

Intro to Data Science for Law Librarians

Australian Law Librarians Association

23 October 2020





Sarah Lin

Information Architect & Digital Librarian
RStudio, PBC

--

sarah.lin@rstudio.com
<http://sarah.rbind.io/>
@sarahemlin
she/her/hers

Agenda

What is data science?

How do you *do* data science?

How does data science relate to *my library*?

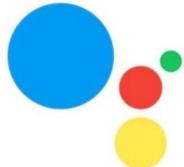
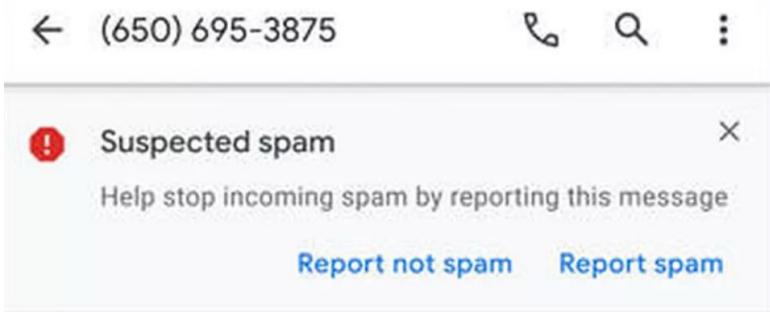
How do I get started?

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



PREDICTIVE POLICING



“Hey Alexa”

“Hey Siri”

“Hey Google”



Frequently bought together



Agenda

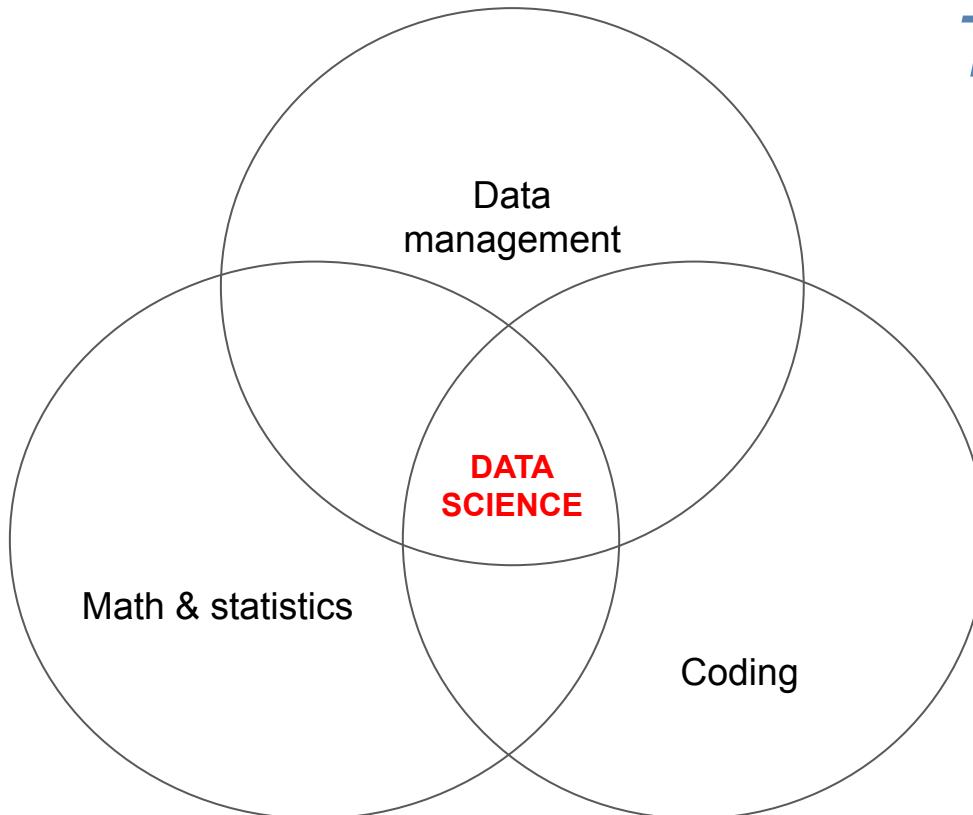
⇒ What is data science?

How do you *do* data science?

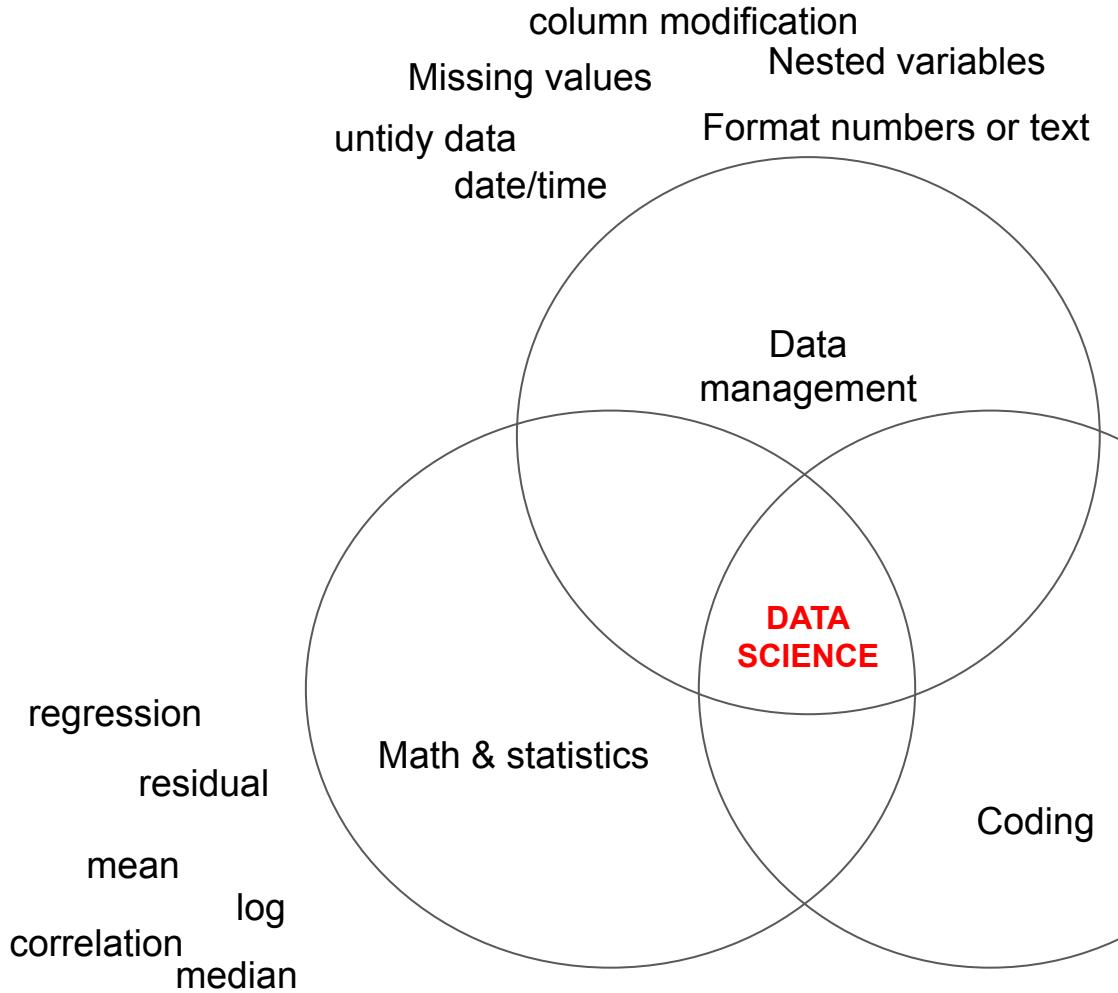
How does data science relate to *my library*?

