

A GIRL GEEK'S GUIDE TO NEW RESEARCH ON INTERACTIVE DATA VISUALIZATION FOR STATISTICS WITH LOTS OF DATA

Di Cook

Econometrics and Business Statistics
Monash University

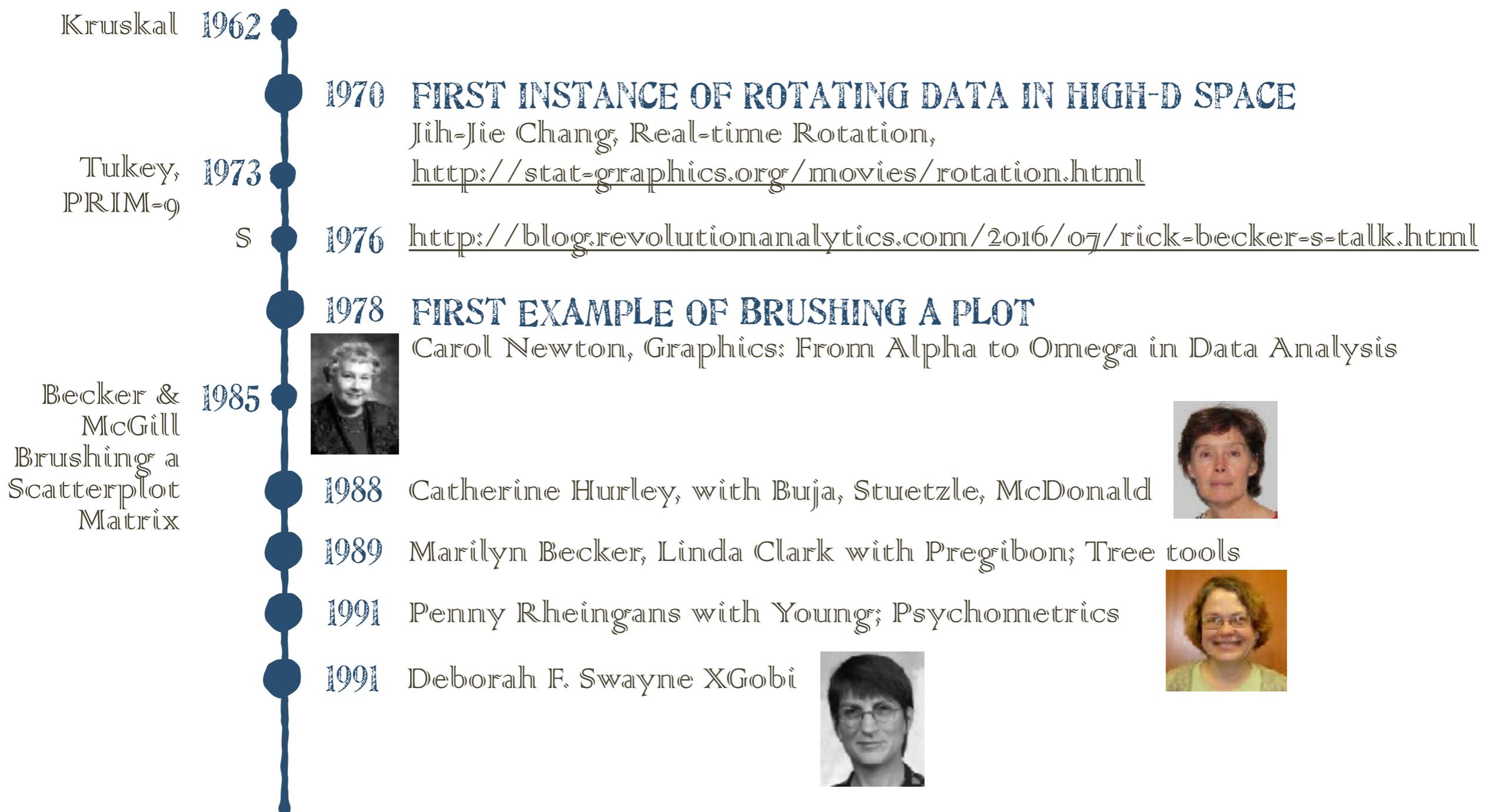


OUTLINE

- ▀ Break-throughs by women in graphics research
- ▀ Graphics research: formal protocol for testing merit of new methods, interactive graphics
- ▀ Hurdles, funding, promotions, journal and R package numbers
- ▀ Support today (R-Ladies, forwards, pyldies, WSDS)



FEMALE PIONEERS



There have always been women involved in graphics research

SKILLS REQUIRED

- ▀ Curiosity, persistence
- ▀ Foundations of statistical theory and methods
- ▀ Computing languages, R, and a lower level or some classical training
- ▀ Experience working with data, e.g. JSM Data Expo's, InfoVis challenges, kaggle
- ▀ Interactive data visualisation methodology
- ▀ Payoff is huge - researchers in this area are in high demand in highly paid interdisciplinary academic research positions, and especially in industry



AREAS OF INTEREST

- R fabulous static plots, and
 - close integration with models and
 - statistical thinking and
 - inference
- Interactive graphics: 10 steps back from software in the 90s
 - Programmability
 - Tight links with modelling
 - Portability
 - Speed and data size
 - Different kinds of linking, brushes
 - Conceptual frameworks



```
head(messy_data)
```

```
## # A tibble: 6 × 22
##   iso2 year m_04 m_514 m_014 m_1524 m_2534 m_3544 m_4554 m_5564 m_65
##   <chr> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1 AD    1989 NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2 AD    1990 NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 3 AD    1991 NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 4 AD    1992 NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 5 AD    1993 NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 6 AD    1994 NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## # ... with 11 more variables: m_u <int>, f_04 <int>, f_514 <int>,
## #   f_014 <int>, f_1524 <int>, f_2534 <int>, f_3544 <int>, f_4554 <int>,
## #   f_5564 <int>, f_65 <int>, m_u <int>
```

Messy data

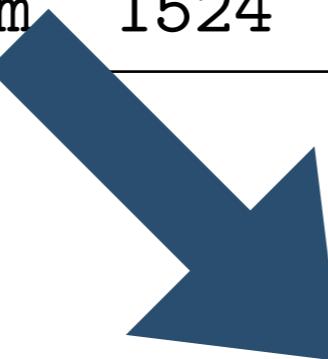
Tidy data

random variables

X₁ X₂ X₃ X₄ X₅

```
## # A tibble: 6 × 5
##   iso2 year gender age count
##   <chr> <int> <chr> <chr> <int>
## 1 AD    1996 m     1524 0
## 2 AD    1997 m     1524 0
## 3 AD    1998 m     1524 0
## 4 AD    1999 m     1524 0
## 5 AD    2000 m     1524 0
## 6 AD    2002 m     1524 0
```

```
## # A tibble: 6 × 5
##       iso2  year gender    age count
##       <chr> <int> <chr> <chr> <int>
## 1     AD    1996     m   1524     0
## 2     AD    1997     m   1524     0
## 3     AD    1998     m   1524     0
## 4     AD    1999     m   1524     0
## 5     AD    2000     m   1524     0
## 6     AD    2002     m   1524     0
```



```
data: tidy_data
layer:
  mapping: x = year,
            y = count, fill = gender
  geom: fill-bar
  facet: age
```

data: tidy_data

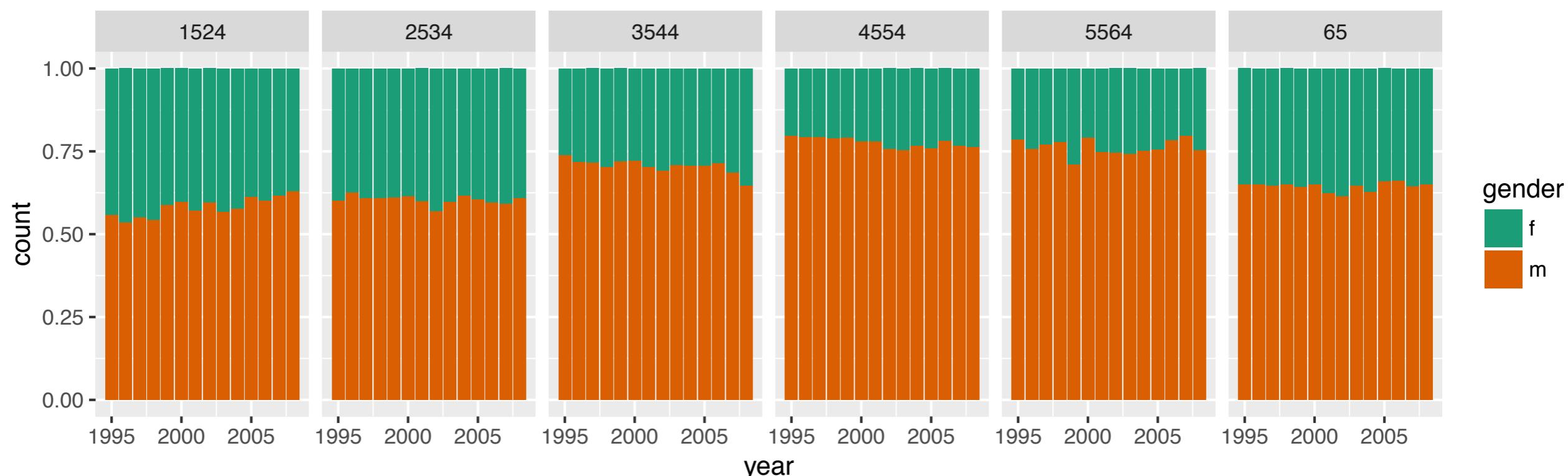
layer:

mapping: x = year,
y = count, fill = gender

geom: fill-bar

facet: age

100% charts



data: tidy_data

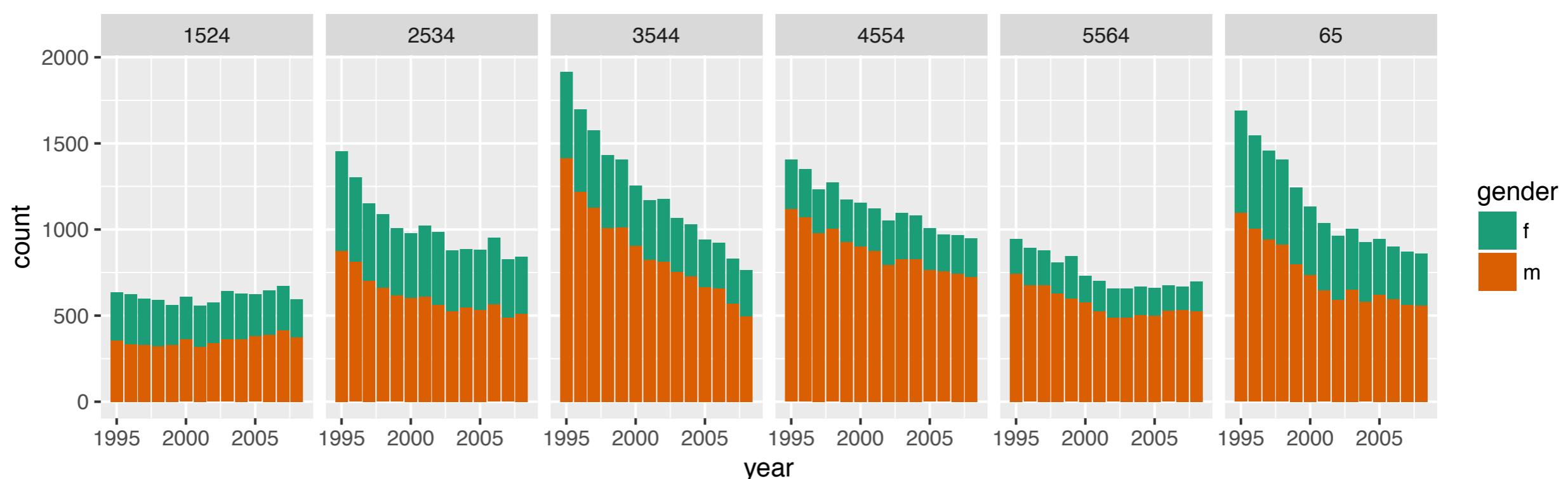
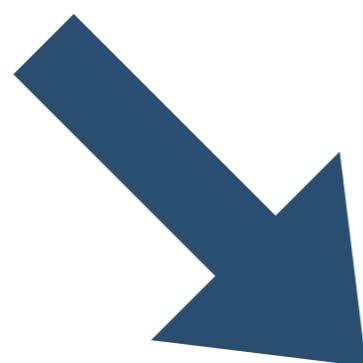
layer:

mapping: x = year,
y = count, fill = gender

geom: bar

facet: age

stacked barcharts



data: tidy_data

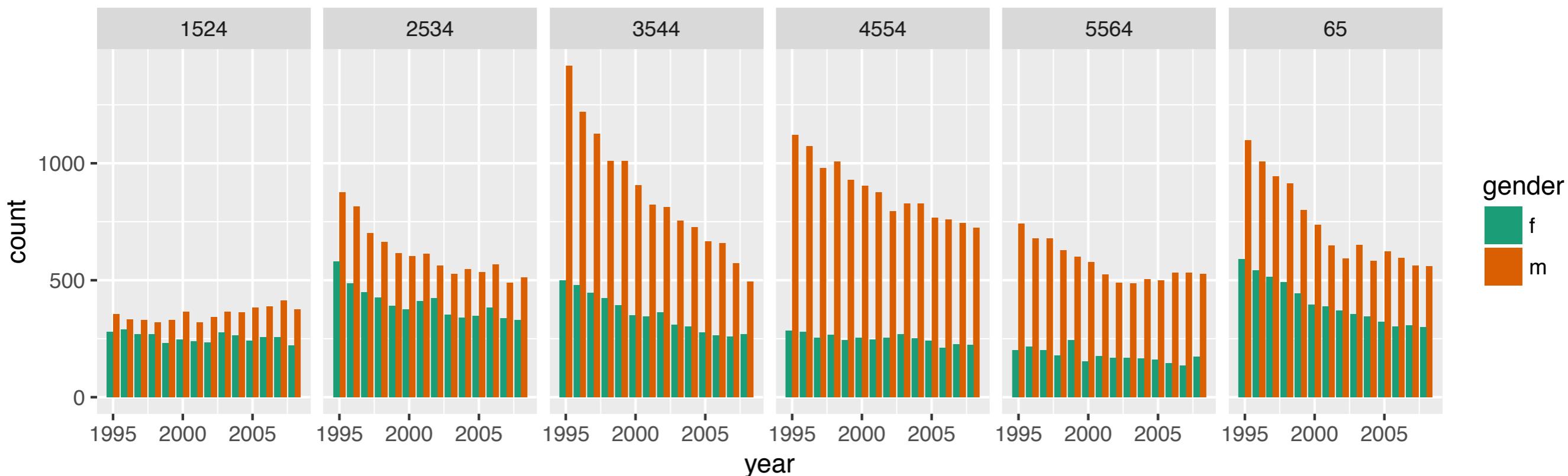
layer:

mapping: x = year,
y = count, fill = gender

geom: dodge-bar

facet: age

side-by-side bar charts



data: tidy_data

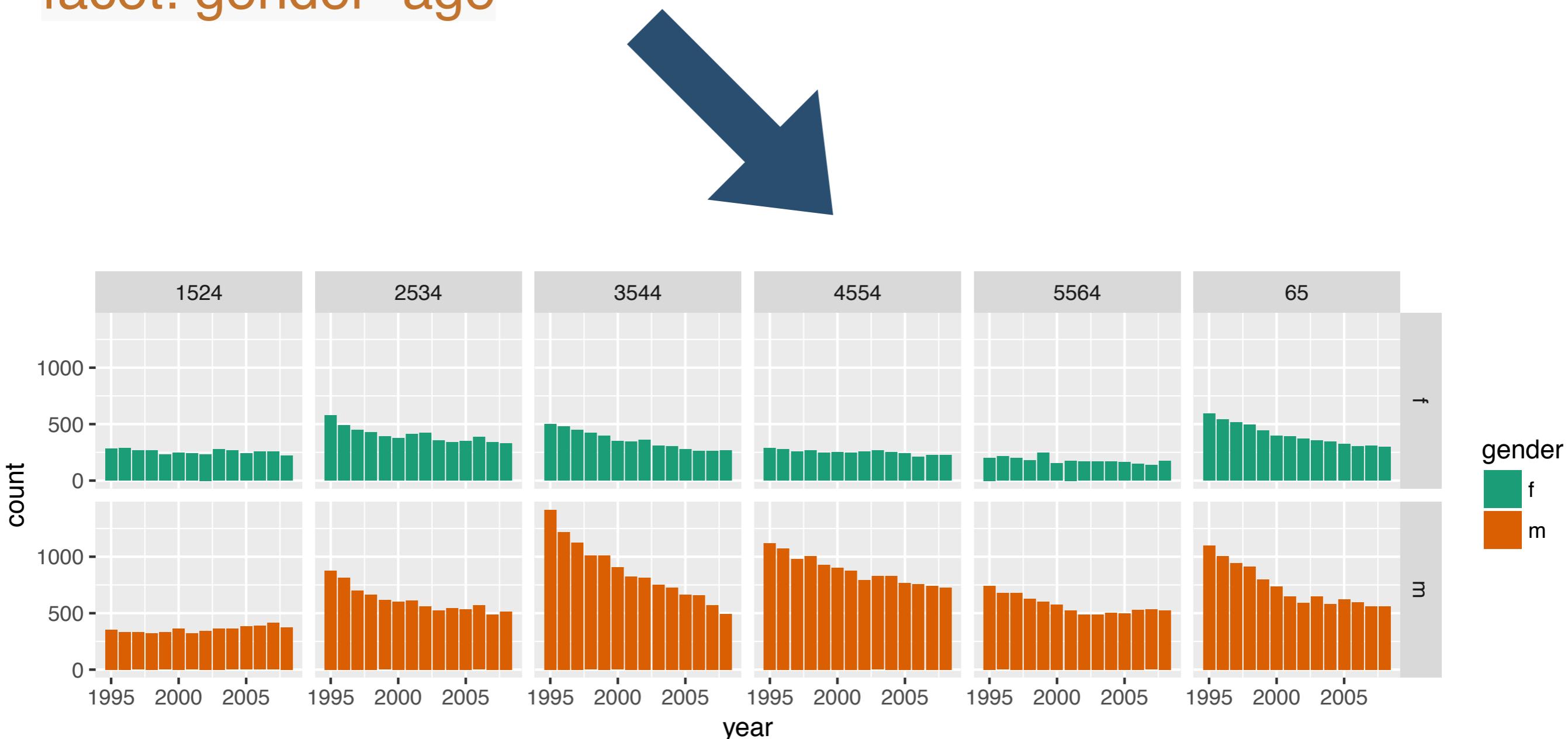
layer:

mapping: x = year,
y = count, fill = gender

geom: bar

facet: gender~age

bar charts



data: tidy_data

layer:

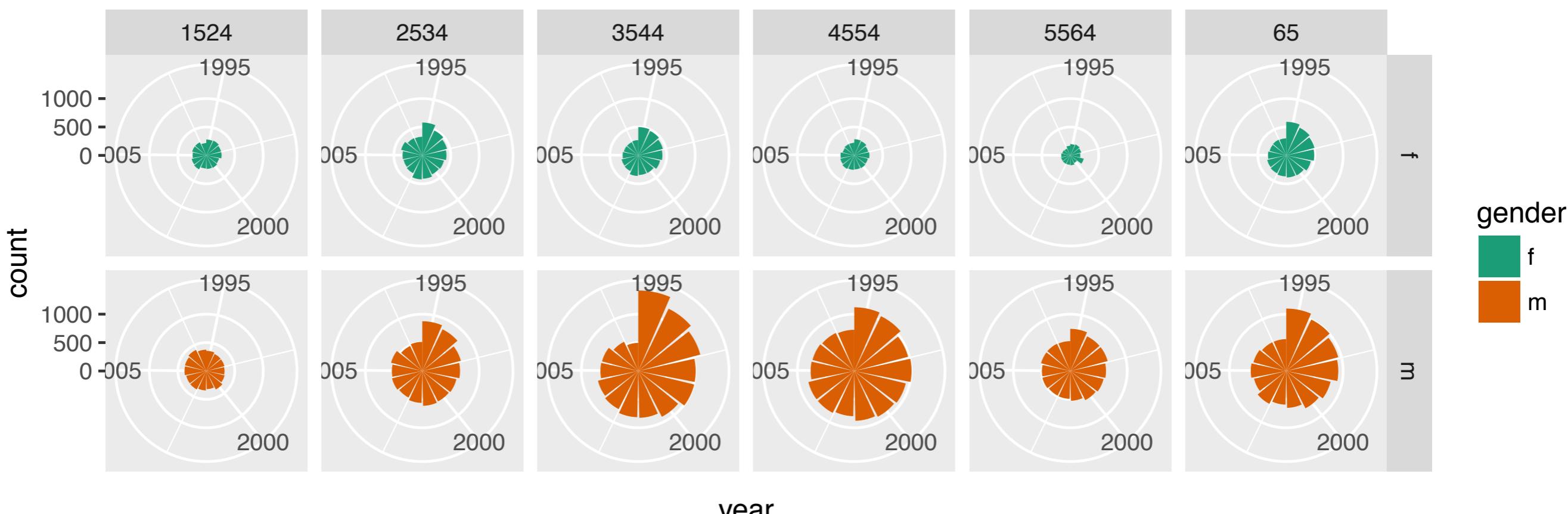
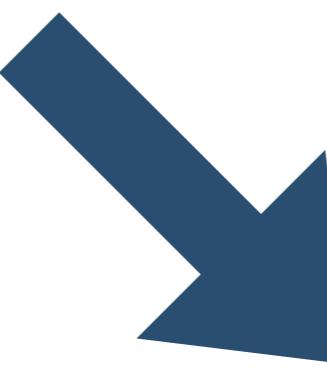
mapping: x = year,
y = count, fill = gender

