



# Data Science Fundamentals

## December 21, 2020

Eugene Guidice, Moderator  
Sarah Lin, Presenter



# Data Science Fundamentals



Sarah Lin

Information Architect & Digital Librarian  
RStudio, PBC

--

[sarah@rstudio.com](mailto:sarah@rstudio.com)  
<http://sarah.rbind.io/>  
[@sarahemlin](https://twitter.com/sarahemlin)  
she/her/hers



# Agenda

What is data science?

How do you *do* data science?

How does data science relate to *my library*?

How do I get started?

demonstration

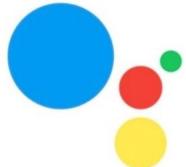
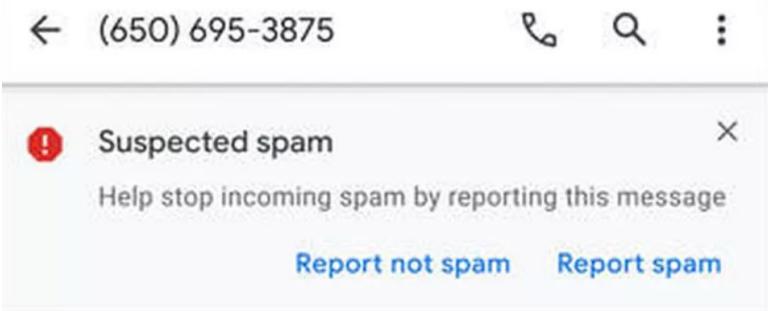
	A	B	C	D	E	F	G	H	I
1	Dragon Name	Dragon Type	Homeland	Colors they can be	Abilities they have	Weaknesses they have	Hardship(s)	Book introduced in	Dies in the Series?
2	Clay	Mudwing	Mud Kingdom	Brown scales with amber and gold underscales	Thick, armored scales, fire (mudwings born in dragonblood colored eggs are fireproof)	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Captured and imprisoned in (Queen) Scarlet's palace, Seawing palace, bitten by Dragonbite viper (survived, thanks to Peril)	Book 1, The Dragonet Prophecy	No
3	Tsunami	Seawing	Kingdom of the Sea	Blue or green or aquamarine scales	Can breath underwater, see in the dark, creat huge waves with one splash of tails and excellent swimmers	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom, other dragons and lack of water	Captured and imprisoned in (Queen) Scarlet's palace and the Rainforest Kingdom	Book 1, The Dragonet Prophecy	No
4	Sunny	Sandwing	Kingdom of Sand	Pale gold or white scales the color of desert sand	Can survive a long time without water, no scorpion like tail, good camouflage in desert and breath fire	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Captured and imprisoned in (Queen) Scarlet's palace, Seawing palace and Burn's Stronghold	Book 1, The Dragonet Prophecy	No
5	Starflight	Nightwing	Old, (as in two thousand years old) Kingdom of Night	Purplish-black scales with scattered silver scales on underside of wings	Breath fire and disappear into dark shadows (can read minds and tell future if born under two full moons)	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Captured and imprisoned in (Queen) Scarlet's palace, Seawing palace	Book 1, The Dragonet Prophecy	No
6	Glory	Rainwing	Rainforest Kingdom	Scales constantly change colors	Scales can change colors to blend into surroundings, prehensile tails for climbing and can shoot deadly venom from fangs	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Captured and imprisoned in (Queen) Scarlet's palace, Seawing palace	Book 1, The Dragonet Prophecy	No
7	Scarlet	Skywing	Sky kingdom	Red-gold or orange scales	Powerful fighters, fliers and can breath fire	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Lost the throne hit in the face with rainwing venom	Book 1, The Dragonet Prophecy	Yes, killed by Ruby
8	Kestral	Skywing	Sky kingdom	Red-gold or orange scales	Powerful fighters, fliers and can breath fire	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Captured by Scarlet	Book 1, The Dragonet Prophecy	Yes, killed by Blister
9	Dune	Sandwing	Kingdom of Sand	Pale gold or white scales the color of desert sand	Can survive a long time without water, no scorpion like tail, good camouflage in desert and breath fire	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Lost a leg	Book 1, The Dragonet Prophecy	Yes, killed by Scarlet
								, The	No



ypDisplay™ Mobile



PREDICTIVE POLICING



“Hey Alexa”

“Hey Siri”

“Hey Google”





# Agenda

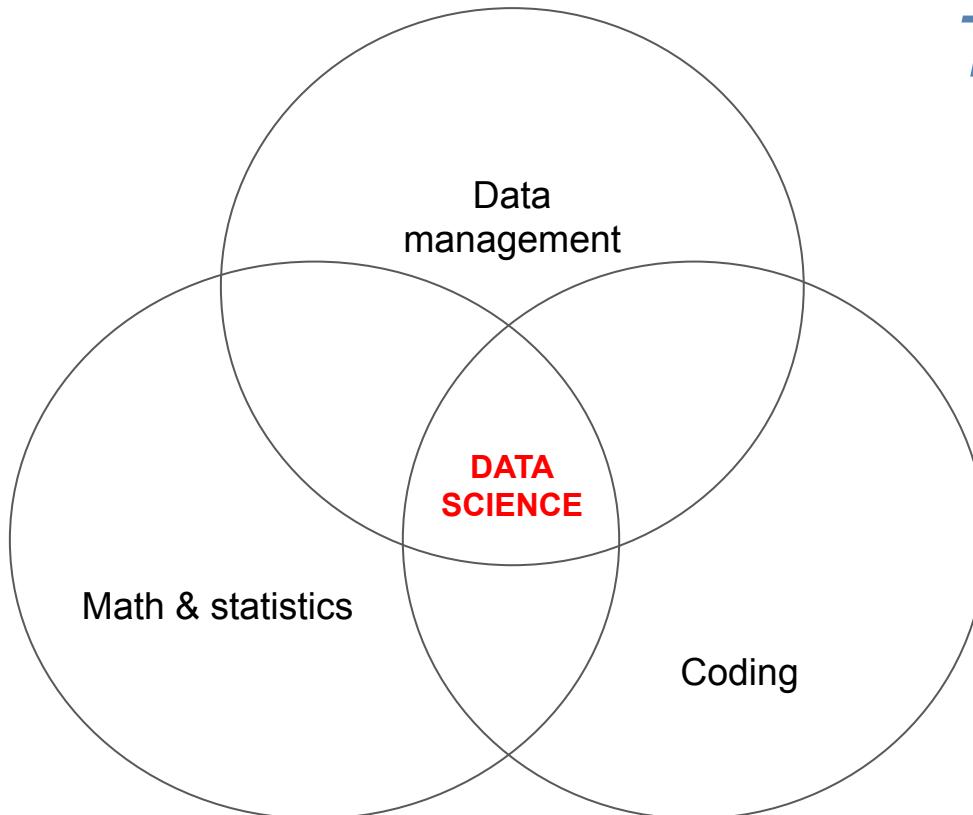
## ⇒ What is data science?

How do you *do* data science?

