Teaching Intro Data Science & Assessing Learning

Preparing to Teach
JSM 2018

Nicholas Horton
Amherst College

@askdrstats

nicholasjhorton

nhorton@amherst.edu

Mine Çetinkaya-Rundel
Duke University + RStudio

@minebocek

mine-cetinkaya-rundel

mine@stat.duke.edu

What is data science?

- Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge.

- We're going to learn to do this in a `tidy` way -- more on that later!

What happens in an intro data science course?

- Will we be doing computing? Yes.
- Is this an intro CS course? No, but many themes are shared.
- Is this an intro stat course? Yes, but it's not your high school statistics course.
- What computing language will we learn? R.
- Why not language X? We can discuss that over ☕.

Poll

Raise your hand if

- you’ve used R
- you’ve used RStudio
- you’ve taught (with) R
- you’ve used R Markdown
- you’ve taught (with) R Markdown
- you’ve used a version control system, e.g. Git and GitHub
- you’ve taught (with) Git and GitHub

12345

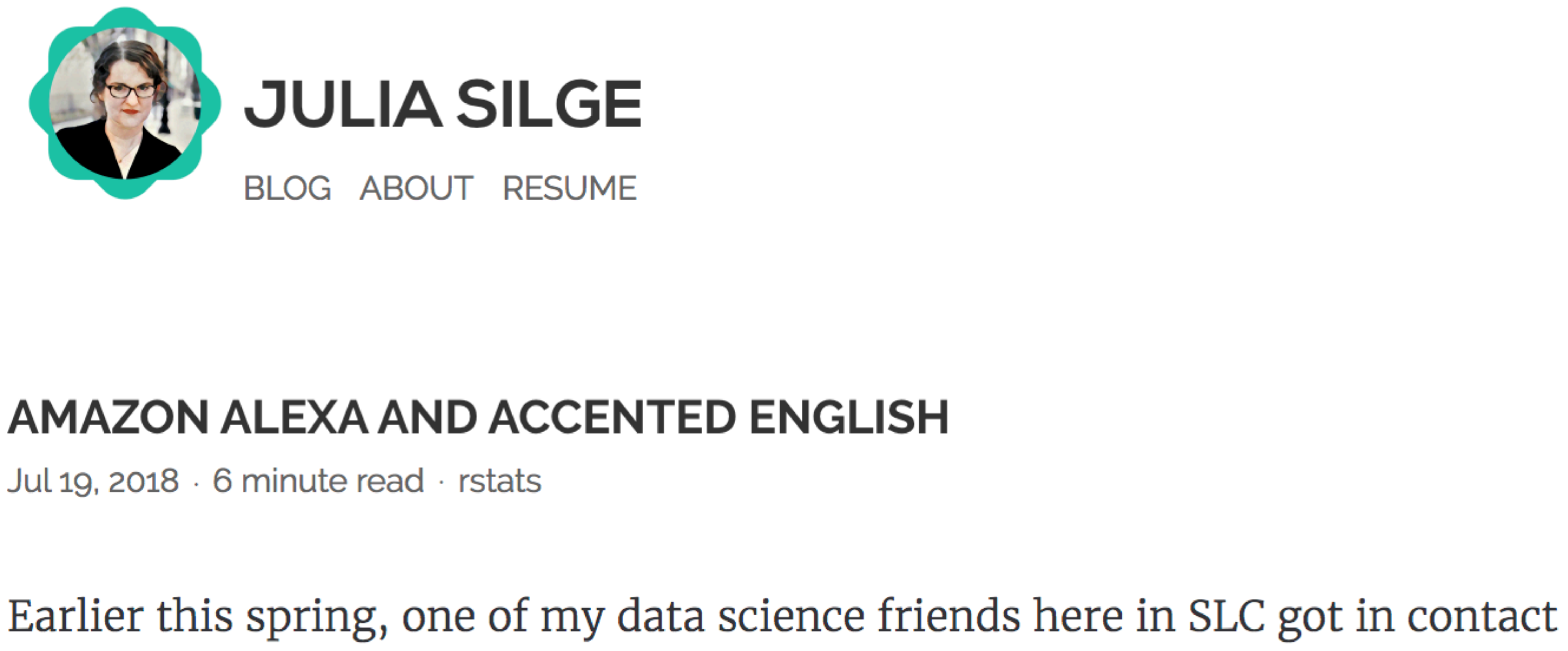
1

cherish day one

minimize
time spent
on course logistics

maximize
time spent
creating a data
visualization

1a. show examples of data in the wild



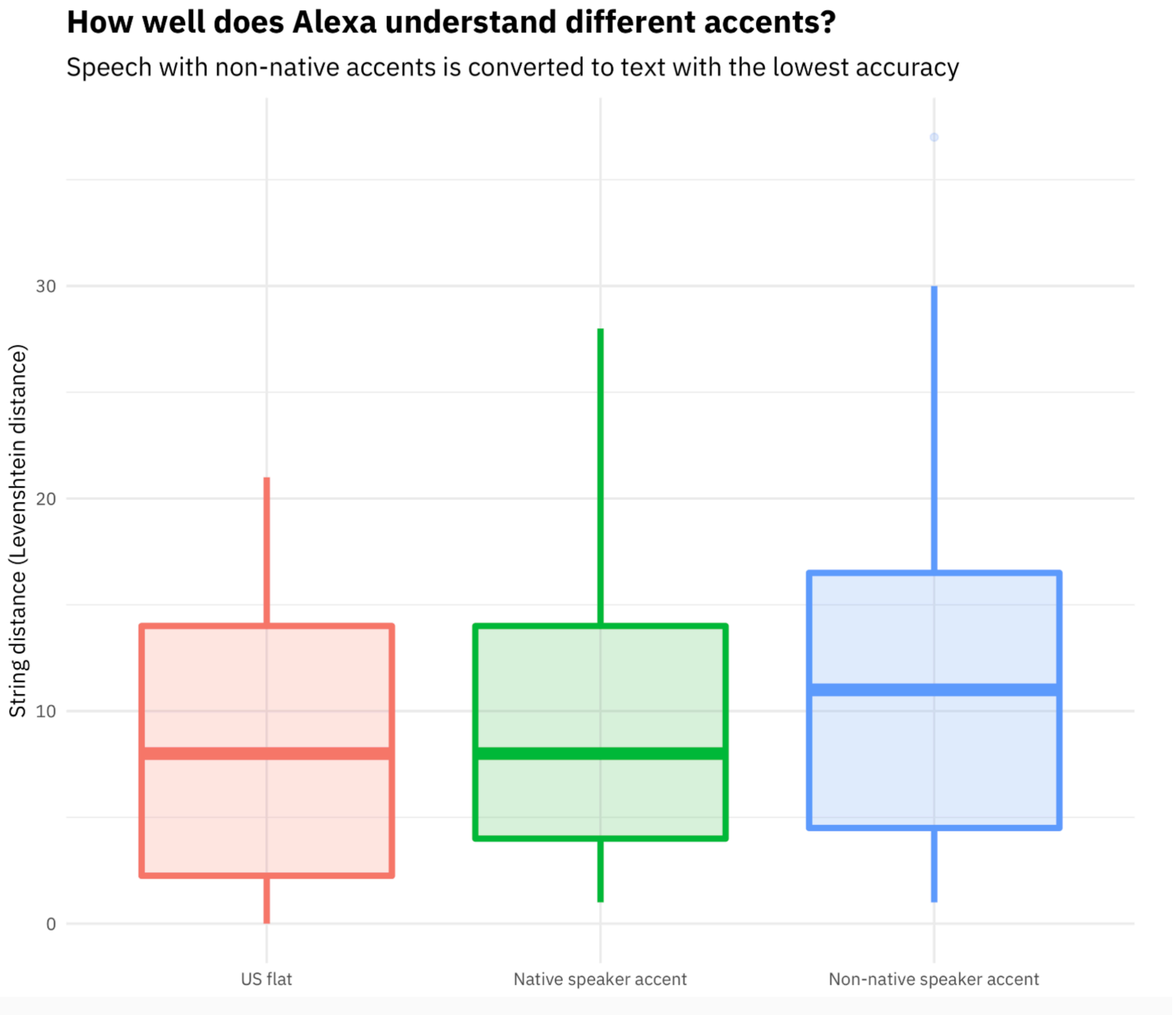
JULIA SILGE

BLOG ABOUT RESUME

AMAZON ALEXA AND ACCENTED ENGLISH

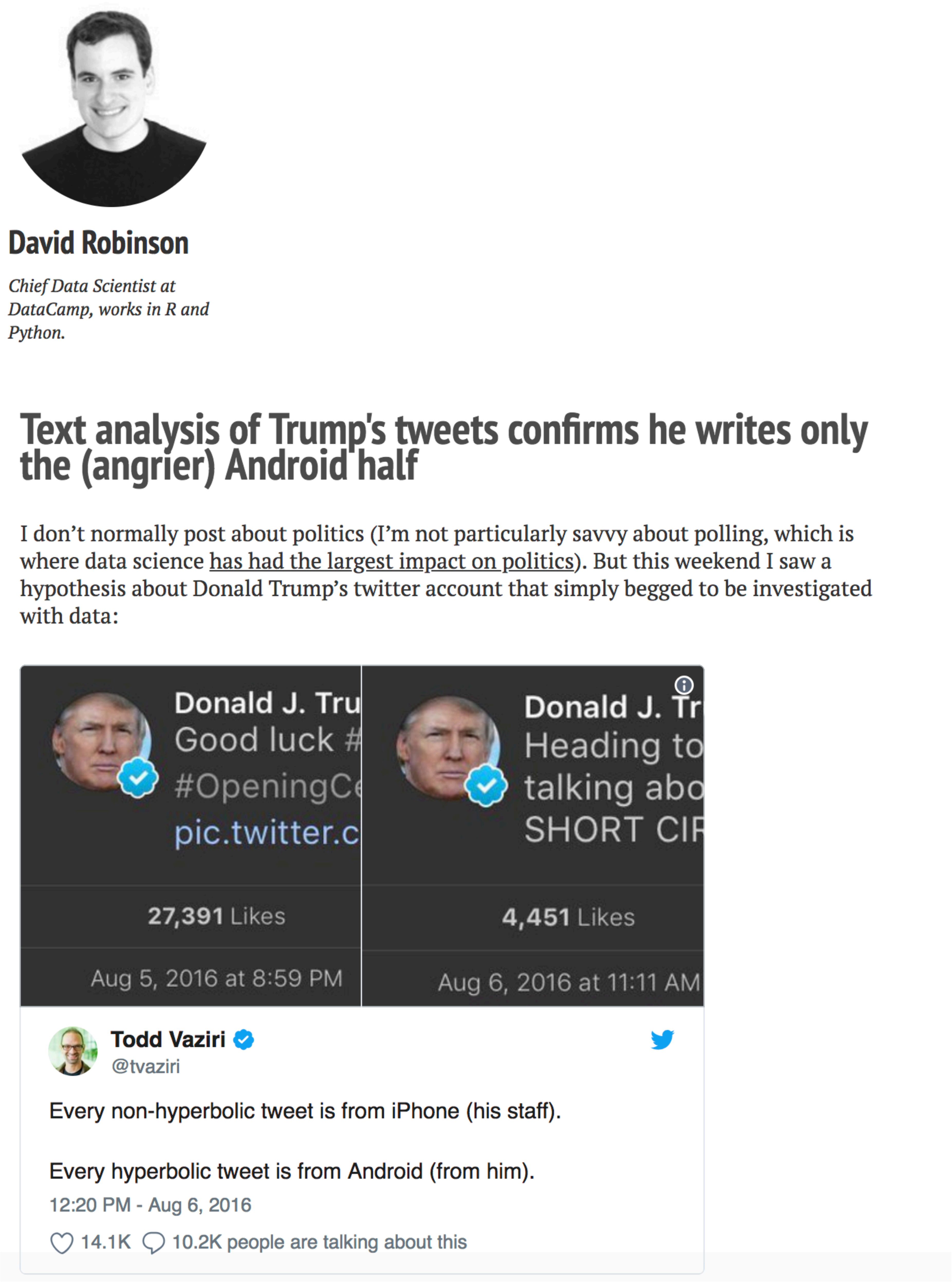
Jul 19, 2018 · 6 minute read · rstats

Earlier this spring, one of my data science friends here in SLC got in contact



How well does Alexa understand different accents?