geom: bar

facet: gender~age

coord: polar

rose plots



data: tidy_data

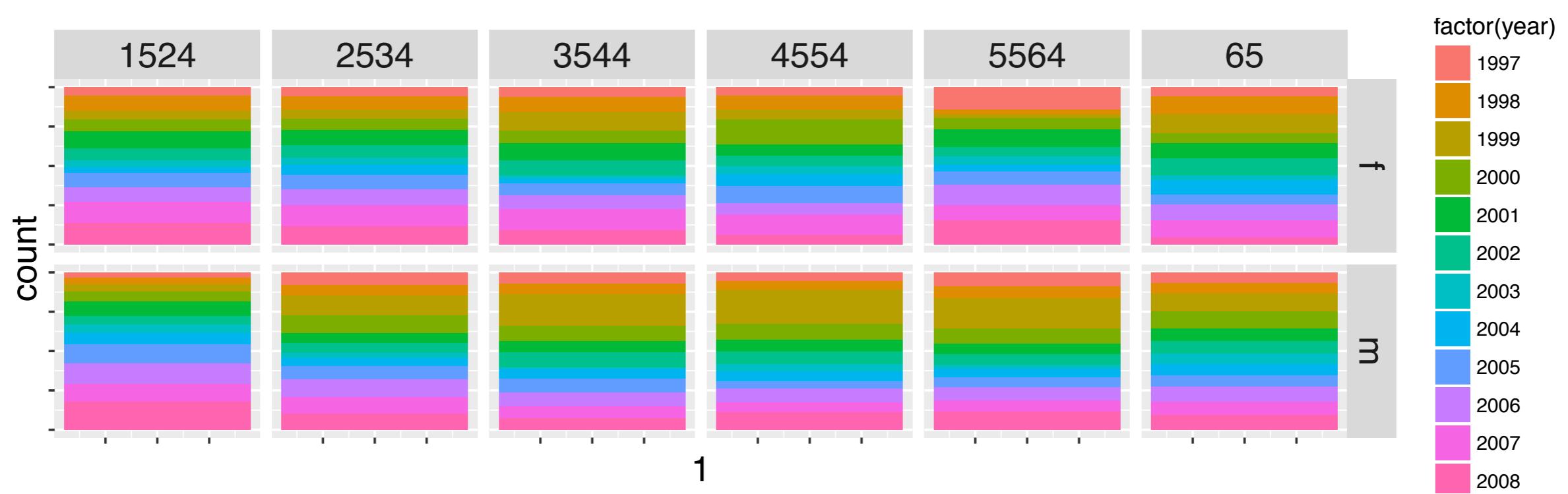
layer:

mapping: x = 1,
y = count, fill = year

geom: fill-bar

facet: gender~age

100% charts



data: tidy_data

layer:

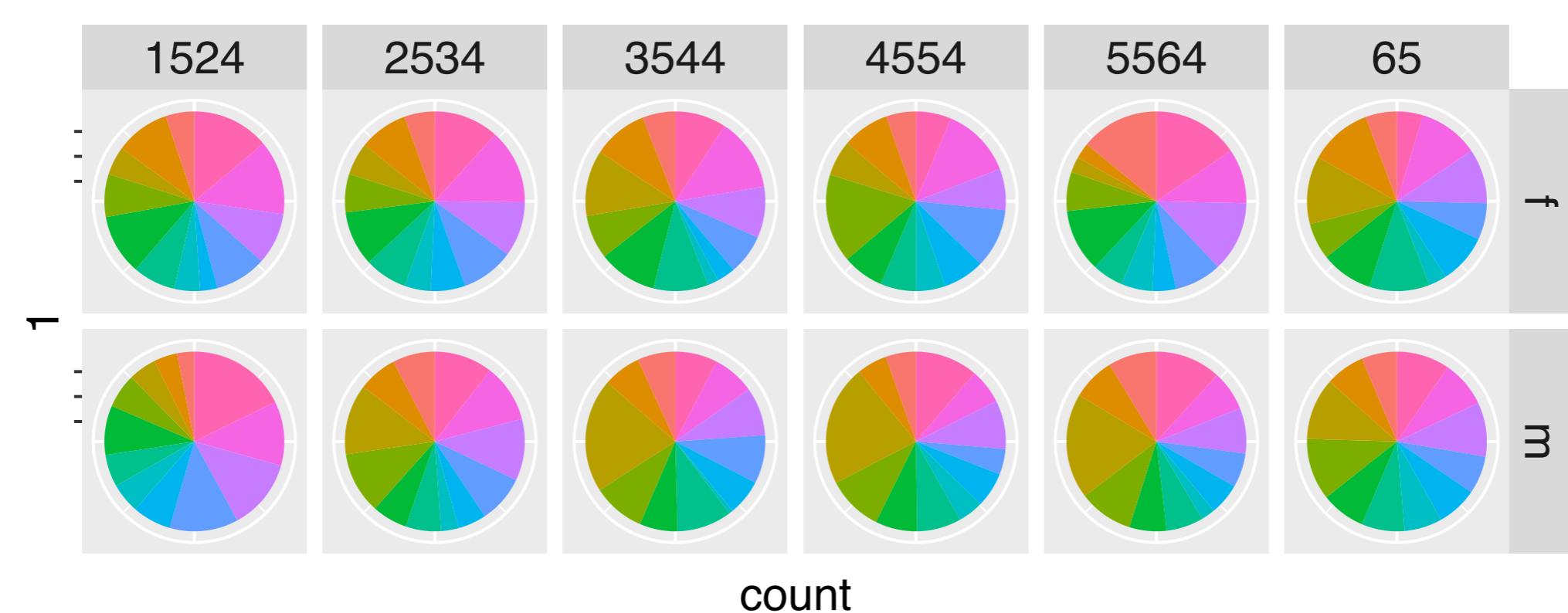
mapping: x = 1,
y = count, fill = year

geom: fill-bar

facet: gender~age

coord: polar

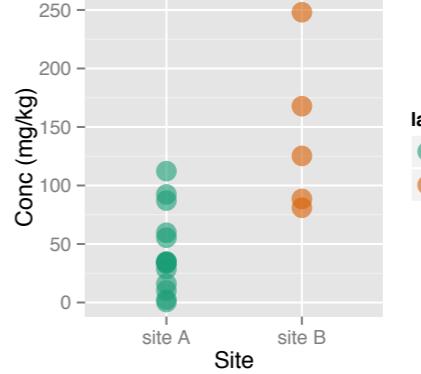
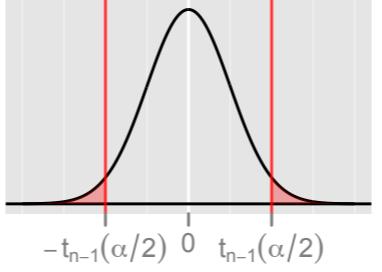
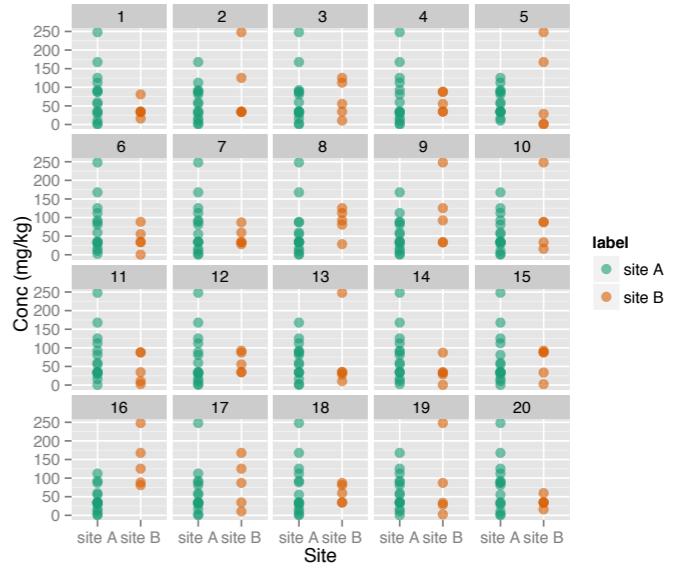
pie charts

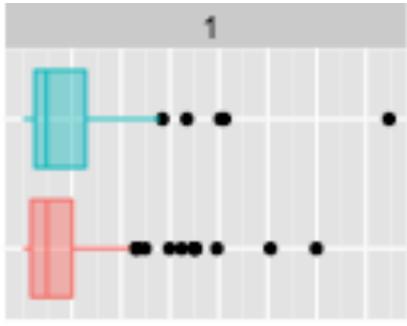




HELEN
GREEN

VISUAL INFERENCE

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



What is the question?

Is there a difference between the two groups

H_0 : no difference, H_1 : difference

What is the data?

Two variables: V1, V2; V1 is categorical

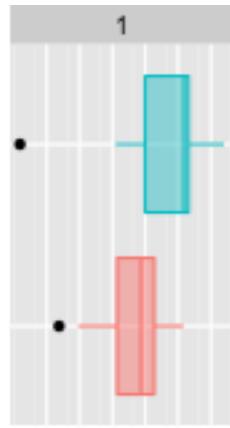
What is the mapping?

`x=V2, y=V1, colour=V1
geom=boxplot`

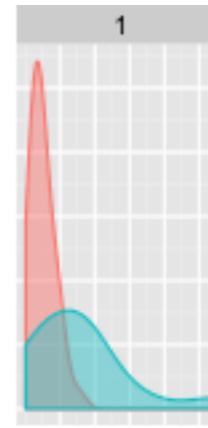
What is a null generating mechanism?

permute the values of V1,
relative to V2

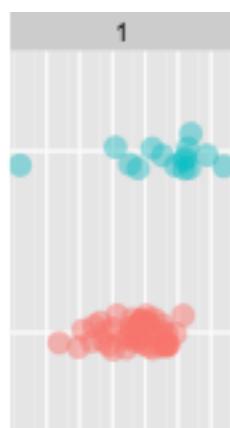
WHICH IS THE BEST DESIGN?



