Statistical Inference by Crowd Sourcing

Di Cook
Econometrics and Business Statistics
Monash University

Joint work with Heike Hofmann, Mahbub Majumder, Debby Swayne, Eun-kyung Lee, Hadley Wickham, Andreas Buja, Lendie Follett, Adam Loy, Susan Vanderplas, Eric Hare, Niladri Roy Chowdhury, Nathaniel Tomasetti, Tengfei Yin

Cornell university, Nov 4, 2015

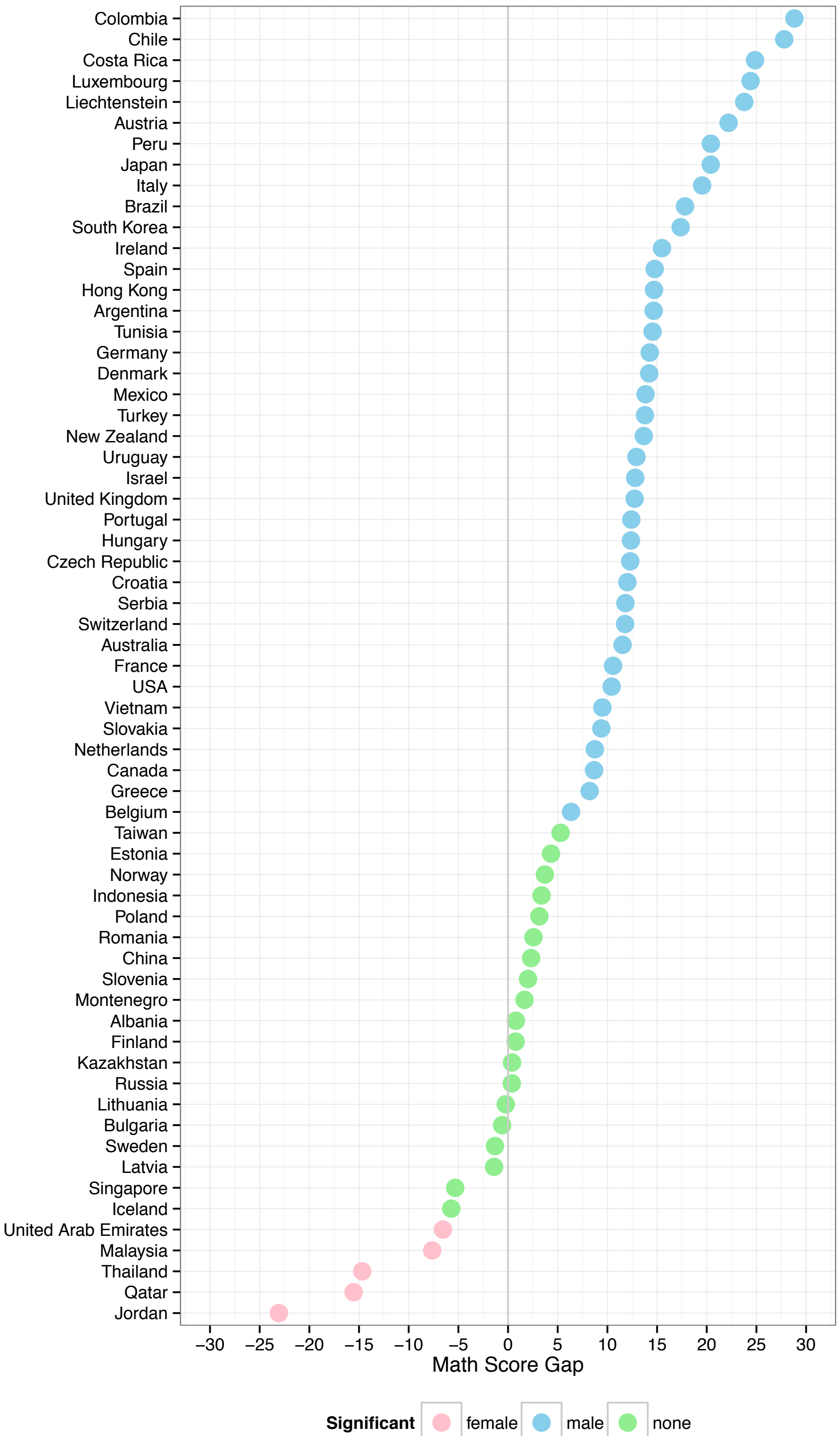
Motivation

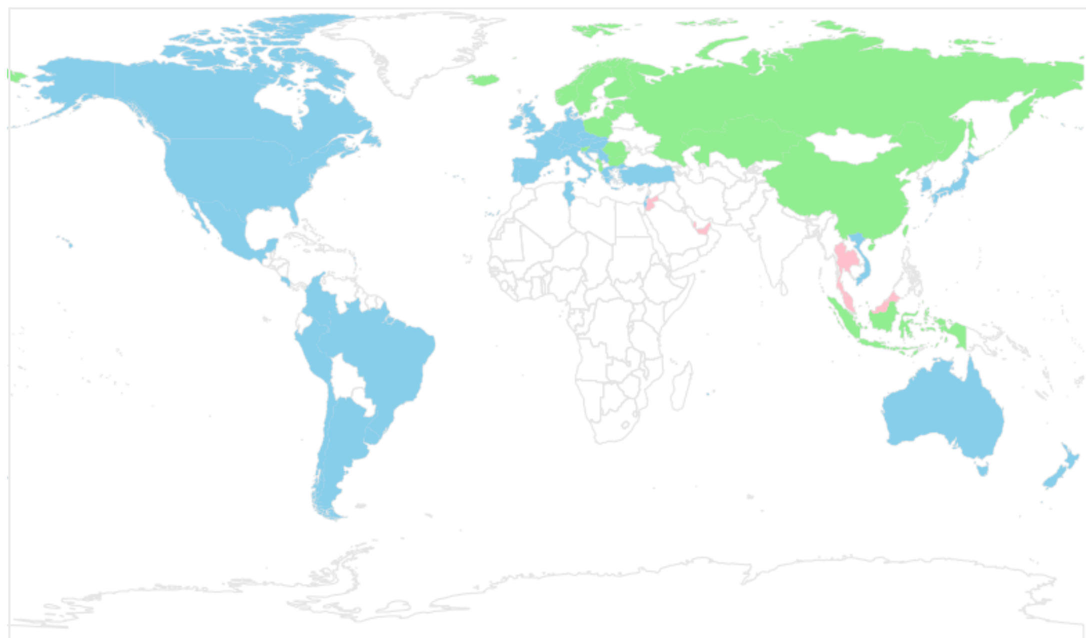
We can learn a lot about our world by making pictures of data, without making formal statistical inference.

- PISA data 2012 measuring 15 year olds workforce readiness, http://www.oecd.org/pisa/pisaproducts/datavisualizationcontest.htm
- US Airline traffic 1986-today http://www.transtats.bts.gov/

Gender & Math

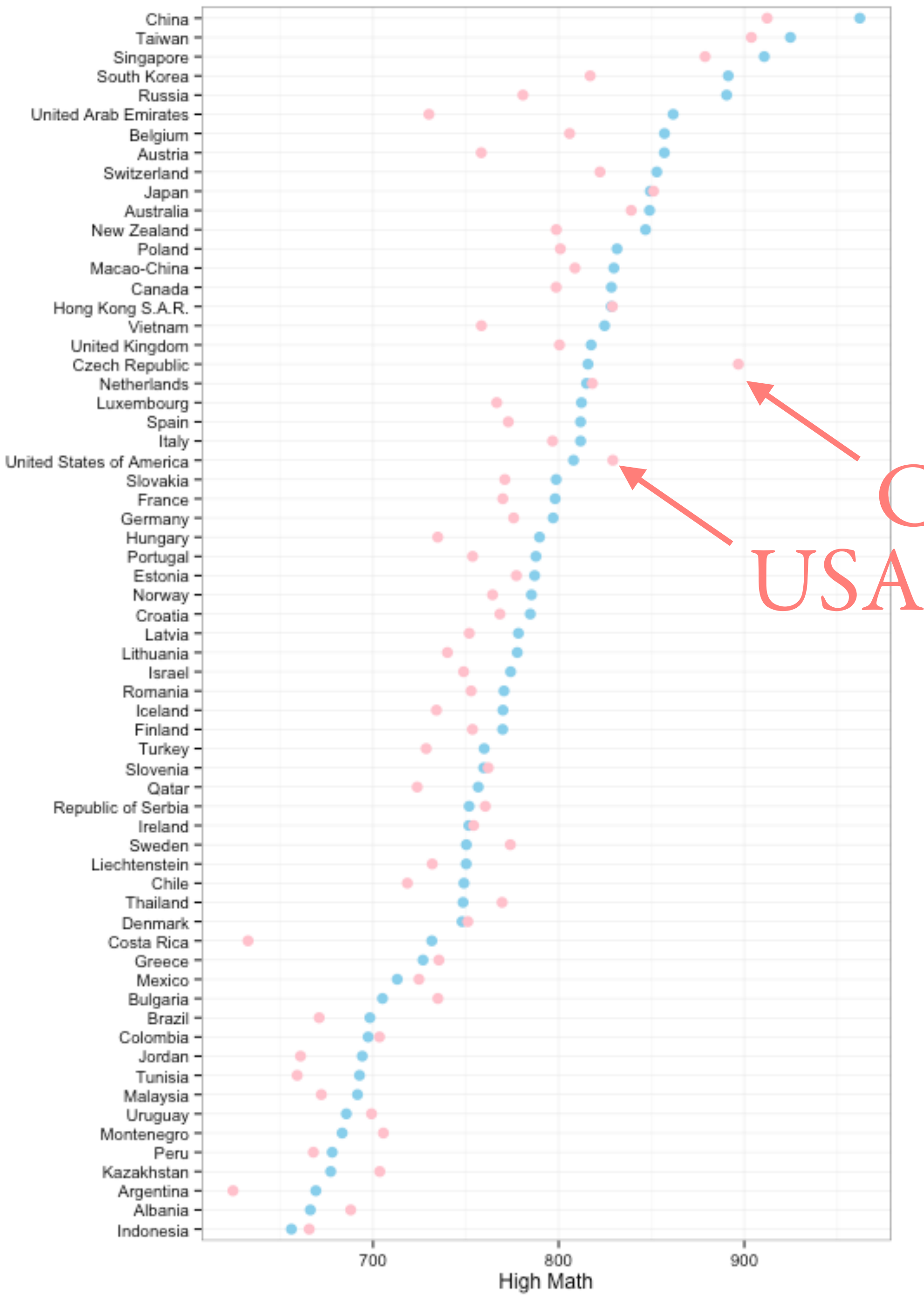
1. Compute weighted means by country and by gender.
2. Show mean difference by country
3. t-test of difference (unadjusted)