Speech with non-native accents is converted to text with the lowest accuracy



David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:

Donald J. Tru
Good luck #
#OpeningCe
pic.twitter.c

27,391 Likes

Aug 5, 2016 at 8:59 PM

Donald J. Tr
Heading to
talking abo
SHORT CIF

4,451 Likes

Aug 6, 2016 at 11:11 AM

Todd Vaziri
@tvaziri

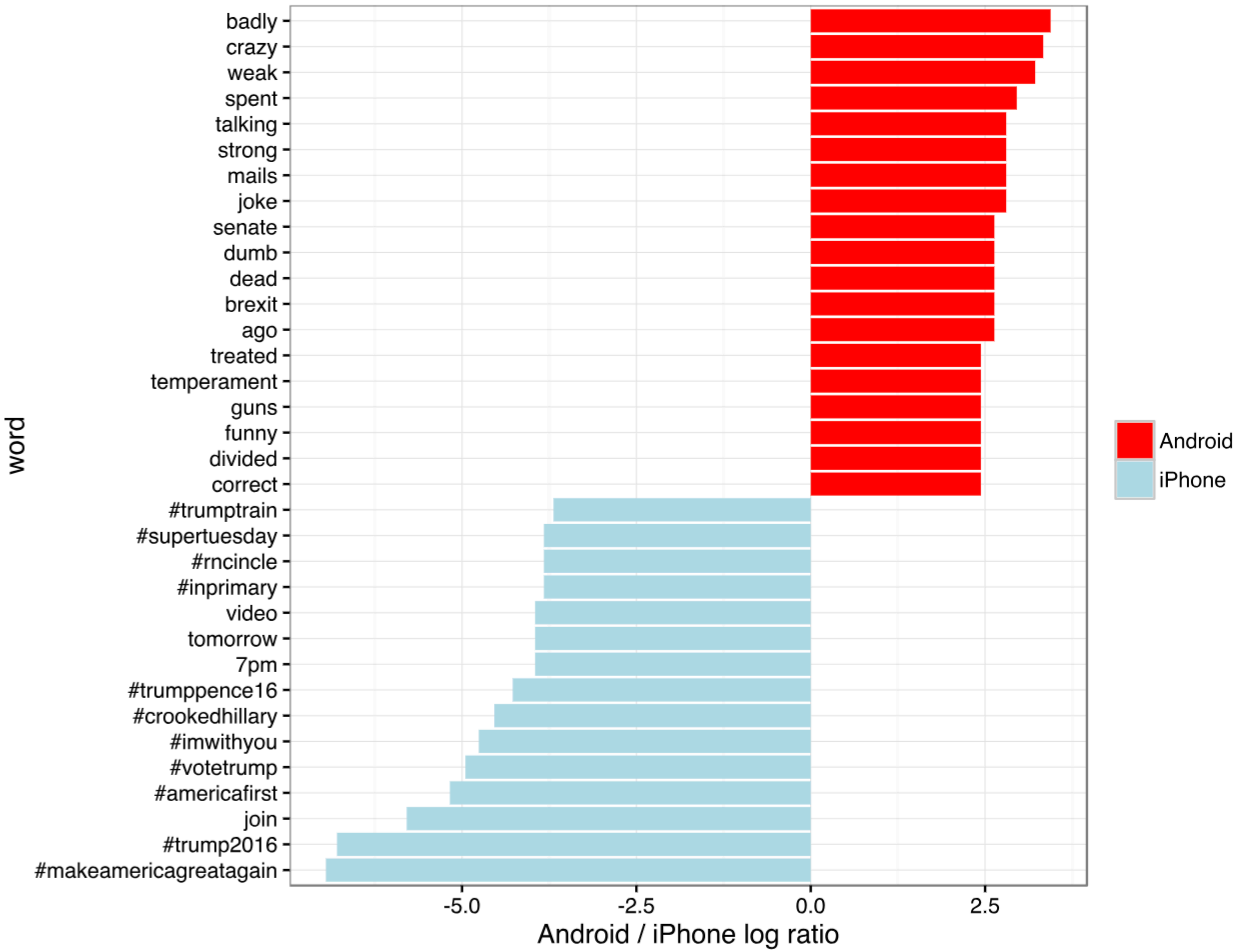
Every non-hyperbolic tweet is from iPhone (his staff).

Every hyperbolic tweet is from Android (from him).

12:20 PM - Aug 6, 2016

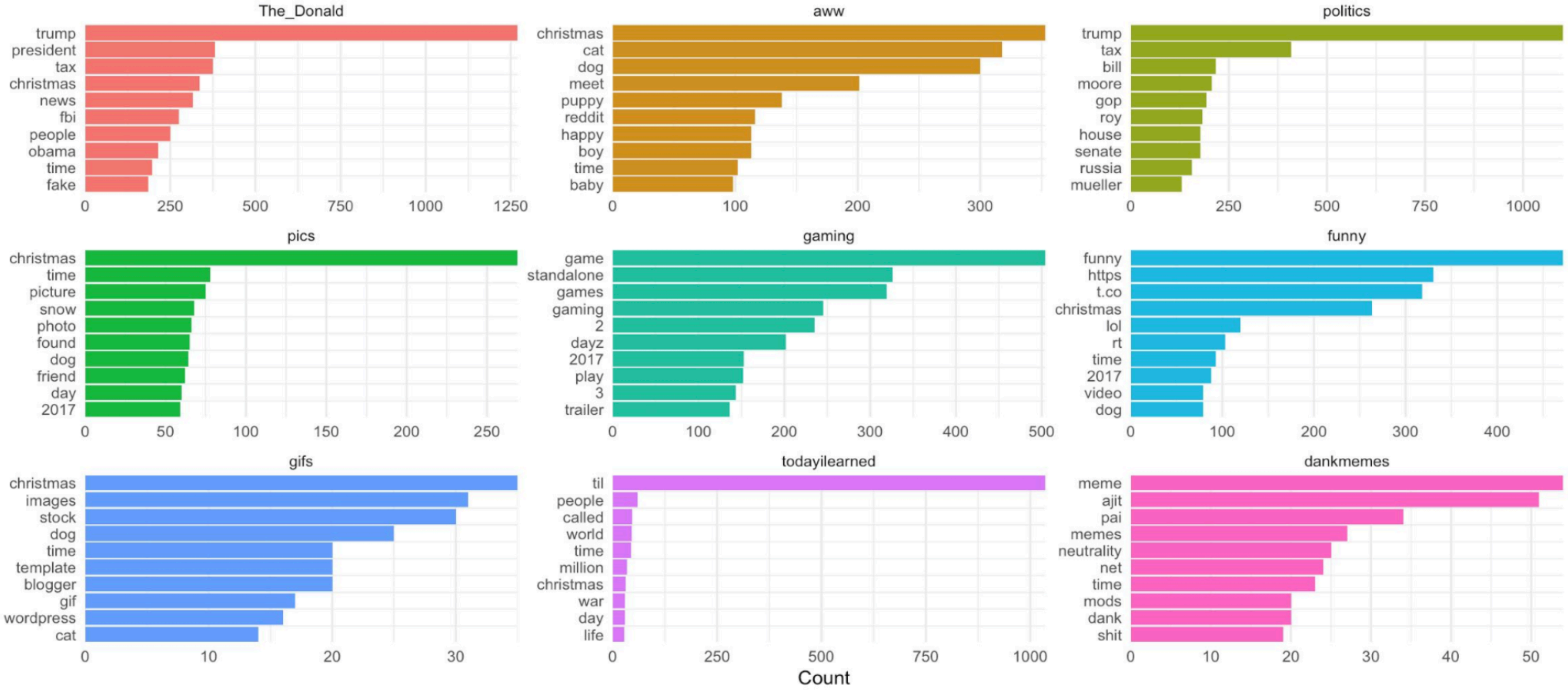
♡ 14.1K 💬 10.2K people are talking about this

