

Practical tools for exploring data and models

Hadley Alexander Wickham

“The process of data analysis is one of parallel evolution. Interrelated aspects of the analysis evolve together, each affecting the others.”

– Paul Velleman, 1997

“Interrelated aspects of the analysis
evolve together”

Questions

Form reshape

Views ggplot2

Models classifly, clusterfly, meifly

A grammar of graphics: past, *present*, and future

Past





“If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum.”

Euclid, ~300 BC

$$m(\sum x) = \sum(mx)$$

The grammar of graphics

- An abstraction which makes thinking, reasoning and communicating graphics easier
- Developed by Leland Wilkinson, particularly in “The Grammar of Graphics” 1999/2005

Present

ggplot2

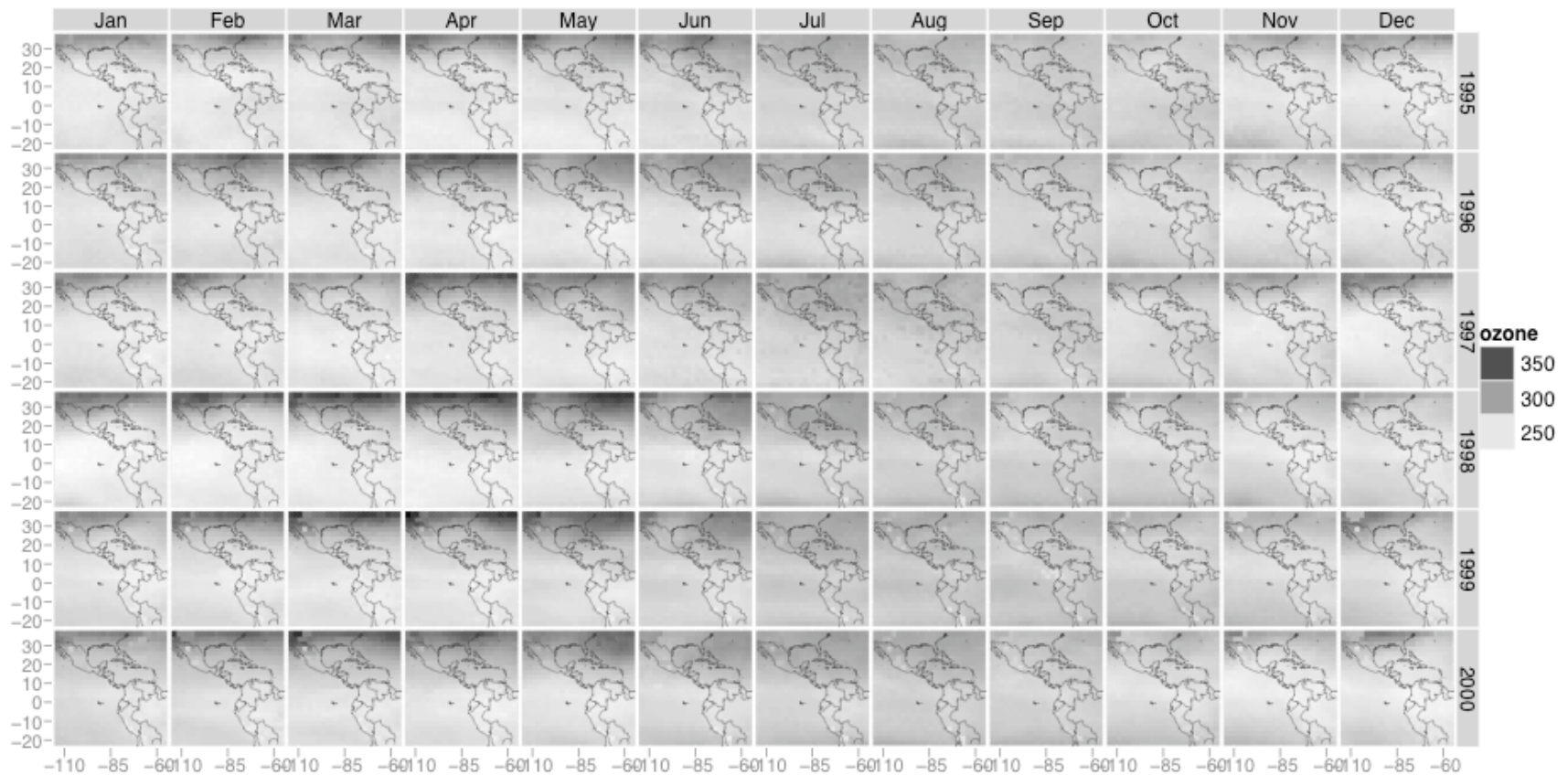
- High-level package for creating statistical graphics.
A rich set of components + user friendly wrappers
- Inspired by “The Grammar of Graphics”
Leland Wilkinson 1999
- John Chambers award in 2006
- Philosophy of ggplot
- Examples from a recent paper
- New methods facilitated by ggplot

