

GOING BEYOND R

DATA VISUALIZATION

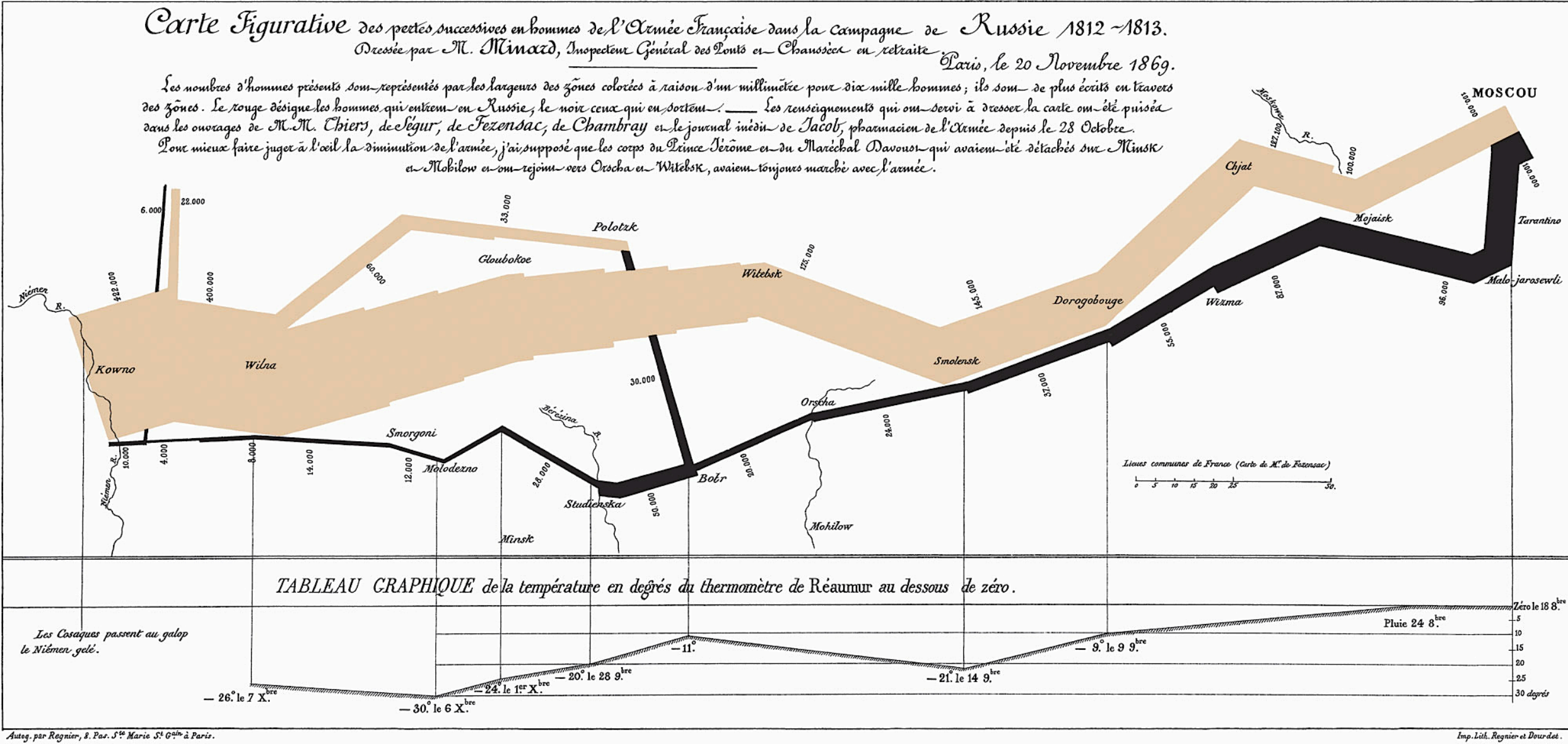
THANKS TO...

- ▶ This presentation draws liberally on examples from Edward Tufte's work on data visualization, Lee Wilkinson's work on the grammar of graphics, and some examples from my own work.
- ▶ Graphics are cited throughout, but due to space constraints, not all ideas are, though I try my best to use references for the core concepts.

WHAT'S THE PURPOSE OF VISUALIZING OUR DATA?

- ▶ Humans are highly visual learners and communicators
 - ▶ We can often recognize in a picture what would take months of advanced training to understand statistically or textually
 - ▶ Remember the adage "A picture is worth a thousand words?"