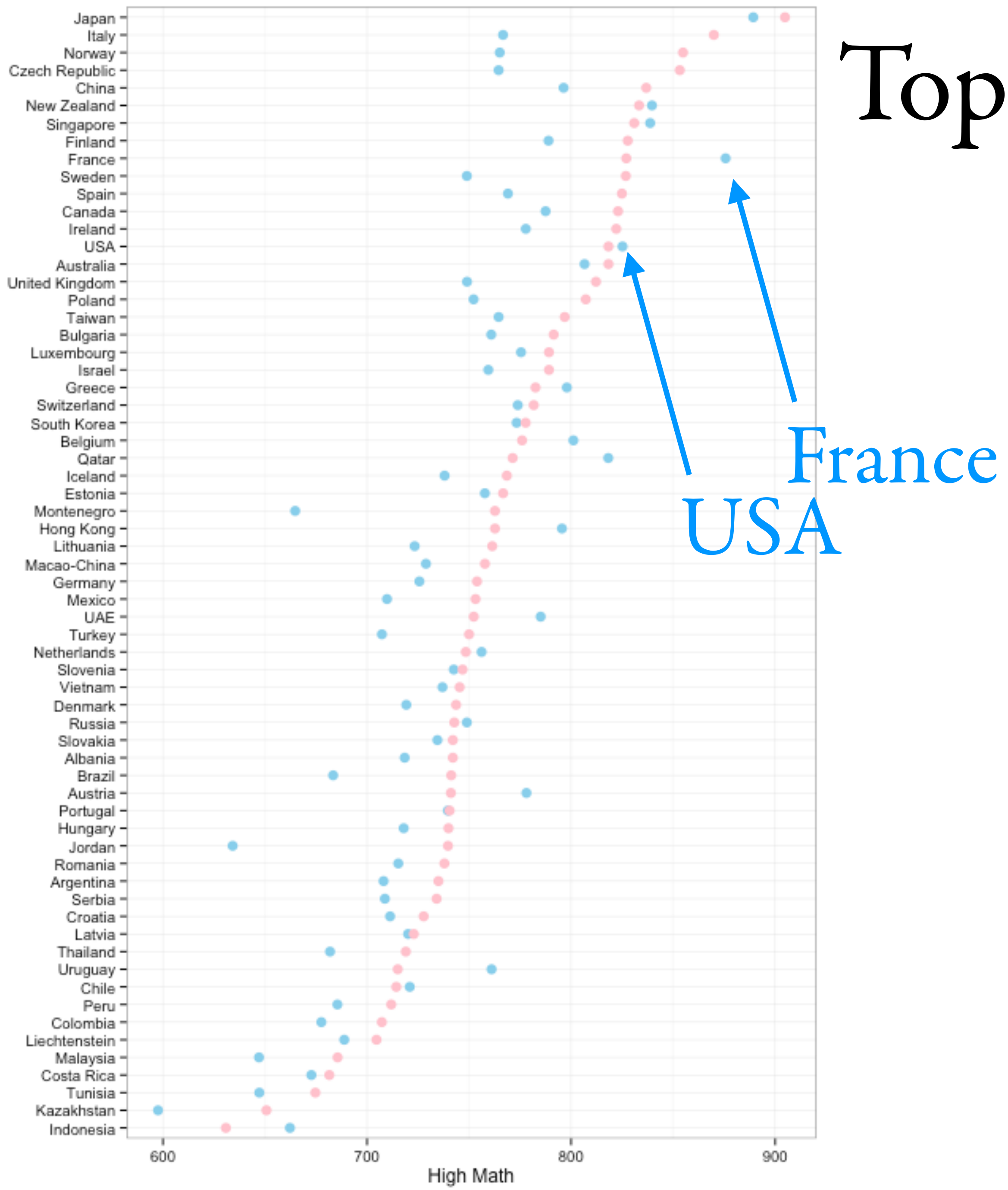


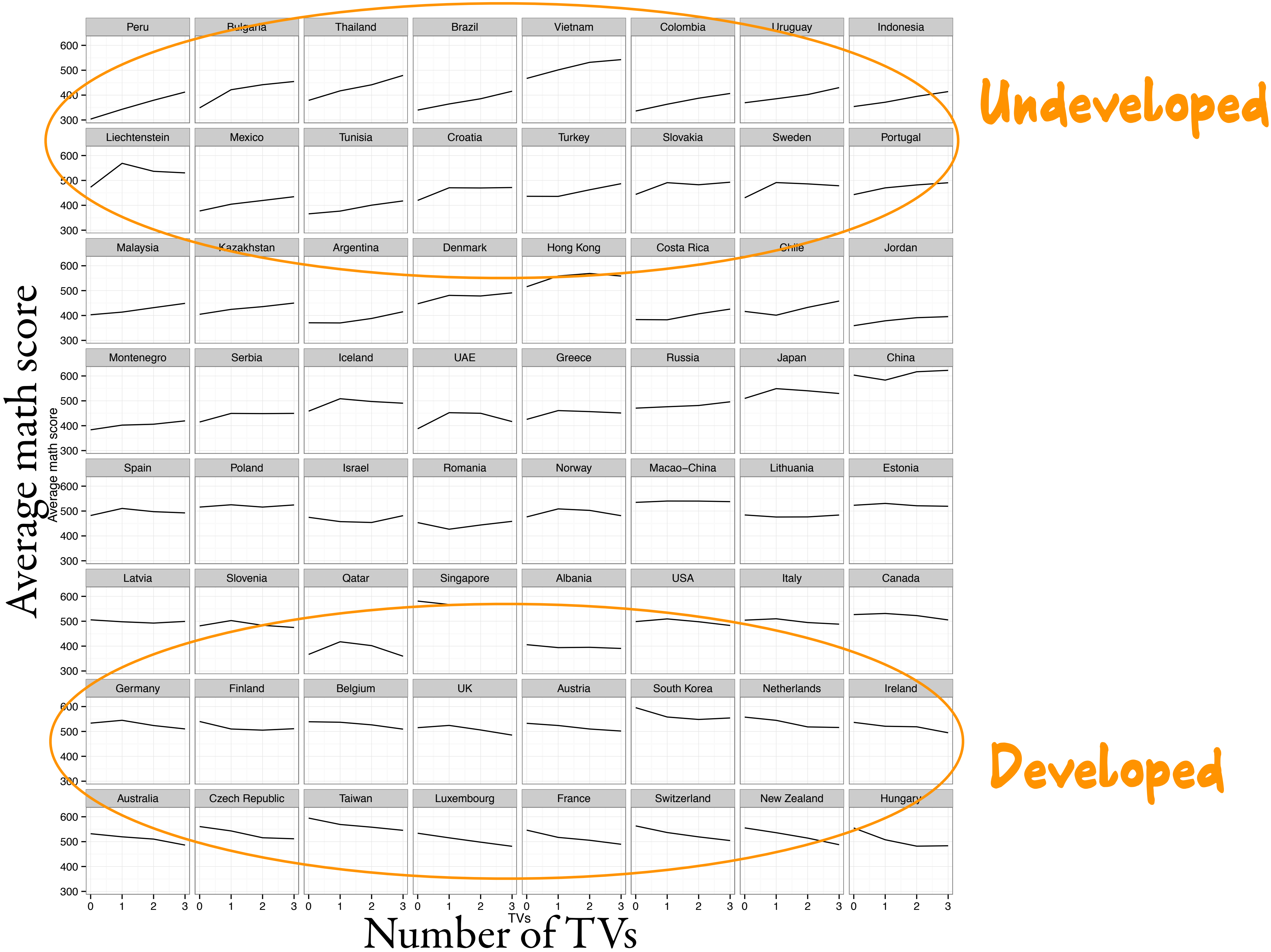
Gender & Reading

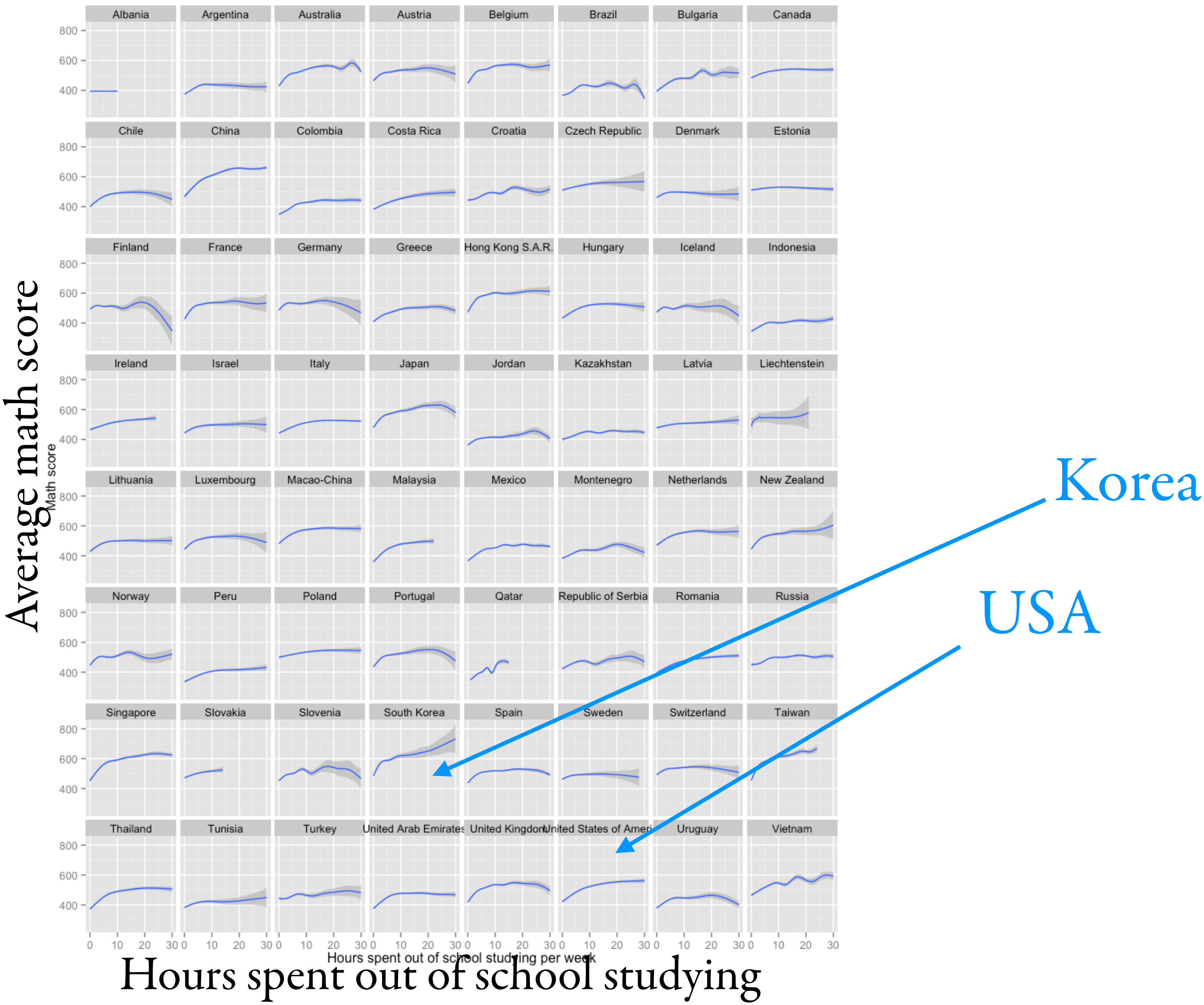
Top individual scores



Top individual scores



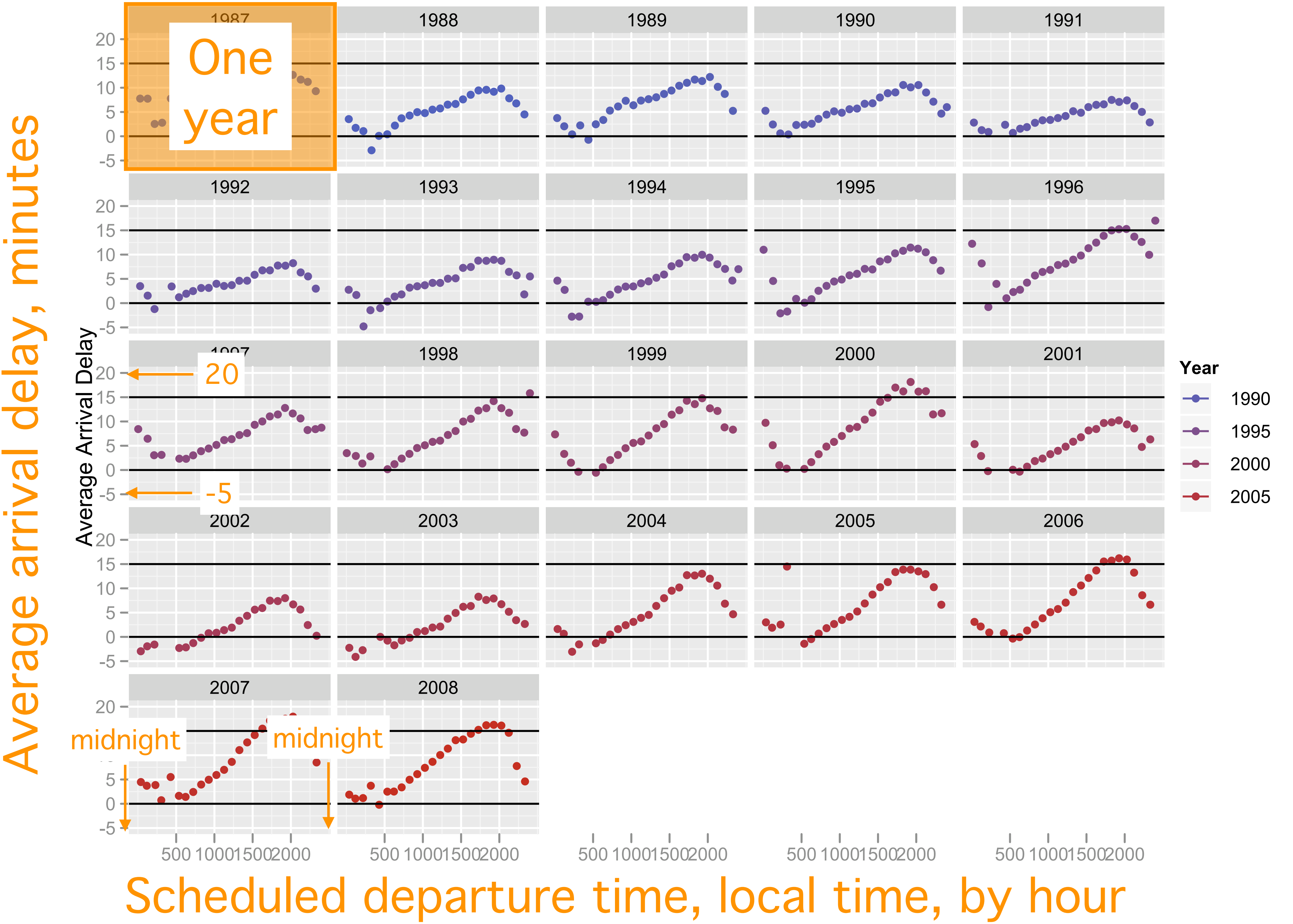

Undeveloped
Developed
Average math score
Number of TVs
Peru
Bulgaria
Thailand
Brazil
Vietnam
Colombia
Uruguay
Indonesia
Liechtenstein
Mexico
Tunisia
Croatia
Turkey
Slovakia
Sweden
Portugal
Malaysia
Kazakhstan
Argentina
Denmark
Hong Kong
Costa Rica
Chile
Jordan
Montenegro
Serbia
Iceland
UAE
Greece
Russia
Japan
China
Spain
Poland
Israel
Romania
Norway
Macao–China
Lithuania
Estonia
Latvia
Slovenia
Qatar
Singapore
Albania
USA
Italy
Canada
Germany
Finland
Belgium
UK
Austria
South Korea
Netherlands
Ireland
Australia
Czech Republic
Taiwan
Luxembourg
France
Switzerland
New Zealand
Hungary
Average math score
TVs



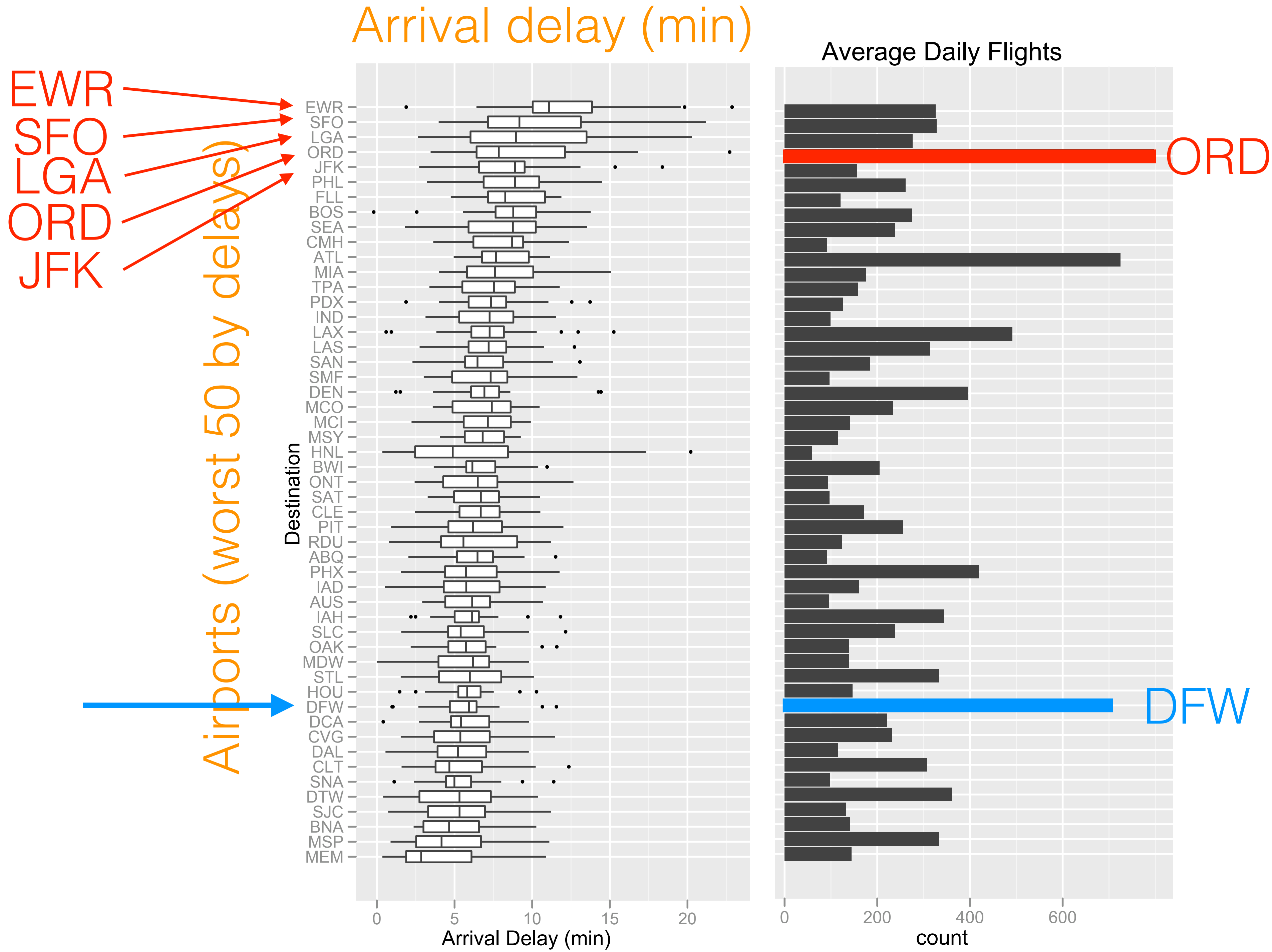
Albania
Argentina
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Colombia
Costa Rica
Croatia
Czech Republic
Denmark
Estonia
Finland
France
Germany
Greece
Hong Kong S.A.R.
Hungary
Iceland
Indonesia
Ireland
Israel
Italy
Japan
Jordan
Kazakhstan
Latvia
Liechtenstein
Lithuania
Luxembourg
Macao-China
Malaysia
Mexico
Montenegro
Netherlands
New Zealand
Norway
Peru
Poland
Portugal
Qatar
Republic of Serbia
Romania
Russia
Singapore
Slovakia
Slovenia
South Korea
Spain
Sweden
Switzerland
Taiwan
Thailand
Tunisia
Turkey
United Arab Emirates
United Kingdom
United States of Ameri
Uruguay
Vietnam
Math score
Hours spent out of school studying per week
800
600
400
0 10 20 30
Average math score
Hours spent out of school studying
Korea
USA

PISA tests

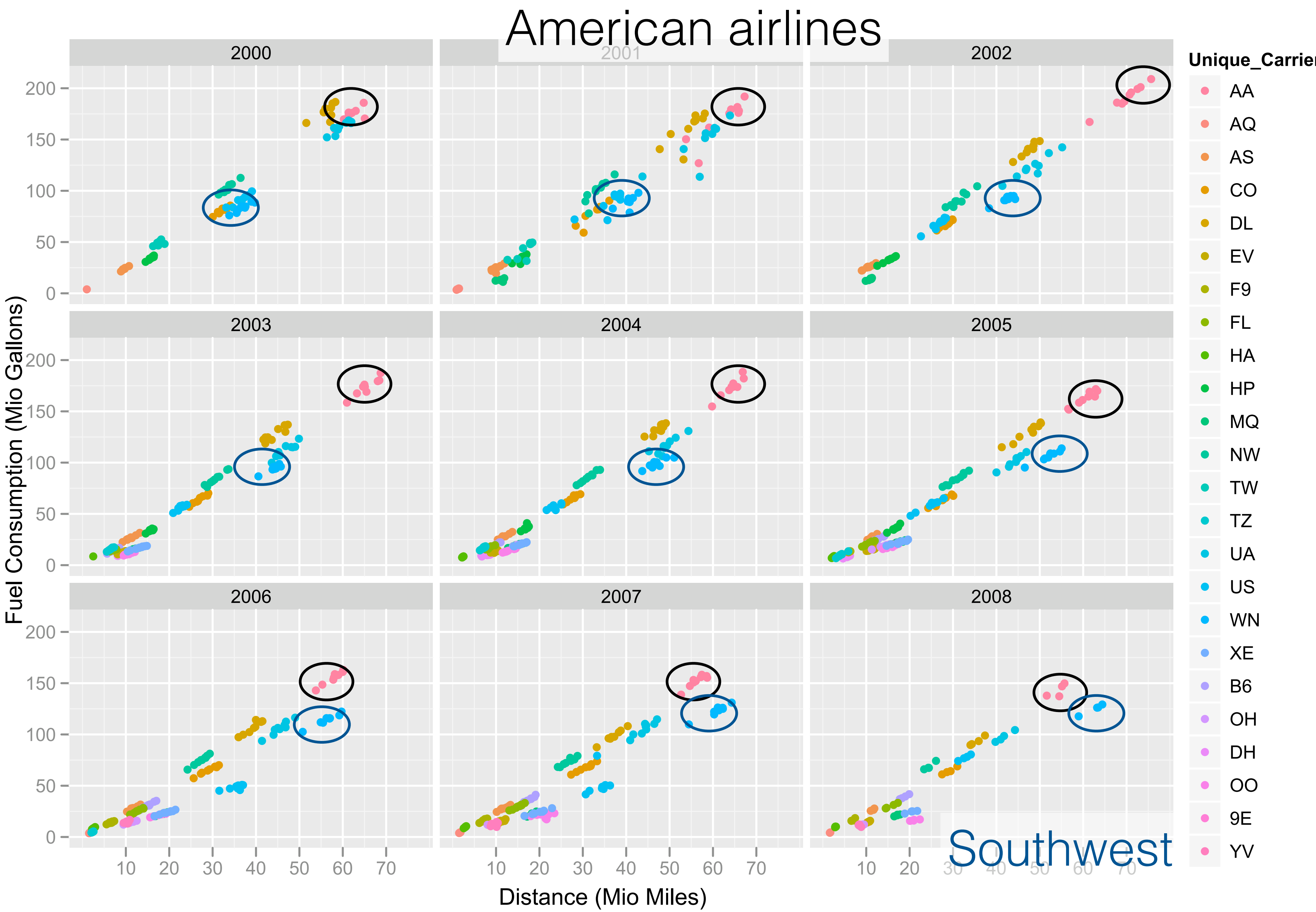
1. The gender gap in math is not universal but the reading gap is, in favor of girls.
2. Time spent out of school studying is important, but only up to a point.
3. On average, more TVs yield higher math scores in the developing world, but lower in the developed countries.
4. Albania is different!



One year
1988
1989
1990
1991
1992
1993
1994
1995
1996
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
20
-5
midnight
midnight
Average Arrival Delay
Average arrival delay, minutes
Year
1990
1995
2000
2005
500 1000 1500 2000
Scheduled departure time, local time, by hour



Arrival delay (min)
Average Daily Flights
EWR
SFO
LGA
ORD
JFK
Airports (worst 50 by delays)
Destination
EWR
SFO
LGA
ORD
JFK
PHL
FLL
BOS
SEA
CMH
ATL
MIA
TPA
PDX
IND
LAX
LAS
SAN
SMF
DEN
MCO
MCI
MSY
HNL
BWI
ONT
SAT
CLE
PIT
RDU
ABQ
PHX
IAD
AUS
IAH
SLC
OAK
MDW
STL
HOU
DFW
DCA
CVG
DAL
CLT
SNA
DTW
SJC
BNA
MSP
MEM
0
5
10
15
20
Arrival Delay (min)
0
200
400
600
count
ORD
DFW

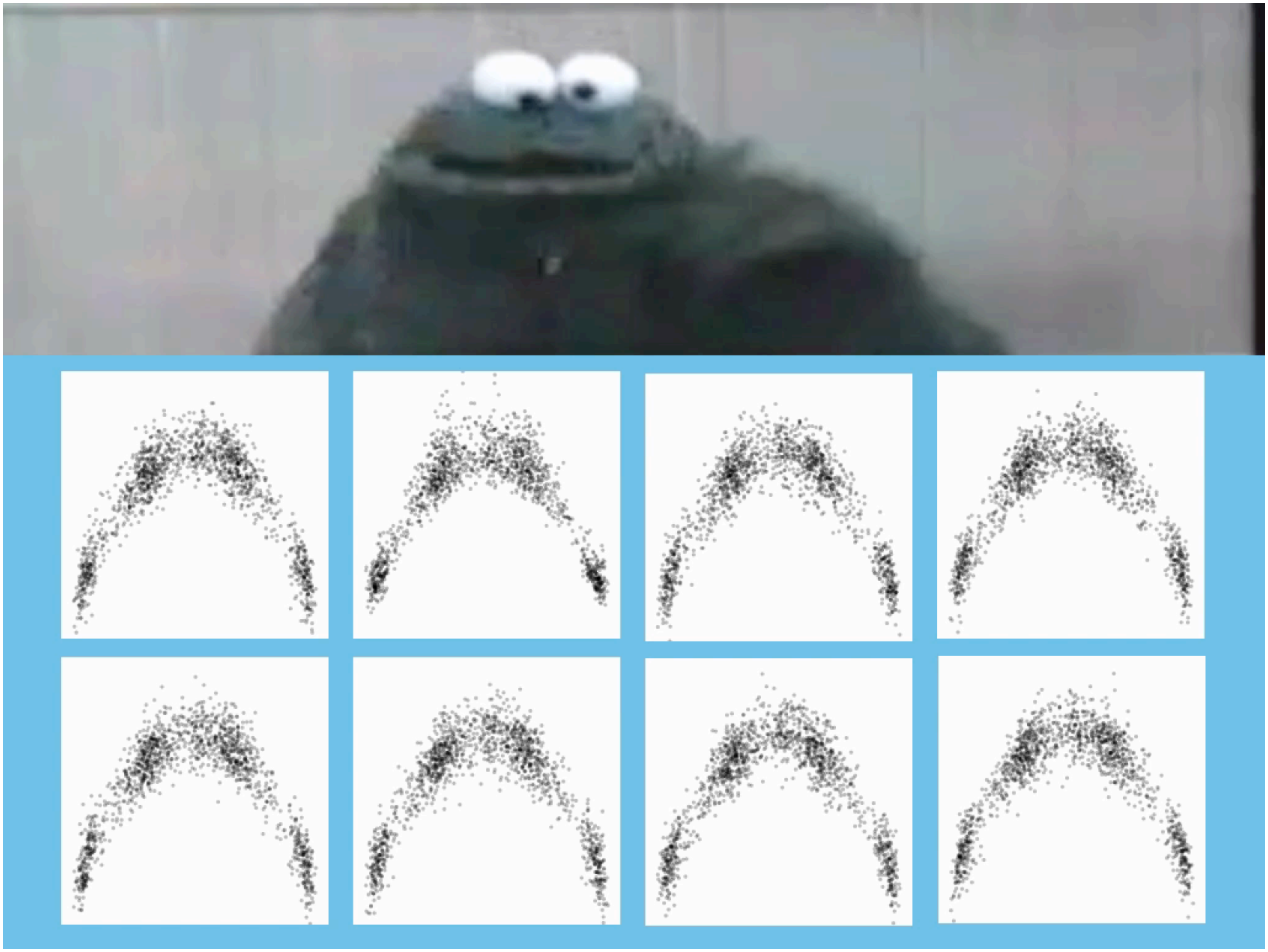

American airlines
2000
2001
2002
2003
2004
2005
2006
2007
2008
Fuel Consumption (Mio Gallons)
Distance (Mio Miles)
0
50
100
150
200
10
20
30
40
50
60
70
Unique_Carrier
AA
AQ
AS
CO
DL
EV
F9
FL
HA
HP
MQ
NW
TW
TZ
UA
US
WN
XE
B6
OH
DH
OO
9E
YV
Southwest

US Air Traffic

1. Fly early in the day
2. Avoid EWR, JFK, ORD, ...
3. American Airlines was in trouble, filed for bankruptcy Nov 29, 2013

Statisticians discover, explore, and also can be skeptical

Exploratory and inferential ARE NOT mutually exclusive.