Philosophy

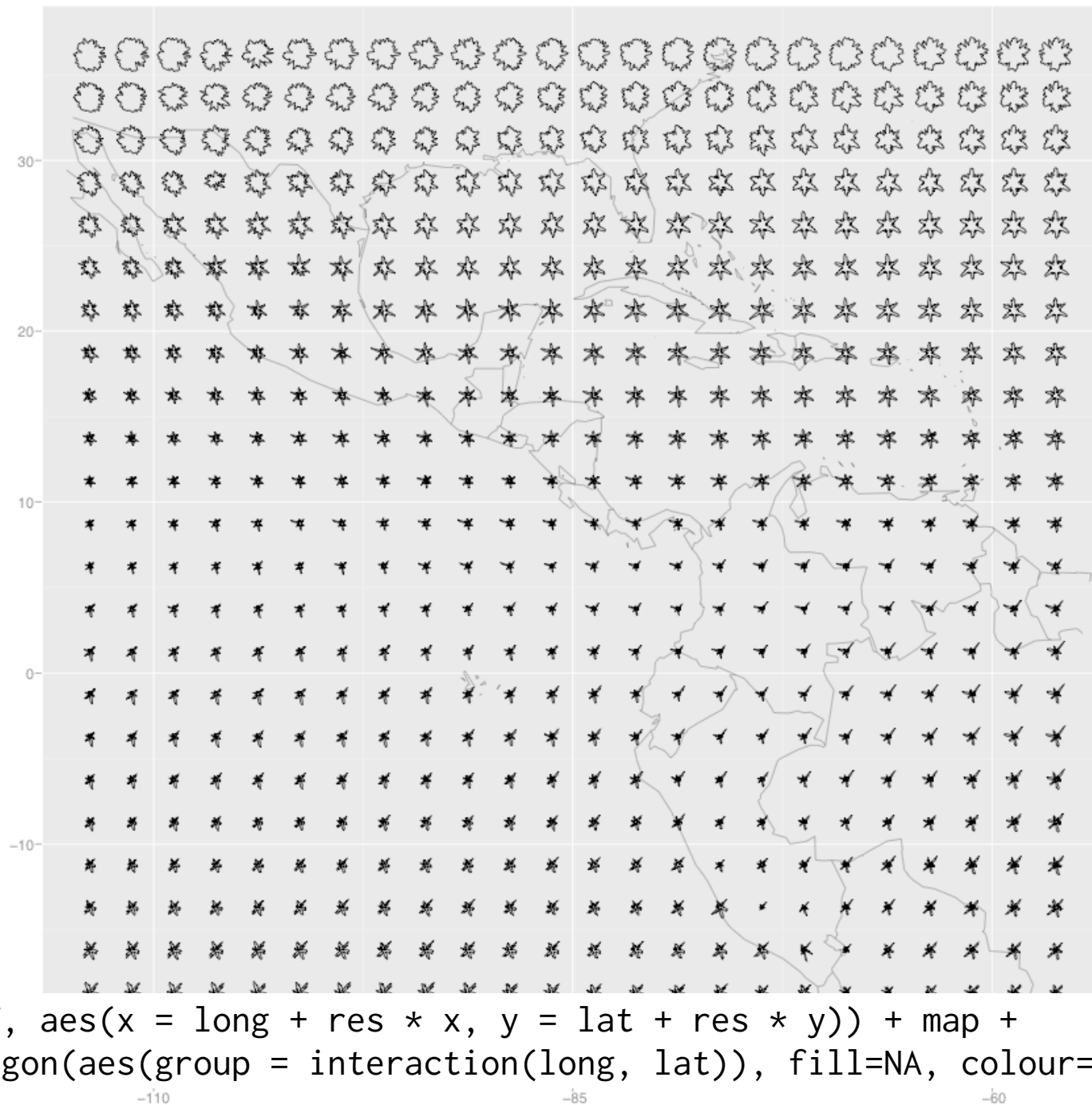
- Make graphics easier
- Use the grammar to facilitate research into new types of display
- Continuum of expertise:
 - start simple by using the results of the theory
 - grow in power by understanding the theory
 - begin to contribute new components
- Orthogonal components and minimal special cases should make learning easy(er?)