 - ▶ Your visuals should be worth a lot more than that

TUFTE: "THE MOST INFORMATIVE DATA VISUALIZATION EVER"



TWO PURPOSES

- ▶ There are two major reasons we want to visualize our data:
 - ▶ Exploratory data analysis (getting to know our own data)
 - ▶ Story-telling (helping others get to know our data)
- ▶ **Using graphics to explore your own data is a best practice**
 - ▶ *In quantitative analysis*, focus on statistical significance can blind us to the real findings, and can cause inadvertent “p-hacking”
 - ▶ *In qualitative analysis*, the huge scope of the data (and often inductive nature of the research) can lead to a similar phenomenon: “cherry-picking” of particular quotations or events that support one’s arguments

EXAMPLE (1)

```
summary(lm(data=dfw, comp~attr+dom+moth+fem))
```

```
##
```

```
## Call: summary(lm(data=dfmw, comp~attr+dom+moth+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

```
##
```

```
## Residuals:
```

```
## Call: summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
## Linear model using ordinary least squares
```

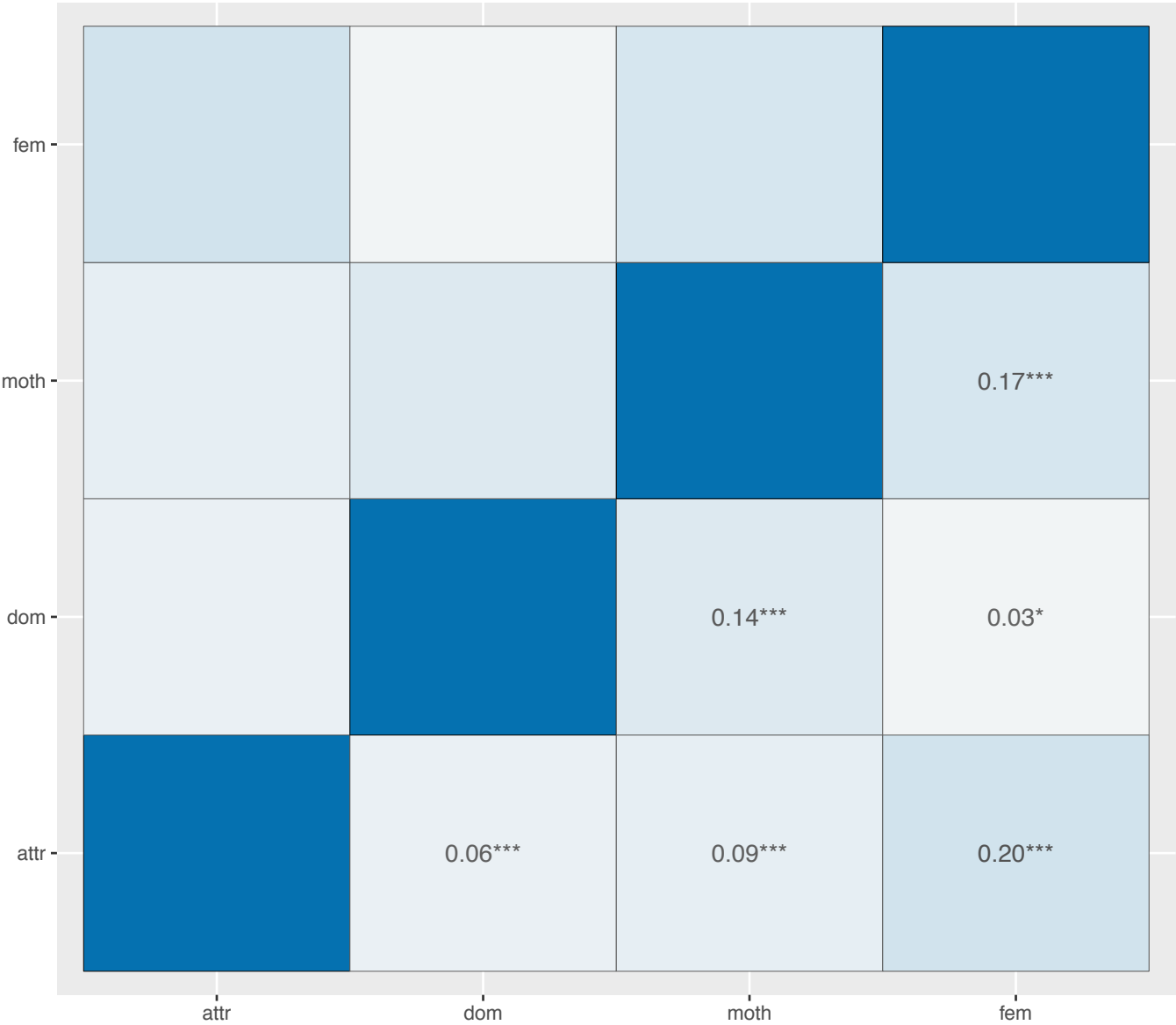
```
summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
##
## Call:
## lm(formula = comp ~ attr + dom + fath + fem, data = dfmm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.5492 -0.7935  0.0126  0.9902  3.5104
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.54347    0.14215   31.963 < 2e-16 ***
## attr         0.06529    0.02001    3.263  0.00112 **
## dom          0.05198    0.02488    2.089  0.03681 *
## fath         0.13626    0.02523    5.401 7.35e-08 ***
## fem         -0.26193    0.02427   -10.794 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.353 on 2196 degrees of freedom
## (2350 observations deleted due to missingness)
## Multiple R-squared:  0.07896,    Adjusted R-squared:  0.07728
## F-statistic: 47.06 on 4 and 2196 DF,  p-value: < 2.2e-16
```

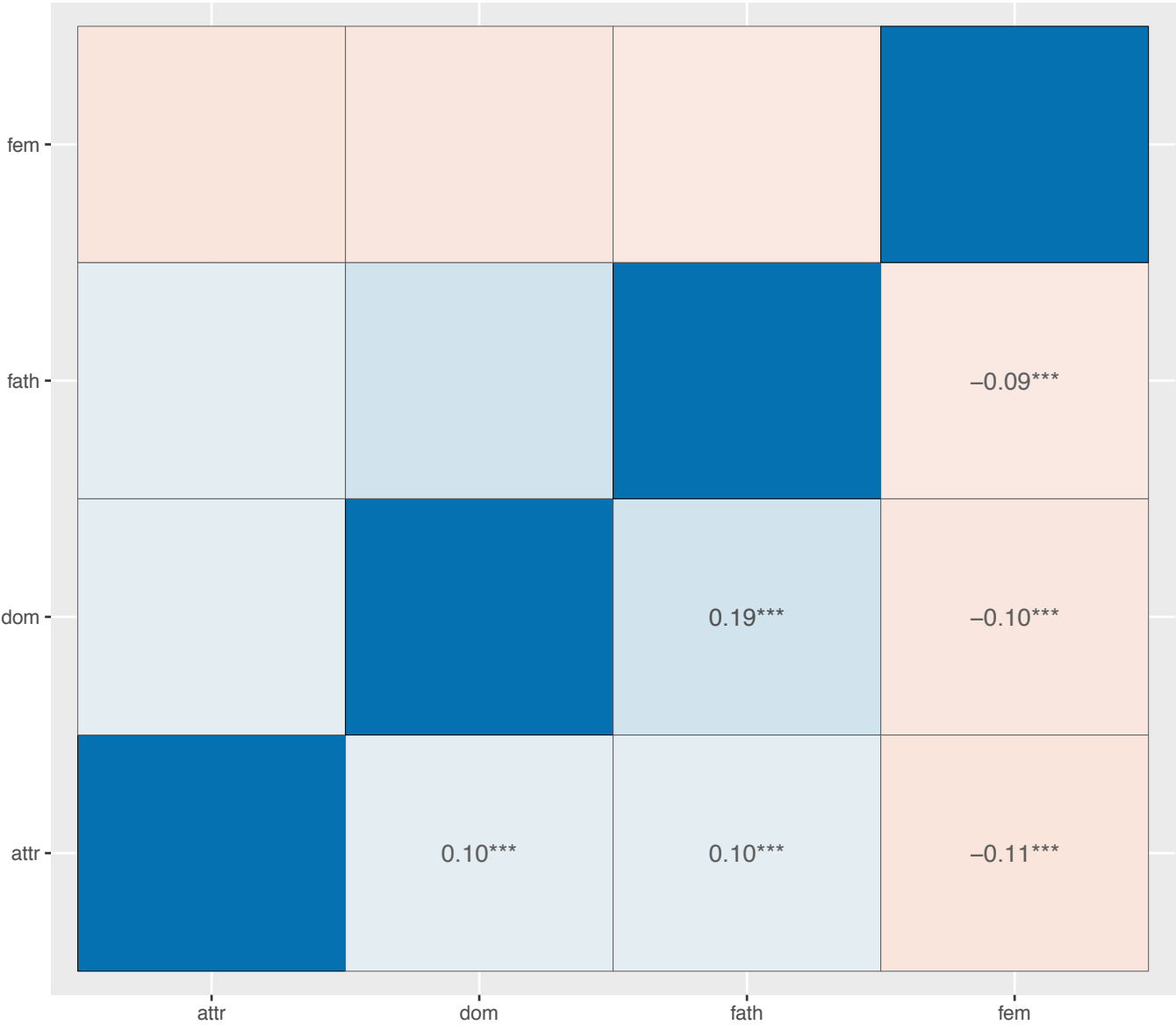
EXAMPLE (1)

Trait Correlation Coefficients by Sex

Females



Males



BEGINNINGS

EXAMPLE (2)

Afraid of my Debt, and not having money to live or someone to take care of my daughter

Mostly because of negative experiences within the local Democratic Party. Also, the office I was originally interested in (School Board) is less attractive and unpaid and the other position I am interested in (County Commissioner) is currently held by someone I admire and who does great work. /

(1) Pay for City Council (contemplated office) too low to do as only job, but job too large to do well on part-time basis (esp. with children still at home) / (2) Disdain for ratio of fundraising/political shmoozing to engagement with actual policy issues

1. Friendly incumbents in office / 2. I hold elected positions in my local, regional, state, and national union. So, my time is spread pretty thin. / / However, I hope that when an opening develops I will be able to stop some of my other work and run for office.

1. Haven't had the right opportunity in my districts yet. / / 2. Need to be more financially stable

1. It's a blood sport where I live and the process of running feels too daunting. 2. I like to have some time for myself and with my kids, spouse and friends. 3. While I think I would be a good legislator, my current job (in policy advocacy) allows me to fight for the things I care about without having to horse trade with people I don't respect.