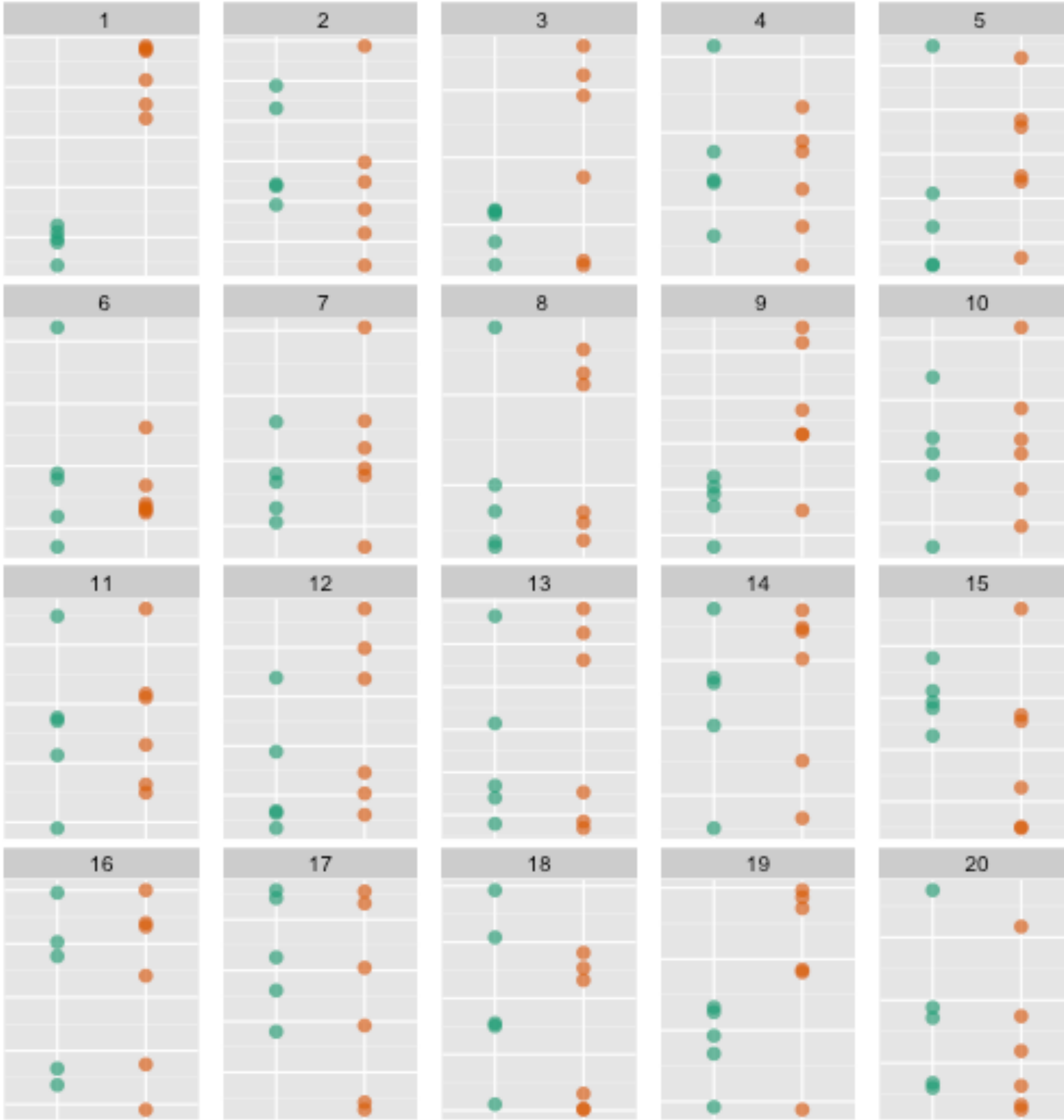


Video courtesy of Hadley Wickham

http://bit.ly/visinf-cornell

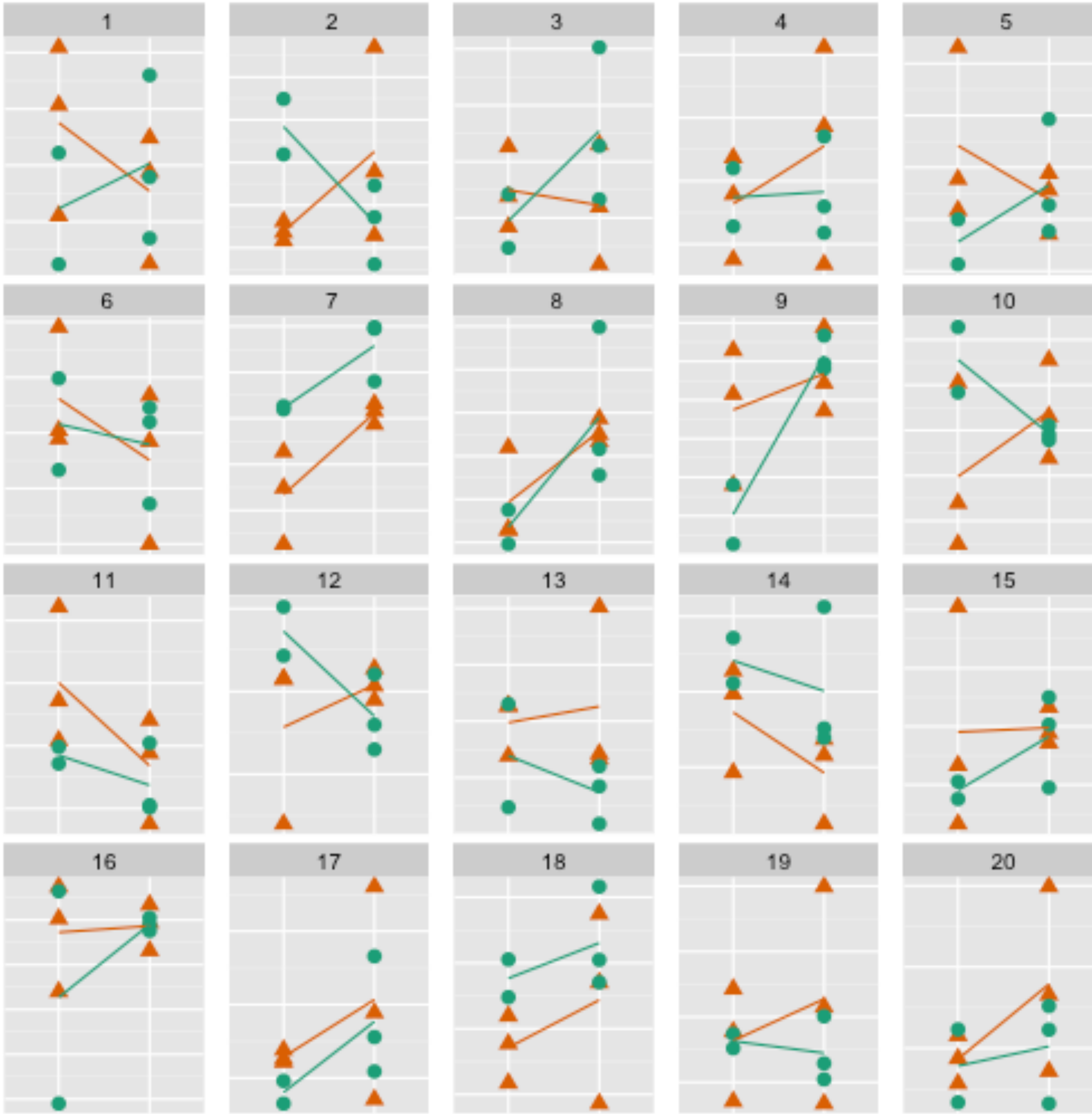
(https://visnut.wufoo.com/forms/cornell-university-seminar/)

1

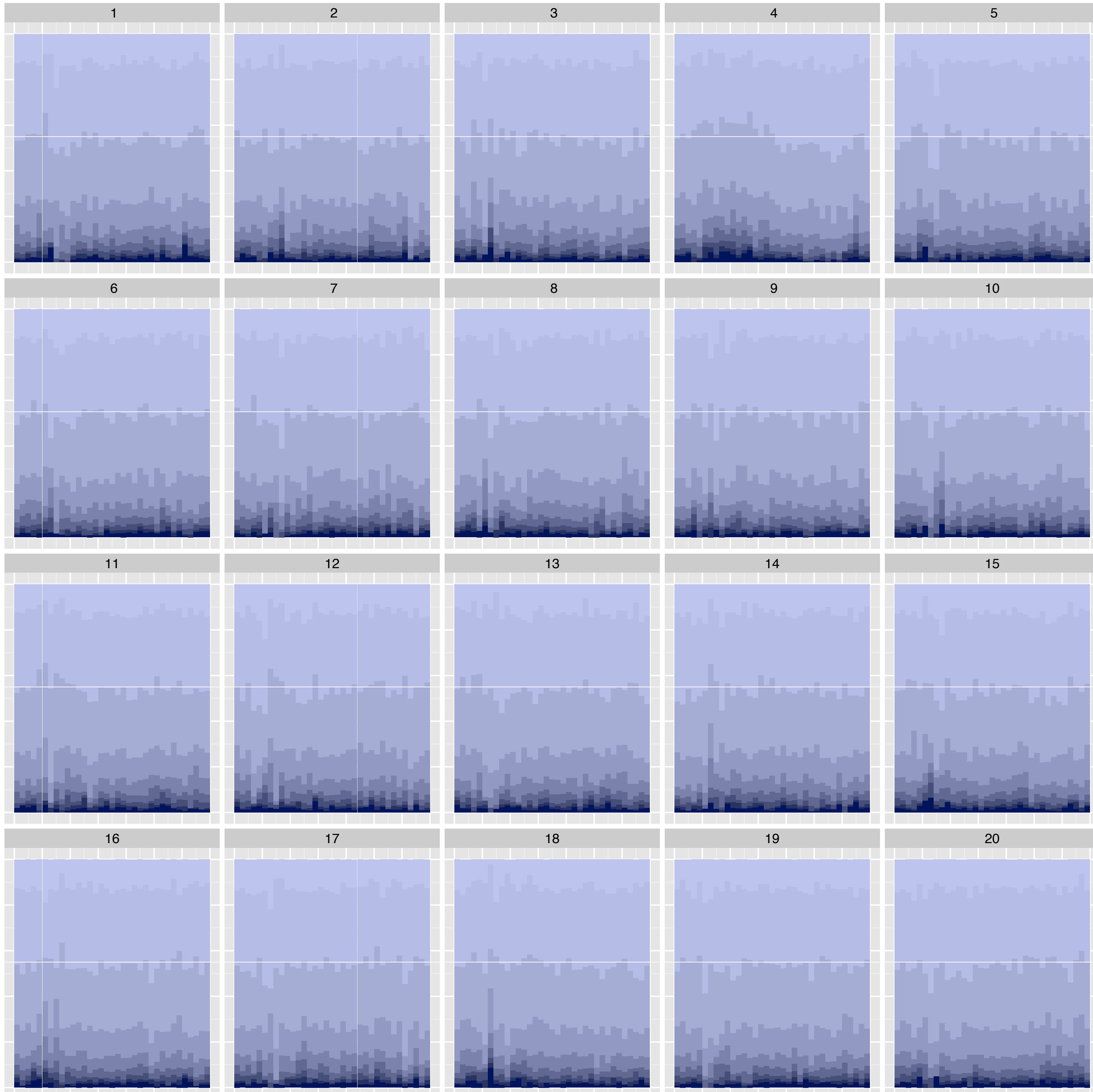


2

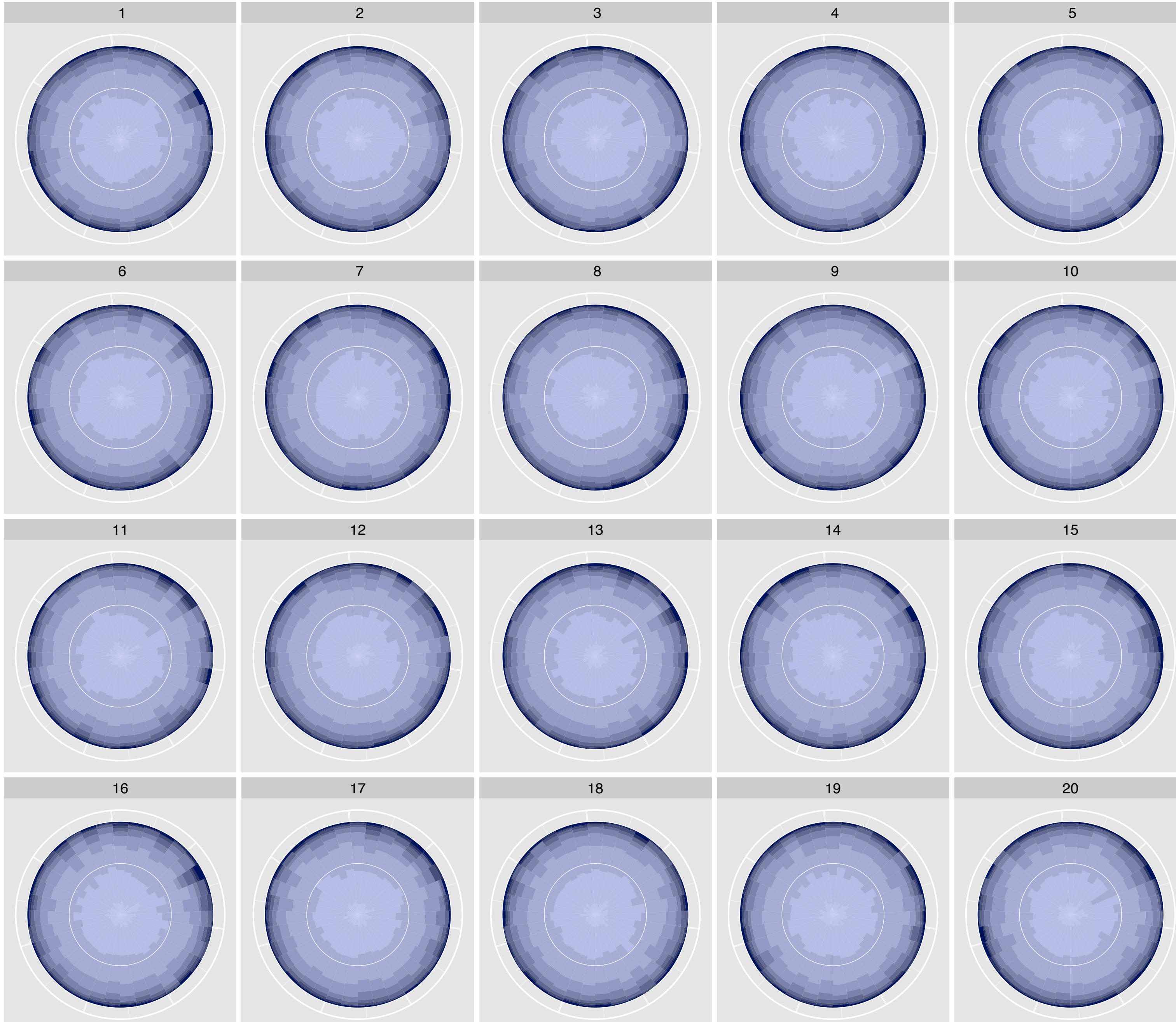
In which of these plots is the green line the steepest, and the spread of the green points relatively small?



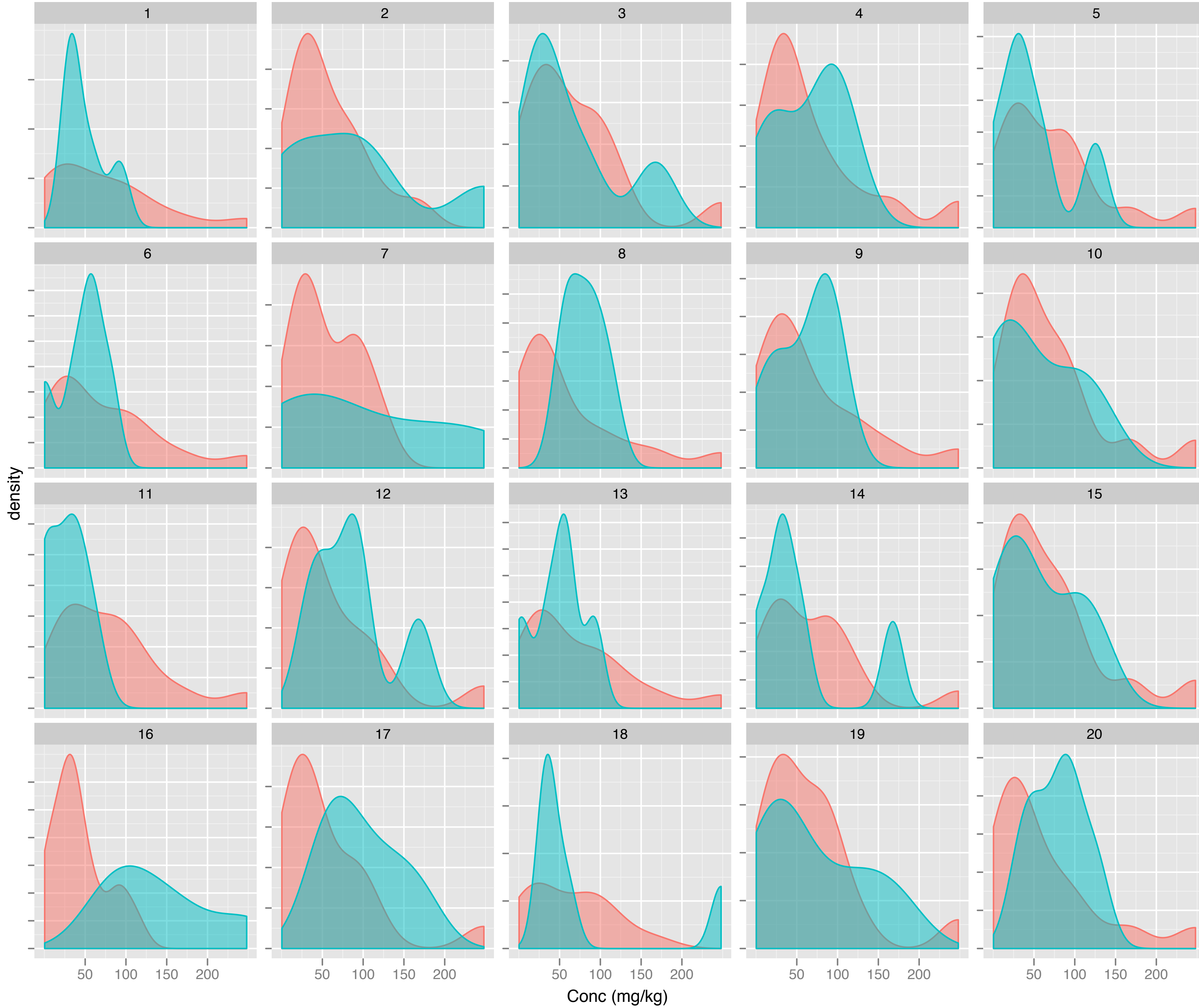
3



4

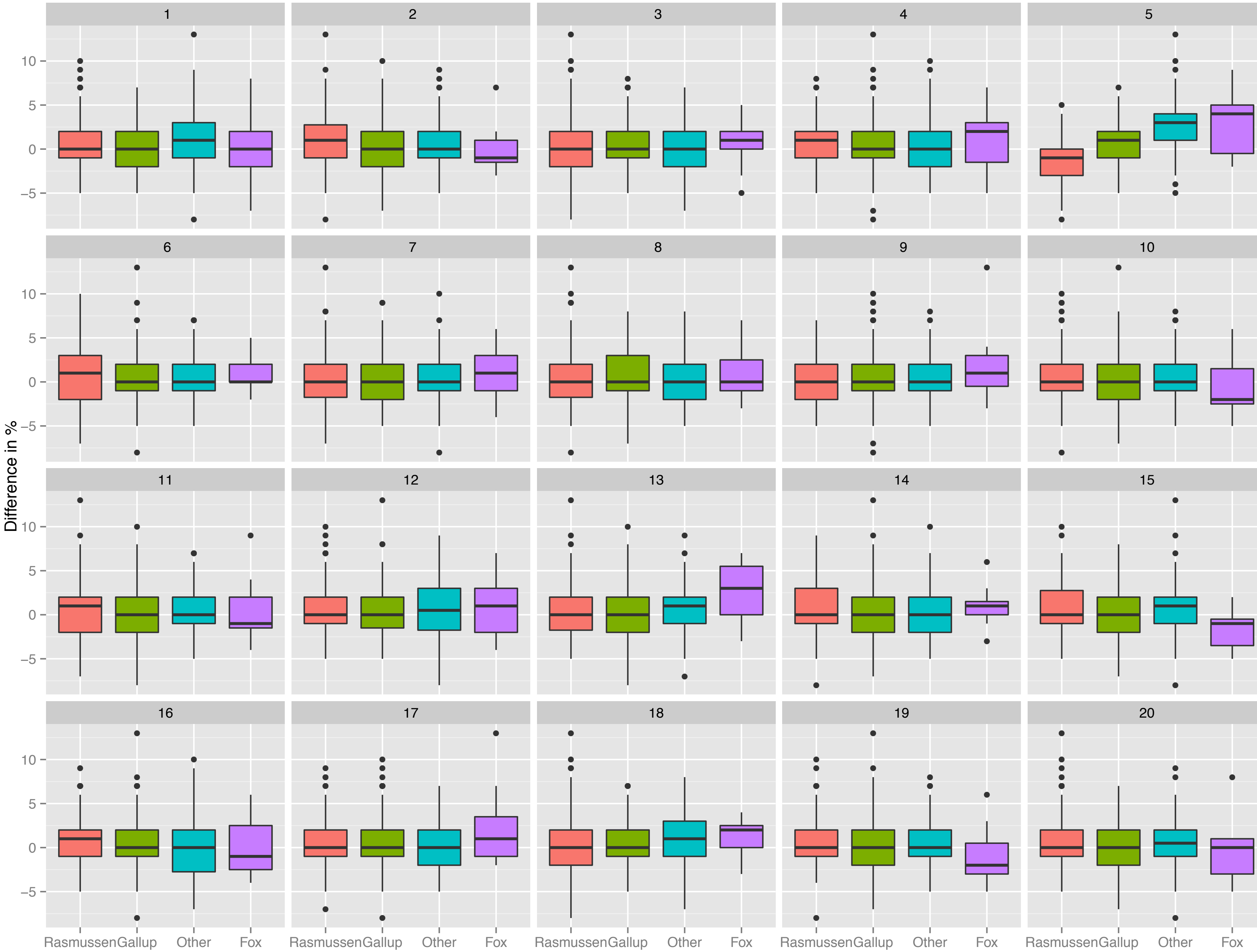


5



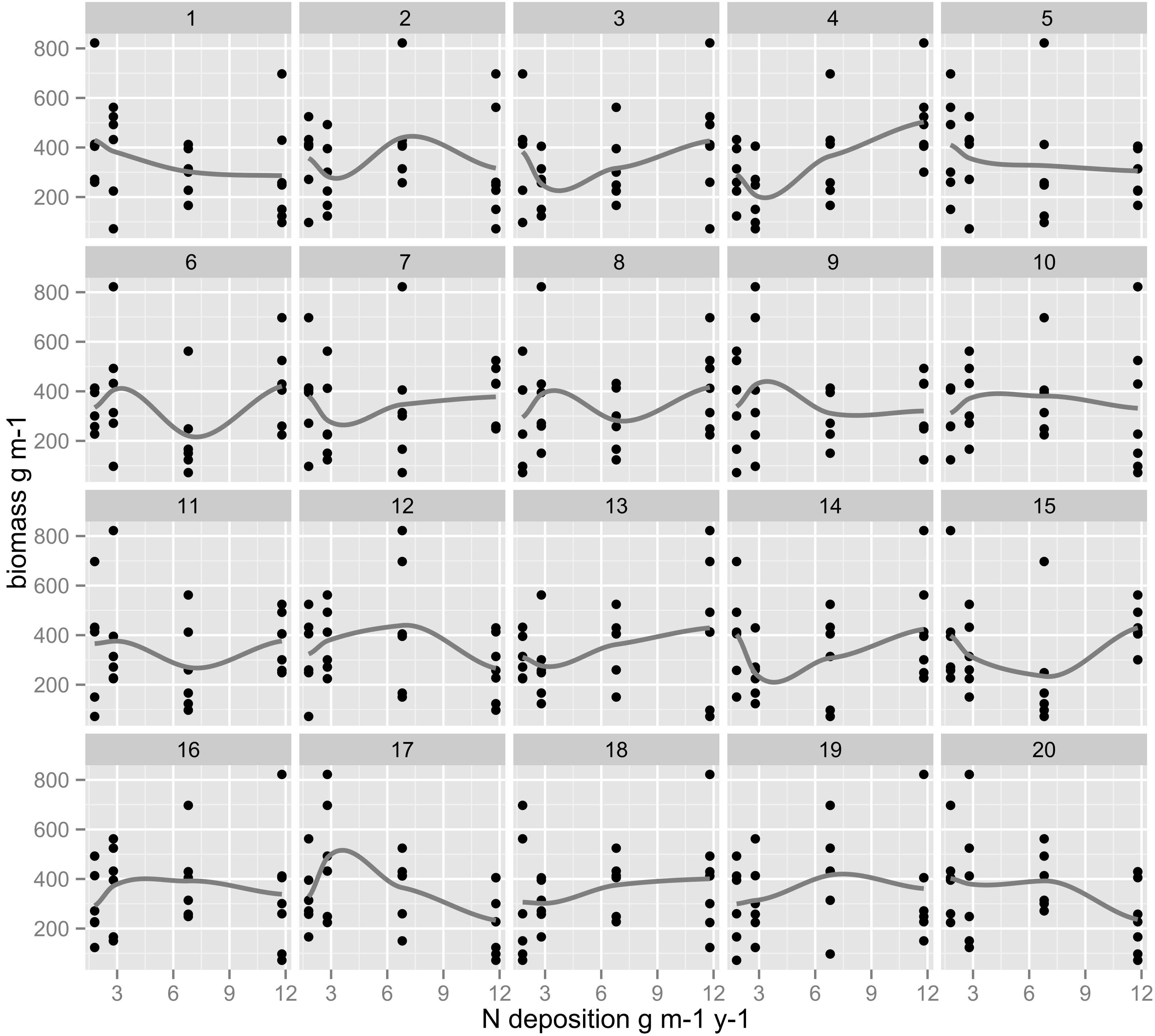
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
density
50 100 150 200
Conc (mg/kg)

