Really? Using the nullabor package to learn if what we see is really there

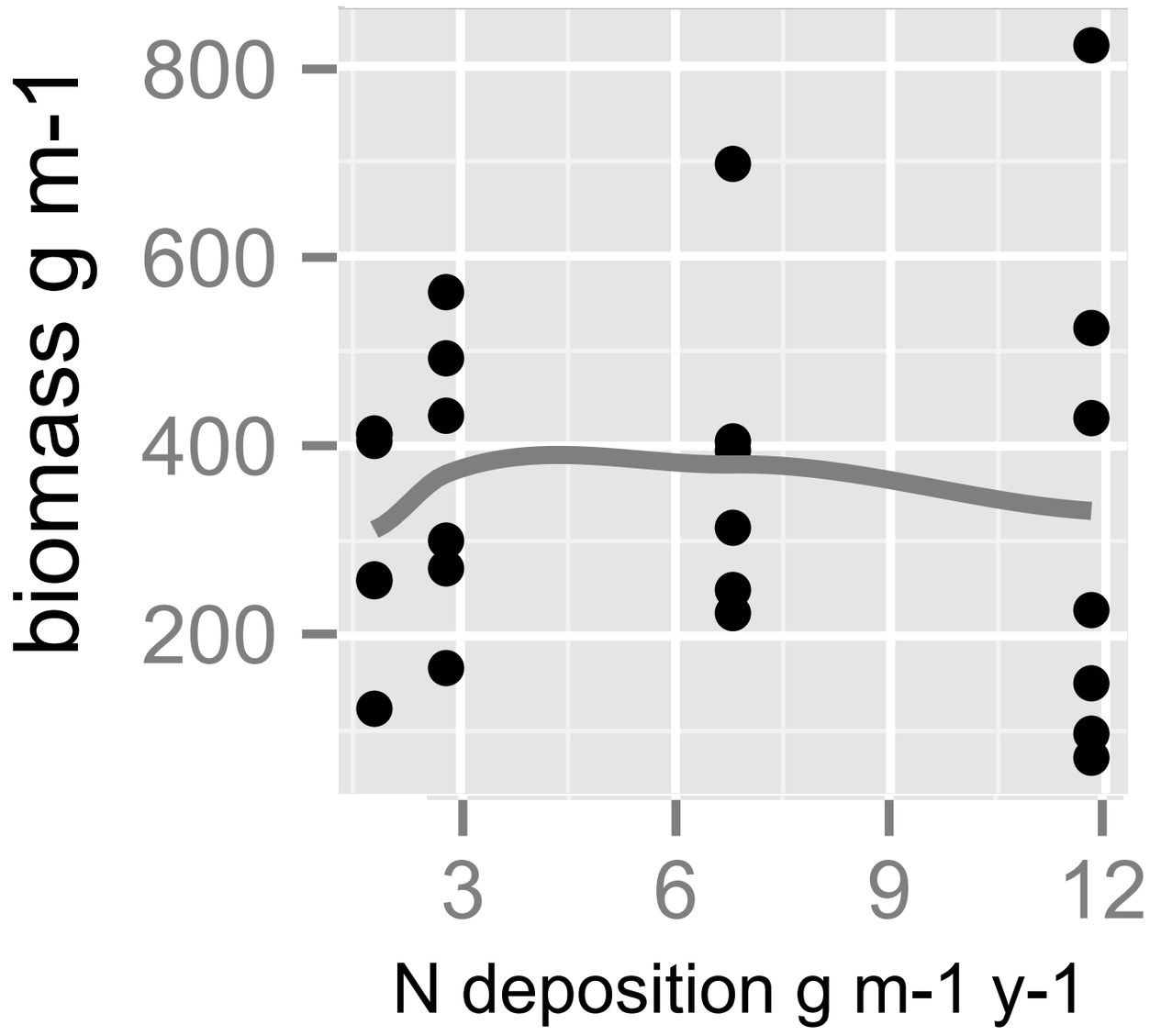
Di Cook, Monash University
Joint with Hadley Wickham, Heike Hofmann,
Niladri Roy Chowdhury, Mahbub Majumder

