EDA: A Historical Perspective and a Path Forward

Di Cook
Econometrics and Business Statistics
Monash University

http://www.dicook.org/files/JSM18_slides.pdf

Outline

- What is EDA? What are the origins. How is it related to data science?

- The role of interactive graphics in EDA.

- The emergence of reproducible research process, principles and tools.

- Can we really capture the full analysis process?

What is EDA?

“Playing in the sand with your data.”

“Relax the focus on the problem statement and explore broadly different aspects of the data. Modern exploratory data analysis software is designed to make this process as fruitful as possible. It is a highly interactive, real-time, dynamic, and visual process, having evolved along with computers.”

Cook & Swayne 2007 - and Buja

Origins of EDA

“Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.”

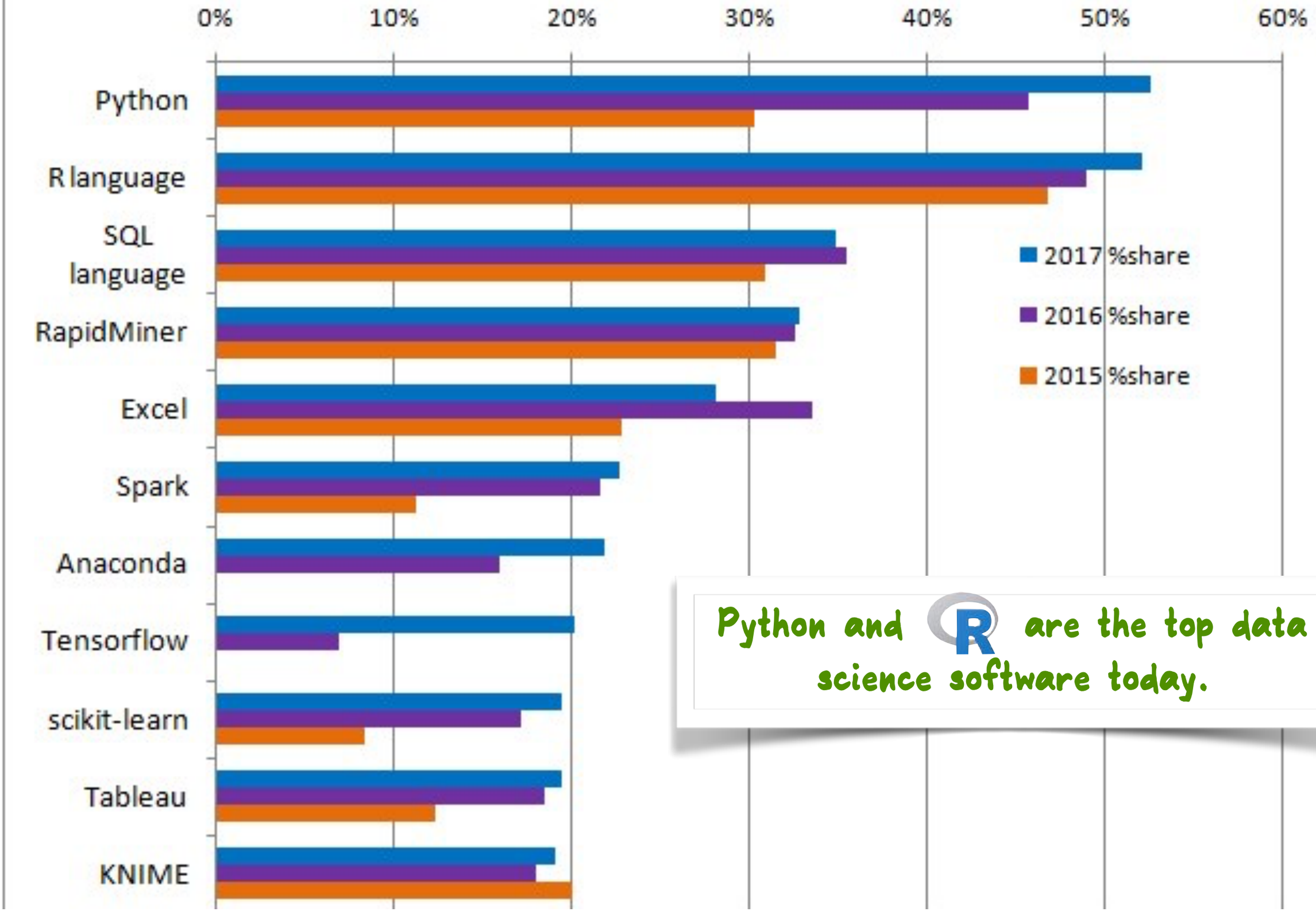
wikipedia

Tukey's championing of EDA encouraged the development of statistical computing packages, especially S at Bell Labs. The S programming language inspired the systems 'S'-PLUS and R.

wikipedia

KDnuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017

http://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html



Python and R are the top data science software today.

EDA within Statistics field

My classic example

- **When I first began teaching at Iowa State University, I was desperately looking for data that I could use for examples in class.**
- These two books **contain data, AND their stories**
  - Andrews and Herzeberg (1985) "***Data: A Collection of Problems from Many Fields for the Student and Research Worker***" - you can find the iris data here
  - Bryant, P. G. and Smith, M. A. (1995) "***Practical Data Analysis: Case Studies in Business Statistics***" for a new to USA resident this data jumped out: **RESTAURANT TIPPING**

Restaurant tipping

- Case study: What factors affect tip rate?
- Make a regression model

```
Tip ~ Total bill + Size of Party + Gender of
billpayer + Smoking section + Day of the
Week + Time of Day
```

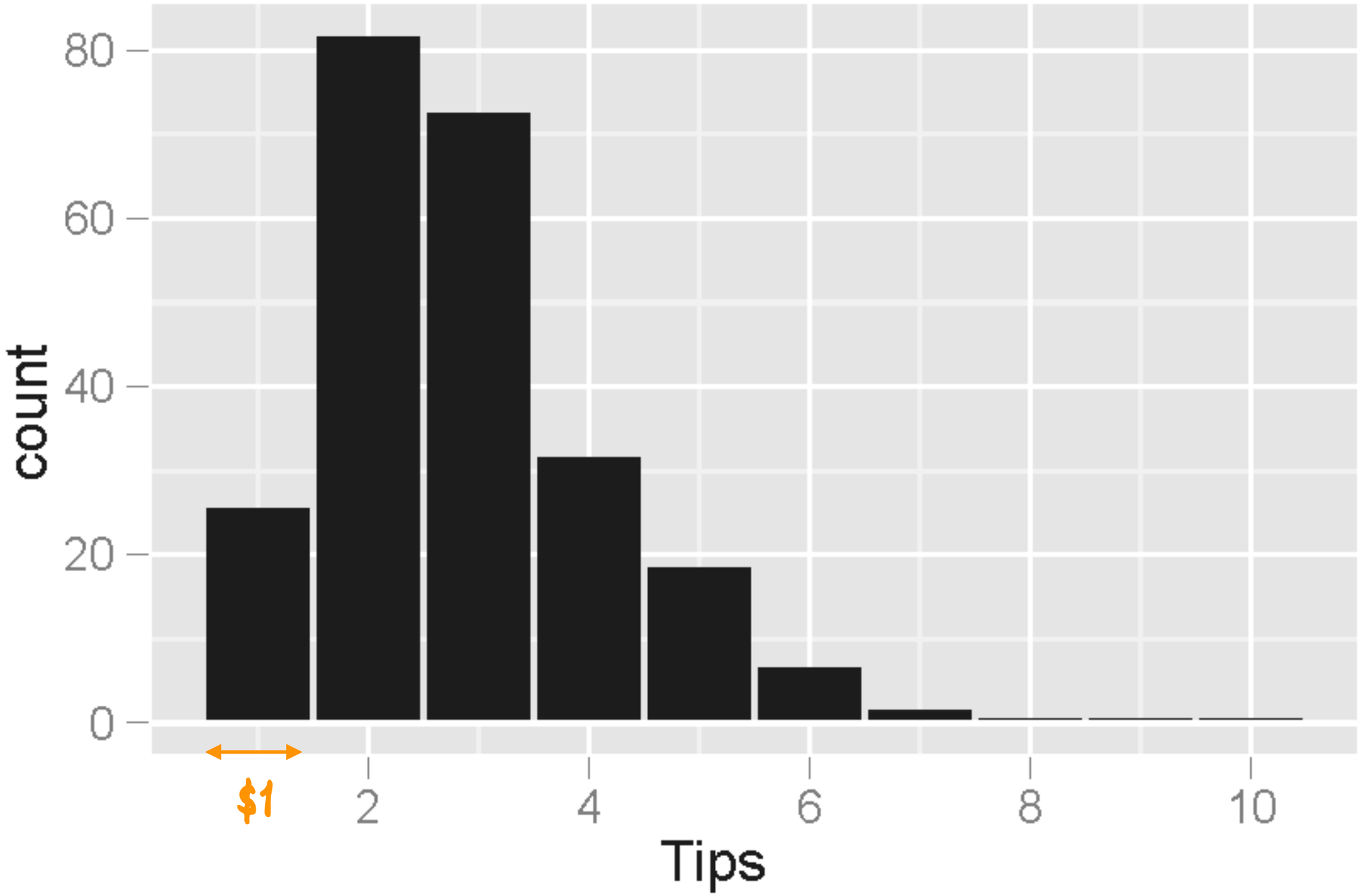
- The effective result is

```
Tip Rate ~ 0.18 + 0.01 x Size of Party
R² = 0.02
```

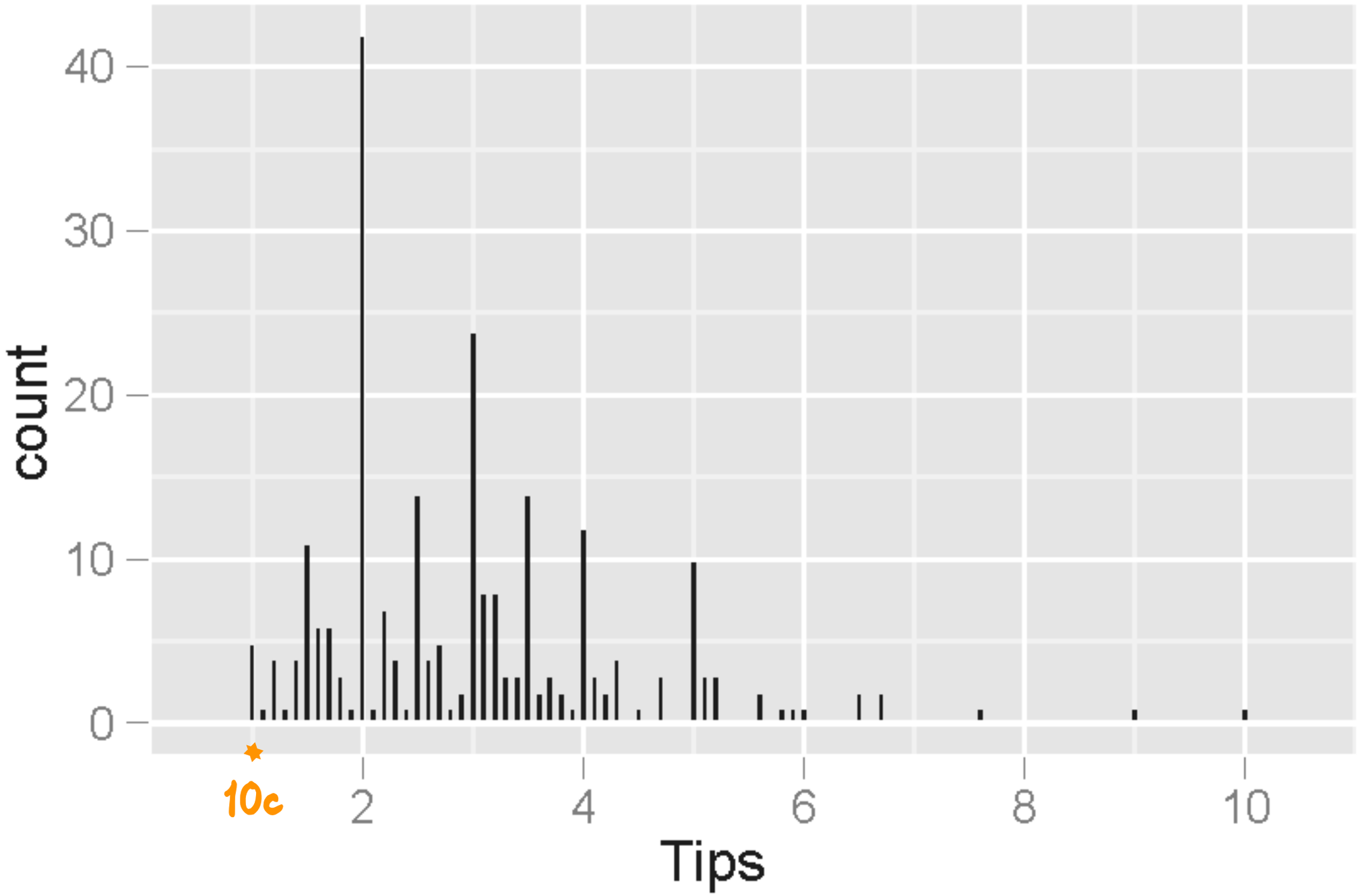
I never just run a regression model. I always look at the data with plots.

I never just run a regression model. I always look at the data with plots.

