DATA SCIENCE, OPEN SOURCE, & RADICAL TRANSPARENCY

Andrew Heiss, PhD
Georgia State University
October 10, 2019
@andrewheiss

DATA, DATA SCIENCE, & PUBLIC SERVICE

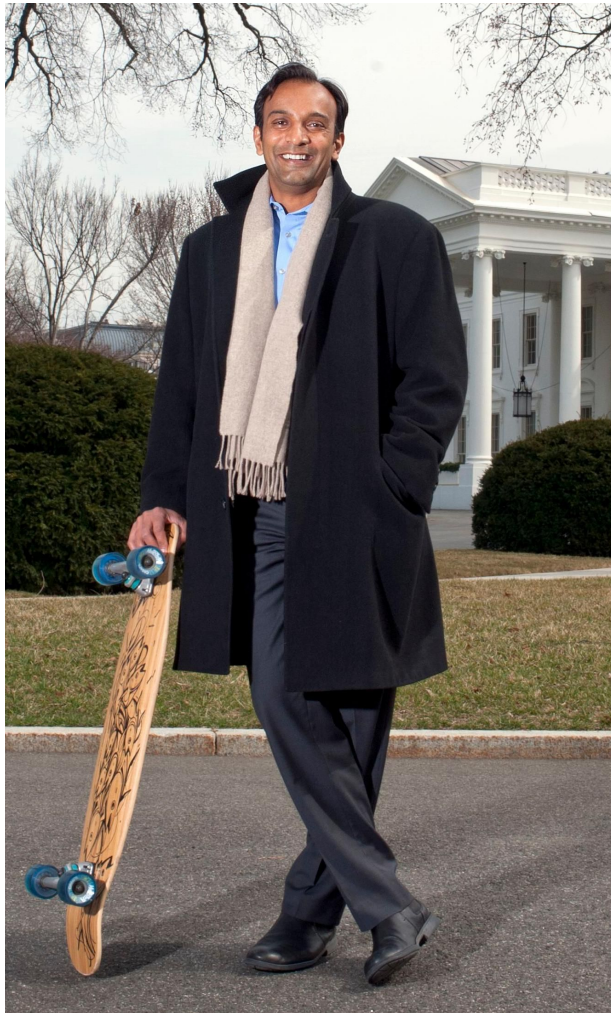
SUSAN CRAWFORD IDEAS 08.22.18 07:00 AM

WHY UNIVERSITIES NEED 'PUBLIC INTEREST TECHNOLOGY' COURSES

POLICYMAKERS AT ALL levels of government are struggling to thoughtfully harness data in the service of public values. Many public servants grew up in an era of firmly separate disciplines: You were either an engineer or an economist, either a programmer or a social worker, but never both. In an era in which data is everything, the risks to core democratic principles—equity, fairness, support for the most vulnerable, delivery of effective government services—caused by technological illiteracy in policymakers, and policy illiteracy in computer scientists, are staggering.

field aimed at addressing precisely this gap in interdisciplinary opportunities. This new area, "public interest technology," is still being defined; it encompasses designing public policy and laws with an awareness of how technology actually works, as well as ensuring that technology is being used to serve public values of fairness and equity. It means consciously thinking about the welfare of society in general, rather than the incentives of a single company.

DATA AND GOVERNMENT



"To responsibly unleash the power of data to benefit all Americans"

The White House

Office of the Press Secretary

For Immediate Release June 30, 2016

FACT SHEET: Launching the Data-Driven Justice Initiative: Disrupting the Cycle of Incarceration

"[O]ur criminal justice system isn't as smart as it should be. It's not keeping us as safe as it should be. It is not as fair as it should be.
Mass incarceration makes our country worse off, and we need to do something about it." –

President Barack Obama, July 14, 2015


M. HARRISON
NEWS WITH A TWIST

SUNLIGHT FOUNDATION

OPEN KNOWLEDGE INTERNATIONAL

U.S. CITY OPEN DATA CENSUS

POWERED BY OPEN DATA CENSUS

Atlanta, GA

0% Open

0% Score

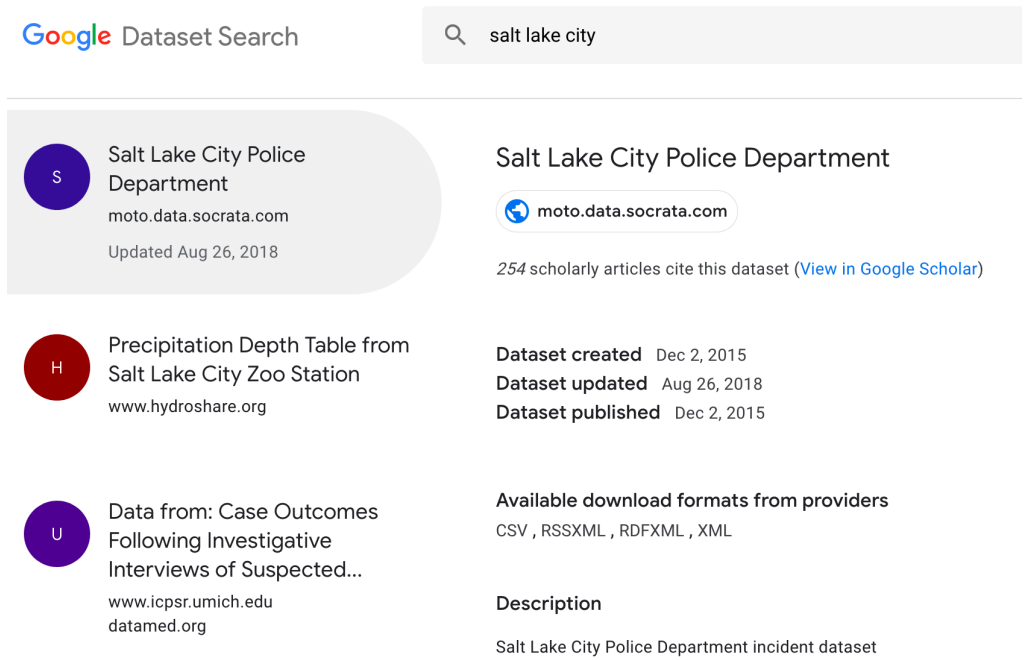
Breakdown

Dataset	Breakdown	Year	Score
Budget			
Business Listings			
Code Violations			
Construction Permits			
Crime Reports			
Emergency Calls			
Employee Salaries			
Lobbyist Activity			
Parcels			
Police Use-of-Force			
Procurement Contracts			
Property Assessment			
Property Transfers			
Public Facilities			
Restaurant Inspections			
Service Requests			
Spending			

Google Dataset Search Beta

Search for Datasets

Try boston education data or weather site:noaa.gov

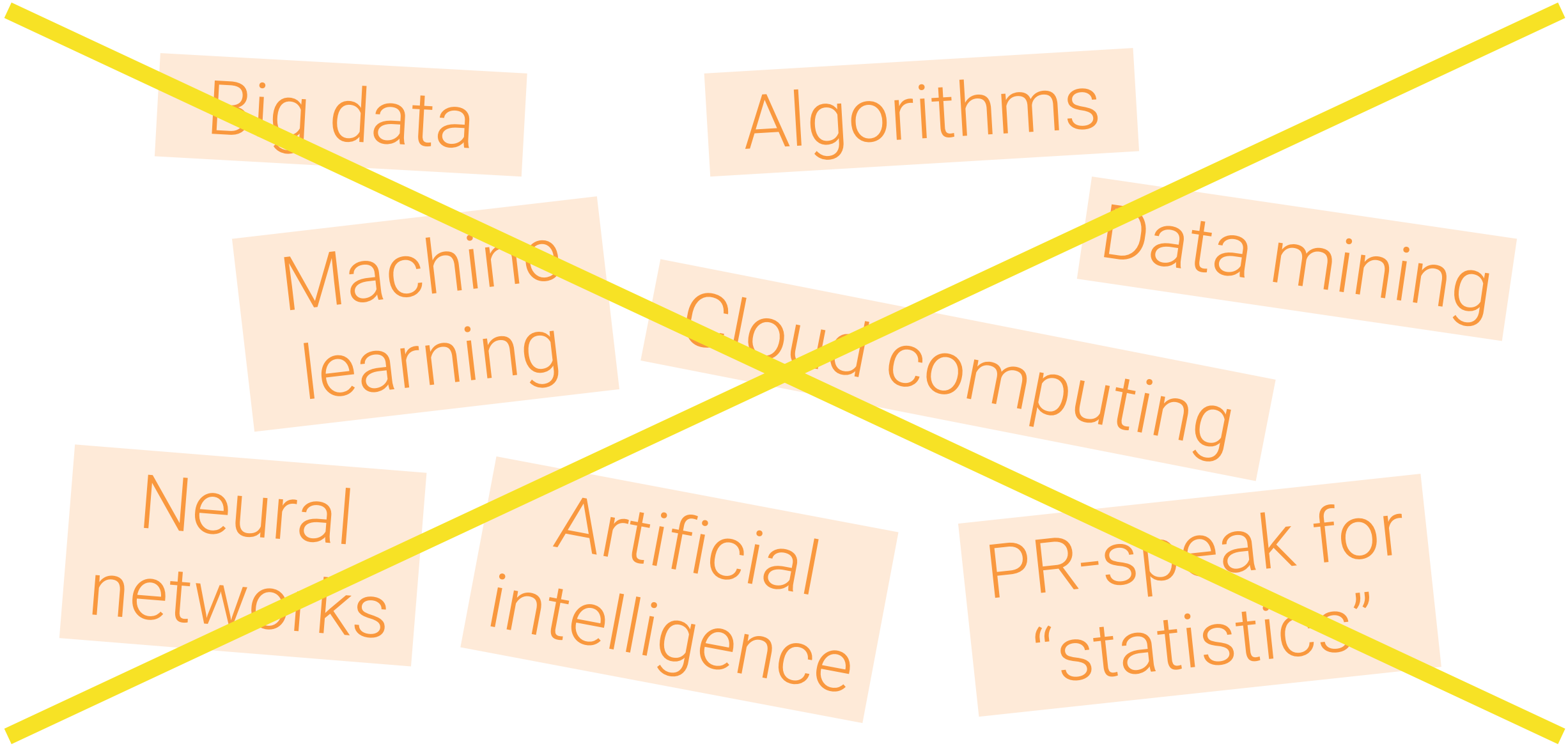


How do you use all this data to make the world better?