6



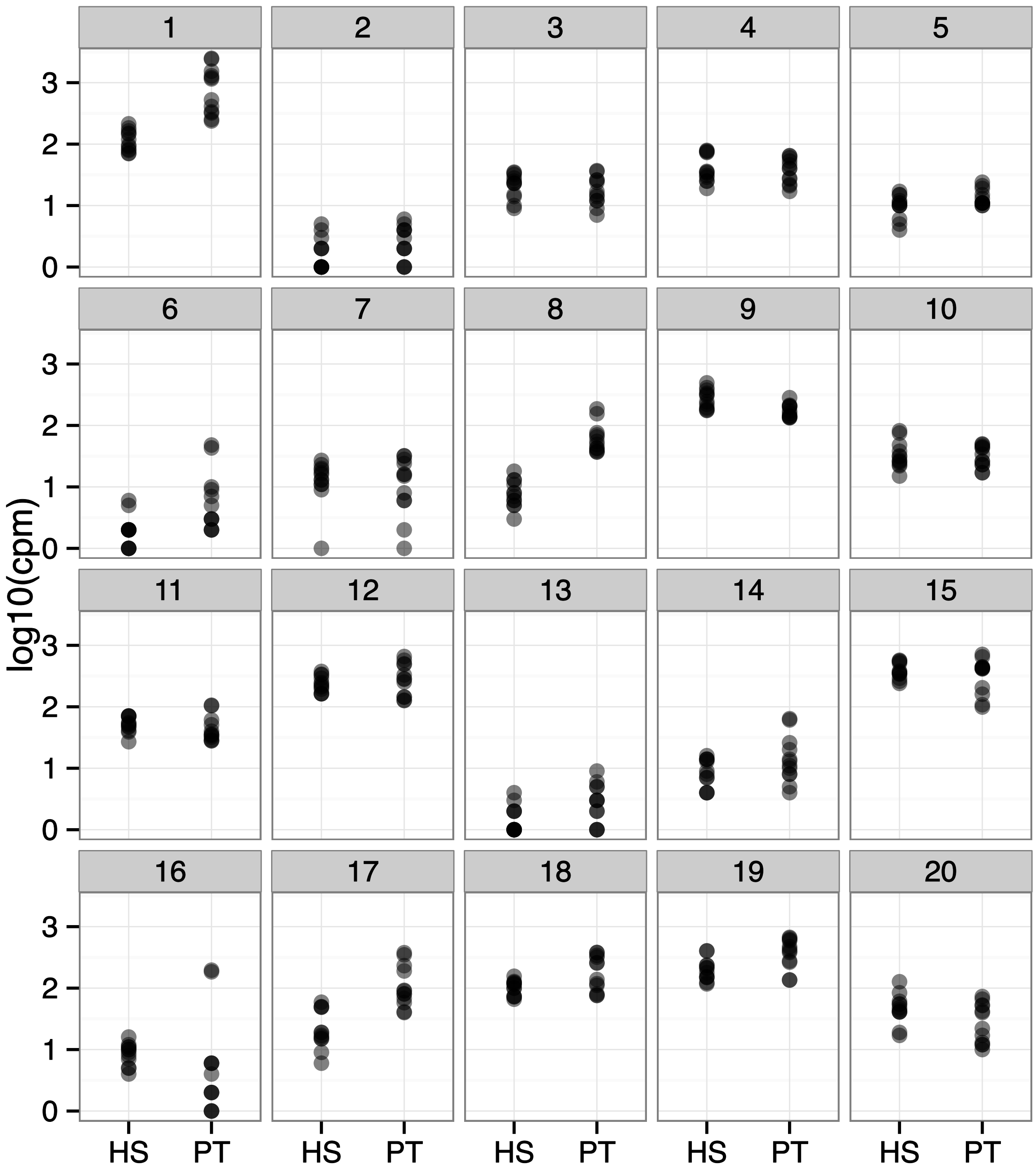
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
10
5
0
−5
Difference in %
Rasmussen
Gallup
Other
Fox

7

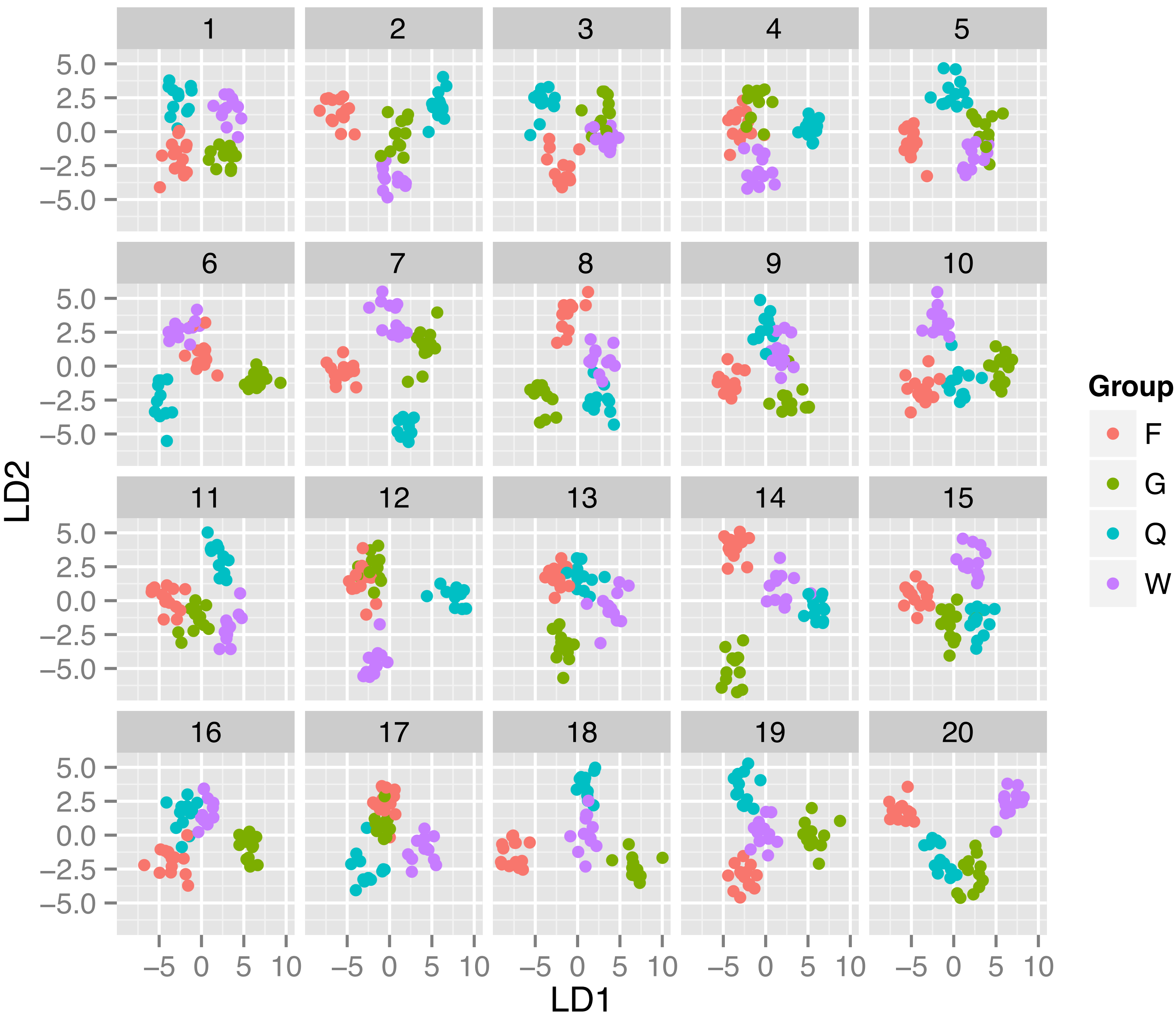


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
800
600
400
200
3
6
9
12
biomass g m-1
N deposition g m-1 y-1

8



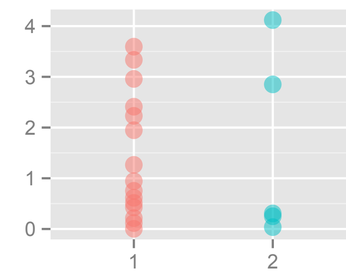
9


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
5.0
2.5
0.0
−2.5
−5.0
−5 0 5 10
LD2
LD1
Group
F
G
Q
W

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Protocols

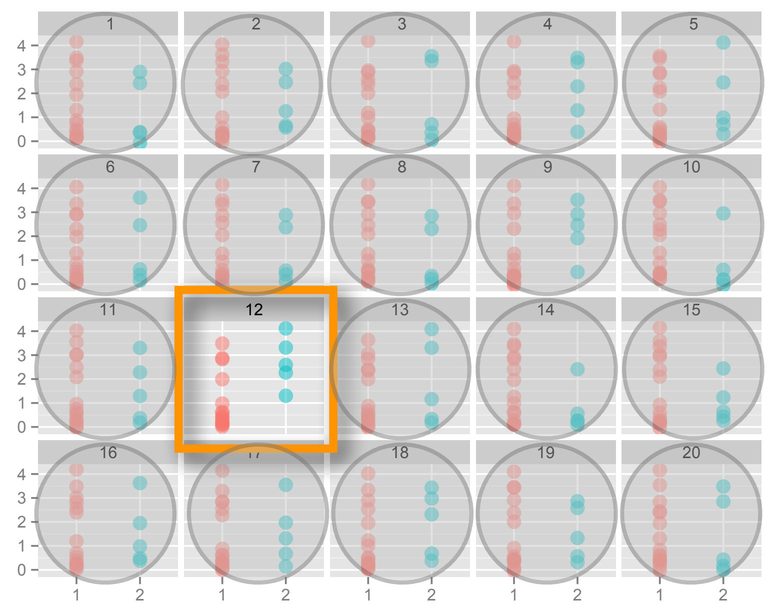
- Rorschach: Show many pictures of data with “nothing” happening, pictures from a null distribution



- Lineup: Embed the plot of the data among plots of data generated from the null distribution

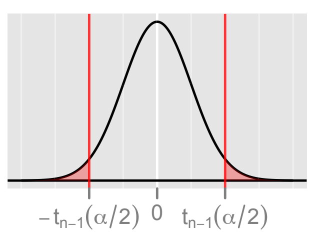
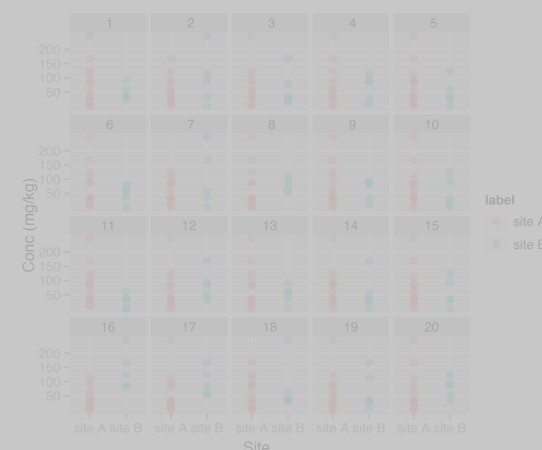
Data plot

Null plots



Source: Buja et al (2009) RSPT(A)

Hypothesis testing

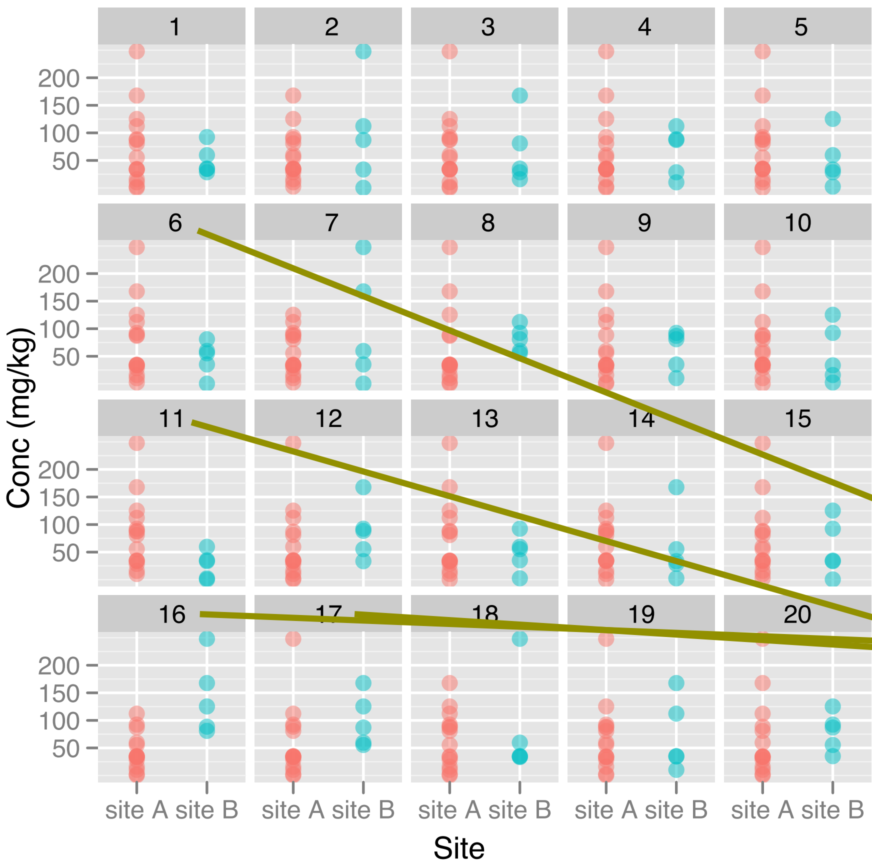
	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	$\downarrow$	$\downarrow$
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T(y) =$ 
	$\downarrow$	$\downarrow$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

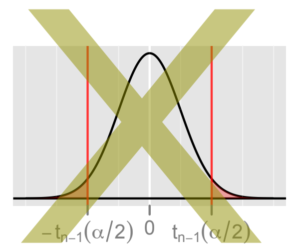
Concepts

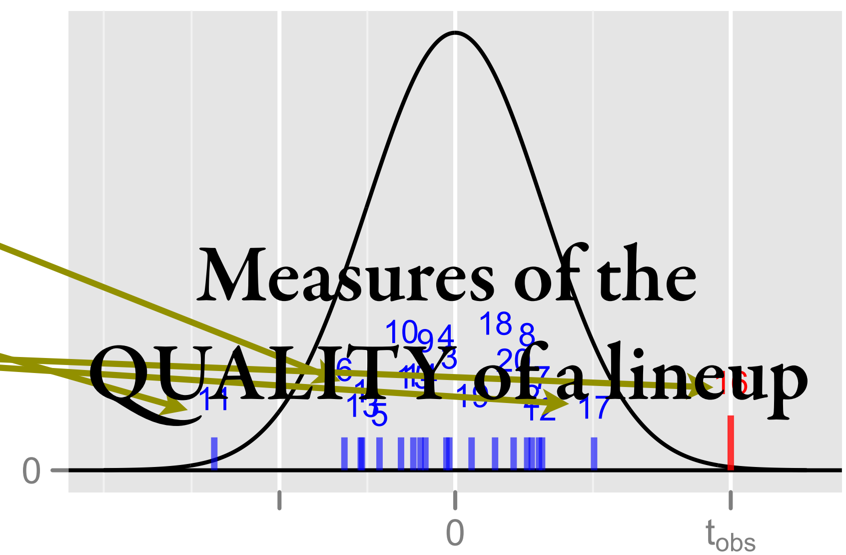
- Plot of data is a test statistic
- Type of plot used typically indicates null/alternative hypothesis, eg scatterplot suggests null hypothesis "no association between two variables"
- Null hypothesis suggests null generating mechanism
- Generate draws from the null, plot, show uninvolved observer
- Data plot detected equivalent to rejection of null, it is extreme relative to the sampling distribution

Finite comparisons

Sampling distribution comparison is against a finite set



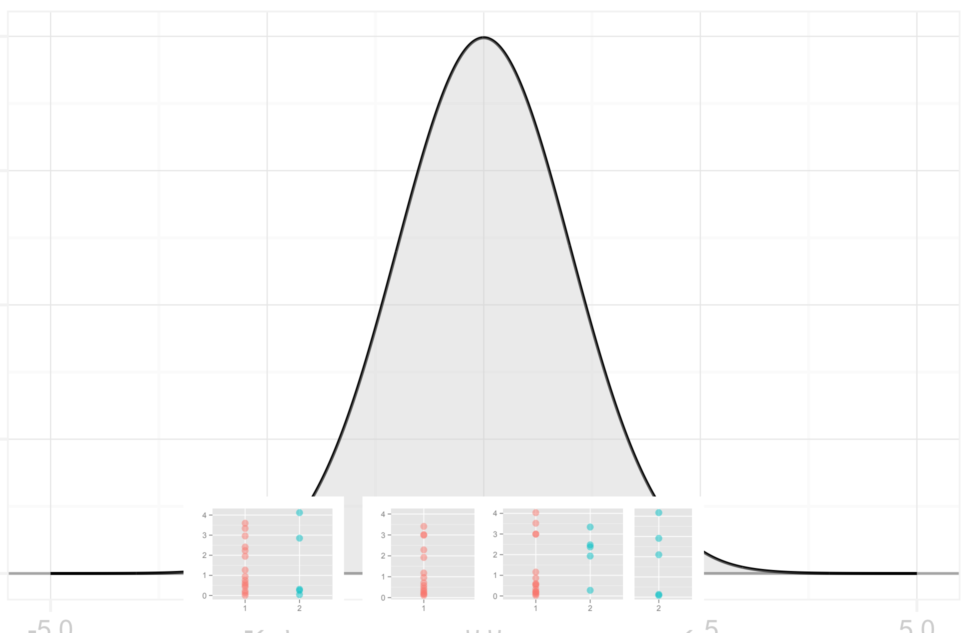




Source: Roy Chowdhury et al (2014)

Null distribution unknown

KEEP IN MIND: In practice, graphics is being used when there is no quantification of a sampling distribution. All we have is (m-1) representatives from whatever that distribution is.



