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

Hadley Wickham
Iowa State University
http://had.co.nz/

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

"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

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

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
- One of the cornerstones of my research (I’ll talk about the others later)

Present

ggplot

- 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

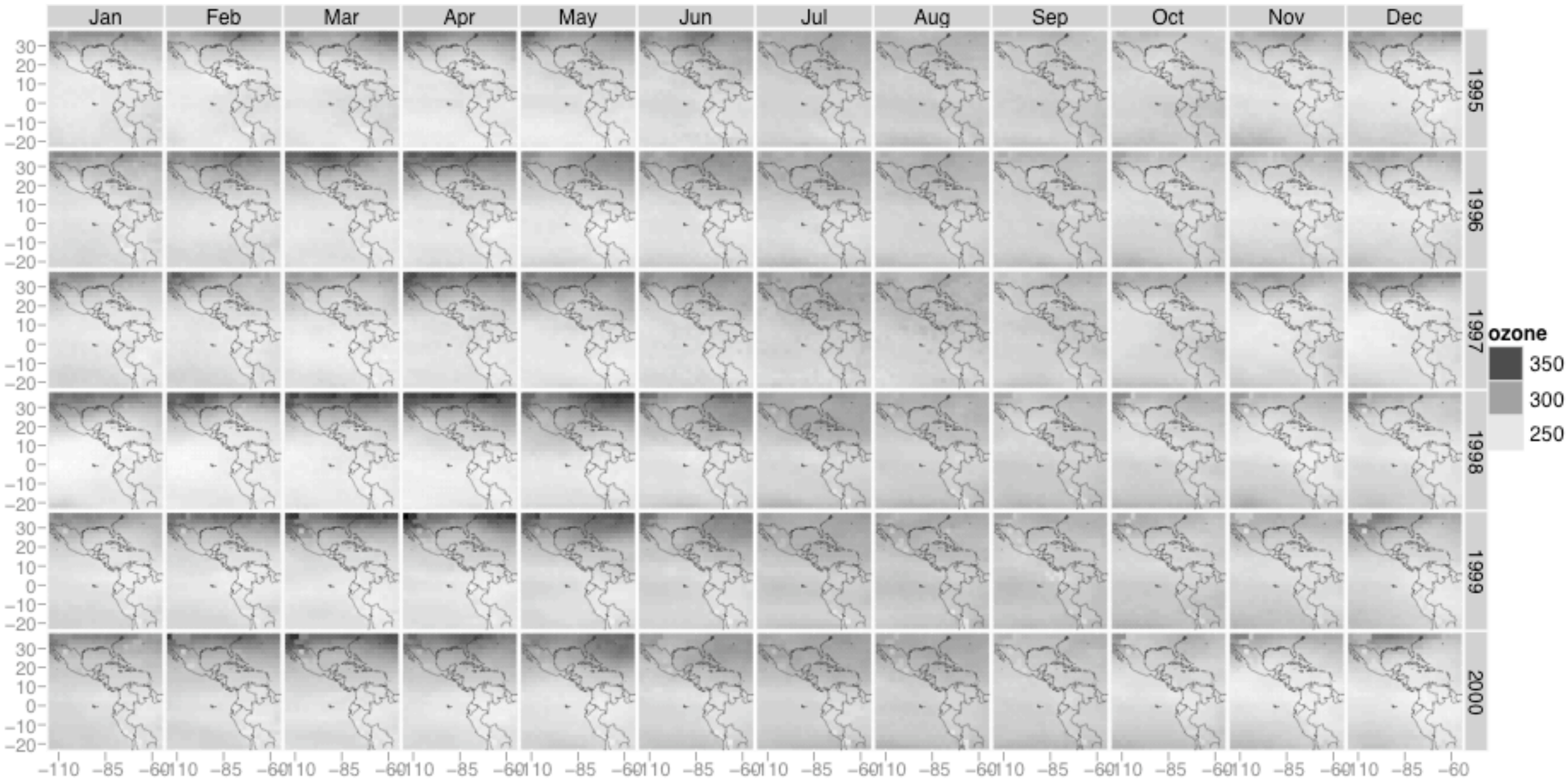
- Philosophy of ggplot
- Examples from a recent paper
- New methods facilitated by ggplot

Philosophy

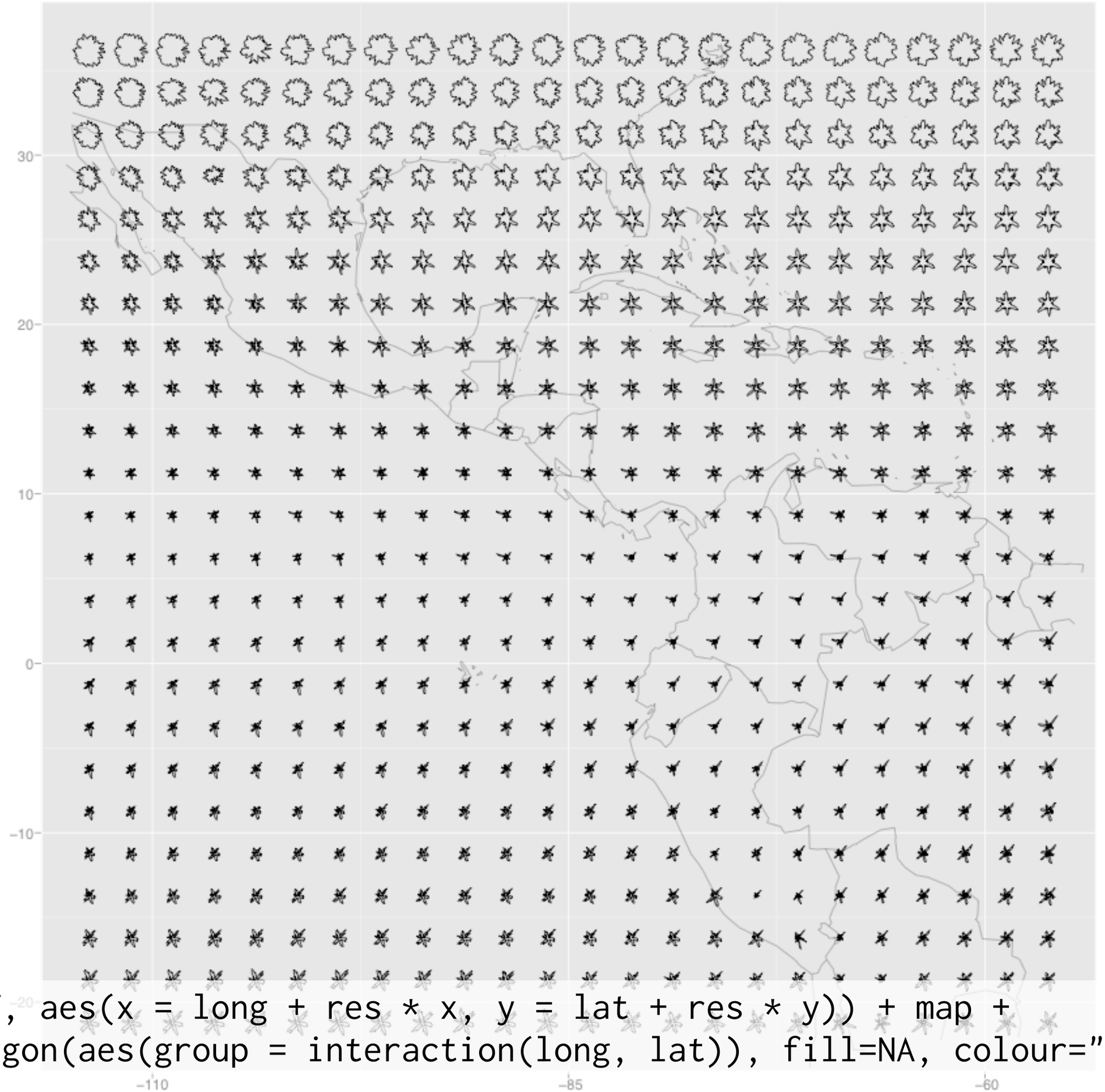
- Make graphics easier
- Use the grammar to facilitate research into new types of display
- Continuum of expertise:
  - start simple by using the results of the theory
  - grow in power by understanding the theory
  - begin to contribute new components
- Orthogonal components and minimal special cases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