WHAT IS "STATISTICS"?

Collecting and analyzing data from a representative sample in order to make inferences about a whole population

WHAT IS "DATA SCIENCE"?

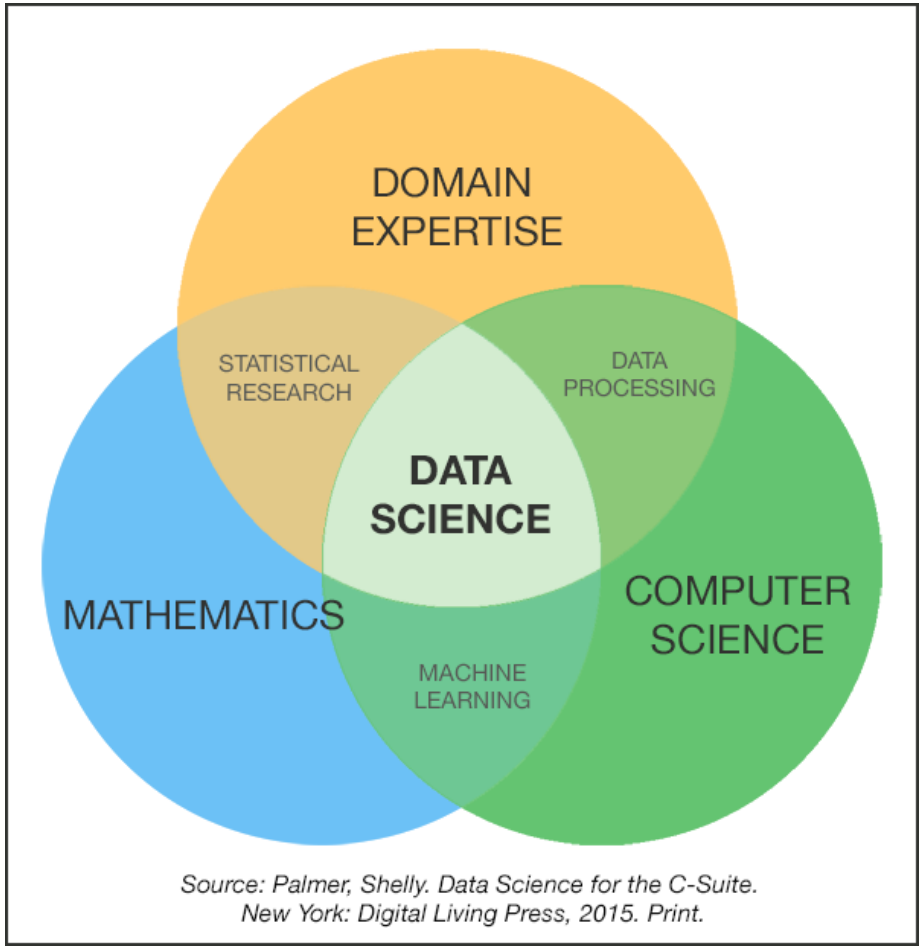


WHAT IS "DATA SCIENCE"?

Turning raw data into understanding, insight, and knowledge

Collect | **Analyze** | **Communicate**

WHAT'S THE DIFFERENCE?



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Statistics

Collect

Analyze

Communicate

HOW DO I LEARN HOW TO DO THIS STUFF?

Find excuses to use data science methods!