Significance

- What is the *p*-value?
- 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

Source: Majumder et al (2013) JASA

Let's check

- For each of the lineups shown earlier, let's calculate the *p*-value
- Measures the significance of the structure in the data plot

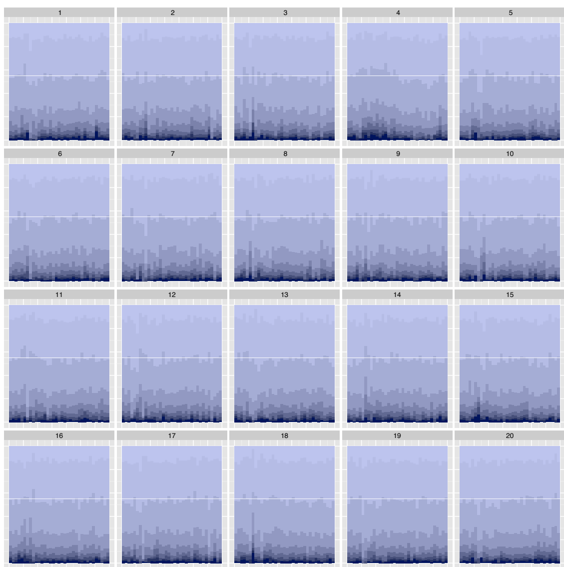
Power

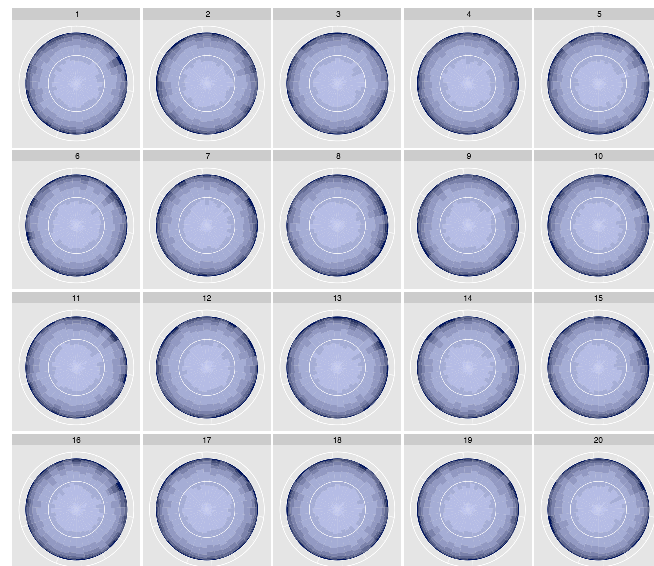
- What is the power of the test?
- There is always a choice of type of plot and graphical elements to use. Some will work better than others. This is analogous to the power of a statistical test.
- Signal strength will be defined as "proportion of observers who identify the data plot".
- Enables the comparison of different plot designs.
- Signal strength equals power, when only plot design changes.

Source: Hofmann et al (2012) InfoVis

Comparing plot designs

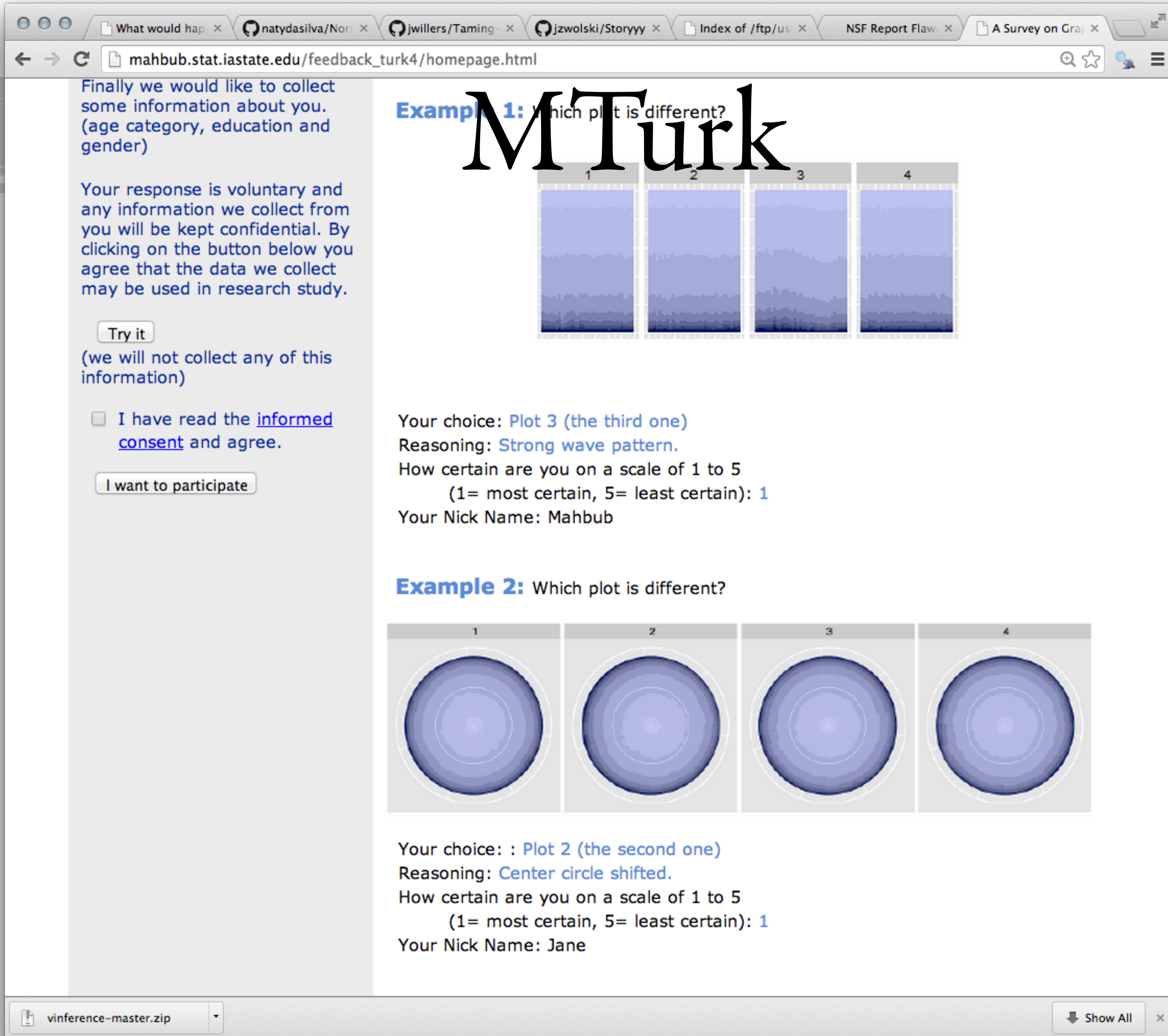
All flights in and out of Seattle/Tacoma International Airport (SEA) between July 2008 and June 2011. How does wind direction (and strength) affect airline traffic? Euclidean or polar coordinates?





Signal strength: Proportion of people who selected the actual data plot.

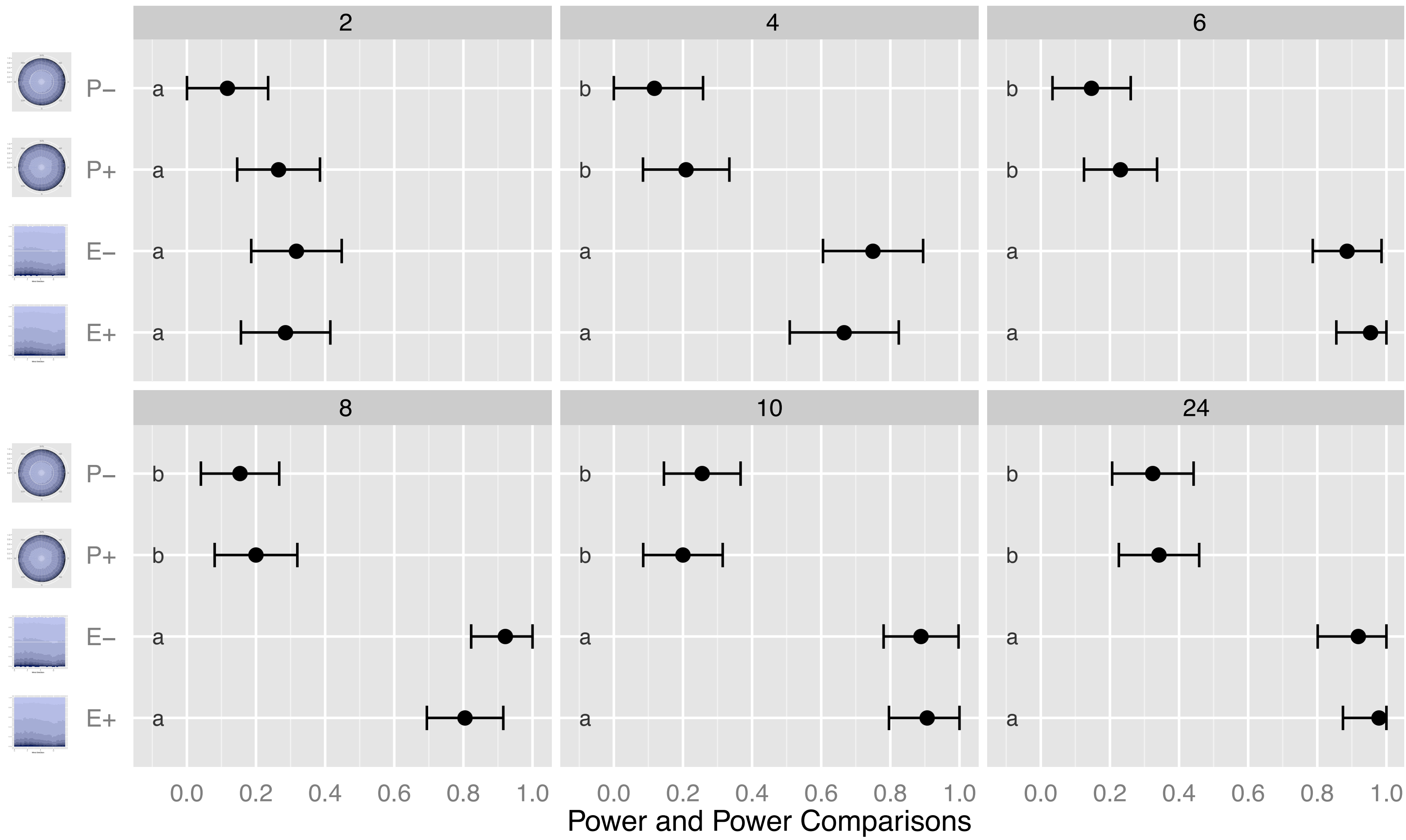
MTurk



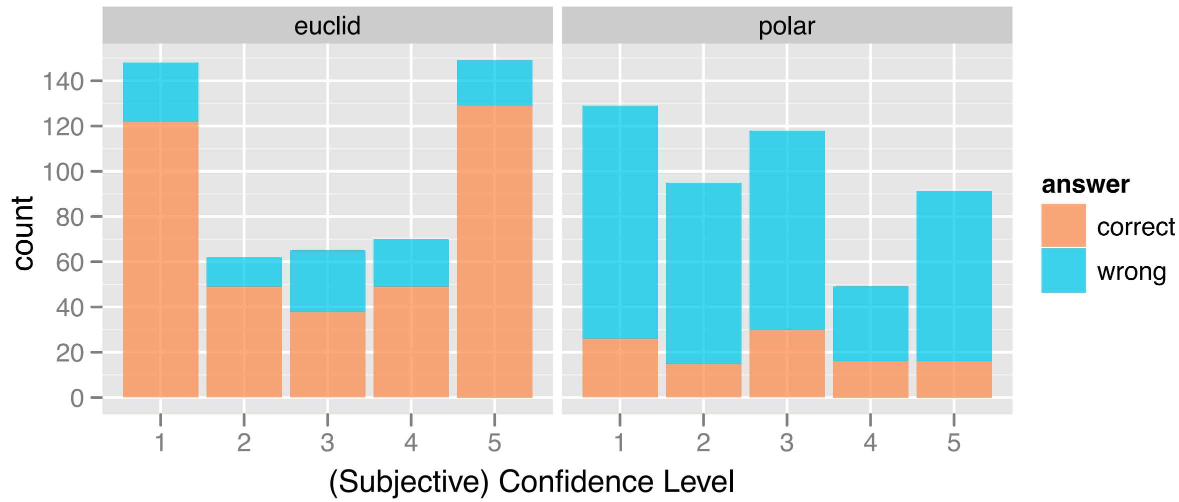
mahbub.stat.iastate.edu/feedback_turk4/homepage.html
Finally we would like to collect some information about you. (age category, education and gender)
Your response is voluntary and any information we collect from you will be kept confidential. By clicking on the button below you agree that the data we collect may be used in research study.
Try it
(we will not collect any of this information)
I have read the informed consent and agree.
I want to participate
Example 1: Which plot is different?
1
2
3
4
Your choice: Plot 3 (the third one)
Reasoning: Strong wave pattern.
How certain are you on a scale of 1 to 5
(1= most certain, 5= least certain): 1
Your Nick Name: Mahbub
Example 2: Which plot is different?
1
2
3
4
Your choice: : Plot 2 (the second one)
Reasoning: Center circle shifted.
How certain are you on a scale of 1 to 5
(1= most certain, 5= least certain): 1
Your Nick Name: Jane
vinference-master.zip
Show All

Experiment 4

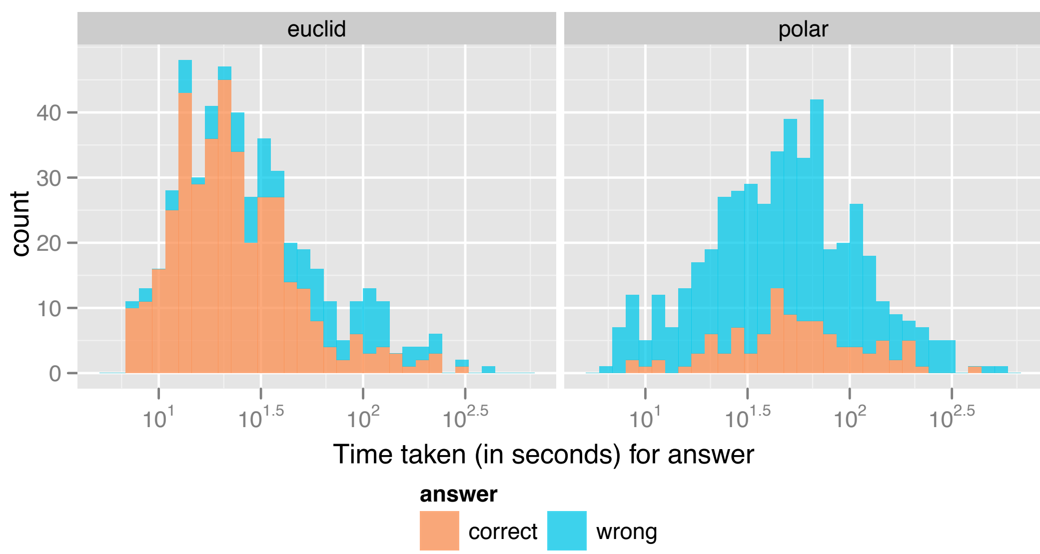
http://mahbub.stat.iastate.edu/feedback_turk4/homepage.html



Additional information

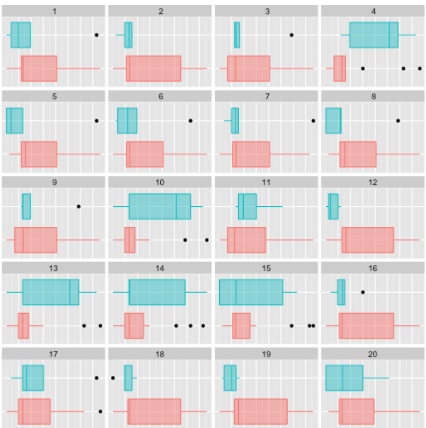
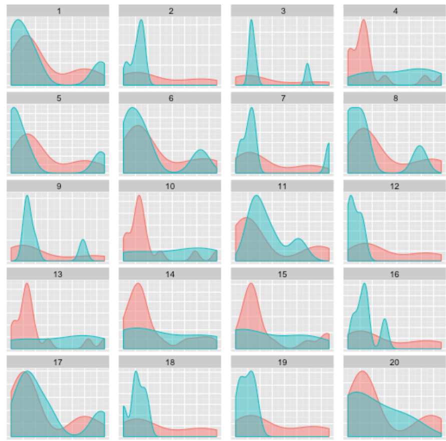
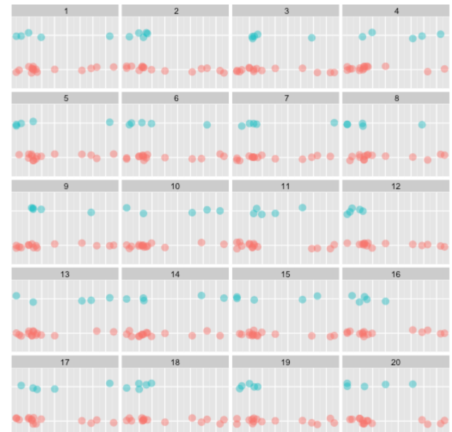


Self-reported confidence in choice



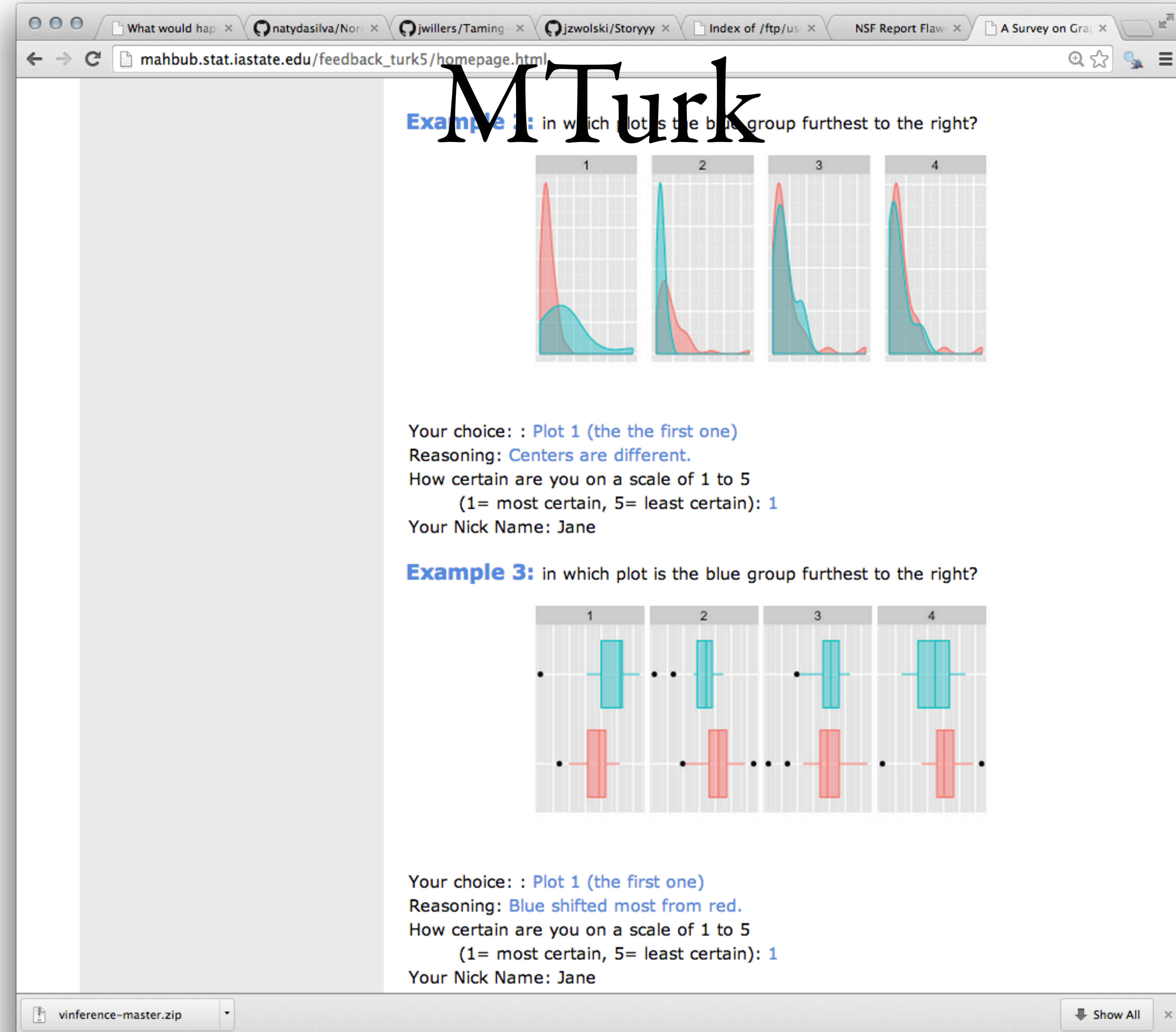
Time taken to answer

Experiment 5


Which style is best for comparing two groups?
boxplots, density, dotplots, histogram