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

```
204 #readr guesses the column type based on the first 1k rows, but sometimes
205 #the data poses a challenge
206 challenge <- read_csv(readr_example("challenge.csv"))
207 #to see the problems listed out
208 problems(challenge)
209 #tweak the x column
210 challenge <- read_csv(
211   readr_example("challenge.csv"),
212   col_types = cols(
213     x = col_double(),
214     y = col_character()
215   )
216 )
217 #but that leaves the y column weird--it looked like a date
218 challenge <- read_csv(
219   readr_example("challenge.csv"),
220
230:1 (Top Level) ◊
```

```
>
> library(tidyverse)
— Attaching packages ————— tidyverse 1.2.1 —
✓ ggplot2 3.2.1    ✓ purrr  0.3.3
✓ tibble  2.1.3    ✓ dplyr  0.8.3
✓ tidyr   1.0.0    ✓ stringr 1.4.0
✓ readr   1.3.1    ✓ forcats 0.4.0
— Conflicts ————— tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()   masks stats::lag()
> |
```

Name	Size	Modified
.RData	37.1 MB	Nov 19, 2019, 8:57 AM
.Rhistory	18.5 KB	Dec 4, 2019, 12:22 PM
challenge.rds	31.9 KB	Nov 30, 2019, 11:30 AM
diamonds.csv	2.3 MB	Nov 20, 2019, 6:39 PM
diamonds.pdf	12.2 KB	Nov 20, 2019, 6:39 PM
diamonds.R	131 B	Nov 20, 2019, 6:40 PM
Notebook.R	6.8 KB	Dec 2, 2019, 1:54 PM
R4DS.Rproj	205 B	Feb 5, 2020, 3:23 PM



