What's missing?

Reduce bias by addressing data gaps in your analysis process

Sarah Moir

Senior Staff Technical Writer | Splunk

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results may differ materially. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only, and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionalities described or to include any such feature or functionality in a future release.

Sarah Moir

Senior Staff Technical Writer | Splunk

splunk> .conf20

Agenda

What I'll cover

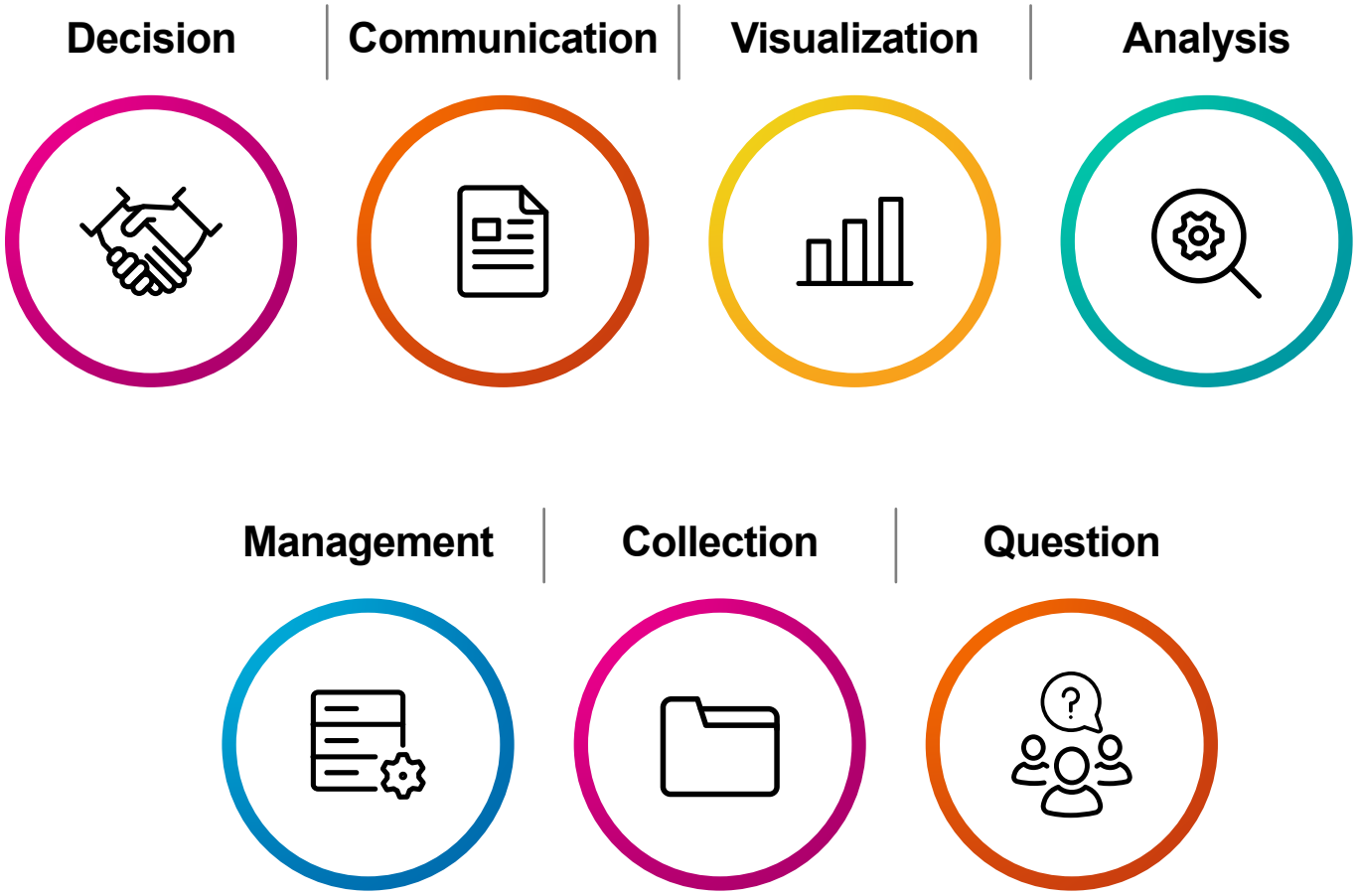
1. Why missing data matters
2. Stages of data analysis
3. Take action

What is missing data?