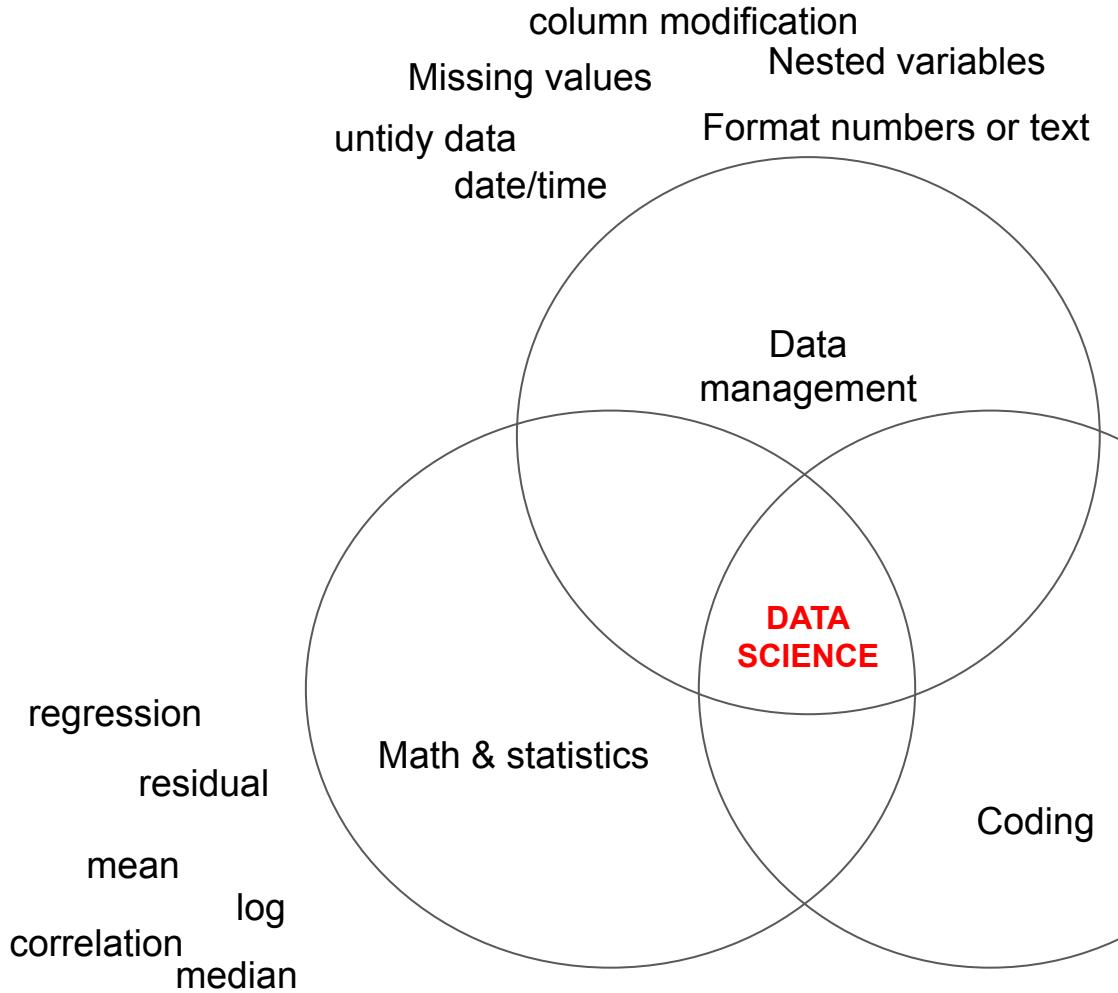
How does data science relate to *my library*?

How do I get started?

*The science of  
learning from  
data*



# *The science of learning from data*



# R is an object-oriented programming language

```
die <- c(1, 2, 3, 4, 5, 6)
die
## 1 2 3 4 5 6

die - 1
## 0 1 2 3 4 5

die / 2
## 0.5 1.0 1.5 2.0 2.5 3.0

die * die
## 1 4 9 16 25 36

m <- matrix(die, nrow = 2)
m
##      [,1] [,2] [,3]
## [1,]     1     3     5
## [2,]     2     4     6
```

```
dice <- sample(die, size = 2, replace =
TRUE)
dice
## 2 4

sum(dice)
## 6
```



The screenshot shows the RStudio IDE interface with the following components:

- Top Bar:** Contains icons for file operations (New, Open, Save, Print, Find, Copy, Paste), Go to file/function, Addins, and the R4DS logo.
- Code Editor:** Displays an R script titled "Notebook.R" with code related to reading CSV files. The code includes comments about `readr` guessing column types and handling a challenge dataset.
- Console:** Shows the R session output. It starts with workspace loading information, then attaches the tidyverse package, lists its dependencies, and shows conflicts between dplyr and stats packages.
- Environment:** Shows the global environment with various data objects listed:

Data	Description
batters	19428 obs. of 3 variables
batting	105861 obs. of 22 variables
by_day	336776 obs. of 19 variables
by_dest	336776 obs. of 19 variables
dec25	719 obs. of 19 variables
delay	96 obs. of 4 variables
delays	4037 obs. of 3 variables
flights_in_air	336776 obs. of 3 variables
flights_sml	336776 obs. of 7 variables
jan1	842 obs. of 19 variables
mid_min	336776 obs. of 20 variables
most_delayed	336776 obs. of 2 variables
not_cancelled	327346 obs. of 19 variables
selection	336776 obs. of 4 variables

- File Browser:** Shows the file structure under the "R4DS" project root, including .RData, .Rhistory, challenge.rds, diamonds.csv, diamonds.pdf, diamonds.R, Notebook.R, and R4DS.Rproj.



DOWNLOAD SUPPORT DOCS COMMUNITY



Products ▾

Solutions ▾

Customers

Resources ▾

About ▾

Pricing

# R & Python

The best data science is done in code

LEARN MORE



# Agenda

What is data science?

## ⇒ **How do you do data science?**

How does data science relate to *my library*?

What are the benefits of learning data science for law librarians?

How do I get started?