Photo by Lyn Cook

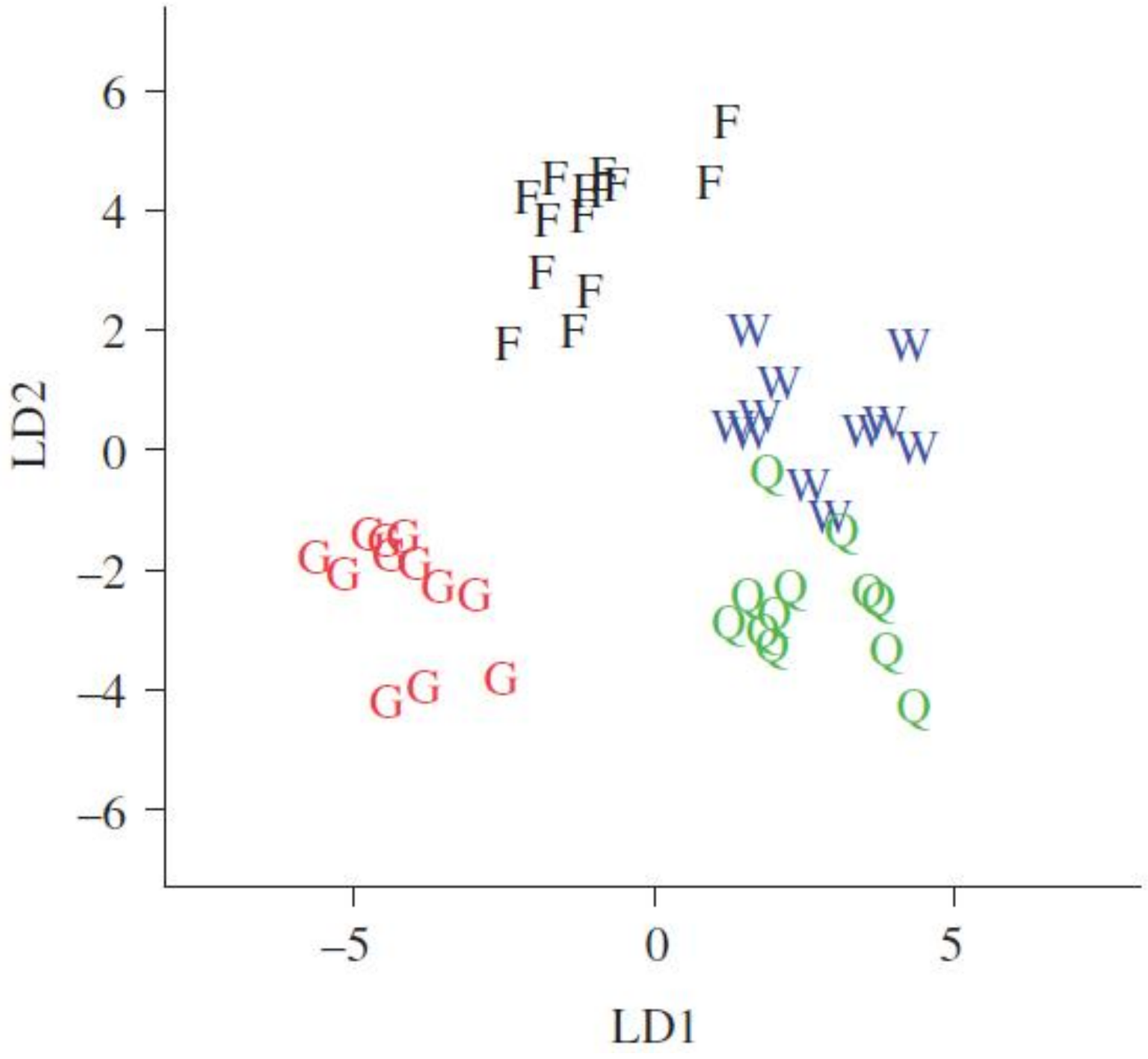
Outline

- Why?
- lineup, rorschach functions
- null generating mechanisms
- p-values
- metrics