Examples

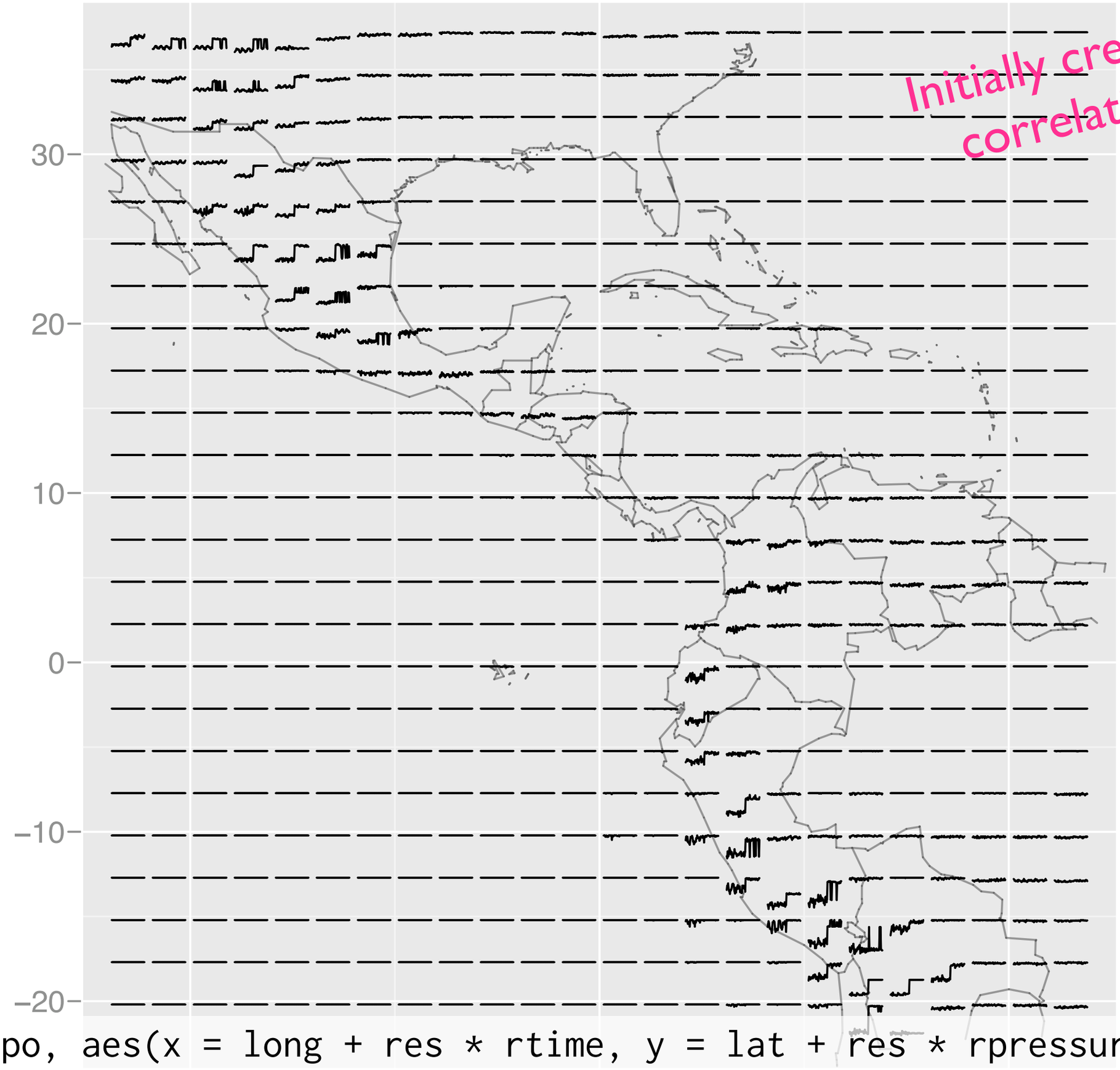
- J. Hobbs, H. Wickham, H. Hofmann, and D. Cook. **Glaciers melt as mountains warm: A graphical case study.** Computational Statistics. Special issue for ASA Statistical Computing and Graphics Data Expo 2006.
- Exploratory graphics created with GGobi, Mondrian, Manet, Gauguin and R, but needed consistent 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
```



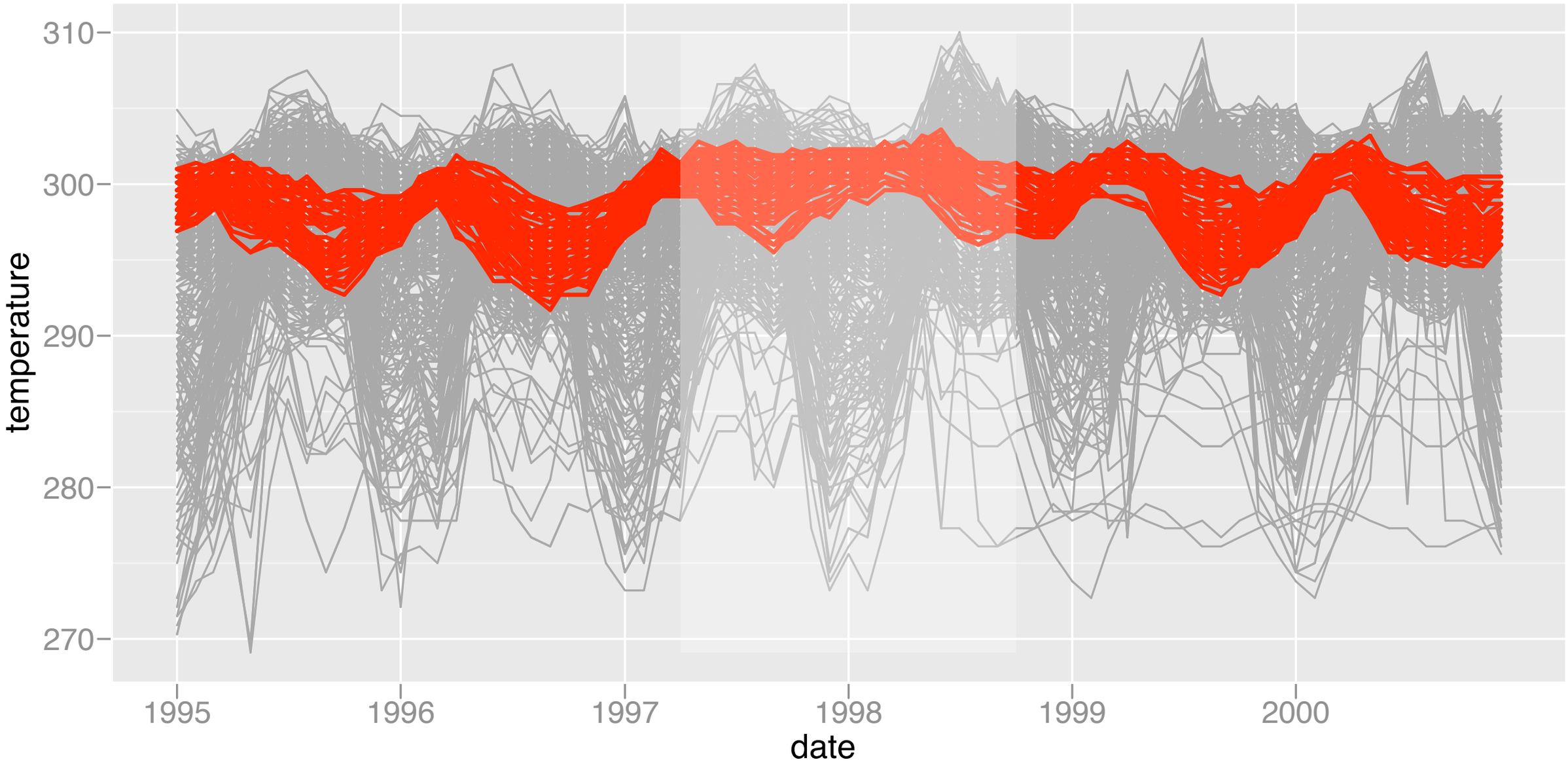