1. Time Commitment. I don't feel like I can give the time my potential constituents deserve given that I work full time in a city 60+ miles from where I live. I'm the primary household earner. I have to work. Also concerns about giving my family the time it deserves. / 2. Nastiness of politics. I'm not confident I have thick enough skin. People are awful to elected officials regardless of whether the elected official deserves it.

1. Too intrusive in my personal life. / 2. Too much focus on fundraising. / 3. Am not 100% aligned with Democratic Party platform (I'm more liberal)

Afraid of the personal scrutiny, dislike fundraising and not good at answering questions on the fly. Also hard to keep integrity when so much compromise is required.

After participating in the Emerge Program I better understand the complexities of fundraising. I live in a big city and the money needed is enormous and I don't have name recognition yet.

At 27 years old, I am working to establish myself in the community and build connections. I am also working to gain political experience through Emerge and working with active campaigns. In two years, I plan on running and feel I will have the experience and name recognition needed to run at that point.

At 28, I've considered myself too young to seriously run for office and have dedicated myself to learning the nuts and bolts of both campaigns and public policy by working on campaigns or for the government.

Because I am 23-years-old

Because I don't want my past, private, and personal life in the spotlight. I also don't subscribe to the democratic party nor my union status quo. It's too constricting and short sided. And the fundraising portion is horrible. If there were someone else in charge of the funding--it would be a lot easier.

Because I had a baby the year i would have probably run for office and now am not sure if i plan to stay in Arizona long term. /

Because I was turned off by the amount of self I felt like I had to sacrifice. I was also discouraged by how difficult it is to break into city politics. Finally, I would like to have impact on issues that are larger than street pavement and light repair.

Because the Democratic party does not seek or encourage women with disabilities to run. Because when my name is floated the first question is always "Is she healthy enough" not "is she qualified."

Because the seat that I want is occupied by a party leader/fellow Emerge alum. I have been pushed to run for other offices but I find them less of a fit with my interests, skills and schedule.

Been building professional career

Been in medical school so don't have time yet.

Busy raising a family and working. I live in a rural, Republican dominated district so a Dem is unlikely to win. I can only run for county or statewide office.

Can't balance family with legislative session, not interested in any local office

Childcare

Children too young, financial not yet independent. But am transitioning into being ED of the non-profit I founded

Concerned about the economic impact on my family as I am the primary wage earner. Secondary concern at having my personal life or private medical decisions publicly trashed by negative campaigning.

Conflict with my job.

current democrat will not step down from office held for 30 years

Currently running.

Decided not for me

Decided to dedicate more time to my family.

Determined that I'm better behind the scenes and don't want to run. Also DC does not have actual voting rights and the City Council is a mess.

Didn't think I had what it takes to run for office. Just finished emerge and there were no positions for me to run for during this election cycle.