Examples

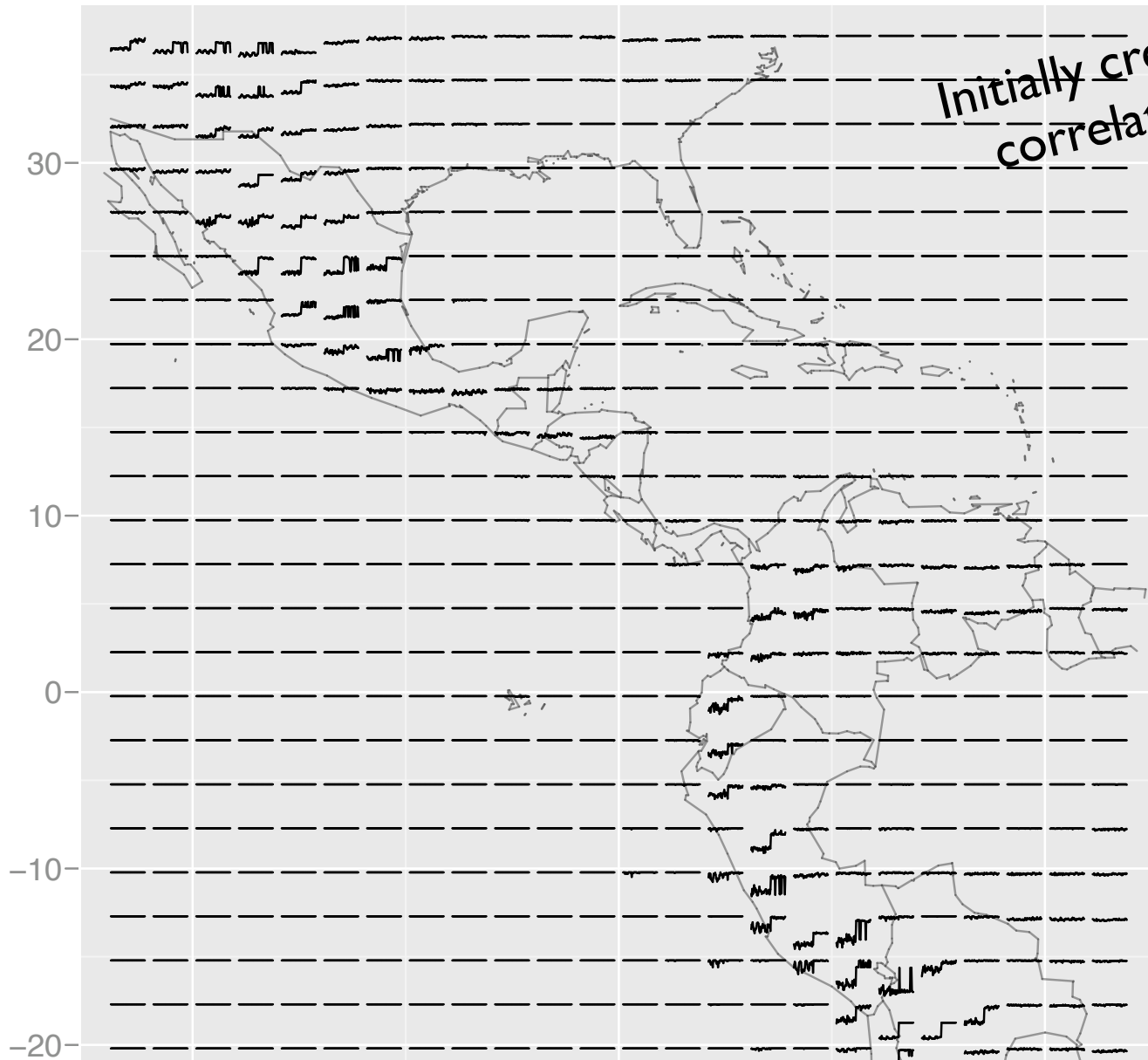
- J. Hobbs, H. Wickham, H. Hofmann, and D. Cook.
Glaciers melt as mountains warm: A graphical case study. Computational Statistics. Special issue for ASA Statistical Computing and Graphics Data Expo 2006.
- Exploratory graphics created with GGobi, Mondrian, Manet, Gauguin and R, but needed consistent high-quality graphics that work in black and white for publication
- So... used ggplot to recreate the graphics



```
qplot(long, lat, data = expo, geom="tile", fill = ozone,
  facets = year ~ month) +
scale_fill_gradient(low="white", high="black") + map
```



```
ggplot(df, aes(x = long + res * x, y = lat + res * y)) + map +  
geom_polygon(aes(group = interaction(long, lat)), fill=NA, colour="black")
```