```
ggplot(df, aes(x = long + res * x, y = lat + res * y)) + map +
geom_polygon(aes(group = interaction(long, lat)), fill=NA, colour="black")
```




```
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))
)
```




```
qplot(date, temperature, data=clustered, group=id, geom="line")
+ pacific + elnino(clustered$temperature)
```

```
pacific <- brush(cluster %in% c(5,6))

brush <- function(condition, background = "grey60", brush = "red") {
  cond_string <- deparse(substitute(condition), width=500)
  colour <- paste(
    "ifelse(", cond_string, ", '", brush, "', '", background, "')", sep=""
  )
  order <- paste(
    "ifelse(", cond_string, ", 2, 1)", sep=""
  )
  size <- paste(
    "ifelse(", cond_string, ", 2, 1)", sep=""
  )
  list(
    aes_string(colour = colour, order = order, size=size),
    scale_colour_identity(),
    scale_size_identity()
  )
}
```

temperature
surftemp
ozone
cloudlow
cloudmid
cloudhigh
1
2
3
4
5
30
20
10
0
−10
−20

temperature
surftemp
ozone
cloudlow
cloudmid
cloudhigh
1
2
3
4
5
30
20
10
0
−10
−20
ggplot(clustered, aes(x = long, y = lat))
+ geom_tile(aes(width = 2.5, height = 2.5,
fill = factor(cluster)))
+ facet_grid(cluster ~ .)
+ map
+ scale_fill_brewer(palette="Spectral")
qplot(date, value, data = clusterm, group = id,
geom = "line", facets = cluster ~ variable,
colour = factor(cluster))
+ scale_y_continuous("", breaks=NA)
+ scale_colour_brewer(palette="Spectral")

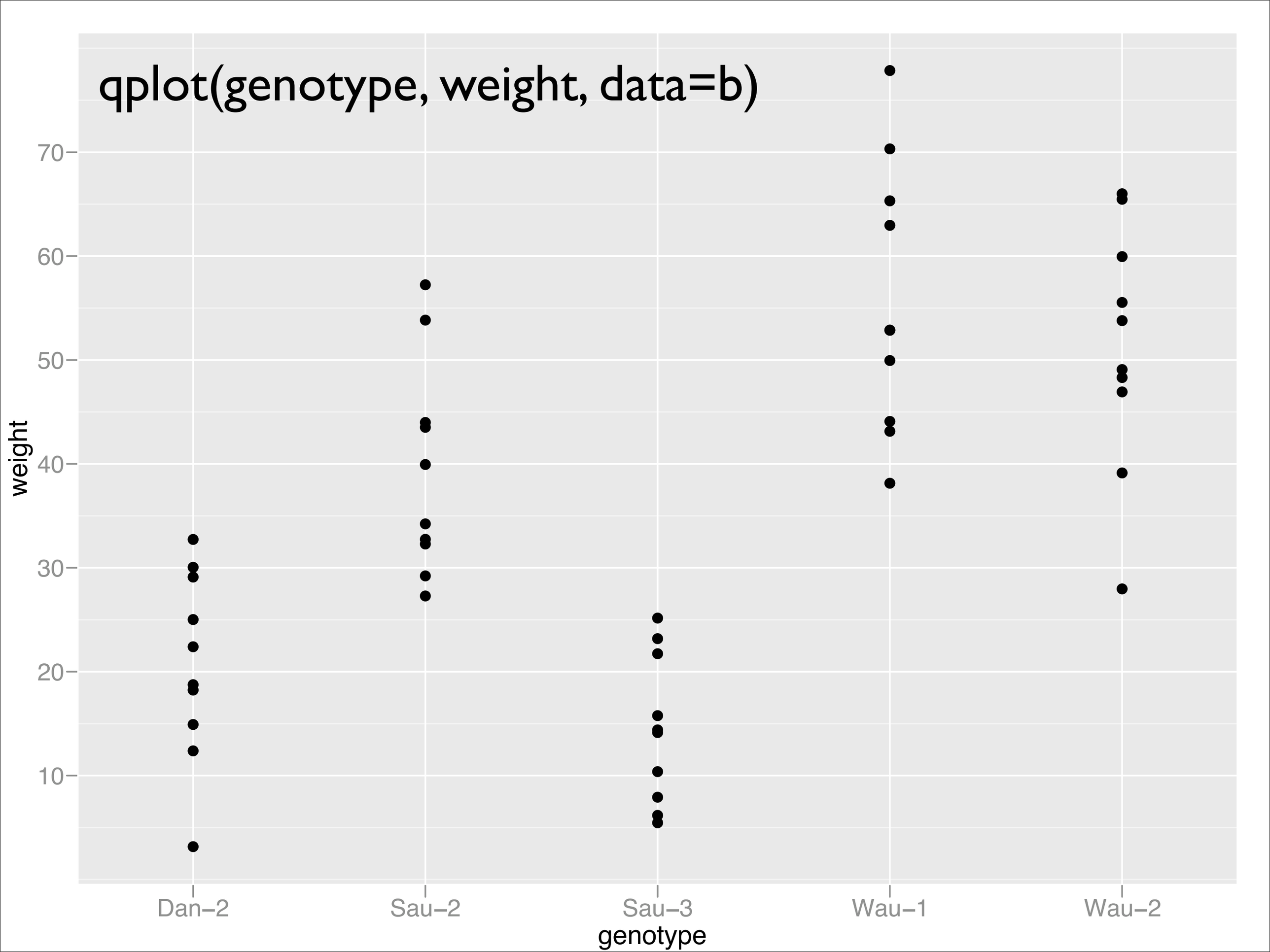
New methods

- Supplemental statistical summaries
- Iterating between graphics and models