Why Missing Data Matters

- All data analysis can be biased
- Data-driven decisions feel safe
- Data-driven decisions can be imperfect
- Identify gaps, identify bias

Data Can Go Missing At Any Stage Of The Data Analysis Process



.conf20

splunk>

Decide with the Data

Oregon Health Authority

Deciding with missing data

Is indoor dining safe? Oregon's data can't say

By Erin Ross (OPB)

Portland, Ore. Aug. 7, 2020 6 a.m.

OHA says data shows "no significant transmission" at Oregon bars and restaurants, but they're also not tracking it specifically.

https://www.opb.org/article/2020/08/07/bar-restaurant-coronavirus-safe-oregon/

Oregon Health Authority

Deciding with missing data

- Acknowledge it when making your decision
- Plan to address missing data in the future

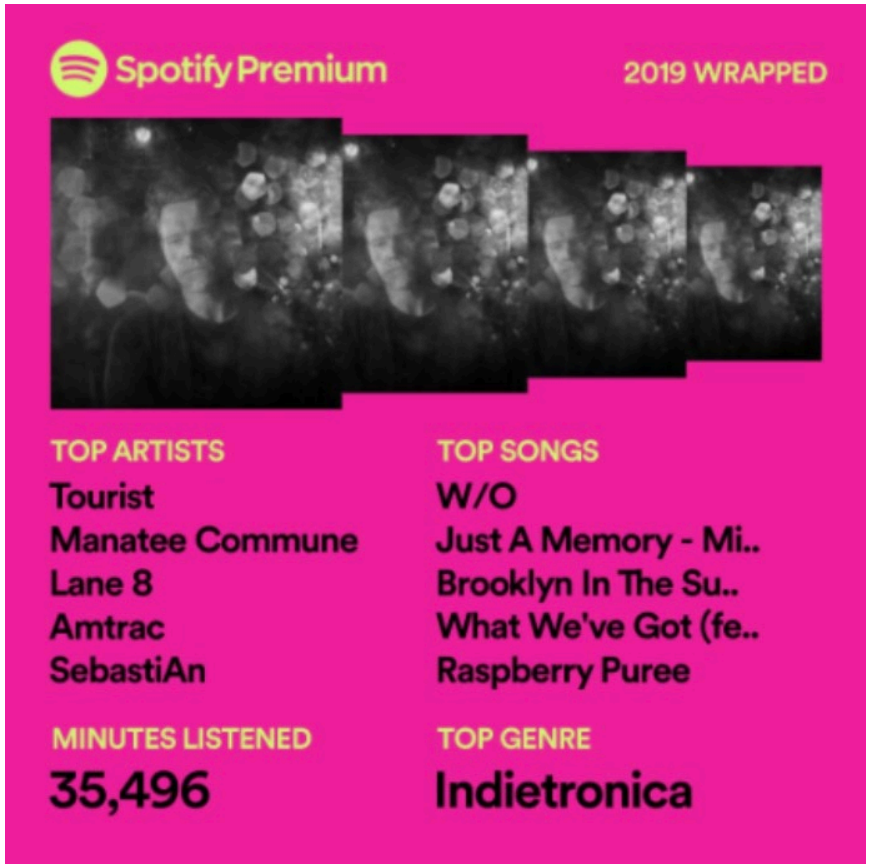
.conf20

splunk>

Communicate the Data

Spotify Wrapped

Communicating missing data



Spotify Wrapped

Communicating missing data

Why are my 2019 Artist Wrapped stats different than the stats I see in Spotify for Artists?

Your Wrapped stats apply to the time period from January 1st to October 31st 2019.