Data collection + tidying



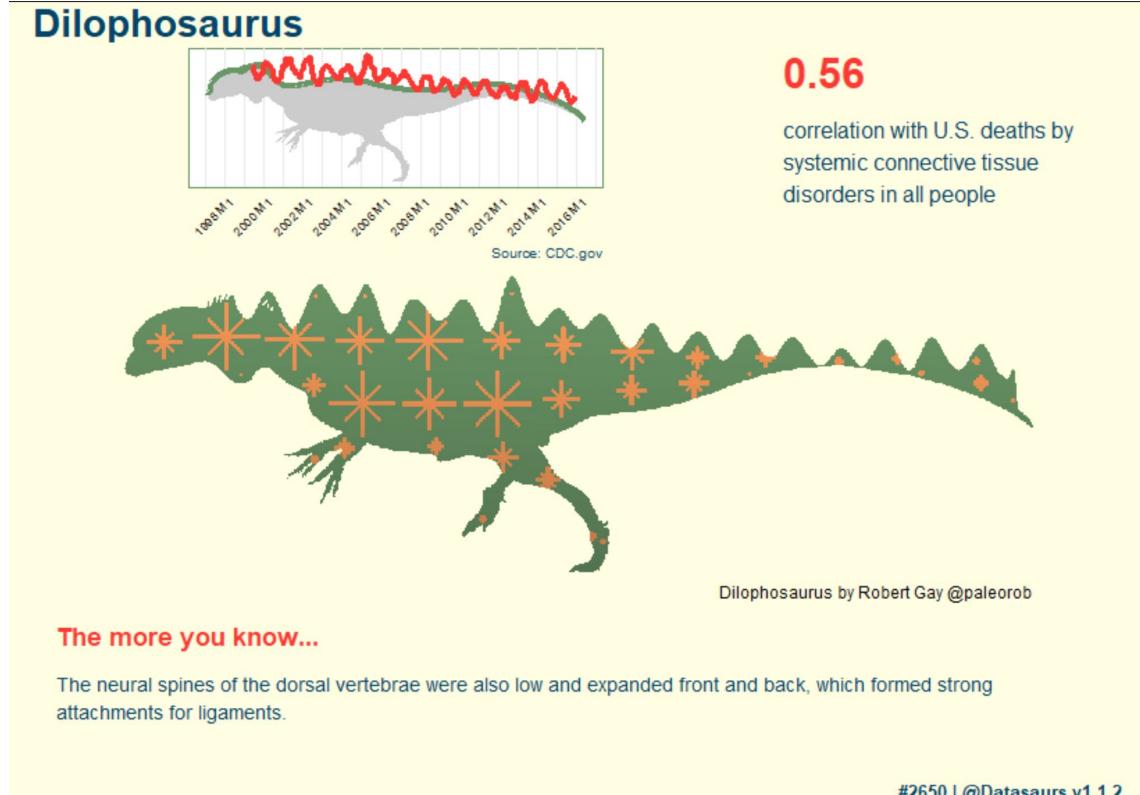
Visualize



Statistical analysis

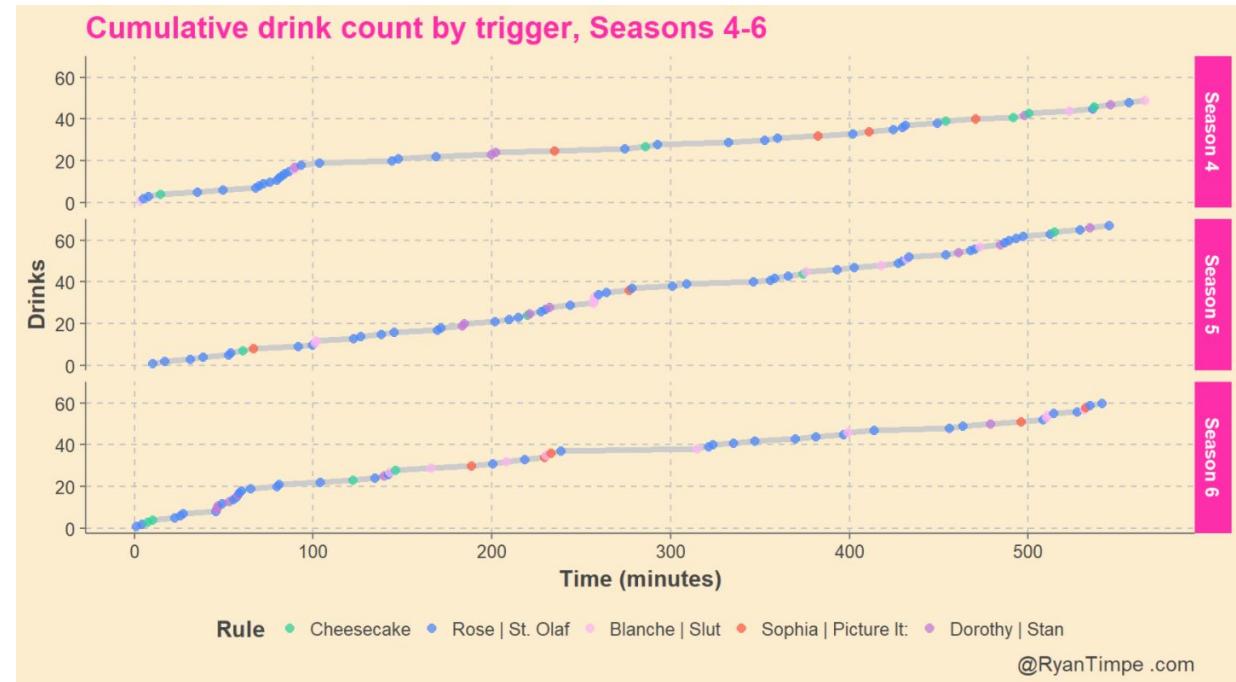


Making meaning



# Data science + text

Data collection + tidying



*text processing*

Visualize

Statistical analysis

Making meaning

# Choices in Text Processing

Unnesting

Word embeddings

Tokenizing

**Sentiment lexicons**

Stopwords

Training & testing datasets

N-grams

Classifiers (models)

Tf-idf

# Doing data science in R using the Tidyverse

## California County Hospitalization Forecasts

Select a county to see how modeled number of hospitalizations compare with actual numbers to date and with the number of licensed hospital beds (black box).

Yolo

Current Daily Hospitalizations:

**27 | 122**

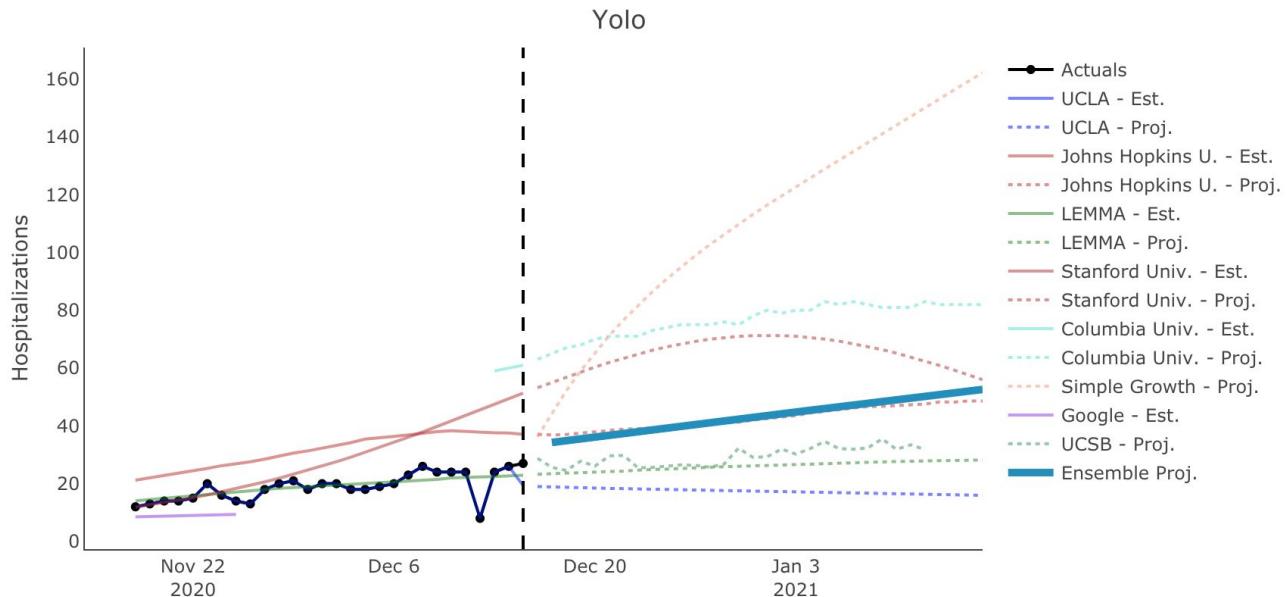
Actuals | Total Beds : 2020-12-15

Projected Daily Total:

**53**

Ensemble Forecast through 2021-01-16

 Download County Hospital Forecasts



 vargovargo update OS gituhbLatest commit ee63121 on Jul 8 

2 contributors



138 lines (138 sloc) | 56.5 KB

[Raw](#)[Blame](#) Search this file...

1	8/8/20	8/9/20	8/10/20	8/11/20	8/12/20	8/13/20	8/14/20	8/15/20	8/16/20
2	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
3	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
4	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
5	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
6	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
7	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
8	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
9	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229
10	0.97060434244736	0.968916802706205	0.967362543186093	0.96591722872997	0.964579006776923	0.963353126571865	0.96223959675998	0.961218728946202	0.960273809162229

