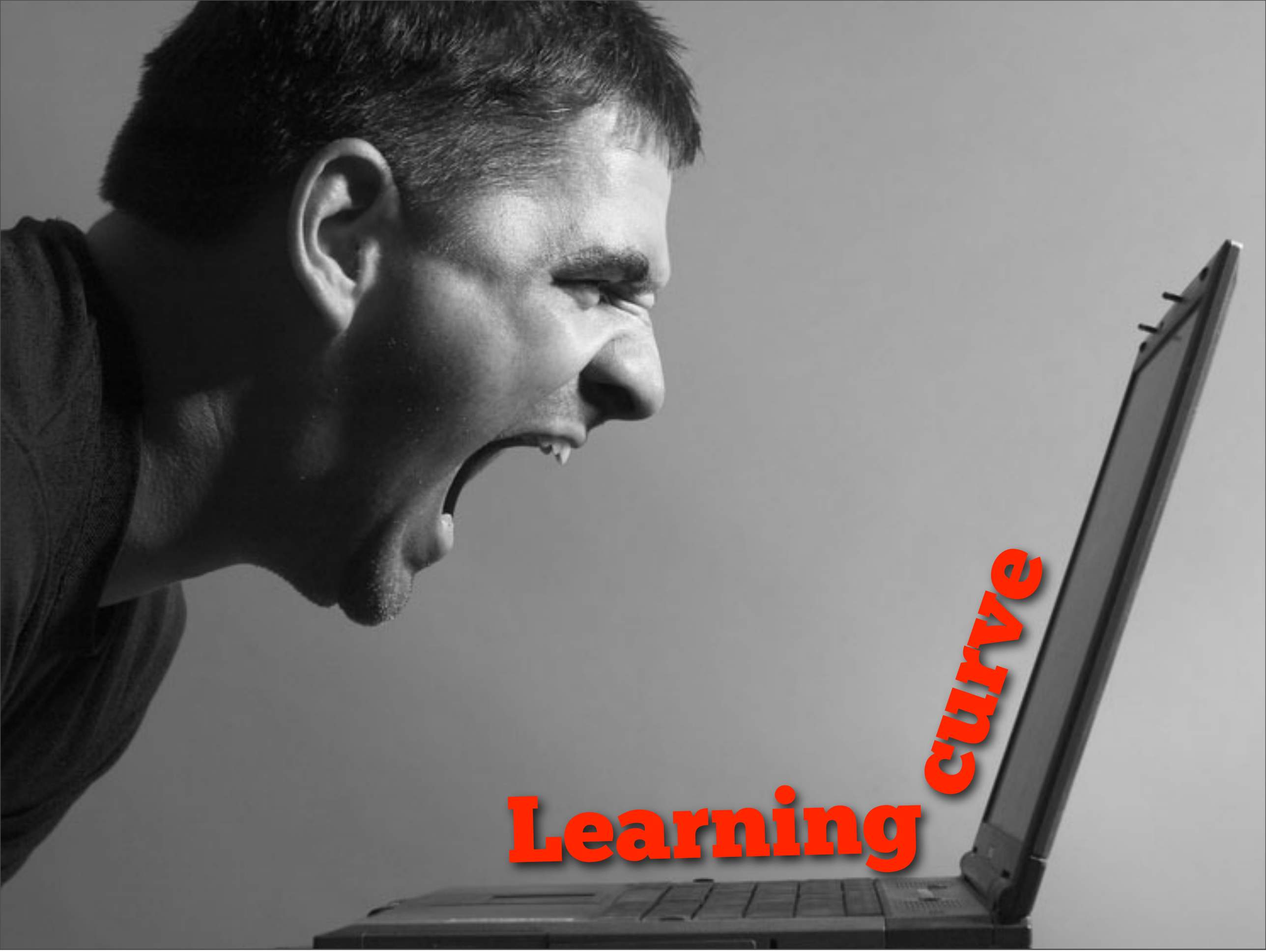
Automation



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

Communication

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


Learning curve

Teaching techniques

1. Start with visualisation

A grammar of graphics means:

scatterplot
histogram

x

aesthetics
facetting

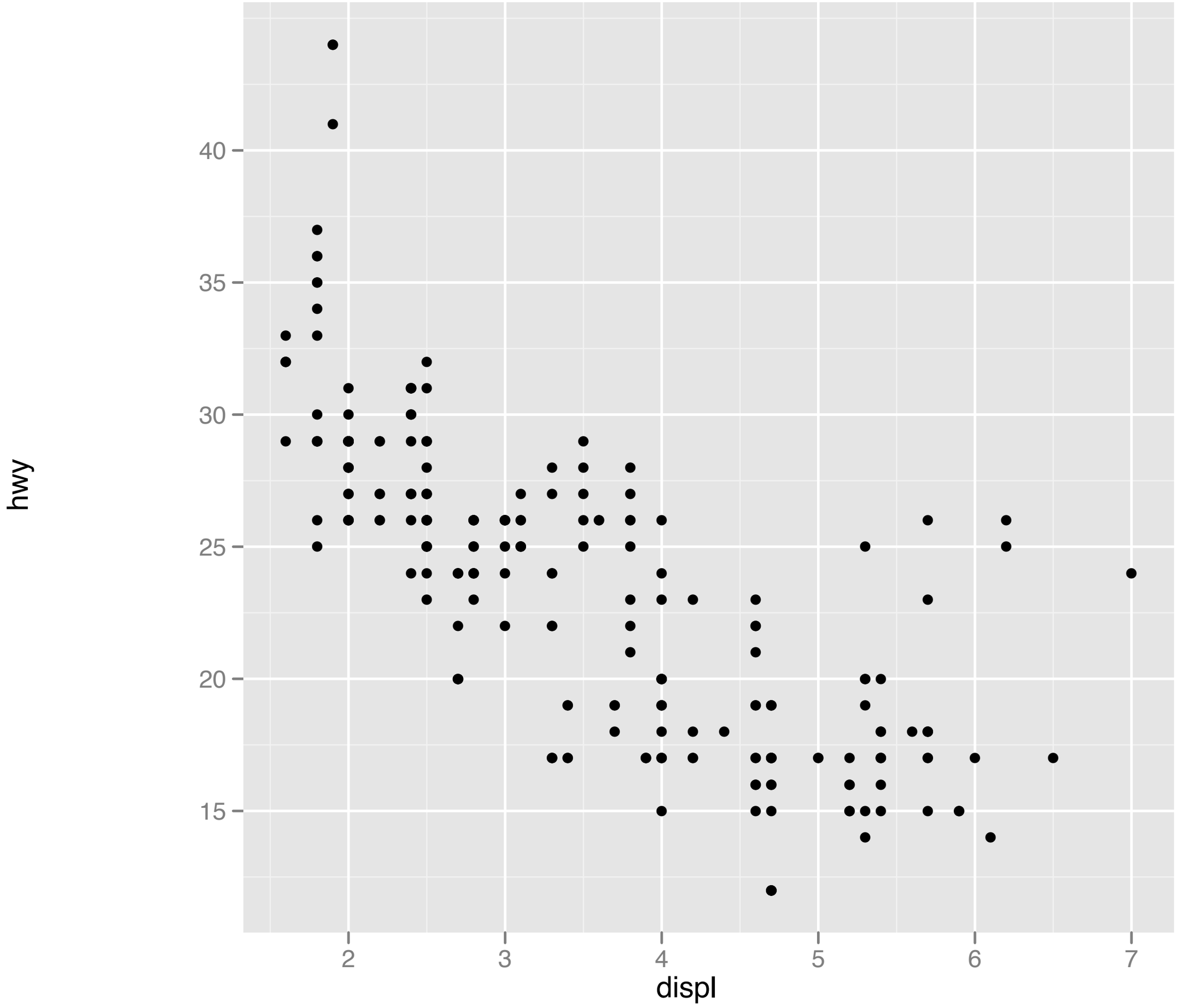
=

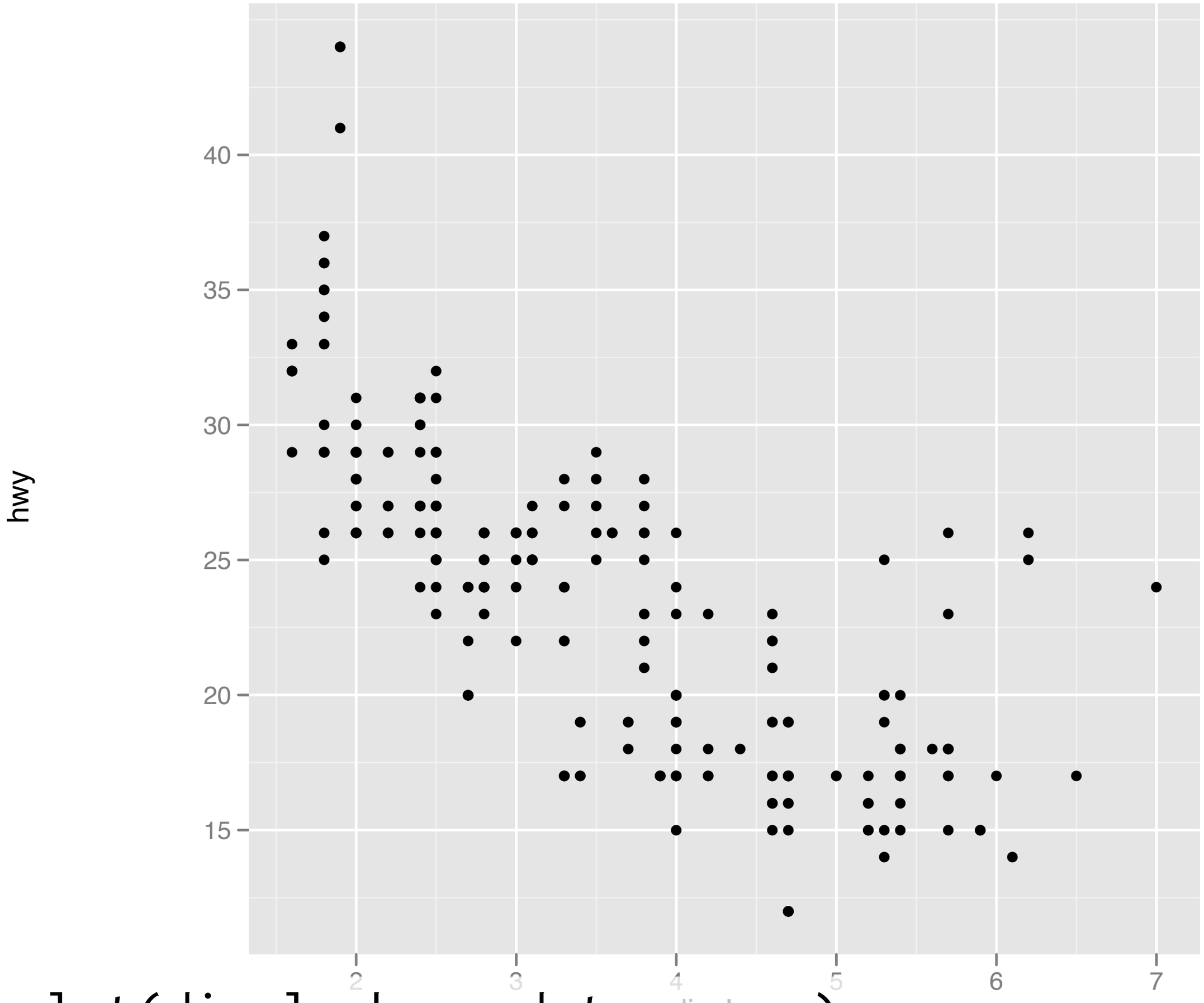
1000's of
possible
graphics

Can later teach other geometric elements:

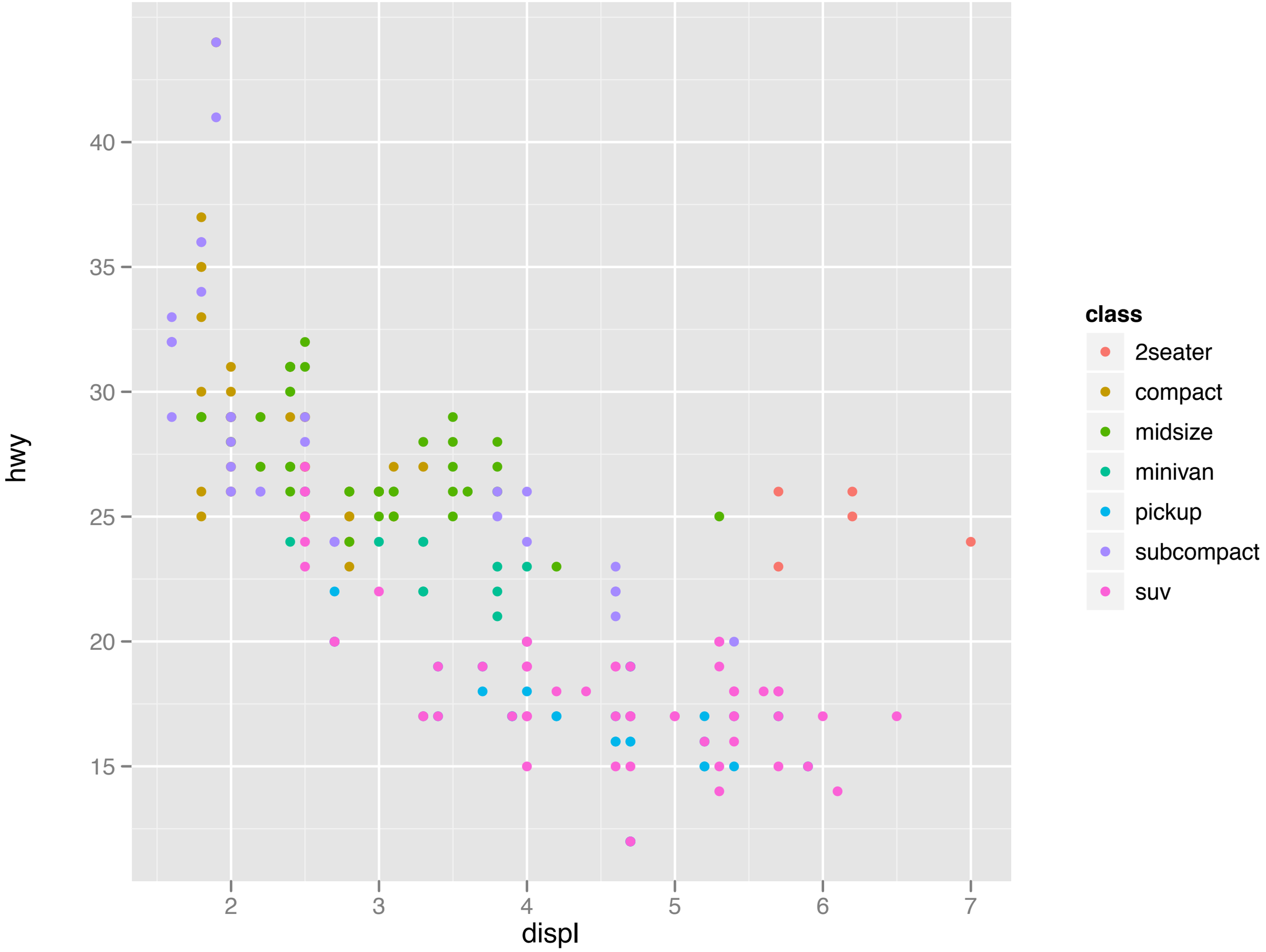
line
path
polygon

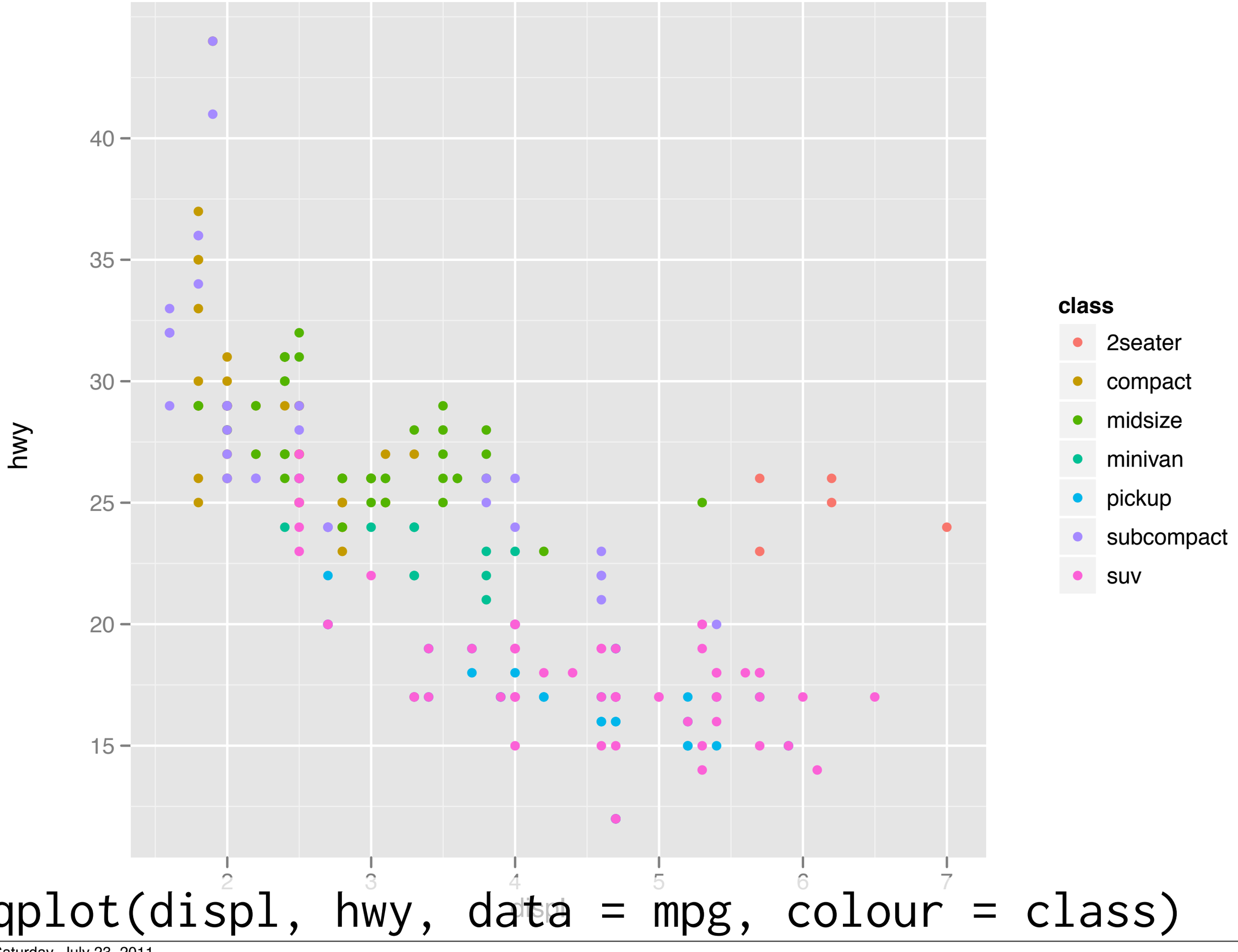
boxplot
smoother
hexagon binning


