
AALL
YOUR LEGAL
KNOWLEDGE
NETWORK™

Intro to Data Science for Law Librarians

Jocelyn Stilwell-Tong, Moderator
Sarah Lin, Presenter

OUR VALUE PROPOSITION

AALL is dedicated to supporting the career development needs of law librarians through quality educational programming and events designed specifically for legal information professionals.

AALL embraces law librarians in all stages of their careers to provide exceptional experiences, tools for success and premier services to support professional growth.

Intro to Data Science for Law Librarians

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

Data is anything you can put into columns and rows

	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 RainforestKingdom	Book 1, The Dragonet Prophecy	No
4	Sunny	Sandwing	Kingdom of Sand	Pale gold orwhite scales the color of desert sand	Can survive a long time without water, no scorpian like tail , good camoflauge in desert and breath fire	Frostbreath, fire, sandwing tail, spears, poison, chakrams, rainwing venom and other dragons	Captured and imprisoned in(Queen) Scarlet's palce, Seawing palce 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 disapear 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 palce, Seawing palce	Book 1, The Dragonet Prophecy	No
6	Glory	Rainwing	Rainforest Kingdom	Scales constantly change colors	Scales can change colors to blend into suroundings, pehensile 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 palce, Seawing palce	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 orwhite scales the color of desert sand	Can survive a long time without water, no scorpian like tail , good camoflauge 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

Google

Kira

ypDisplay℠ Mobile

PREDICTIVE POLICING

neustar®

Real Intelligence. Better Decisions.™

(650) 695-3875

Suspected spam

Help stop incoming spam by reporting this message

Report not spam Report spam

Taleo

Frequently bought together

"Hey Alexa"

"Hey Siri"

"Hey Google"

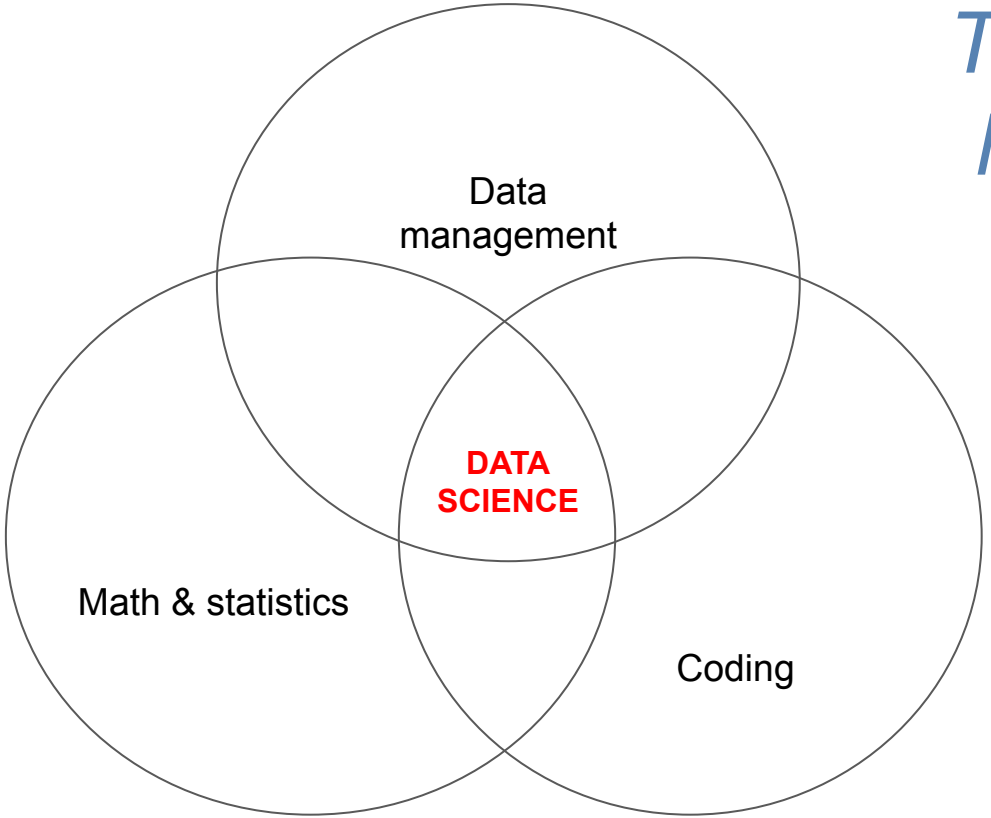
Agenda

⇒ What is data science?

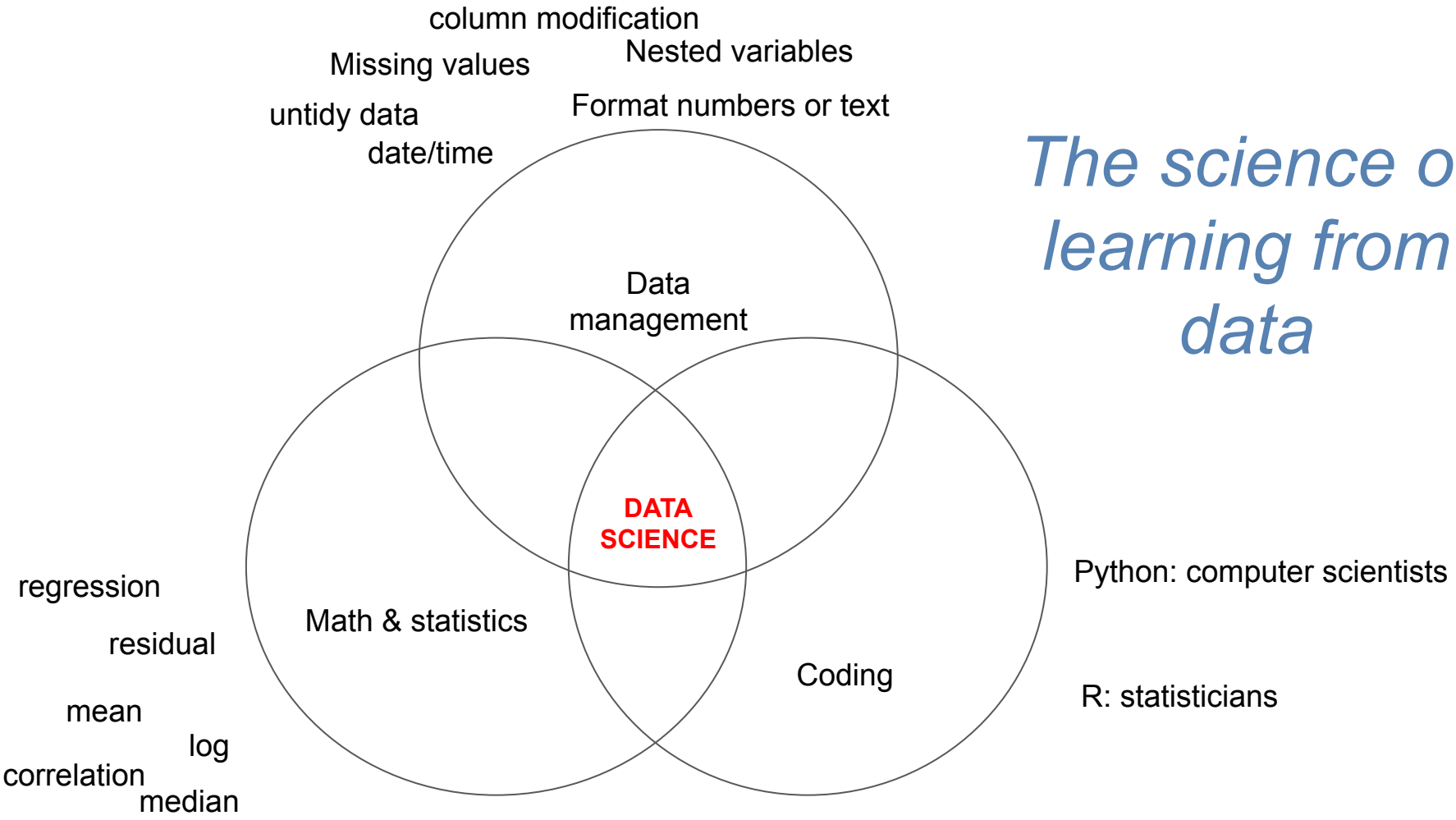
How do you *do* data science?