# Tidied dataset from UCLA

```

grab_ucla_county <- function(State = state_name){

  url <- paste0("https://gist.githubusercontent.com/ZeroWeigh

  if (as.character(url_file_exists(url)[1]) == "TRUE" ) {

    ucla <- jsonlite::fromJSON(url)
    ucla_raw <- tibble::enframe(unlist(ucla))
    n_cols_max <- ucla_raw %>% pull(name) %>% str_split("\\."
    nms_sep <- paste0("name", 1:n_cols_max)
    ucla_cnty <- ucla_raw %>% separate(name, into = nms_sep,
    names(ucla_cnty) <- c("output","type","county","date", "v
    ucla_cnty$date <- as.Date(ucla_cnty$date, format ="%m/%d/
    msg <- paste0("Successfully downloaded Rt data from UCLA

  } else {
    msg <- paste0("Problem with UCLA link to file updates. Da
  }

  print(msg)

  return(ucla_cnty)
}

#### County Hospitalization Projections ####

#Data Prep
county.hosp <- reactive({
  progress <- Progress$new()
  # Make sure it closes when we exit this reactive, even if there's an error
  on.exit(progress$close())
  progress$set(message = "Gathering Hospitalization Forecasts", value = 0)

  cnty <- inputs$select.county.hosp
  progress$inc(3/4)
  # out <- lapply(cnty[1], function(x) get_can_cnty(x))
  # cnty.hosp <- do.call("rbind",out)
  out <- filter(can.county.observed, fips == cnty)

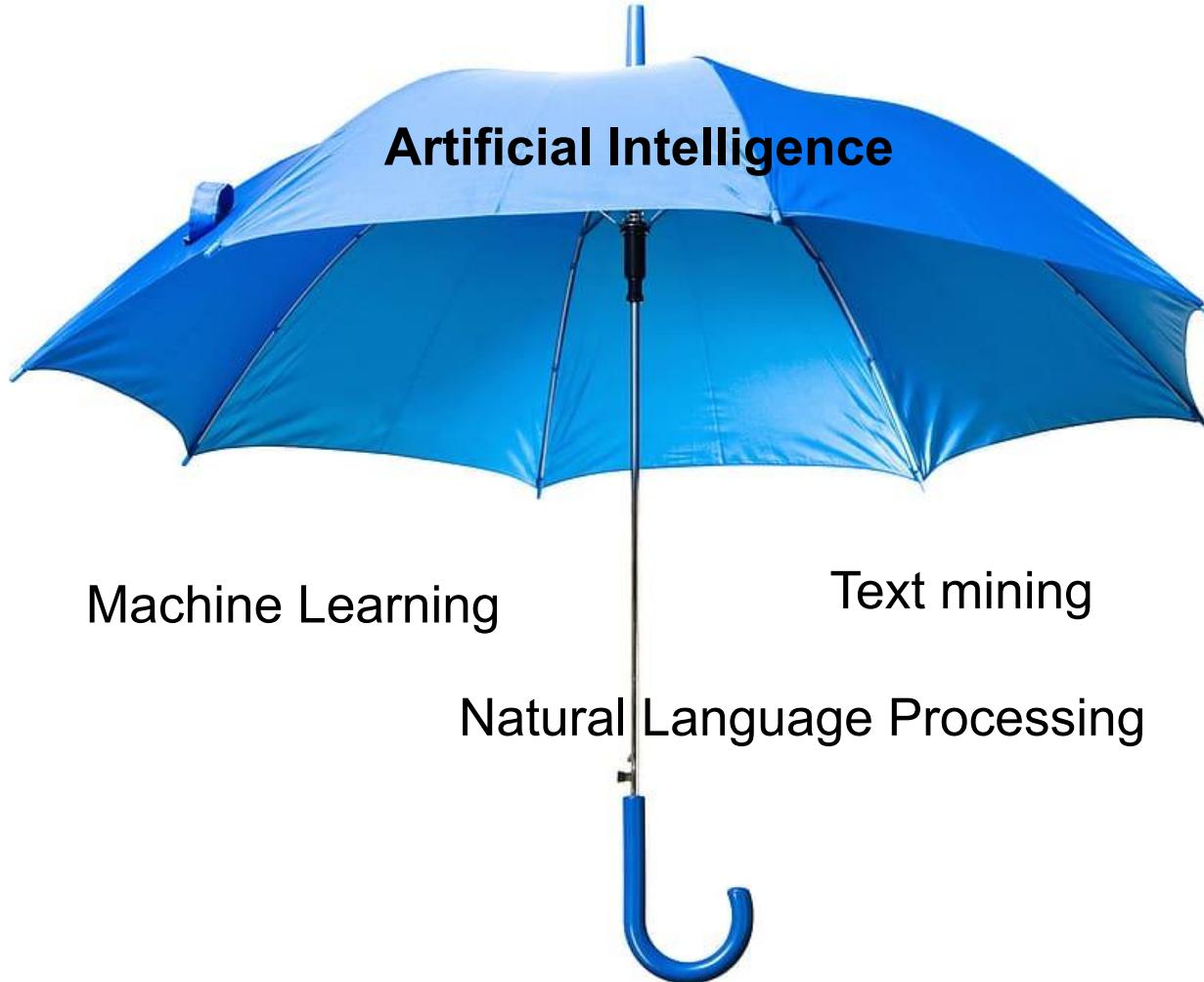
  cnty.hosp <- out %>% select(date,hospitalBedsRequired) %>% as.data.frame()
  progress$inc(1/4)
  return(cnty.hosp)
})

```

Build an interactive app that lets you model data by county against models from other entities like UCLA, John Hopkins, etc.

# DATA





# Machine Learning Process

*Before you start... tidy data*

1. Choose a <statistical> model
2. Process your data (extract/convert/transform variables, i.e. stemming)
3. Sample : choose testing & training datasets  
→ *measure the fit of the model* ←
4. Tune parameters : tweak models/metrics

*Bundle the steps into a workflow and...*

PREDICT!



# Agenda

What is data science?

How do you *do* data science?

**⇒How does data science relate to *my library*?**

How do I get started?