- Tables of plots

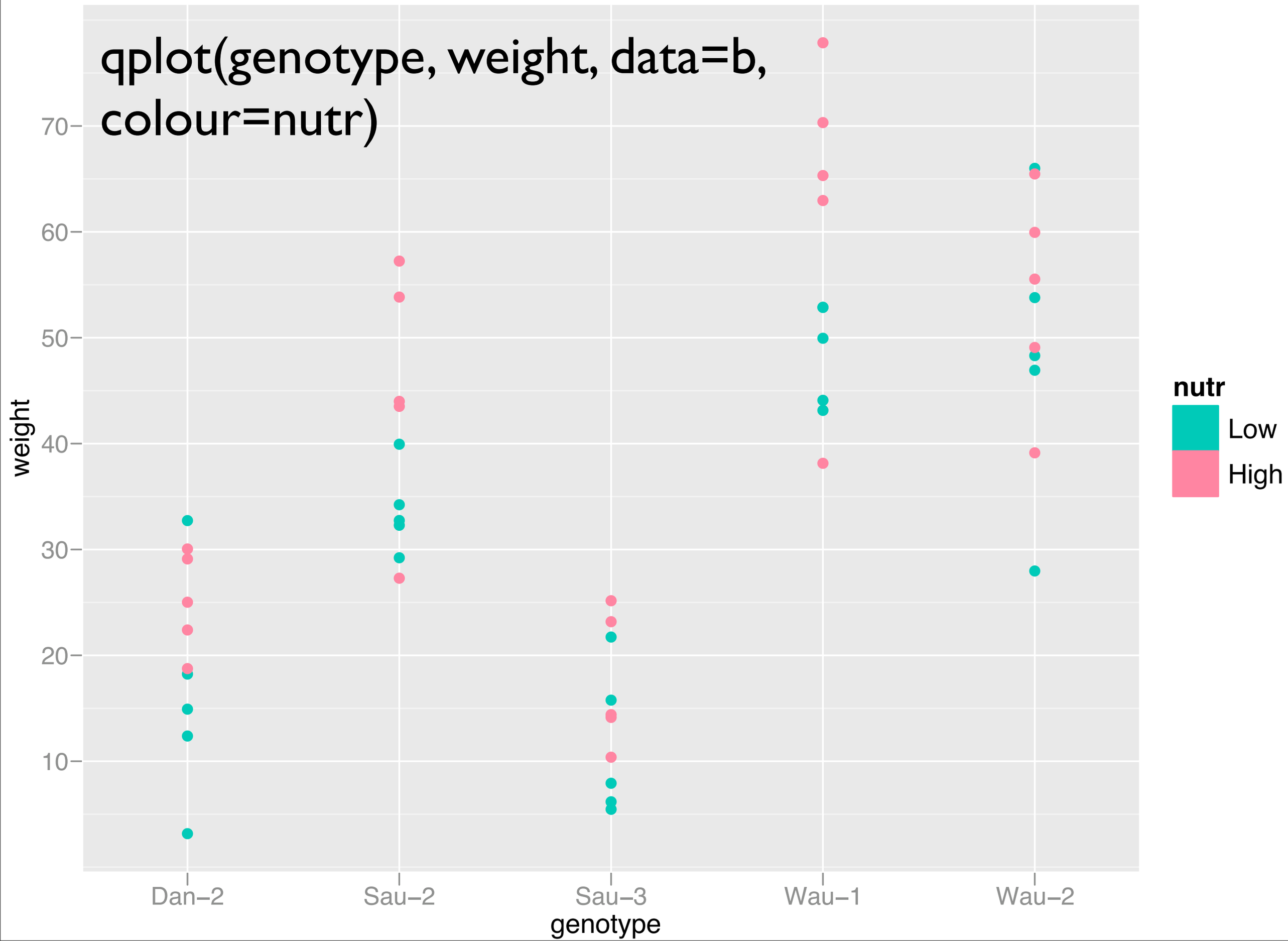
- Inspired by ideas of Tukey (and others)
- Exploratory graphics, not as pretty

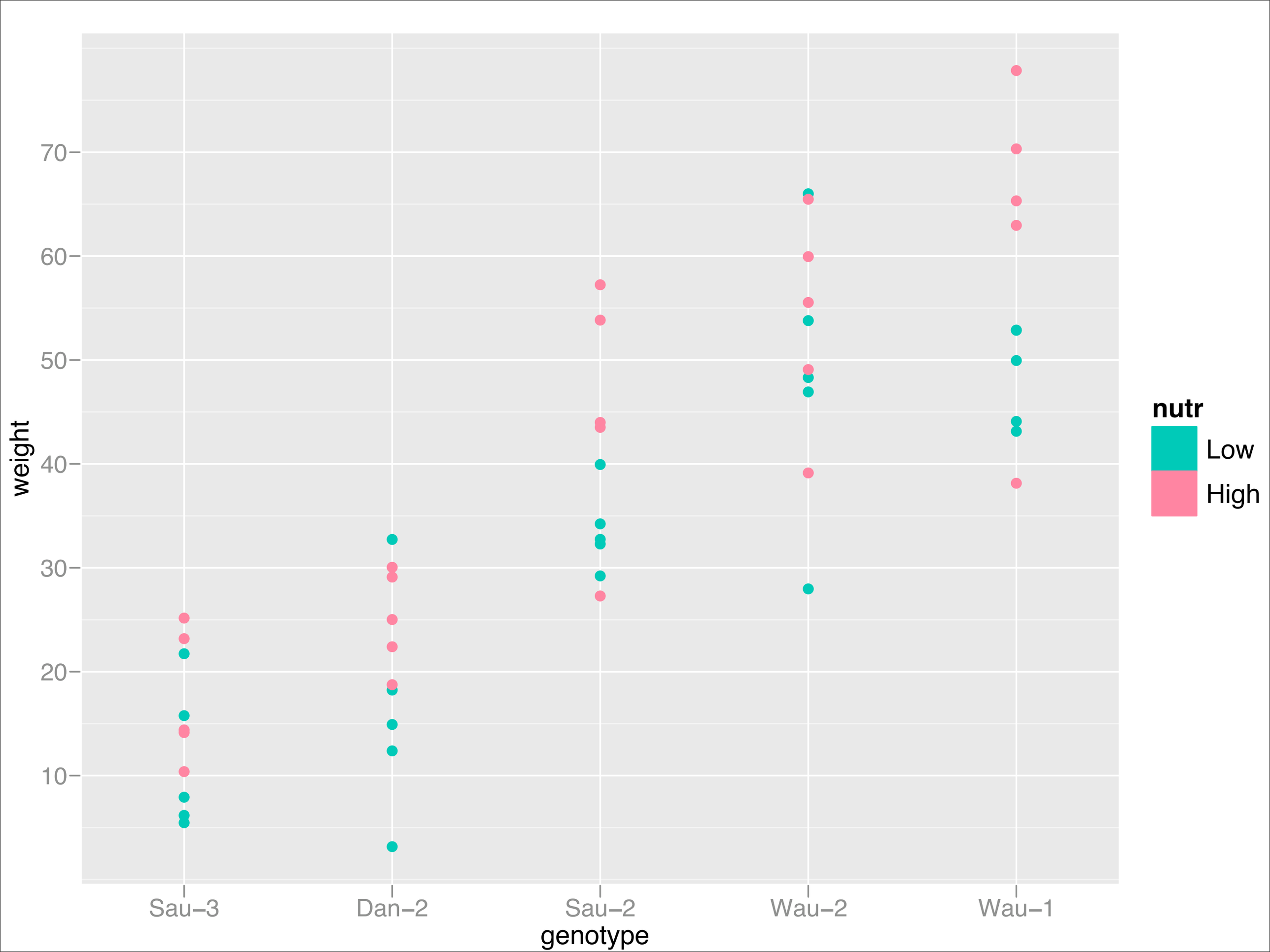
Intro to data

- 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)
weight
70
60
50
40
30
20
10
Dan–2
Sau–2
Sau–3
Wau–1
Wau–2
genotype

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




weight
70
60
50
40
30
20
10
Sau–3
Dan–2
Sau–2
Wau–2
Wau–1
genotype
nutr
Low
High

Comparing means

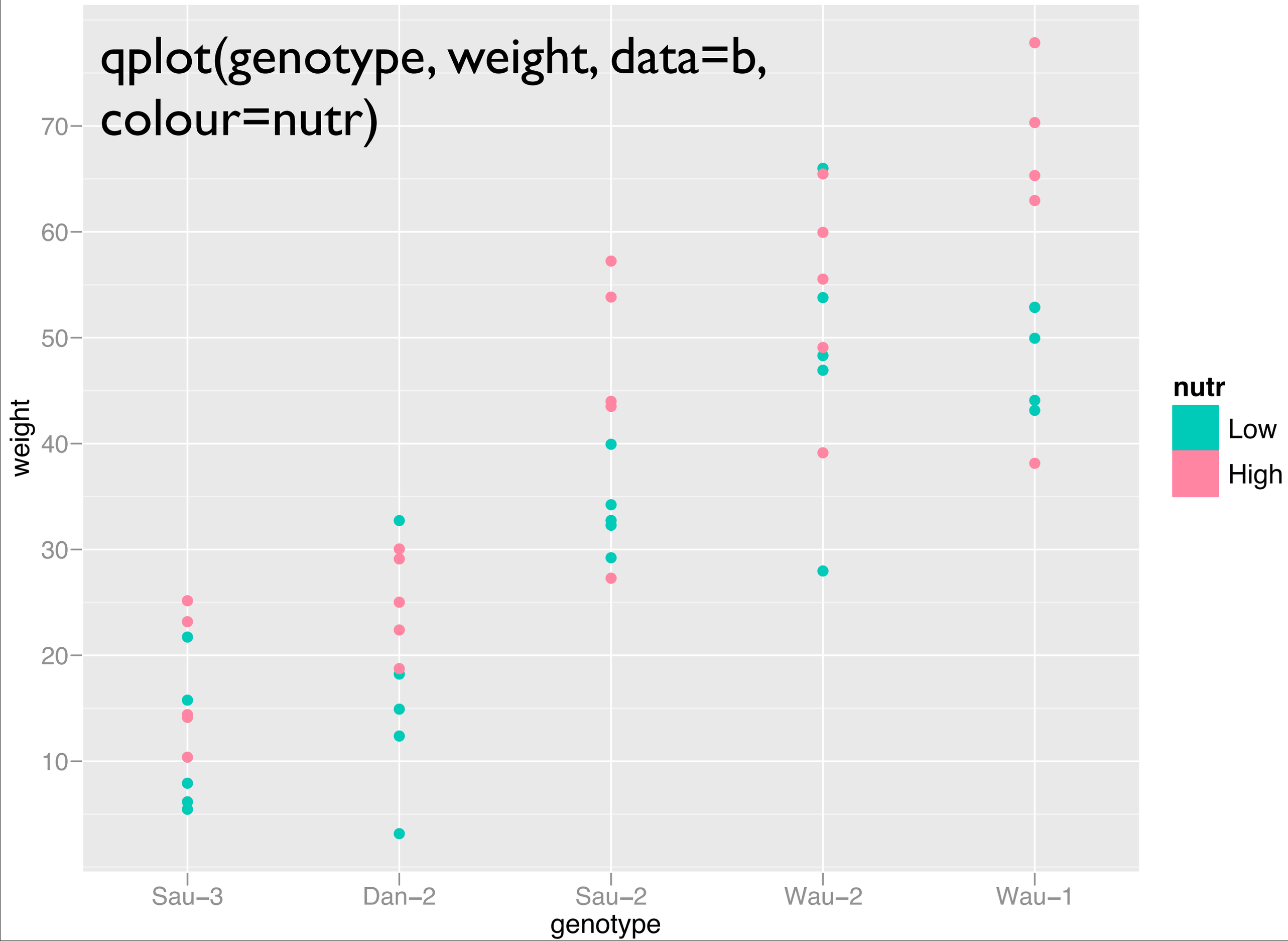
- For inference, interested in comparing the means of the groups
- But hard to do visually - eyes naturally compare ranges
- What can we do? - Visual ANOVA

Supplemental summaries

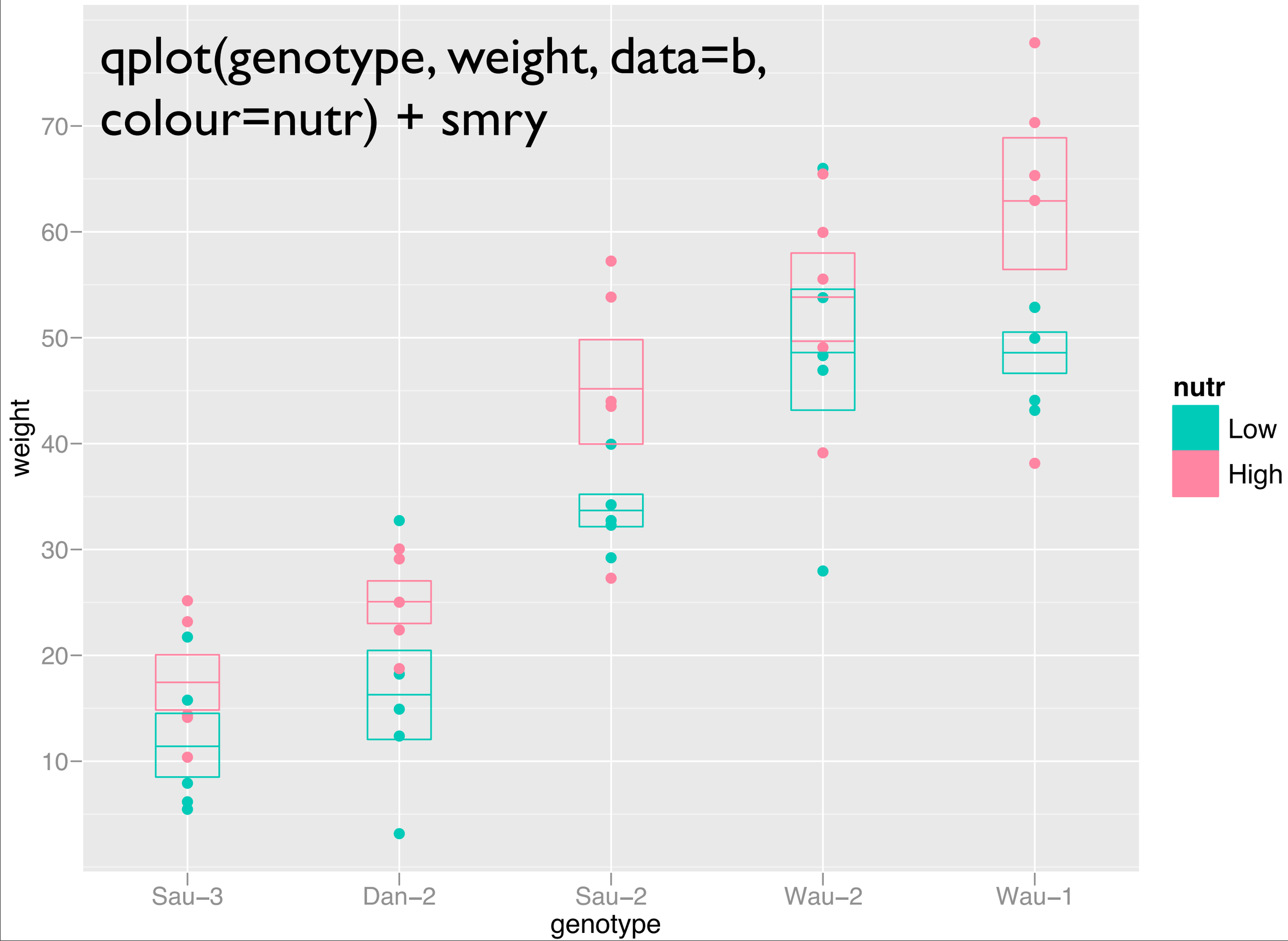
From Hmisc