hwy
40
35
30
25
20
15
2
3
4
5
6
7
displ

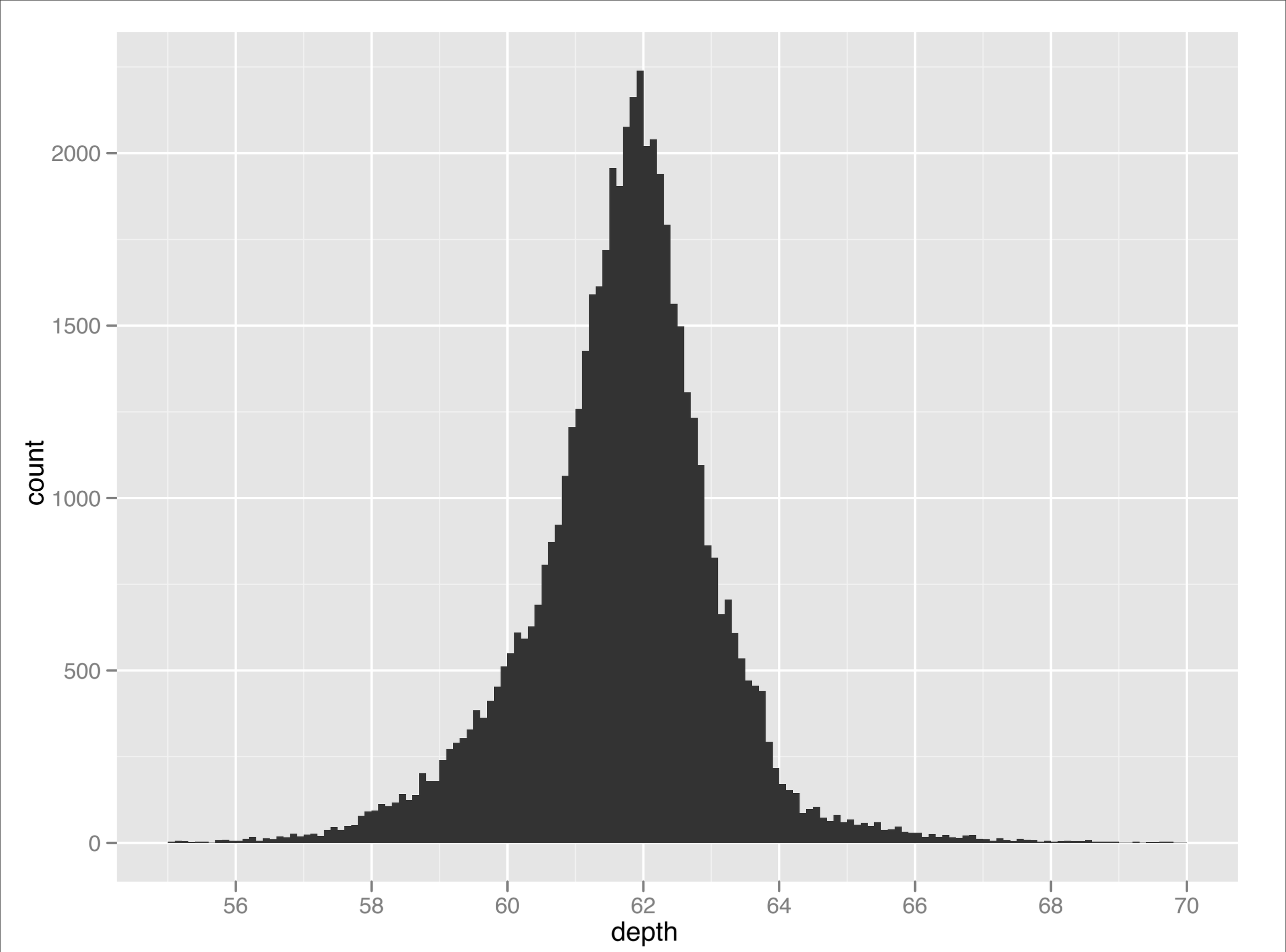


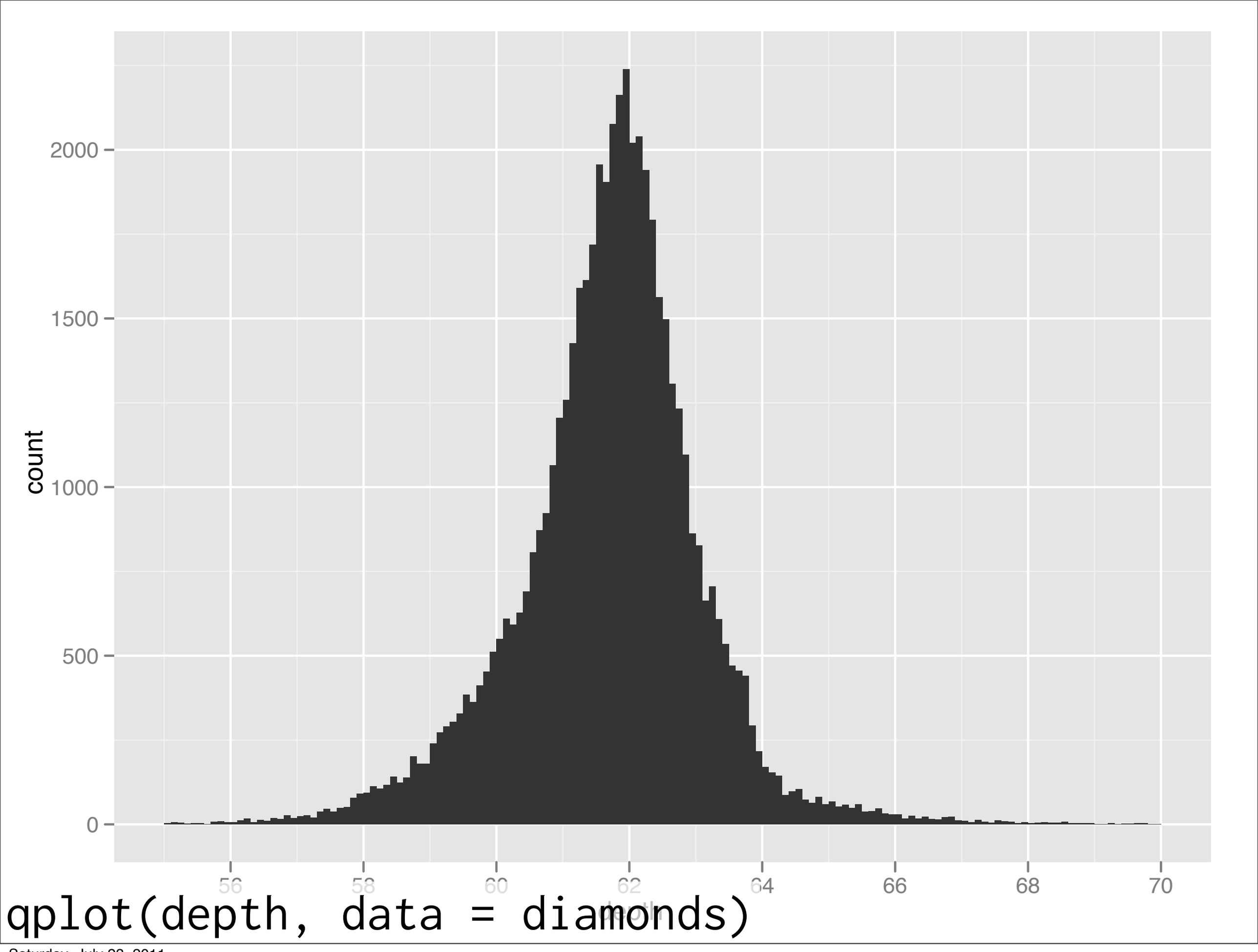

```
qplot(displ, hwy, data = mpg)
```


hwy
40
35
30
25
20
15
2
3
4
5
6
7
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv