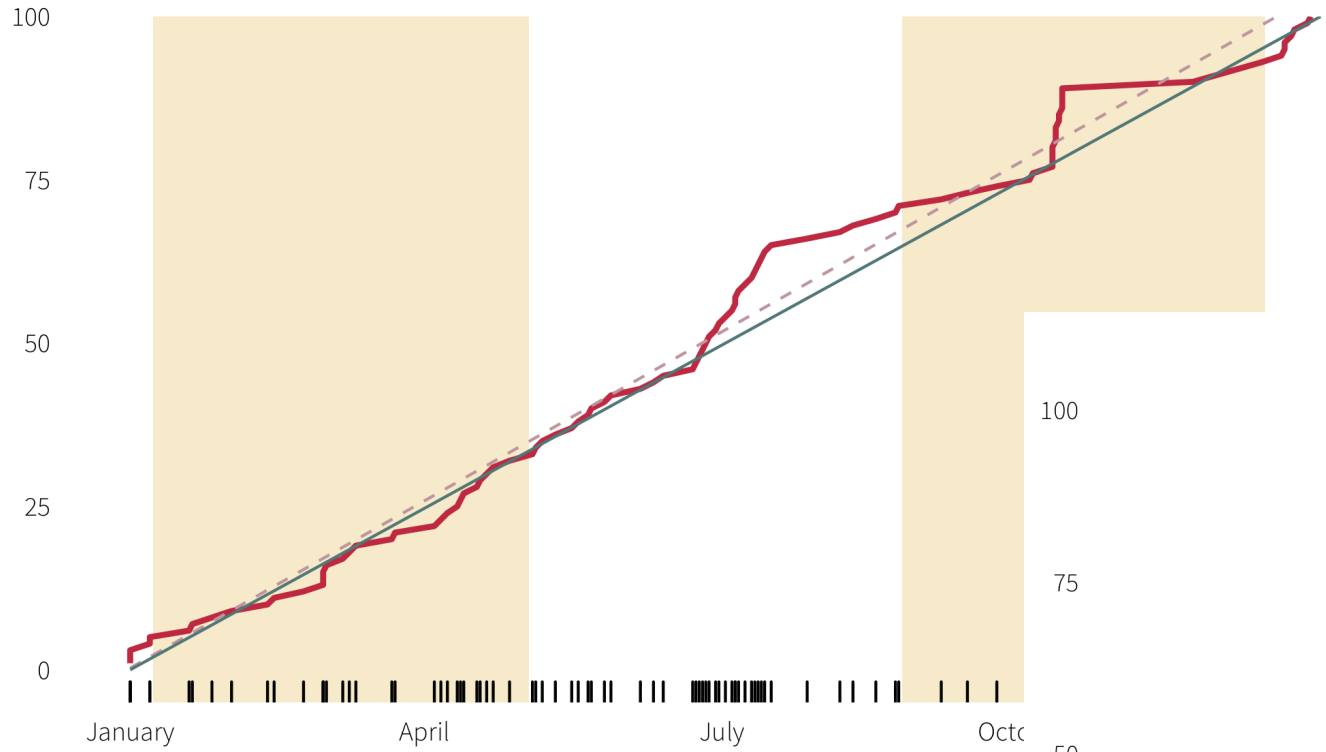
Dumb dinky projects

Data play time

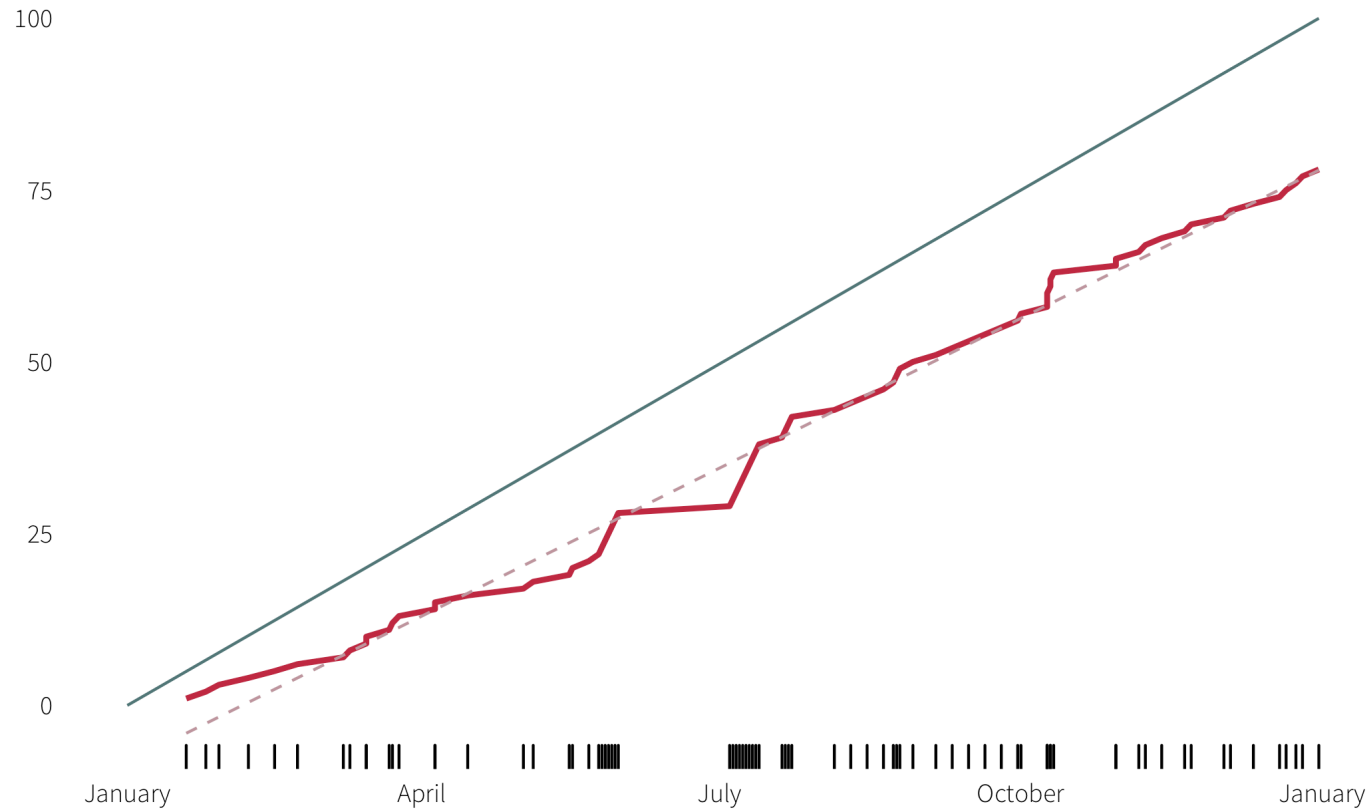
Actual projects

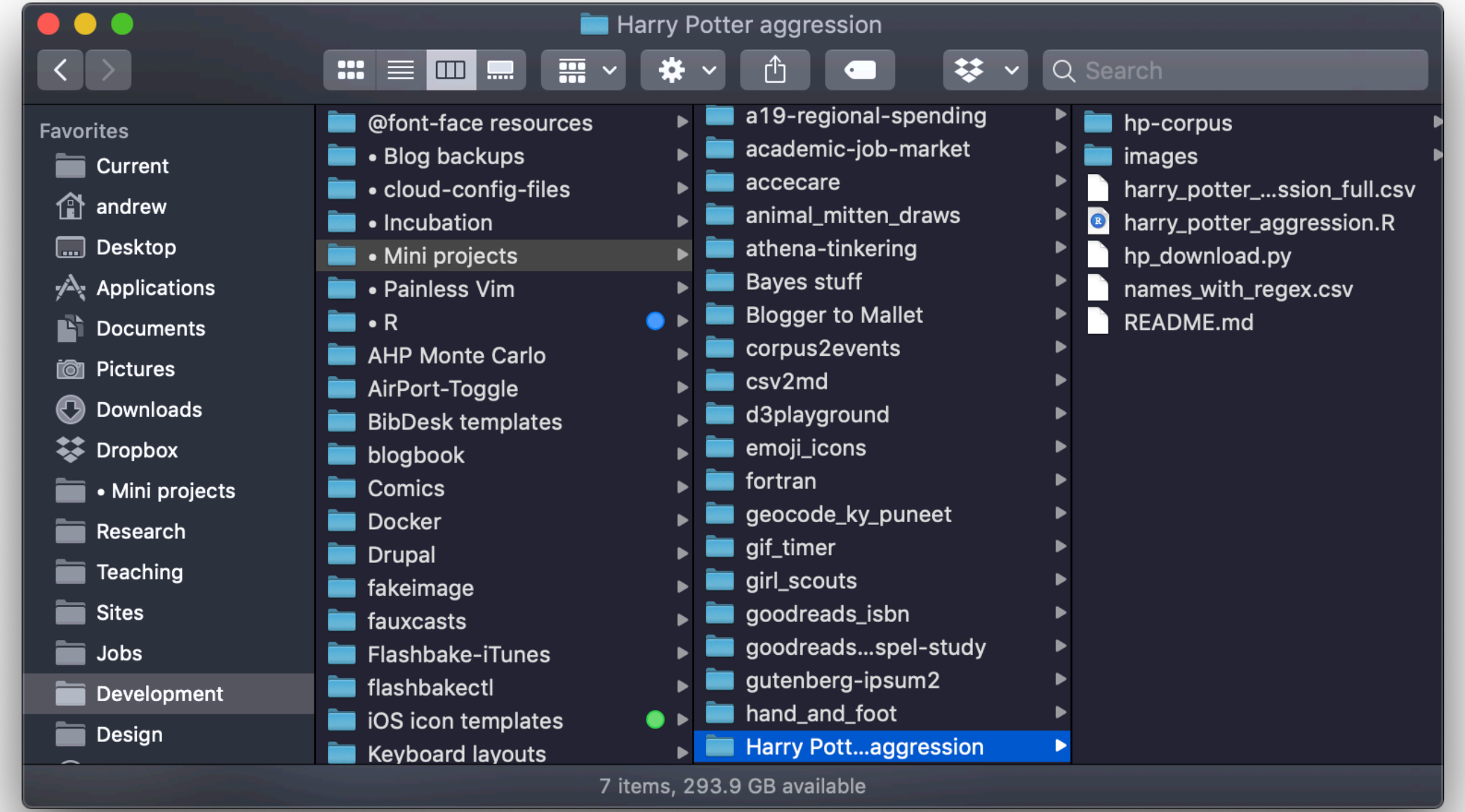
Cumulative number of family walks in 2014

Duke semesters shaded in yellow

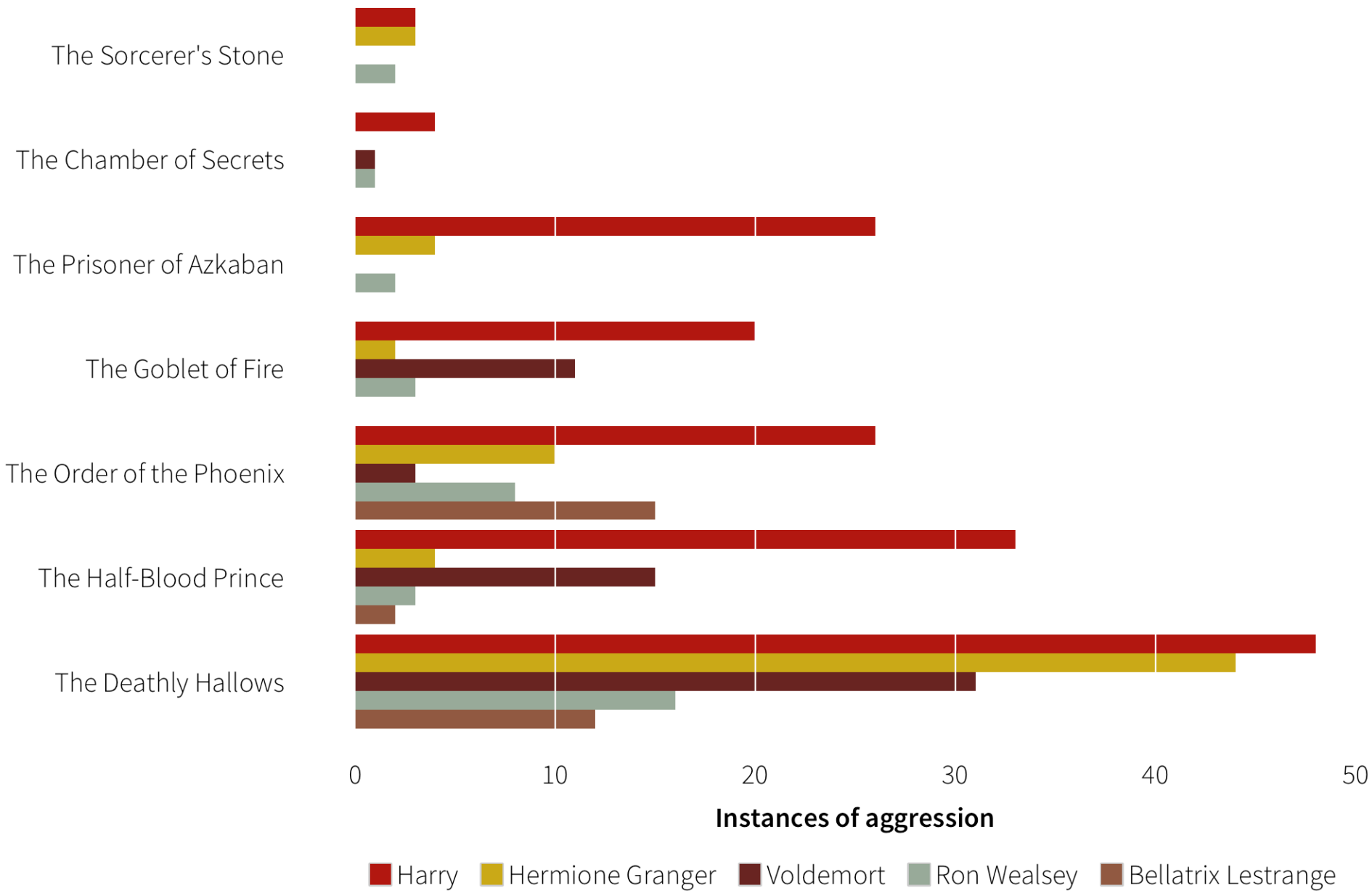


Cumulative number of family walks in 2015




Harry Potter aggression
Search
Favorites
Current
andrew
Desktop
Applications
Documents
Pictures
Downloads
Dropbox
• Mini projects
Research
Teaching
Sites
Jobs
Development
Design
@font-face resources
• Blog backups
• cloud-config-files
• Incubation
• Mini projects
• Painless Vim
• R
AHP Monte Carlo
AirPort-Toggle
BibDesk templates
blogbook
Comics
Docker
Drupal
fakeimage
fauxcasts
Flashbake-iTunes
flashbakectl
iOS icon templates
Keyboard layouts
a19-regional-spending
academic-job-market
accecare
animal_mitten_draws
athena-tinkering
Bayes stuff
Blogger to Mallet
corpus2events
csv2md
d3playground
emoji_icons
fortran
geocode_ky_puneet
gif_timer
girl_scouts
goodreads_isbn
goodreads...spel-study
gutenberg-ipsum2
hand_and_foot
Harry Pott...aggression
hp-corpus
images
harry_potter_...ssion_full.csv
harry_potter_aggression.R
hp_download.py
names_with_regex.csv
README.md
7 items, 293.9 GB available

Most aggressive characters in the Harry Potter series



Take classes!

Teach yourself!