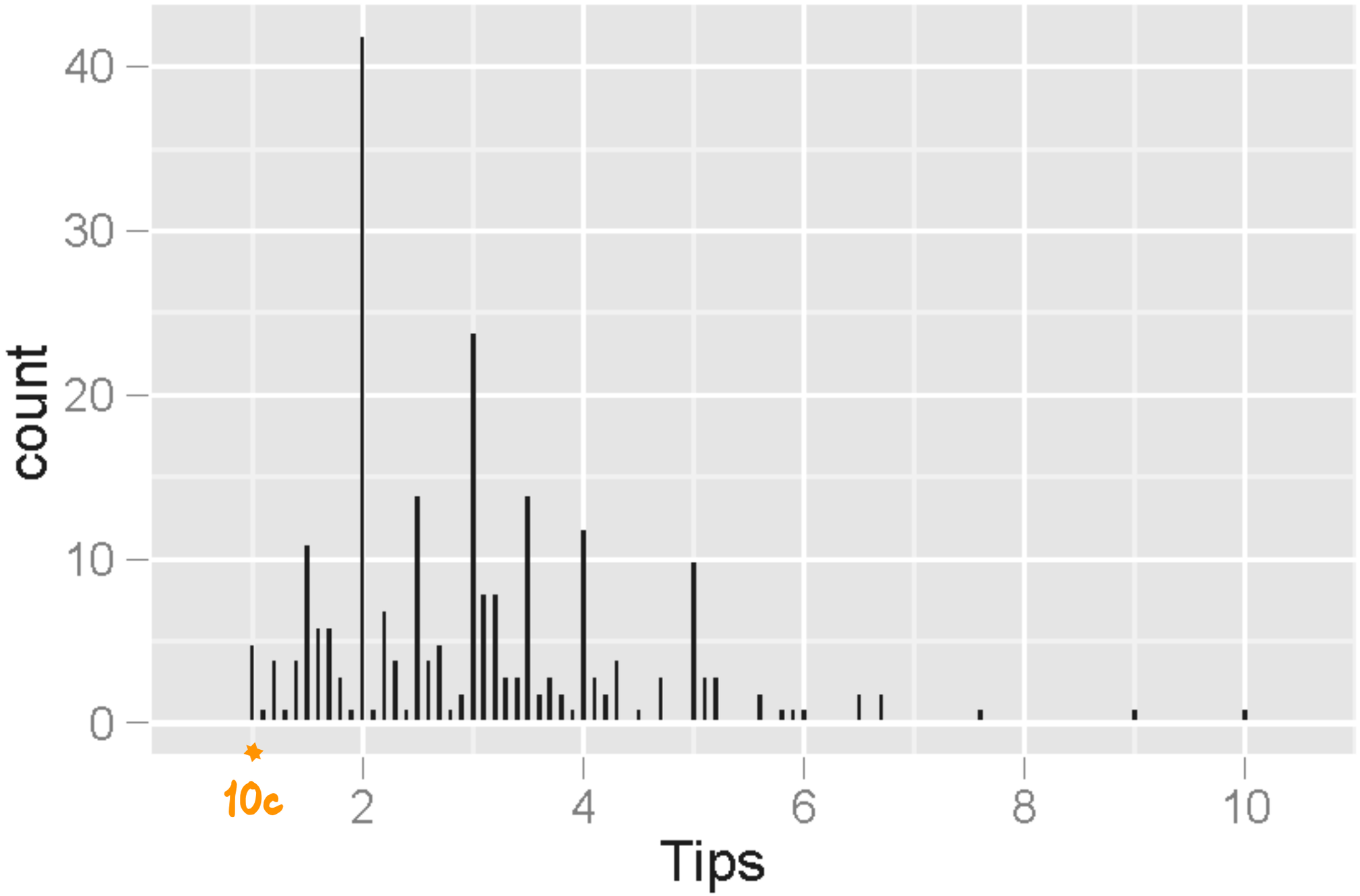
EDA of tipping data ->



Skewed right distribution: more people tip low than high

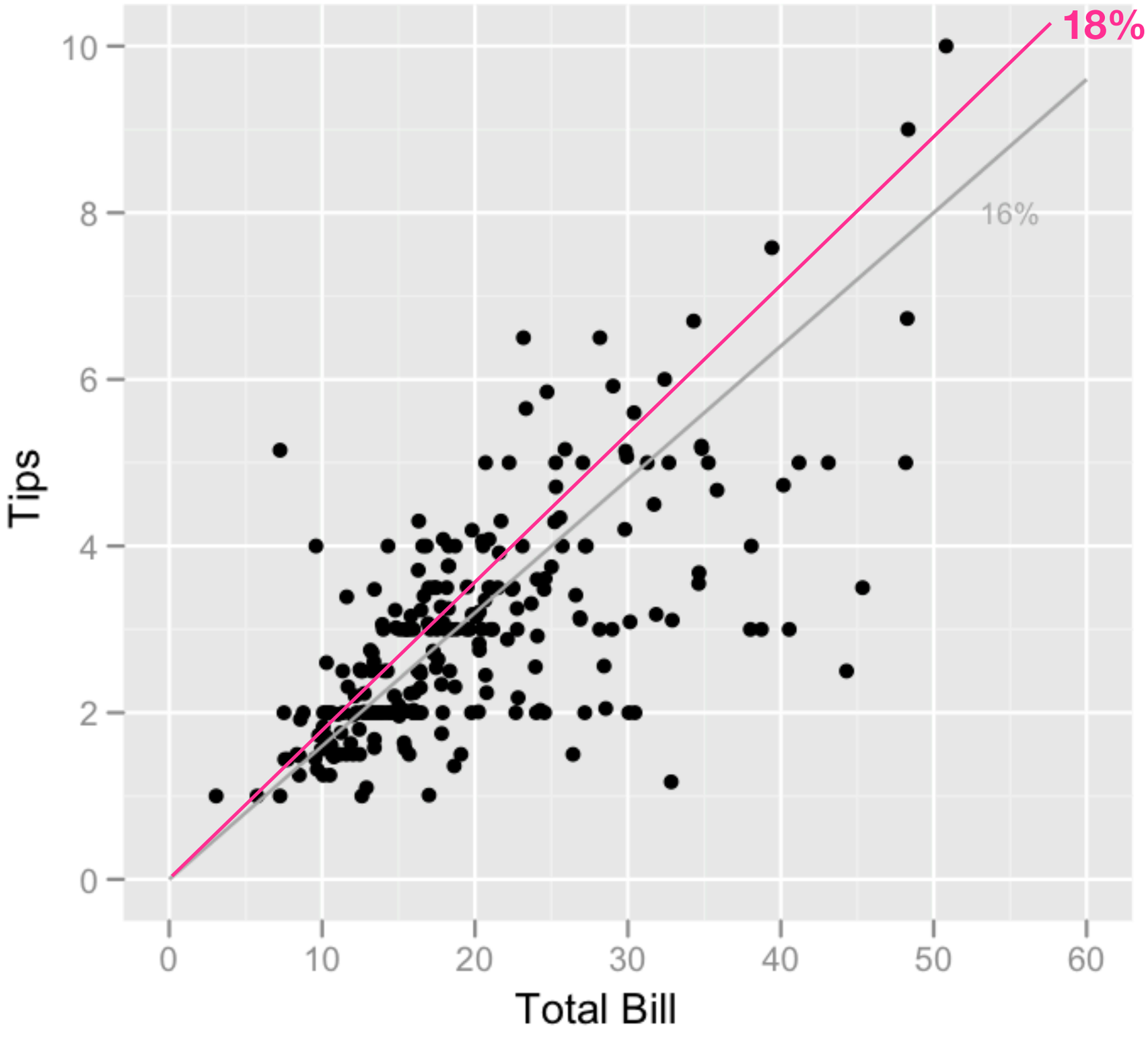


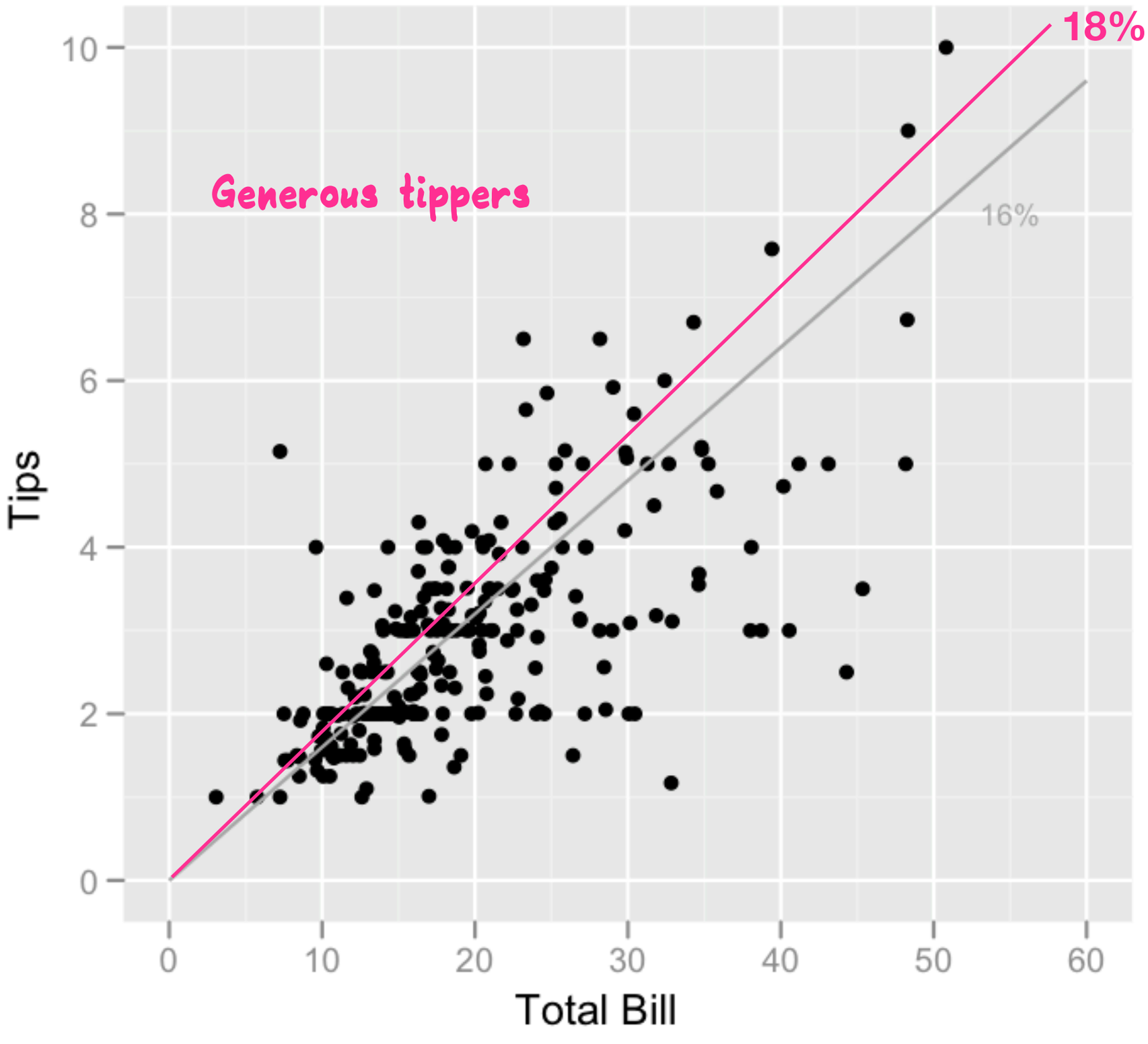
Skewed right distribution: more people tip low than high

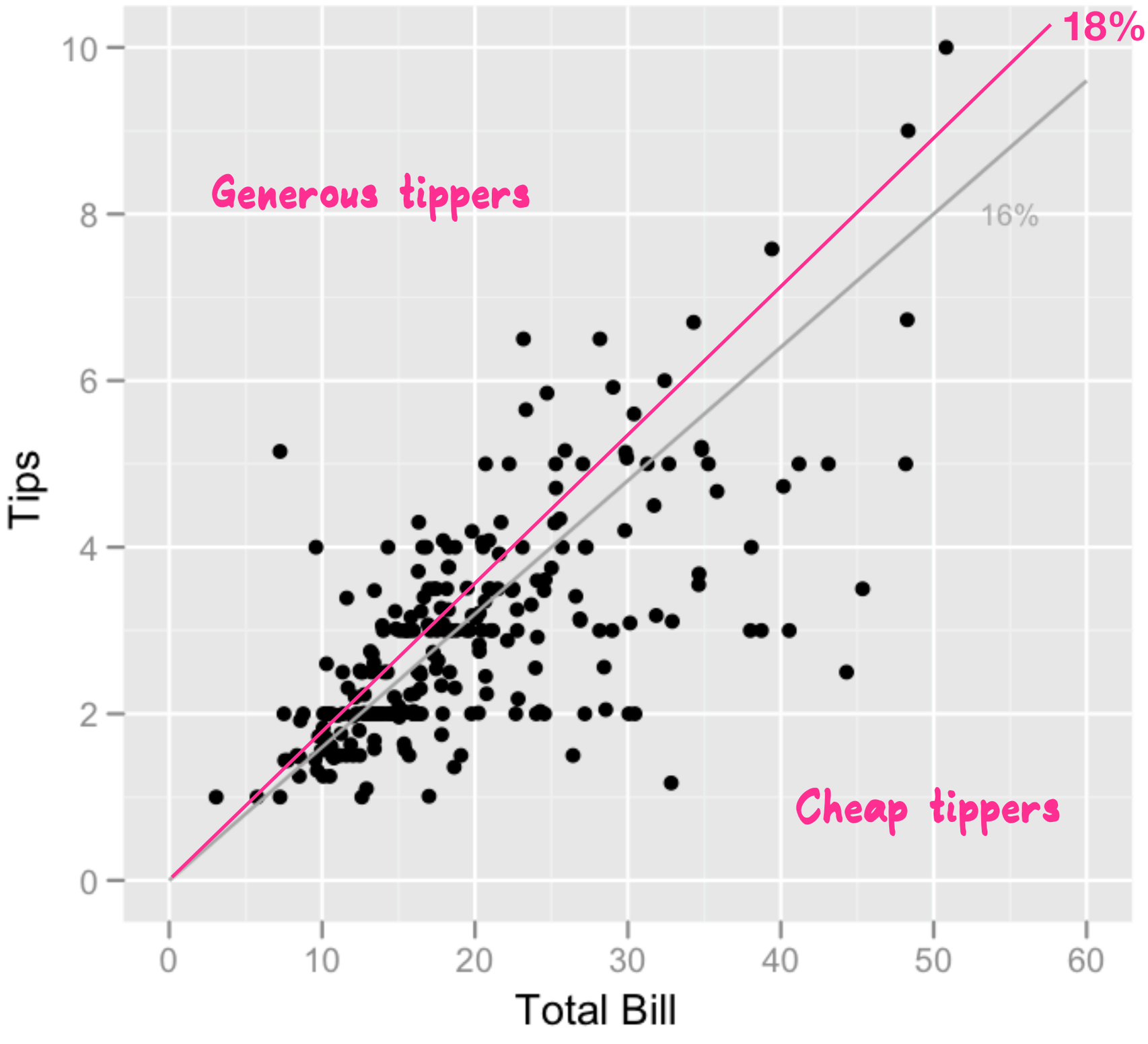


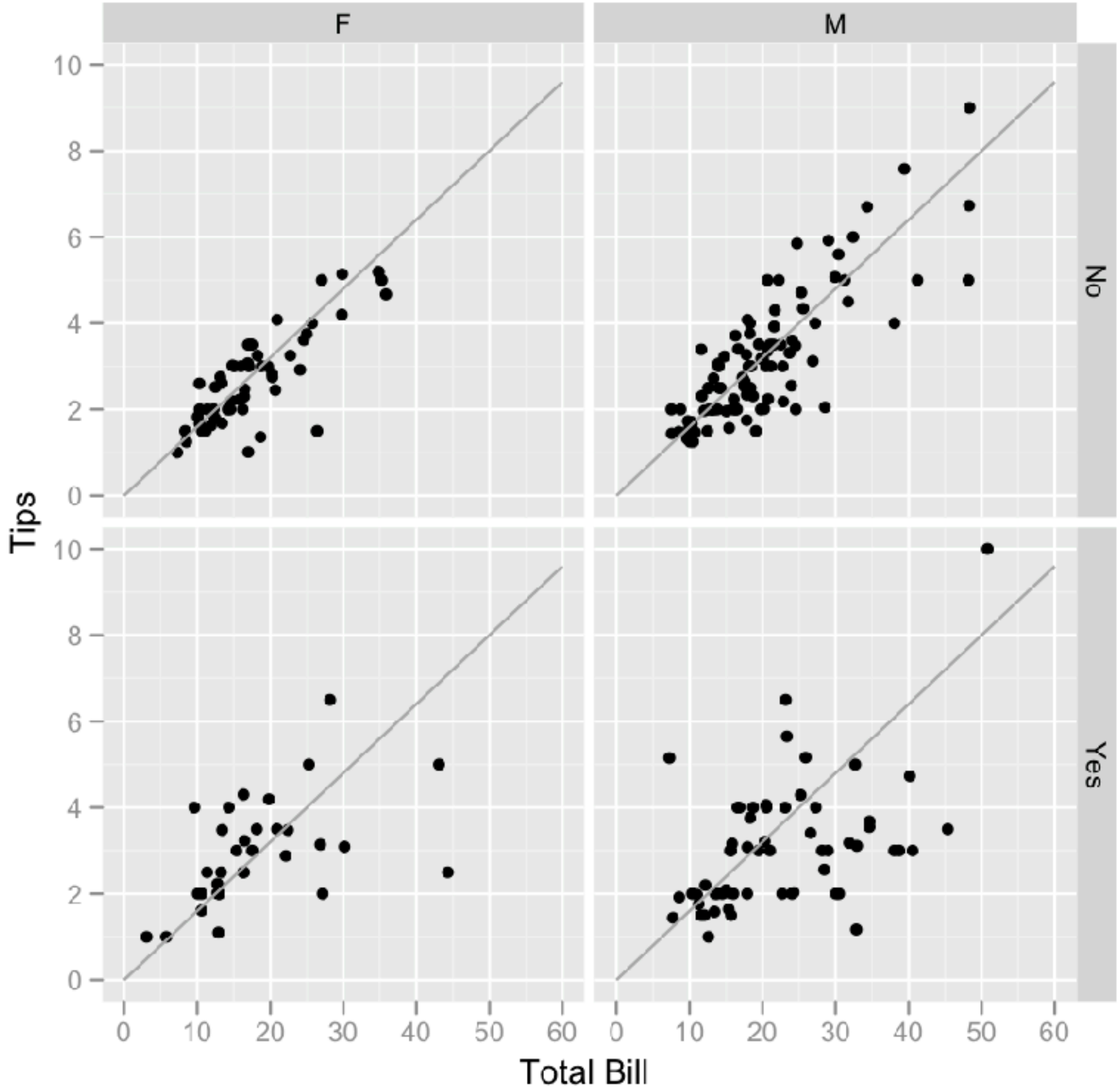
Skewed right distribution: more people tip low than high