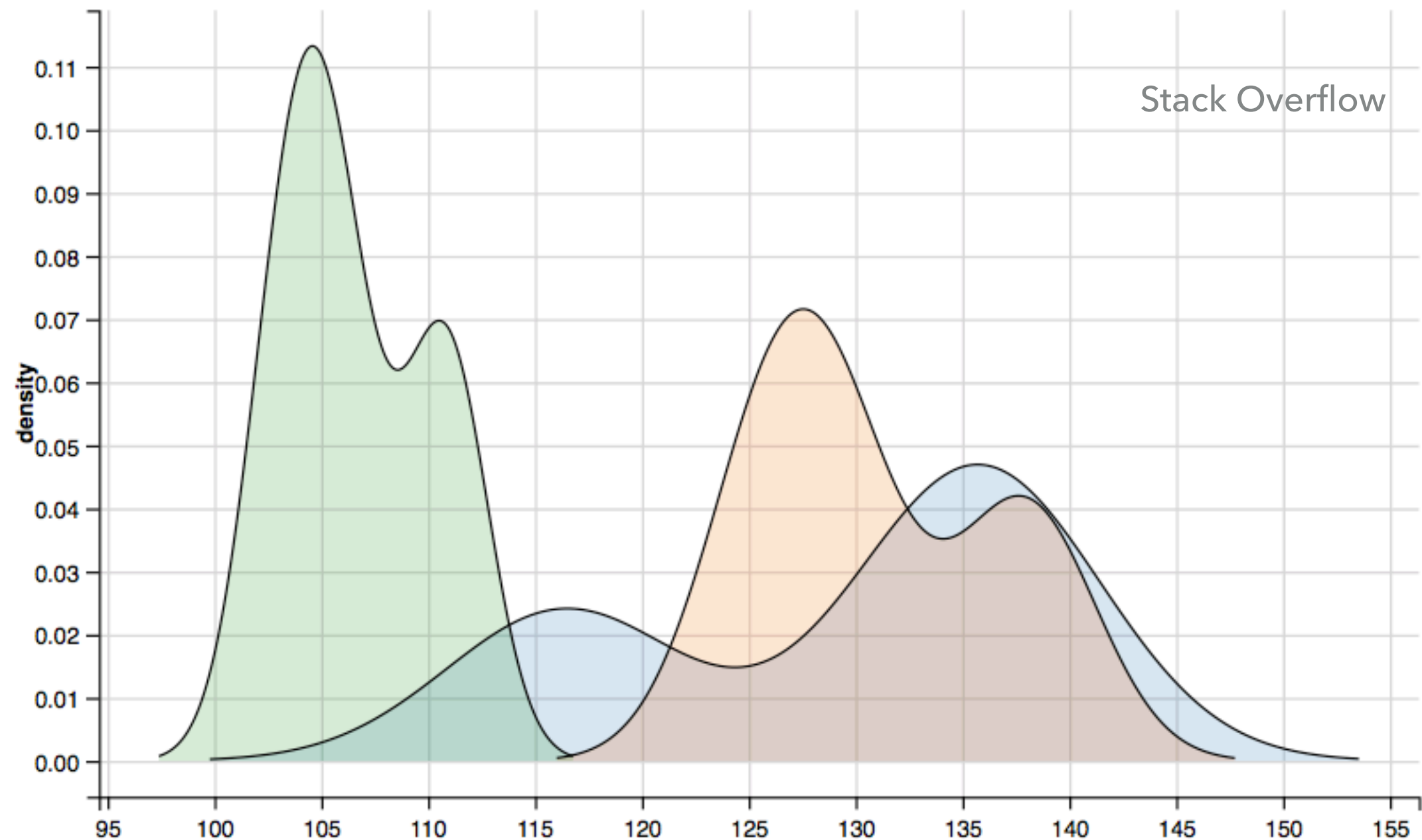


HOW SHOULD WE VISUALIZE A GIVEN STORY?

- ▶ First, start with your data
 - ▶ Univariate analyses
 - ▶ Bivariate analyses
 - ▶ Multivariate analyses
 - ▶ Qualitative/other

UNIVARIATE ANALYSES

- Histograms, frequency plots, density plots, bar charts



UNIVARIATE ANALYSES

- ▶ Bar charts should be reserved for categorical data only

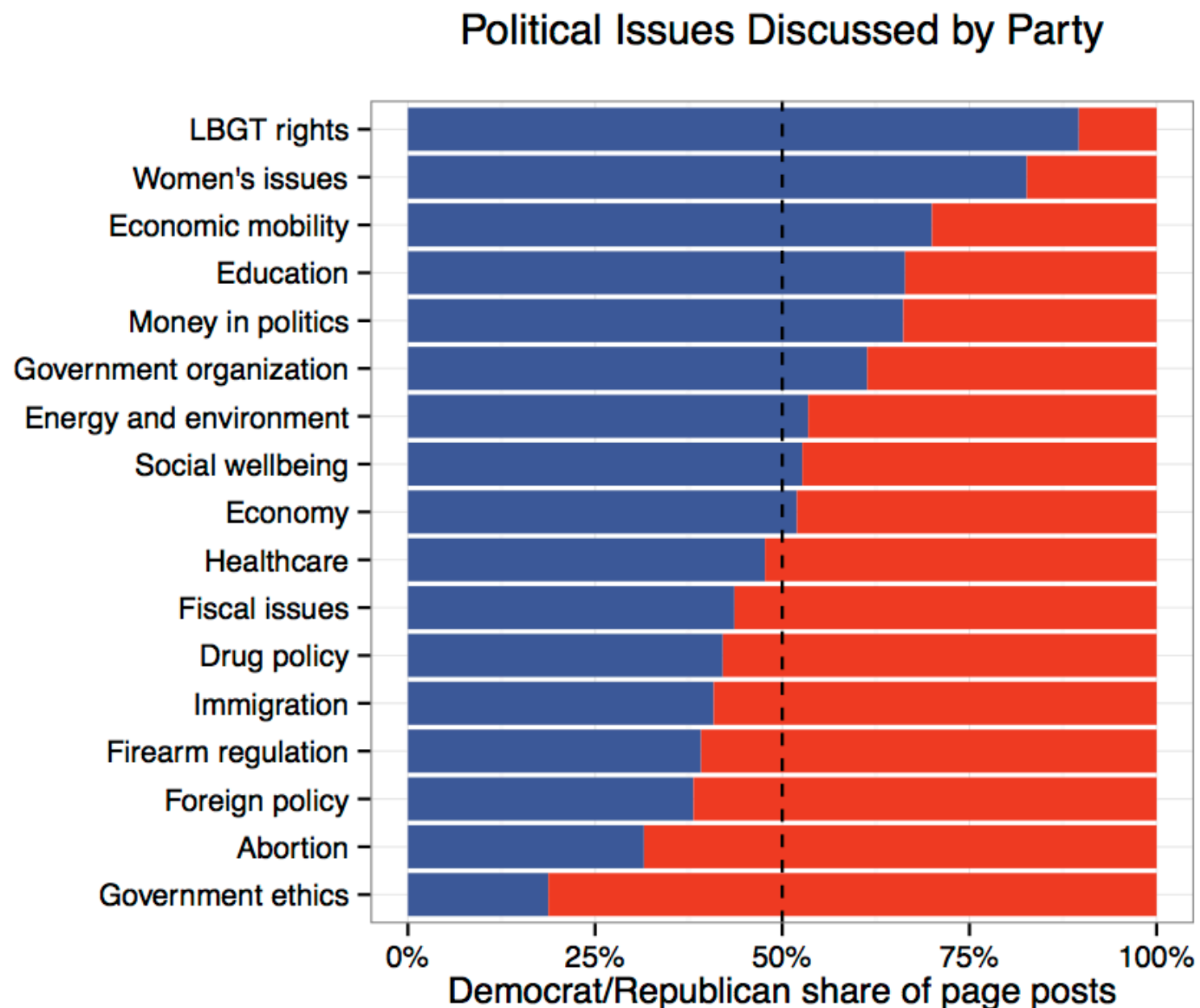


Chart from Solomon
Messing's Wordpress