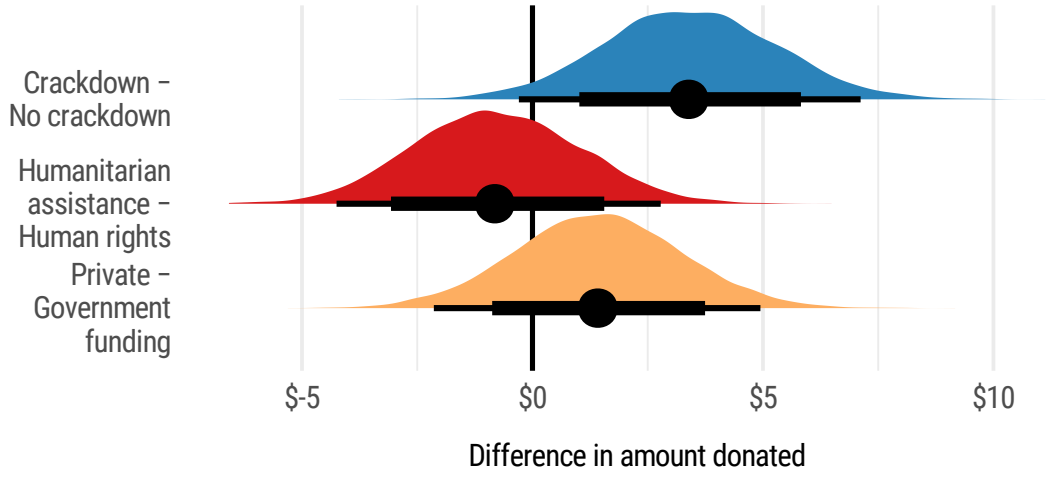
Conference sessions

Coursera + online tutorials

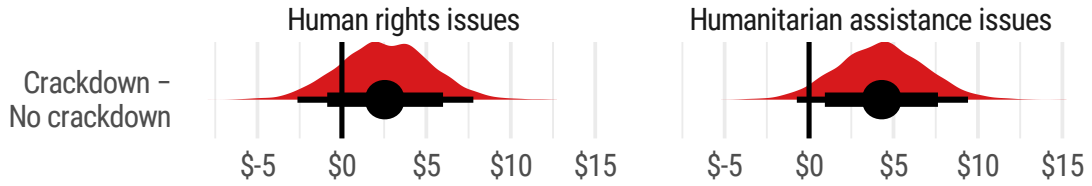
Read books


Topics
egypt (p_w)
peopl (p_w)
egyptian (p_w)
protest (p_w)
tahrir_squar (p_w)
court (p_w)
right (p_w)
case (p_w)
constitu (p_w)
brotherhood (p_w)
militar (p_w)
scaf (p_w)
politic (p_w)
mubarak (p_w)
brotherhood (p_w)
Documents
Maspero interrogation continues, virginity checks case adjourned
December 0 / 2011 ,13 Comments / 3 Views
CAIRO: An investigations judge began Tuesday interrogating 29 defendants
Egypt political forces call for mass 'Eyes of Freedom' rally Friday
Rejection of President Morsi's new constitutional declaration will likely take centre stage in planned
Osman El Sharnou
Egyptian workers
Post–revolutionary Egypt
Environmental issues
Development
Youth in the street
Sexual violence
Christian issues
Religious issues
Morsi and press freedom
Elections
National government
Business and government
Legislation
Draft constitution
Human rights and civil society
Protests and clashes
Police arrests
Police torture
SCAF
Trials
Al–Ahram English
Daily News Egypt
Egypt Independent

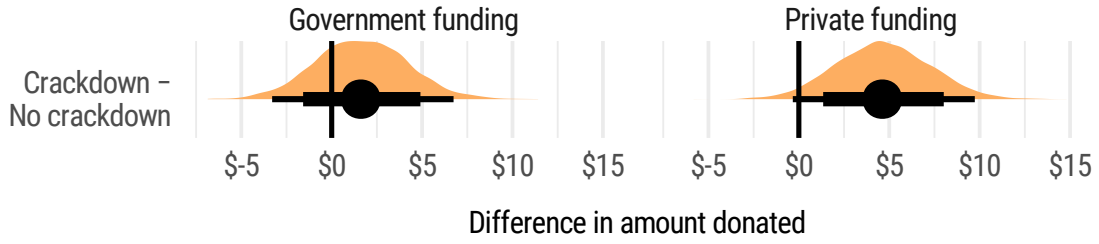
A



B



C



D

$$x_{\text{group 1, group 2}} \sim \text{Student } t(\nu, \mu, \sigma) \quad \text{[likelihood]}$$

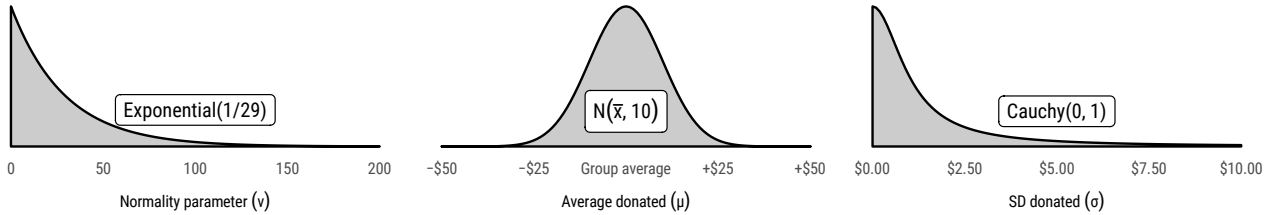
$$\text{Difference} = x_{\text{group 2}} - x_{\text{group 1}} \quad \text{[difference in means]}$$

$$x : \text{Mean amount donated}$$

$$\nu \sim \text{Exponential}(1/29) \quad \text{[prior normality]}$$

$$\mu_{\text{group 1, group 2}} \sim \mathcal{N}(\bar{x}_{\text{group 1, group 2}}, 10) \quad \text{[prior donation mean per group]}$$

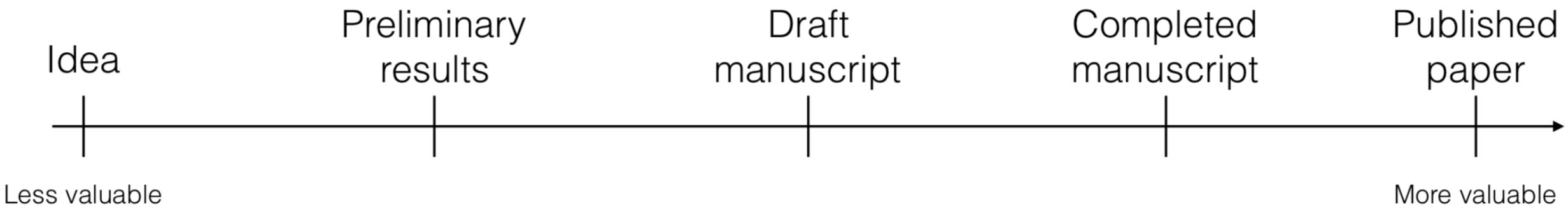
$$\sigma_{\text{group 1, group 2}} \sim \text{Cauchy}(0, 1) \quad \text{[prior donation sd per group]}$$



ds4ps.org

RADICAL TRANSPARENCY & PUBLIC WORK

How we normally think of our work and goals



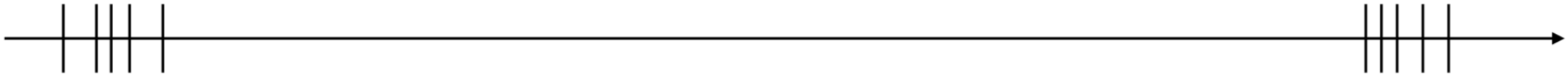
How we should think of our work and goals

Anything still on your computer

(Data, code, results, draft, finished paper)

Anything out in the world

(Paper, preprint, product, blog post, open source, tweet)



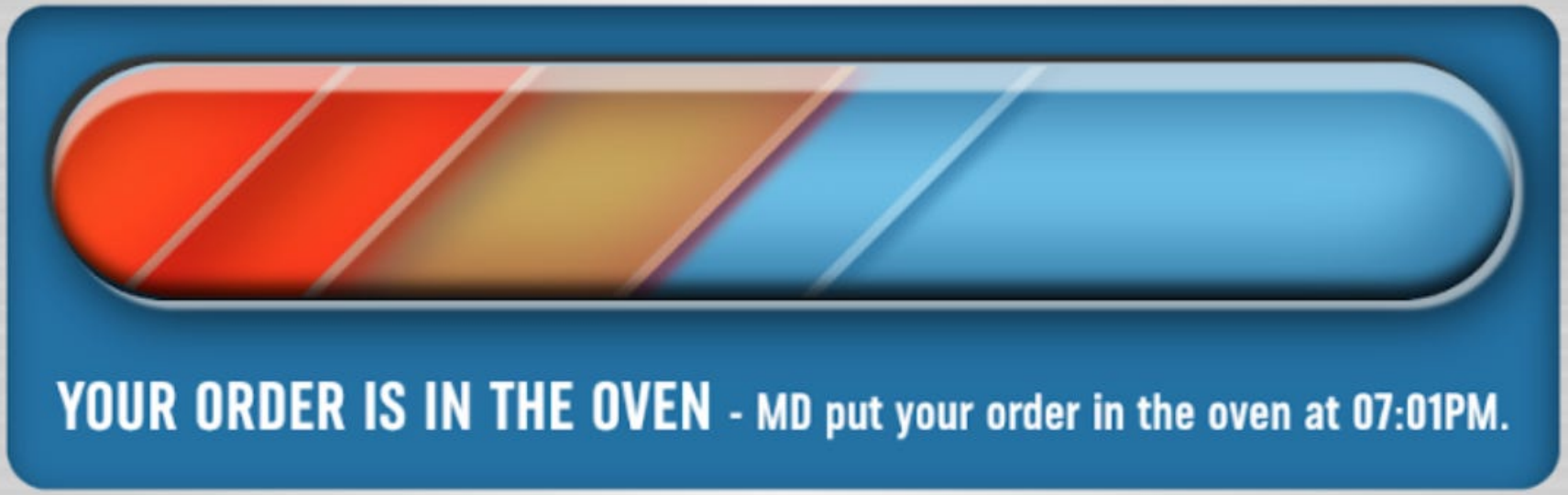
Less valuable

More valuable

Track Your Order

DOMINO'S TRACKER®

Know the status of your order, from the moment it's placed to the second it leaves our store for delivery or is ready to be picked up.