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



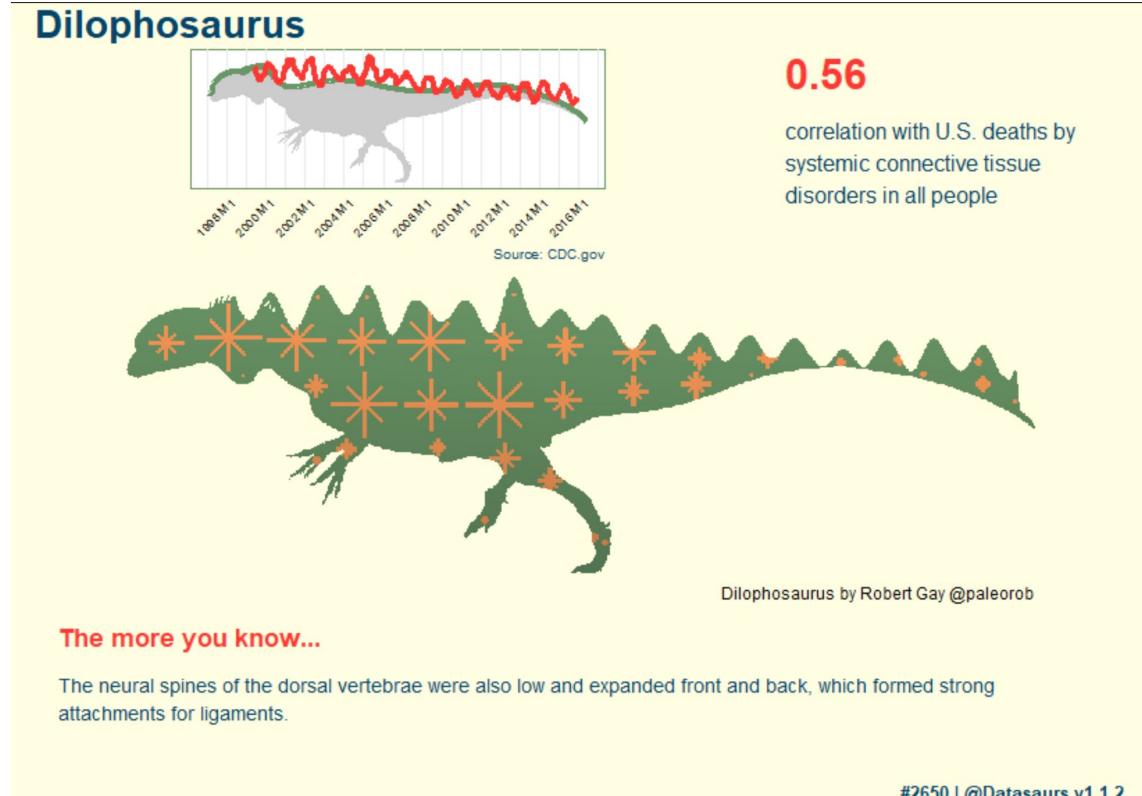
Visualize



Statistical analysis

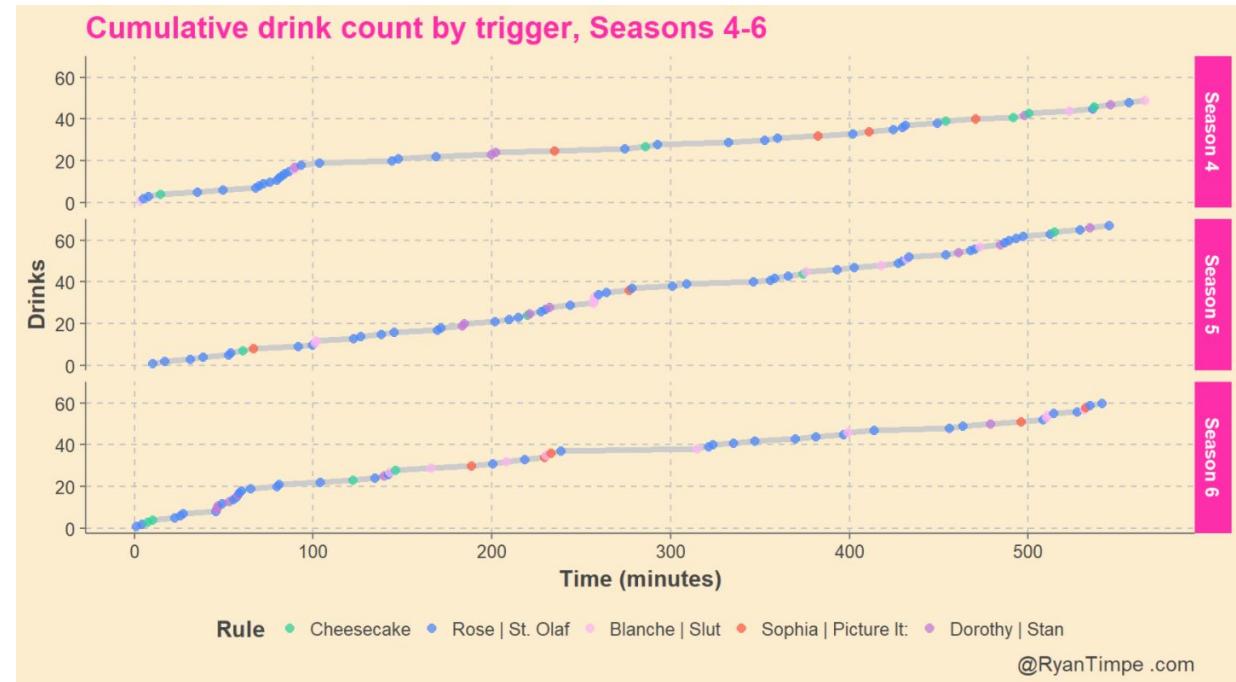


Making meaning



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

COVID-19 Open-source Infection Dynamics

Source Code Twitter

COVOID | SEIR Model

Introduction Parameterise model Define interventions Run simulation Report canvas Animation

Parameterise model

Simulation settings

Location: Australia

Initial number susceptible: 25498681

Age distribution

Initial conditions

Initial number exposed: 1000

Age distribution of exposed cases

Match national age distribution (radio button selected)

Initial number infectious: 200

SEIR model parameters

The initial unmitigated reproduction number (R_0)

Duration of the latent period in days ($1/\sigma$)

SEIR Model Compartmental diagram

Duration of infectious period in days ($1/\gamma$)

SEIR Model Compartmental diagram

```
graph LR; E[E] -- "0.016" --> I[I]; I -- "0.182" --> R[R];
```



Download Files

File type

Major topic / Special groupings

Standard Projections
(Estimates and Projection variants)

[Population data](#) [Fertility data](#) [Mortality data](#) [Migration data](#) [Annual and single age data](#)

CSV format

Probabilistic Projections
(including prediction intervals)

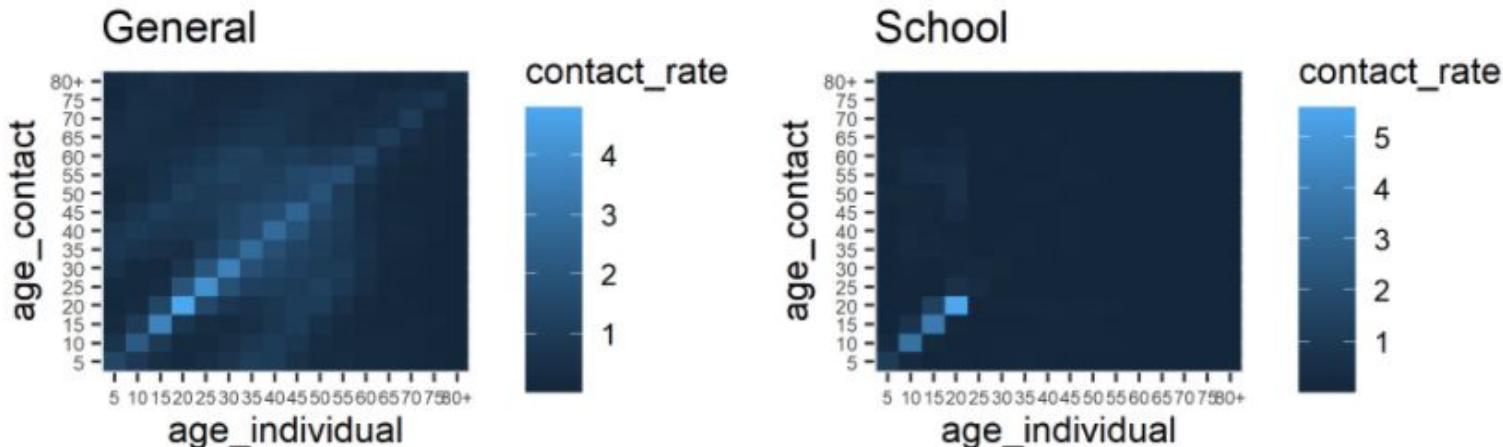
Special Aggregates

Metadata

Archive (previous revisions)