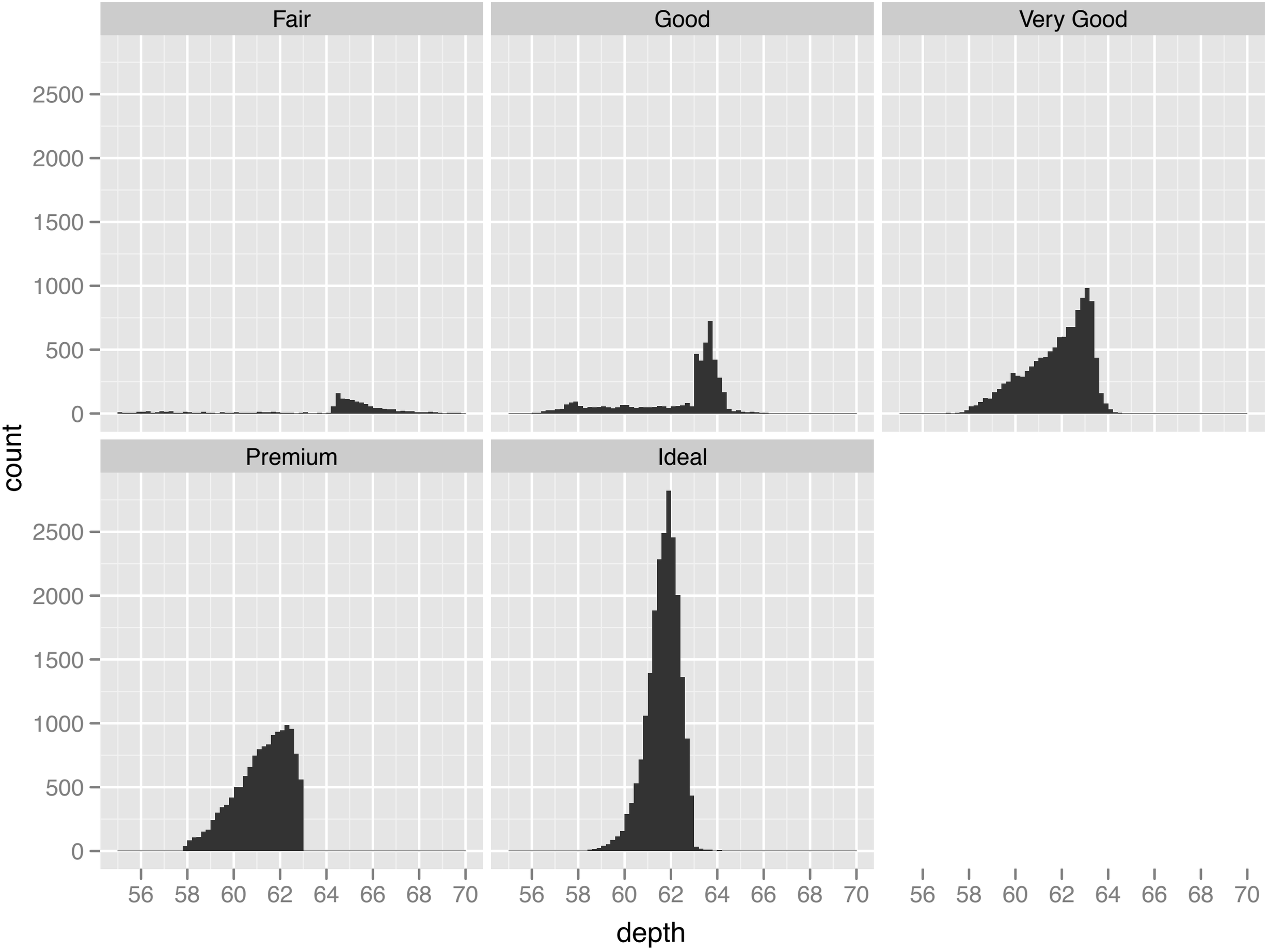


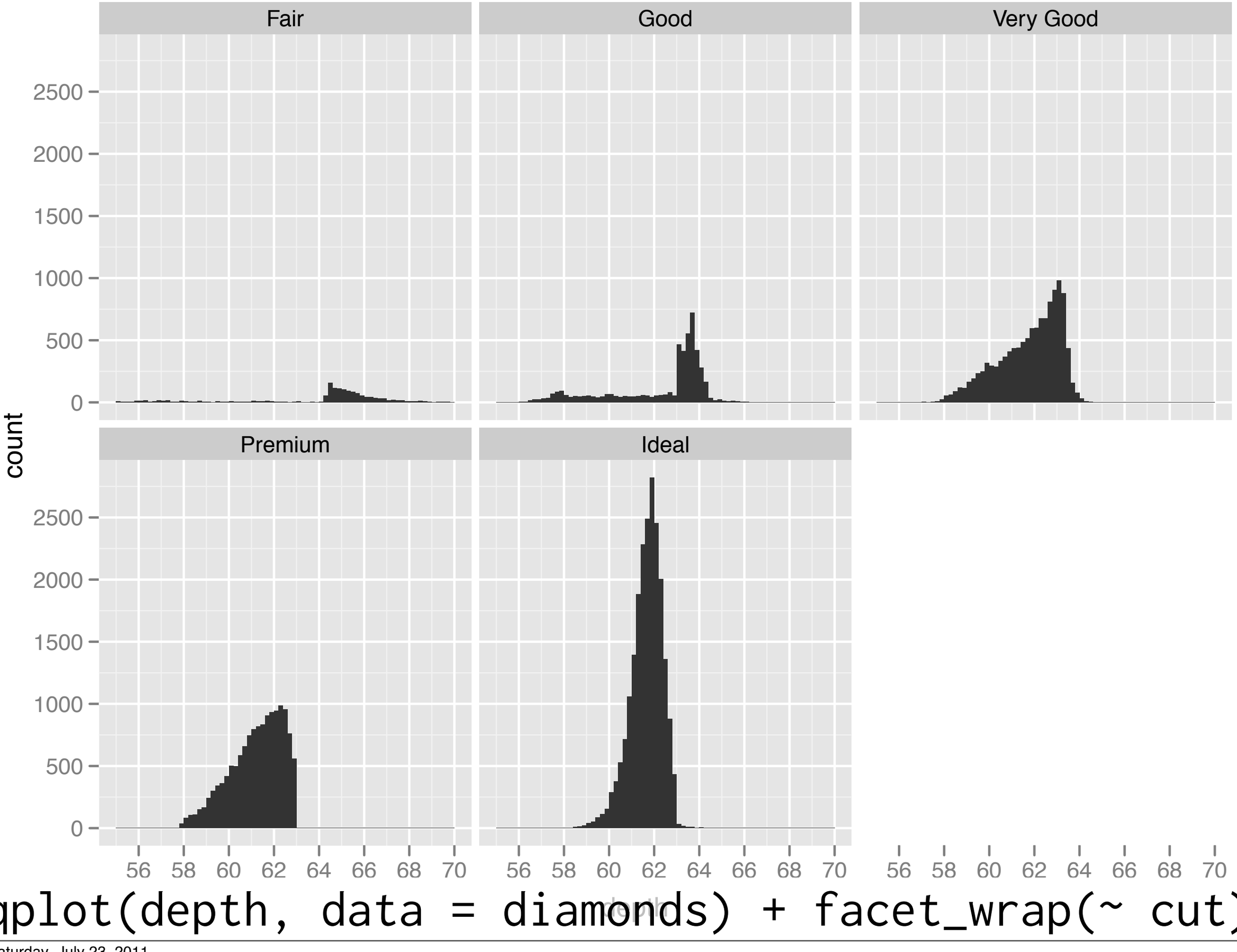

```
qplot(displ, hwy, data = mpg, colour = class)
```