Which are the words most likely to be from Android and most likely from iPhone?



How to Succeed on Reddit

Team InterstellR



Most frequent words within popular subreddits

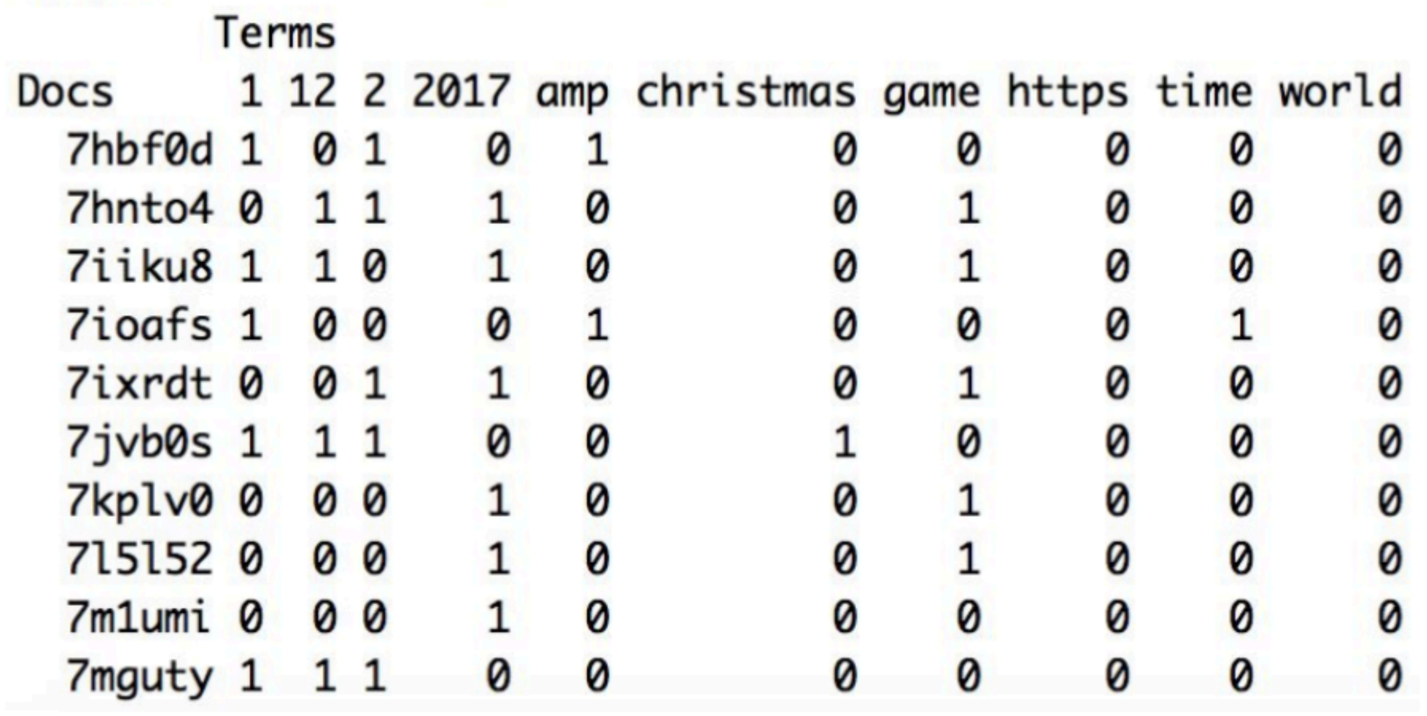
in December 2017

How to Succeed on Reddit

Team InterstellR

Modeling Popularity

- Multivariate linear model
- Target: score
- Predictors: subreddit, sentiment, dog_cat, [text features], ...
- Stepwise selection by AIC
- R^2 = 0.177