BIVARIATE ANALYSES

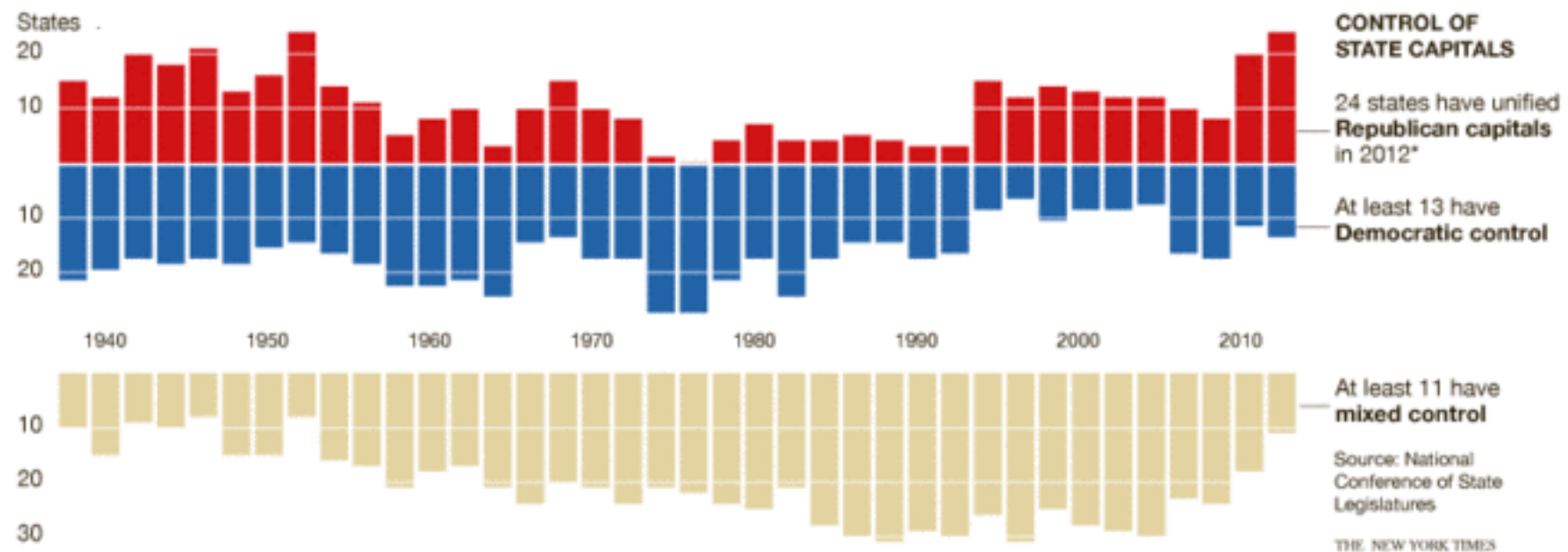
- ▶ Stacked bar, box plot, coefficient plot, scatterplot, line or multi-line plot, etc.

Complete Control

There are now more state capitals dominated by a single party — where one party controls the legislature and the governor's office — than at any time since 1952.

* Virginia is counted as unified Republican because its State Senate is tied and its tiebreaker, the lieutenant governor, is a Republican.