How does data science relate to *my library*?

How do I get started?



The science of learning from data

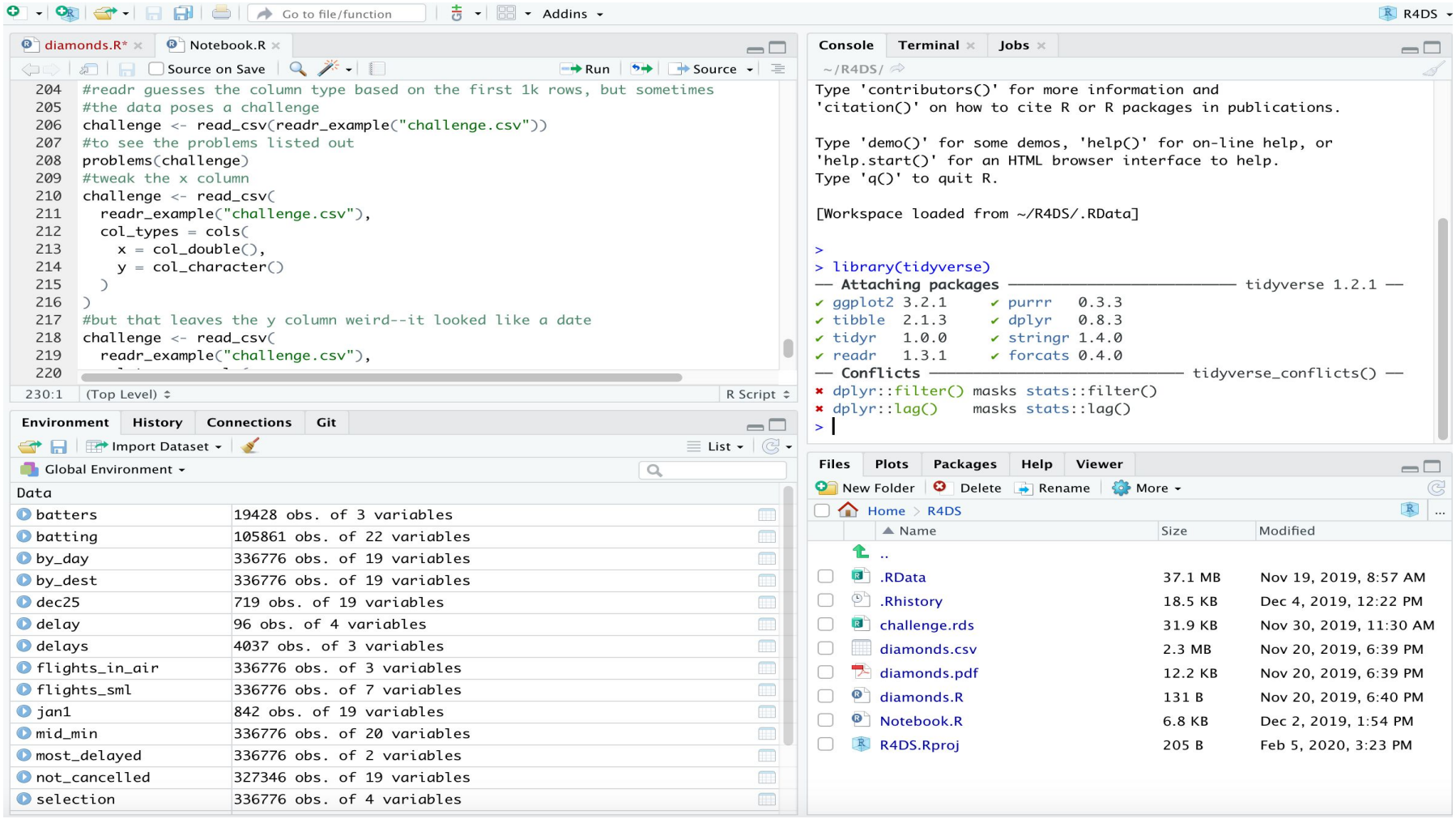


The science of learning from data

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
```


Go to file/function
Addins
R4DS
diamonds.R*
Notebook.R
Source on Save
Run
Source
#readr guesses the column type based on the first 1k rows, but sometimes
#the data poses a challenge
challenge <- read_csv(readr_example("challenge.csv"))
#to see the problems listed out
problems(challenge)
#tweak the x column
challenge <- read_csv(
readr_example("challenge.csv"),
col_types = cols(
x = col_double(),
y = col_character()
)
)
#but that leaves the y column weird--it looked like a date
challenge <- read_csv(
readr_example("challenge.csv"),
230:1
(Top Level)
R Script
Console
Terminal
Jobs
~/R4DS/
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/R4DS/.RData]
> 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()
Environment
History
Connections
Git
Import Dataset
List