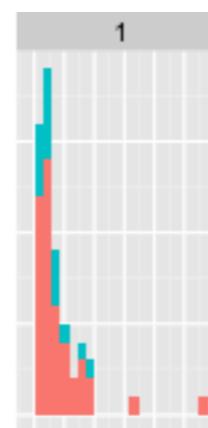
mapping:
 $x = V_2$
 $y = V_1$
geom: boxplot



mapping:
 $x = V_1$
color = V_2
geom: density



mapping:
 $x = V_2$
 $y = V_1$
geom: dotplot

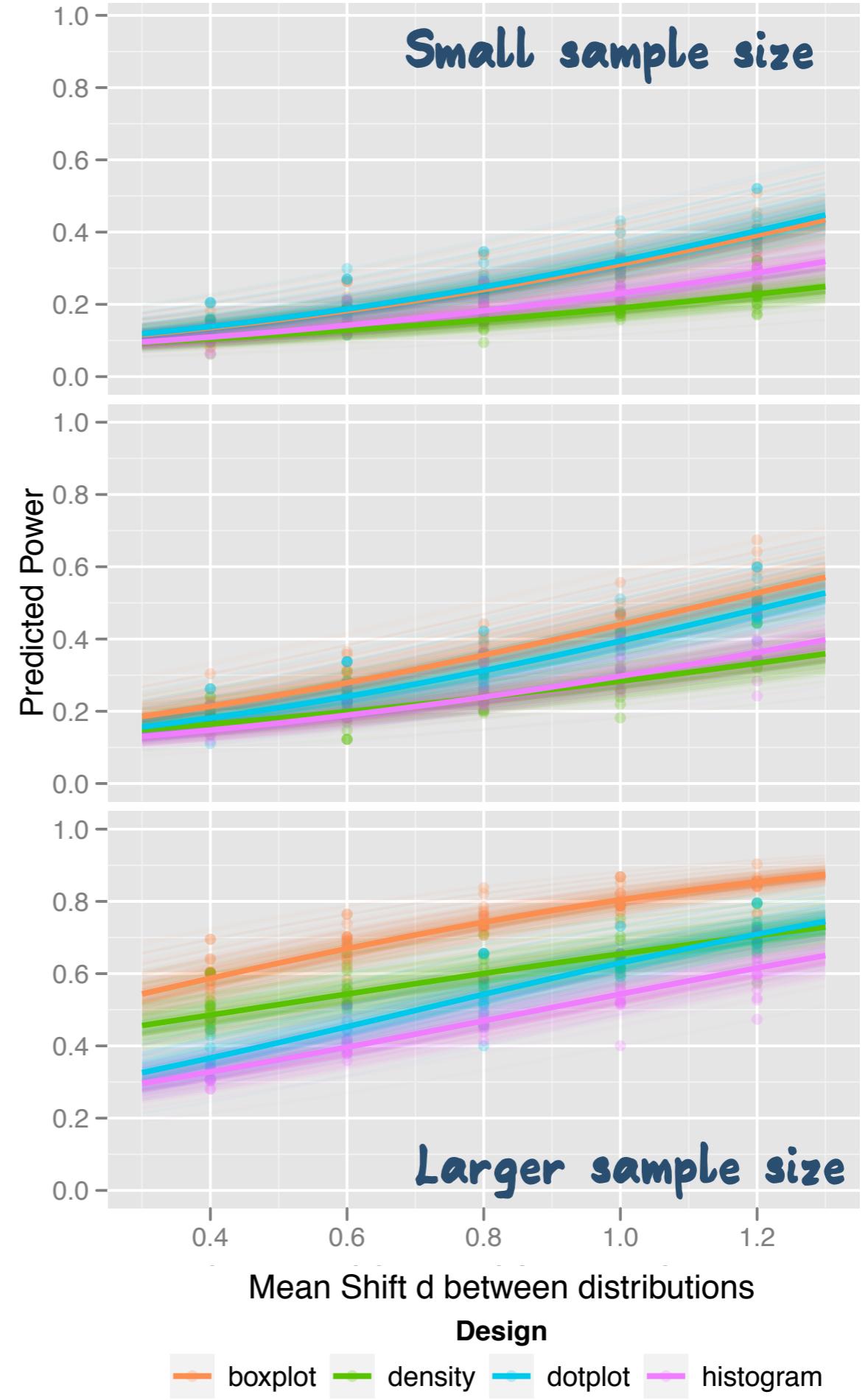


mapping:
 $x = V_1$
color = V_2
geom: histogram

Compute power for each design.

The design which allows reader to detect the difference more frequently is the most powerful statistic.

People detect the data plot significantly more often when side-by-side boxplots are used, except when the sample size is small, and then side-by-side dotplots win



INTERACTIVE GRAPHICS

 There have been a lot of advances in interactive graphics in the last couple of years, e.g. `plotly` (Sievert), `ggvis` (Wickham), `htmlwidgets` (RStudio), `rbokeh` (Hafen), `crosstalk` (Cheng), `rCharts` (Vaidyanathan), ...

 BUT we are still decades behind the state-of-the-art of the 90s

 The gold standards were `XLispStat`, `DataDesk`, `XGobi/GGobi`





<http://stat-graphics.org/movies/xgobi.html>

PLOTLY

- Takes a `ggplot2` object and converts to javascript
- Mouse over labels are easy
- Brushing of groups of points possible, but harder
- Linking between plots possible, a little trickier
- Tours, rotation in high-d, possible via animation, not live projection computation

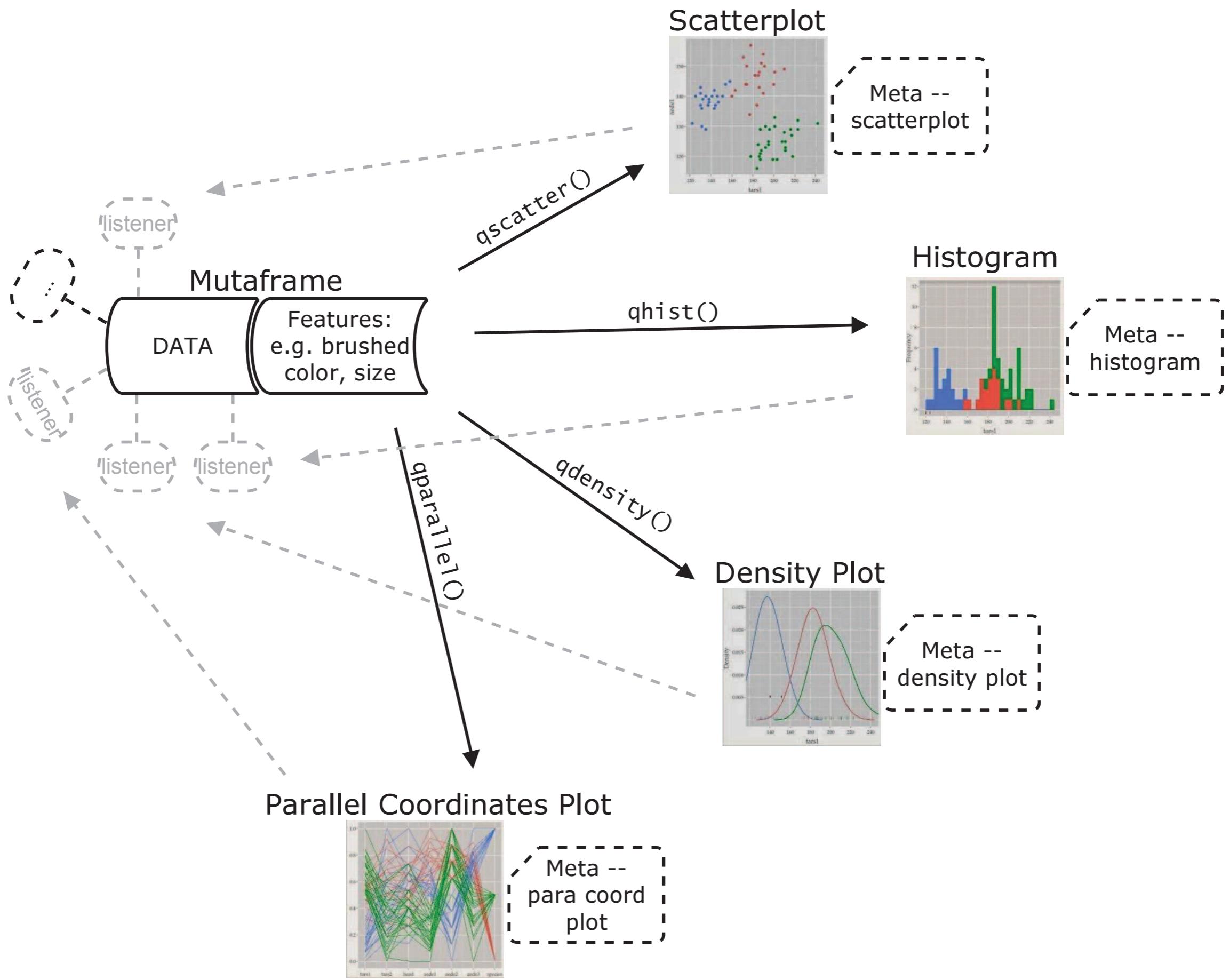
See Australian election app <https://vimeo.com/167367369>

CRANVAS

- qt widget set - speed, but trouble with portability
- R wrappers to qt functionality
- Fixed set of plot types, but new could be programmed directly in R
- Have to program plots from first principles
- Brushes and linking defined
- Linking by mutable objects (plumber), and reference classes (ObjectSignals)

Xie, Cheng, Hofmann, Cook, Schloerke, Vendettuoli,
built on foundation by Lawrence, Wickham





```
library(cranvas)
data(nrcstat)

nrcstat$Inst.Prg <- paste(nrcstat$Institution, nrcstat$ProgramName)
nrcdist <- dist(scale(nrcstat[,c(20,21,26,30,32,33,34,36,41,43,46)]))
nrc.hc <- hclust(nrcdist, method="ward")
plot(nrc.hc)

nrc.clust <- data.frame(nrc, cl1=cutree(nrc.hc, k=12),
```