- ```
  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)



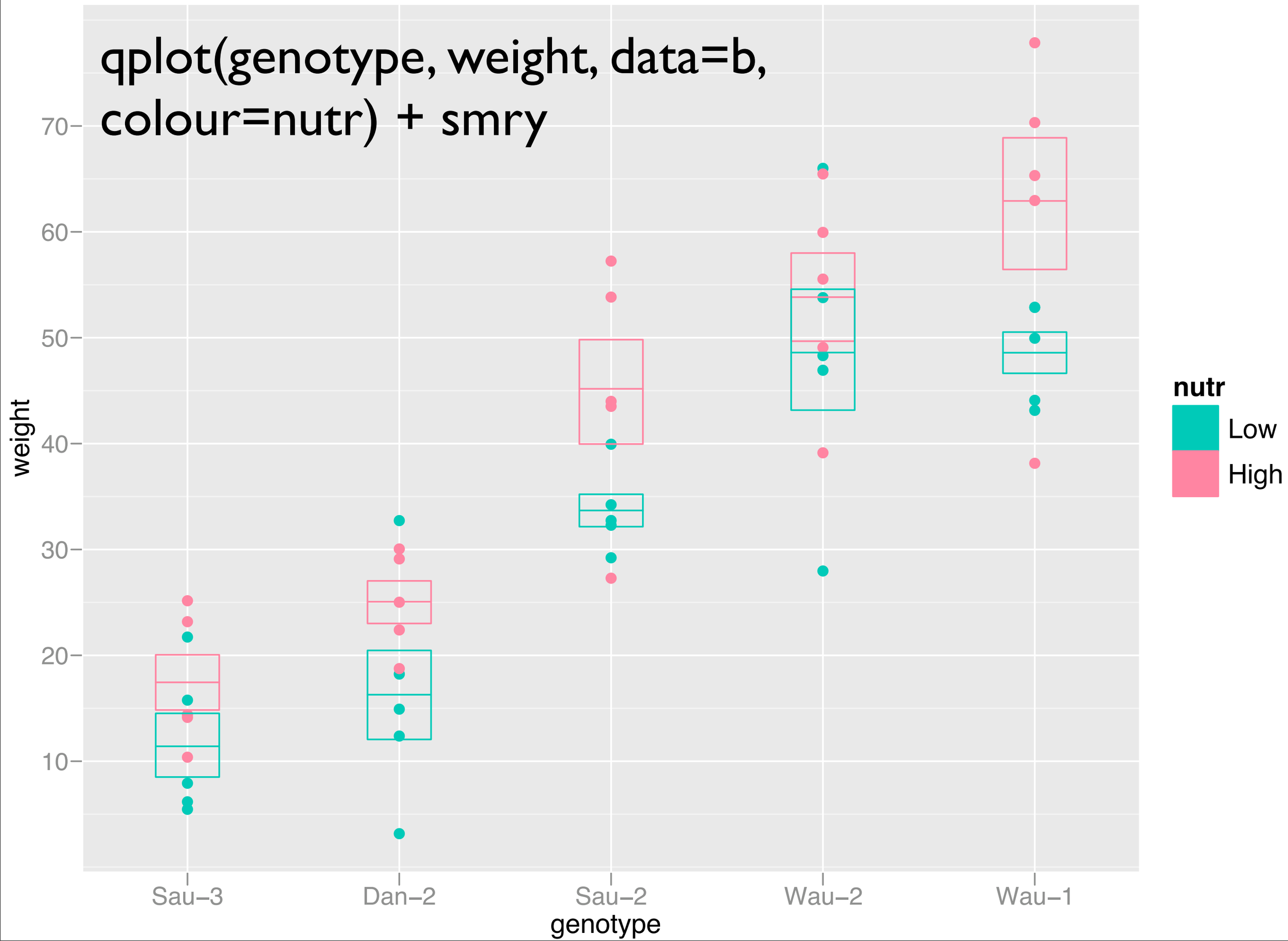
qplot(genotype, weight, data=b, colour=nutr) + smry