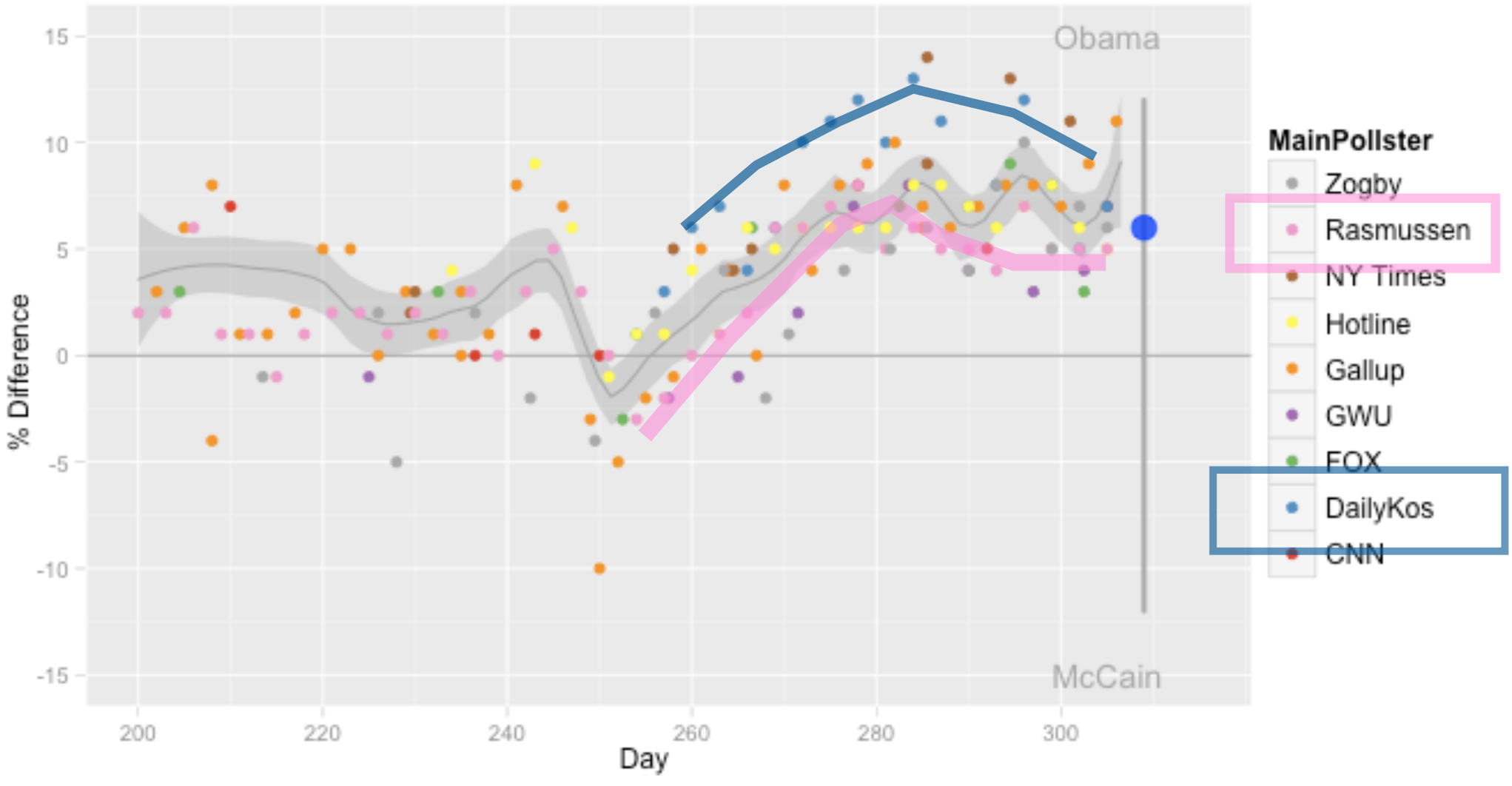
"Biomass really looks related to nitrogen deposition, but none of my tests show a significant relationship!" Ecologist colleague



"These four species of wasps have very different gene expression patterns" Published paper 2010



"Is it possible that the pollsters are systematicably biased?" Our US election monitoring



Why inference?

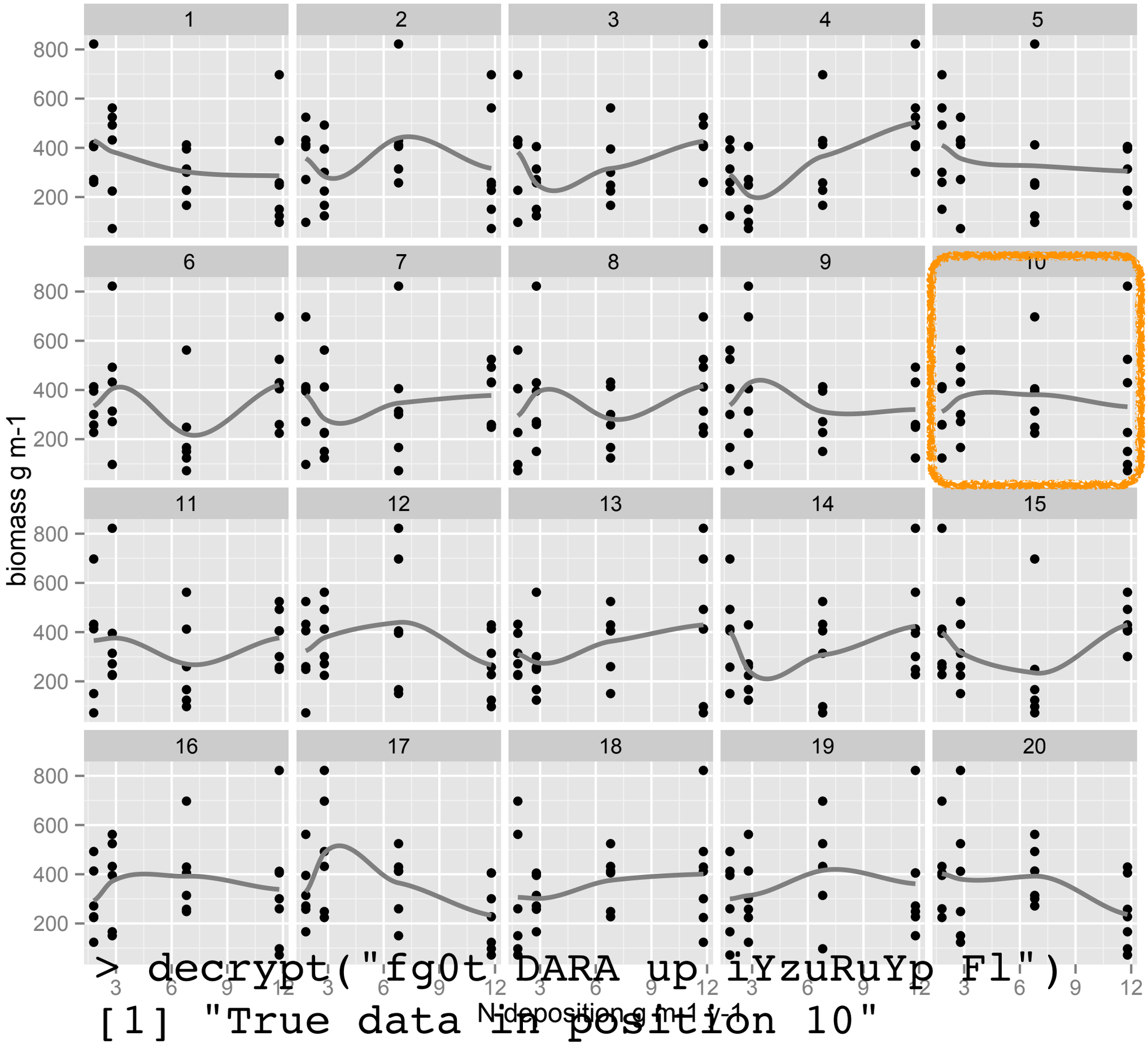
- Plots of data allow us to uncover the unexpected, but it needs to be calibrated against what might be seen by chance, if there really is no underlying pattern
- Classical statistical inference allows computing probabilities of this being a likely value of a statistic if there really is no structure