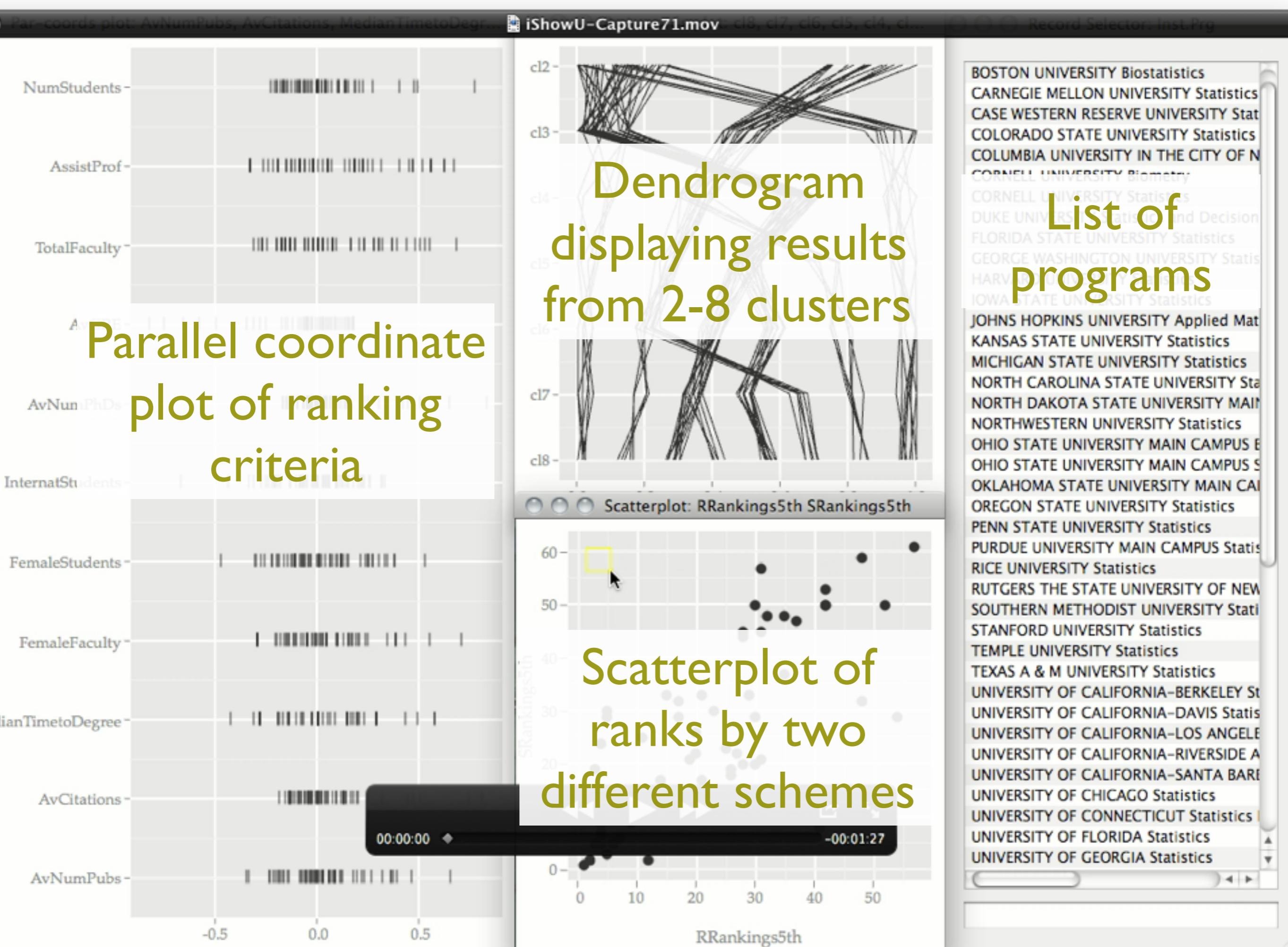
Set up data for cranvas

```
cl2=cutree(nrc.hc, k=13), cl3=cutree(nrc.hc, k=14),
cl4=cutree(nrc.hc, k=15), cl5=cutree(nrc.hc, k=16),
cl6=cutree(nrc.hc, k=17), cl7=cutree(nrc.hc, k=18),
cl8=cutree(nrc.hc, k=19))

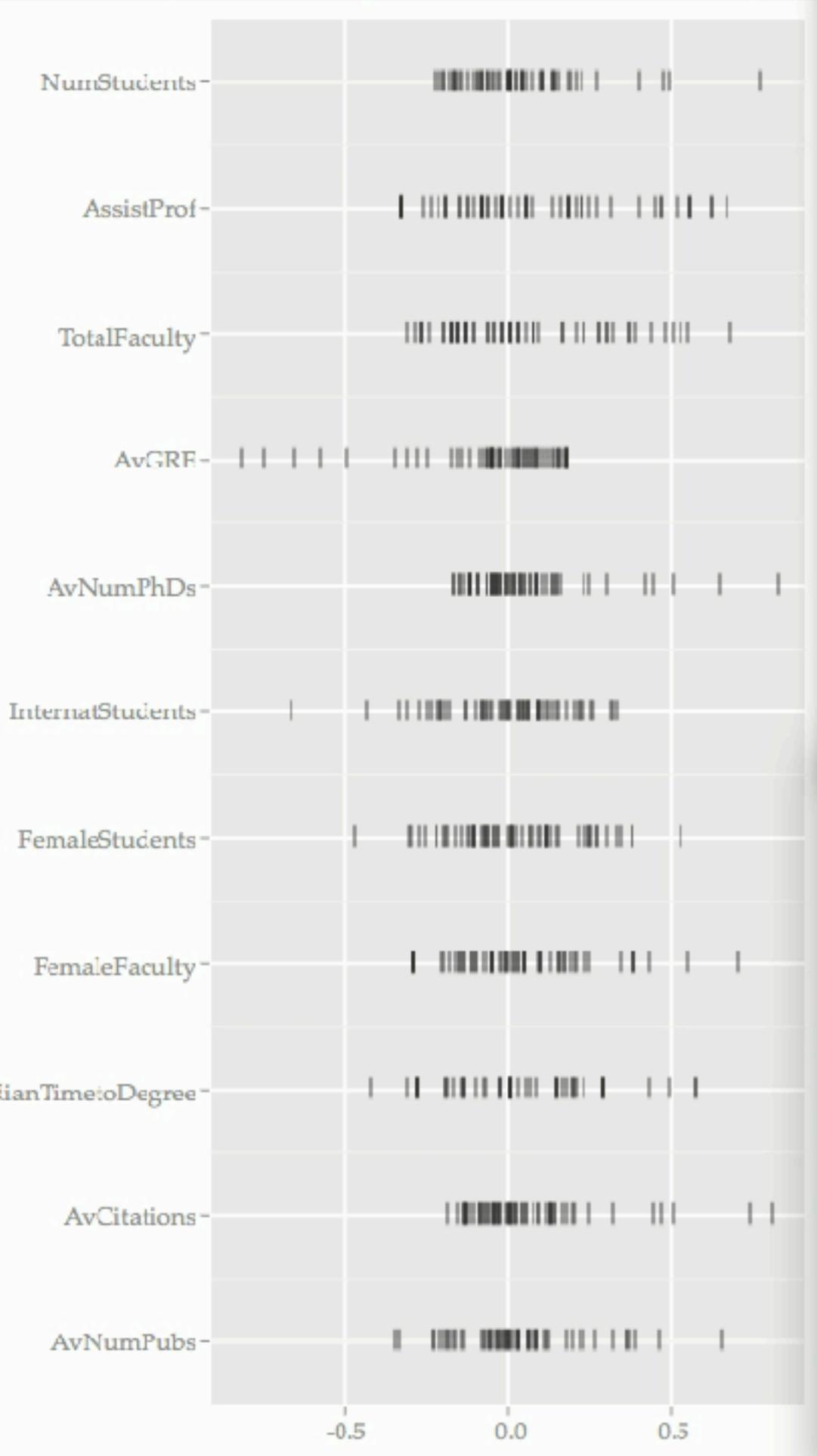
qnrc.clust <- qdata(nrc.clust)
qparallel(c(20,21,26,30,32,33,34,36,41,43,46), data=qnrc.clust,
          center = median, horizontal=T, glyph = "tick")
qparallel(80:74, data=qnrc.clust, horizontal=T,
          jitter=c("cl2","cl3","cl4","cl5","cl6","cl7","cl8"))
qscatter(RRankings5th, SRankings5th, data=qnrc.clust)
record_selector(Inst.Prg, qnrc.clust)
```

Load library, data, add better id variable

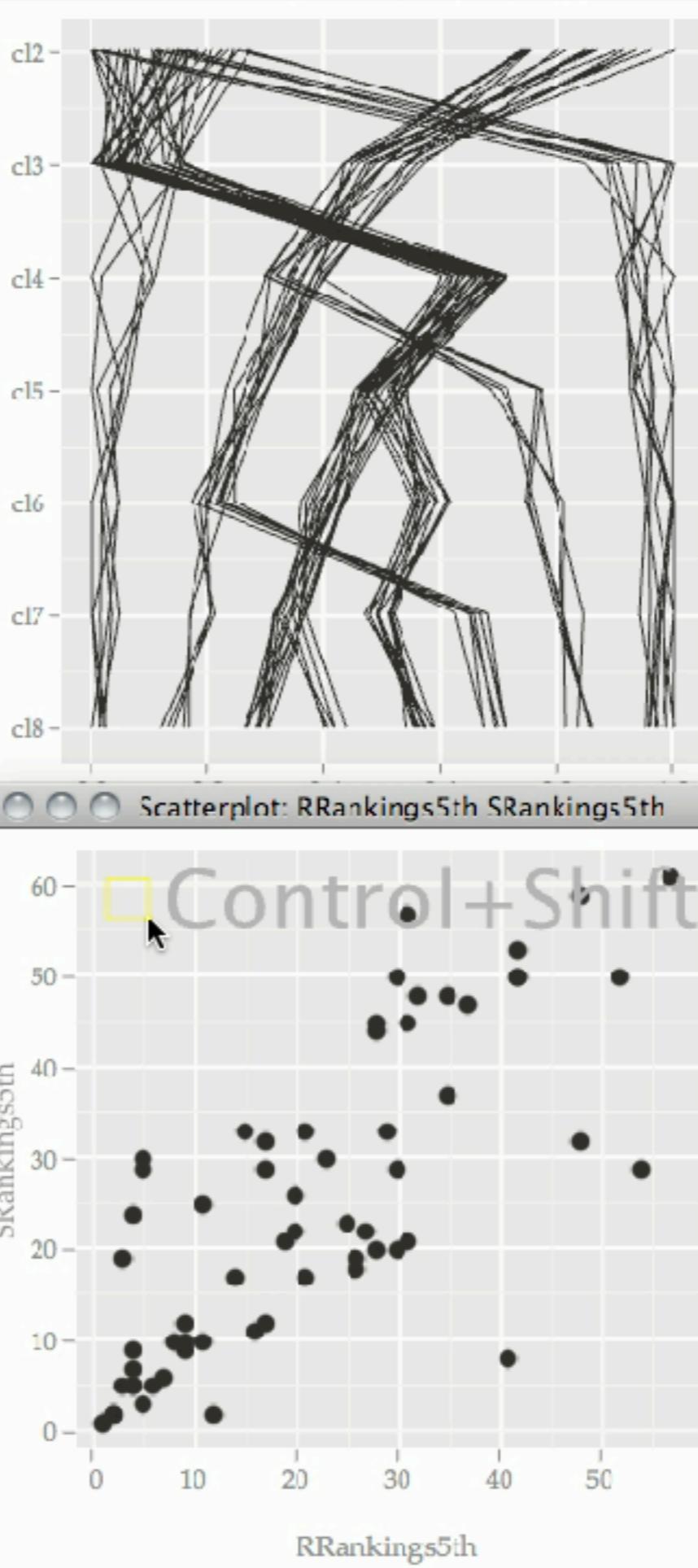
Use Euclidean distance
Open a pcp on ranking
criteria, pcp for
dendrogram, scatterplot,
and label browser



Par-coords plot: AvNumPubs, AvCitations, MedianTimetoDegr...



Par-coords plot: cl8, cl7, cl6, cl5, cl4, cl...