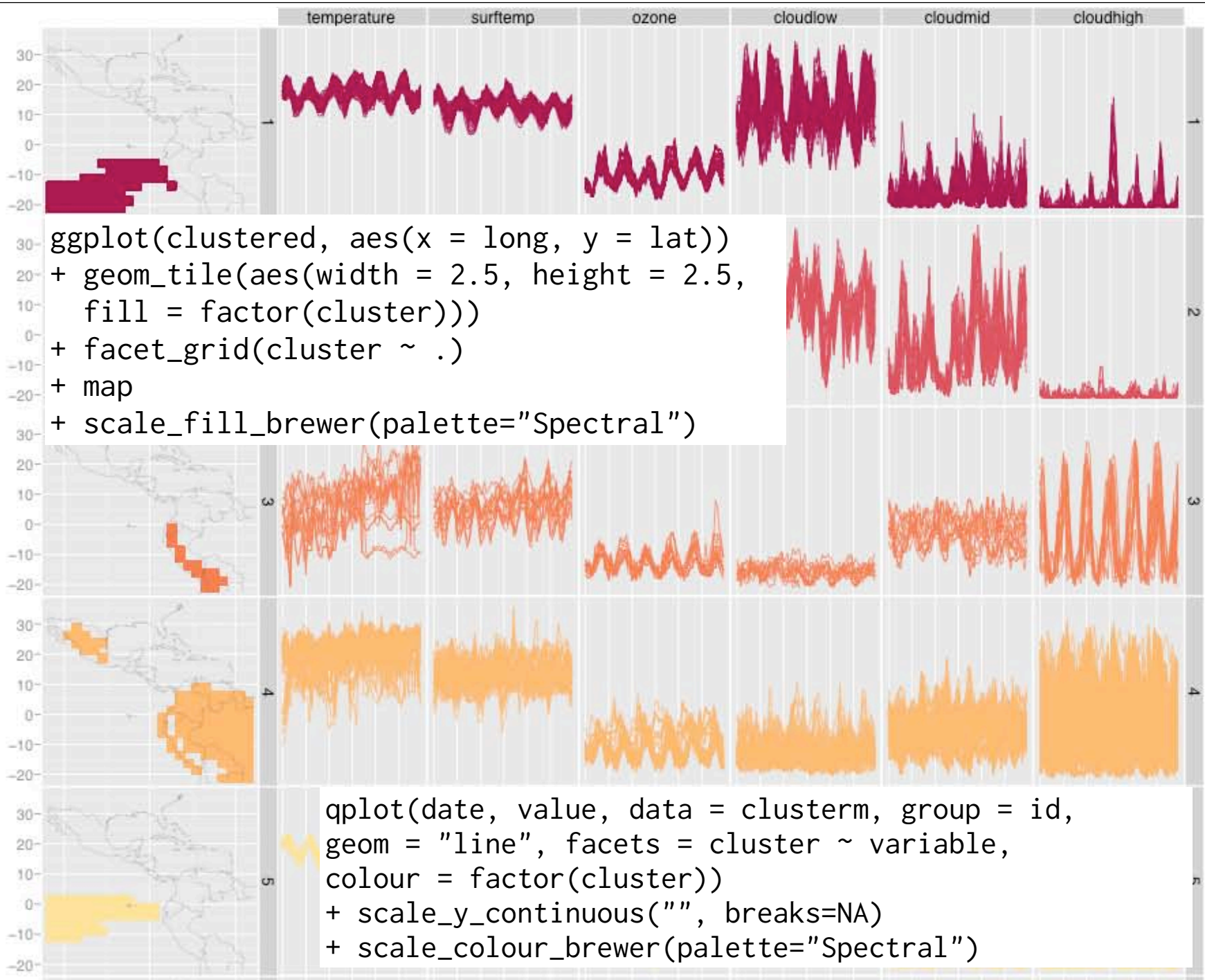


Initially created with
correlation tour

```
ggplot(rexpo, aes(x = long + res * rtime, y = lat + res * rpressure))  
+ map + geom_line(aes(group = id))
```

```
library(maps)
outlines <- as.data.frame(map("world",xlim=-c(113.8, 56.2),ylim=c(-21.2, 36.2)))

map <- c(
  geom_path(aes(x = x, y = y), data = outlines, colour = alpha("grey20", 0.2)),
  scale_x_continuous("", limits = c(-113.8, -56.2), breaks = c(-110, -85, -60)),
  scale_y_continuous("", limits = c(-21.2, 36.2))
)
```