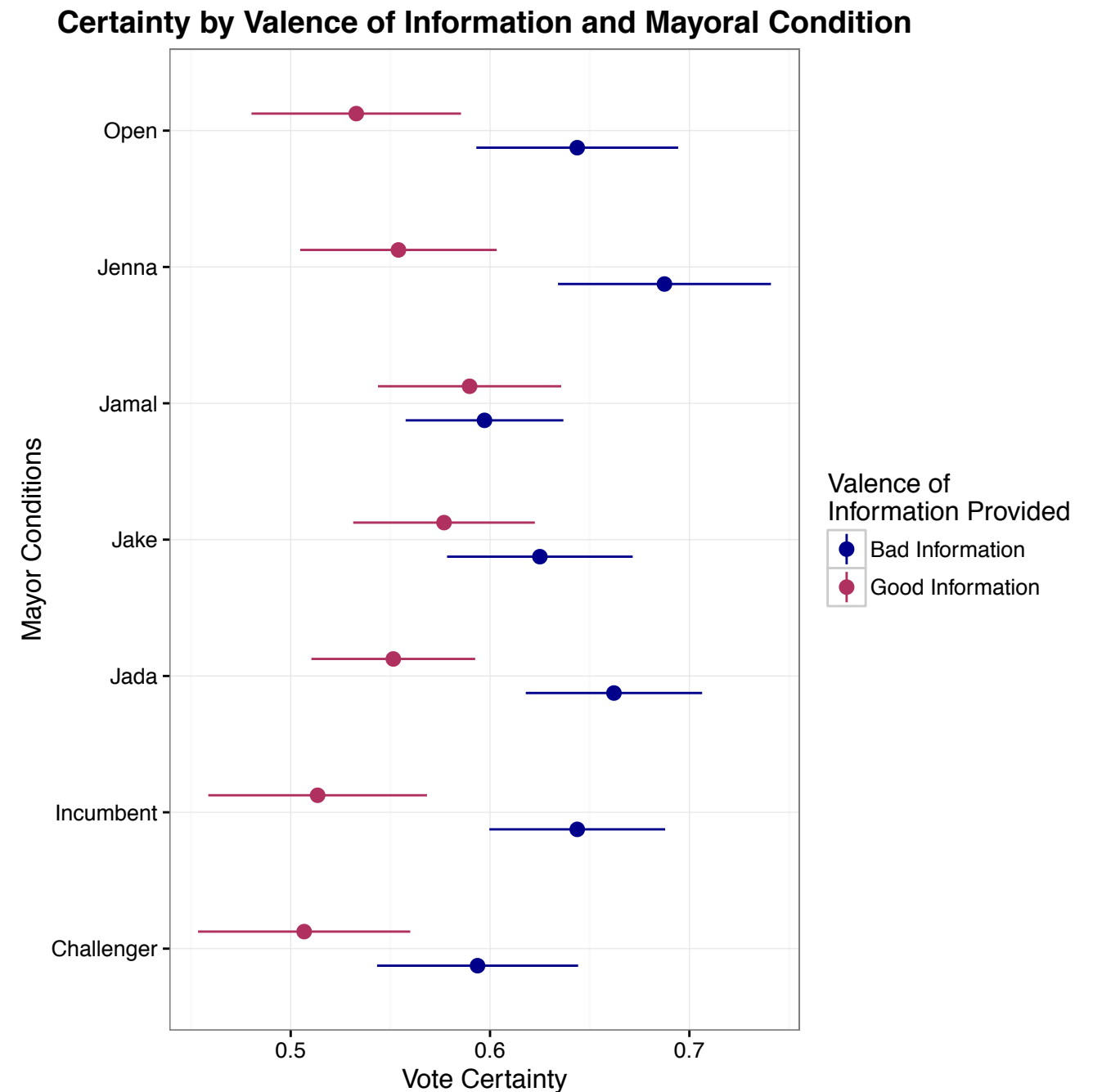
† Early results appeared to show that New York had unified Democratic control, but votes are still being counted in many races.



NY Times

BIVARIATE ANALYSES

- ▶ When your X variable is categorical or ordinal, you'll need a stacked bar, box plot, or coefficient plot



BIVARIATE ANALYSES

- Analyses with ordinal Y variables are often best presented as mean-centered stacked bars

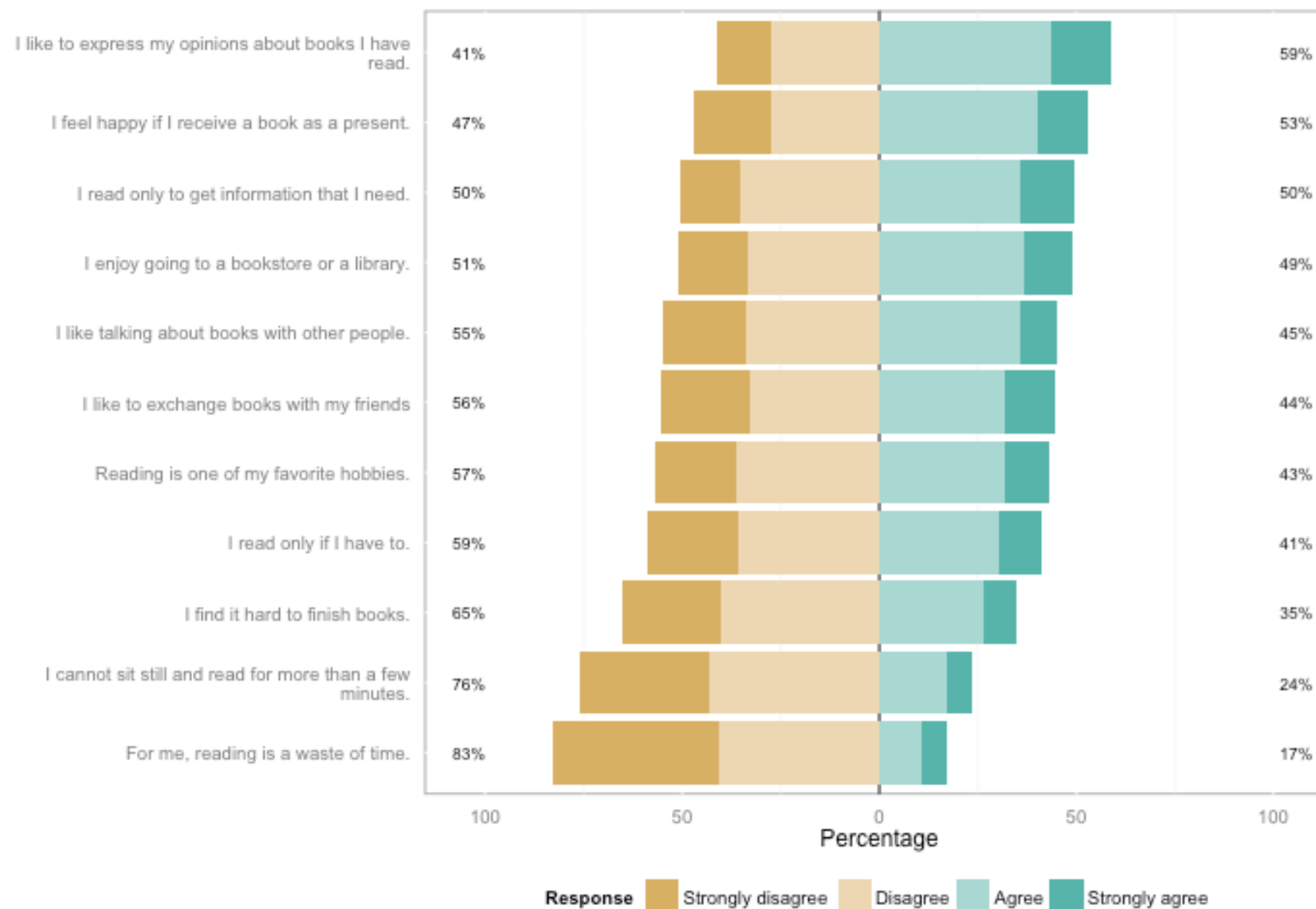


Chart from
Jason Bryer

BIVARIATE ANALYSES

- ▶ Analyses with discrete X variables (for example, by year, month, or day of week) variables allow you some additional options

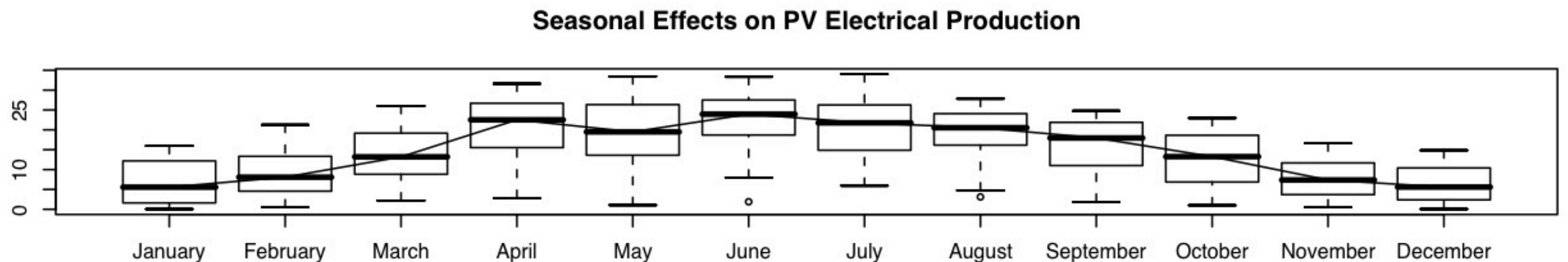