Inference

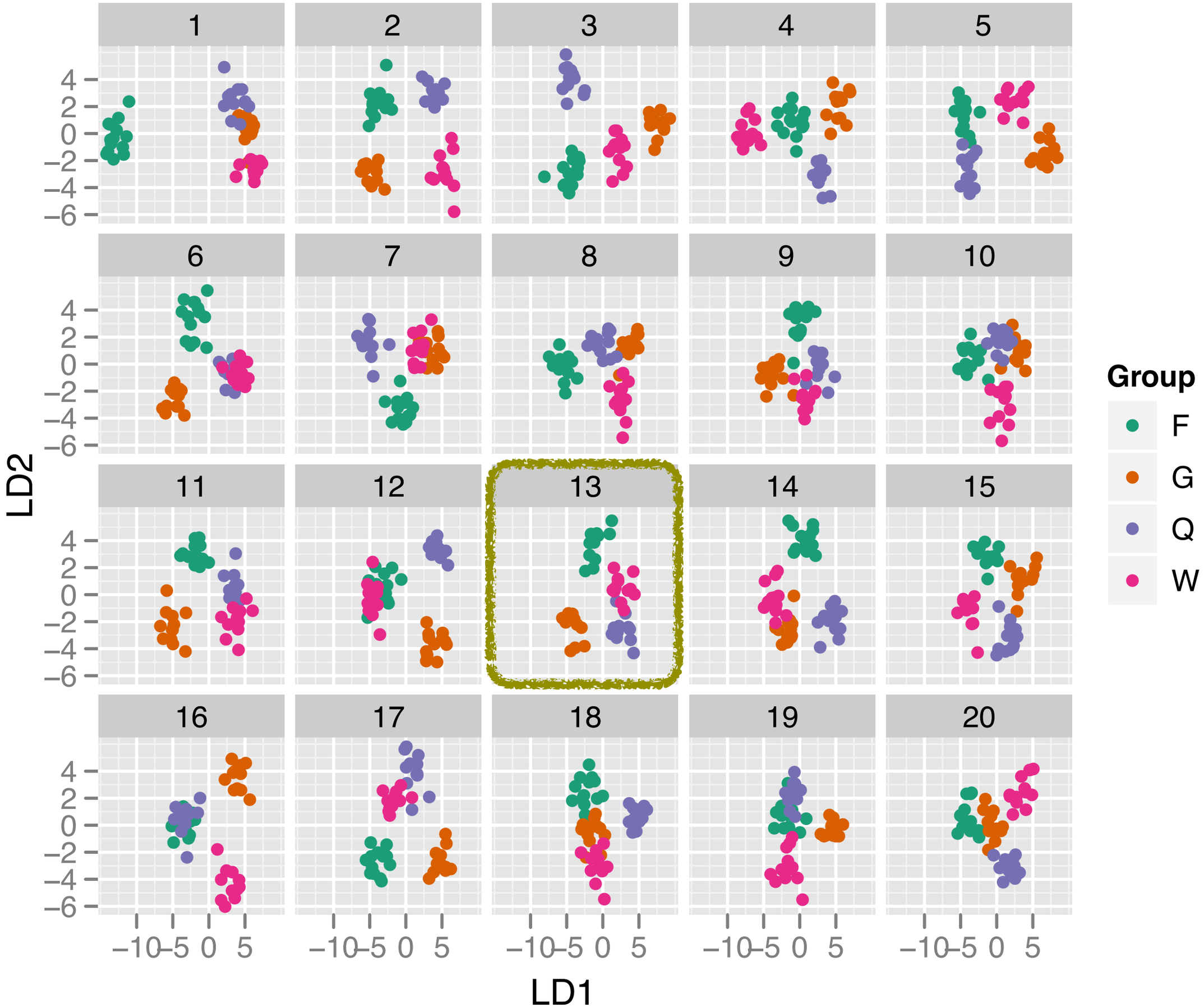
- Once you see it, its too late
- You cannot legitimately test for significance of structure