# How Jared Kushner built a luxury skyscraper using loans meant for job-starved areas

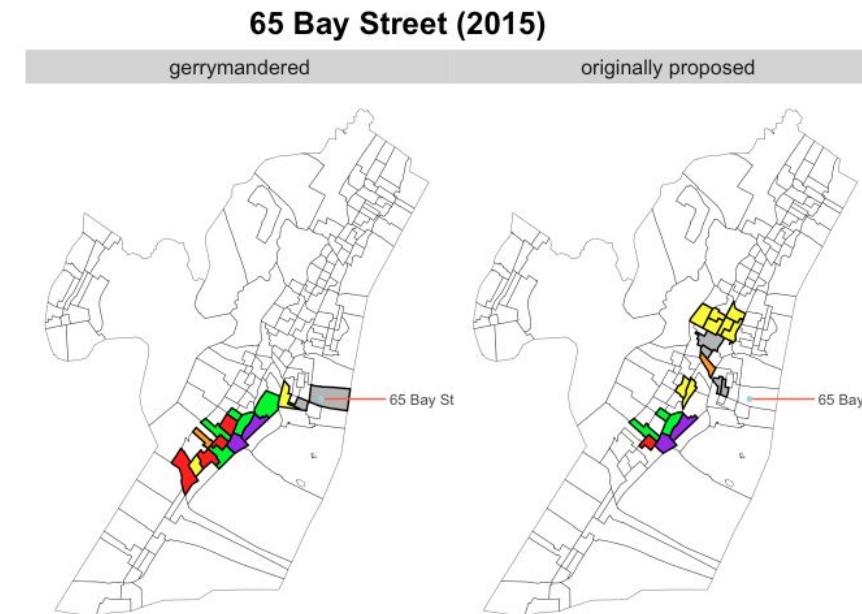
wpinvestigative / kushner\_eb5\_census

Code Issues Pull requests Actions

master kushner\_eb5\_census / data /

andrewbtran so long .ds\_store

..  
callcode.csv  
codes2015.csv  
jcpd-calls-for-service-2014-csv.csv  
jcpd-calls-for-service-2015-csv.csv  
jcpd-calls-for-service-2016.csv  
jcpd-calls-for-service-2017-dashboard.csv  
jerseycity2015shootings-use.csv



```
library(tidy census)
census_api_key(key)

# Get 5-year ACS data based on most recently available data
hudson_unemployment_2015 <- get_acs(geography="tract", endyear=2015, variables= c("B23025_005E",
"B23025_002E"), county = "Hudson", state="NJ")

# Setting up the dataframes (tidying)
hudson_unemployment_2015 <- spread(hudson_unemployment_2015, variable, estimate )

# Figuring out the unemployment rate from the estimates
hudson_unemployment_2015 $per_un <-
round(hudson_unemployment_2015[, 4]/hudson_unemployment_2015[, 3]*100,2)

# Getting rid of the blank census tracts
hudson_unemployment_2015 <- filter(hudson_unemployment_2015, GEOID != "34017006900" &
GEOID != "34017980100" )

# Creating a column based on the rank of the unemployment in the county
hudson_unemployment_2015 $rank <- rank(hudson_unemployment_2015 $per_un)

# Creating an array of census tracts as specified by the developers
proj1 <- c("34017001900", "34017004600", "34017005300", "34017006600", "34017006700", "34017007100")
```

```
hudson_unemployment_2012 <- get_acs(geography="tract", endyear=2015, variables= c("B23025_005E", "B23025_002E"), county  
= "Hudson", state="NJ")  
  
hudson_unemployment_2012 <- spread(hudson_unemployment_2012, variable, estimate )  
hudson_unemployment_2012 $per_un <- round(hudson_unemployment_2012[, 4]/hudson_unemployment_2012[, 3]*100,2)  
hudson_unemployment_2012 <- filter(hudson_unemployment_2012, GEOID != "34017006900" & GEOID != "34017980100")  
hudson_unemployment_2012_sm <- filter(hudson_unemployment_2012, GEOID %in% proj1)  
colnames(hudson_unemployment_2012_sm) <- c("GEOID", "name", "total", "unemployed", "unemp_rate")  
hudson_unemployment_2012_sm $un_rate <- hudson_unemployment_2012_sm $unemp_rate$B23025_005  
hudson_unemployment_2012_sm $radius <- "gerrymandered"  
proj1_half <- c("34017000300", "34017000400", "34017000500", "34017000600", "34017000700", "34017000800", "34017001100",  
"34017001201", "34017001202", "34017001500", "34017002200", "34017002300", "34017002500", "34017002600", "34017003000",  
"34017003100", "34017003200", "34017003300", "34017004400", "34017004500", "34017004600", "34017005000", "34017005200",  
"34017005300", "34017005500")  
  
hudson_unemployment_2012_sm_half <- filter(hudson_unemployment_2012, GEOID %in% proj1_half)  
colnames(hudson_unemployment_2012_sm_half) <- c("GEOID", "name", "total", "unemployed", "unemp_rate")  
hudson_unemployment_2012_sm_half $un_rate <- hudson_unemployment_2012_sm_half $unemp_rate$B23025_005  
hudson_unemployment_2012_sm_half $radius <- "originally proposed"  
hudson_unemployment_sm <- rbind(hudson_unemployment_2012_sm, hudson_unemployment_2012_sm_half)  
  
nj_hf <- fortify(nj_h, region="GEOID")  
nj_hud <- left_join(nj_hf, hudson_unemployment_sm, by=c("id"="GEOID"))  
nj_hud <- filter(nj_hud, !is.na(radius))  
nj_hud$nj_mid <- cut(nj_hud$un_rate, 6)
```

```
nj_map <- ggplot()  
#nj_map <- nj_map + geom_polygon(data=nj_hf, aes(x=long, y=lat, group=group), fill=NA, color="black", size=.1)  
nj_map <- nj_map + geom_polygon(data=nj_hud, aes(x=long, y=lat, group=group, fill=factor(nj_mid)), color="black",  
size=.5)  
nj_map <- nj_map + facet_wrap(~radius)  
nj_map <- nj_map + coord_map()  
nj_map <- nj_map + scale_fill_manual(drop=FALSE, values=c("gray", "yellow", "orange", "red", "green", "purple"),  
na.value="#EEEEEE", name="Unemployment rate")  
  
nj_map <- nj_map + theme_nothing(legend=TRUE)  
nj_map <- nj_map + labs(x=NULL, y=NULL, title="65 Bay Street (2015)")  
nj_map <- nj_map + theme(panel.grid.major = element_line(colour = NA))  
nj_map <- nj_map + theme(text = element_text(size=15))  
nj_map <- nj_map + theme(plot.title=element_text(face="bold", hjust=.4))  
nj_map <- nj_map + theme(plot.subtitle=element_text(face="italic", size=9, margin=margin(l=20)))  
nj_map <- nj_map + theme(plot.caption=element_text(size=12, margin=margin(t=12), color="#7a7d7e", hjust=0))  
nj_map <- nj_map + theme(legend.key.size = unit(1, "cm"))  
nj_map <- nj_map + annotate("segment", x = -74.03577, xend = -74.005, y = 40.72, yend = 40.72, colour = "tomato",  
size=.5)  
nj_map <- nj_map + annotate("point", x = -74.03577, y = 40.72, colour = "lightblue", size = 1)  
nj_map <- nj_map + annotate("text", x = -73.98217, y = 40.72, label = "65 Bay Street", size=3, colour="gray30")
```

[This data is published under an [Attribution-NonCommercial-ShareAlike 4.0 International \(CC BY-NC-SA 4.0\) license](#)]

## About this story

---

[Jared Kushner and his partners used a program meant for job-starved areas to build a luxury skyscraper](#)

## About the folders in this repo

---

- [data](#) - Raw and summarized data on shootings and service calls from the Jersey City open data portal
- [map\\_output](#) - PDF output of maps made with ggplot2 used as basis for WP graphics
- [markdown](#) - R Markdown files used to generate exploratory data analysis
- [scripts](#) - Exploratory data analysis and visualizations scripts

## Notebooks

---

- [Profiles on Jersey city project tracts](#) - Aggregated information geolocated to specific census tracts to assist with reporting
- [Comparing economic conditions in Jersey City](#) - Exploratory data visualizations
- [Comparing economic conditions in Jersey City \(2\)](#) - More focused exploratory data visualizations
- [Comparing economic conditions in Jersey city \(3\)](#) - Comparing the developer's original flawed pitch to the state-adjusted gerrymandered version

# DIY Census Data Analysis

```
library(tidycensus)  
  
census_api_key("YourCensusKey")  
  
ca_income <- get_acs(state = "CA", geography = "county", variables =  
"B19013_001", geometry = TRUE)
```

# R Packages for Election Data

pollstR  
ggparliament  
ggplot  
leaflet  
tilegramR  
tidycensus  
censusapi  
FiveThirtyEight

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(viridis)
library(ggthemes)

# this next code will need to be adapted if you don't have a ../data/ folder...
# alternatively, if that link stops working, there's a static copy at http://ellisp.github
www <- "http://projects.fivethirtyeight.com/general-model/president_general_polls_2016.cs
download.file(www, destfile = "../data/polls.csv")