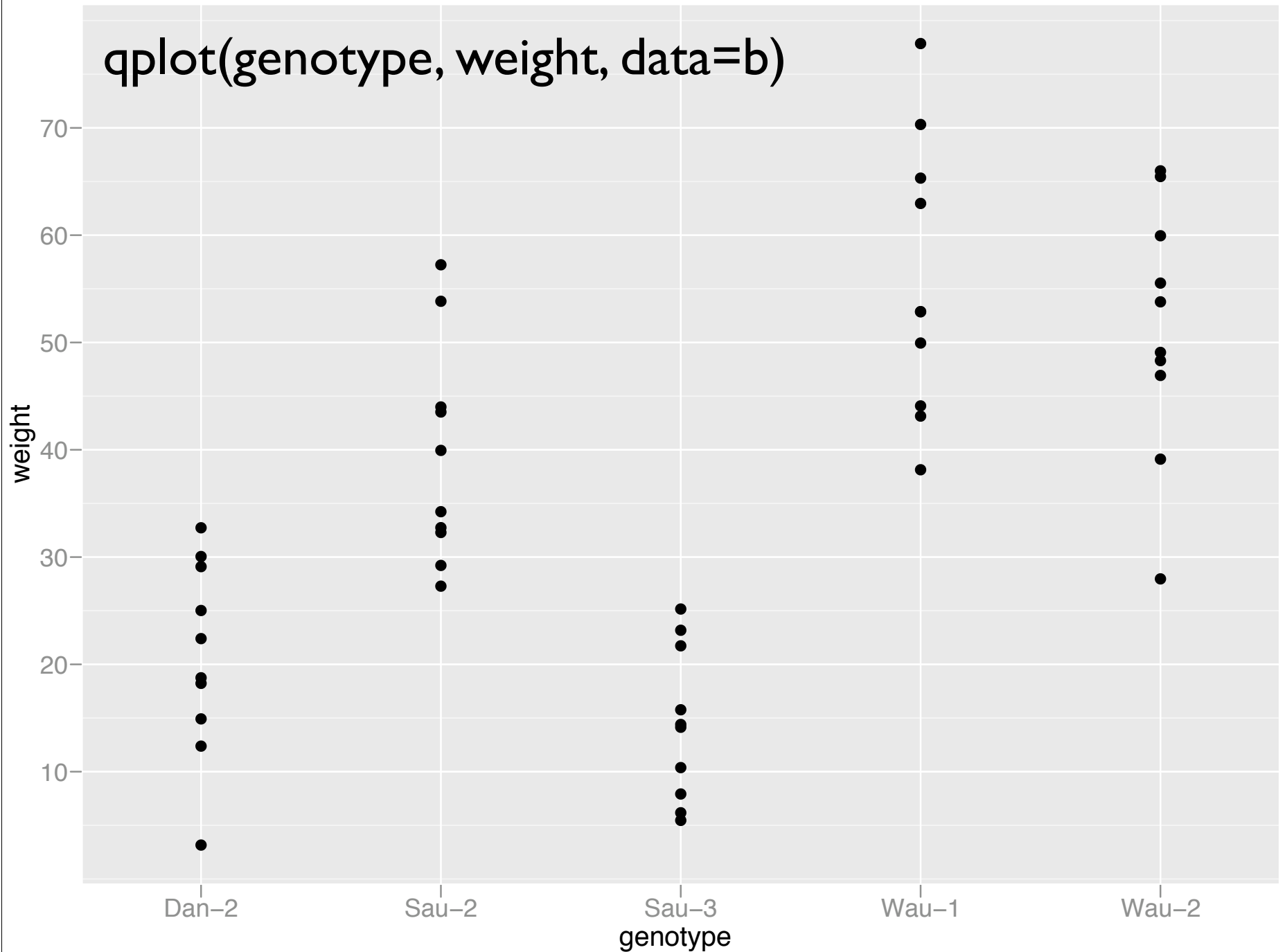
New methods

- Supplemental statistical summaries
- Iterating between graphics and models
- Inspired by ideas of Tukey (and others)
- Exploratory graphics, not as pretty

Intro to data

- Response of trees to gypsy moth attack
- 5 genotypes of tree: Dan-2, Sau-2, Sau-3, Wau-1, Wau-2
- 2 treatments: NGM / GM
- 2 nutrient levels: low / high
- 5 reps
- Measured: weight, N, tannin, salicylates

qplot(genotype, weight, data=b)



```
qplot(genotype, weight, data=b,  
colour=nutr)
```

weight

nutr

Low

High

Dan-2

Sau-2

Sau-3

Wau-1

Wau-2

genotype

70

60

50

40

30

20

10

qplot(genotype, weight, data=b,
colour=nutr)

```
qplot(reorder(genotype, weight), weight,  
data=b, colour=nutr)
```

weight

nutr

Low

High

Sau-3

Dan-2

Sau-2

Wau-2

Wau-1

genotype

70

60

50

40

30

20

10

Comparing means

- For inference, interested in comparing the means of the groups
- But this is hard to do visually as eyes naturally compare ranges
- What can we do?

Supplemental summaries

From Hmisc

- ```
smry <- stat_summary(
 fun="mean_cl_boot", conf.int=0.68,
 geom="crossbar", width=0.3
)
```
- Adds another layer with summary statistics:  
mean + bootstrap estimate of standard error
- Motivation: still exploratory, so minimise  
distributional assumptions, will model explicitly  
later

```
qplot(genotype, weight, data=b,
colour=nutr)
```