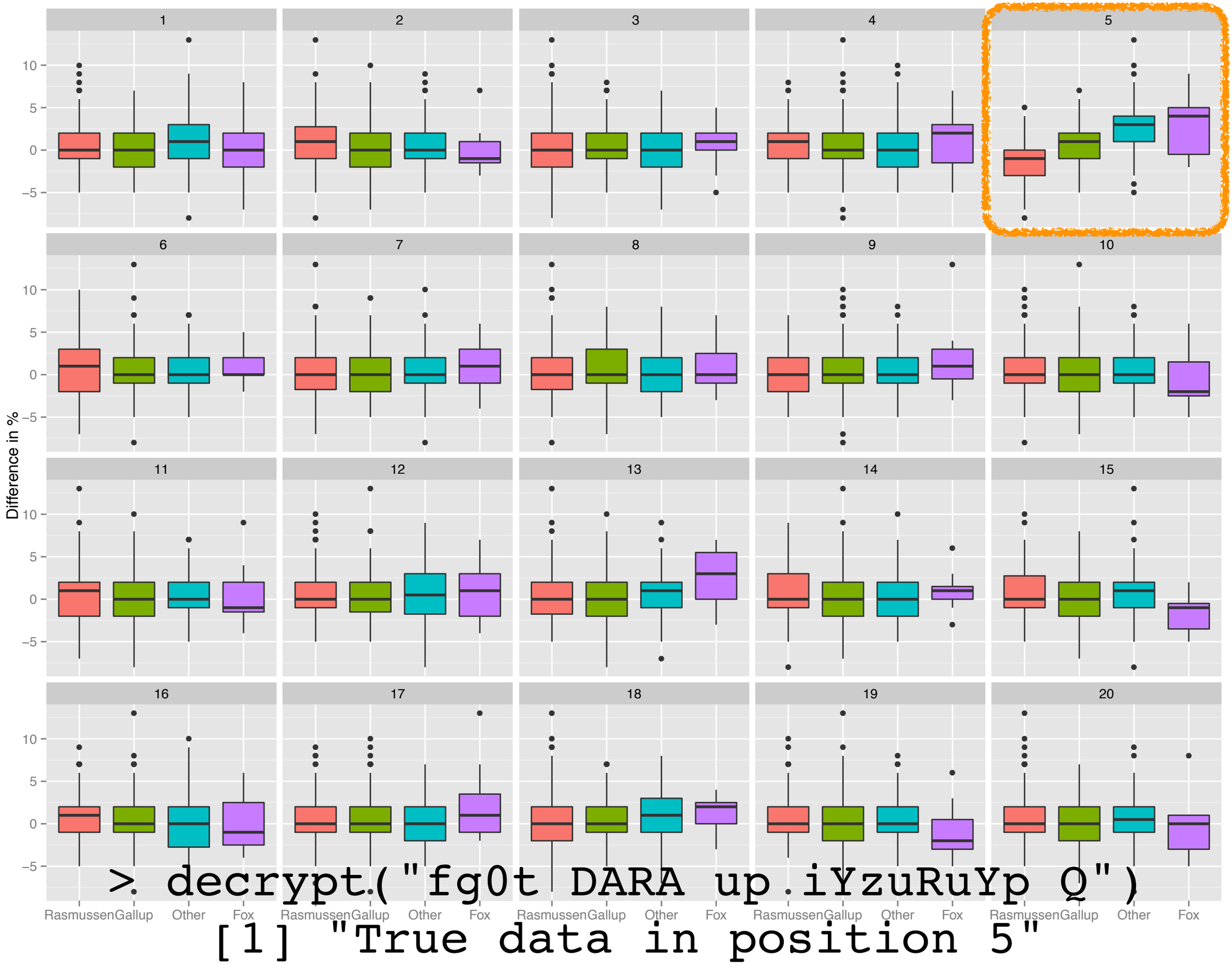
nullabor

- Lineup protocol: Plots your data among a field of "null" plots
- Puts it in the context of what it might look like if there is really no structure
- Encrypts the location of the data plot




```
> decrypt("fg0t DARA up iYzuRuYp Fl")
[1] "True data in position 10"
```


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
4
2
0
−2
−4
−6
−10
−5
0
5
LD1
LD2
Group
F
G
Q
W