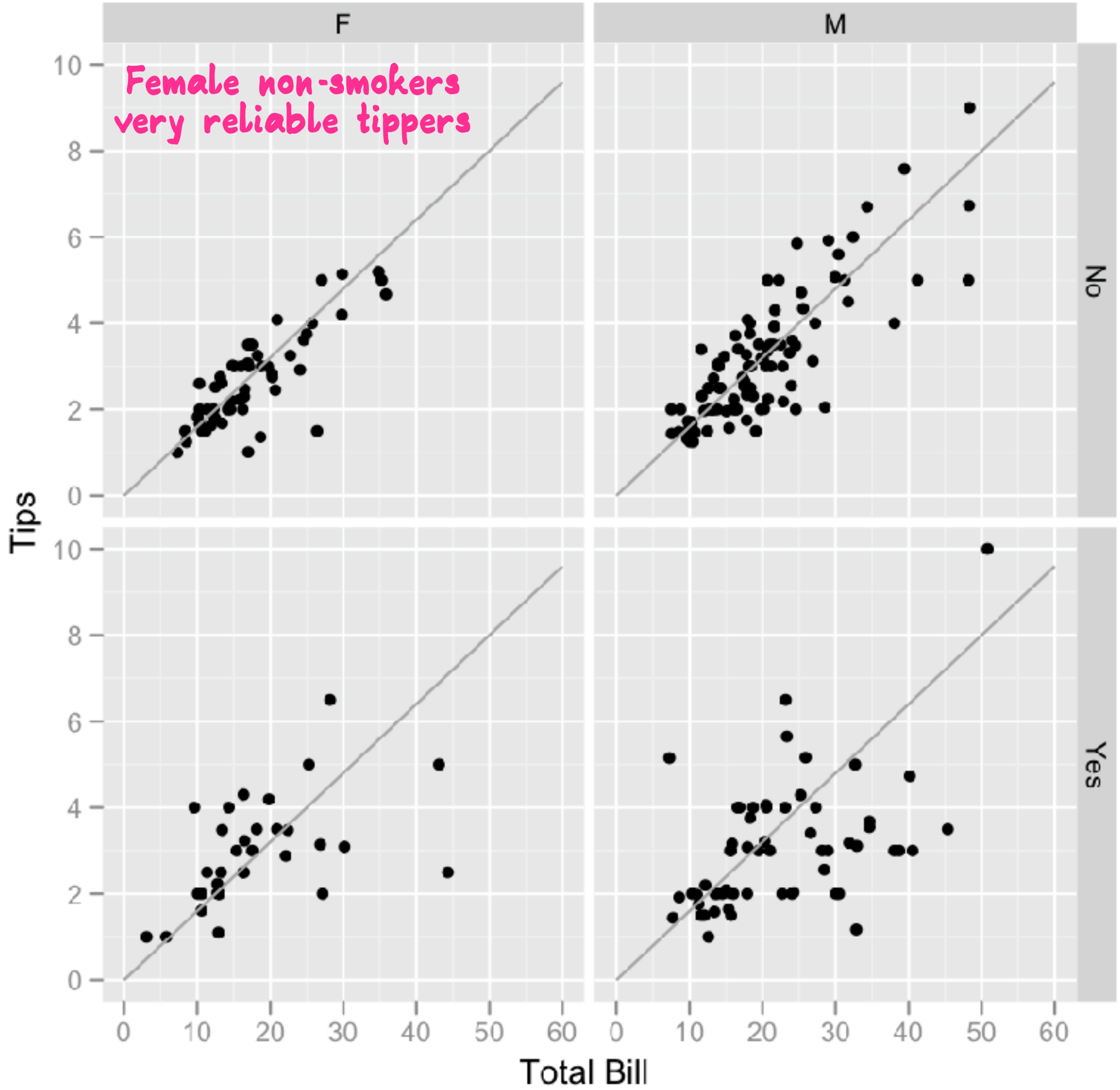
Multimodal - Regular peaks: people tend to round their tips to full $ or half-dollar amounts

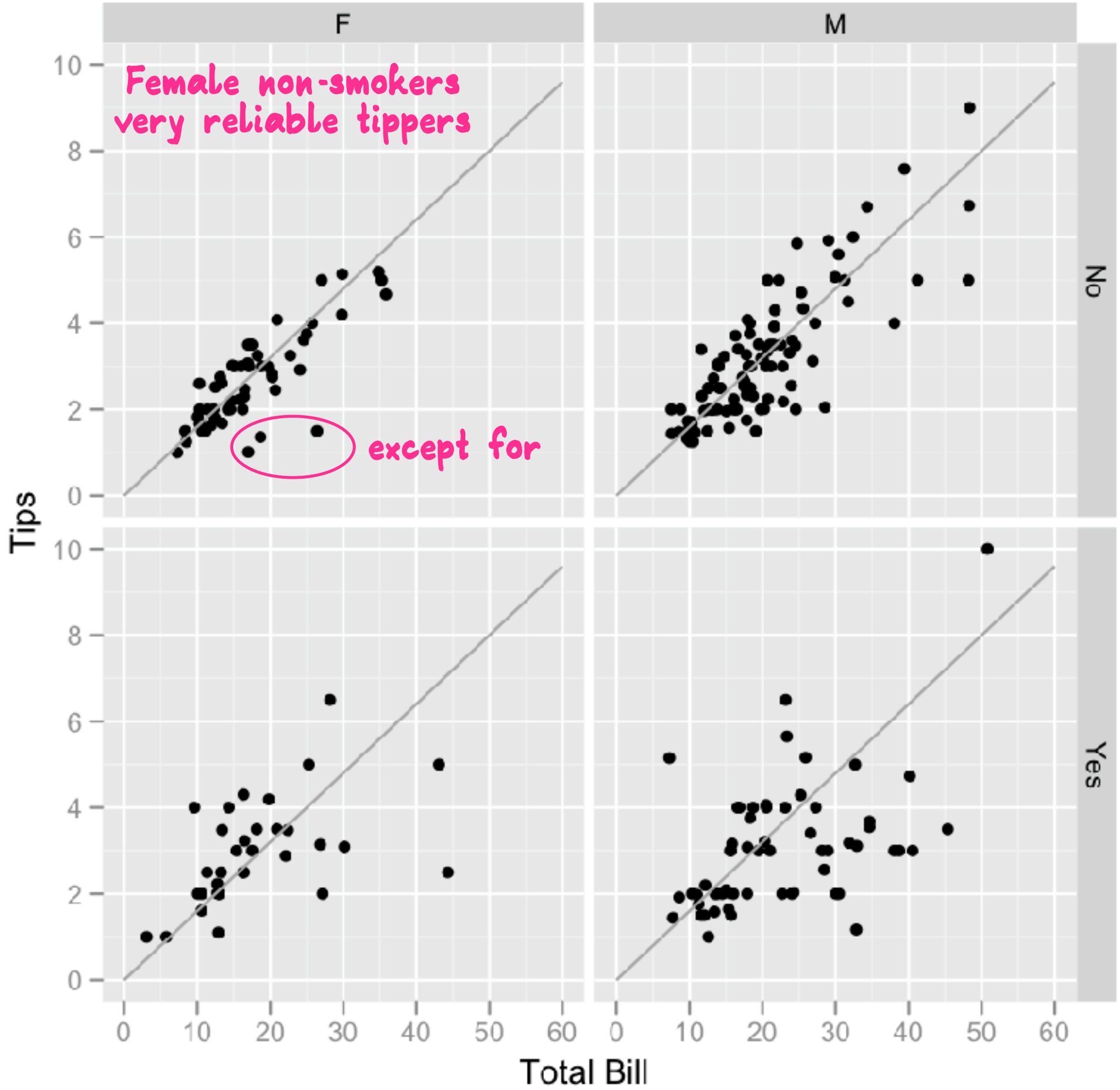

18%
16%
10
8
6
4
2
0
Tips
0
10
20
30
40
50
60
Total Bill

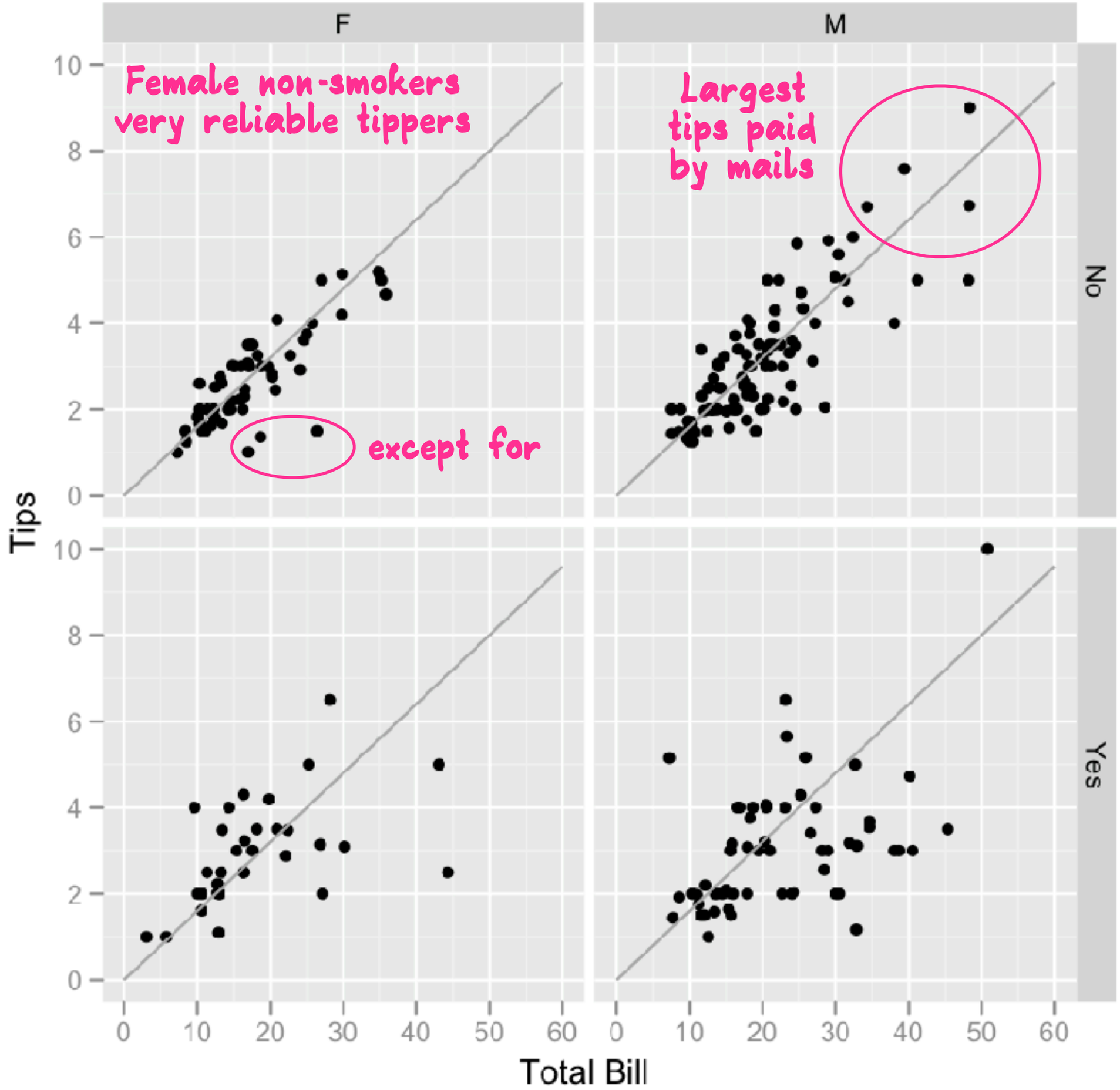

18%
Generous tippers
16%
10
8
6
4
2
0
Tips
0
10
20
30
40
50
60
Total Bill

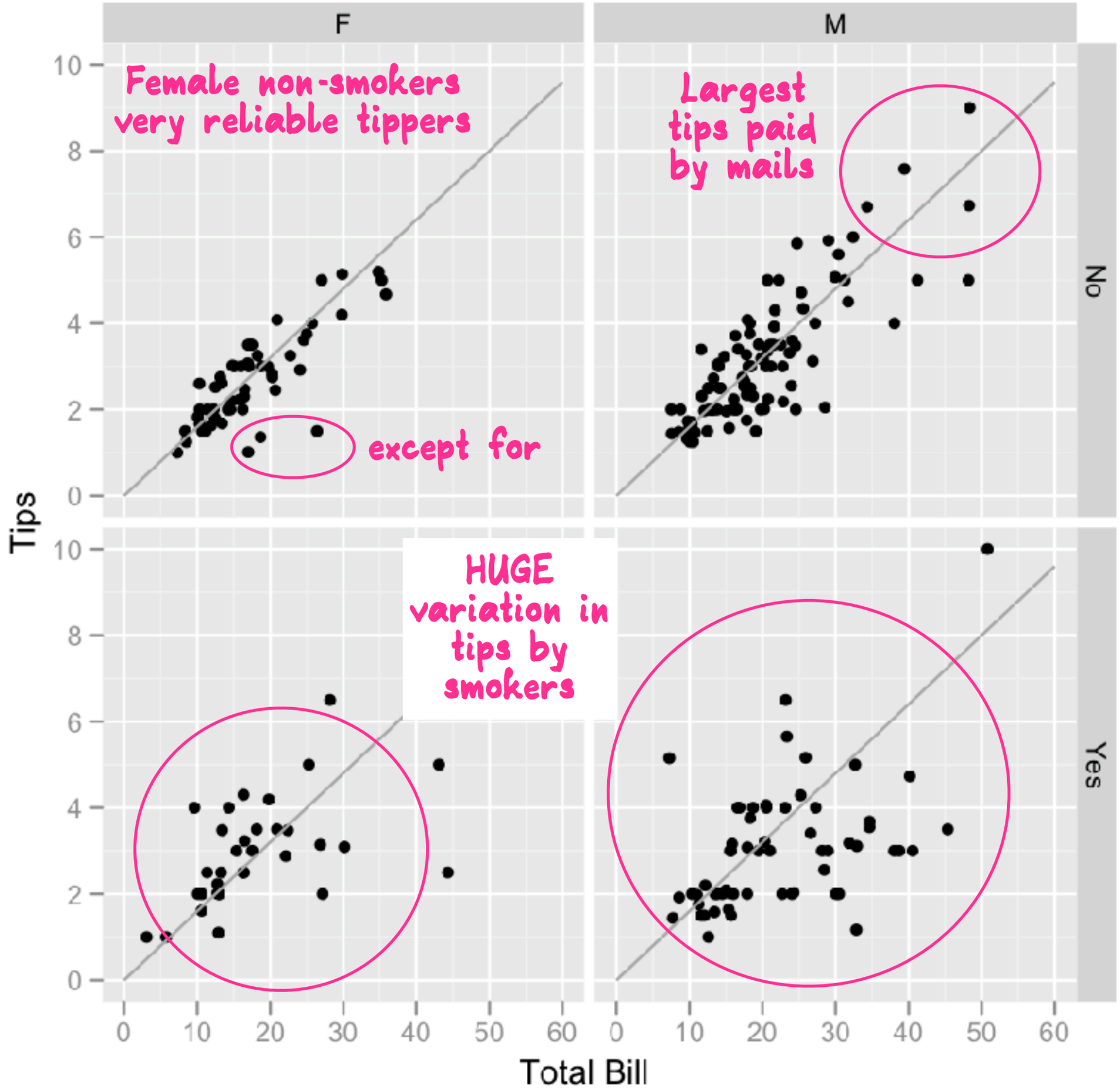

18%
16%
Generous tippers
Cheap tippers
Tips
Total Bill
0
2
4
6
8
10
0
10
20
30
40
50
60