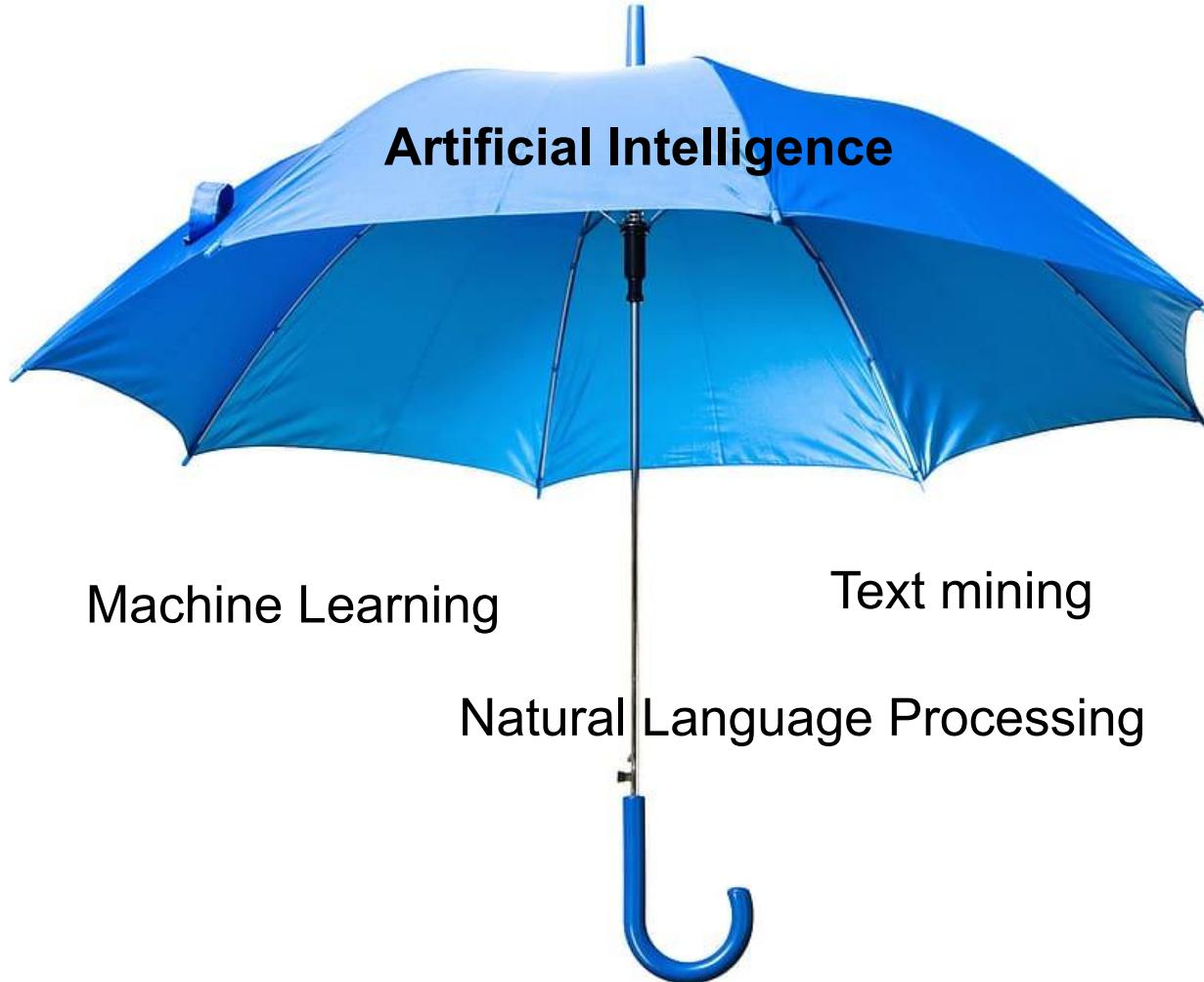
```
library(covoid)
cm_oz <- import_contact_matrix("Australia","general")
p_age_oz <- import_age_distribution("Australia")
# initial conditions
S <- p_age_oz*4e6
I <- c(0,0,0,0,5,3,1,2,0,0,0,0)
R <- rep(0,length(S))
```

```
cm_oz_all <- import_contact_matrix("Australia","general")
cm_oz_sch <- import_contact_matrix("Australia","school")
# separate out school and general population contact rates
cm_oz_all <- cm_oz_all - cm_oz_sch
p_all <- plot(cm_oz_all) + labs(title = "General") +
  theme(axis.text.x = element_text(size=6, angle=0),
        axis.text.y = element_text(size=6))
p_sch <- plot(cm_oz_sch) + labs(title = "School") +
  theme(axis.text.x = element_text(size=6, angle=0),
        axis.text.y = element_text(size=6))
gridExtra::grid.arrange(p_all,p_sch,ncol=2)
```



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

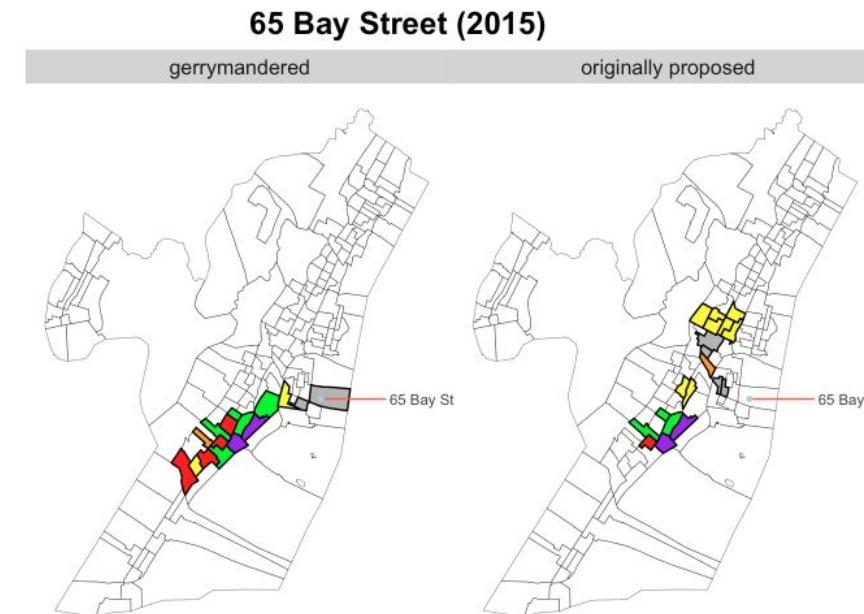
wpinvestigative / kushner_eb5_census

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=ni_hf, aes(x=long, y=lat, group=group), fill=NA, color="black", size=.1)
nj_map <- nj_map + geom_polygon(data=ni_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(knitr)
library(dplyr)
library(eechidna)
data(tpp16)
data(abs2016)