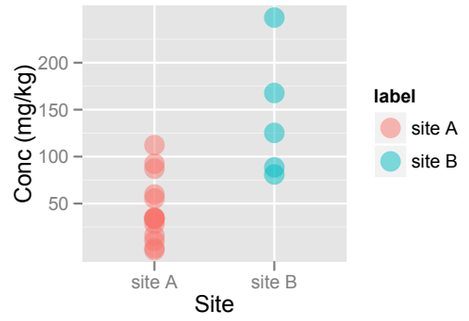
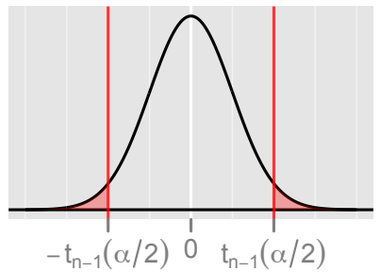
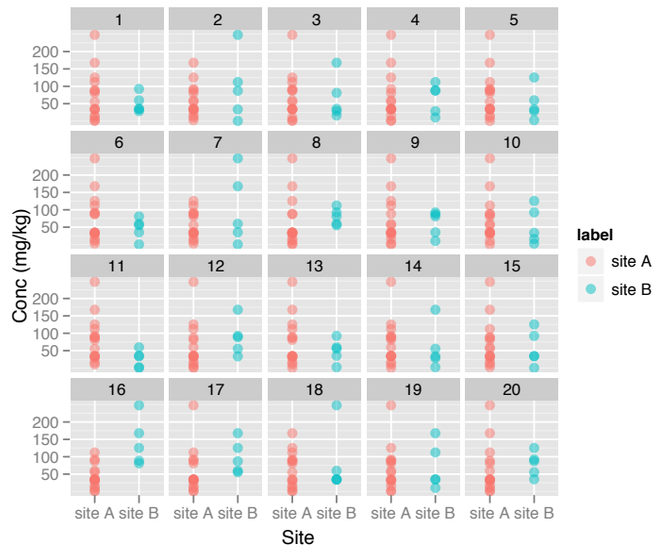


> decrypt("fg0t DARA up iYzuRuYp Q")
[1] "True data in position 5"

nullabor functions

- `lineup:` Generates a lineup using one of the given null generating mechanisms
- `pvisuals:` Compute p-values after showing to impartial jurers
- `distmet:` empirical distribution of distance between data plot and null plots

Demo

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	↓	↓
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t);$ 	$f_{T(y)}(t);$ 
	↓	↓
Reject H_0 if	observed T is extreme	observed plot is identifiable

Visual p-values

- For one observer, the probability of randomly selecting the data plot is 1/m, where m is the number of plots in the lineup.

- With multiple observers, the p-value is estimated by

Number of independent observers

$$P(X \geq x) = 1 - Binom_{K,1/m}(x - 1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^{i} \left(\frac{m-1}{m}\right)^{K-i}$$

Number of observers choosing data plot

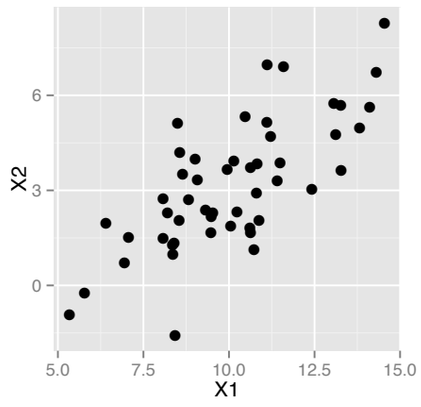
Null generators

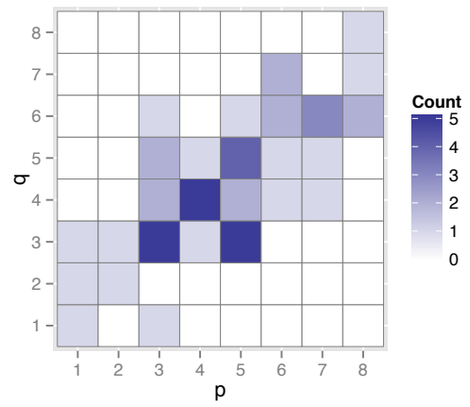
- `null_dist`: Null hypothesis: variable has specified distribution
- `null_lm`: Null hypothesis: variable is linear combination of predictors, comes with different residual generators
- `null_permute`: Null hypothesis: variable is independent of others

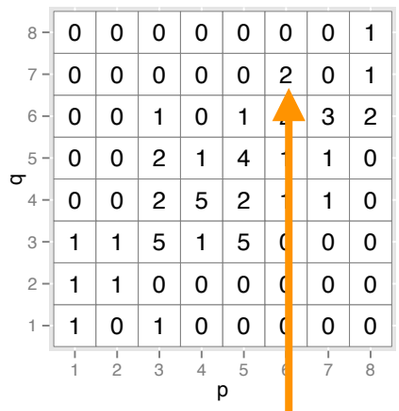
Distance metrics

Can we measure how different the data plot is from the null plots?