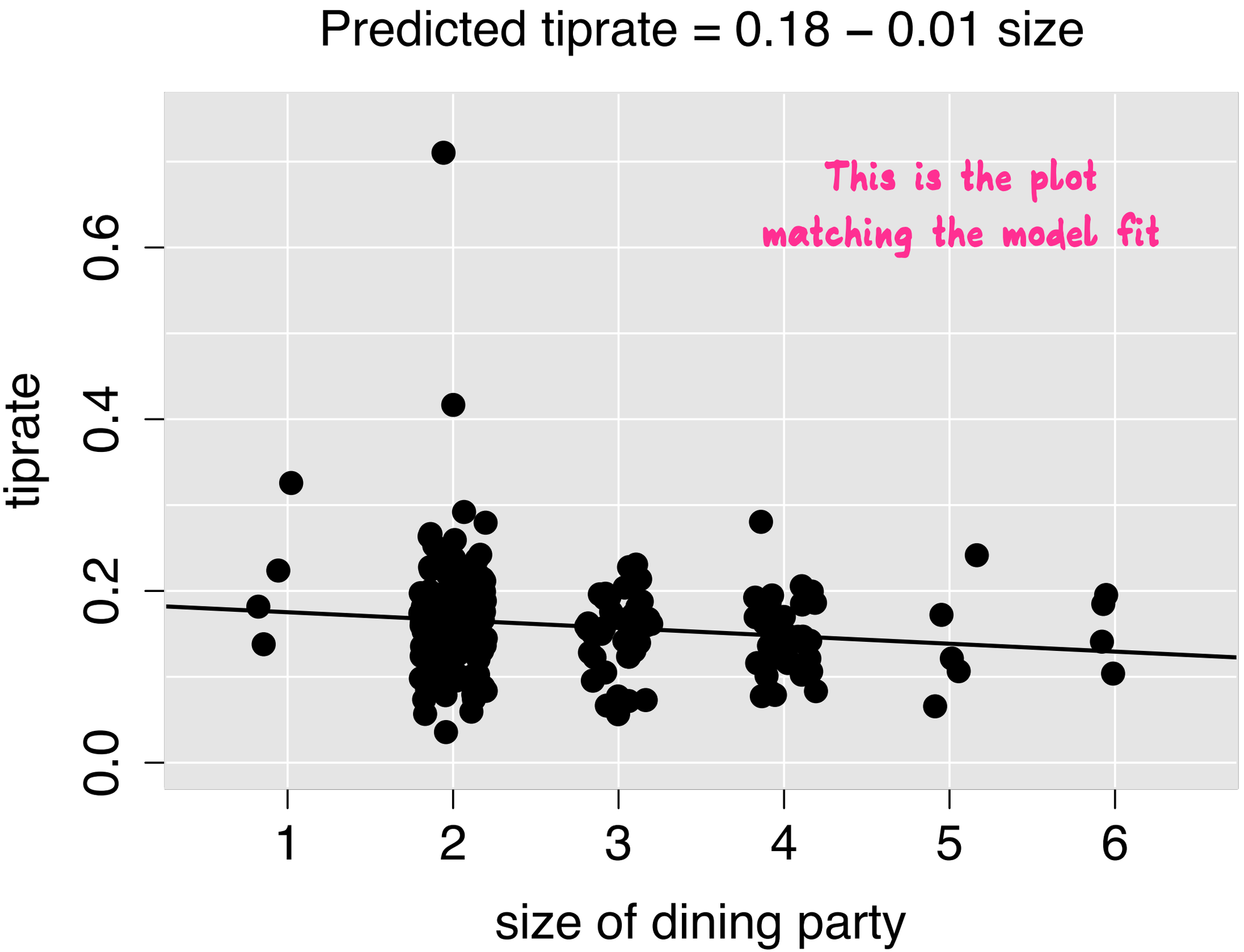

F
M
No
Yes
Tip
Total Bill
0
2
4
6
8
10
0
10
20
30
40
50
60


F
M
No
Yes
Female non-smokers
very reliable tippers
Tip
Total Bill
0
2
4
6
8
10
0
10
20
30
40
50
60


F
M
No
Yes
Female non-smokers
very reliable tippers
except for
Tip
Total Bill
0
2
4
6
8
10
0
10
20
30
40
50
60


F
M
No
Yes
Female non-smokers
very reliable tippers
except for
Largest
tips paid
by mails
Tip
Total Bill
0
2
4
6
8
10
0
10
20
30
40
50
60


F
M
No
Yes
Female non-smokers
very reliable tippers
except for
Largest
tips paid
by mails
HUGE
variation in
tips by
smokers
Tips
Total Bill
0
2
4
6
8
10
0
10
20
30
40
50
60


Predicted tiprate = 0.18 – 0.01 size
This is the plot
matching the model fit
tiprate
0.0
0.2
0.4
0.6
1
2
3
4
5
6
size of dining party

EDA vs model

- Model is used as a hypothesis testing tool to find factors affecting tipping. It's a weak model.

- EDA using lots of plots, told me a lot more about tipping **behaviour.**

Sources: plots from wikipedia entry on EDA, written by me

Data is everywhere today

No longer need to rely of specialist compendiums of data. Data is all around you.

Data is everywhere today

No longer need to rely of specialist compendiums of data. Data is all around you.

Example: pedestrian sensors in Melbourne, see Earo Wang's Graphics Award talk 10:30 today, session 167

A problem for statistics

We are advising the world based on weak models which might work when you are trying to etch out a profit when margins are low, but completely back-fires when you talk about people. We are more alike than different and weak models exaggerate divisions.

A problem for statistics

We are advising the world based on weak models which might work when you are trying to etch out a profit when margins are low, but completely back-fires when you talk about people. We are more alike than different and weak models exaggerate divisions.

examples: math gap, c-section, ...

The role of interactive graphics in EDA

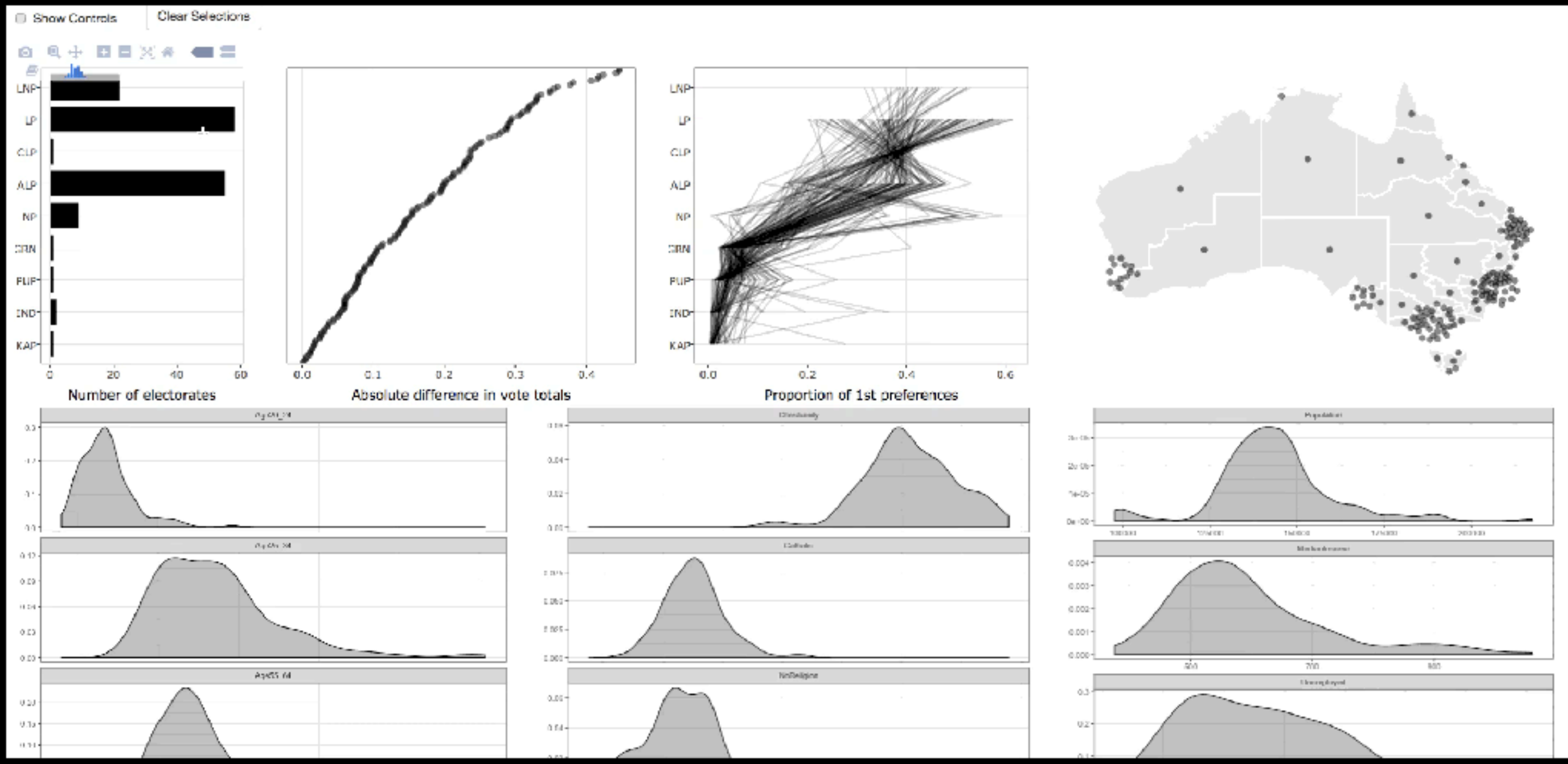
- Tukey's EDA **for consumption** was mostly pencil and paper techniques, like stem-and-leaf plots

- His **research** was on specialist graphics devices. Check out the ASA video lending library e.g. http://stat-graphics.org/movies/prim9.html

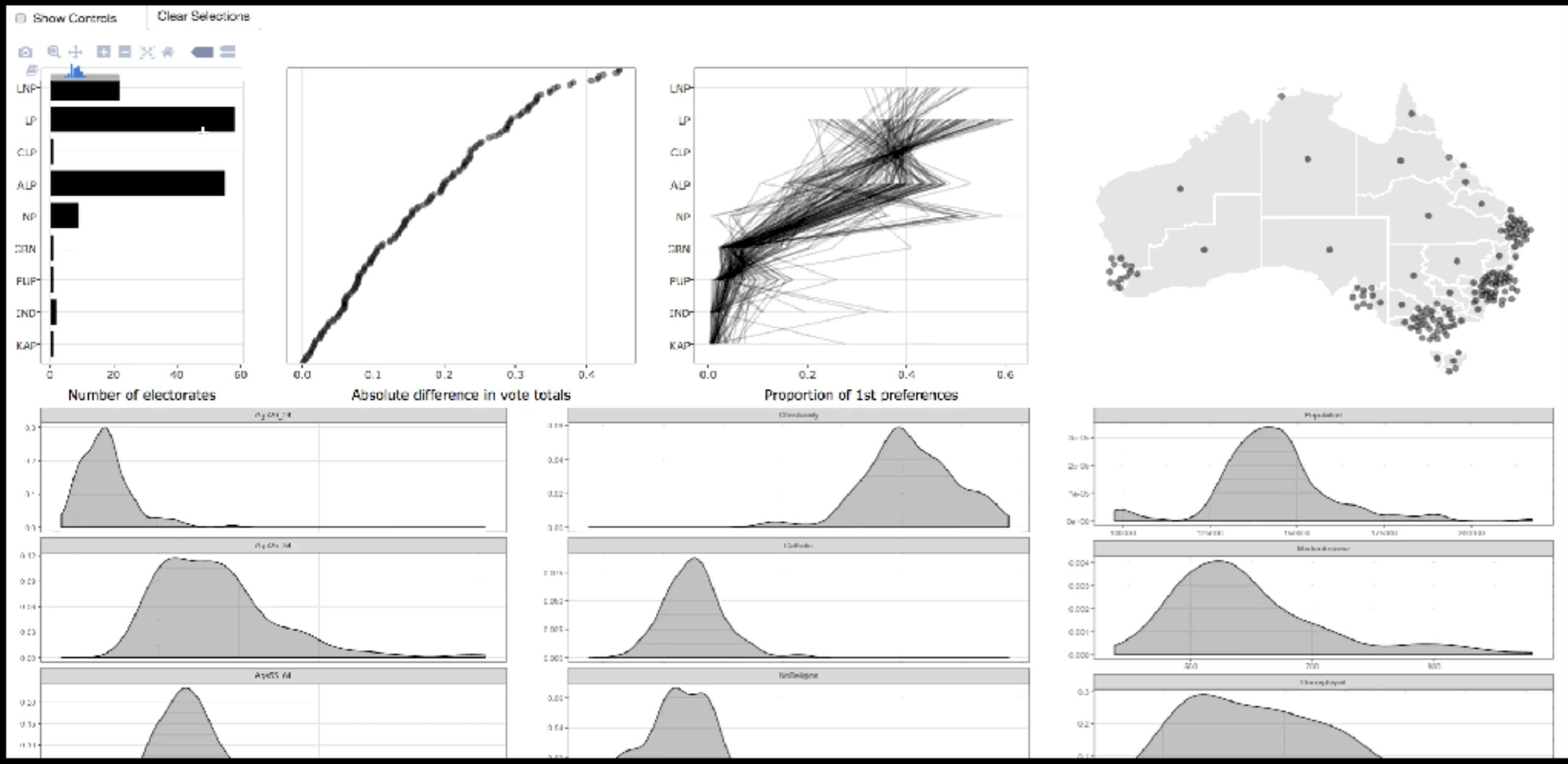
- We **all have access** to these specialist devices today, so we can all do highly interactive exploratory data analysis

What is interactive graphics?