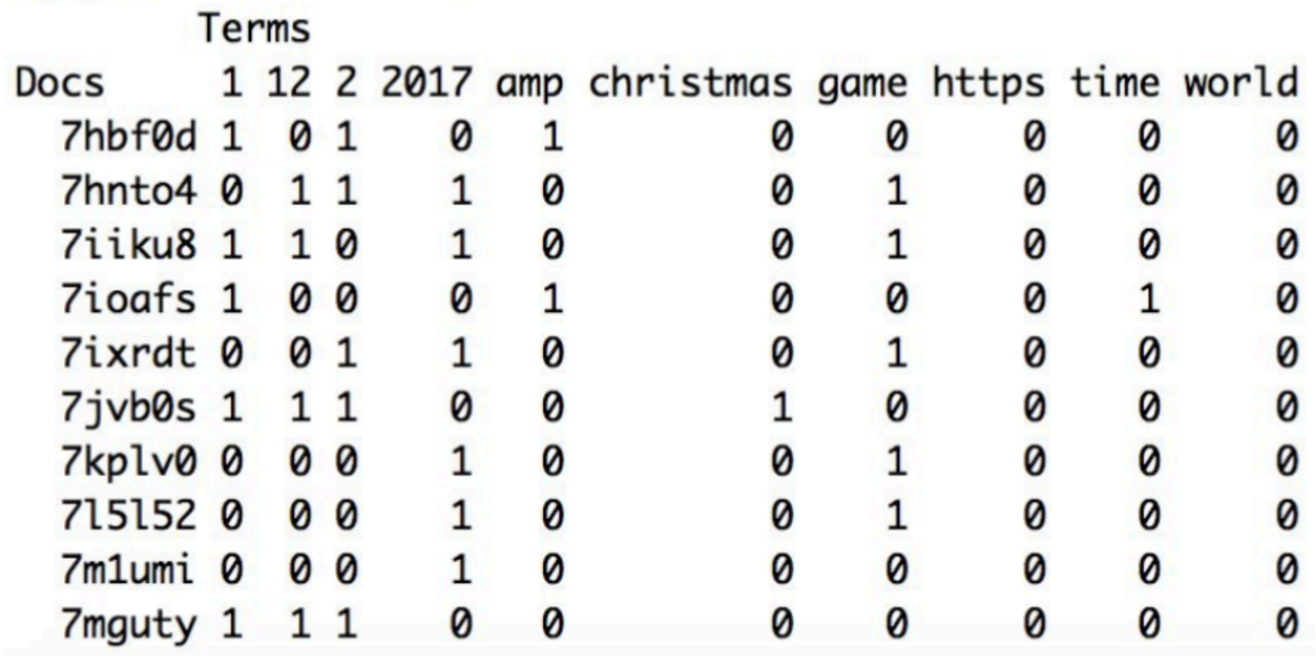
```
        Terms
Docs     1 12 2 2017 amp christmas game https time world
  7hbf0d 1  0 1    0   1         0    0     0    0     0
  7hnto4 0  1 1    1   0         0    1     0    0     0
  7iiku8 1  1 0    1   0         0    1     0    0     0
  7ioafs 1  0 0    0   1         0    0     0    1     0
  7ixrdt 0  0 1    1   0         0    1     0    0     0
  7jvb0s 1  1 1    0   0         1    0     0    0     0
  7kplv0 0  0 0    1   0         0    1     0    0     0
  7l5l52 0  0 0    1   0         0    1     0    0     0
  7m1umi 0  0 0    1   0         0    0     0    0     0
  7mguty 1  1 1    0   0         0    0     0    0     0
```

How to Succeed on Reddit

Team InterstellR

Modeling Popularity

- Multivariate linear model
- Target: score
- Predictors: subreddit, sentiment, dog_cat, [text features], ...
- Stepwise selection by AIC
- R^2 = 0.177



```
         Terms
Docs      1 12 2 2017 amp christmas game https time world
  7hbf0d 1  0 1    0   1         0    0     0    0     0
  7hnto4 0  1 1    1   0         0    1     0    0     0
  7iiku8 1  1 0    1   0         0    1     0    0     0
  7ioafs 1  0 0    0   1         0    0     0    1     0
  7ixrdt 0  0 1    1   0         0    1     0    0     0
  7jvb0s 1  1 1    0   0         1    0     0    0     0
  7kplv0 0  0 0    1   0         0    1     0    0     0
  7l5l52 0  0 0    1   0         0    1     0    0     0
  7m1umi 0  0 0    1   0         0    0     0    0     0
  7mguty 1  1 1    0   0         0    0     0    0     0
```

Conclusions

1. Be negative
2. Dogs and cats are both good choices
3. Post on /r/gifs
4. Don’t talk about December, games, and don’t ask questions
5. Do talk about home and news
6. Don’t use a linear model to predict Reddit post scores!

1b. have students to create a visualization

don’t start like this

- Install R
- Install RStudio
- Install the following packages:
  - rmarkdown
  - tidyverse
  - …
- Load these packages
- Install git

instead do this

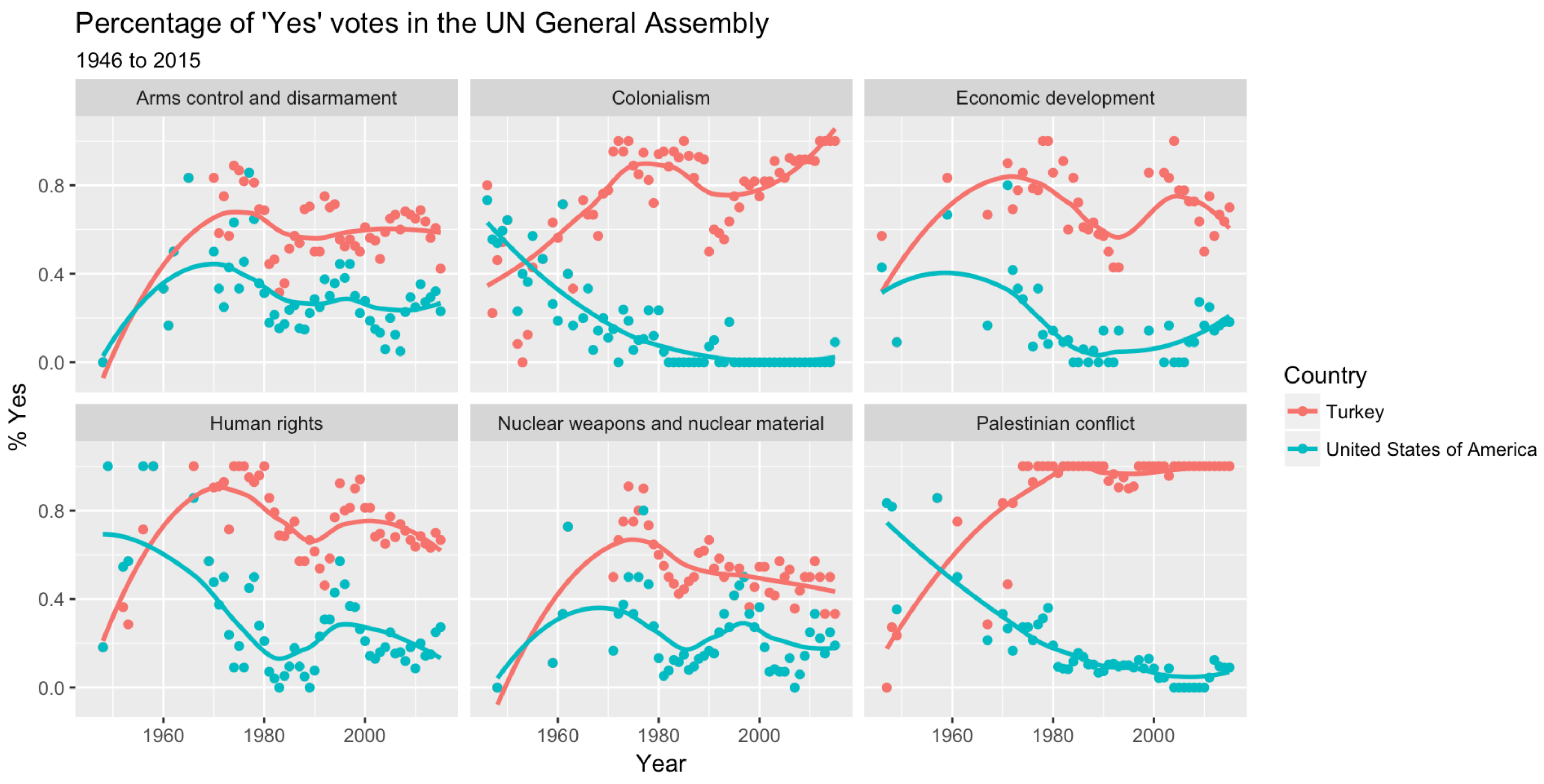
- Go to rstudio.cloud (or some other server based solution)
- Log in with your ID & pass