Record Selector: Inst.Prg

BOSTON UNIVERSITY Biostatistics
 CARNEGIE MELLON UNIVERSITY Statistics
 CASE WESTERN RESERVE UNIVERSITY Stat
 COLORADO STATE UNIVERSITY Statistics
 COLUMBIA UNIVERSITY IN THE CITY OF N
 CORNELL UNIVERSITY Biometry
 CORNELL UNIVERSITY Statistics
 DUKE UNIVERSITY Statistics and Decision
 FLORIDA STATE UNIVERSITY Statistics
 GEORGE WASHINGTON UNIVERSITY Statis
 HARVARD UNIVERSITY Statistics
 IOWA STATE UNIVERSITY Statistics
 JOHNS HOPKINS UNIVERSITY Applied Mat
 KANSAS STATE UNIVERSITY Statistics
 MICHIGAN STATE UNIVERSITY Statistics
 NORTH CAROLINA STATE UNIVERSITY Sta
 NORTH DAKOTA STATE UNIVERSITY MAIN
 NORTHWESTERN UNIVERSITY Statistics
 OHIO STATE UNIVERSITY MAIN CAMPUS E
 OHIO STATE UNIVERSITY MAIN CAMPUS S
 OKLAHOMA STATE UNIVERSITY MAIN CAI
 OREGON STATE UNIVERSITY Statistics
 PENN STATE UNIVERSITY Statistics
 PURDUE UNIVERSITY MAIN CAMPUS Statis
 RICE UNIVERSITY Statistics
 RUTGERS THE STATE UNIVERSITY OF NEW
 SOUTHERN METHODIST UNIVERSITY Stati
 STANFORD UNIVERSITY Statistics
 TEMPLE UNIVERSITY Statistics
 TEXAS A & M UNIVERSITY Statistics
 UNIVERSITY OF CALIFORNIA-BERKELEY St
 UNIVERSITY OF CALIFORNIA-DAVIS Statis
 UNIVERSITY OF CALIFORNIA-LOS ANGELE
 UNIVERSITY OF CALIFORNIA-RIVERSIDE A
 UNIVERSITY OF CALIFORNIA-SANTA BAR
 UNIVERSITY OF CHICAGO Statistics
 UNIVERSITY OF CONNECTICUT Statistics
 UNIVERSITY OF FLORIDA Statistics
 UNIVERSITY OF GEORGIA Statistics

INTERACTIVE GRAPHICS



cranvas not portable!



More detailed wish list for research activity are
at [**https://github.com/dcook/useR.2015**](https://github.com/dcook/useR.2015)

(Keynote talk at useR! 2015 Denmark)



Group started discussing more complete,
portable system options at [**https://github.com/
ropenscilabs/rivis**](https://github.com/ropenscilabs/rivis)

- keen to get people involved,
open discussion, no real leader, but support for anyone
interested in the area.



Why get involved?

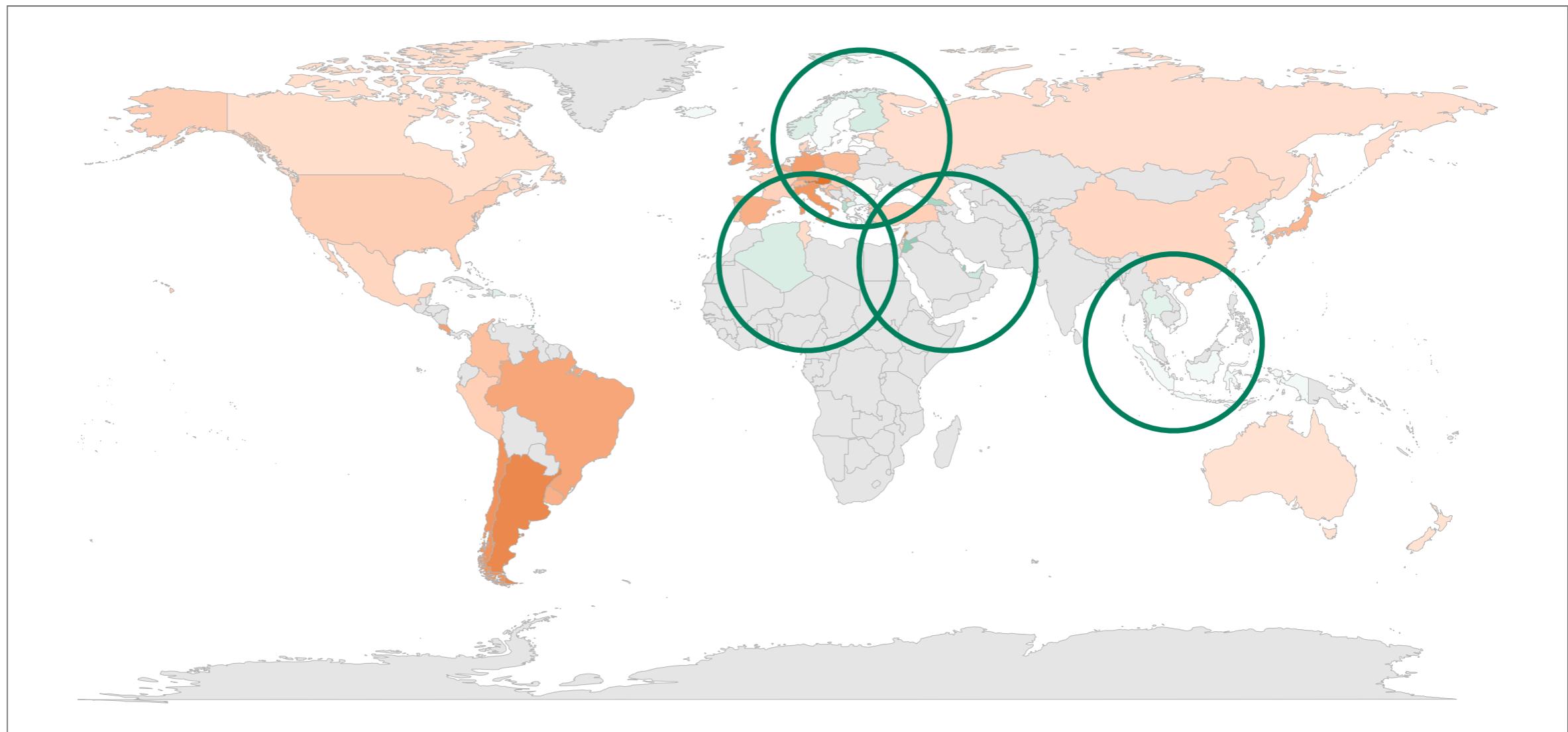


Myth busting

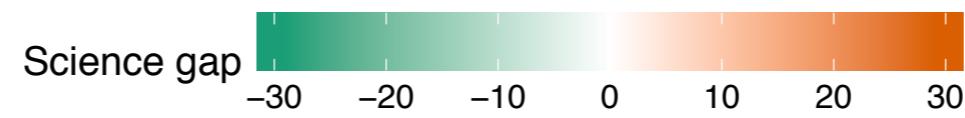
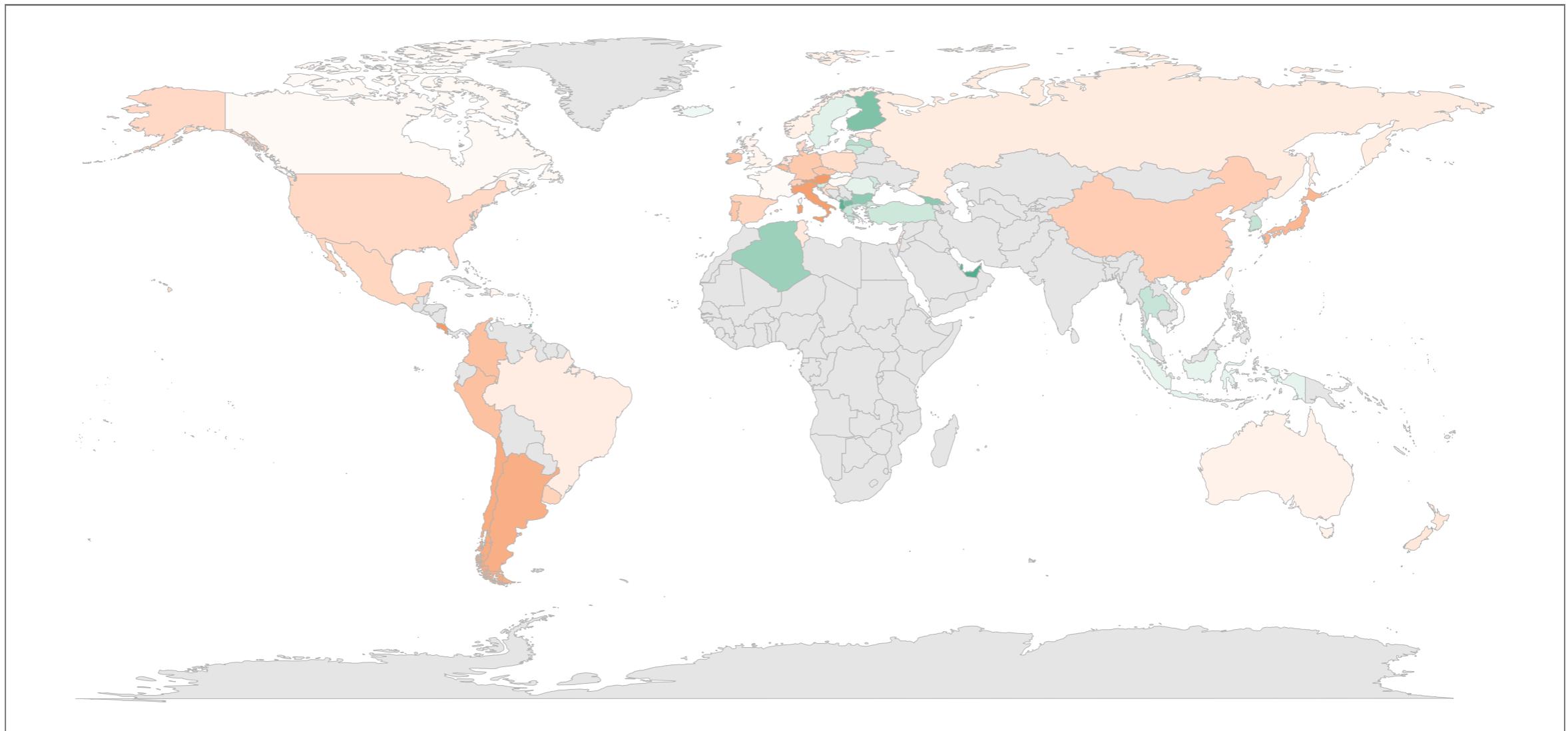