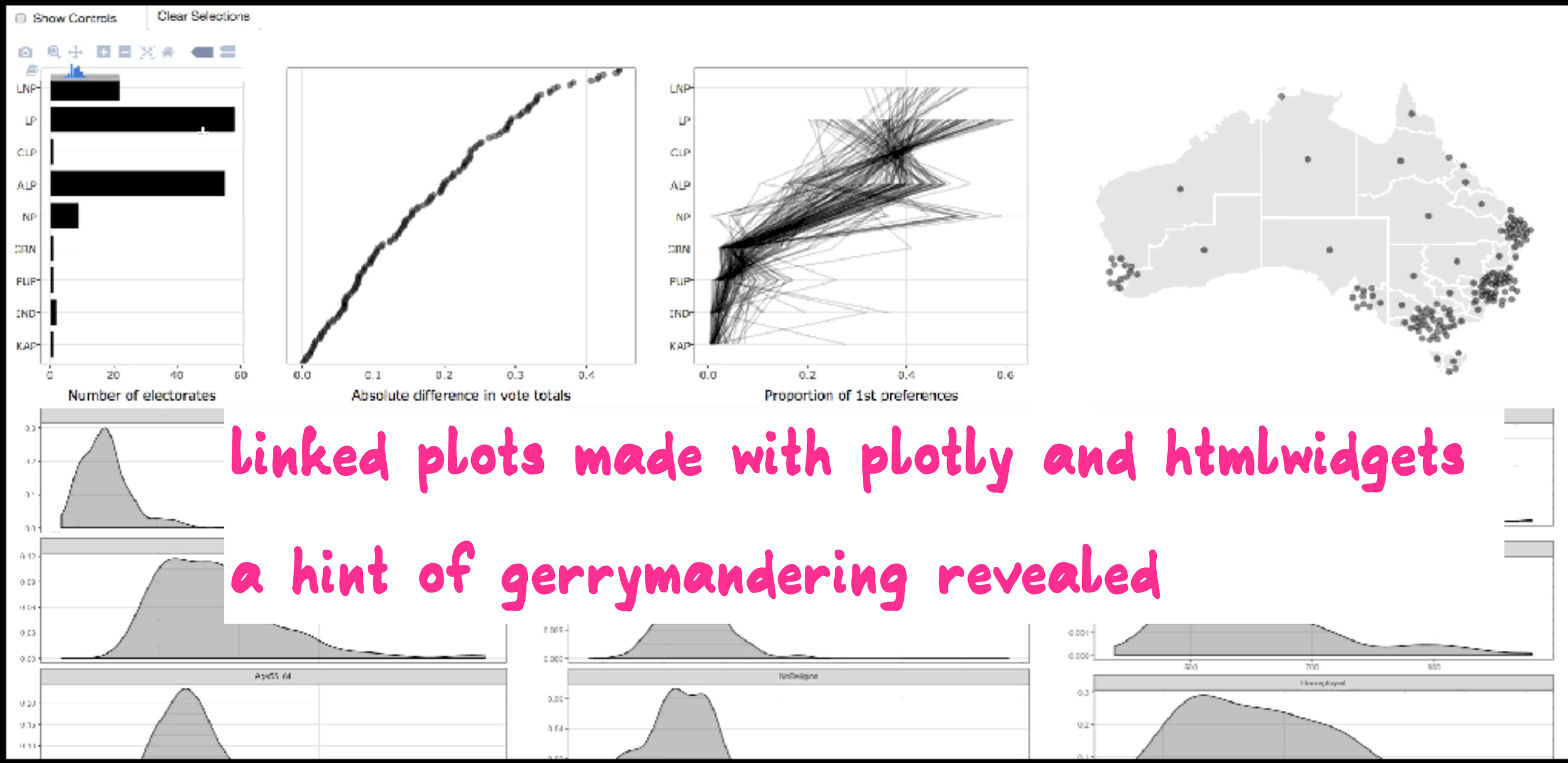
- Mouse-over - tooltips and labels EVERYONE HAS THIS
- Selection and and pan/zoom - ALMOST EVERYONE HAS THIS
- Linking between plots: highlighting and labelling - VERY FEW HAVE THESE
- Dynamic and controlled rotation of high-dimensional quantitative data - ALMOST NO-ONE HAS THIS



eechidna: Exploring Election and Census Highly Informative Data Nationally for Australia `https://github.com/ropenscilabs/eechidna`



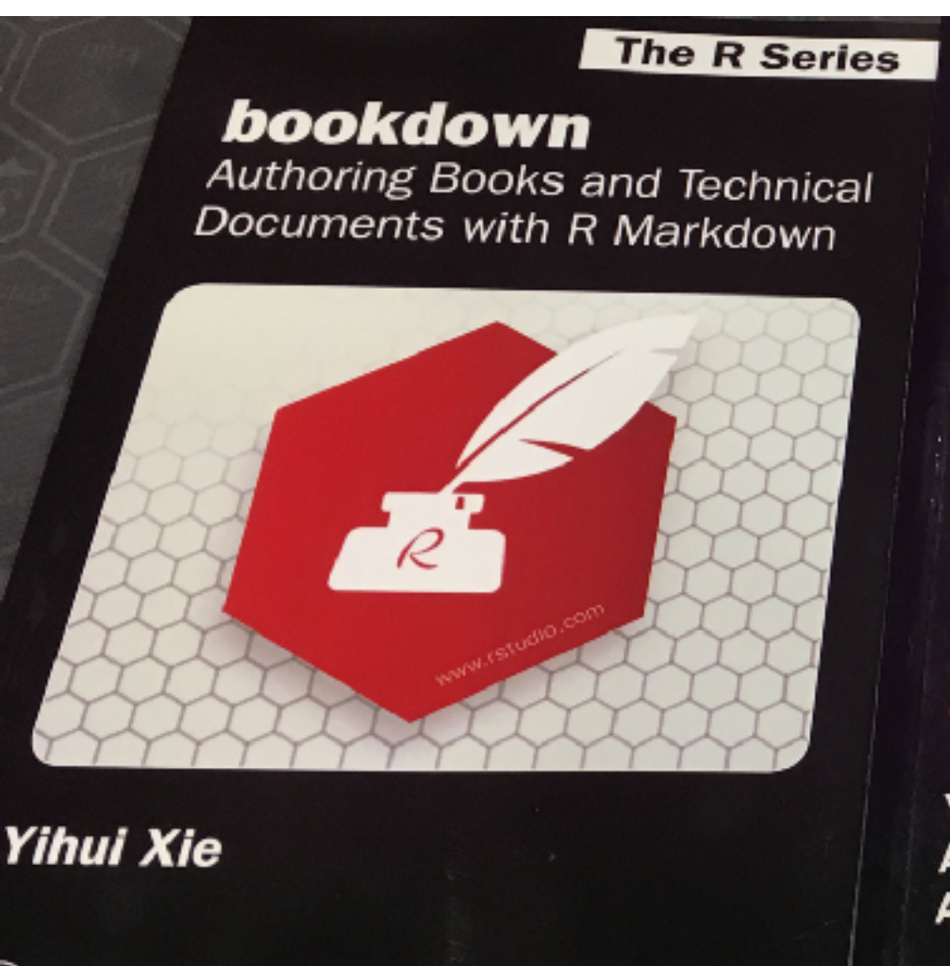
eechidna: Exploring Election and Census Highly Informative Data Nationally for Australia `https://github.com/ropenscilabs/eechidna`

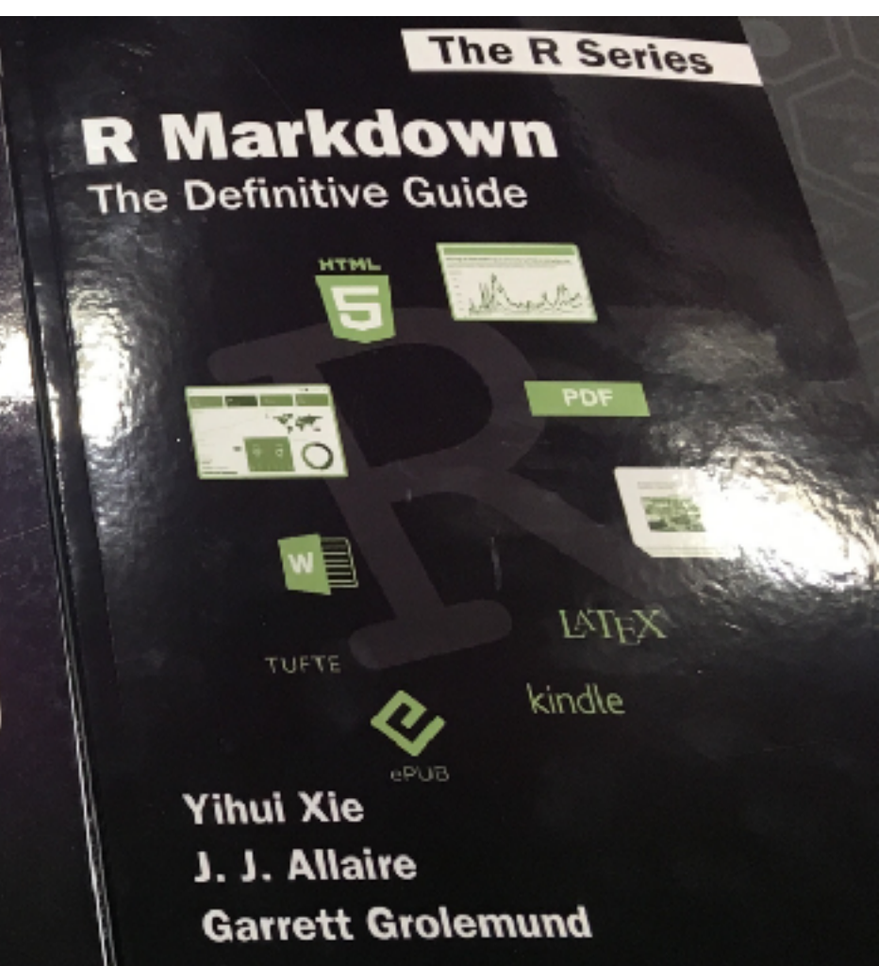


eechidna: Exploring Election and Census Highly Informative Data Nationally for Australia `https://github.com/ropenscilabs/eechidna`

Emergence of reproducibility with

- Literate statistical practice: ESS (Rossini, 2001, Emacs interface to R), Sweave (Leisch, 2002, Sweave, Part I: Mixing R and LATEX), knitr (Xie, 2013, Dynamic Documents with R and knitr), R Markdown (Allaire, Xie and many more, 2014, Dynamic Documents for R)
- compendium (Gentleman & Temple Lang, 2004, Statistical Analyses and Reproducible Research), docker (Hykes, 2013, containerization)

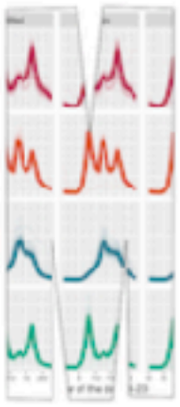




ebsmonash.shinyapps.io/lecture1b_intro/#section-making-a

Hark Media - stickers | useR! 2018 registrations | South West Ro... | YesBookit | ScholarOne Manuscripts | ebsintranet | Access via Monash | GitHub | Twitter | Gender Decod...s in job ads

R generation - Thieme - 2018 -... | JSM 2018 Online Program | Reproducibility - Wikipedia | (12) GUI and Command-line - C... | ETC1010: Data Modelling and C...



ETC1010: Data Modelling and Computing

Professor Di Cook, EBS, Monash U.

What is R Markdown?

Code chunks

Different types of documents

Making a report

Lab quiz

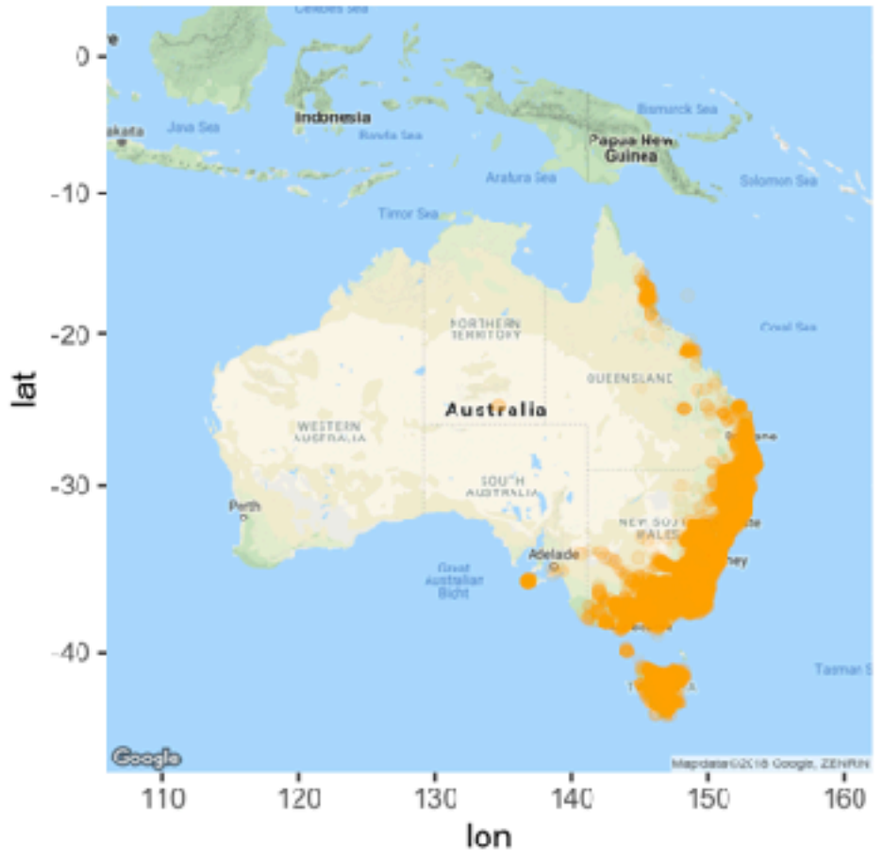
Share and share alike

Start Over

longitude

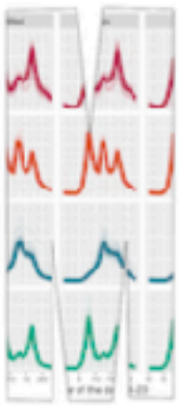
The locations would be even more recognisable if we added a real map underneath. One way this can be done is to use a google map. The `ggmap` package has an interface to extracting google maps. Install it and then grab a map of Australia with this code.

```
library(ggmap)
oz <- get_map(location=c(lon=133.8807, lat=-23.6980), zoom=4)
ggmap(oz) +
  geom_point(data=platydata, aes(x=longitude, y=latitude),
             alpha=0.1, colour="orange")
```



Write a couple of paragraphs about the locations of platypus in Australia, based on the map that you have created.

Temporal trend



ETC1010: Data Modelling and Computing

Professor Di Cook, EBS, Monash U.