Global Environment
Data
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
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
R4DS
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

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

↓

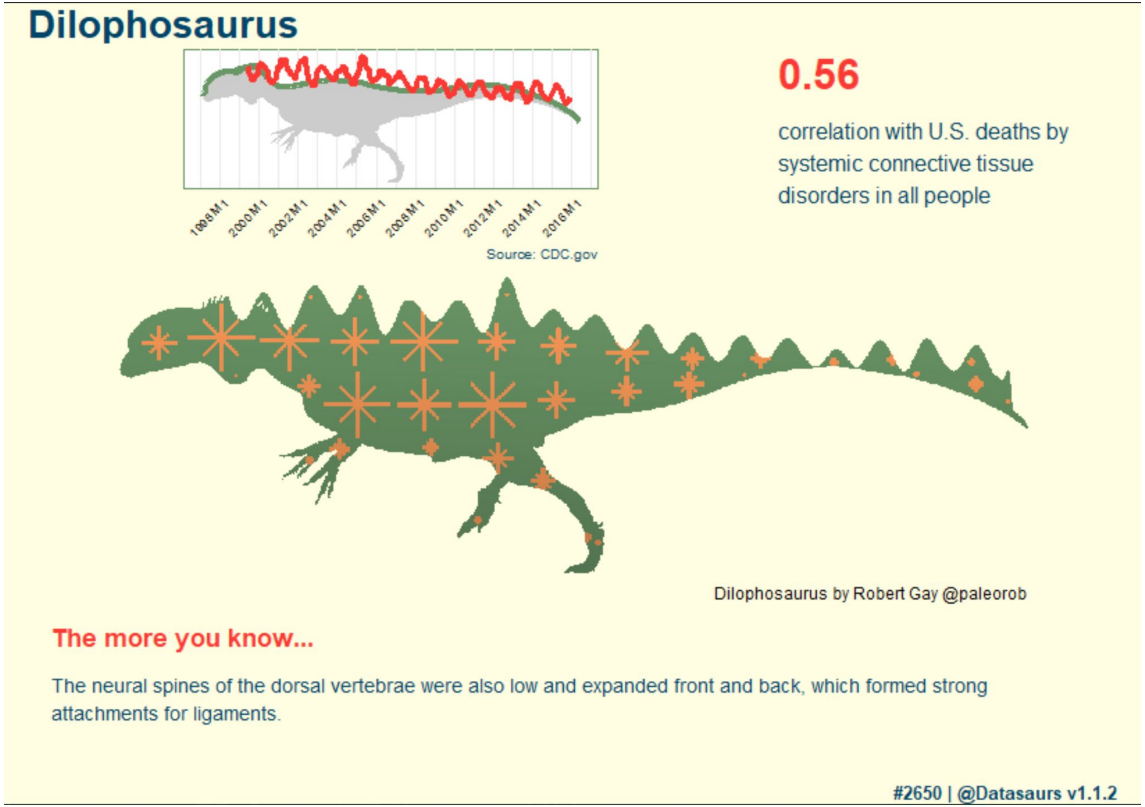
Visualize

↓

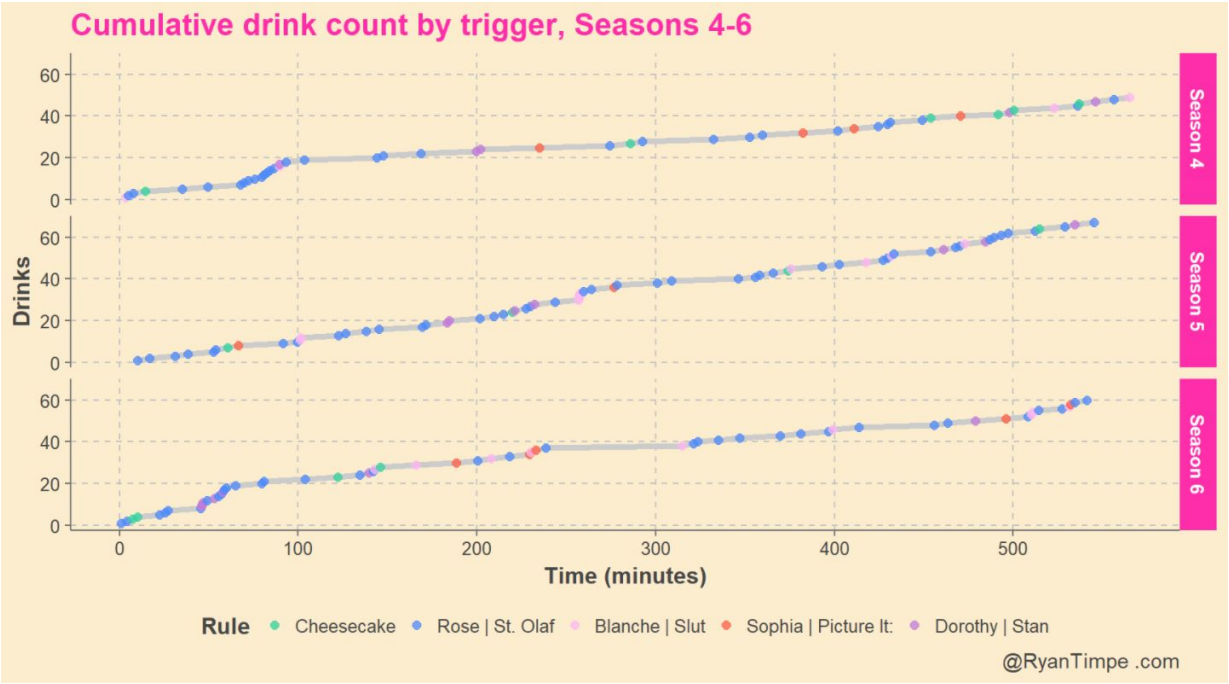
Statistical analysis

↓

Making meaning



Data science + text



Data collection + tidying

↓

text processing

↓

Visualize

↓

Statistical analysis

↓

Making meaning

Choices in Text Processing

Unnesting

Tokenizing

Stopwords

N-grams

Tf-idf

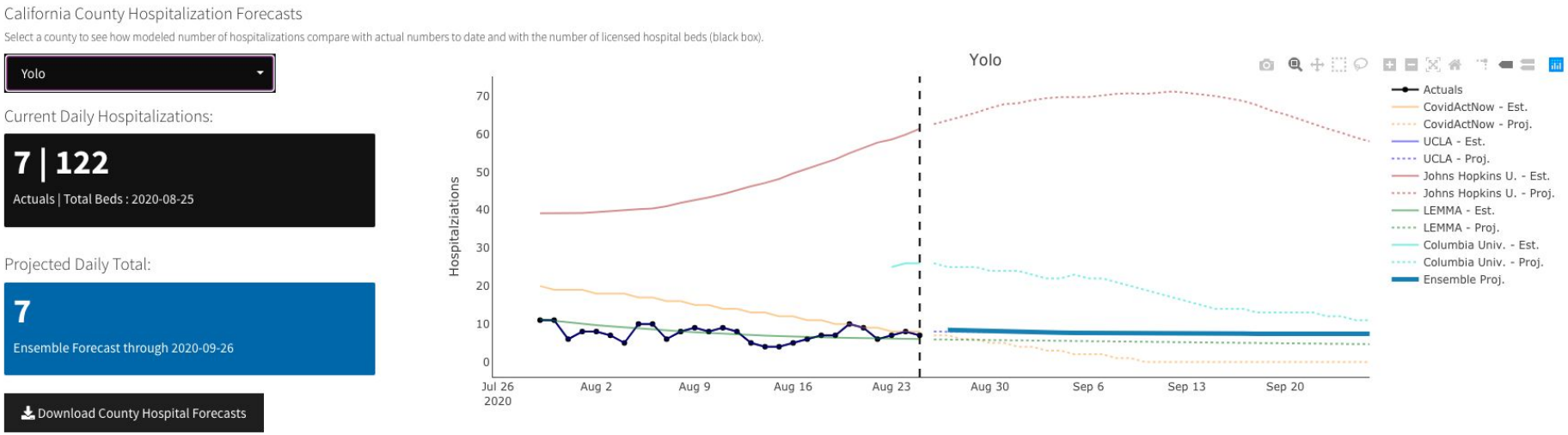
Word embeddings

Sentiment lexicons

Training & testing datasets

Classifiers (models)

Doing data science in R using the Tidyverse



master ▾ CalCAT / data / CA / ucla_state.csv Go to file ···

vargovargo update OS gituhb Latest commit ee63121 on Jul 8 History

2 contributors

138 lines (138 sloc) | 56.5 KB Raw Blame

Search this file...

	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