weight

**nutr**

Low

High

Sau-3

Dan-2

Sau-2

Wau-2

Wau-1

genotype

70

60

50

40

30

20

10



qplot(genotype, weight, data=b,  
colour=nutr) + smry

weight

**nutr**  
Low  
High

Sau-3

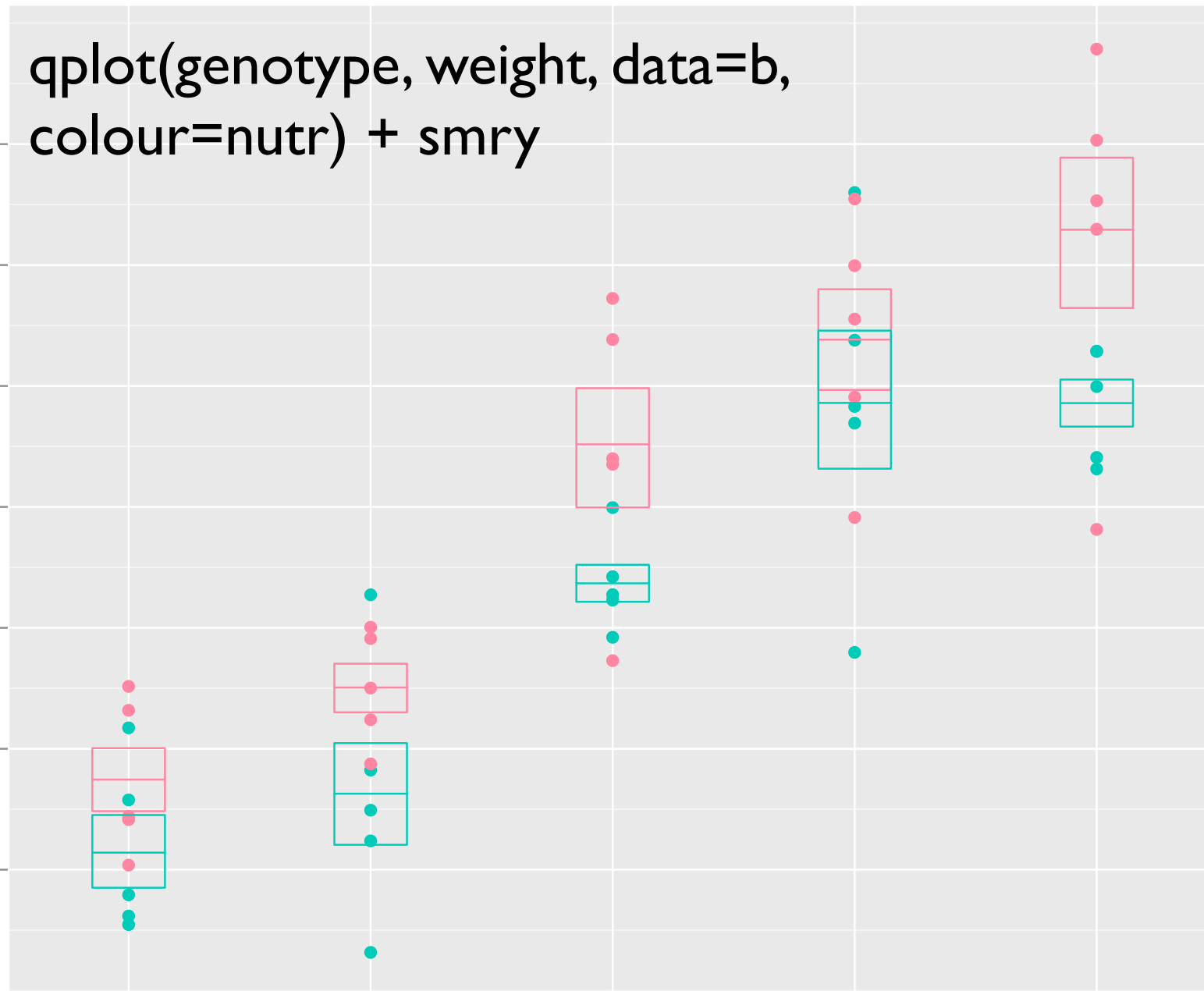
Dan-2

Sau-2

Wau-2

Wau-1

genotype



# Iterating graphics and modelling

- Clearly strong genotype effect. Is there a nutr effect? Is there a nutr-genotype interaction?
- Hard to see from this plot - what if we remove the genotype main effect? What if we remove the nutr main effect?
- How does this compare an ANOVA?