PATENT PENDING


MENU
Neighborhood Improvement Tracker
Search for an address or intersection:
ex '8515 W Chicago', 'Van Dyke & Harper'
Residential Demolitions
Completed (source)
Contracted (source)
Pipeline (source)
Commercial Demolitions
Completed (source)
Contracted (source)
Pipeline (source)
DLBA Sold Properties
Partner & Project Sales (source)
Own It Now (source)
Closed Auctions (source)
Side Lots (source)
Currently for sale (source)
DLBA Owned Structure (source)
DLBA Owned Properties
DLBA Owned Vacant Land (source)
Side Lots For Sale (source)
Building Permits
Building Permits (source)
Dalrymple St
Bordeau St
Welton St
Woodrow St
Lipton Ave
Chenier St
Belleterre Ave
Dundee St
Grand River Ave
W Chicago St
Nardin Park
Yosemite St
Yellowstone St
Cascade Ave
W Boston Blvd
Joy Rd
© Mapbox © OpenStreetMap Improve this map

Benefits of working in public

Build reputation

Learn more

Grow community

Early feedback on ideas

Validation

2016-17

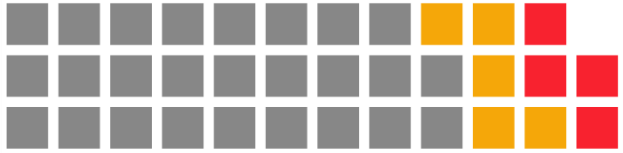
Political science (43)

Public administration and policy (41)

2017-18

Political science (11)

Public administration and policy (31)



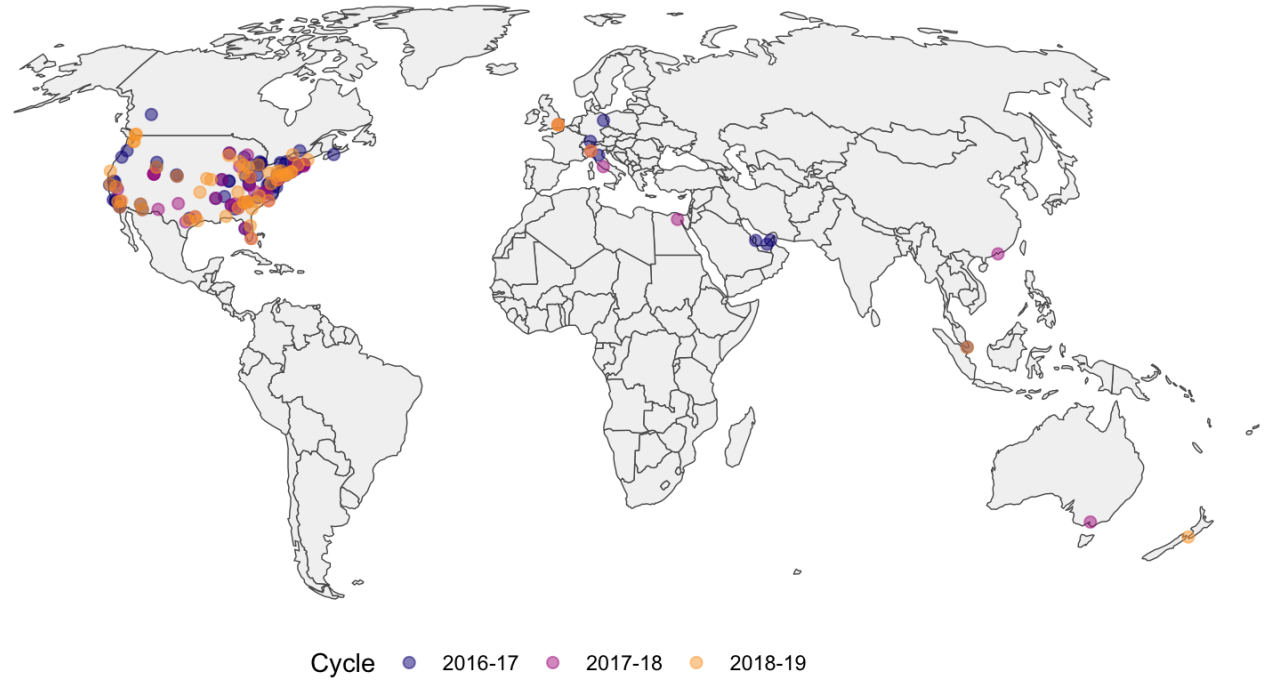
2018-19

Political science (37)

Public administration and policy (23)

Nothing Skype, no flyout Flyout, no offer Visiting offer Tenure-track offer

One box = one job posting



Andrew Heiss

International NGOs, nonprofit management, authoritarianism, data science, and R

ORCID iD: 0000-0002-3948-3914

PGP public • PGP fingerprint:
4AA2 FA83 A8B2 05A4 E30F
610D 1382 6216 9178 36AB

Monday, December 17, 2018

The academic job search finally comes to an end

I am *so beyond thrilled* to announce that I'll be joining the Andrew Young School of Policy Studies at Georgia State University in Fall 2019 as an assistant professor in the Department of Public Management and Policy. I'll be teaching classes in statistics/data science, economics, and nonprofit management in beautiful downtown Atlanta, and we'll be moving back to the South. I am so so excited about this! The Andrew Young School does amazing work in public policy, administration, and nonprofit management, and I'll be working with phenomenal colleagues and students. I still can't believe this is real.

Part of the reason I'm in shock is that for the past 2.5 years, I've been ripped apart and destroyed by the academic job market. This job market is a horrendous beast of a thing. It is soul-crushing and dream-shattering and a constant stream of rejection. While facing rejection is good and builds grit etc., etc., in reality it's awful.

In an effort to stay On Brand™, here are a bunch of fancy graphs and numbers showing what it's been like to apply for nearly 200 jobs since August 2016. Unlike many of my other blog posts, I haven't included any of the code to generate these. That code is all available in a GitHub repository (see `README.Rmd`), along with the raw data that I've collected over the past few years (for the morbidly curious).

Application count and outcomes

Between August 31, 2016 and November 18, 2018, I applied for 186 tenure-track and non-tenure-track academic jobs at R1 schools, liberal arts colleges, and teaching-focused public universities. I was offered one two-year visiting assistant professorship at the Romney

```
---
title: "The academic job search finally comes to an end"
output: github_document
editor_options:
  chunk_output_type: console
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE, fig.retina = 2)
```

> See the [actual blog post](https://www.andrewheiss.com/blog/2018/12/17/academic-job-market-visualized/).

I am *so beyond thrilled* to announce that I'll be joining the [Andrew Young School of Policy Studies](https://aysps.gsu.edu/) a

Part of the reason I'm in shock is that for the past 2.5 years, I've been ripped apart and destroyed by the academic job market

In an effort to stay On Brand™, here are a bunch of fancy graphs and numbers showing what it's been like to apply for nearly 20

```{r load-libraries-data, warning=FALSE, message=FALSE}
library(tidyverse)
library(lubridate)
library(here)
library(sf)
library(waffle)
library(ggstance)
library(scales)
library(countrycode)
# library(mapview)  # For interactive maps!
library(units)
library(patchwork)

# Load jobs data
jobs_clean <- read_csv(here("data", "jobs_clean.csv")) %>%
  mutate_at(vars(`Skype interview`, `Flyout`, contains("ffer")),
            funs(bin = !is.na(.)))
```
```

How to work in public

Blog and tweet

Analyze data

Blog about research

Teach concepts

Make research public

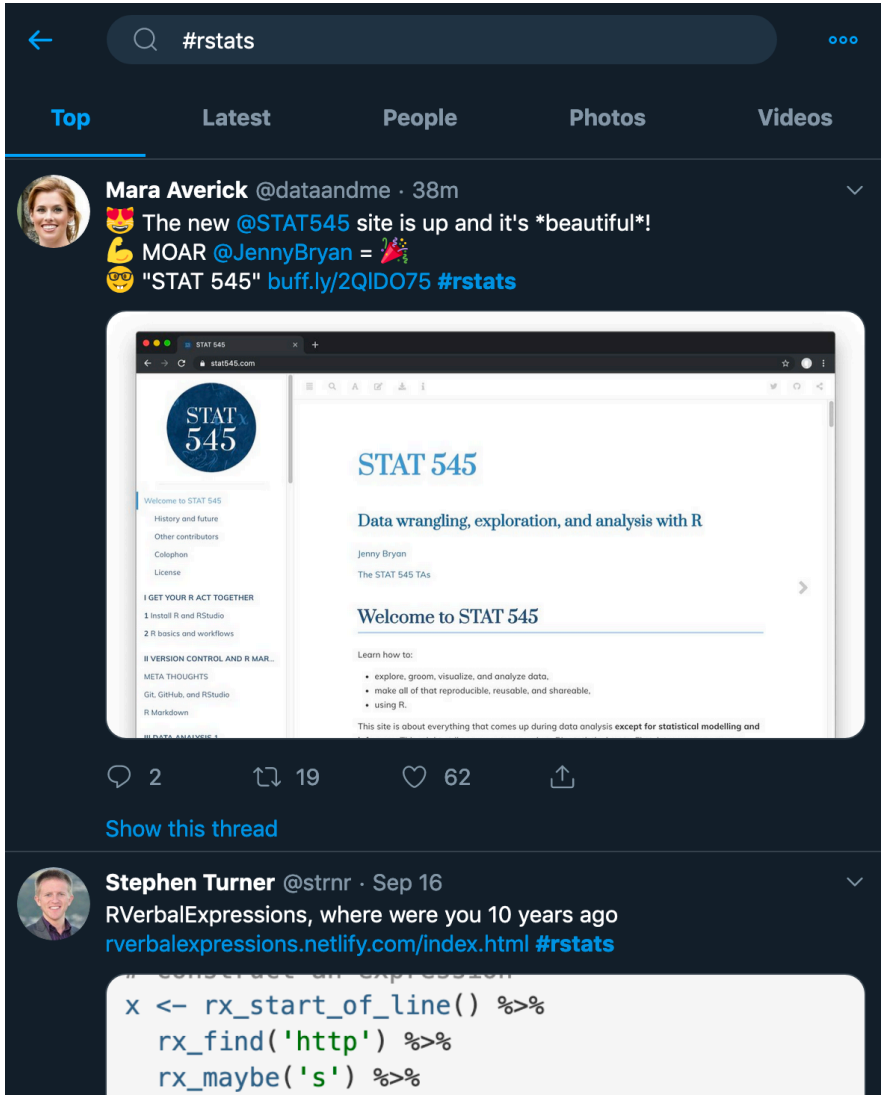
Start a blog

We introduce an R package, **blogdown**, in this short book, to teach you how to create websites using R Markdown and Hugo. If you have experience with creating websites, you may naturally ask what the benefits of using R Markdown are, and how **blogdown** is different from existing popular website platforms, such as WordPress. There are two major highlights of **blogdown**:

1. It produces a static website, meaning the website only consists of static files such as HTML, CSS, JavaScript, and images, etc. You can host the website on any web server (see Chapter 3 for details). The website does not require server-side scripts such as PHP or databases like WordPress does. It is just one folder of static files. We will explain more benefits of static websites in Chapter 2, when we introduce the static website generator Hugo.

2. The website is generated from R Markdown documents (R is optional, i.e., you can use plain Markdown documents without R code chunks). This brings a huge amount of benefits, especially if your website is related to data analysis or (R) programming. Being able to use Markdown implies simplicity and more importantly, *portability* (e.g., you are giving yourself the chance to convert your blog posts to PDF and publish to journals or even books in the future). R Markdown gives you the benefits of dynamic documents — all your results, such as tables, graphics, and inline values, can

Finding online communities



Analyze data

Saturday, August 26, 2017

Quickly play with Polity IV and OECD data (and see the danger of US democracy)

The Polity IV Project released new data yesterday, with democratization scores for 169 countries up to 2016. I wanted to check if the ongoing erosion of US democratic institutions since the 2016 elections registered in the US's Polity score, and, lo and behold, it did! We dropped from our solid, historically consistent 10 to an 8.

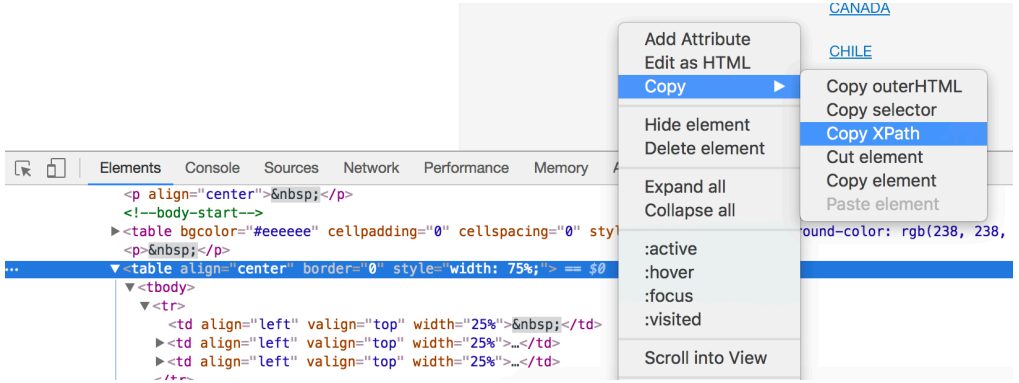
But is that bad? How does that compare to other advanced democracies, like countries in the OECD?

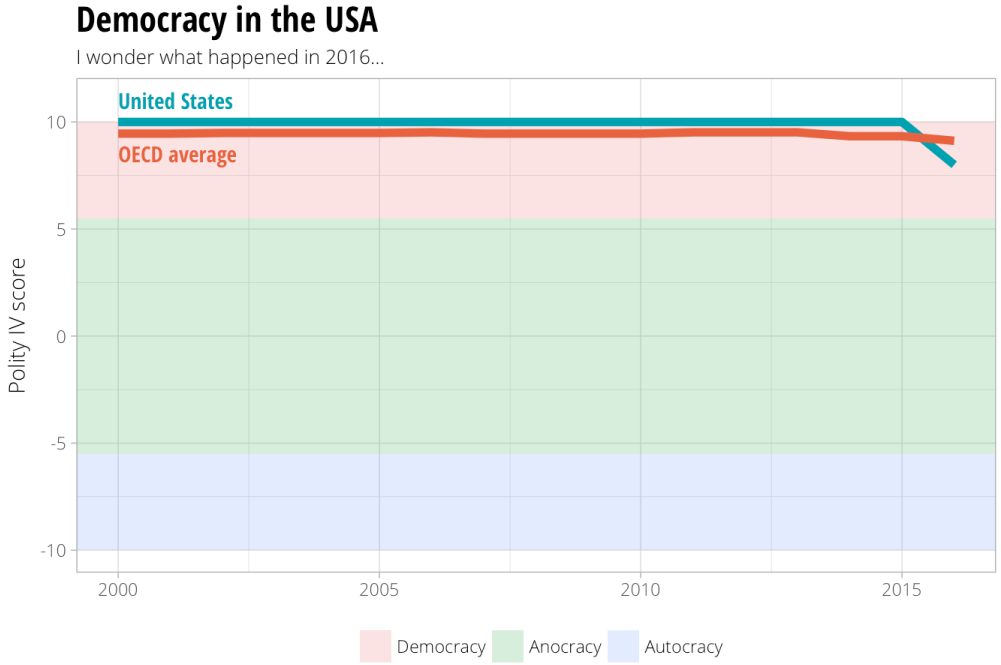
What follows below shows how relatively easy it is to quickly and reproducibly grab the new data, graph it, and compare scores across countries. (This notebook is also in a GitHub repository.)

Before we start, we'll load all the libraries we'll need:

```
library(tidyverse)      # dplyr, ggplot, etc.
library(readxl)         # Read Excel files
library(forcats)        # Deal with factors
library(countrycode)    # Deal with country codes and names
library(rvest)          # Scrape websites
library(httr)           # Download stuff
library(ggrepel)        # Place non-overlapping labels on plots
```

First, we have to download the new Polity data. We could navigate to the Polity IV data page and download the data manually, but that's not scriptable. Instead, we can use `GET()` from





Source: Polity IV Project

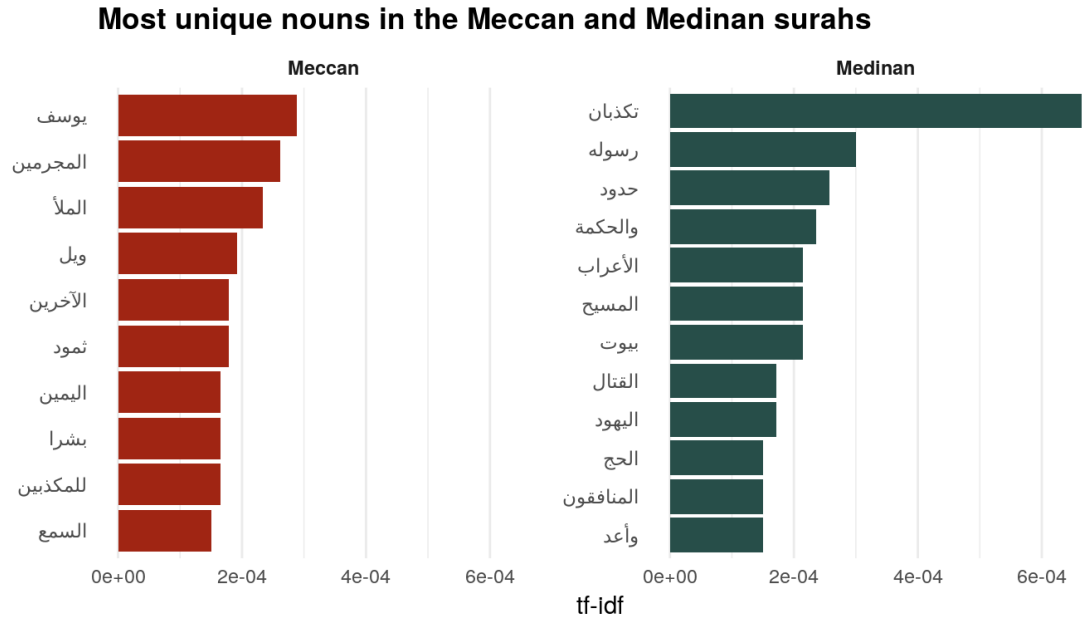
Friday, December 28, 2018

Tidy text, parts of speech, and unique words in the Qur'an

(See this notebook on GitHub)

As I showed in a previous blog post, the **cleanNLP** package is a phenomenal frontend for natural language processing in R. Rather than learn the exact syntax for NLP packages like spaCy or CoreNLP, you can use a consistent set of functions and let **cleanNLP** handle the API translation behind the scenes for you.

Previously, I used spaCy to tag the parts of speech in the Four Gospels to find the most distinctive nouns and verbs in the Gospel of John. Here, I'll show a quick example of how to use CoreNLP to tag parts of speech in Arabic. CoreNLP is far far far slower than spaCy, but it can handle languages like Arabic and Chinese, which is pretty magical.



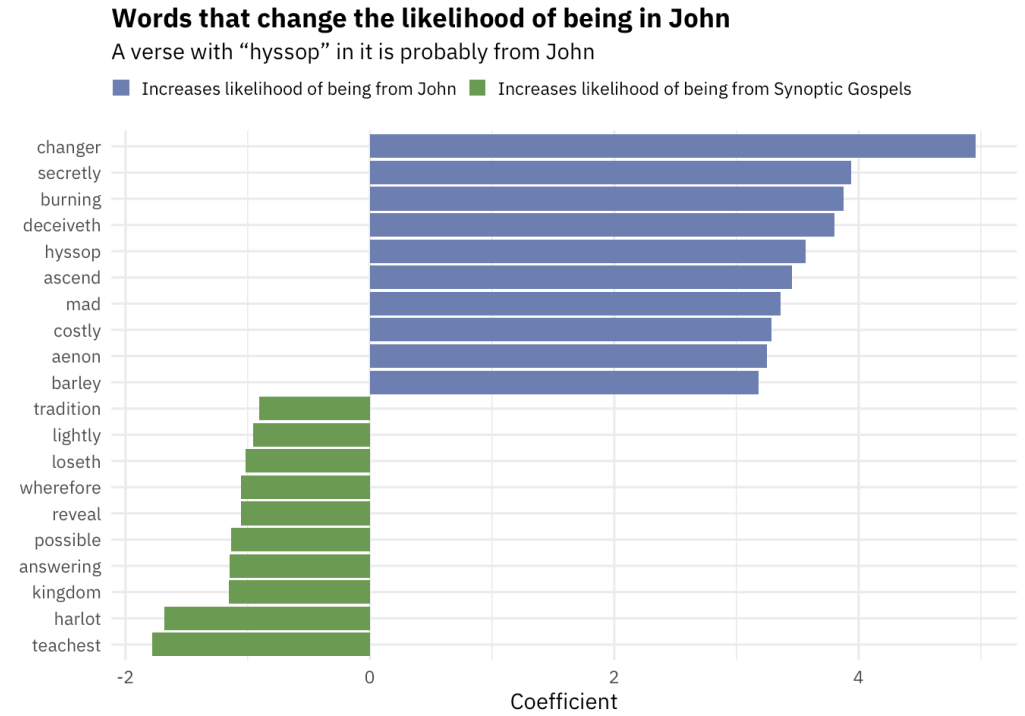
Wednesday, December 26, 2018

Tidy text, parts of speech, and unique words in the Bible

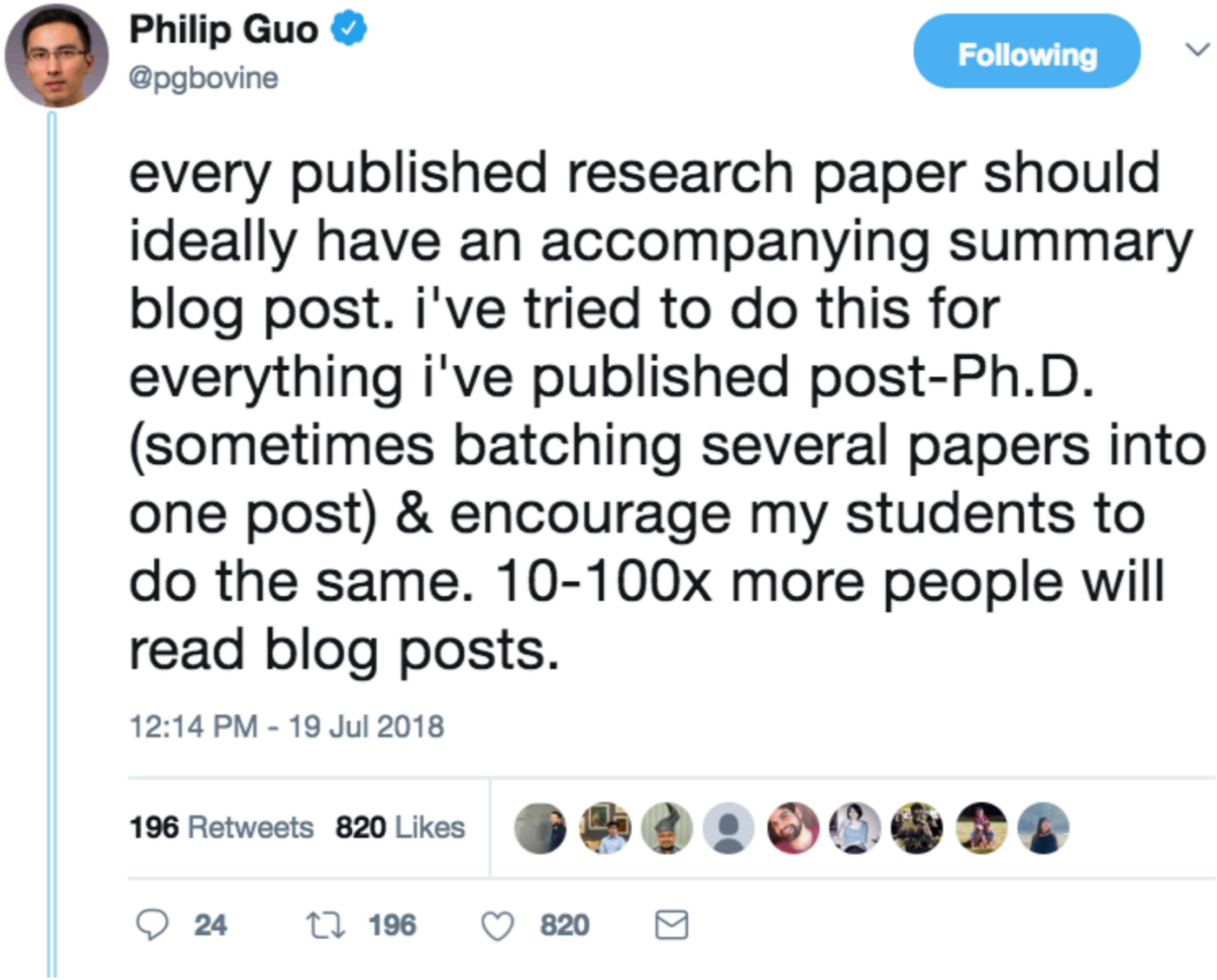
(See this notebook on GitHub)

As part of my goal to read some sort of religiously themed book every day (what I've read so far), I've been reading Eric Huntsman's new *Becoming the Beloved Disciple*, a close reading of the Gospel of John from an LDS perspective.

Near the beginning, Huntsman discusses several word frequencies that make John unique compared to the synoptic gospels of Matthew, Mark, and Luke (which all draw on the same Q source). For instance, Huntsman states that John focuses more on themes of discipleship



Blog about research



Teach a concept

Tuesday, January 29, 2019

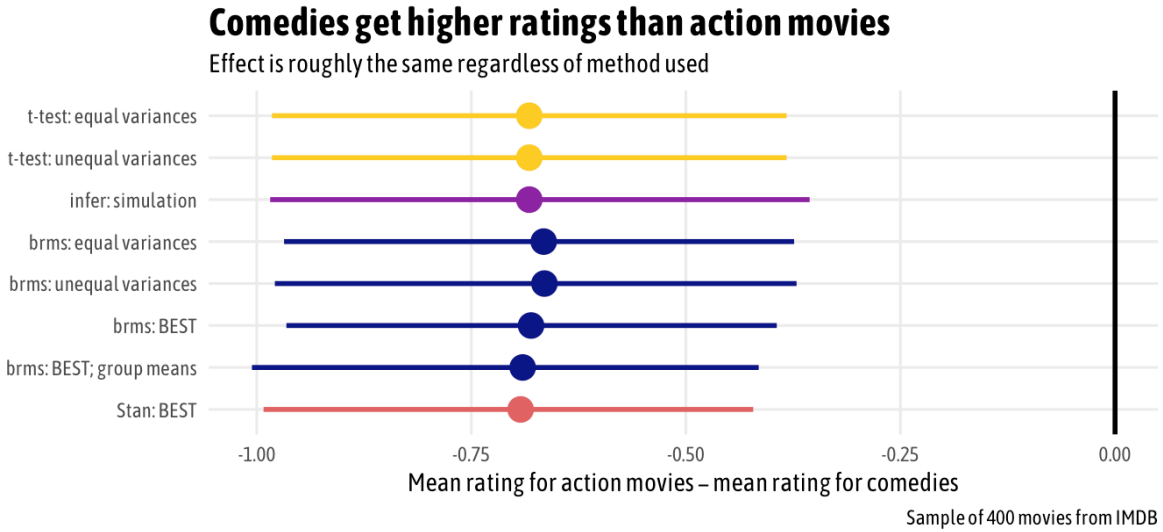
Half a dozen frequentist and Bayesian ways to measure the difference in means in two groups

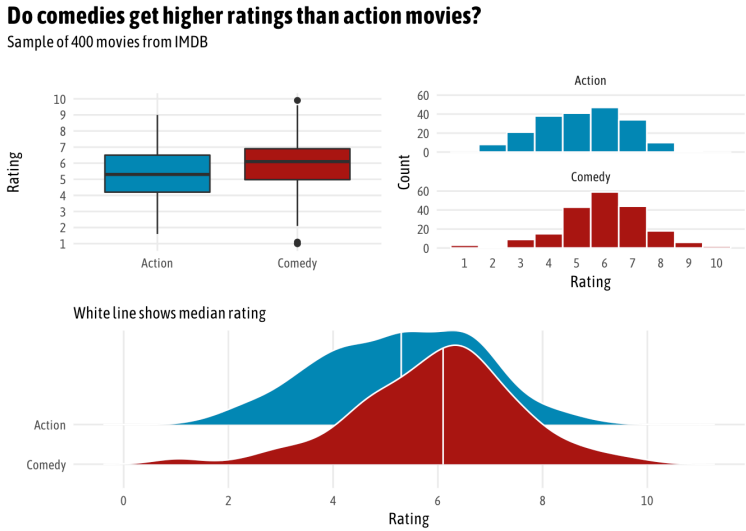
(See this notebook on GitHub)

Taking a sample from two groups from a population and seeing if there's a significant or substantial difference between them is a standard task in statistics. Measuring performance on a test before and after some sort of intervention, measuring average GDP in two different continents, measuring average height in two groups of flowers, etc.—we like to know if any group differences we see are attributable to chance / measurement error, or if they're real.