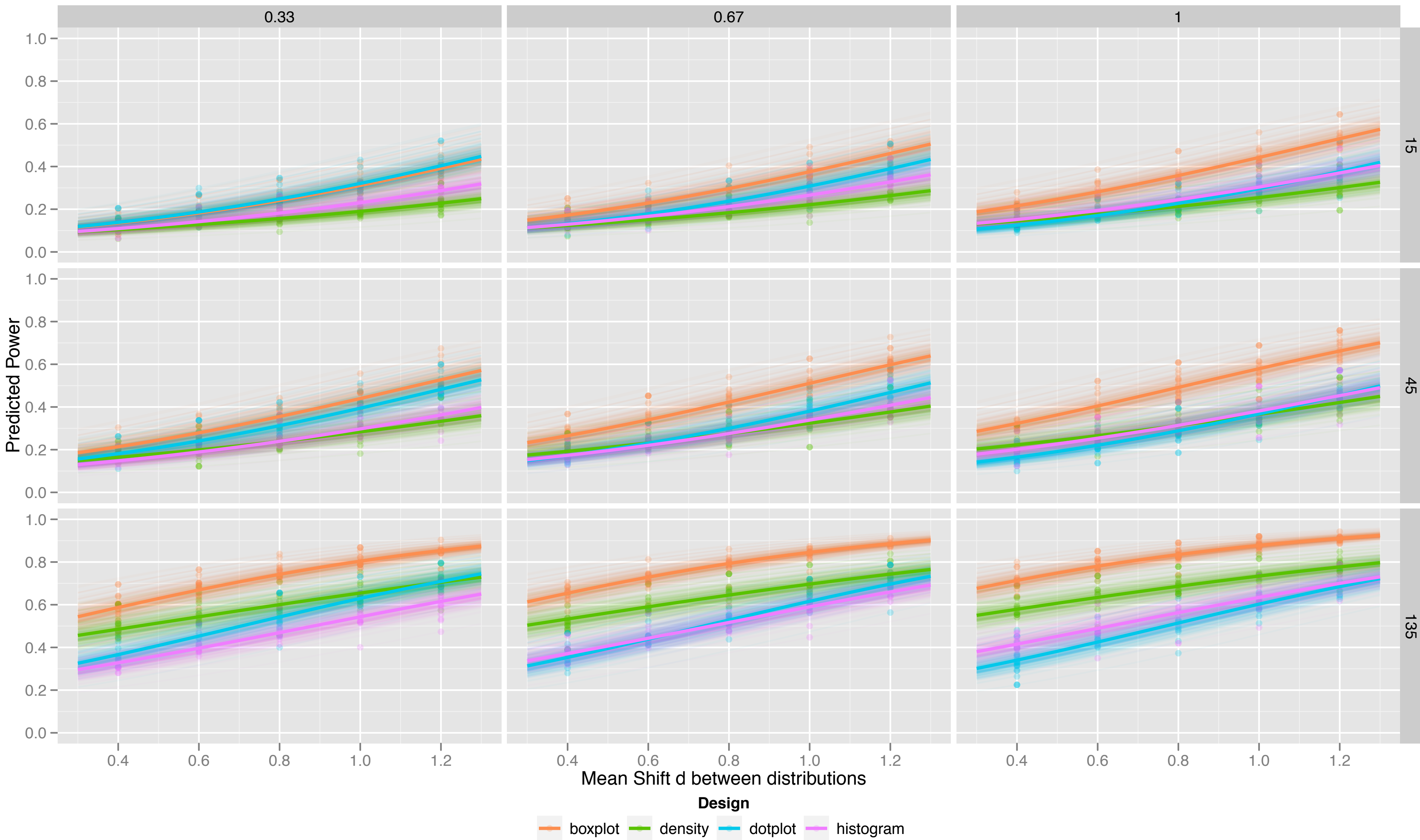
MTurk



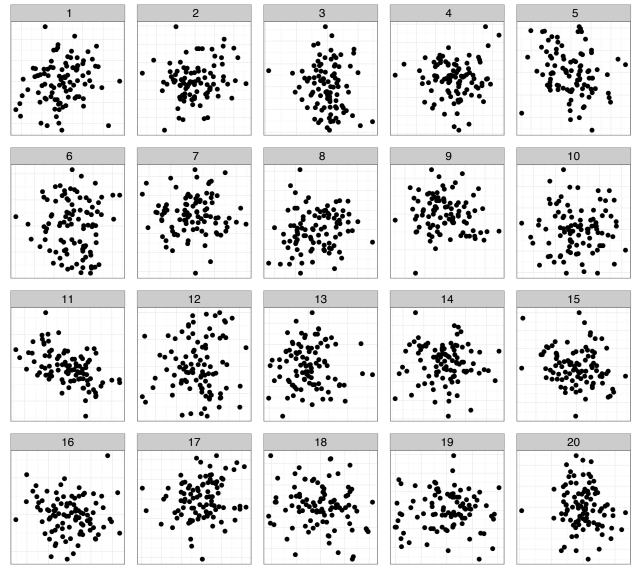
mahbub.stat.iastate.edu/feedback_turk5/homepage.html
Your choice: : Plot 1 (the the first one)
Reasoning: Centers are different.
How certain are you on a scale of 1 to 5
(1= most certain, 5= least certain): 1
Your Nick Name: Jane
Example 3: in which plot is the blue group furthest to the right?
1
2
3
4
Your choice: : Plot 1 (the first one)
Reasoning: Blue shifted most from red.
How certain are you on a scale of 1 to 5
(1= most certain, 5= least certain): 1
Your Nick Name: Jane
vinference-master.zip
Show All

Experiment 5

http://mahbub.stat.iastate.edu/feedback_turk5/homepage.html



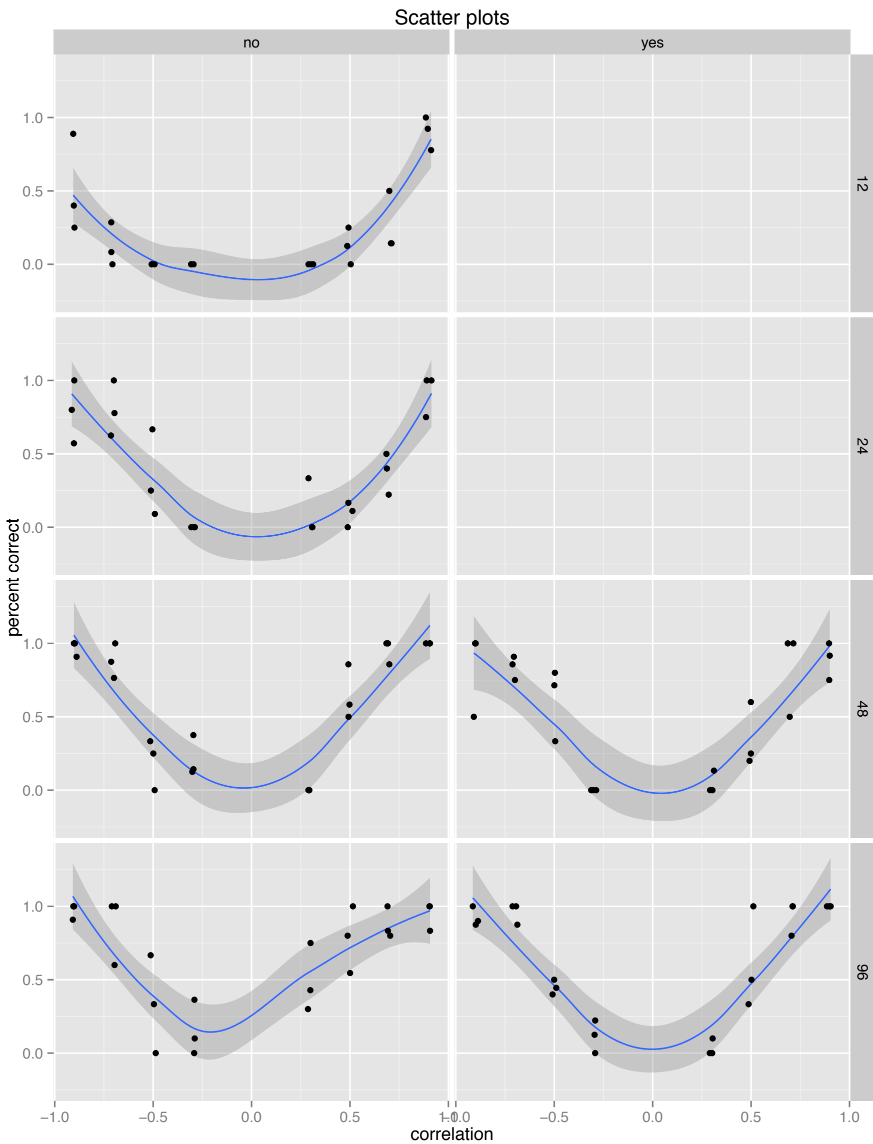
Experiment 18



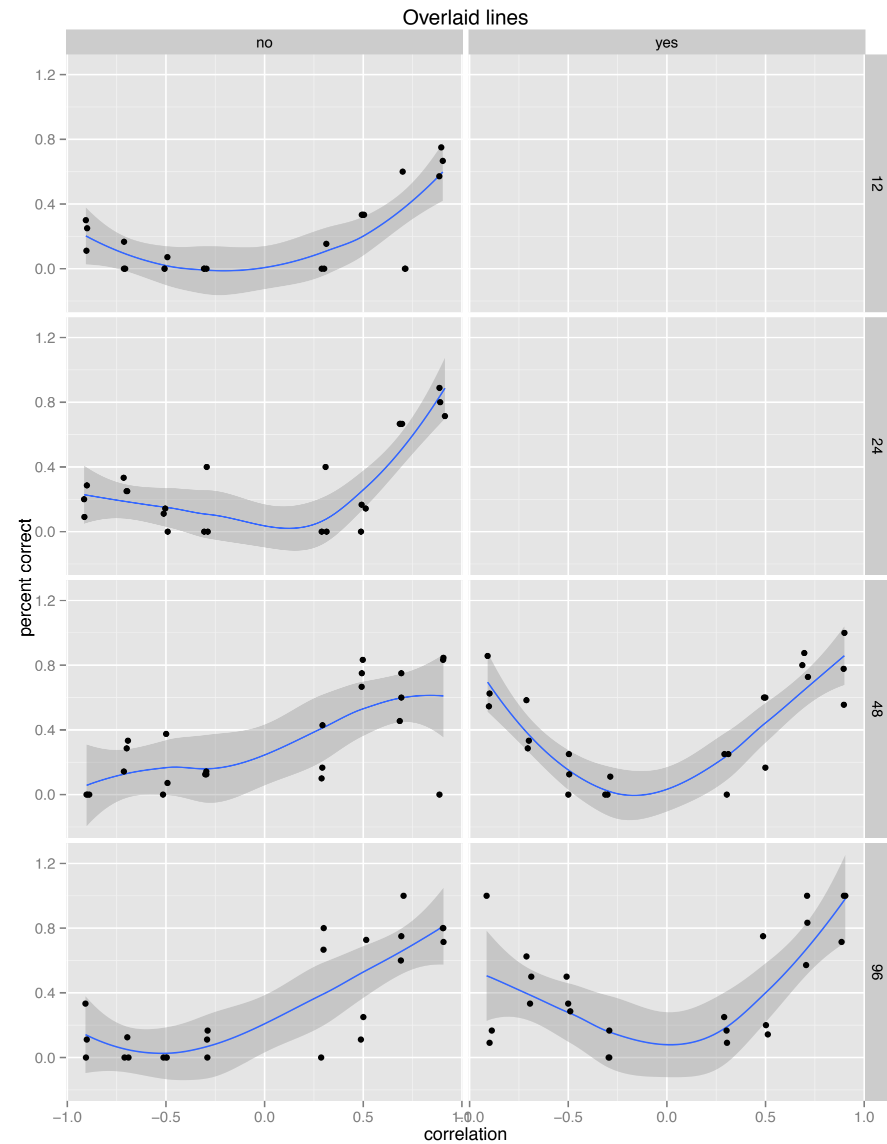
For reading association between two time series, is it better to display as overlaid line plots, or as a scatterplot?

http://104.236.245.153:8080/mahbub/turk18/index.html

Scatterplots



Line plot



Proportion selected data plot

-1 <— 0 —> 1   -1 <— 0 —> 1   -1 <— 0 —> 1   -1 <— 0 —> 1

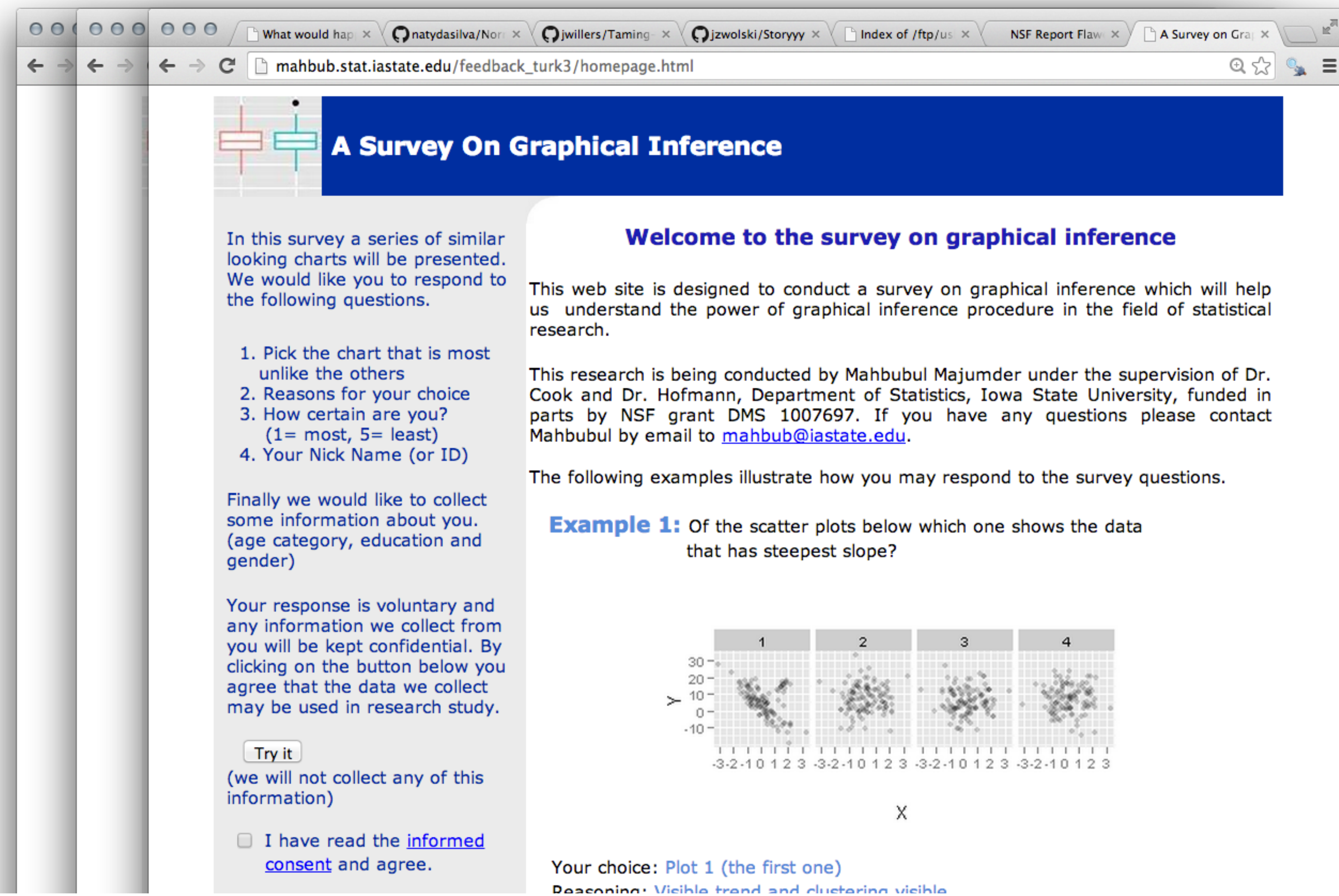
Scatterplots universally better, line plots are difficult when correlation is negative

Our experiments

- Experiments 1, 2, 3 compare lineup protocol with relevant classical test, result published in JASA
- Experiments 4, 5 examine plot design: cartesian vs polar, side-by-side boxplots vs dot plots
- Experiment 6 compares variations on boxplots: notched, violin, vase, beeswarm, ...
- Experiment 7 assesses large p, small n effects
- Experiment 9 tests for presence of any structure in an RNA-seq experiment
- .....

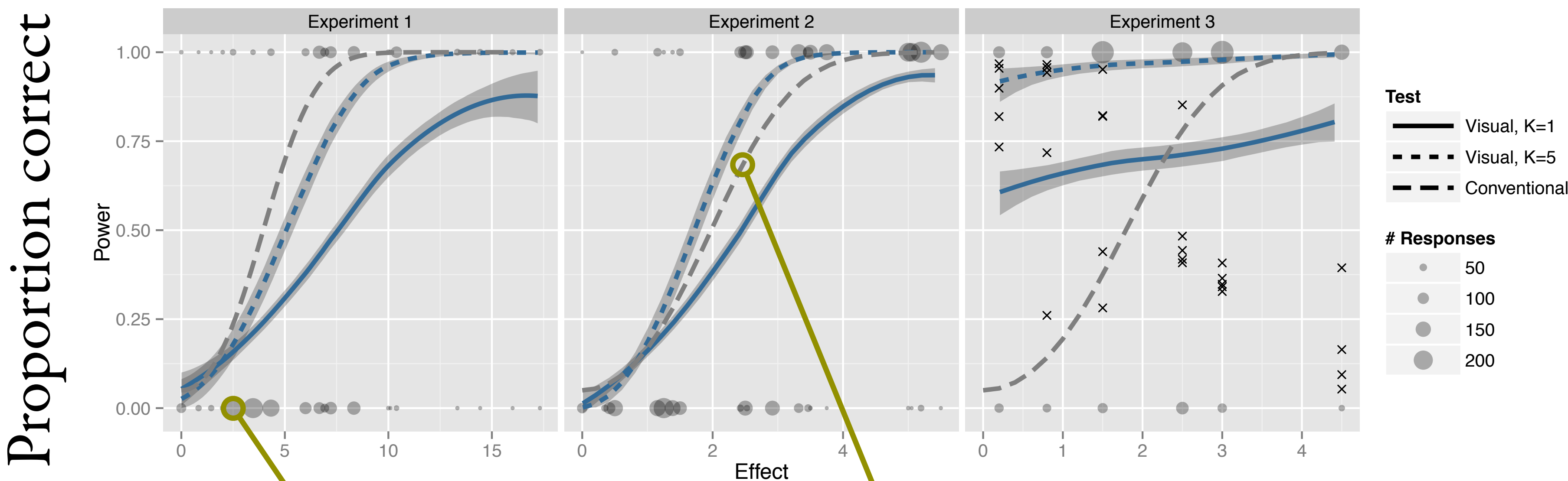
Experiments 1, 2, 3

http://mahbub.stat.iastate.edu/feedback_turk/homepage.html



Experiments 1, 2, 3

http://mahbub.stat.iastate.edu/feedback_turk/homepage.html



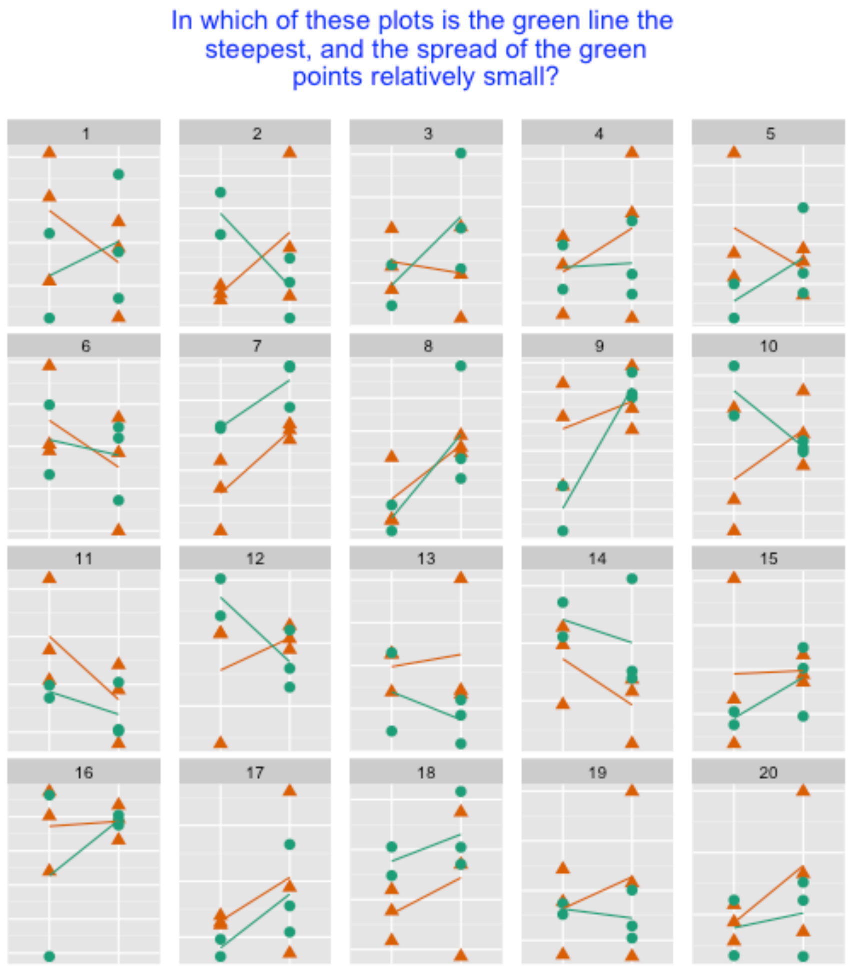
Effect: $E = \sqrt{n} \cdot \beta / \sigma$

- Power matches conventional test in form, if difference exists people see it
- Pooling results from (5) observers improves the power, and it is possible to beat conventional test power
- People beat conventional test when data was contaminated