# download data
polls_orig <- read.csv("../data/polls.csv", stringsAsFactors = FALSE)

table(table(polls_orig$poll_id))
##      3
## 3067

table(polls_orig$type)
## now-cast polls-only polls-plus
## 3067      3067      3067
```

Measure

Arrest risk ratio

Year

2015

Age

Adult

Race

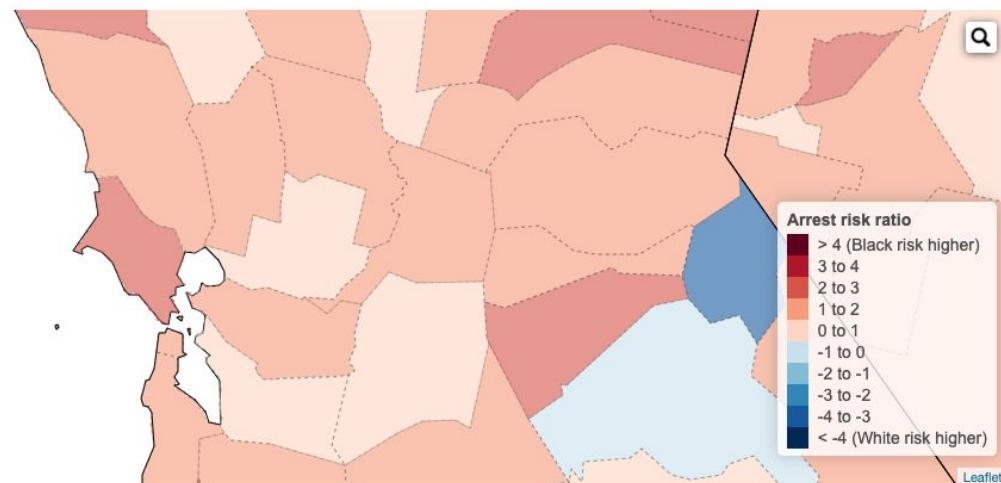
Black

County

California, Yolo

[Reset zoom](#)

## Adult Black / White Arrest Risk Ratio



Measure

Arrest risk ratio

Year

2015

Age

Juvenile

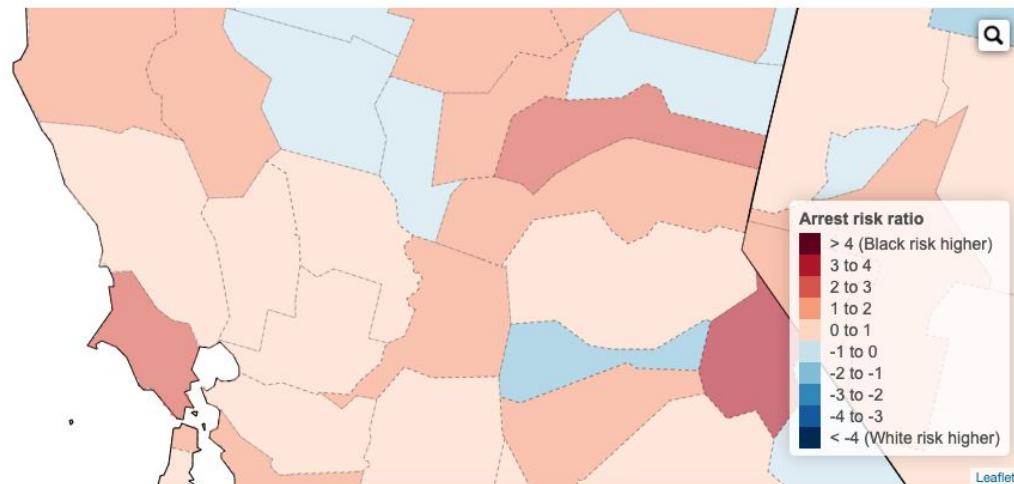
Race

Black

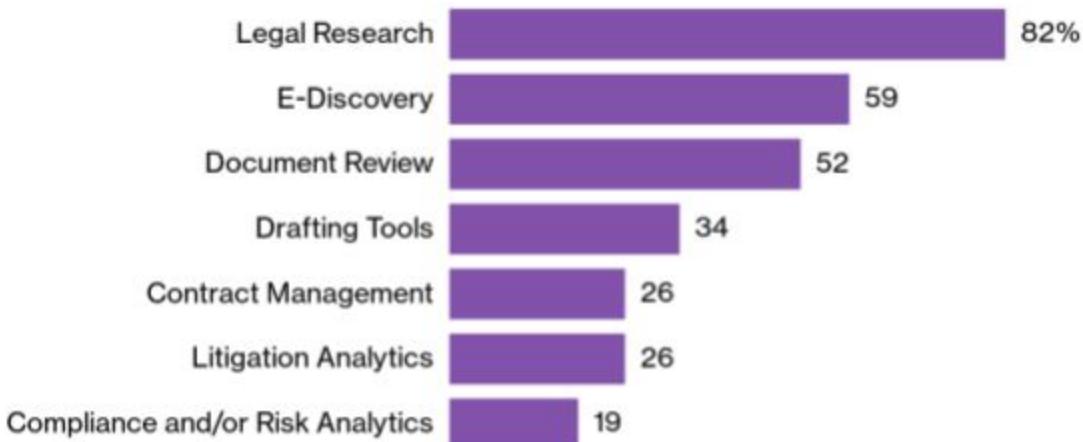
County

California, Yolo

## Juvenile Black / White Arrest Risk Ratio

[Reset zoom](#)

## Some Types of AI-Driven Tech Are Used Heavily

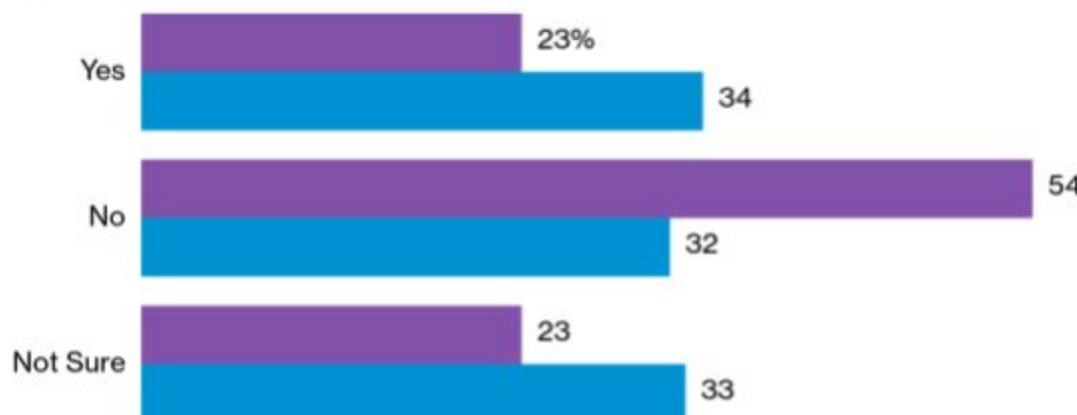


Source: Bloomberg Law Legal Technology Survey (2020), n=331 in-house and law firm respondents

Respondents were asked which types of legal tech their organization is currently using and were asked to select all that applied. List filtered for types of legal tech that commonly employ AI or machine learning algorithms.

## Is Your Organization Using AI or Machine Learning Tools?

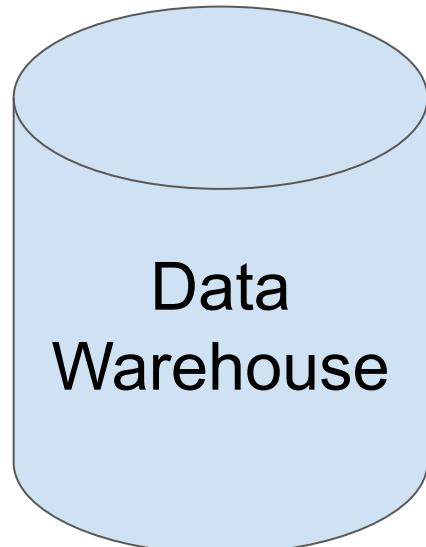
■ 2019 ■ 2020



Sources: Bloomberg Law Legal Technology Survey (2020); Bloomberg Legal Operations & Technology Survey (2019)  
2020 Survey, n=331 in-house and law firm respondents; 2019 Survey, n=489 in-house and law firm respondents

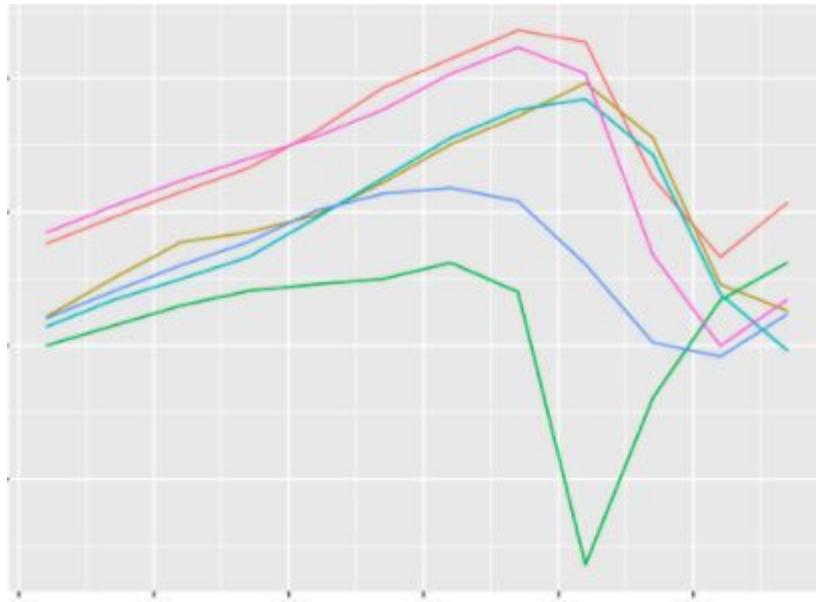
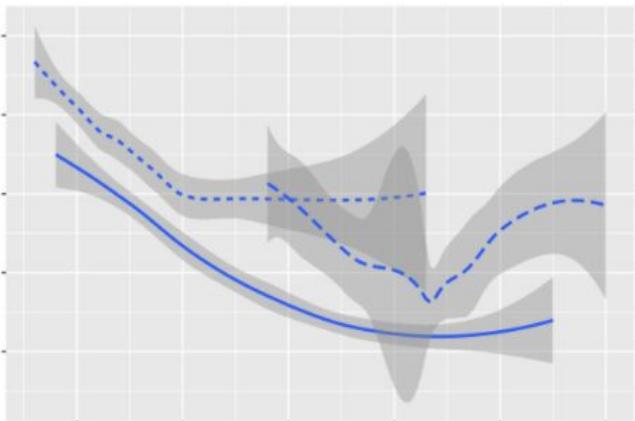
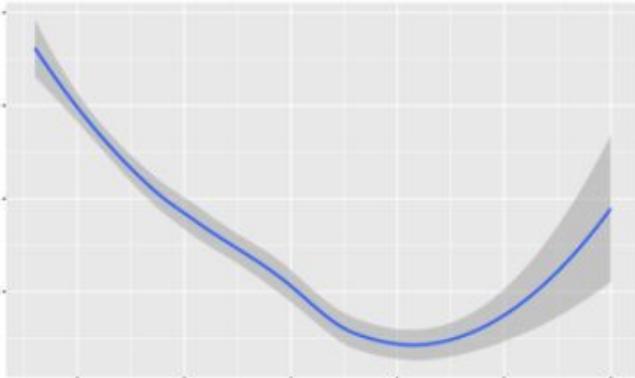
Bloomberg Law

Data stored,  
often in the cloud



Data  
collected  
from  
software  
applications

Data  
extracted  
for analysis



This document computes total viewership of the RViews posts (i.e., those articles published under [rviews.rstudio.com](https://rviews.rstudio.com)). The tables generated are `reactables` and allow you to search, sort, and in some cases drill down into them as you like.

Three tabs provide different views of post popularity:

**1. Total Views.** This tab counts [rviews.rstudio.com](https://rviews.rstudio.com) pageviews and users from the day it was posted until now.

**2. Windowed Views.** This tab counts [rviews.rstudio.com](https://rviews.rstudio.com) pageviews and users only within a 15 day window of when the post was published. This metric helps us compare the popularity of posts, regardless of how long they've been on the site.

Total Views    Windowed Views    Sources    Medium

Statistics in this table are aggregated from 2020-04-10 through 2020-12-16.

Search

Title	Post Date	Users	Page Views	Unique Page Views	Mean Time on Page	Mean Bounce Rate
<a href="#">Plotting Surfaces with R</a>	2020-12-14	981	1,148	1,055	30	42%
<a href="#">Learn and Teach R</a>	2020-12-02	2,556	3,248	2,713	52	26%
<a href="#">COVID-19 Mobility Data</a>	2020-11-25	598	721	632	33	61%
<a href="#">October 2020: "Top 40" New CRAN Packages</a>	2020-11-19	3,771	5,490	4,136	73	34%
<a href="#">ROC Day at BARUG</a>	2020-11-12	375	491	407	61	55%
<a href="#">Biologically Plausible Fake Survival Data</a>	2020-11-02	439	507	458	34	45%
<a href="#">Sept 2020: "Top 40" New CRAN Packages</a>	2020-10-23	5,893	8,284	6,522	57	35%

# Tettra usage 2019-2020

Sarah Lin

12/01/2020

## Examining Tettra Usage

Hiring a librarian in late September 2019 was geared towards improving findability of internal information at RStudio. A portion of internal information is stored in our Tettra wiki software. Over the last year, has engagement with Tettra increased? How do we measure engagement?

Text

### Content

Is Tettra just for benefits information?

#### Top 25 pages over the last year

Show 10 entries

Search:

	PageTitle	totalViews
1	Solutions Engineering	549
2	AWS Subaccounts: Accessing with assumeroles	367
3	Home Office Favorites	345
4	RStudio Hardware Standards	313
5	CS Guidelines - When to Go to Whom	283
6	Company Holidays	277
7	2020 Product Launches	242
8	2020 Beta/Experimental Programs - CS Involvement	224
9	Focus Time	218
10	Support Cheat Sheet	215

Showing 1 to 10 of 25 entries

Previous 1 2 3 Next

Data analysis