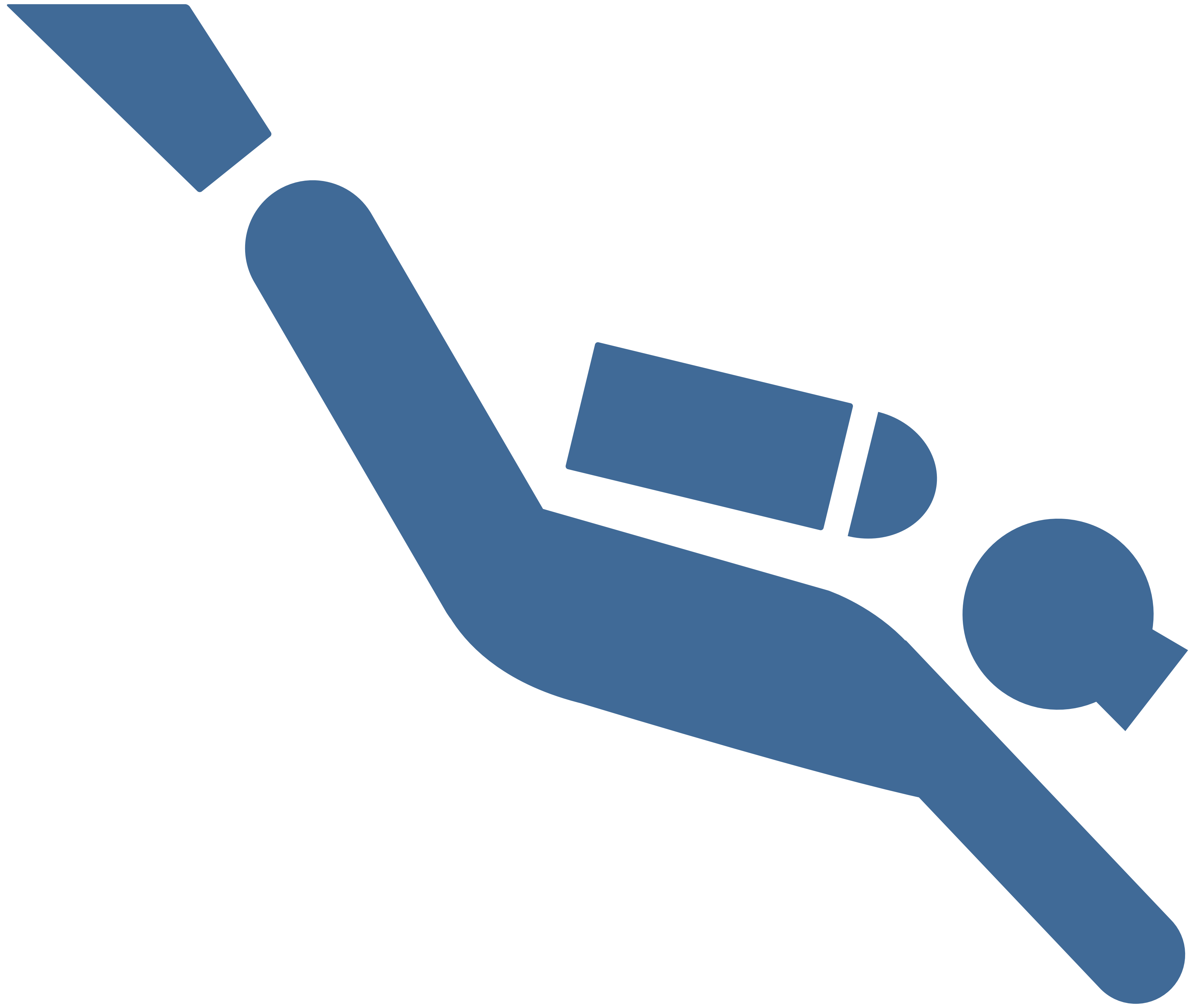
```
> hello R!
```

don’t start with this

```
class(mtcars$mpg)
#> [1] "numeric"
mean(mtcars$mpg)
#> [1] 20.09062
median(mtcars$mpg)
#> [1] 19.2
sd(mtcars$mpg)
#> [1] 6.026948
```

instead do this





- Go to bit.ly/ptt-rscloud and create an account to join the RStudio Cloud workspace for this workshop.
- Open the project called UN Votes.
- Open the R Markdown document called un-votes.Rmd, knit the document, view the result.
- Then, change countries plotted and knit again.