# Join 2016 election and Census
data16 <- left_join(tpp16 %>% select(LNP_Percent, UniqueID), abs2016, by = "UniqueID")

# See what it looks like
data16 %>%
  select(LNP_Percent, DivisionNm, Population, Area, AusCitizen, MedianPersonalIncome, Renting) %>
  head() %>%
  kable
```

LNP_Percent	DivisionNm	Population	Area	AusCitizen	MedianPersonalIncome	Renting
41.54	CANBERRA	196037	2004.2730	85.47927	695.4141	29.74457
36.11	FENNER	202955	460.3633	83.64317	668.7201	33.79648
51.44	BANKS	155806	49.4460	80.66827	440.7940	31.32000
41.70	BARTON	172850	39.6466	72.40498	436.0027	37.23520
59.72	BENNELONG	168948	58.6052	74.59277	495.5510	36.09845
66.45	BEROWRA	145136	749.6359	87.35600	542.7789	14.50767

R Packages for Election Data

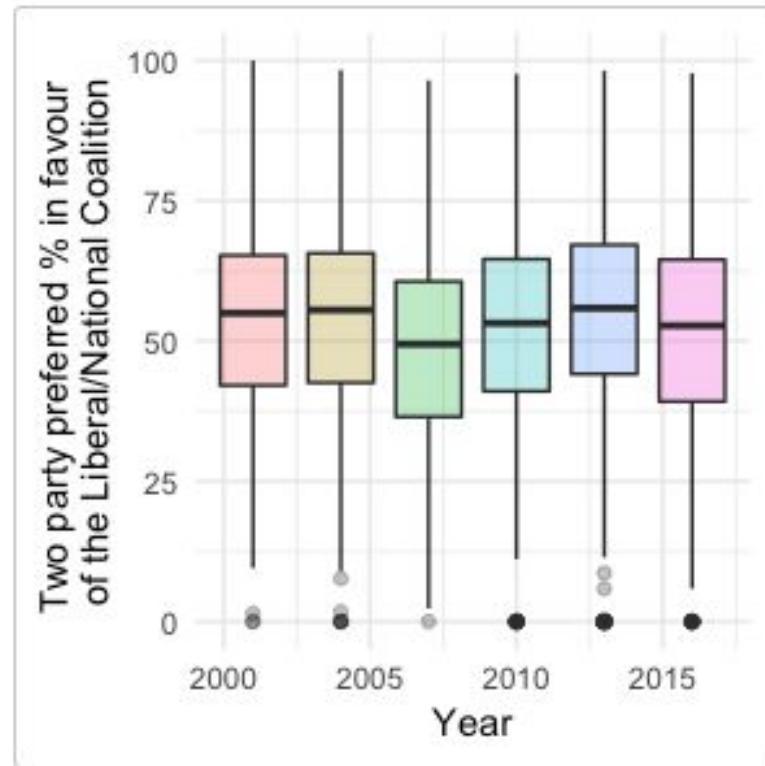
eechidna

ggparliament

ggplot

leaflet

tileR



Measure

Arrest risk ratio

Year

2015

Age

Adult

Race

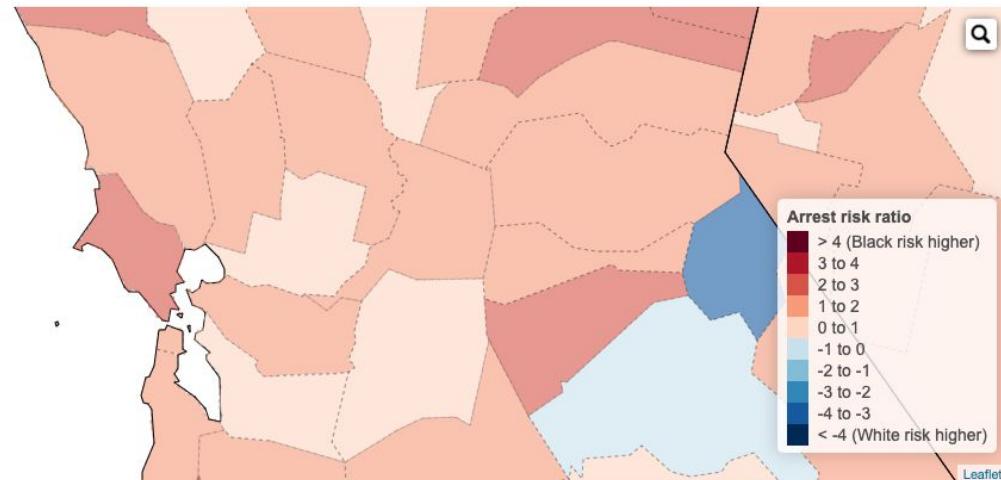
Black

County

California, Yolo

[Reset zoom](#)

Adult Black / White Arrest Risk Ratio



Measure

Arrest risk ratio

Year

2015

Age

Juvenile