```r
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)
}
```

```r
#### 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)

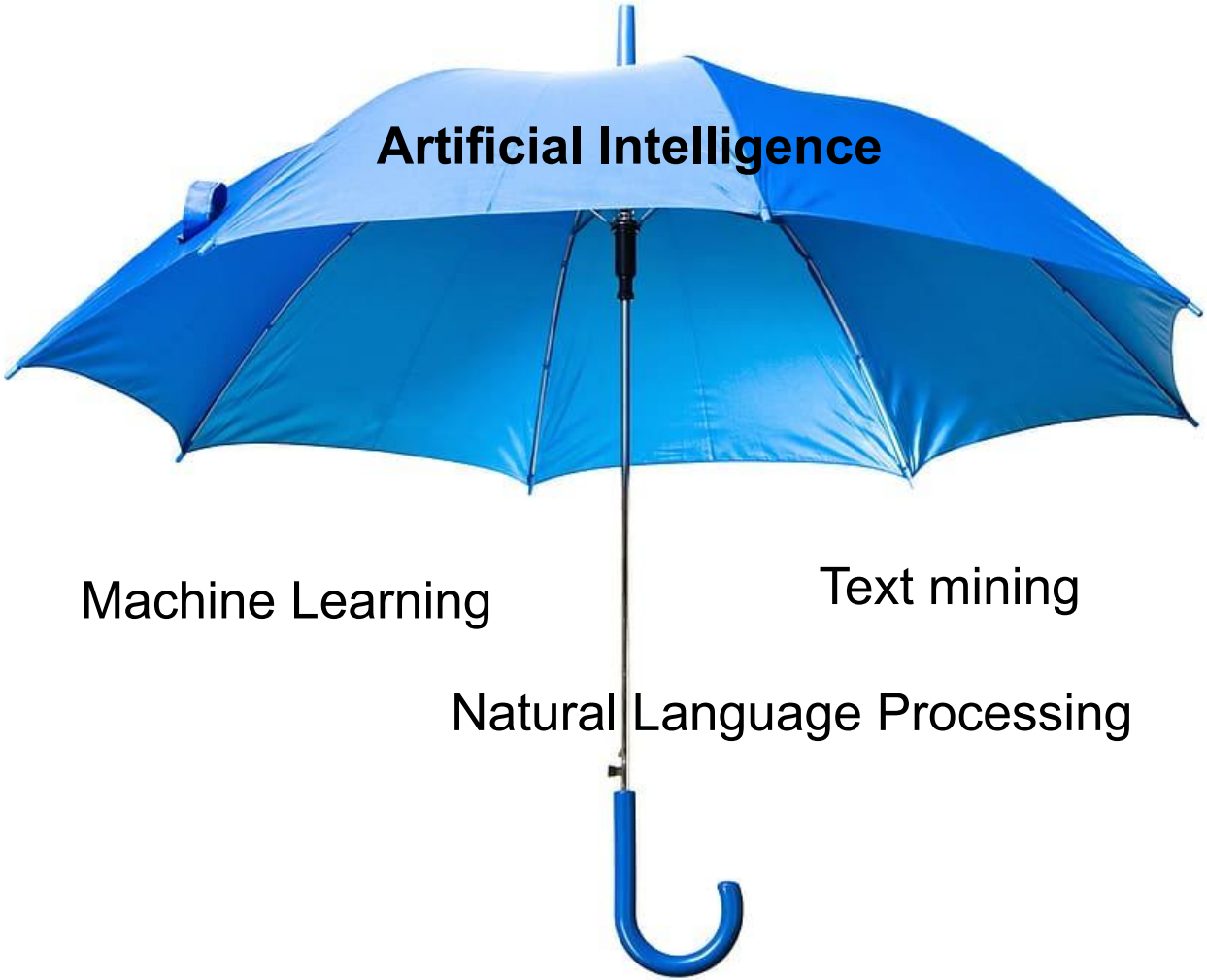
  cnty <- input$select.county.hosp
  progress$inc(3/4)
  # out <- lapply(cnty[1], function(x) get_can_cnty(x))
  # cnty.hosp <- do.call("rbind",out)
  out <- filter(can.county.observed, fips ==  cnty)

  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


Artificial Intelligence
Machine Learning
Text mining
Natural Language Processing

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?