gender/diversity (in)balance

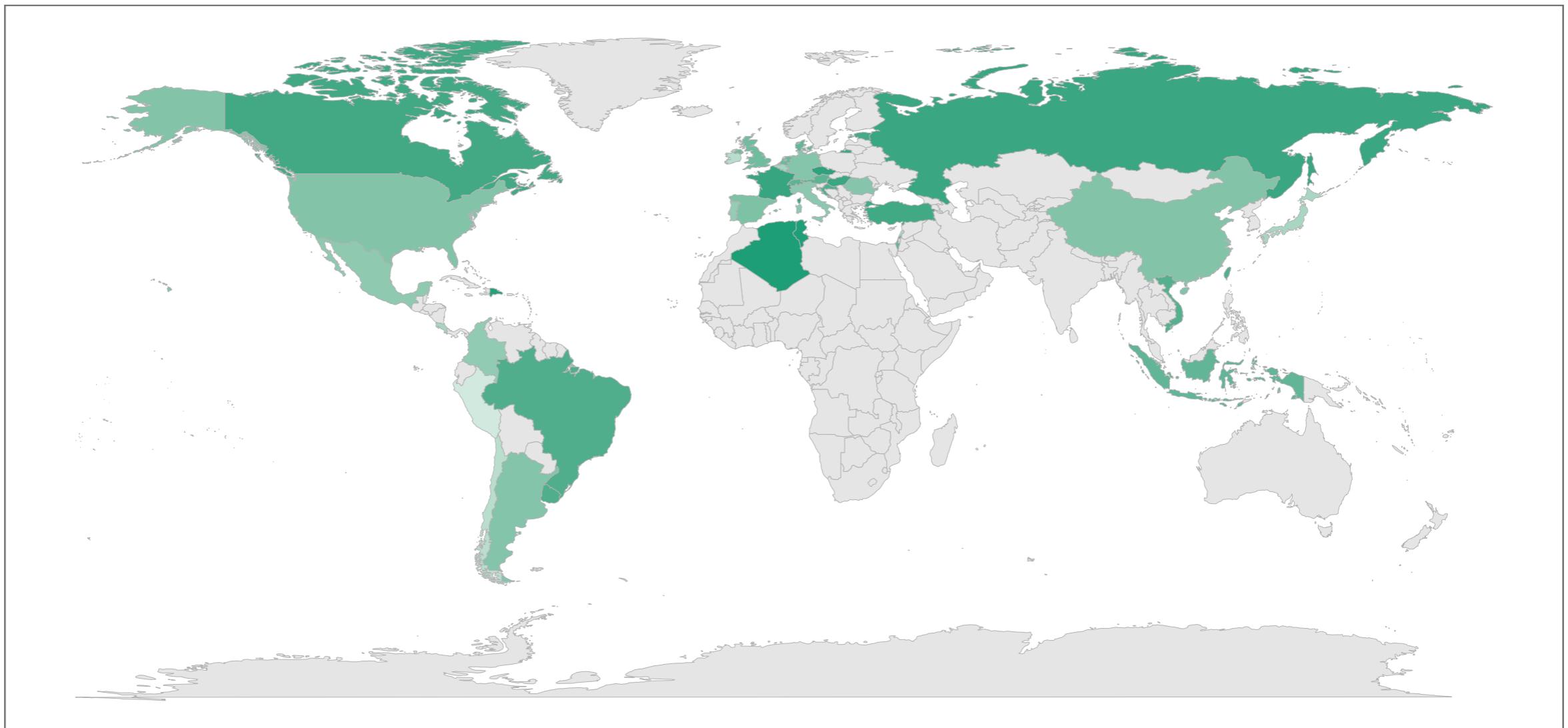
PISA - MATH

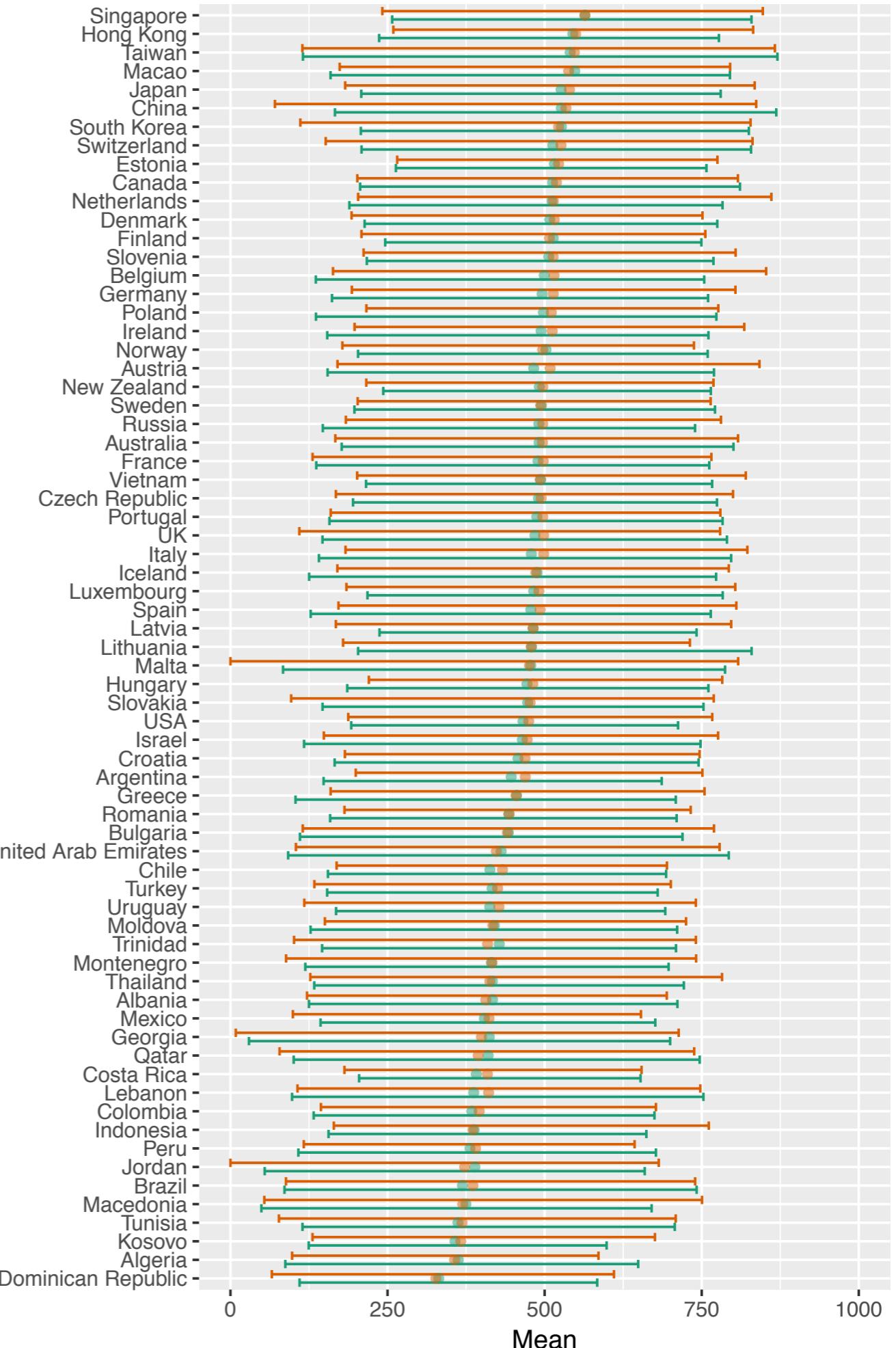
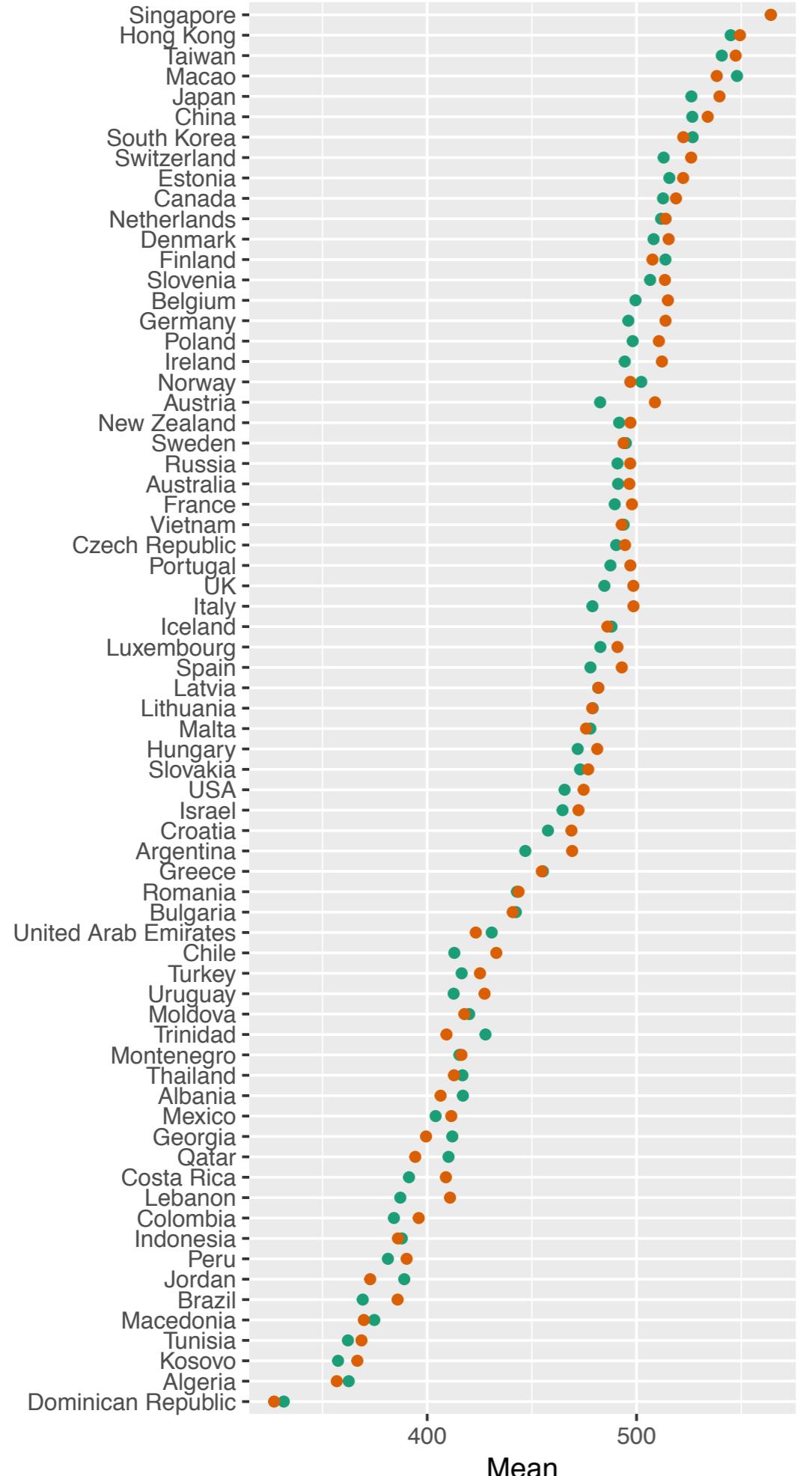


PISA - SCIENCE



PISA - READING





You as an individual
matter more than your
demographic!