```

1  ---
2  title: "Tettra usage 2019-2020"
3  author: "Sarah Lin"
4  date: "9/2/2020"
5  output: html_document
6  editor_options:
7    chunk_output_type: inline
8  ---
9
10 ````{r setup, include=FALSE}
11 knitr::opts_chunk$set(echo = FALSE)
12 #install.packages("googlesheets4")
13 #load libraries
14 library(googlesheets4) #googlesheets4 automatically opens your browser and asks you to
15 authenticate
16 library(tidyverse)
17 library(lubridate)
18 library(scales)
19 library(DT)
20 #library(curl)
21 #handle_setopt(http_version = 2)
22 ``
23 ## Examining Tettra Usage
24
25 Hiring a librarian in late September 2019 was geared towards improving findability of
internal information at RStudio. A portion of internal information is stored in our
Tettra wiki software. Over the last year, has engagement with Tettra increased? How
do we measure engagement?
26
27 # Content
28 Is Tettra just for benefits information? Here are the top 25 pages over the last year.
29
30 ````{r include=FALSE}
31 page_names <- read_sheet("https://docs.google.com/spreadsheets/d/1GHDPP7k4miLWbSTGtHY0A
EAzPp0NaZ4hCyBI7LUpp_0/edit#gid=0")

```

**Access to raw data is limited**

32

```

32 pages <- page_names %>
33   select(c(Month, Views, PageTitle, Name))
34
35
36 pages1 <- pages %>%
37   group_by(PageTitle) %>%
38   summarise(sum(Views))
39
40 total_views <- pages1 %>%
41   arrange(desc(pages1$sum(Views)))
42
43 top25 <- total_views %>%
44   slice_head(n = 25)
45 ...
46 ````{r}
47 datatable(top25)

```

## Shareable code

PageTitle	sum(Views)
Solutions Engineering	439
AWS Subaccounts: Accessing with assumeroles	317
Home Office Favorites	306
RStudio Hardware Standards	273
Company Holidays	189
2020 Beta/Experimental Programs – CS Involvement	188
Do I contact DevOps or IT for my technical needs?	187
Travel to Boston	186
2020 Product Marketing Planning and Resources	179
What do I contact DevOps for?	173

1-10 of 25 rows

Previous 1 2 3 Next

48

49

50 Tettra is definitely used for HR-type information, but a significant number of pages related to information about working at RStudio as well as specific team-related information were viewed over the last year.

51

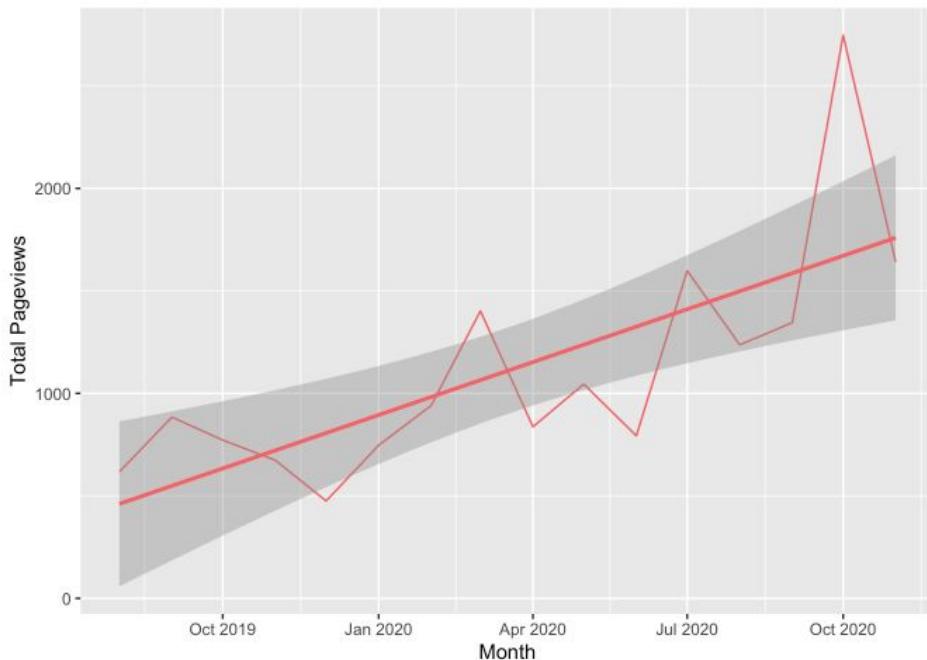
# Pageviews

Pageviews is the total count of pages viewed per month. If one person visits the same page 10 times in one month, that's counted as 10 pageviews, so it's a very high level, generic count of Tettra usage. In July 2020, Amanda migrated the Support Wiki from GitHub to Tettra, which explains the high number of pages created that month.

## Total pageviews over the past year

The trend is definitely strongly increasing.

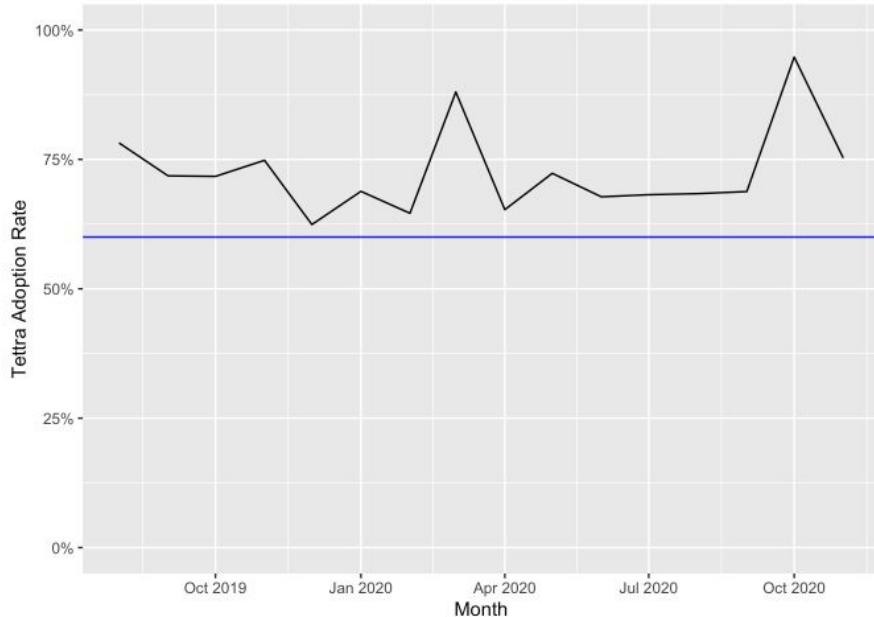
```
## `geom_smooth()` using formula 'y ~ x'
```