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




```
b$weight2 <- resid(lm(weight ~ genotype, data=b))
qplot(genotype, weight2, data=b, colour=nutr) + smry
```


```
b$weight3 <- resid(lm(weight ~ genotype + nutr, data=b))
qplot(genotype, weight3, data=b, colour=nutr) + smry
```

```
               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
```

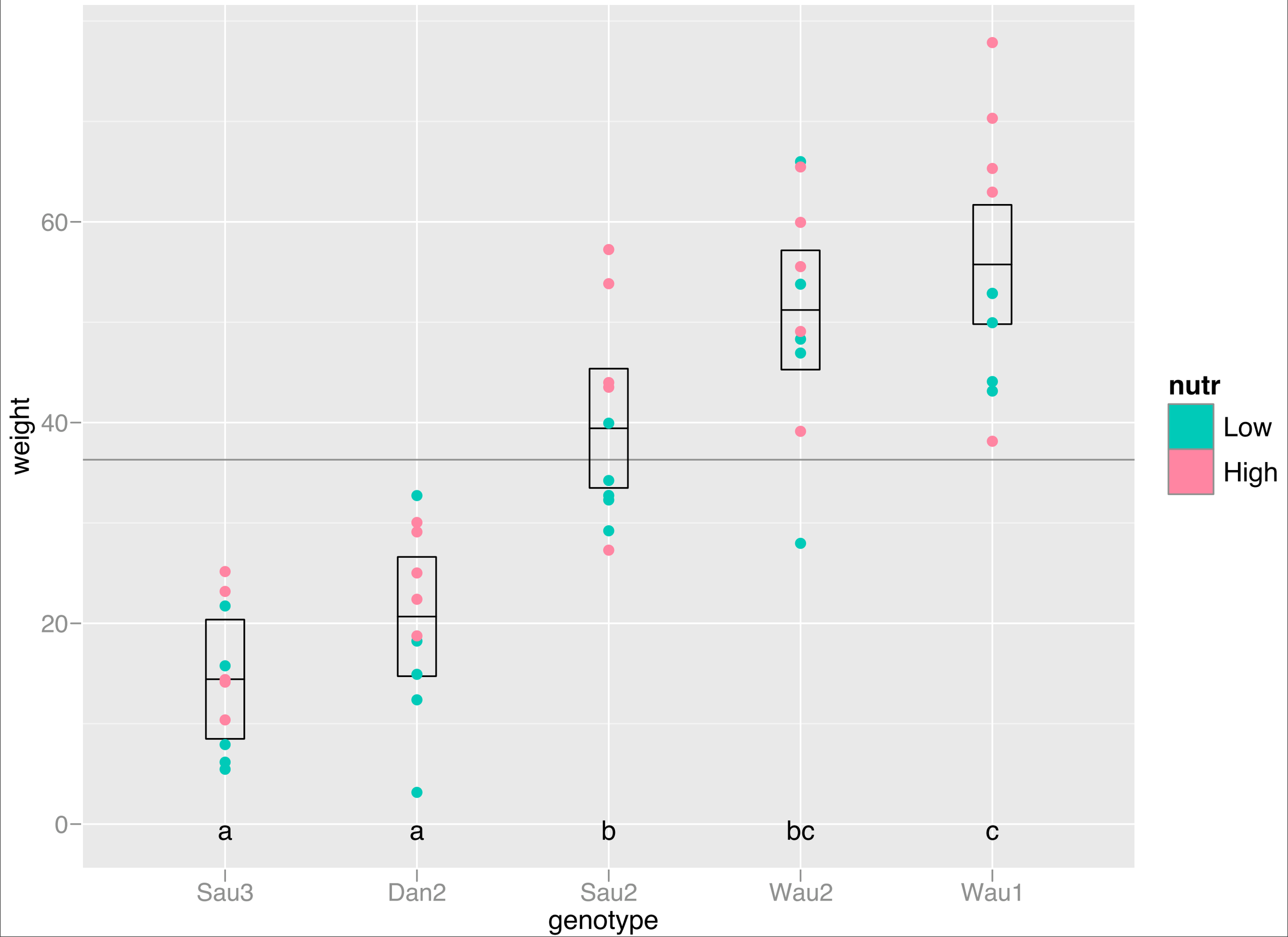
```
anova(lm(weight ~ genotype * nutr, data=b))
```