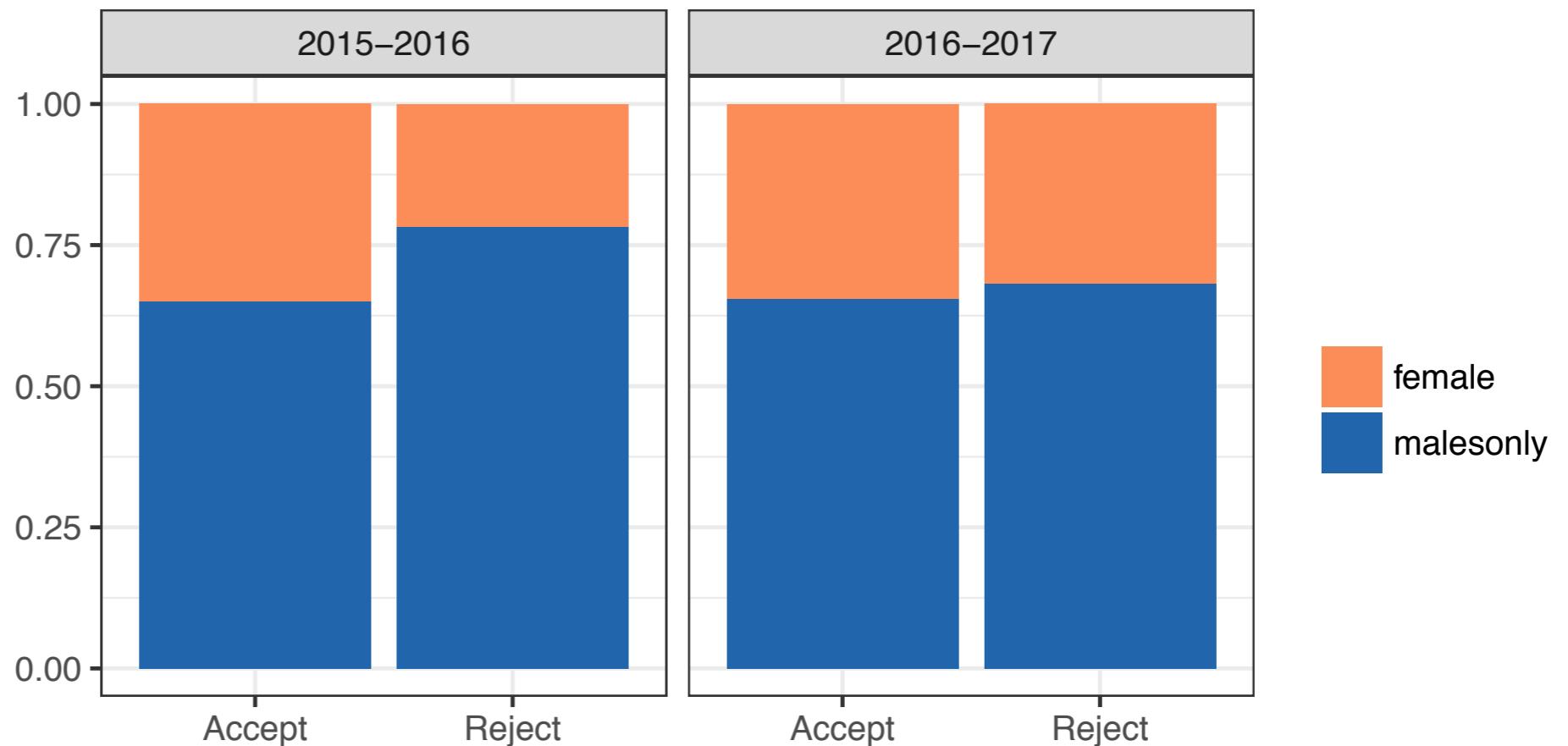
The perception of ACTUAL
gender difference is
magnified way out of all
proportion, EXCEPT FOR
READING!

PUBLISHING



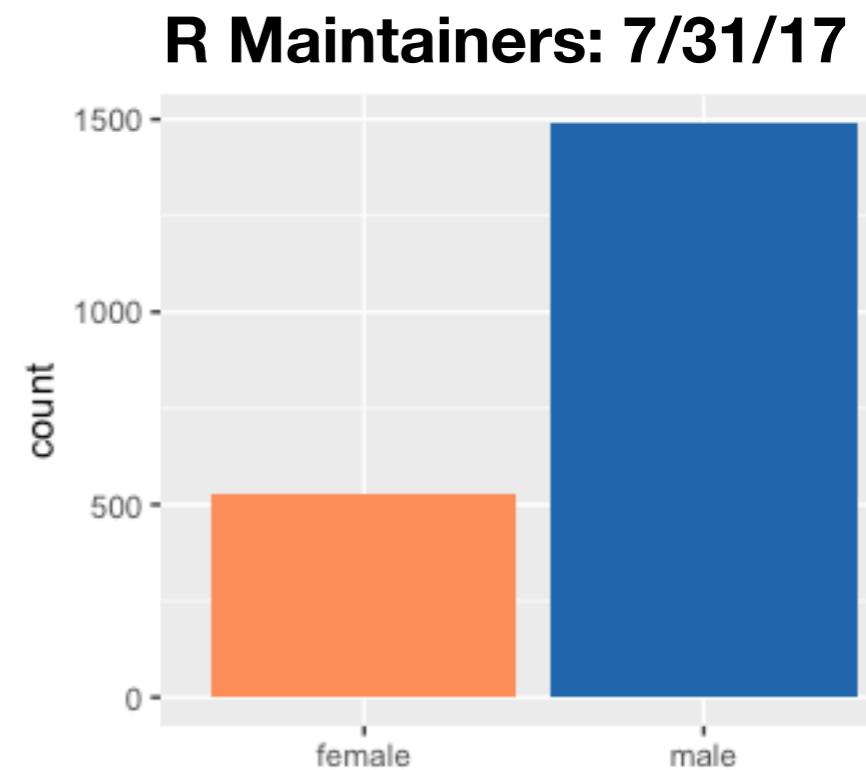
Proportion of papers with at least one female author in JCGS (our flagship journal) is low

Year	Prop female
2015-2016	0.25
2016-2017	0.33

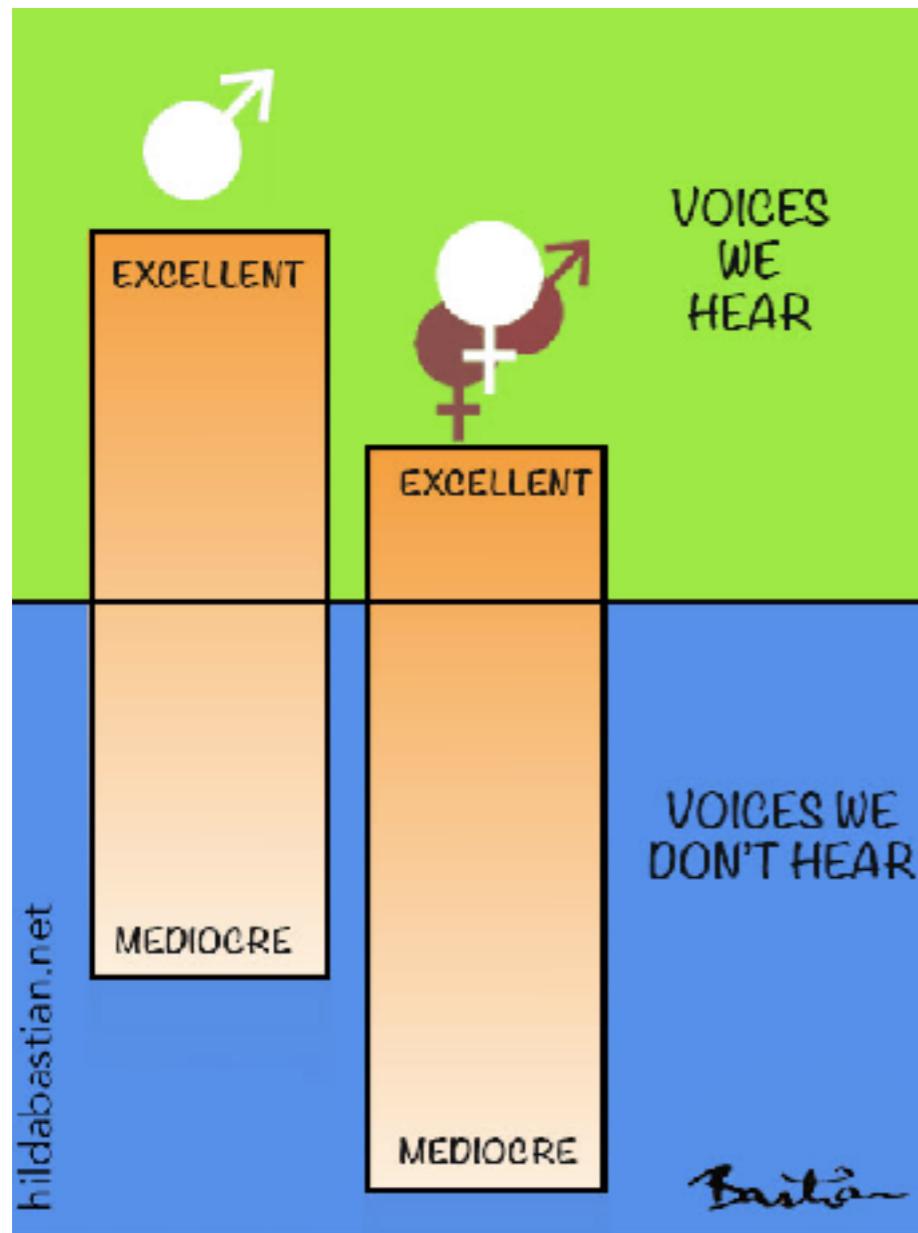


SOFTWARE

- Proportion of female R package authors and maintainers is low
- 2010 survey of 1087 maintainers of packages on CRAN/Bioconductor, 9% of package authors were women
- Mar 2016 genderizer analysis of R package maintainers, 11% women
- Jul 31 2017 genderizer analysis of R package maintainers, 26% women!!!!



ABSOLUTELY MAYBE



LACK OF DIVERSITY CAN BE A
GREAT WAY TO OVER-SAMPLE
FOR MEDIOCRITY!

Hilda Bastian's blog

This cartoon comes from Hull 22 post "Why Pockets and Waves of De-Feminization in Science's Past Matter Now"

 Get published

 Get research funds

 Get promoted

 You are not alone

SUMMARY

- ▀ Data visualisation is a hot topic
- ▀ Lots of open problems in interactive graphics, and in thinking of plots at another type of statistic in order to do inference
- ▀ Great opportunities to make big contributions in data science
- ▀ Lots of room to improve gender imbalance in academic research, and in producing computer software



CONTACT DETAILS

 email: dicook@monash.edu

 github: dicook

 twitter: @visnurt

 Slides available at github.com/dicook/JSM17

THANKS FOR COMING!