```
119 * ````{r include = FALSE}
120 #read in the PageViews-2019-2020 Google Sheet
121 views <- read_sheet("https://docs.google.com/spreadsheets/d/1zuZvXI965VQ1qzyJ0
122 _HwEp84ZNx1z9fDvK9Qh8u-S_s/edit#gid=0")
123 print(views)
124 ````
125
126 Total pageviews over the past year is definitely trending upwards.
127
128 * ````{r}
129 ggplot(data = views, aes(by_month, total_pages_viewed, color = "red")) +
130   geom_line() +
131   geom_smooth(method = "lm")
132 ````
```

# Engagement Summary

Tettra is clearing being used, and used a lot by some employees. Another measure of engagement is to look at the adoption rate: what percentage of employees are engaging with Tettra, both on the whole and month-by-month?



```
254 ## Engagement Summary
255
256 Tettra is clearing being used, and used a lot by some employees. Another
measure of engagement is to look at the adoption rate: what percentage of
employees are engaging with Tettra, both on the whole and month-by-month?
257
258 ``{r include=FALSE}
259 # views is the table of total views per month
260 employees <- read_sheet("https://docs.google.com/spreadsheets/d/1mRYd3bH75wfW8
3sR9f5_Nlg64JpQtNiVbF2TGsjgFRg/edit#gid=0")
261
262
263 ``{r}
264
265 # views$total_people is the count of employees who've viewed at least one page
266 # employees$employee_count is the count of people employed by month
267
268 # adoption rate = number of readers / number of employees OR
269 # adoption rate = views$total_people / employees$employee_count
270
271 utilization <- employees %>%
272   mutate(
273     viewers = views$total_people,
274     adoption = views$total_people / employees$employee_count)
275
276 ggplot() +
277   geom_line(mapping = aes(x = month, y = adoption), data = utilization) +
278   ylim(0,1) +
279   labs(x = "Month", y = "Tettra Adoption Rate")
280
281 ...
```

# *Bias limitations in ALL [legal research] databases*

Databases utilize:

- Text mining
- Natural language processing
- Machine learning
- Relevance ranking

Based upon:

- Raw data
- Data tidying
- Textual analysis
- modeling

## (C) PERSONS AND ENTITIES PROTECTED, k3920-k3934

Add to favorites

Copy link

Select all content • No items selected • Clear Selection Collapse All

Specify Content to Search

3920 In general

3921 Non-citizens; aliens

3922 Children and minors

3923 Unborn children; fetuses

3924 Prisoners, detainees, and convicts

3925 Armed services personnel

3926 Organizations and associations

3927 Business organizations; corporations

3928 Government entities

3929 —In general

3930 —Employees and officials

3931 —Foreign governments

3932 Other particular persons and entities

[Home](#) / Constitutional Law

## Constitutional Law



### Amendment Process

- + [Bill of Rights](#)
- + [Congressional Duties & Powers](#)
- + [Elections, Terms & Voting](#)
- + [Equal Protection](#)

### General Overview

[Income Tax](#)[Involuntary Servitude](#)[Privileges & Immunities](#)[Prohibition](#)[Qualifications for Federal Office](#)

- + [Relations Among Governments](#)

[Separation of Powers](#)[State Constitutional Operation](#)

- + [State Sovereign Immunity](#)
- + [Substantive Due Process](#)
- + [Supremacy Clause](#)
- + [The Judiciary](#)
- + [The Presidency](#)



**Quick Links:** Commodities Stocks Rates & Bonds Currencies Futures Economics Fixed Income ETFs Sectors Watchlist Wealth Election 2020 Covid-19

JPY X-RATE 103.34 +0.230 +0.22% EUR-USD 1.22 -0.003 -0.22% Gold Spot \$/Oz 1881.45 -3.97 -0.21% Bloomberg Barclays Global Aggregate Bond Index 557.57 +1.578 +0.28% WTI Fut

## LATEST

Politics

**Republican Threatens 'Brawl' Over Biden's EPA, Interior Picks**

25m ago

Prognosis

**December Is Set to Be Covid's Deadliest Month in the U.S. by Far**

27m ago

Politics

**Bidens to Get Vaccine Monday, Harris and Husband Following Week**

updated 16m ago

Get the latest from the topics you follow

**+ Add my topics**



Business

**N.Y. Daily Covid Cases Top 12,000, Most Since Pandemic Began**

updated an hour ago

More Than 11 Million People Have Been Vaccinated: Covid-19 Tracker

Some U.S. States Say Feds Cut Vaccine Allocations by 40%

Pence Gets Televised Vaccination to Show Confidence in Shot

Markets

**Stocks Drop From Records With Congress in Talks: Markets Wrap**

Get caught up with what's moving markets.

updated 17 minutes ago

Politics

**Senator Threatens Stopgap Over Stimulus Checks: Congress Update**

All the latest news and headlines from the last-minute talks to provide relief for millions of Americans.

updated 42 minutes ago

Politics

**Top Court Rejects Challenge to Trump Census Bid as Premature**

The court's ruling doesn't mean Trump will succeed in his effort.

updated 2 hours ago

Markets

**Near-Disaster in U.S. Treasuries Lights a Fresh Fire for Reform**

The market's brush with disaster in March has reawakened calls to overhaul the system and ease pressure on the Federal Reserve



## Bloomberg Opinion

Sarah Green Carmichael  
**It Always Pays to Take Sabbaticals and Vacations**

Alex Webb  
**Cyberpunk Fiasco Is Going to Make Gaming More Expensive**

Dan Wang  
**New U.S. Restrictions Will Help Make China Great Again**



# Agenda

What is data science?

How do you *do* data science?

How does data science relate to *my library*?

What are the benefits of learning data science for law librarians?

⇒ How do I get started?

# *Focus on improving each component of data science*

## **Data tidying skills**

Excel  
OpenRefine

## **Coding skills**

MARC & MarcEdit  
R or Python

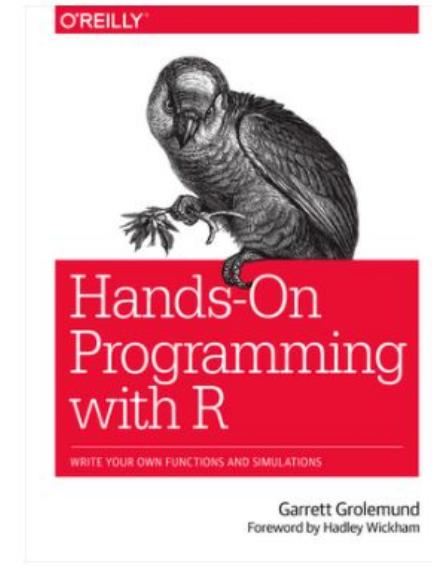
Statistical literacy  
Data visualization

## **Math & Statistical skills**

Feeling adventurous?



A promotional poster for "R for Excel Users" by Julia Lowndes & Allison Horst. The background is a sunset over the San Francisco skyline. On the left, there are two purple cartoon characters wearing hats: one with "I ❤️ SF" and another with "I ❤️ rstats". There are also two small circular photos of real people, Julia and Allison. The title "R for Excel Users" is in large white font. Below it, the authors' names are listed. At the bottom, the "rstudio::conf" logo is shown with the text "SAN FRANCISCO // JANUARY 27 - 30, 2020" and "from RStudio".



# Questions?

*Find me at*

[sarah@rstudio.com](mailto:sarah@rstudio.com)

<http://sarah.rbind.io/>

[linkedin.com/sarahemlin](https://www.linkedin.com/in/sarahemlin)

[@sarahemlin](#) on Twitter & Github