Investigations

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

wpinvestigative / kushner_eb5_census

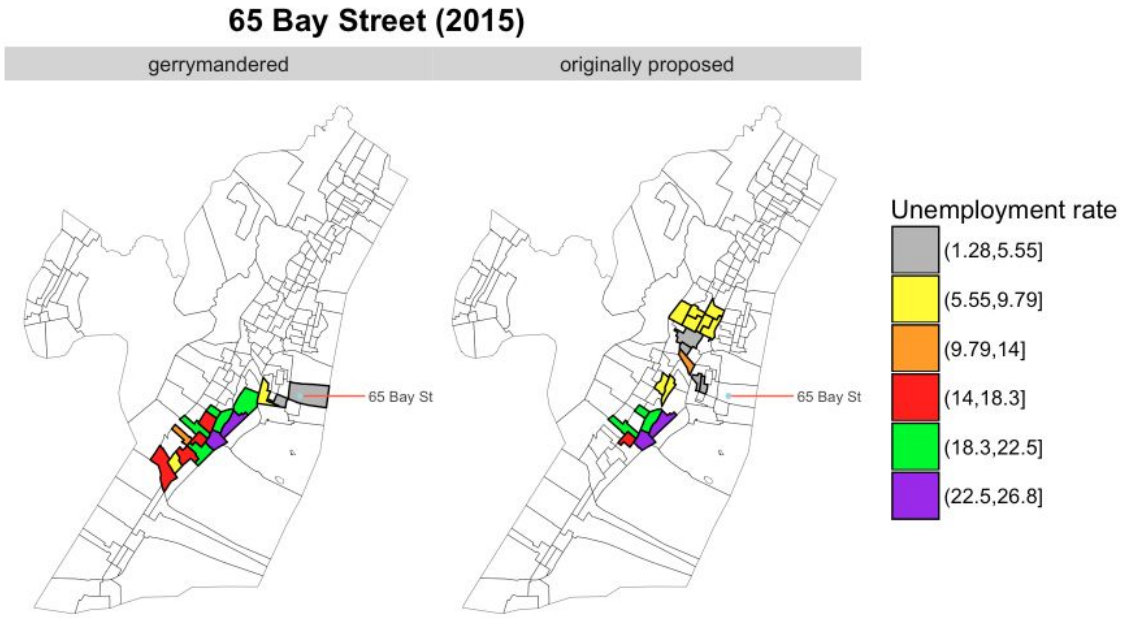
<> Code Issues Pull requests Actions

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(tidycensus)
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")
```

readme.MD

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.githu
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
```