Subject choices

Conventional test

Experiment 9



4 observers
3 chose the data plot
p-value is 0

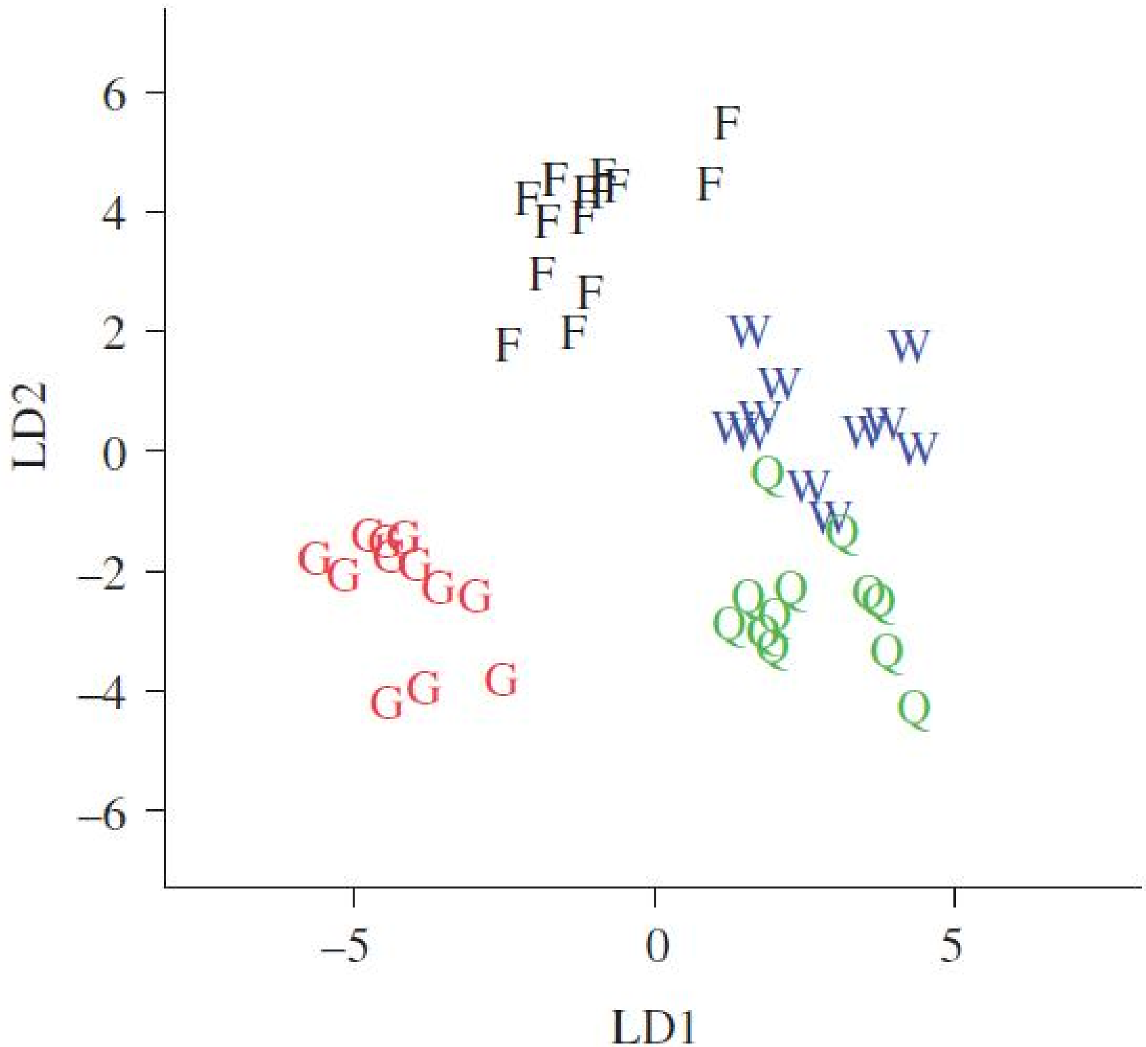
10 observers
8 chose the data plot
p-value is 0

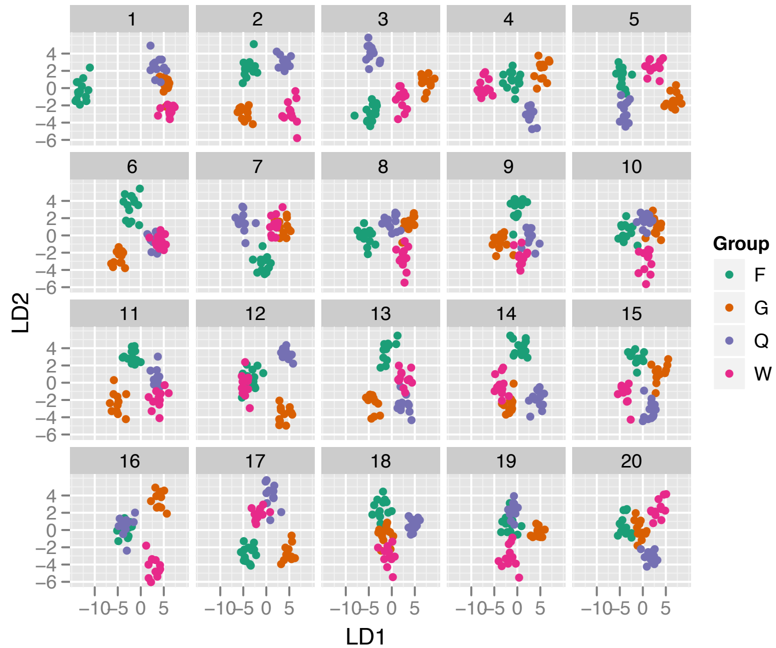
There is some significant structure in the data!

Experiment 7

- 40 oligos (variables)
- 48 wasps (cases)
- 4 types of wasps
- Best LDA 2D separation of four groups

(Toth et al, 2010)



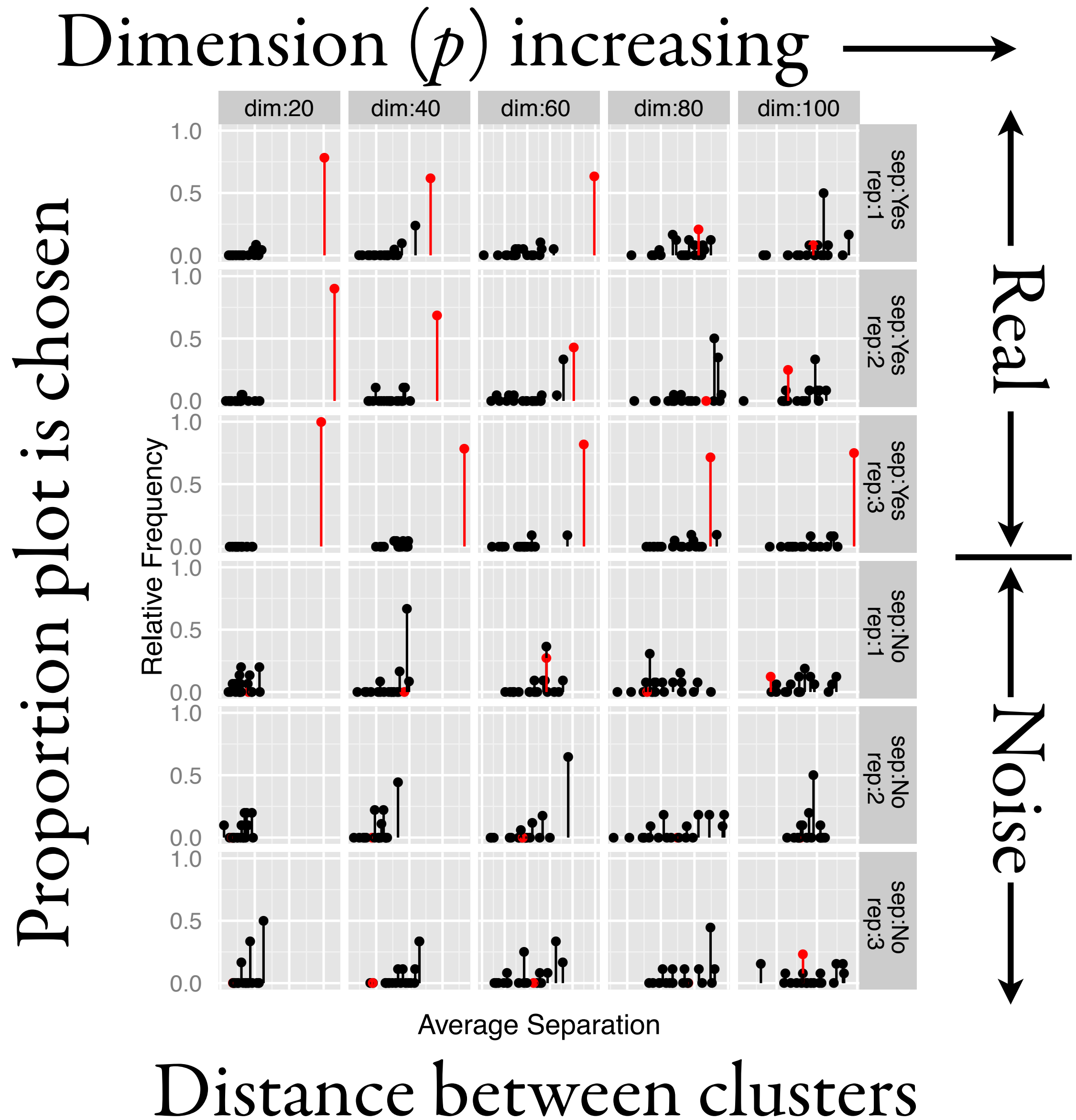


Wasps data plot was not detectable from null plots

Separation is not real

Data	Replicate	Num Subjects	Detection rate	p-value
Wasps	1	25	0.0000	1.0000
	2	13	0.0000	1.0000
	3	27	0.0000	1.0000
Purely noise	1	19	0.2632	0.0002
	2	18	0.0000	1.0000
	3	14	0.0000	1.0000

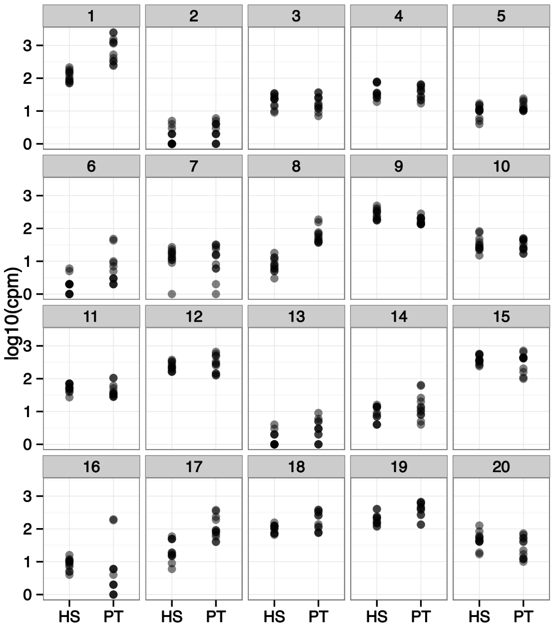
What are people choosing?



One pin = one plot in a lineup

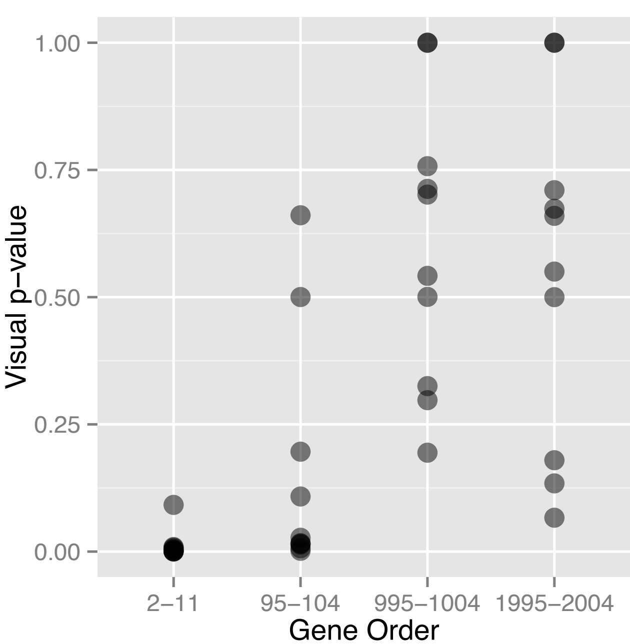
Red means data plot

People tend to pick plot with biggest difference.



Experiment 12

Provide an additional decision making method for deciding on the number of significant genes. Beyond FDR and ROC curves.

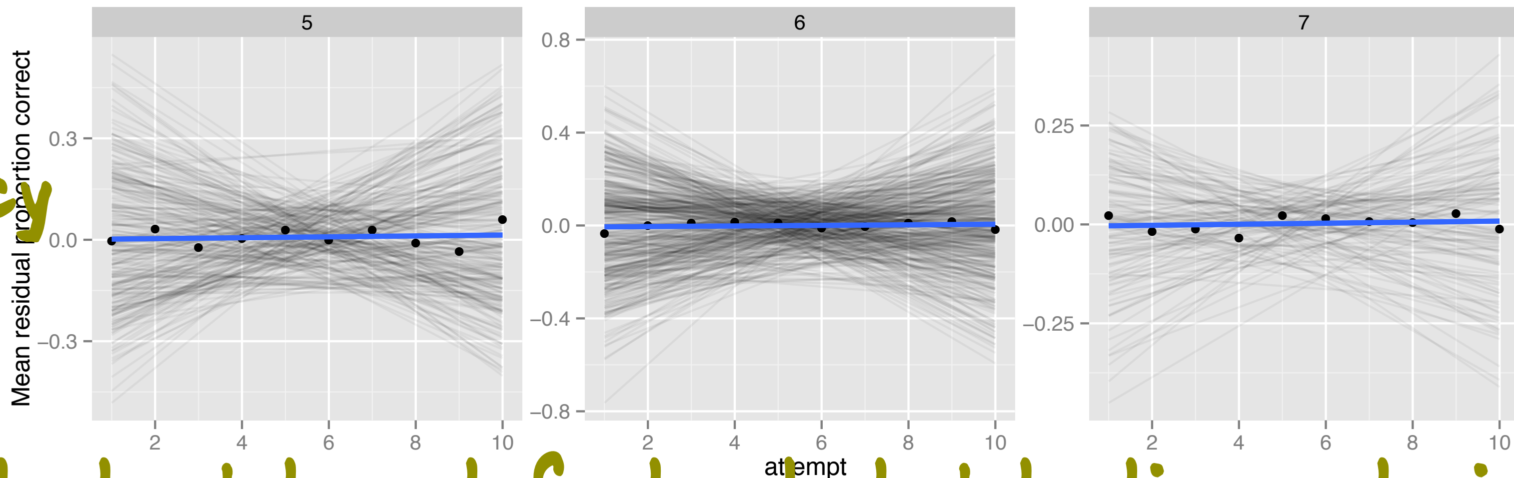


Pulled data from published study, stated ~3000 significant genes. Re-did the analysis, using permutation to provide nulls. Focused on top 10, 100, 1000, 2000. Computed visual p-values.

Suggests closer to 100 significant genes.

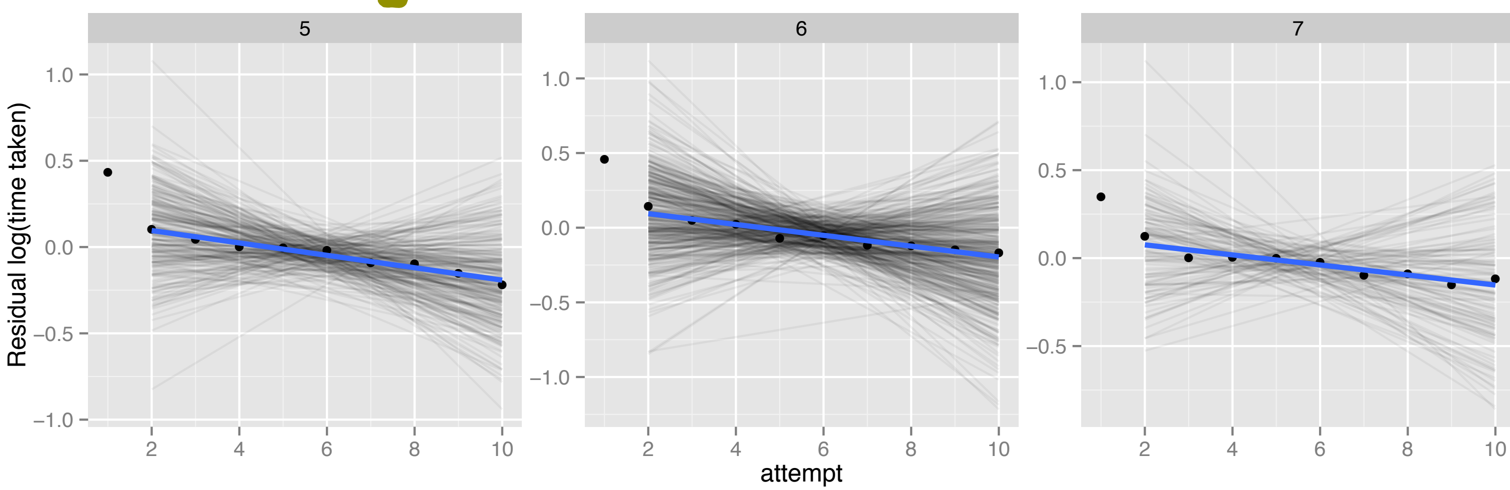
Learning trends

Accuracy

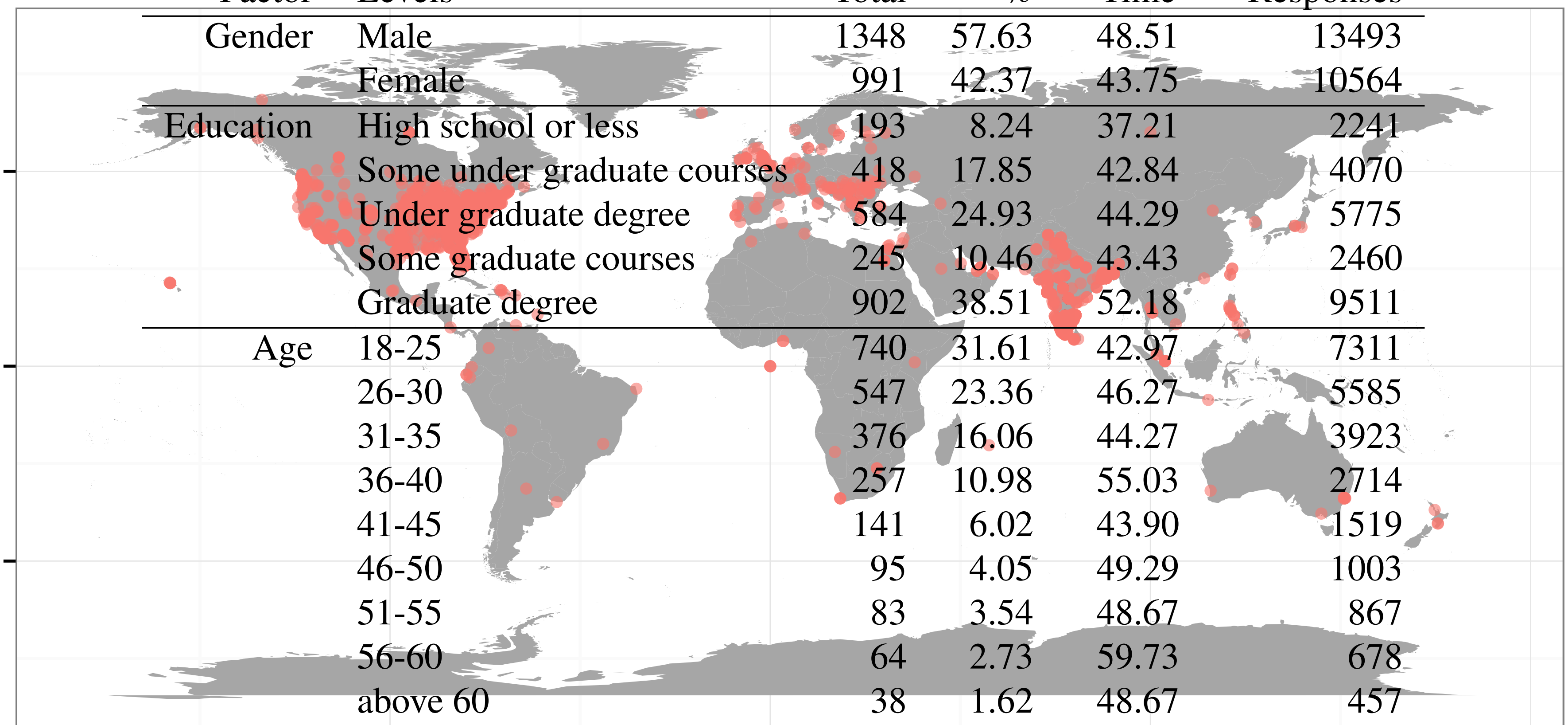


Subjects tend to get faster but detection rate is same

Speed



Turkers



Factor	Levels	Participants Total	Participants %	Average Time	Number of Responses
Gender	Male	1348	57.63	48.51	13493
	Female	991	42.37	43.75	10564
Education	High school or less	193	8.24	37.21	2241
	Some under graduate courses	418	17.85	42.84	4070
	Under graduate degree	584	24.93	44.29	5775
	Some graduate courses	245	10.46	43.43	2460
	Graduate degree	902	38.51	52.18	9511
Age	18-25	740	31.61	42.97	7311
	26-30	547	23.36	46.27	5585
	31-35	376	16.06	44.27	3923
	36-40	257	10.98	55.03	2714
	41-45	141	6.02	43.90	1519
	46-50	95	4.05	49.29	1003
	51-55	83	3.54	48.67	867
	56-60	64	2.73	59.73	678
	above 60	38	1.62	48.67	457
Country	United States	1087	46.83	39.64	10769
	India	980	42.22	52.63	10227
	Rest of the world	254	10.94	46.86	2819

R Package

- Nullabor package on CRAN
- When you plot your data, plot it first in a lineup, so you can be the unbiased observer

```
> lineup(null_permute("Obama.Romney"),
tracking.polls[,c(9,11)])
> decrypt("fg0t DARA up iYzuRuYp Q")
[1] "True data in position 5"
```

- Several null generating procedures included

Summary

- Visual inference offers quantitative assessment of significance, especially when there is no formal test available.
- With increasing size of data, statistical significance often can be obtained, but it is really practical significance, or effect size that is important to know.
- Visual inference protocols may be useful for introducing statistical thinking in introductory classes.
- Lineup protocol can help to decide best plot design for communication purposes.