(a) Dataset X with two variables X_1 and X_2

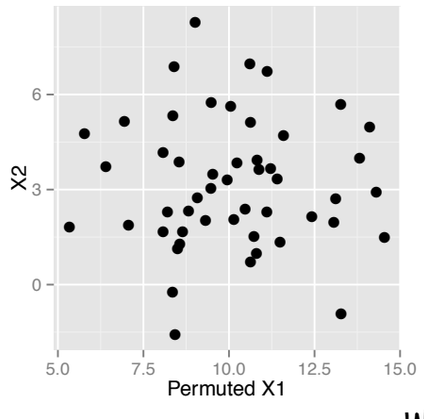


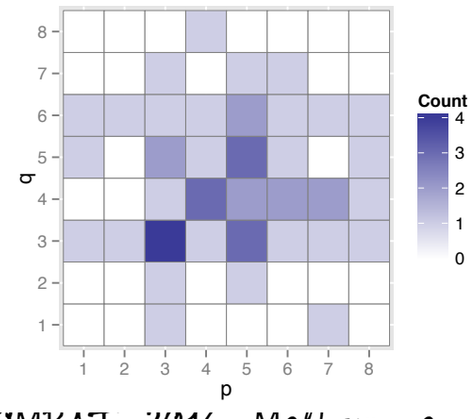


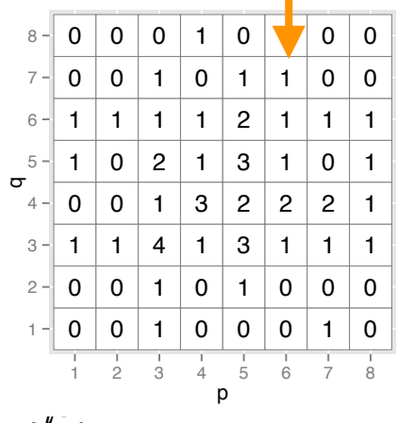


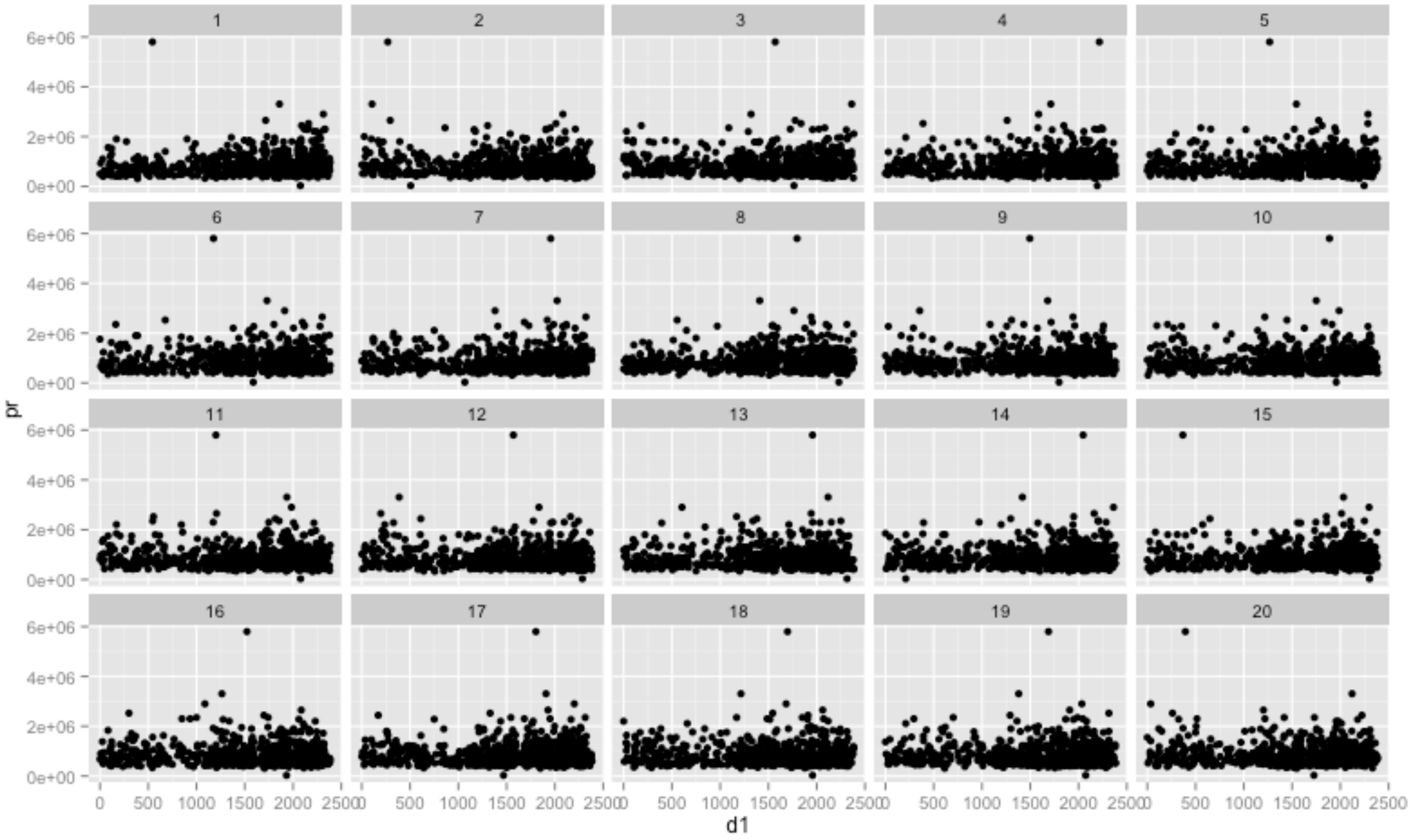
Difference

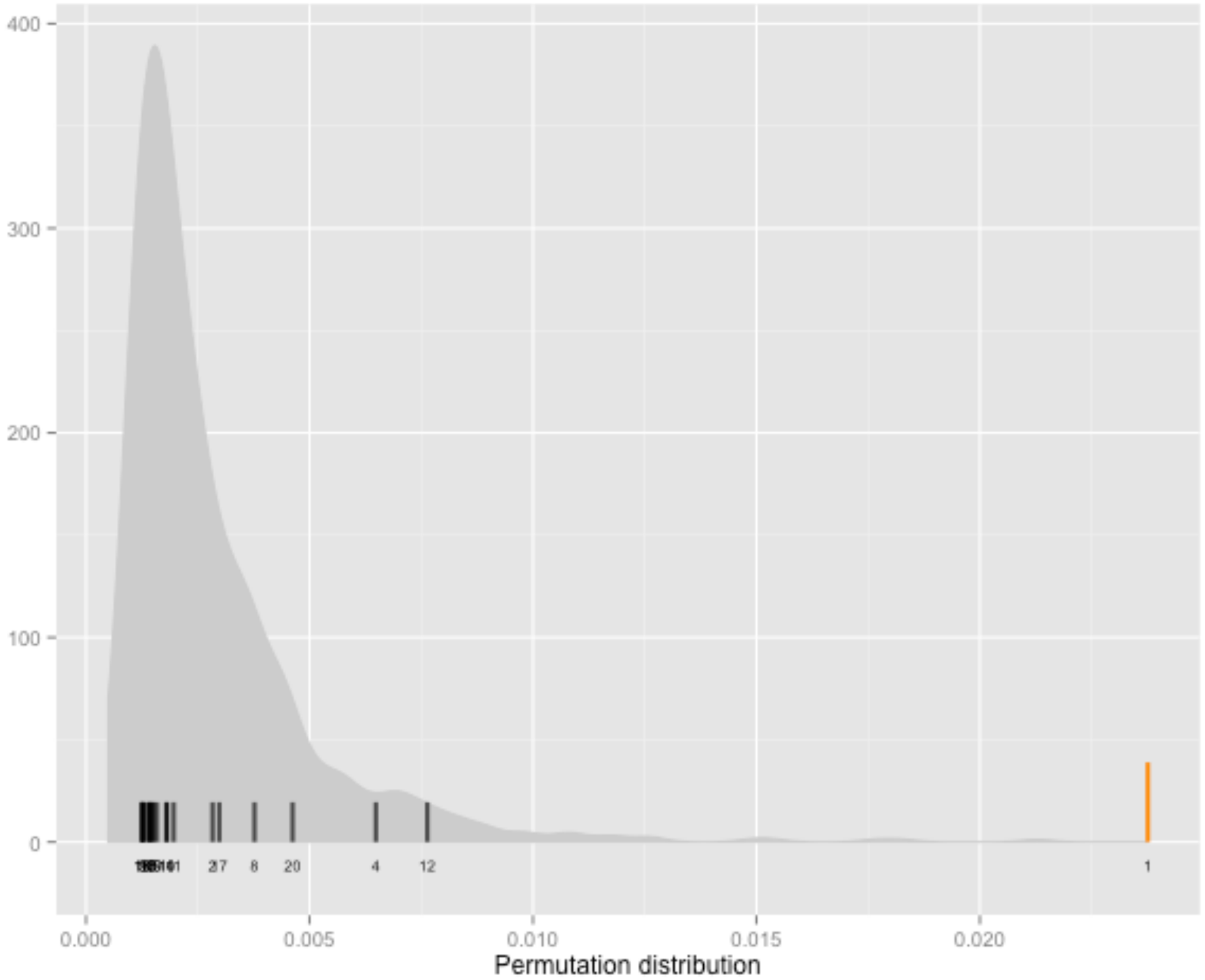
(b) Dataset Y with permuted X_1 and original X_2








1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
6e+06
4e+06
2e+06
0e+00
pr
0
500
1000
1500
2000
2500
d1


400
300
200
100
0
27
8
20
4
12
1
0.000
0.005
0.010
0.015
0.020
Permutation distribution

Summary

- Really useful package
- Helps to adjust our expectations, dampen surprise, support surprise
- Calibrate your eyes on what randomness looks like