Racial Disparities in Police Arrests Map

Home | Map | FAQ

Measure: Arrest risk ratio

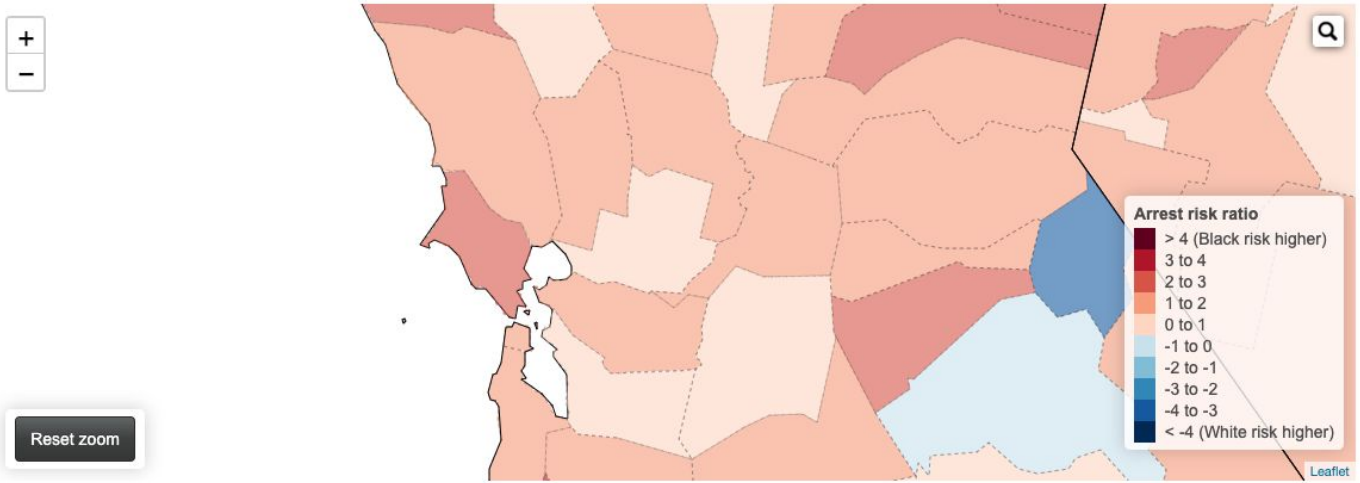
Year: 2015

Age: Adult

Race: Black

County: California, Yolo

Adult Black / White Arrest Risk Ratio



Racial Disparities in Police Arrests Map

Home | Map | FAQ

Measure: Arrest risk ratio

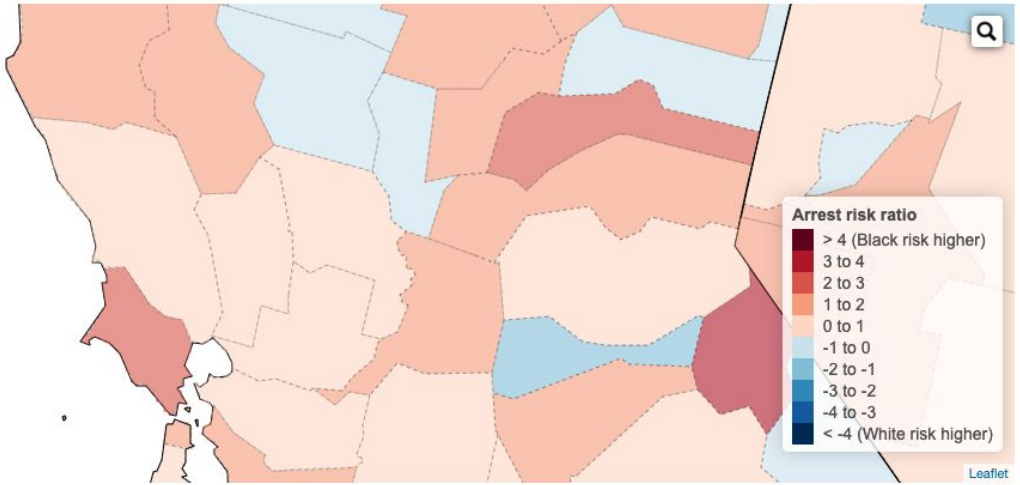
Year: 2015

Age: Juvenile

Race: Black

County: California, Yolo

Juvenile Black / White Arrest Risk Ratio

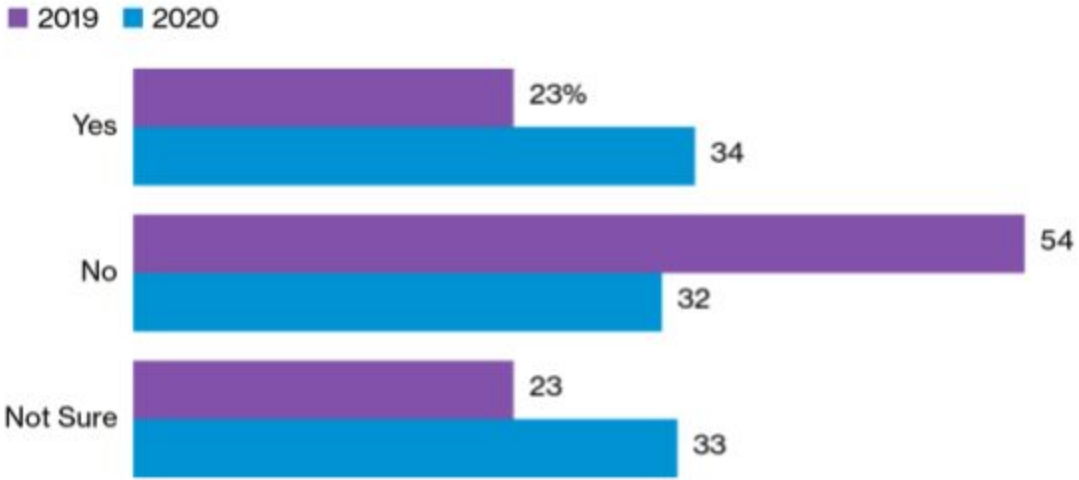


Some Types of AI-Driven Tech Are Used Heavily

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.

Bloomberg Law

Is Your Organization Using AI or Machine Learning Tools?

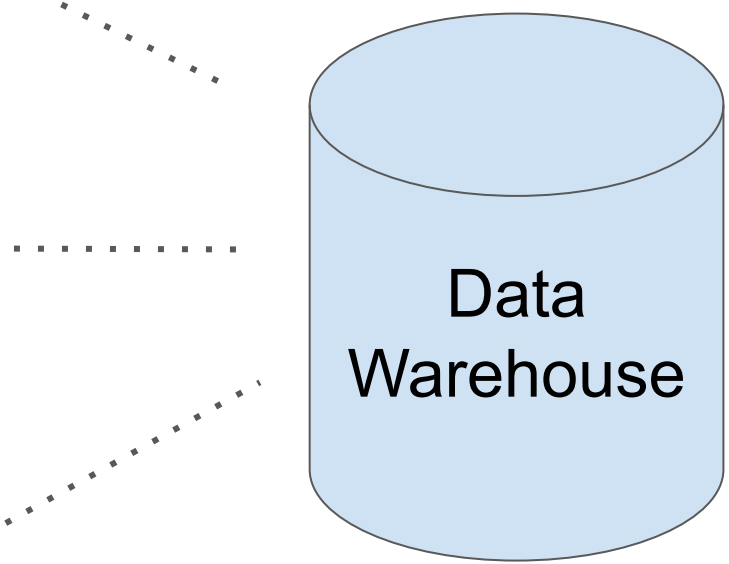


Sources: Bloomberg Law Legal Technology Survey (2020); Bloomberg Legal Operations & Technology Survey (2019)
2020 Survey, n=331 in-house and law firm respondents; 2019 Survey, n=489 in-house and law firm respondents

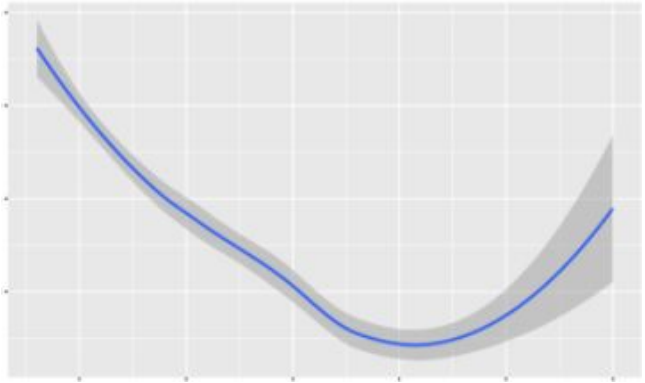
Bloomberg Law

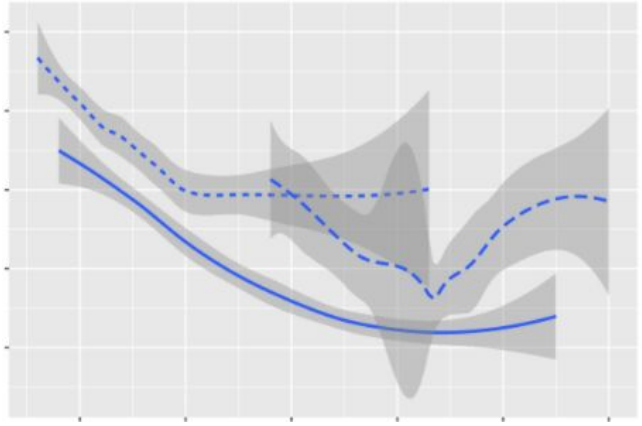
Data stored,
often in the cloud

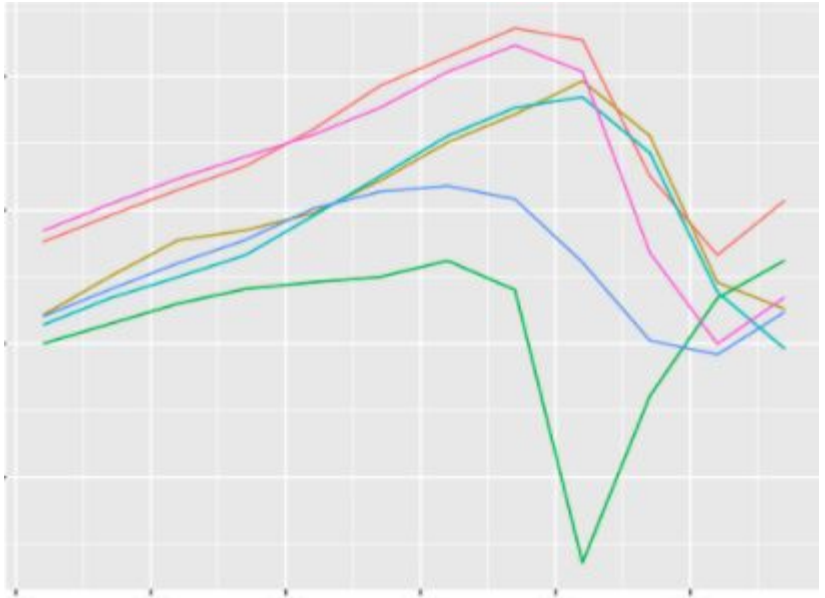
Data
collected
from
software
applications



Data
extracted
for analysis







RViews Article Viewership Metrics

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 2020-09-01.