qplot(genotype, weight, data=b,  
colour=nutr) + smry

weight

**nutr**  
Low  
High

Sau-3

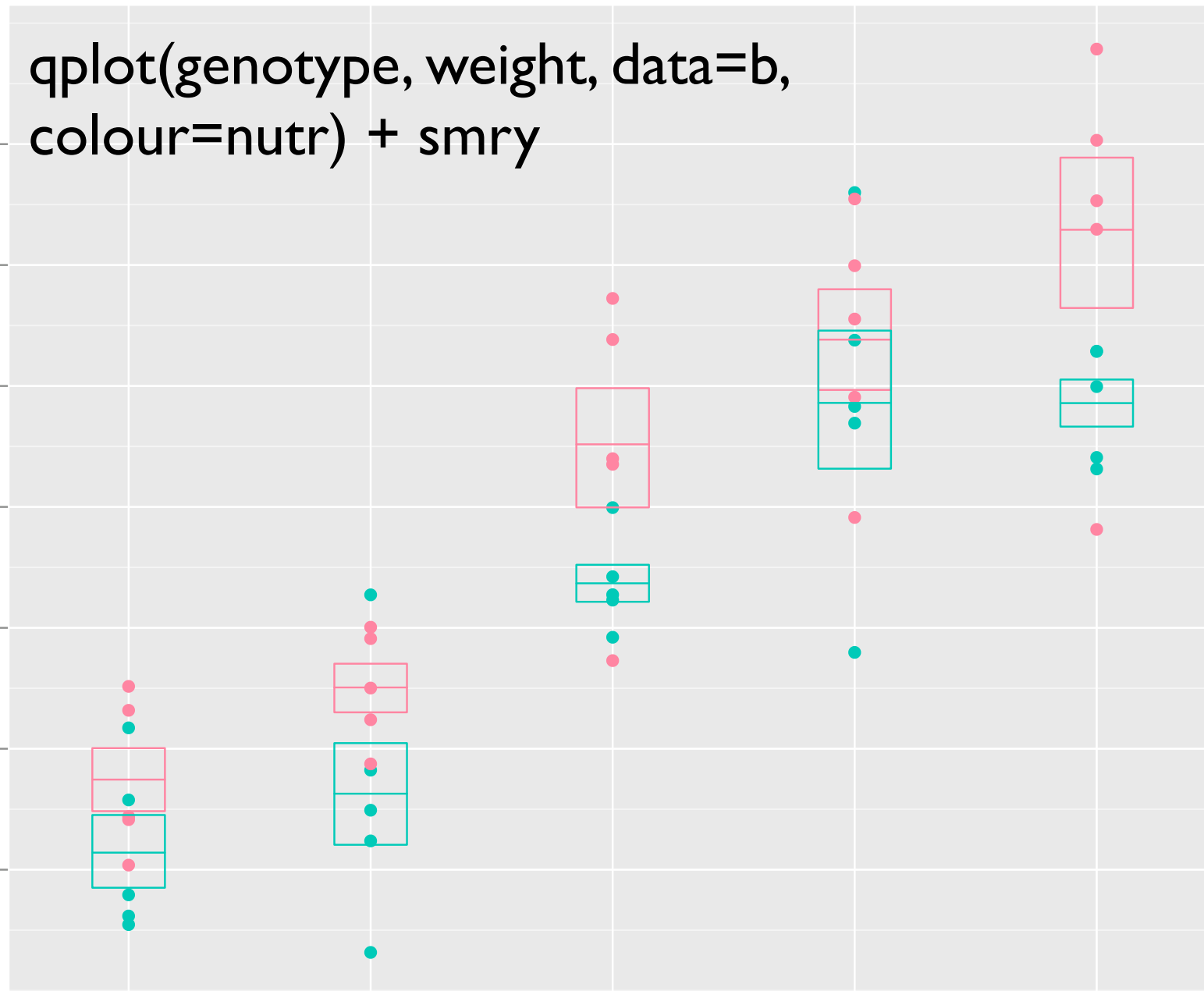
Dan-2

Sau-2

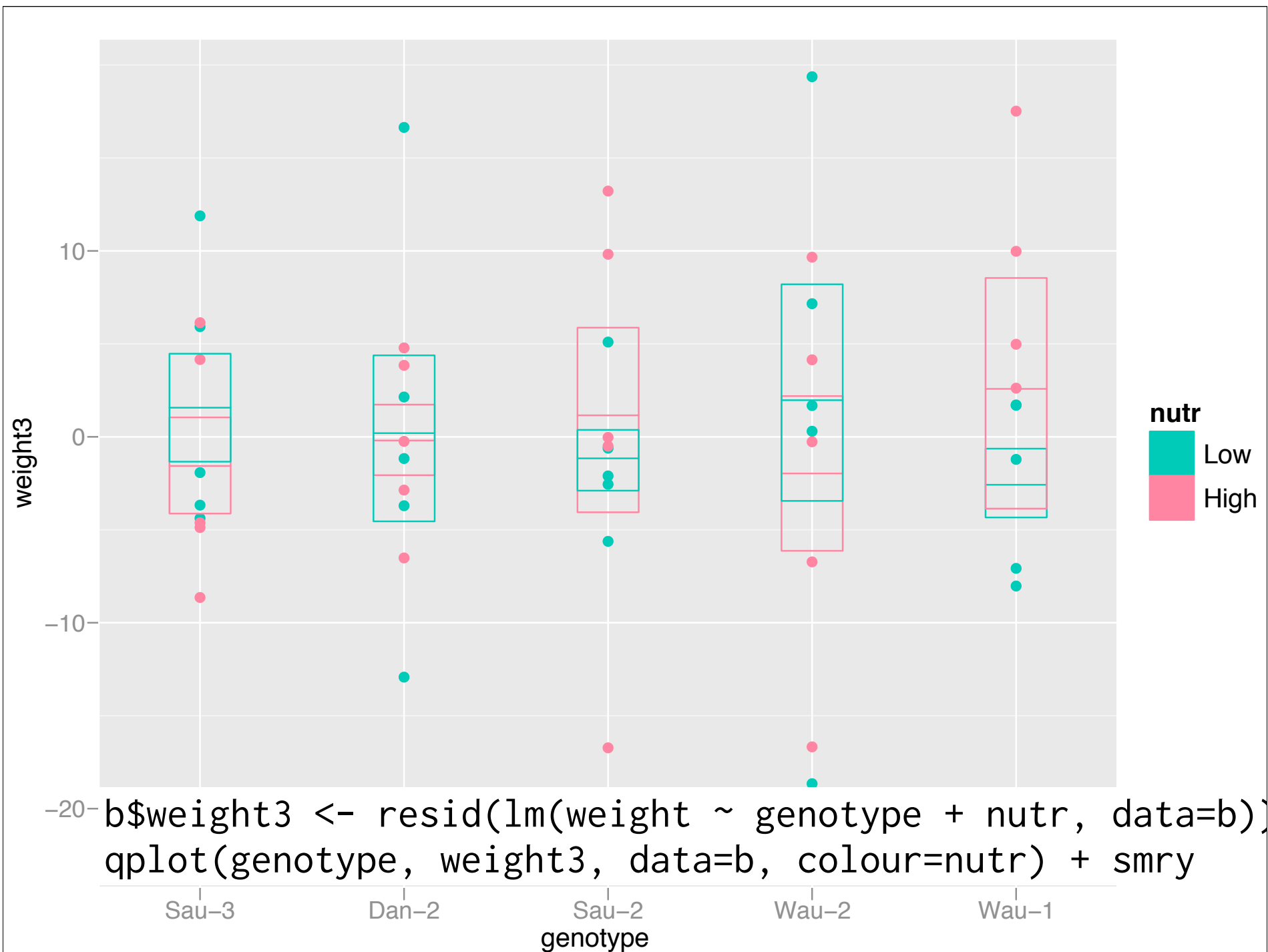
Wau-2

Wau-1

genotype







|               | Df | Sum Sq | Mean Sq | F value | Pr(>F)  |     |
|---------------|----|--------|---------|---------|---------|-----|
| genotype      | 4  | 13331  | 3333    | 36.22   | 8.4e-13 | *** |
| nutr          | 1  | 1053   | 1053    | 11.44   | 0.0016  | **  |
| genotype:nutr | 4  | 144    | 36      | 0.39    | 0.8141  |     |
| Residuals     | 40 | 3681   | 92      |         |         |     |

```
anova(lm(weight ~ genotype * nutr, data=b))
```