Classical frequentist statistics typically measures the difference between groups with a t-test, but t-tests are 100+ years old and statistical methods have advanced a lot since 1908. Nowadays, we can use simulation and/or Bayesian methods to get richer information about the differences between two groups without worrying so much about the assumptions and preconditions for classical t-tests.

Mostly as a resource to future me, here are a bunch of different ways to measure the difference in means in two groups. I've done them all in real life projects, but I'm tired of constantly searching my computer for the code to do them:)





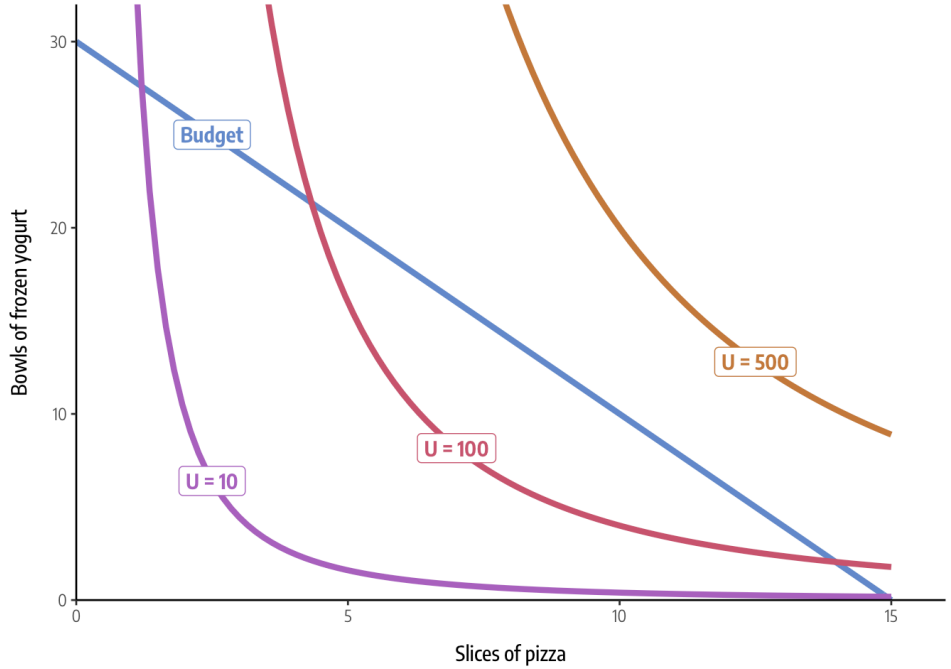
Saturday, February 16, 2019

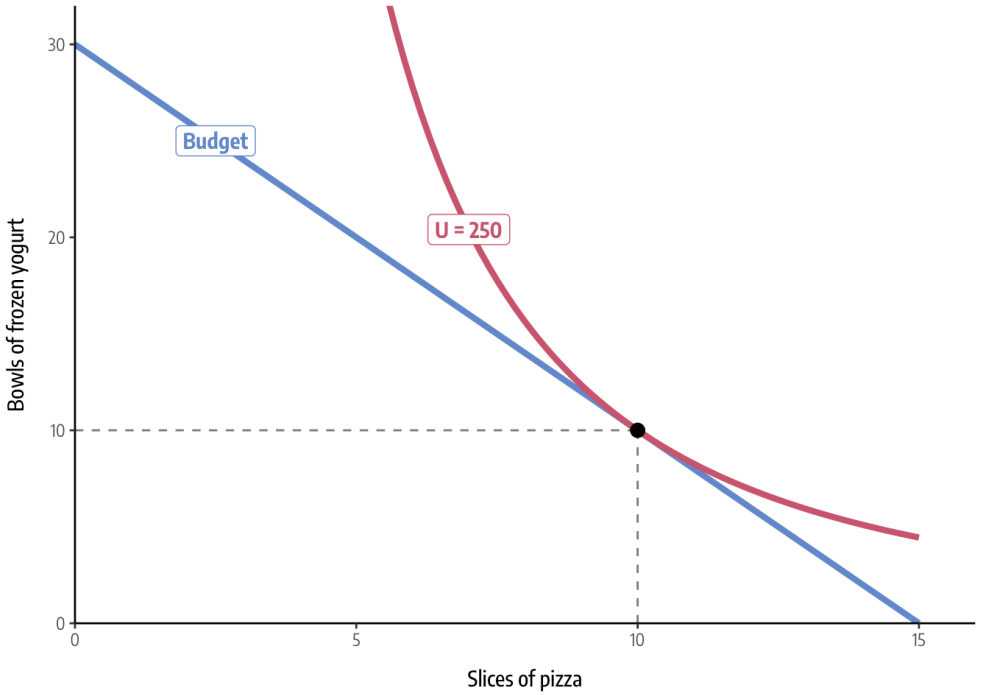
Chidi's budget and utility: doing algebra and calculus with R and yacas

(See this notebook on GitHub)

A year ago, I wrote about how to use R to solve a typical microeconomics problem: finding the optimal price and quantity of some product given its demand and cost. Doing this involves setting the first derivatives of two functions equal to each other and using algebra to find where they cross. I showed how to use neat functions like `Deriv::Deriv()` and `splinefun()` and make fancy plots showing supply and demand and it's pretty cool. I wrote it mostly because I was teaching an introductory microeconomics course and wanted an easy, generalizable, and manual math-less way to make these plots for my students' exercises and problem sets, and it works great.

I'm teaching microeconomics again this year and decided to tackle a trickier problem that involves curvier curves, more variables, and more math. And the results are even cooler and open the door for more doing math and symbolic algebra directly with R.



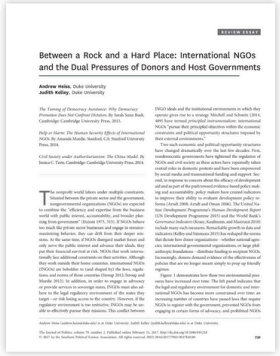


Make research public

Andrew Heiss, "Taking Control of Regulations: How International Advocacy NGOs Shape the Regulatory Environments of their Target Countries," *Interest Groups and Advocacy* (forthcoming), doi: 10.1057/s41309-019-00061-0

Presented at the workshop on Interest Groups, International Organizations, and Global Problem-Solving Capacity, Stockholm University, Sweden, June 2018, organized by Elizabeth Bloodgood and Lisa Dellmuth

Full details | PDF | Code

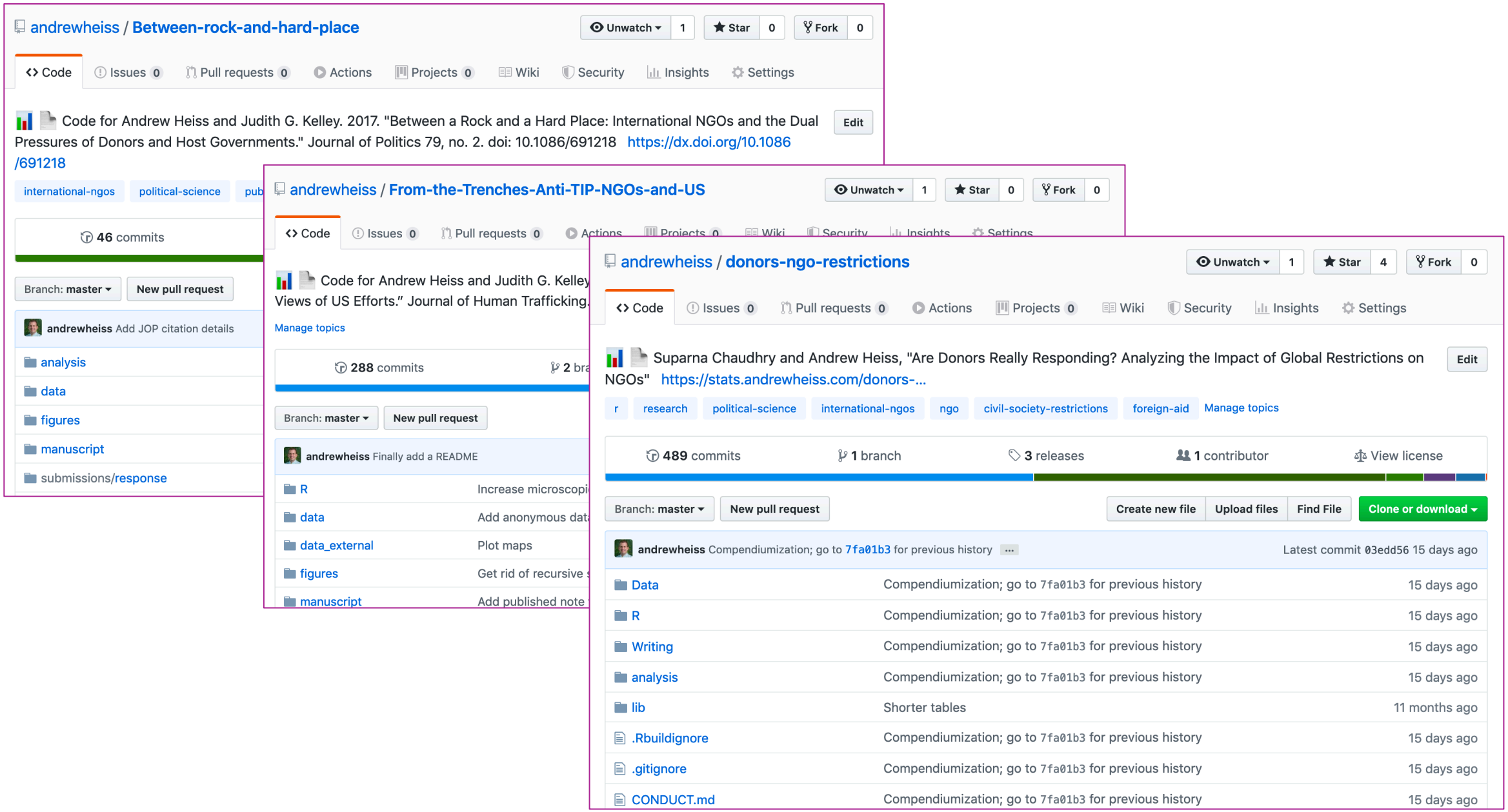