What is R Markdown?
Code chunks
Different types of documents
Making a report
Lab quiz
Share and share alike

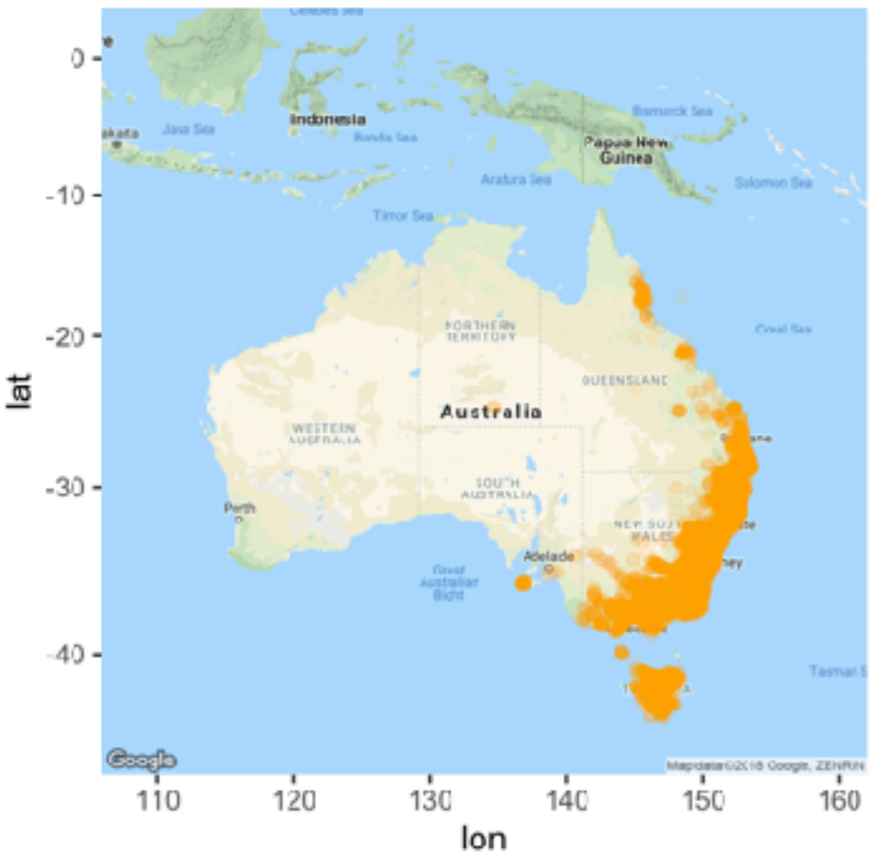
Start Over

rmarkdown: code, results and explanation in one place

longitude

The locations would be even more recognisable if we added a real map underneath. One way this can be done is to use a google map. The `ggmap` package has an interface to extracting google maps. Install it and then grab a map of Australia with this code.

```
library(ggmap)
oz <- get_map(location=c(lon=133.8807, lat=-23.6980), zoom=4)
ggmap(oz) +
  geom_point(data=platydata, aes(x=longitude, y=latitude),
             alpha=0.1, colour="orange")
```



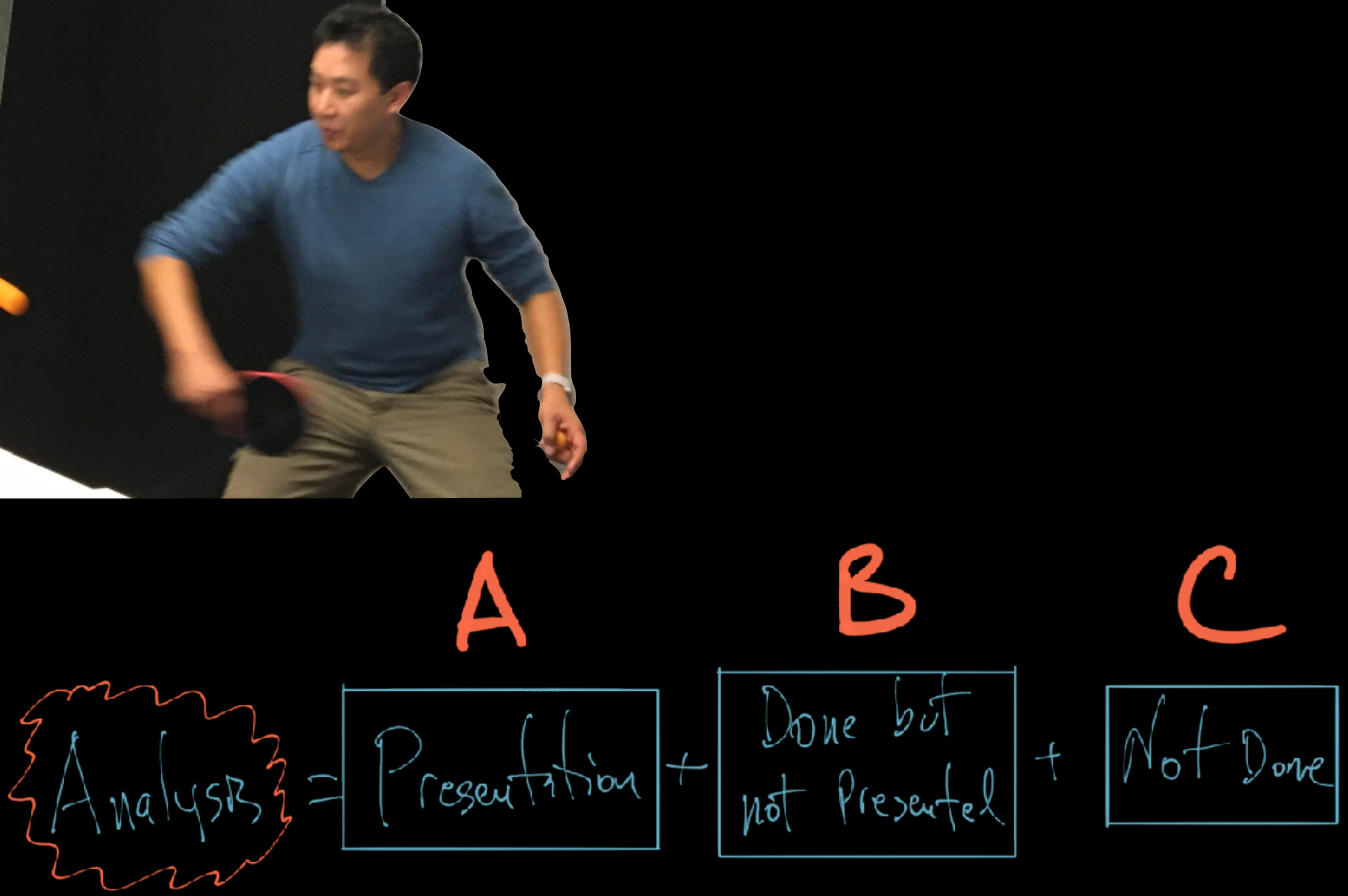
Write a couple of paragraphs about the locations of platypus in Australia, based on the map that you have created.

Temporal trend

This is amazing! It has changed my world, and made it much easier to write papers, remember what we did when the revision was needed.

My teaching philosophy has changed to be: whatever I can do, you can do too.

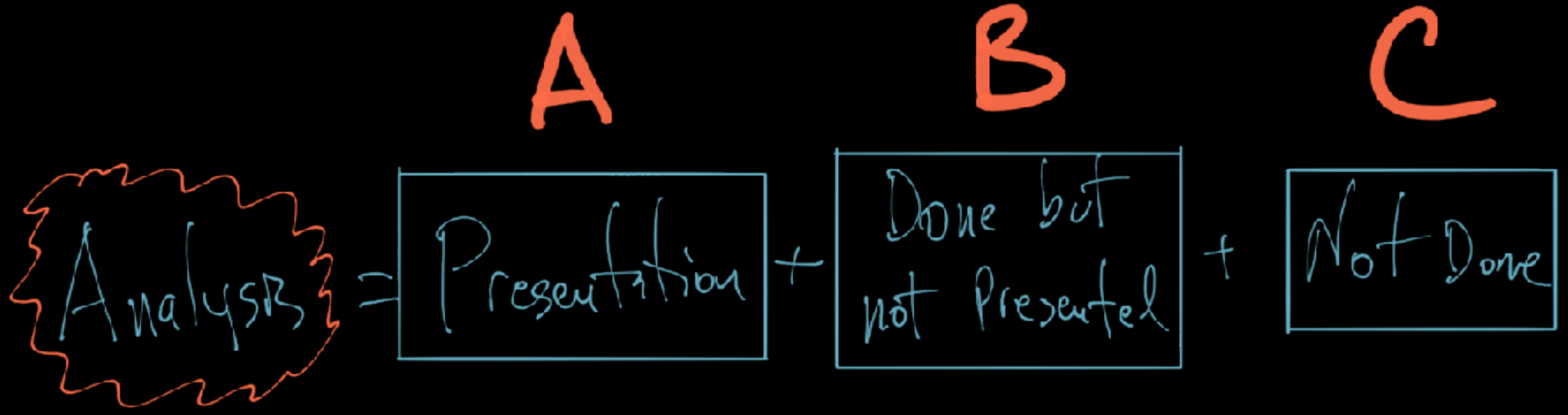
But its not full reproducibility…



https://simplystatistics.org/2018/06/04/trustworthy-data-analysis/

Roger Peng

(in action at the ping pong tables at Monash)



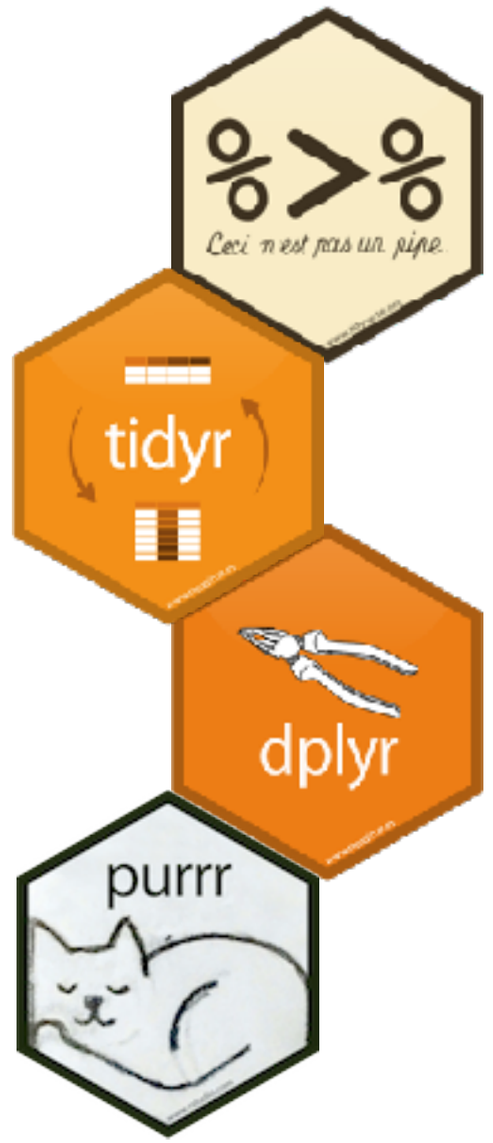
https://simplystatistics.org/2018/06/04/trustworthy-data-analysis/

What did you throw out?

Can we capture the full analysis process?

- Rmarkdown helps, enormously. Because the analysis is **scripted**.
- The %>% operator encourages readable code, and more importantly a data analysis pipeline.
- Tidyverse makes wrangling more explicit.

```
gapminder %>%
  mutate(year = year-1950) %>%
  group_by(continent, country) %>%
  nest() %>%
  mutate(model = purrr::map(data,
                ~ lm(lifeExp ~ year, data = .))) %>%
  unnest(model %>% purrr::map(broom::tidy)) %>%
  select(continent, country, term, estimate) %>%
  spread(term, estimate) %>%
  ggplot(aes(x=`(Intercept)`, y=year,
     colour=continent, label=country)) +
  geom_point()
```



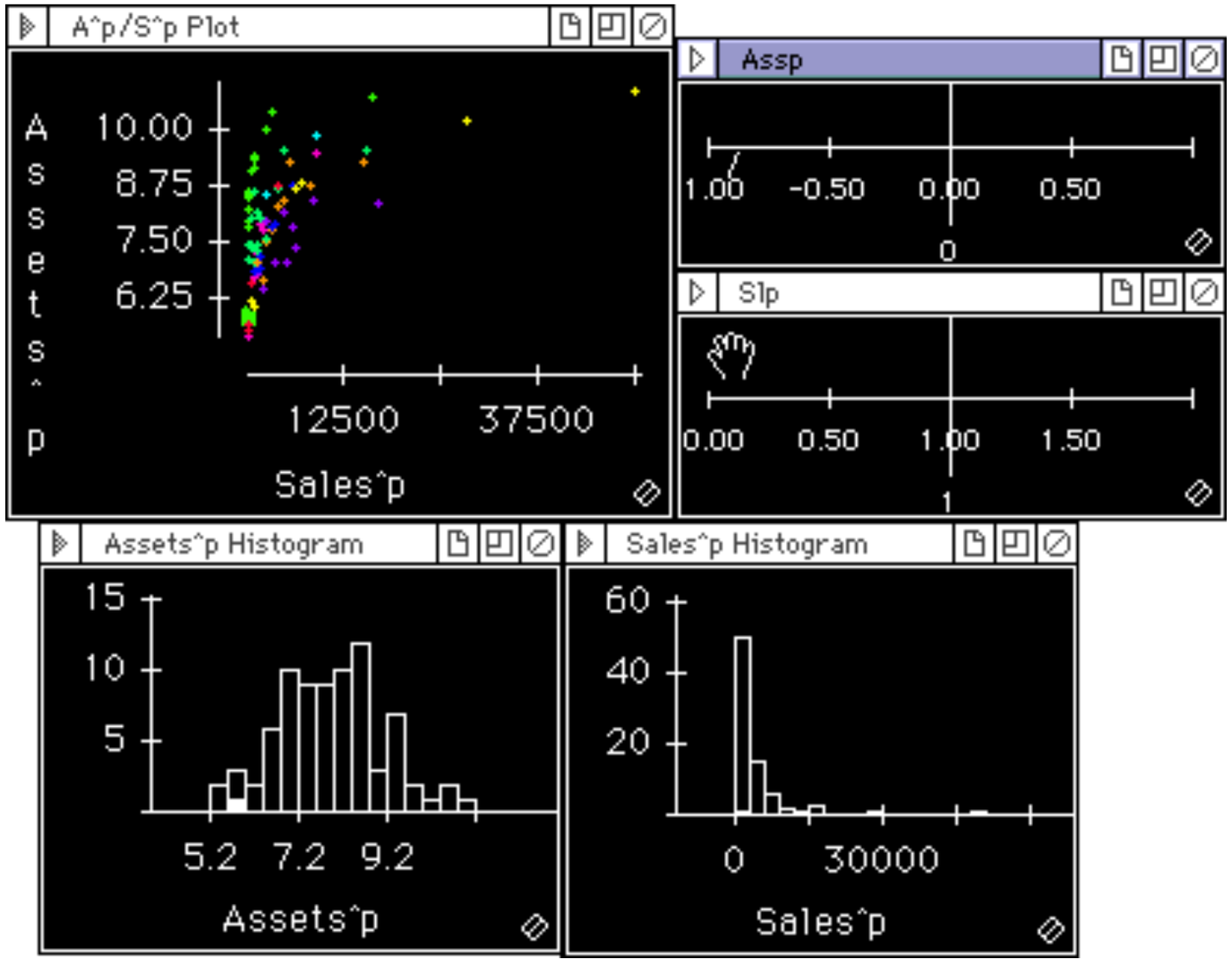
Interactive graphics

- Interactive graphics are **not scripted**

- Interactive graphics interface with you using a Graphical User Interface (GUI) as opposed to a command line (see Unwin and Hofmann for a discussion on relative merits.)