Resources:
Cloud computing for teaching

- **Frictionless onboarding to data science with RStudio Cloud** (Çetinkaya-Rundel)
  - Nitty-gritty of setting up your course on RStudio Cloud
  - Video and slides: https://www.causeweb.org/cause/ecots/ecots18/tech-talk/4
- **Infrastructure and tools for teaching computing throughout the statistical curriculum** (Çetinkaya-Rundel and Rundel)
  - Overview of cloud computing resources for teaching
  - Part of the Practical Data Science for Stats collection
  - https://peerj.com/preprints/3181/

2

rethink , don’t just add

don’t start with this

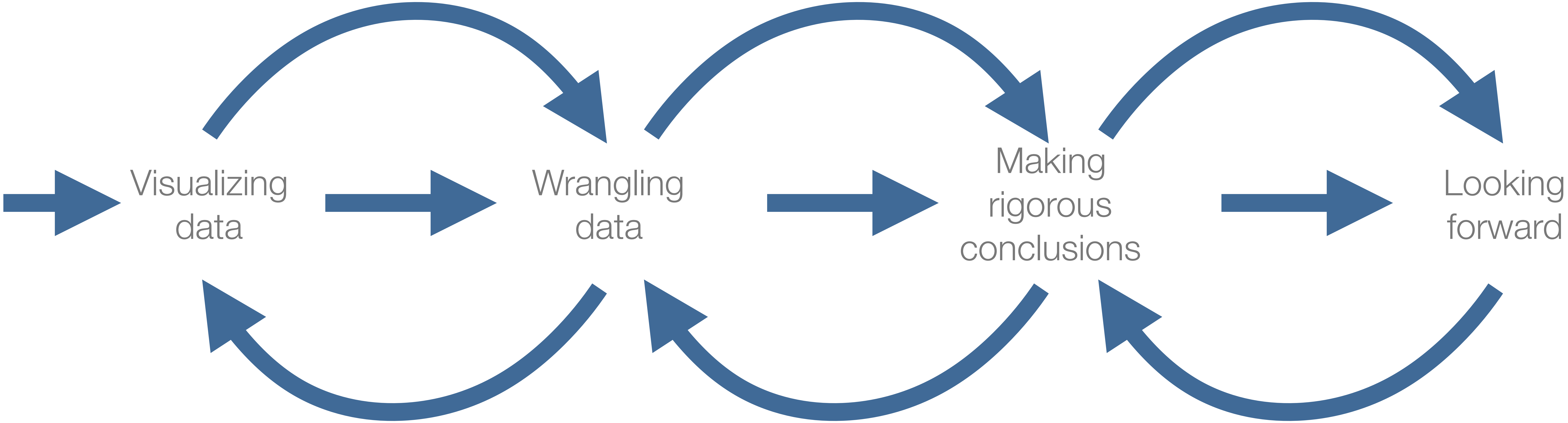
- Exploratory data analysis
- Study design
- Probability
- Random variables
- Central Limit Theorem
- One sample mean HT and CI
- One sample proportion HT and CI
- Two sample mean HT and CI
- Two sample proportion HT and CI
- Chi-square test
- ANOVA
- Simple linear regression

and add all this

+ R
+ R Markdown
+ git / GitHub
+ data scraping
+ iteration
+ working with non-rectangular data
+ interactive visualization

…

curriculum



Visualizing data

Wrangling data

Making rigorous conclusions

Looking forward

Fundamentals of data & data viz, revision exercises, confounding variables and Simpson's paradox (and git/GitHub)

Tidy data, data frames vs. summary tables, recoding and transforming variables, web scraping and iteration

Building and selecting models, visualizing interactions, prediction and model validation, inference via simulation & discussion of CLT

Interactive visualization and reporting with Shiny, Bayesian inference, text analysis, ???