count
2000
1500
1000
500
0
56
58
60
62
64
66
68
70
depth




```
qplot(depth, data = diamonds)
```


Fair
Good
Very Good
Premium
Ideal
count
depth
0
500
1000
1500
2000
2500
56
58
60
62
64
66
68
70




```
qplot(depth, data = diamonds) + facet_wrap(~ cut)
```

2. Motivate everything with a real problem

Baseball

NBA play-by-play

Diamond prices

Baby names

Airline delays

Card counting

Fuel economy

Movie rankings

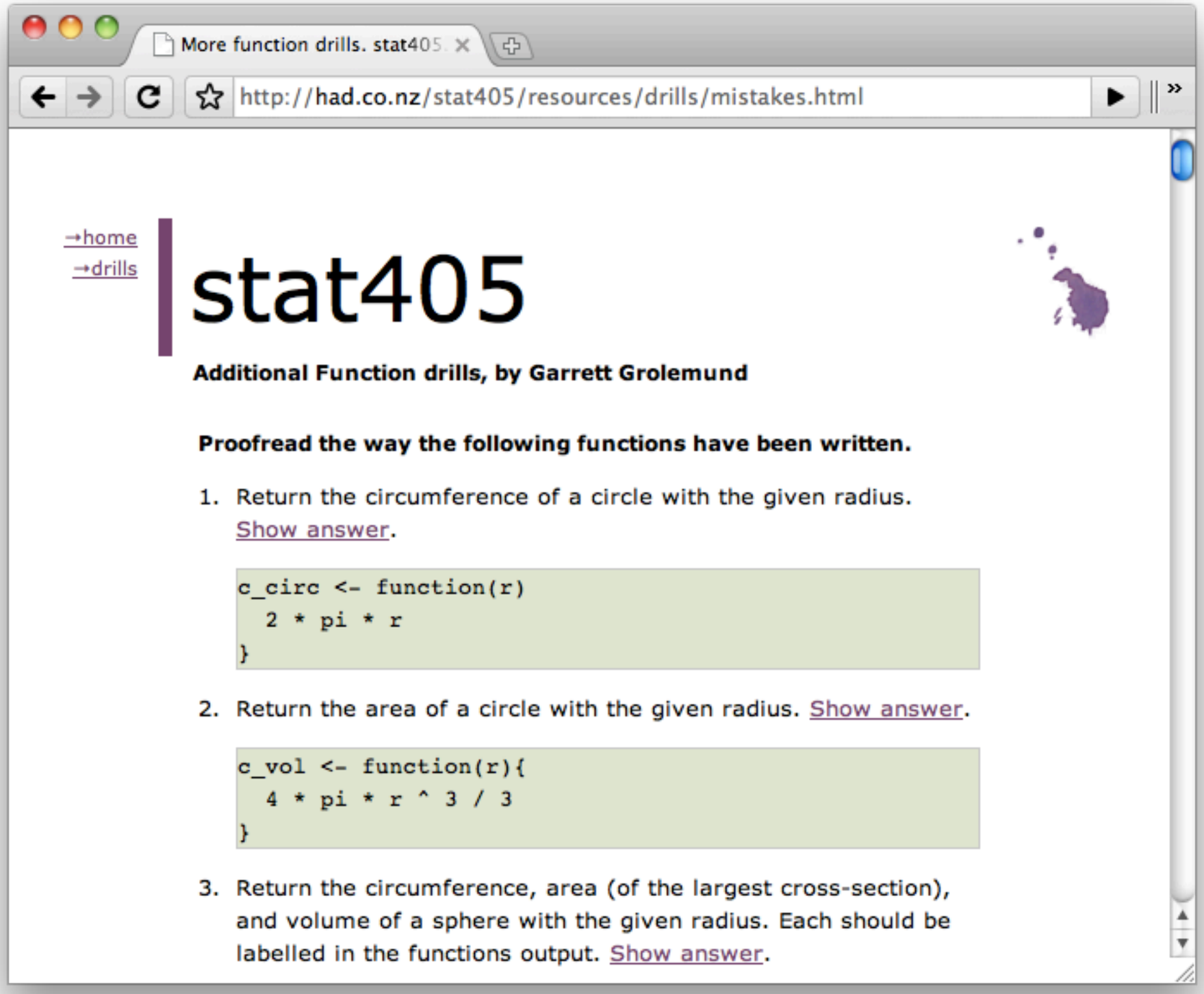
http://github.com/hadley

The problem motivates the tools.