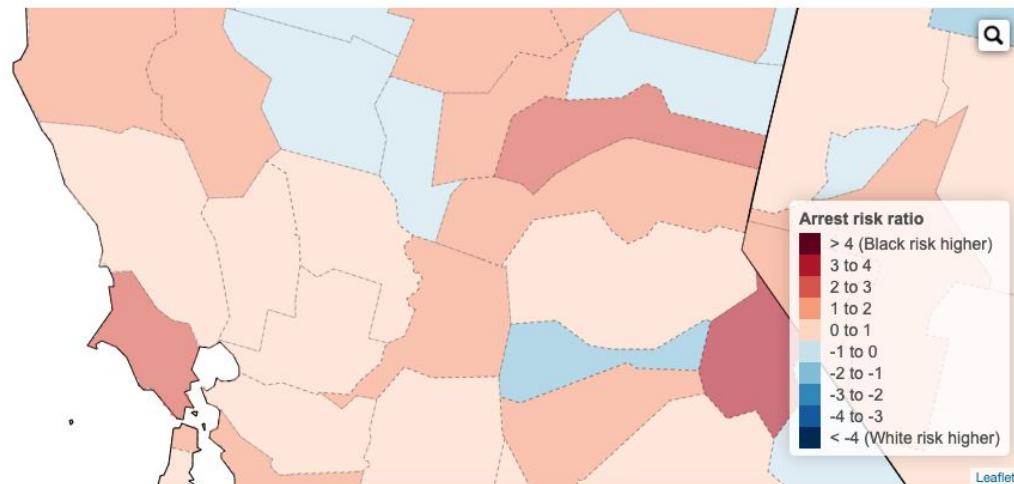
Race

Black

County

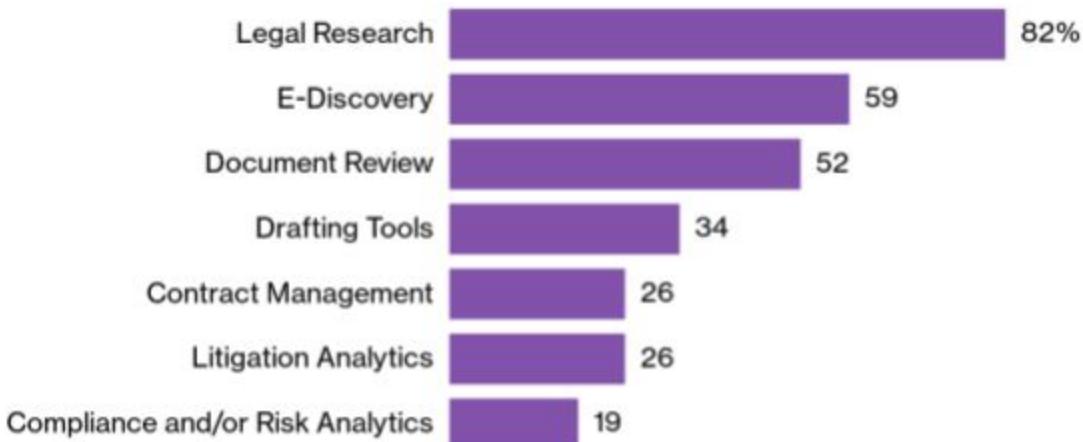
California, Yolo

Juvenile Black / White Arrest Risk Ratio

[Reset zoom](#)

Leaflet

Some Types of AI-Driven Tech Are Used Heavily

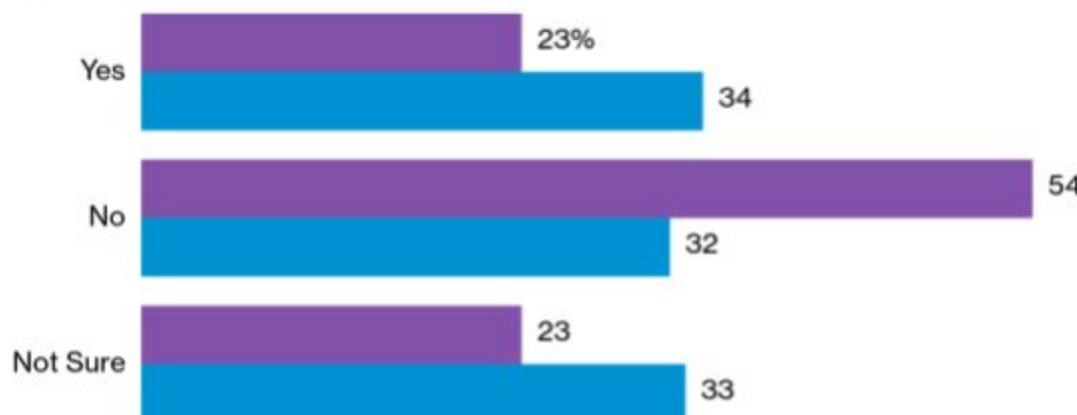


Source: Bloomberg Law Legal Technology Survey (2020), n=331 in-house and law firm respondents

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?

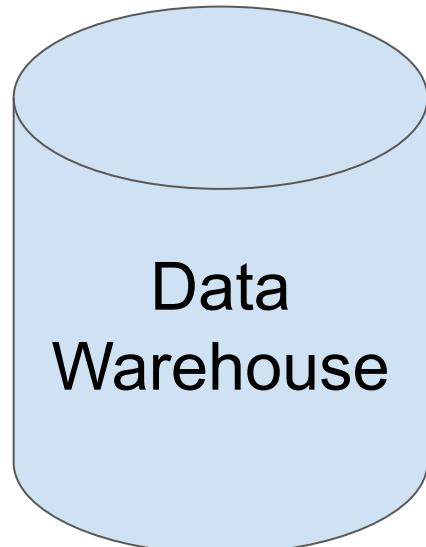
■ 2019 ■ 2020



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

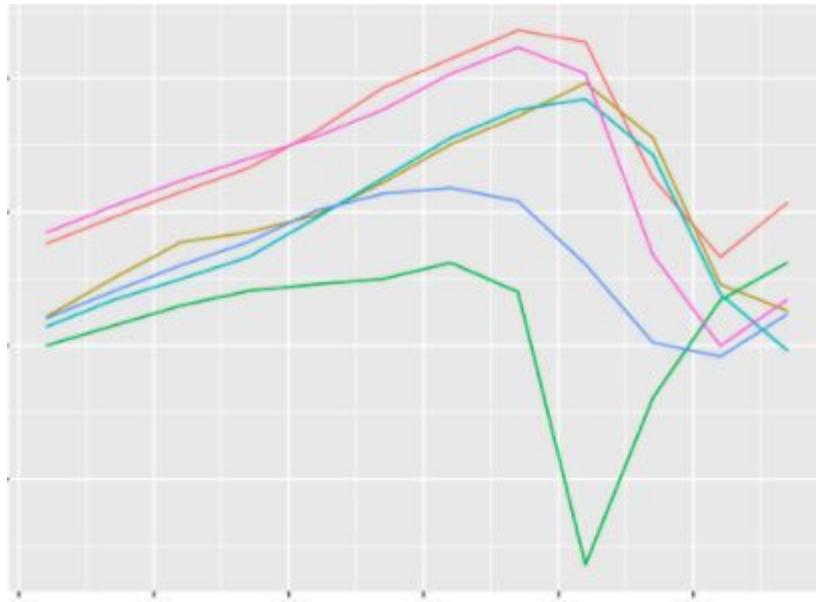
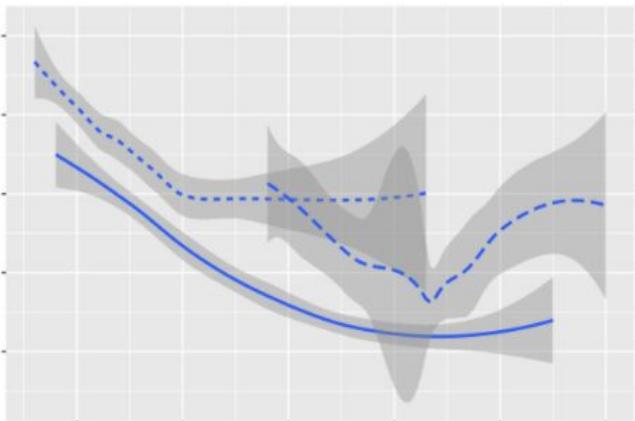
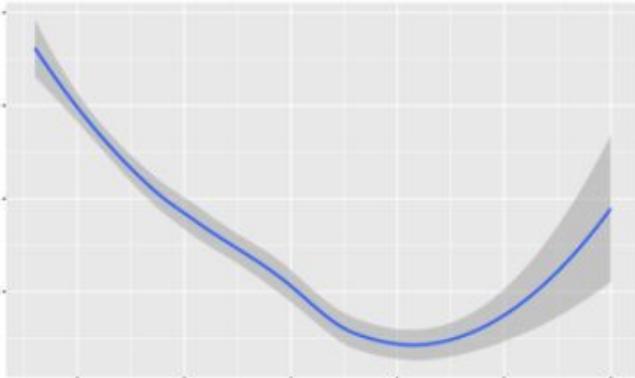
Bloomberg Law

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](#)). 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](#) pageviews and users from the day it was posted until now.

2. Windowed Views. This tab counts [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 2020-10-19.

Search

Title	Post Date	Users	Page Views	Unique Page Views	Mean Time on Page	Mean Bounce Rate
Help Delphi's COVIDcast Project fight the pandemic!	2020-10-13	306	350	313	23	52%
Fake Survival Data for the Disease Progression Model	2020-10-08	612	723	669	24	45%
R Package Integration with Modern Reusable C++ Code Using Rcpp - Part 6	2020-09-28	424	538	446	33	28%
August 2020: "Top 40" New CRAN Packages	2020-09-22	4,985	6,986	5,411	64	34%
Some Thoughts on R / Medicine 2020	2020-09-16	638	730	662	32	49%

Tettra usage 2019-2020

Sarah Lin

9/2/2020

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? Here are the top 25 pages over the last year.

Show 10 entries

Search:

	PageTitle	sum(Views)
1	Solutions Engineering	439
2	AWS Subaccounts: Accessing with assumeroles	317
3	Home Office Favorites	306
4	RStudio Hardware Standards	273
5	Company Holidays	189
6	2020 Beta/Experimental Programs - CS Involvement	188
7	Do I contact DevOps or IT for my technical needs?	187
8	Travel to Boston	186
9	2020 Product Marketing Planning and Resources	179
10	What do I contact DevOps for?	173

Showing 1 to 10 of 25 entries

Previous 1 2 3 Next

Data
analysis

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.