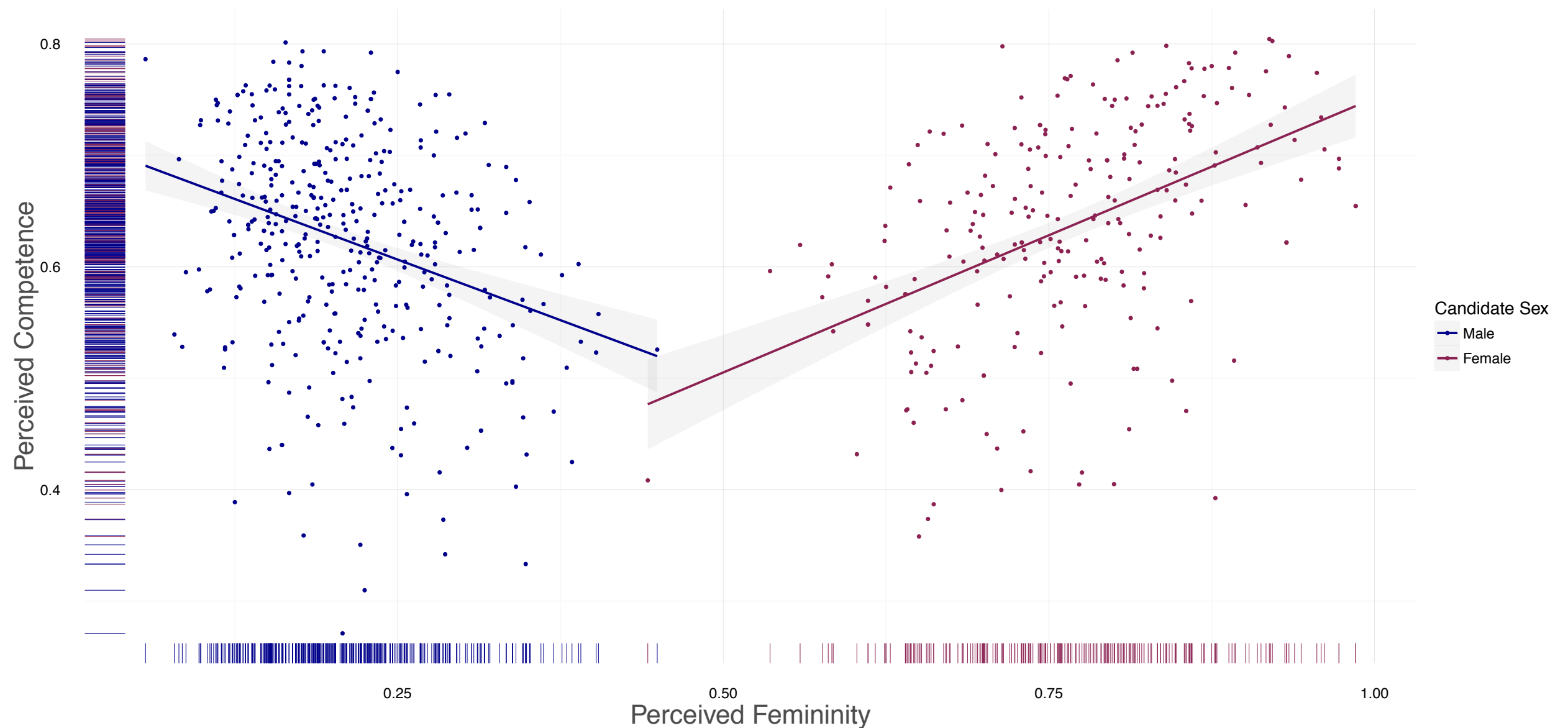


Chart from the
Personality Project

BIVARIATE ANALYSES

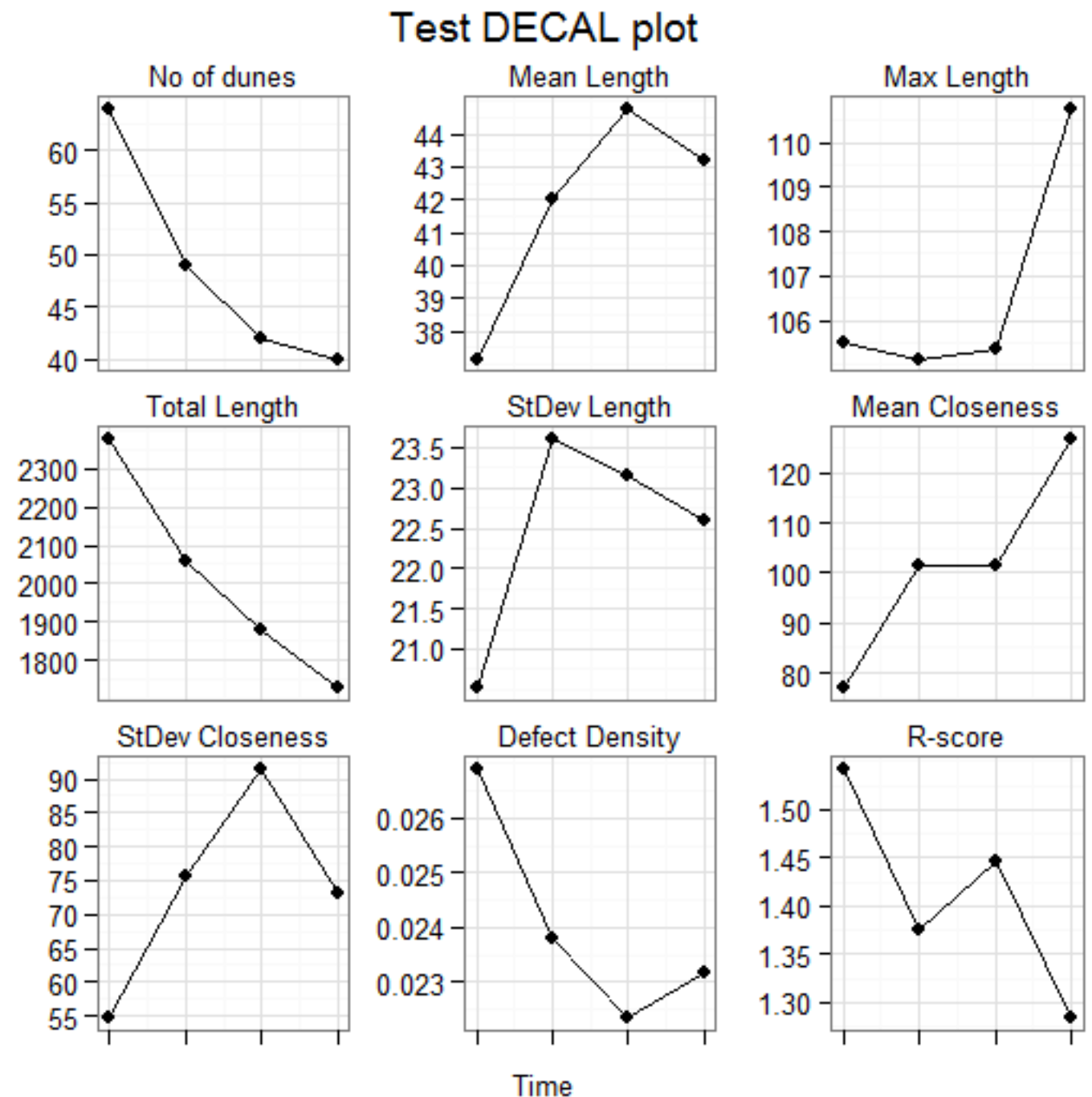
- ▶ With continuous X variables, the scatterplot is your best friend



MULTIVARIATE ANALYSES

- ▶ Multivariate analyses are more complicated to display—if you can plot it as a coefficient plot, or “facet” it into subsets, that’s often best (otherwise, consider reverting to a regression table)

Chart from RT Wilson



PLOTTING WITH GGLOT (AND BEYOND)

- ▶ Visuals have three key components in ggplot:
 1. **Data** (the dataframe or matrix you're using)
 2. **Objects** (basically, the type of plot—line, bar, etc.)
 - ▶ Raw data (think scatterplot points)
 - ▶ Summary data (think regression lines)
 3. **Aesthetics** (color, size, shape, etc.)

PLOTTING WITH GGPLOT (AND BEYOND)

- ▶ Decisions about **objects** and **aesthetics** depend on your answers to two questions:
 - ▶ What is your core argument? (I.e., why do you even need a graph in the first place?)
 - ▶ Who is your audience? (A journal article, a conference presentation, a TED talk?)
 - ▶ The former probably want to see more of the raw data (that is, you're using a chart so you can show more *complexity* and more *raw data* than a table of figures would allow), the latter probably want to see the most digestible version of the story