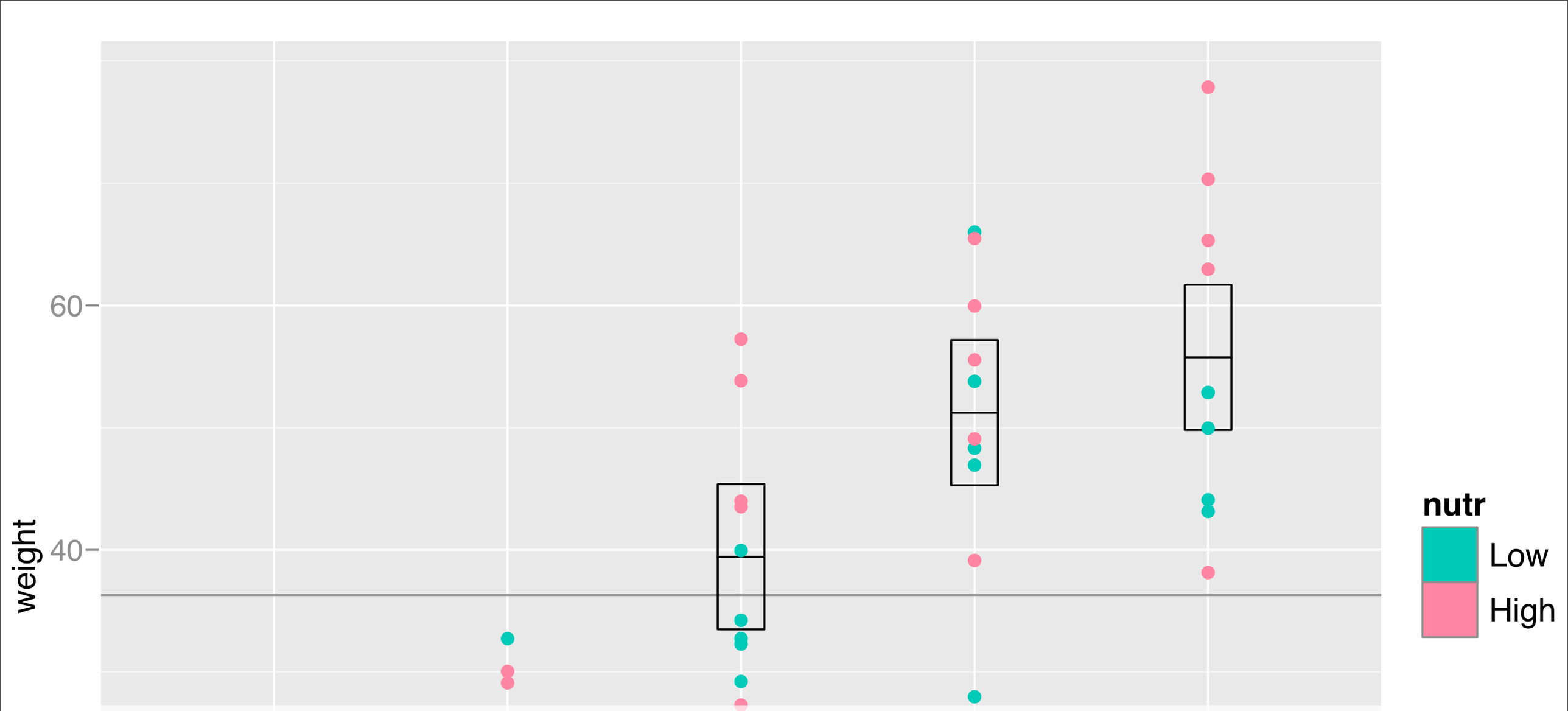
Graphics → Model

- 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


60
40
20
0
weight
a
a
b
bc
c
Sau3
Dan2
Sau2
Wau2
Wau1
genotype
nutr
Low
High




```
ggplot(b, aes(x=genotype, y=weight))
+ geom_hline(intercept = mean(b$weight))
+ geom_crossbar(aes(y=fit, min=lower,max=upper),
  data=geffect)
+ geom_point(aes(colour = nutr))
+ geom_text(aes(label = group), data=geffect)
```

Tables of plots

- Often interested in marginal, as well as conditional, relationships
- Or comparing one subset to the whole, rather than to other subsets
- Like in contingency table, we often want to see margins as well

Dan–2
Sau–2
Sau–3
Wau–1
Wau–2
High
Low
n
4.5
4.0
3.5
3.0
2.5
2.0
GM
NGM
trt
trt
NGM
GM

Dan–2
Sau–2
Sau–3
Wau–1
Wau–2
(all)
High
Low
(all)
n
4.5
4.0
3.5
3.0
2.5
2.0
GM
NGM
trt
trt
NGM
GM

Arranging plots

- Facilitate comparisons of interest
- Small differences need to be closer together (big differences can be far apart)
- Connections to model?

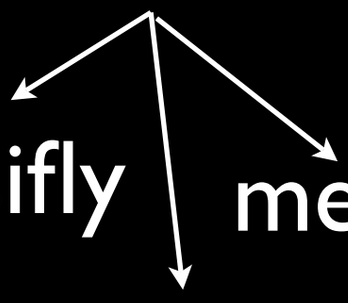
Summary

- 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

Future

GGobi

rggobi



classifly

meifly

clusterfly

faceoff

geozoo

ggplot2

reshape

fda

hints

localmds

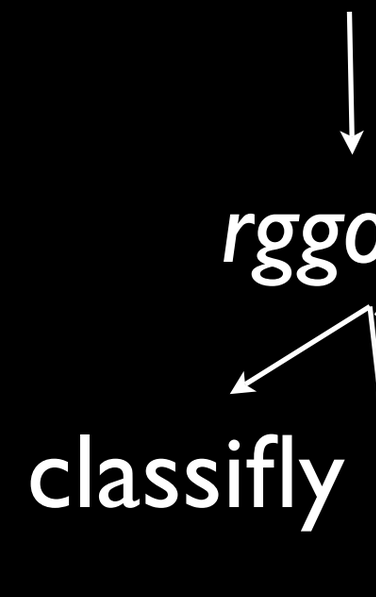
lvboxplot

scagnostics

DescribeDisplay

Bio- and bibliographic tools for statistics

GGobi



rggobi

classifly

meifly

clusterfly

faceoff

geozoo

ggplot2

reshape

fda

hints

localmds

lvboxplot

scagnostics

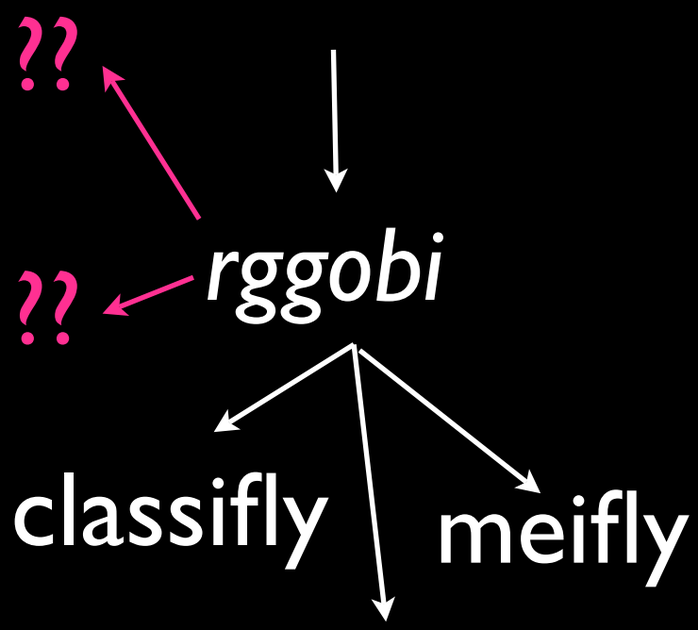
DescribeDisplay

New methods

Foundations of statistical graphics

Dissemination and outreach

GGobi



faceoff *geozoo*

ggplot2

A grammar of interactive graphics

A grammar of graphics for categorical data

reshape2

fda

hints

localmds

lvboxplot

scagnostics

DescribeDisplay

New methods

Foundations of statistical graphics

Dissemination and outreach

Questions?