Andrew Heiss and Judith G. Kelley, "Between a Rock and a Hard Place: International NGOs and the Dual Pressures of Donors and Host Governments," *Journal of Politics* 79, no. 2 (April 2017): 732–41, doi: 10.1086/691218.

Full details | Ungated | Gated | Code

Andrew Heiss and Judith G. Kelley, "From the Trenches: A Global Survey of Anti-TIP NGOs and their Views of US Efforts," *Journal of Human Trafficking* 3 (2017), doi: 10.1080/23322705.2016.1199241.

Full details | Ungated | Gated | Code


andrewheiss / Between-rock-and-hard-place
Unwatch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
Settings
Code for Andrew Heiss and Judith G. Kelley. 2017. "Between a Rock and a Hard Place: International NGOs and the Dual Pressures of Donors and Host Governments." Journal of Politics 79, no. 2. doi: 10.1086/691218 https://dx.doi.org/10.1086/691218
Edit
international-ngos
political-science
46 commits
Branch: master
New pull request
andrewheiss Add JOP citation details
analysis
data
figures
manuscript
submissions/response
andrewheiss / From-the-Trenches-Anti-TIP-NGOs-and-US
Unwatch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
Settings
Code for Andrew Heiss and Judith G. Kelley
Views of US Efforts." Journal of Human Trafficking.
Manage topics
288 commits
2 bra
Branch: master
New pull request
andrewheiss Finally add a README
R
Increase microscopi
data
Add anonymous dat
data_external
Plot maps
figures
Get rid of recursive
manuscript
Add published note
andrewheiss / donors-ngo-restrictions
Unwatch 1
Star 4
Fork 0
Code
Issues 0
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
Settings
Suparna Chaudhry and Andrew Heiss, "Are Donors Really Responding? Analyzing the Impact of Global Restrictions on NGOs" https://stats.andrewheiss.com/donors-...
Edit
r
research
political-science
international-ngos
ngo
civil-society-restrictions
foreign-aid
Manage topics
489 commits
1 branch
3 releases
1 contributor
View license
Branch: master
New pull request
Create new file
Upload files
Find File
Clone or download
andrewheiss Compendiumization; go to 7fa01b3 for previous history
Latest commit 03edd56 15 days ago
Data
Compendiumization; go to 7fa01b3 for previous history
15 days ago
R
Compendiumization; go to 7fa01b3 for previous history
15 days ago
Writing
Compendiumization; go to 7fa01b3 for previous history
15 days ago
analysis
Compendiumization; go to 7fa01b3 for previous history
15 days ago
lib
Shorter tables
11 months ago
.Rbuildignore
Compendiumization; go to 7fa01b3 for previous history
15 days ago
.gitignore
Compendiumization; go to 7fa01b3 for previous history
15 days ago
CONDUCT.md
Compendiumization; go to 7fa01b3 for previous history
15 days ago