Search

Title	Post Date	Users	Page Views	Unique Page Views	Mean Time on Page	Mean Bounce Rate
Crowd Counting Consortium Crowd Data and Shiny Dashboard	2020-08-31	64	74	64	138	67%
July 2020: "Top 40" New CRAN Packages	2020-08-27	1,713	2,346	1,804	404	34%
R Package Integration with Modern Reusable C++ Code Using Rcpp - Part 5	2020-08-24	777	947	791	106	26%
R Package Integration with Modern Reusable C++ Code Using Rcpp - Part 4	2020-08-18	393	444	407	170	35%
Monitor COVID-19 at the COVID-19 Forecast Hub	2020-08-10	1,073	1,224	1,114	124	67%
Party with R: How the Community Enabled Us to Write a Book	2020-08-03	1,633	1,798	1,691	174	50%
R Package Integration with Modern Reusable C++ Code Using Rcpp - Part 3	2020-07-31	478	558	490	105	37%

Tettra usage 2019-2020

Sarah Lin

9/2/2020

Examing 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?

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

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.

Text

Data analysis

tettra-usage-2019-2020.Rmd

Knit | Insert | Run

```
---
title: "Tettra usage 2019-2020"
author: "Sarah Lin"
date: "9/2/2020"
output: html_document
editor_options:
  chunk_output_type: inline
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE)
#install.packages("googlesheets4")
#load libraries
library(googlesheets4) #googlesheets4 automatically opens your browser and asks you to authenticate
library(tidyverse)
library(lubridate)
library(scales)
library(DT)
#library(curl)
#handle_setopt(http_version = 2)
```

Examing 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?

Content
Is Tettra just for benefits information? Here are the top 25 pages over the last year.

```{r include=FALSE}
page_names <- read_sheet("https://docs.google.com/spreadsheets/d/1GHDPp7k4miLWbSTGtHY0AEAzPp0NaZ4hCyBI7LUpp_0/edit#gid=0")
```

Access to raw data is limited

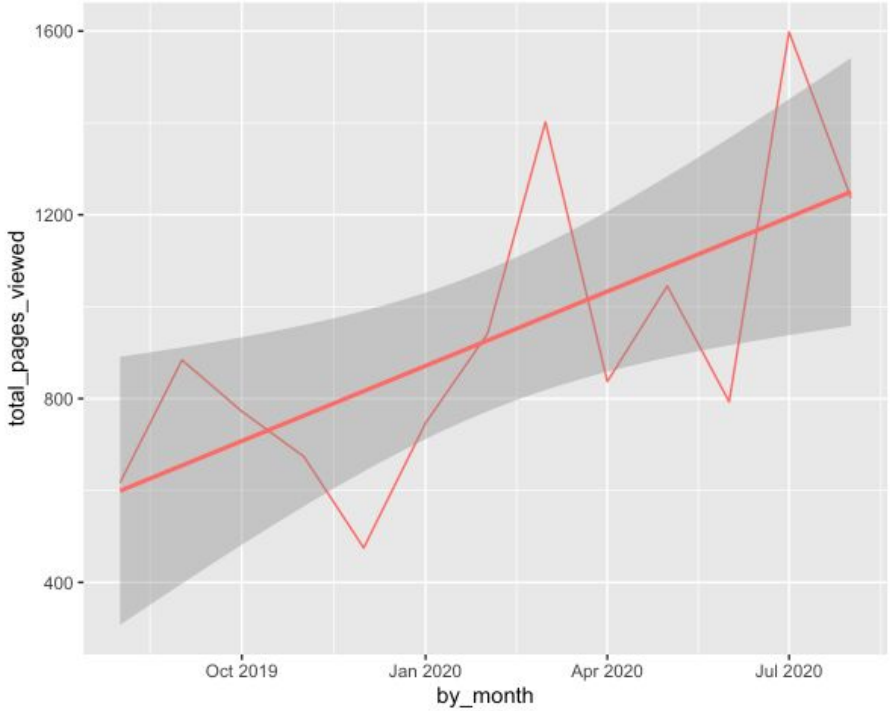
```
pages <- page_names %>%
  select(c(Month, Views, PageTitle, Name))

pages1 <- pages %>%
  group_by(PageTitle) %>%
  summarise(sum(Views))

total_views <-  pages1 %>%
  arrange(desc(pages1$`sum(Views)`))

top25 <- total_views %>%
  slice_head(n = 25)
```

Shareable code

```{r}
datatable(top25)
```

PageTitle <chr>	sum(Views) <dbl>
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

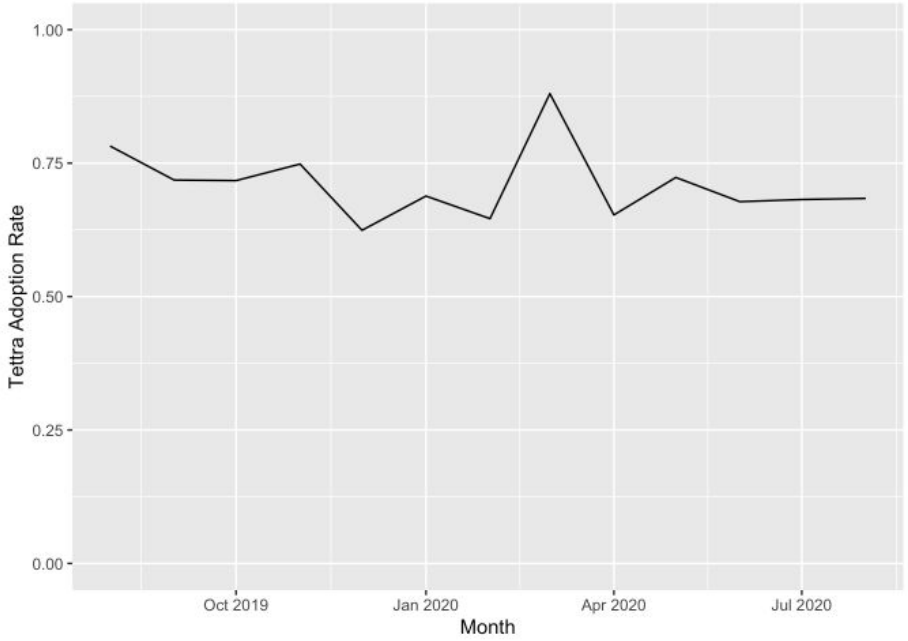
```

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.
```
```

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'
```




````
```{r include = FALSE}
#read in the PageViews-2019-2020 Google Sheet
views <- read_sheet("https://docs.google.com/spreadsheets/d/1zuZvXI965VQ1qzyJ0_HwEp84ZNx1z9fDvK9Qh8u-S_s/edit#gid=0")

print(views)
```

Total pageviews over the past year is definitely trending upwards.