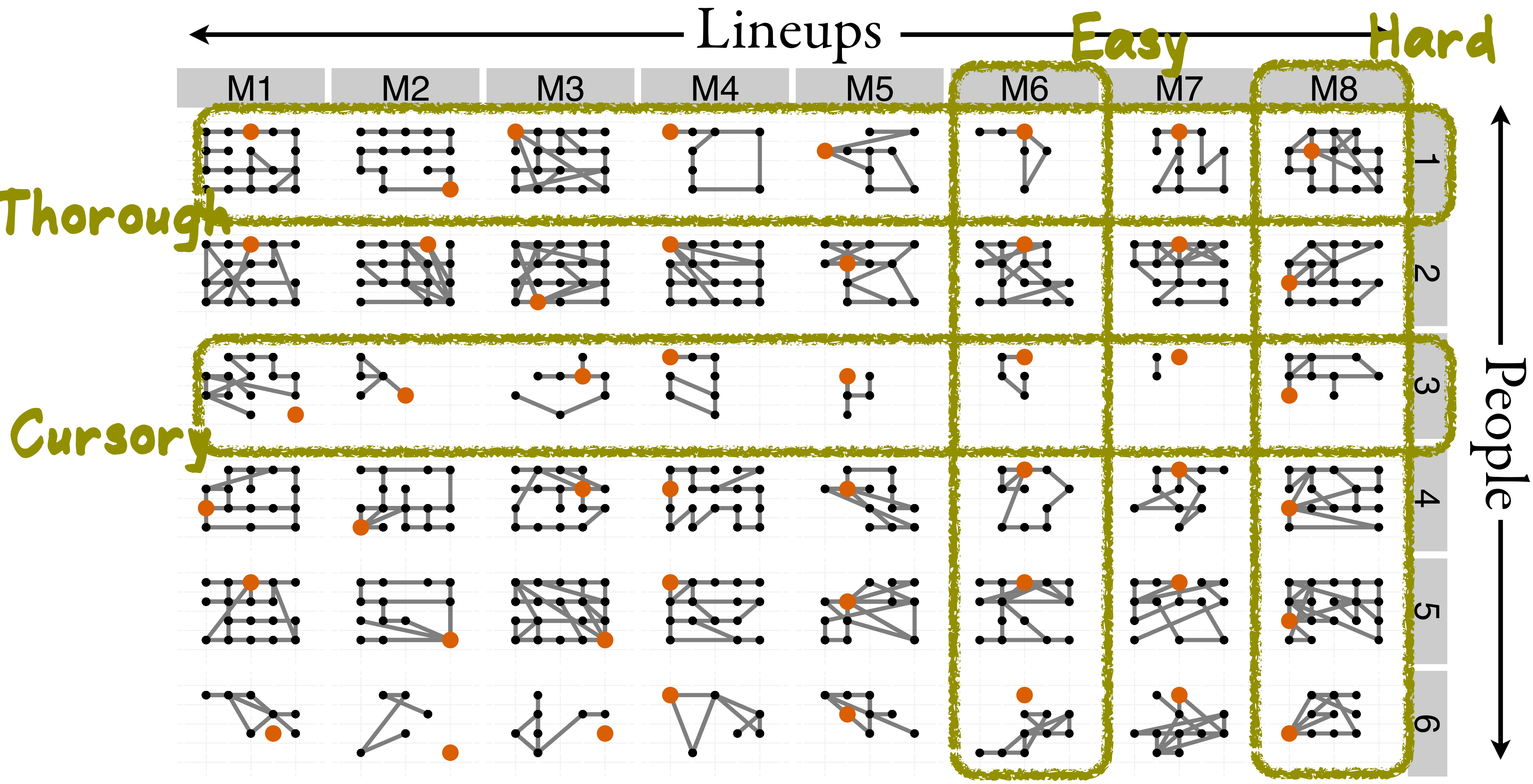
Acknowledgements

Plots produced using R package **`ggplot2`** by Hadley Wickham

Projection pursuit done using R package **`tourr`** by Wickham, Cook, with PDA index from Lee

National Science Foundation grant DMS 1007697

Eye-tracking experiment



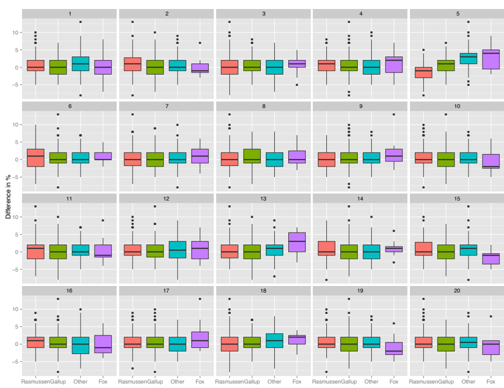
Some foundations...

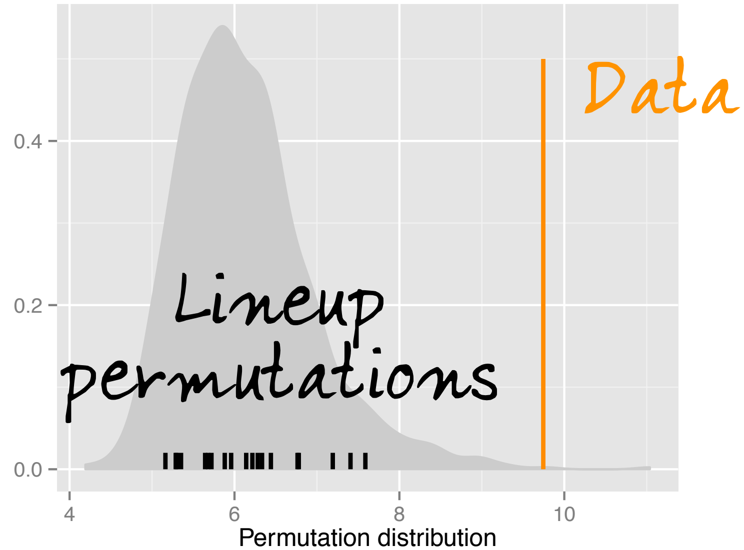
- Scott et al (1954): Generated synthetic plates to compare with real astronomical plates, acknowledged in Brillinger's (2005) Neyman lecture.
- Daniel (1976) had 40 pages of null plots for industrial applications.
- Diaconis (1983) describes 'magical thinking'.
- Buja et al (1988) describe 'Informal Statistical Inference' in association with the software Dataviewer.
- Gelman (2004) simulate data from statistical models.
- Davies (2008) suggest viewing null data sets.

Metrics

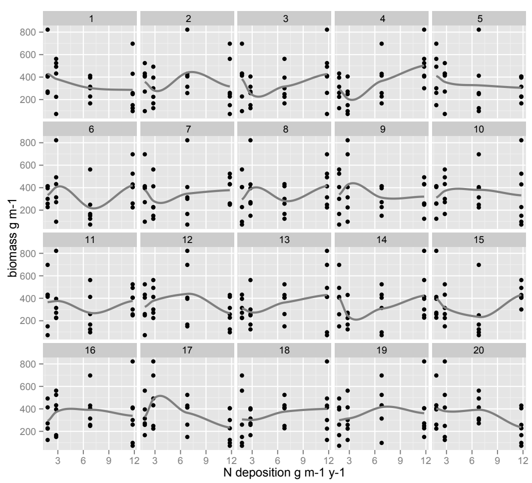
- Can we measure the structure in plots??
- Scagnostics: outlying, convex, skinny, stringy, monotonic, straight, striated, skewed (Wilkinson et al, 2005)
- Originally ideas from 1980s Tukey and Tukey (cognostics, scagnostics)
- Calculate for all pairs of plots, each plot gets a score for each of these structures

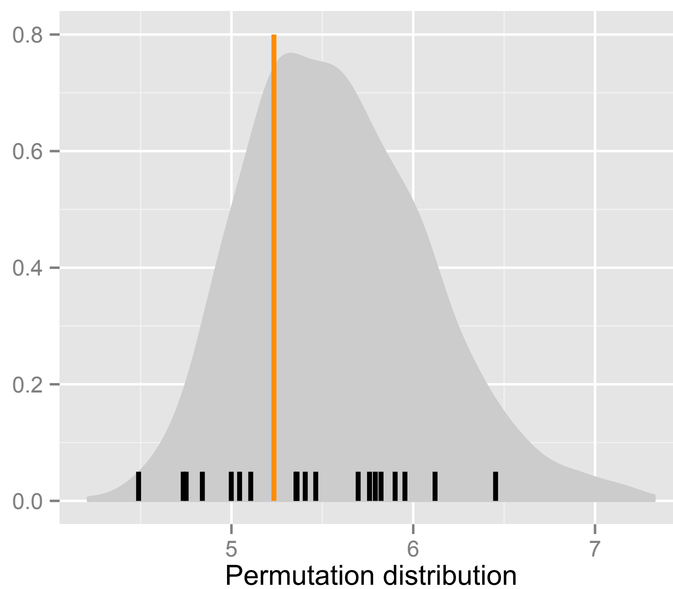
Structure, power and metrics





EASY?

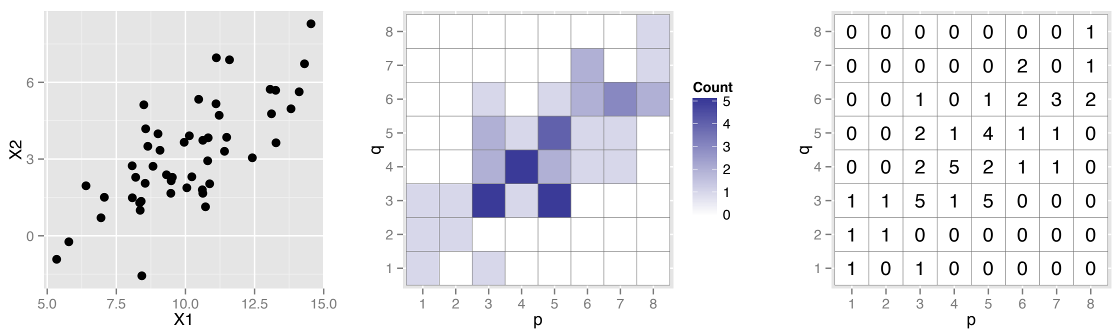




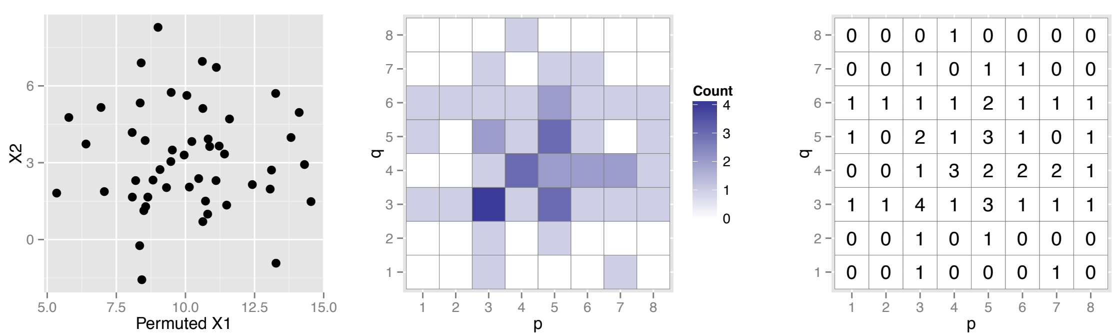
HARD?

Metrics

(a) Dataset X with two variables X_1 and X_2



(b) Dataset Y with permuted X_1 and original X_2



Binned distance

$$d^2_{BN}(X,Y) := ||C_X(X_1,X_2) - C_Y(X_1,X_2)||^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}(C_X(X_{1i},X_{2j}) - C_Y(X_{1i},X_{2j}))^2.$$

Metrics

Boxplots distance:

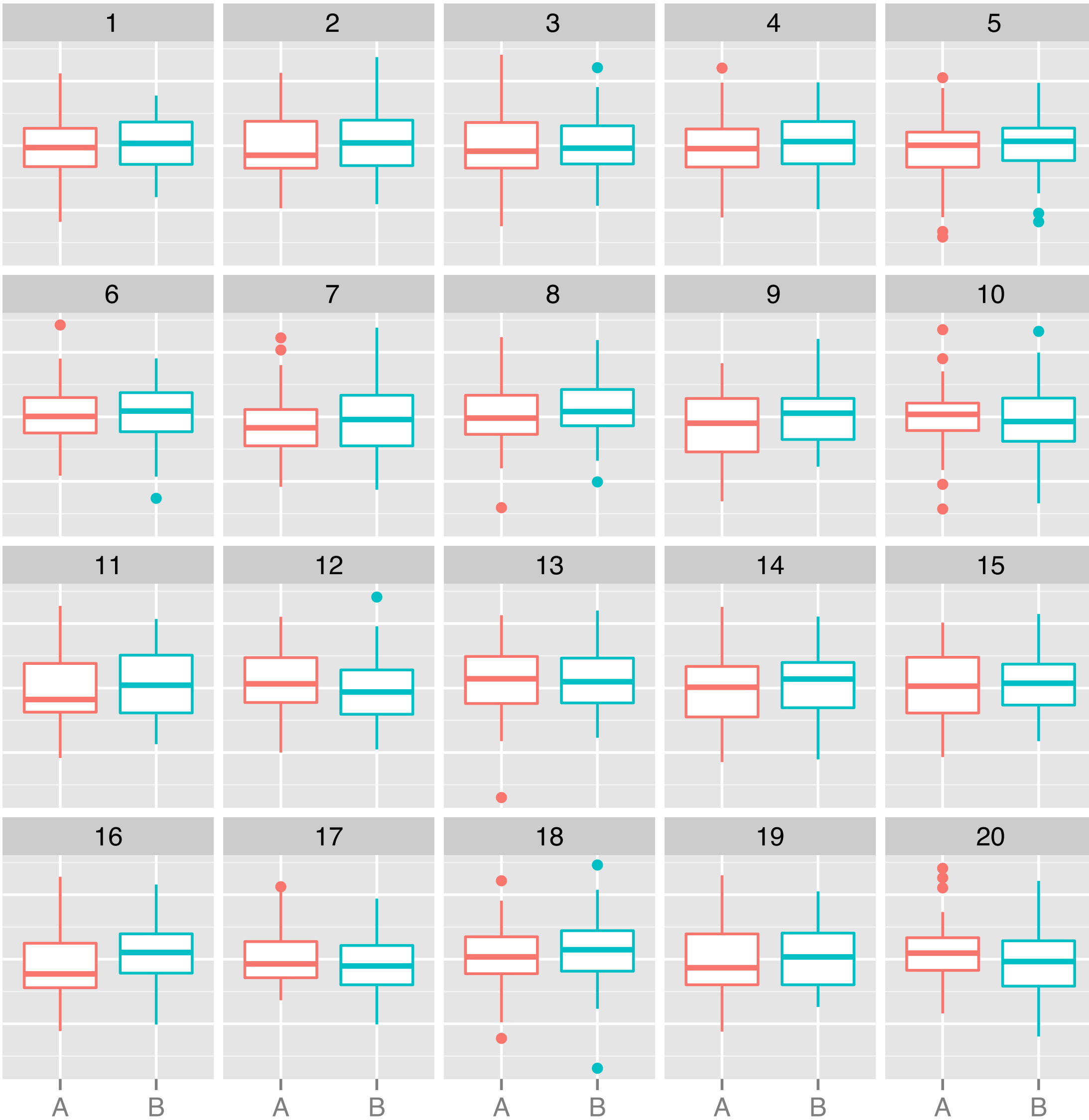
$$d^2_{BX}(X,Y) := ||d_q(X) - d_q(Y)||^2 = \sum_{i=1}^{3} \left(d_q(X)_i - d_q(Y)_i\right)^2 .$$

Regression line:

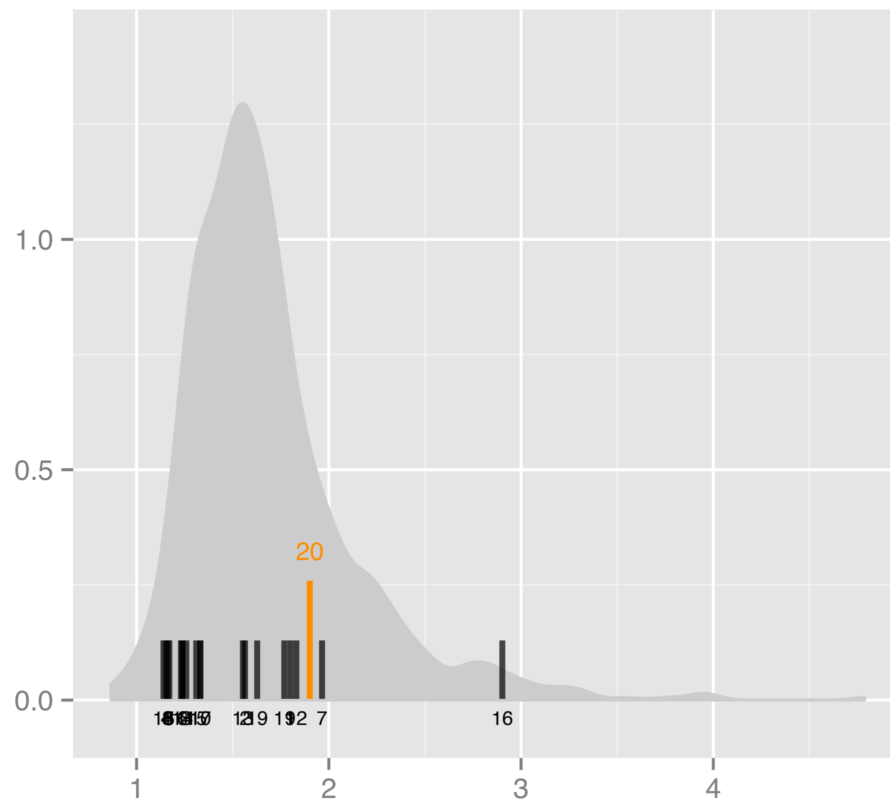
$$\begin{aligned} d^2_{RG}(X,Y) &:= \mathrm{tr}(B(X) - B(Y))'(B(X) - B(Y)) \\ &= \sum_{i=1}^{b}((b_0(X))_i - (b_0(Y))_i)^2 + \sum_{i=1}^{b}((b_1(X))_i - (b_1(Y))_i)^2 \end{aligned}$$

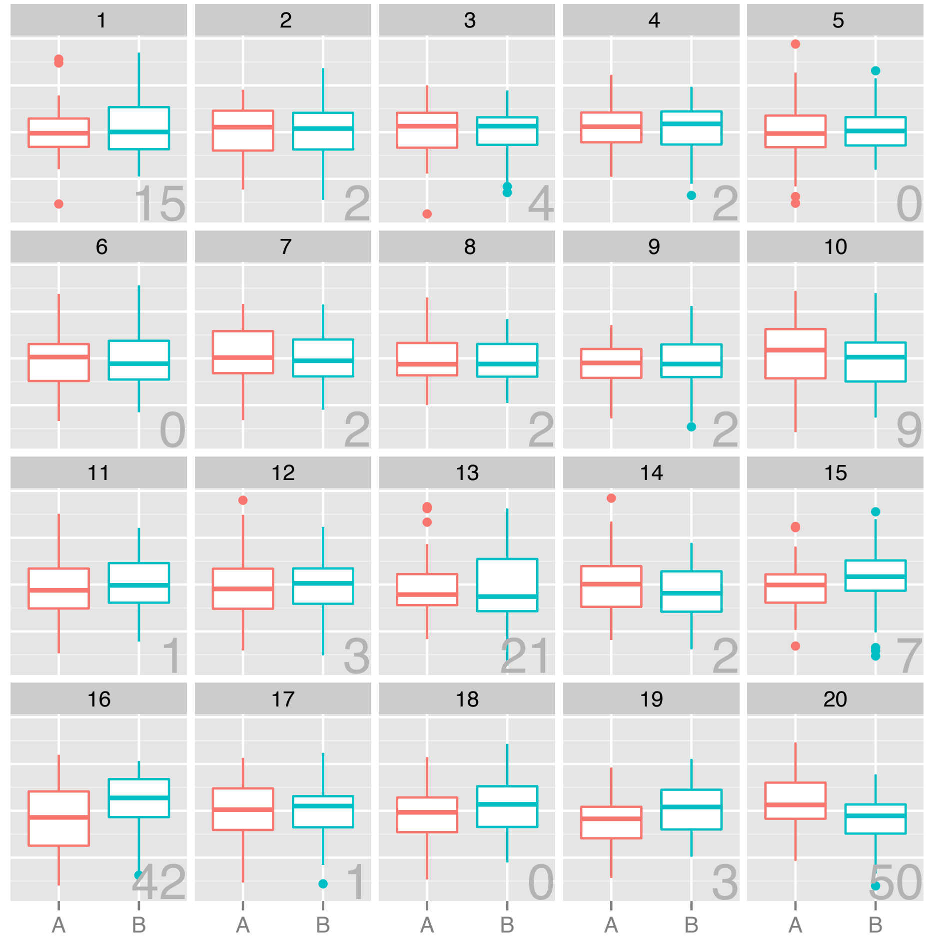
Separation between groups:

$$d^2_{MS}(X,Y) := ||s_m(X) - s_m(Y)||^2 = \sum_{i=1}^{g}((s_m(X))_i - (s_m(Y))_i)^2$$



(b) Boxplot based distance





(b) Boxplot based distance (c) Binned (2,8) distance (d) Binned (2,2) distance