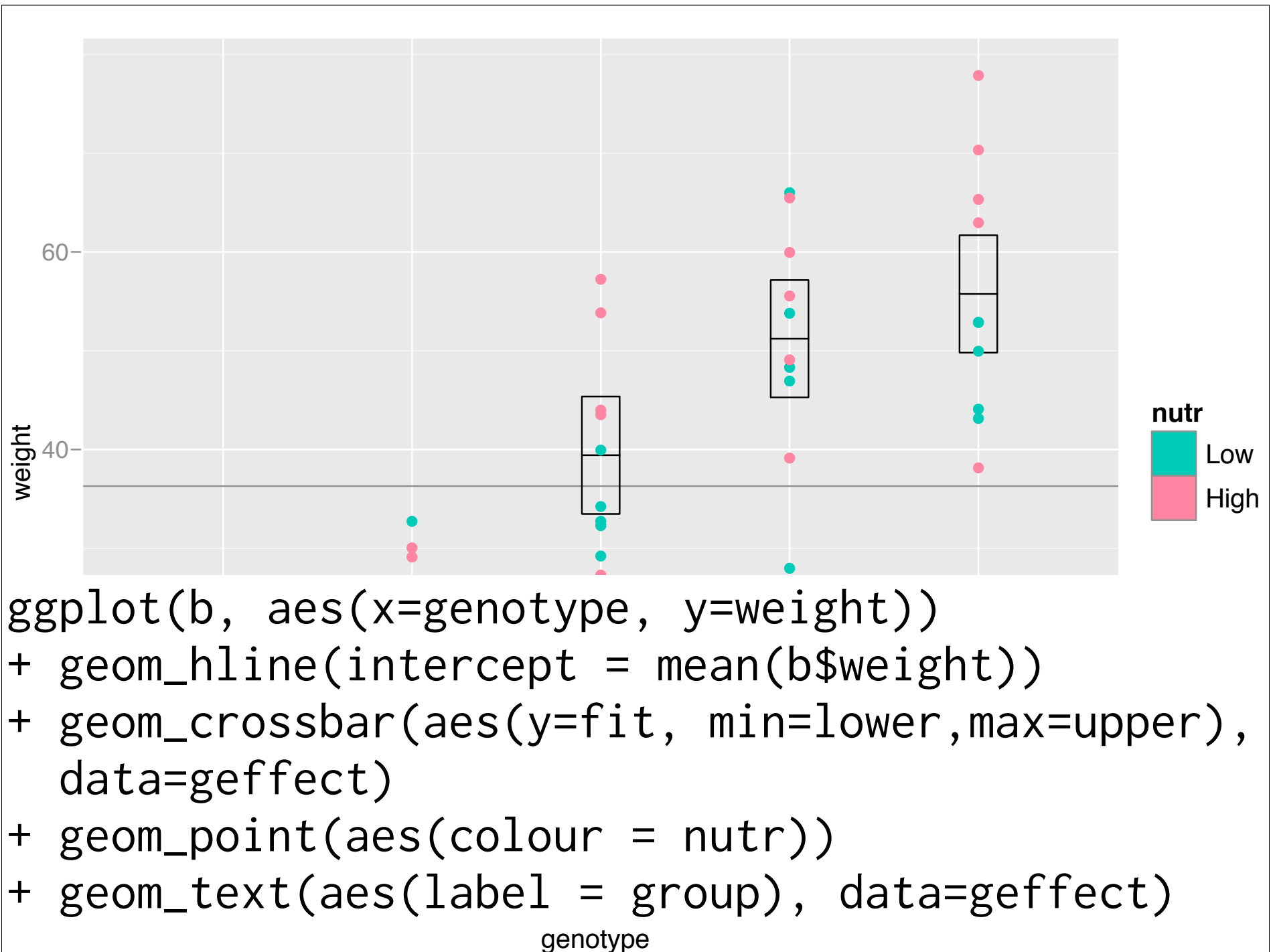
# Graphics → Model

- In the previous example, we used graphics to iteratively build up a model - a la stepwise regression!
- But: here interested in gestalt, not accurate prediction, and must remember that this is just one possible model
- What about model → graphics?

# Model → Graphics

- If we model first, we need graphical tools to summarise model results, e.g. post-hoc comparison of levels
- We can do better than SAS! But it's hard work: effects, multComp and multCompView
- Rich research area





# Summary

- Need to move beyond canned statistical graphics to experimenting with new graphical methods
- Strong links between graphics and models, how can we best use them?
- Static graphics often aren't enough

# Questions?