```

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 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)

```

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

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.

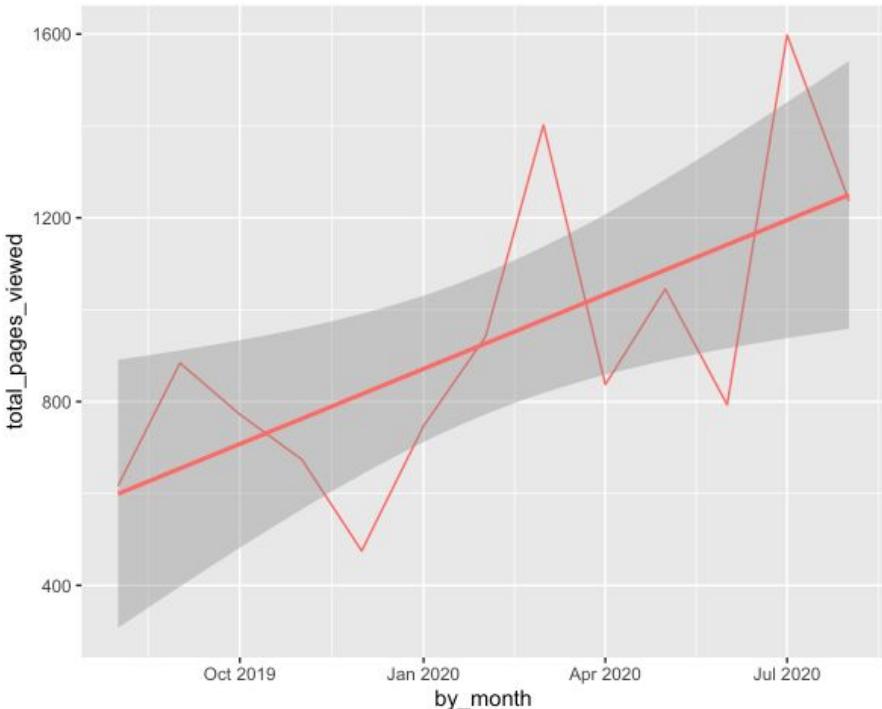
Shareable code

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 is definitely trending upwards.

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



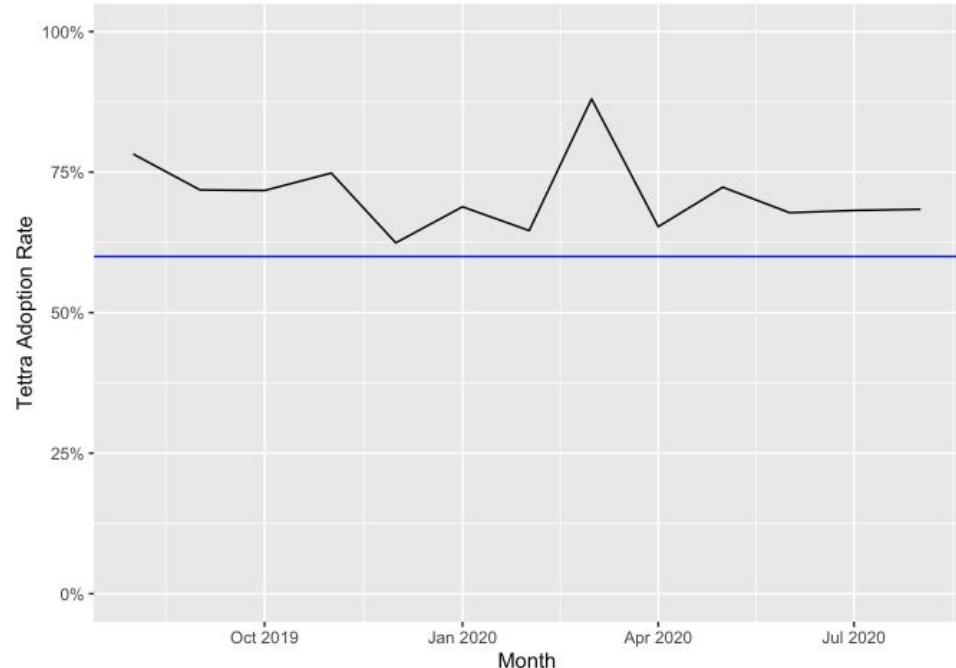
```
119 ````{r include = FALSE}
120 #read in the PageViews-2019-2020 Google Sheet
121 views <- read_sheet("https://docs.google.com/spreadsheets/d/1zuZvXI965VQ1qzyJ0
122 _HwEp84ZNx1z9fDvK9Qh8u-S_s/edit#gid=0")
123 print/views)
124 ````
```

Total pageviews over the past year is definitely trending upwards.

```
128 ````{r}
129 ggplot(data = views, aes(by_month, total_pages_viewed, color = "red"))
130   geom_line() +
131   geom_smooth(method = "lm")
```

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?



```
264 ## Engagement Summary
265
266 Tettra is clearing being used, and used a lot by some employees. Another measure
267 adoption rate: what percentage of employees are engaging with Tettra, both on the
268
269 # views is the table of total views per month
270 employees <- read_sheet("https://docs.google.com/spreadsheets/d/1mRYd3bH75wfW83sR9
271 ")
272
273
274
275 # views$total_people is the count of employees who've viewed at least one page
276 # employees$employee_count is the count of people employed by month
277
278 # adoption rate = number of readers / number of employees OR
279 # adoption rate = views$total_people / employees$employee_count
280
281 utilization <- employees %>%
282   mutate(
283     viewers = views$total_people,
284     adoption = views$total_people / employees$employee_count)
285
286 ggplot() +
287   geom_line(mapping = aes(x = month, y = adoption), data = utilization) +
288   labs(x = "Month", y = "Tettra Adoption Rate") +
289   scale_y_continuous(labels = scales::percent, limits = c(0,1)) +
290   geom_hline(yintercept = .6, color = "blue")
291
```

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](#)



[STOCKS](#)[BONDS](#)[COMMODITIES](#)[CURRENCIES](#)Sponsored by **MERRILL** 

S&P 500 3,452.31 **+0.74%**

Dow Jones 28,378.08 **+0.65%**

FTSE 100 5,889.22 **+0.08%**

Nikkei 225 23,567.04 **-0.44%**

THE WIRE



① 5m ago

VW gauges interest in Ducati brand as a prelude to potential sale: sources

① 6m ago

Canada police officer found not guilty in death of mentally ill black man: CBC

① 7m ago

'For heaven rest us, we're not Asbestos': Canadian town chooses new name

① 8m ago

Oil holds steady; traders look to U.S. stimulus talks for hope



U.S. says Google breakup may be needed to end violations of antitrust law

The U.S. Justice Department filed an antitrust lawsuit against Google, claiming the \$1 trillion company uses its market power to fend off rivals and said nothing was off the table, including a breakup of the internet search and advertising company.

- Regulatory probes Google is facing in Asia and

Trump pushes for major COVID-19 deal over Senate Republican objections

10:47AM EDT

Trump to campaign in battleground Pennsylvania as voting enters final two weeks

10:55AM EDT

Berkshire Hathaway to pay \$4.1 million to settle alleged Iran sanctions violations

10:42AM EDT

U.S. single-family homebuilding, permits surge to more than 13-year high



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

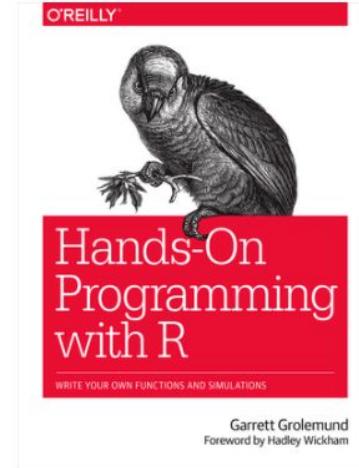
Statistical literacy
Data visualization

Math & Statistical skills

Feeling adventurous?

RYouWithMe: The plan

The result: **RYouWithMe** - a series of online learning resources for using R - designed by us, for us. The series is oriented towards R beginners and aims to provide a solid foundation of R skills. That way, future R-Ladies Sydney activities can be accessible to more of our members. Frustration no more! (or at least minimised!)



Questions?

Find me at

sarah.lin@rstudio.com

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

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

[@sarahemlin](https://twitter.com/sarahemlin) on Twitter & Github