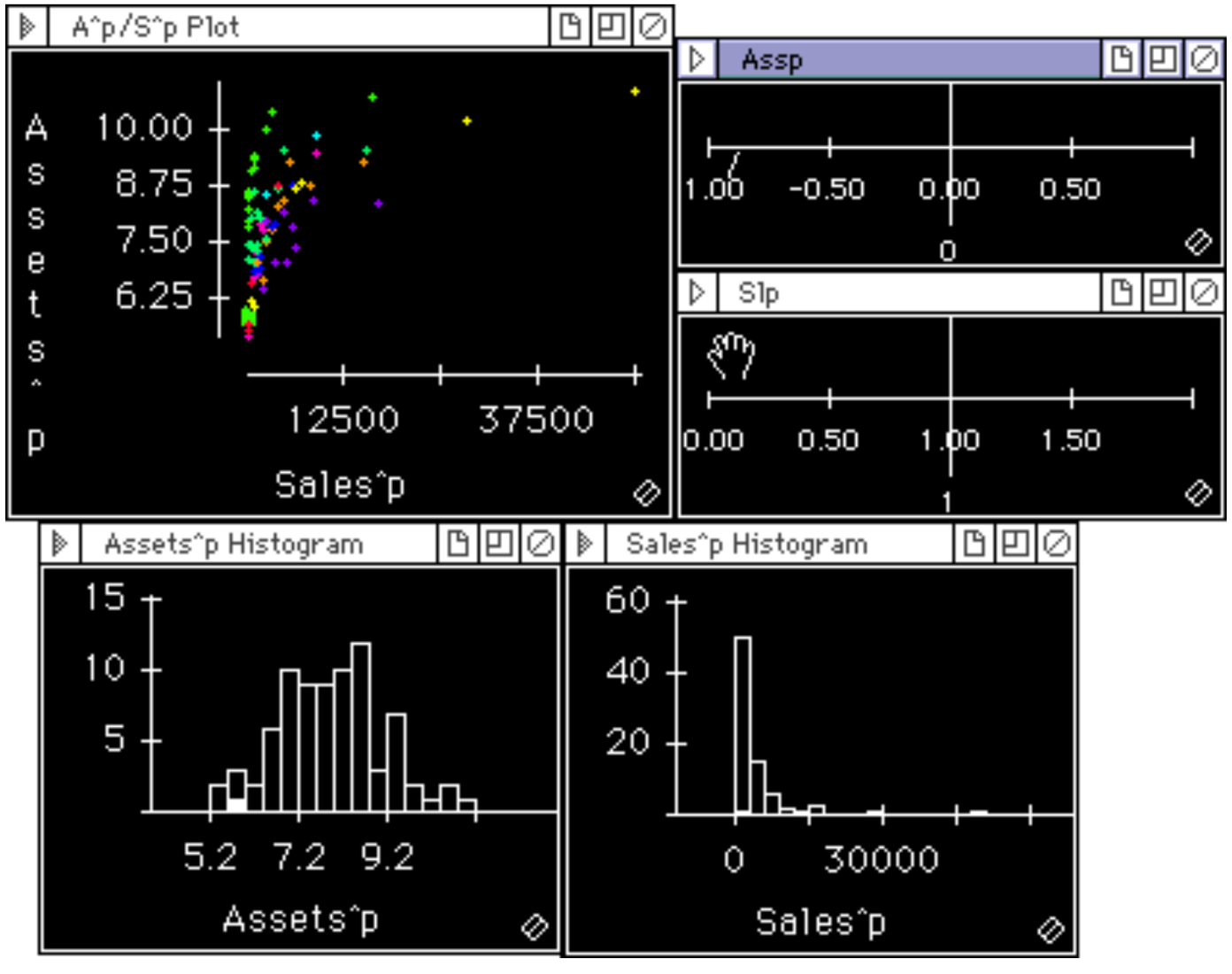
Recording user actions

“Notebook Windows: Record history of analysis…” DataDesk is something of a gold standard in what we want to have now.

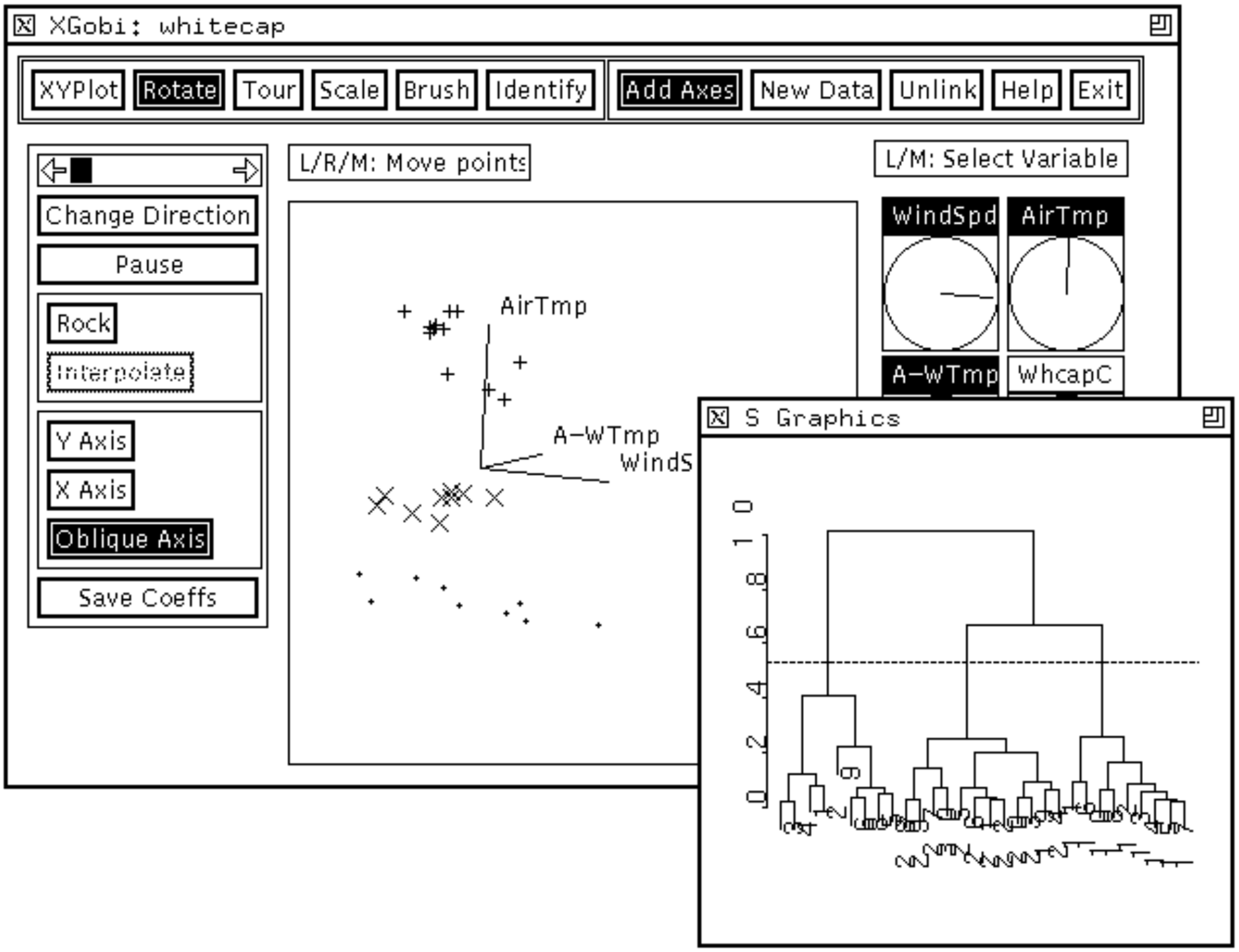


Recording user actions

“Notebook Windows: Record history of analysis…” DataDesk is something of a gold standard in what we want to have now.

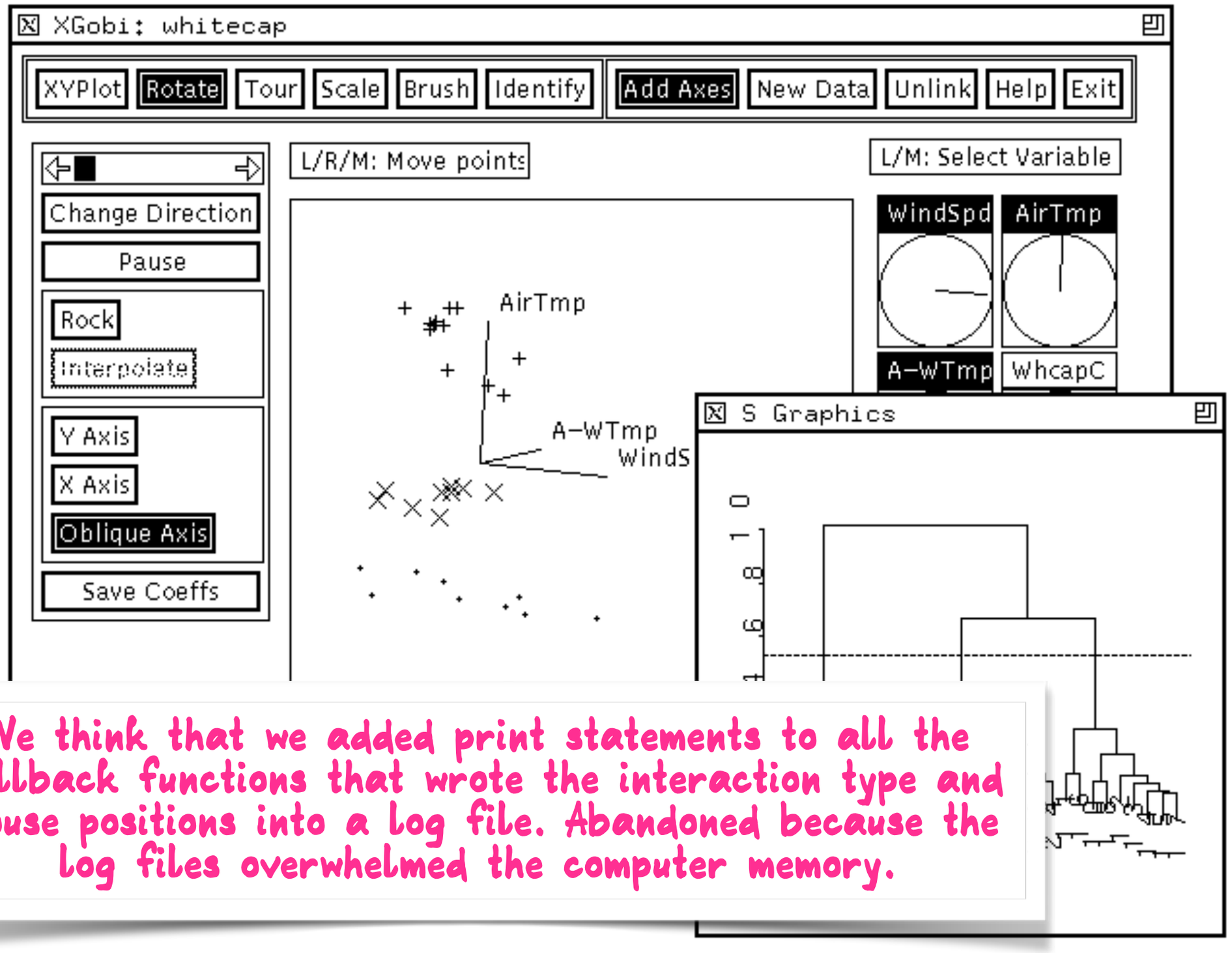


XGobi meets S (Swayne, Hubbell, Buja, 1991)



XClust: XGobi and S Graphics Window

XGobi meets S (Swayne, Hubbell, Buja, 1991)



We think that we added print statements to all the callback functions that wrote the interaction type and mouse positions into a log file. Abandoned because the log files overwhelmed the computer memory.

XClust: XGobi and S Graphics Window

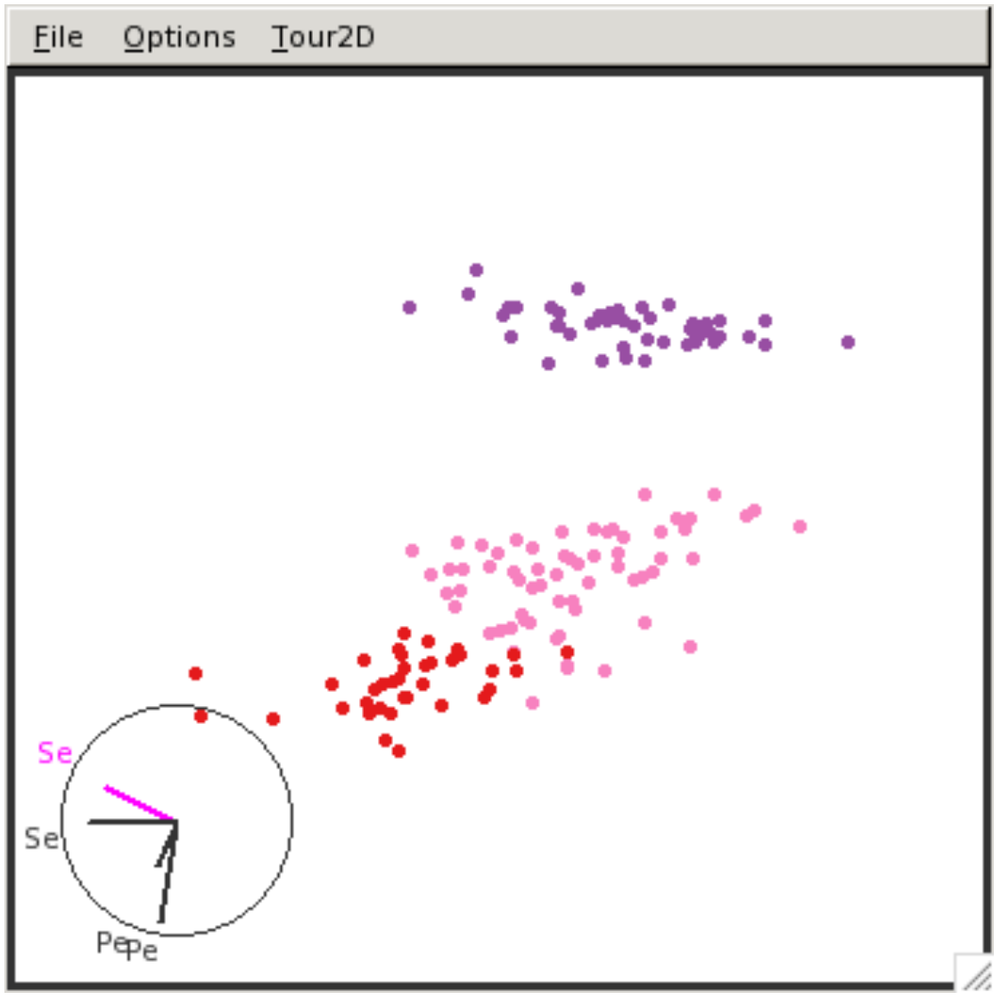
A hiccup has been what to do with the volume of data generated.

GGobi meets R: an extensible environment for interactive dynamic data visualization (Temple Lang, Swayne, 2001)

rggobi (Wickham, Lawrence, Temple Lang, Swayne, 2007)

"The rggobi package provides a command-line interface to GGobi, an interactive and dynamic graphics package."
GGobi web site

```
g <- ggobi(iris)
clustering <- hclust(dist(iris[,1:4]),
  method="average")
glyph_colour(g[1]) <- cuttree(clustering, 3)
```



I really need it!