Practice & Feedback

	Technique
Muscle memory	Drills
Dispositions	Open-ended data analyses

C. Wild and M. Pfannkuch. *Statistical thinking in empirical enquiry*. International Statistical Review, 67(3): 223–248, 1999.


More function drills. stat405
http://had.co.nz/stat405/resources/drills/mistakes.html
→home
→drills
stat405
Additional Function drills, by Garrett Grolemund
Proofread the way the following functions have been written.
1. Return the circumference of a circle with the given radius. Show answer.
c_circ <- function(r)
2 * pi * r
}
2. Return the area of a circle with the given radius. Show answer.
c_vol <- function(r){
4 * pi * r ^ 3 / 3
}
3. Return the circumference, area (of the largest cross-section), and volume of a sphere with the given radius. Each should be labelled in the functions output. Show answer.

More function drills. stat405

http://had.co.nz/stat405/resources/drills/mistakes.html

→home
→drills

stat405

Additional Function drills, by Garrett Grolemund

Proofread the way the following functions have been written.

1. Return the circumference of a circle with the given radius. Hide answer.

```
c_circ <- function(r)
  2 * pi * r
}
```

```
c_circ <- function(r){
  2 * pi * r
}
```

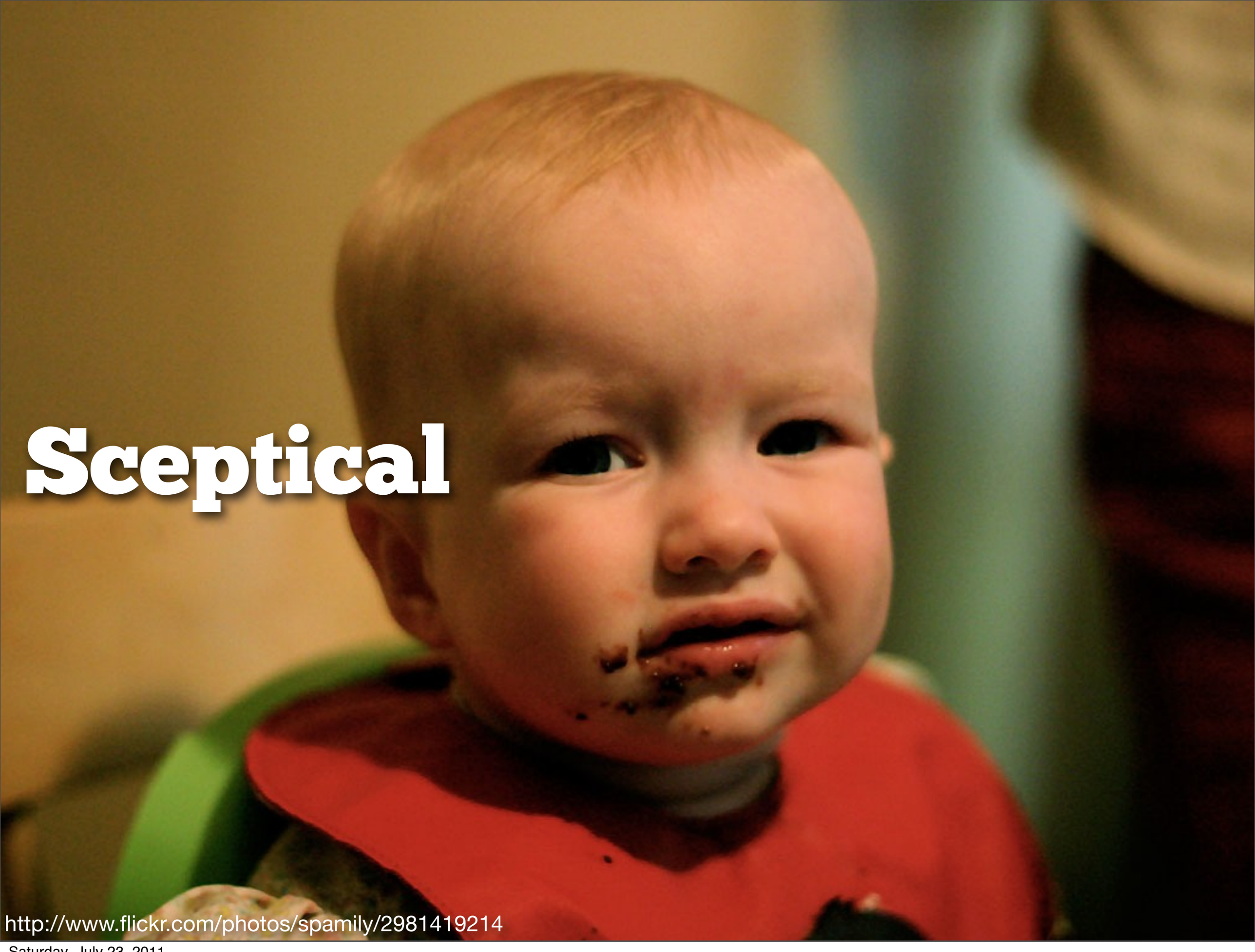
2. Return the area of a circle with the given radius. Show answer.

```
c_vol <- function(r){
  4 * pi * r ^ 3 / 3
}
```

Curious

http://www.flickr.com/photos/tommyhj/145269819

Sceptical



http://www.flickr.com/photos/spamily/2981419214

http://www.flickr.com/photos/kristinroach/3995676135

Conclusions

1. Statistics students need to learn how to program
2. Teach programming by starting with visualisation
3. Motivate every new technique with a real problem
4. Practice low-level skills with drills
5. Give feedback on dispositions