```{r}
ggplot(data = views, aes(by_month, total_pages_viewed, color = "red")) +
  geom_line() +
  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?




```
## 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?

```{r include=FALSE}
# views is the table of total views per month
employees <- read_sheet("https://docs.google.com/spreadsheets/d/1mRYd3bH75wfW83sR9f5_Nlg64JpQtNiVbF2TGsjgFRg/edit#gid=0")
```

```{r}

# views$total_people is the count of employees who've viewed at least one page
# employees$employee_count is the count of people employed by month

# adoption rate = number of readers / number of employees OR
# adoption rate = views$total_people / employees$employee_count

utilization <- employees %>%
  mutate(
    viewers = views$total_people,
    adoption = views$total_people / employees$employee_count)

ggplot() +
  geom_line(mapping = aes(x = month, y = adoption), data = utilization) +
  ylim(0,1) +
  labs(x = "Month", y = "Tettra Adoption Rate")

```
```

~~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

- 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

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

2000 INDEX 1536.78	DOLLAR INDEX SPOT 92.99	USD-JPY X-RATE 104.57	EUR-USD 1.18	Gold Spot $/Oz 1950.86
▼ -5.82 -0.38%	▲ +0.016 +0.02%	▼ -0.170 -0.16%	▼ -0.001 -0.07%	▲ +6.42 +0

LATEST

Technology
WeChat iPhone Downloads Surge in the U.S. Ahead of Trump Ban
1h ago

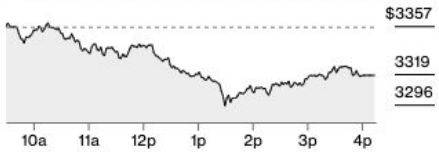
Hyperdrive
Luminar's Young Founder to Have Full Control When Firm Goes Public
1h ago

Business
Business Chiefs Push Cuomo, de Blasio to Bolster NYC After Covid
updated 1h ago

Stocks Slide
S&P 500 Index

3,319.47 USD -37.54 -1.12% ▼

1D 1M 1Y 5Y



As of 04:20 PM EDT

Markets
Stocks Hit Six-Week Low as Tech Slide Accelerates
updated 57 minutes ago

- U.S. Consumer Sentiment Climbed to a Six-Month High in September
- Bond Market Shows U.S. Is Leading in Race to Reflate Economy
- A $700 Million CMBS Portfolio Is On the Brink as Malls Collapse

Markets
Wall Street Loses Faith That Congress Will Rescue States, Cities
Bank of America, which once expected $400 billion in rescue funds, says expectations are

Prognosis
U.S. New Cases Quicken; French Minister Infected: Virus Update
All the latest updates from the global pandemic.

Wealth
Perelman Selling Almost Everything as Pandemic Roils His Empire
Billionaire's artwork, private jet, company stakes are on the market

Technology
U.S. Expels WeChat, TikTok From App Stores on China Concern
New downloads and updates of the WeChat and TikTok app are prohibited starting Sunday.

BloombergOpinion

Tim Culpan
Trump's WeChat Ban Is Just a MAGA Wall in Cyberspace

Brooke Sutherland
Manufacturing Recovery Is a Slow Game of Inches

Agenda

What is data science?

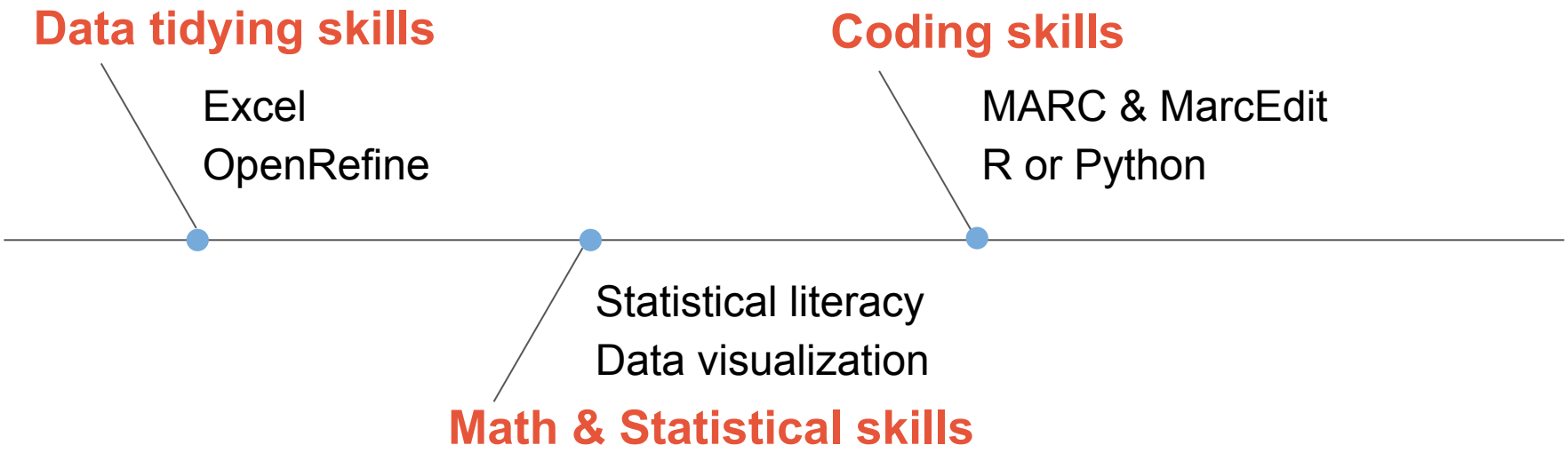
How do you *do* data science?

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



Feeling adventurous?

Questions?

Find me at

sarah.lin@rstudio.com

http://sarah.rbind.io/

linkedin.com/sarahemlin

@sarahemlin on Twitter & Github


AALL
YOUR LEGAL
KNOWLEDGE
NETWORK™