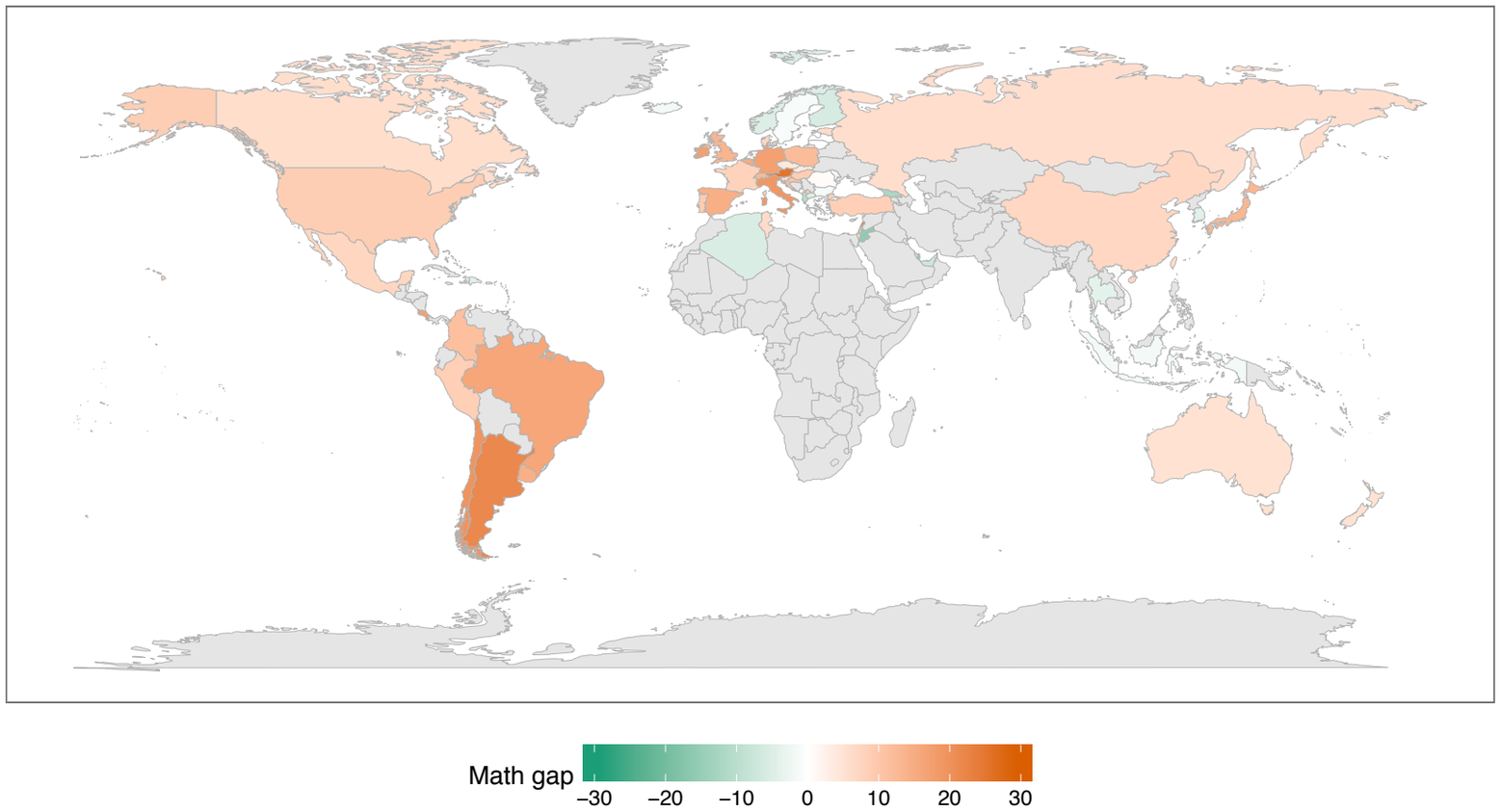
This was an editor’s response to a paper submission in 2015:

“…places it feels an awful lot like a 'fishing expedition' - that is, it reads as if someone has had access to a large dataset and kept playing with it until interesting results popped out.”

I really need it!

This was an editor’s response to a paper submission in 2015:

“…places it feels an awful lot like a 'fishing expedition' - that is, it reads as if someone has had access to a large dataset and kept playing with it until interesting results popped out.”

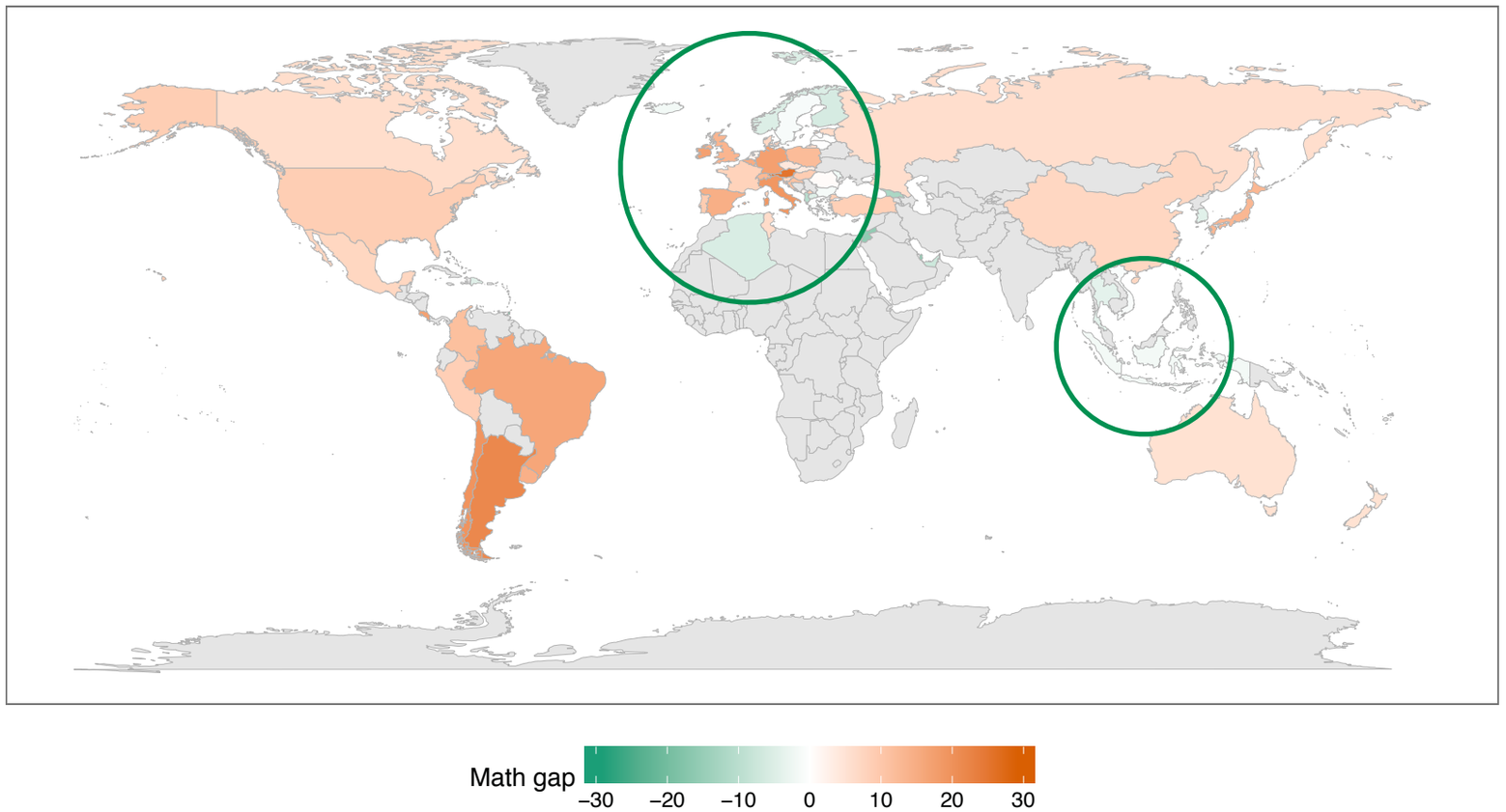


I really need it!

This was an editor’s response to a paper submission in 2015:

“…places it feels an awful lot like a 'fishing expedition' - that is, it reads as if someone has had access to a large dataset and kept playing with it until interesting results popped out.”

math gap is not a gap

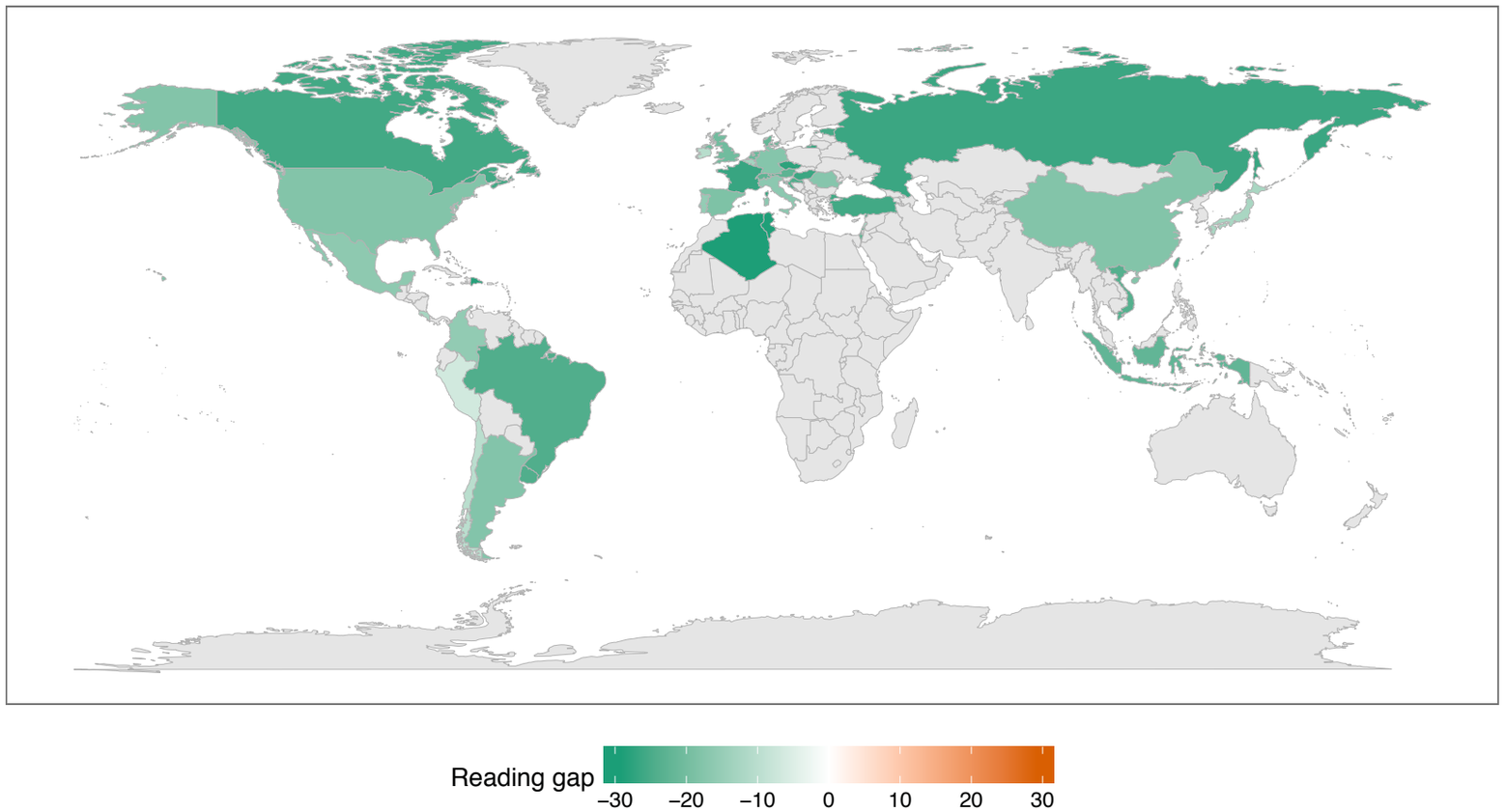


I really need it!

This was an editor’s response to a paper submission in 2015:

“…places it feels an awful lot like a 'fishing expedition' - that is, it reads as if someone has had access to a large dataset and kept playing with it until interesting results popped out.”

math gap is not a gap



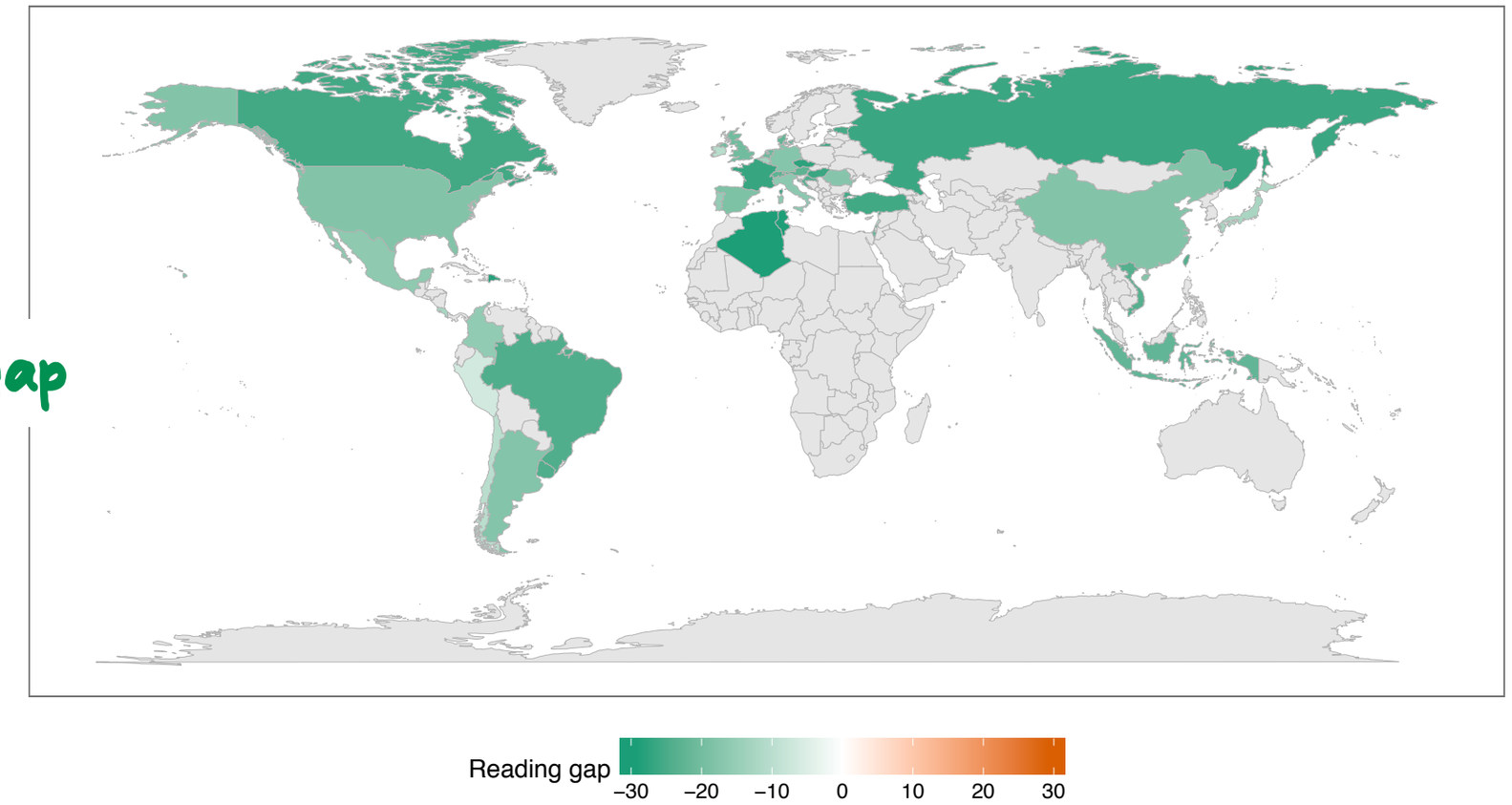
I really need it!

This was an editor’s response to a paper submission in 2015:

“…places it feels an awful lot like a 'fishing expedition' - that is, it reads as if someone has had access to a large dataset and kept playing with it until interesting results popped out.”

math gap is not a gap

reading gap is a real gap



The road forward

- Recording user interactions is a part of comprehensive record of analysis.

- Its not easy, but I these next talks show us that we are getting closer.