3

stick with a consistent grammar

choose a grammar that grows with the complexity of the analysis

but that doesn’t require constantly climbing a steep learning curve

4 teach tools for good science

literate programming

version control

collaboration

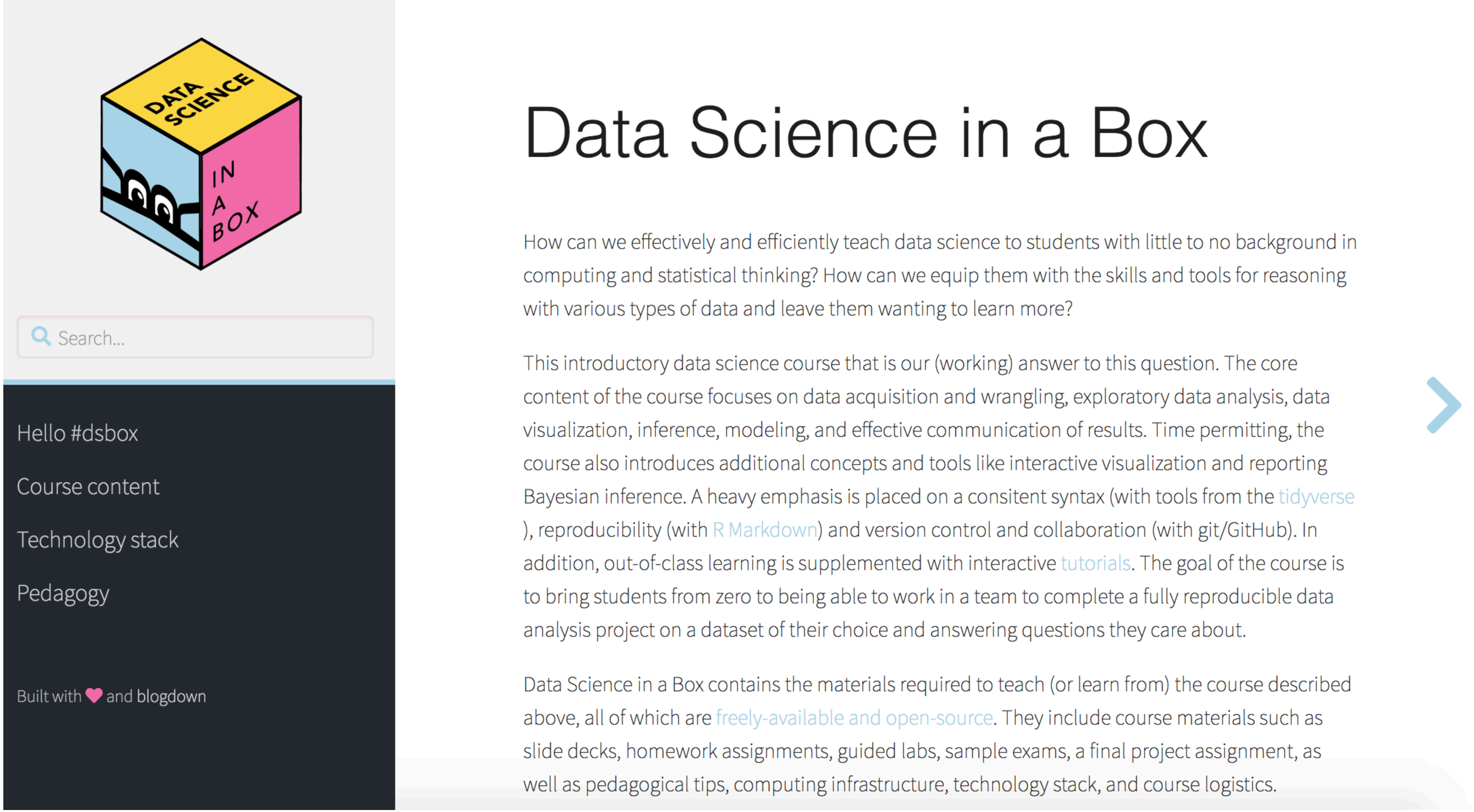
reproducibility

5

use real and relatable examples

bit.ly/dsbox-web

bit.ly/dsbox-repo

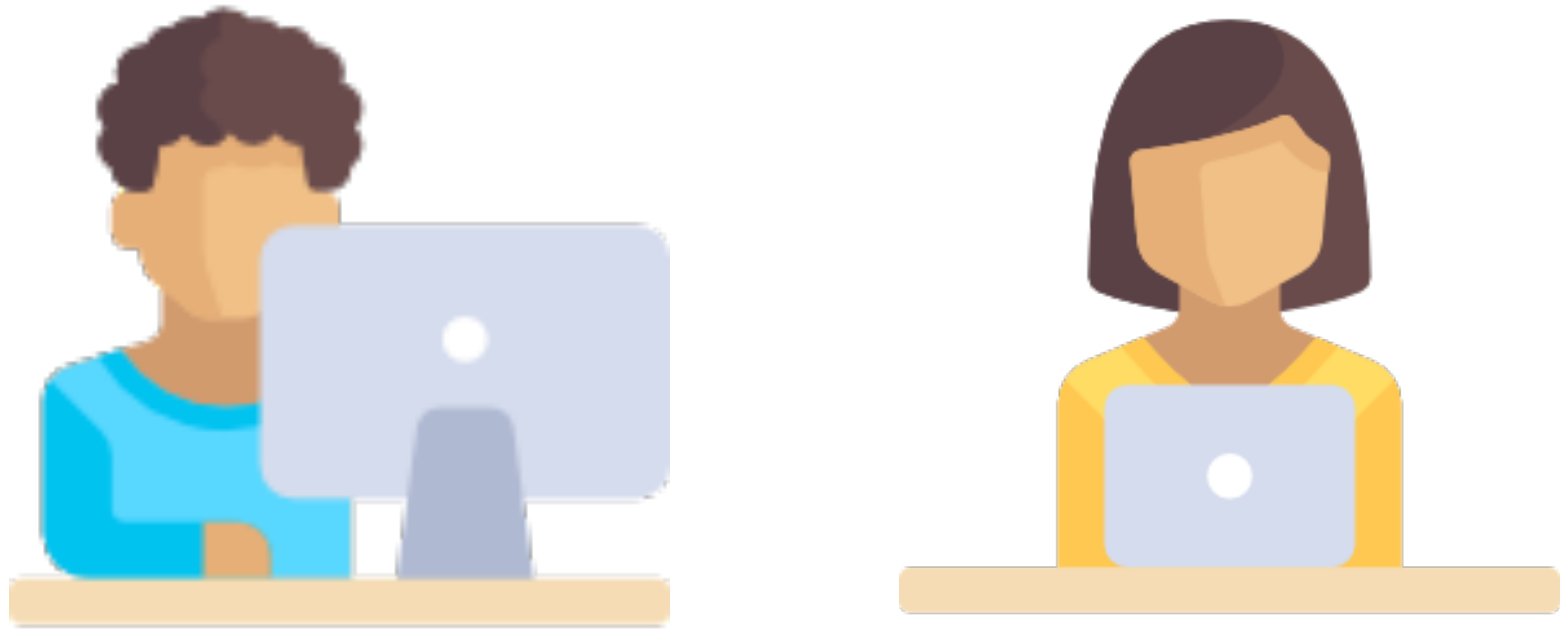


Data Science in a Box

How can we effectively and efficiently teach data science to students with little to no background in computing and statistical thinking? How can we equip them with the skills and tools for reasoning with various types of data and leave them wanting to learn more?

This introductory data science course that is our (working) answer to this question. The core content of the course focuses on data acquisition and wrangling, exploratory data analysis, data visualization, inference, modeling, and effective communication of results. Time permitting, the course also introduces additional concepts and tools like interactive visualization and reporting Bayesian inference. A heavy emphasis is placed on a consitent syntax (with tools from the tidyverse), reproducibility (with R Markdown) and version control and collaboration (with git/GitHub). In addition, out-of-class learning is supplemented with interactive tutorials. The goal of the course is to bring students from zero to being able to work in a team to complete a fully reproducible data analysis project on a dataset of their choice and answering questions they care about.

Data Science in a Box contains the materials required to teach (or learn from) the course described above, all of which are freely-available and open-source. They include course materials such as slide decks, homework assignments, guided labs, sample exams, a final project assignment, as well as pedagogical tips, computing infrastructure, technology stack, and course logistics.



slides	x 26
labs	x 10
assignments	x 6
project	x 1
exams	x 2

course infrastructure

using the tech stack

lesson plans

pedagogical tips

dsbox

Datasets for the Data Science Course in a Box

```
install_github(“rstudio-education/dsbox”)
```