Your Spotify for Artists dashboard shows stats for a number of different time periods, so it's possible the numbers won't match up exactly.

http://web.archive.org/web/20191227230903if_/https://artists.spotify.com/faq/wrapped-2019#how-do-i-get-my-2019-wrapped

Verizon DBIR

Communicating missing data

Figure 2. What tactics are utilized? (Actions)

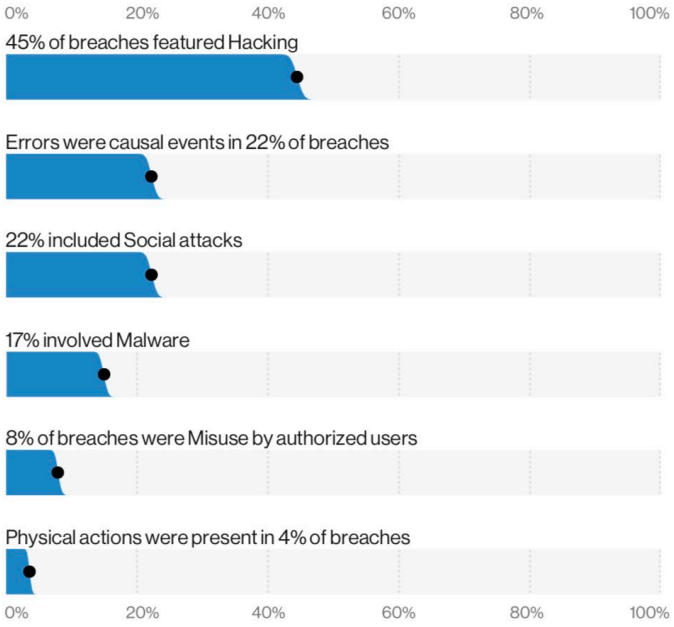
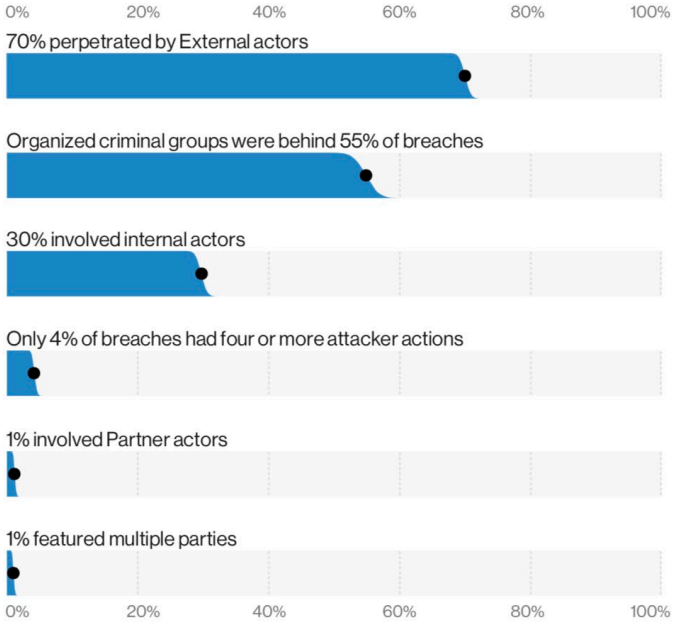


Figure 3. Who's behind the breaches?



https://enterprise.verizon.com/resources/reports/dbir/

Results and analysis

The results found in this and subsequent sections within the report are based on a dataset collected from a variety of sources, including cases provided by the Verizon Threat Research Advisory Center (VTRAC) investigators, cases provided by our external collaborators and publicly disclosed security incidents. The year-to-year data will have new incident and breach sources as we continue to strive to locate and engage with additional organizations that are willing to share information to improve the diversity and coverage of real-world events. This is a sample of convenience,[6] and changes in contributors—both additions and those who were not able to contribute this year—will influence the dataset. Moreover, potential changes in contributors' areas of focus can shift bias in the sample over time. Still other potential factors, such as how we filter and subset the data, can affect these results. All of this means that we are not always researching and analyzing the same population. However, they are all taken into consideration and acknowledged where necessary within the text to provide appropriate context to the reader. Having said that, the consistency and clarity we see in our data year-to-year gives us confidence that while the details may change, the major trends are sound.

Now that we have covered the relevant caveats, we can begin to examine some of the main trends you will see while reading through this report. When looking at Figure 6 below, let's focus for a moment on the Trojan[7] line. When many people think of how hacking attacks play out, they may well envision the attacker dropping a Trojan on a system and then utilizing it as a

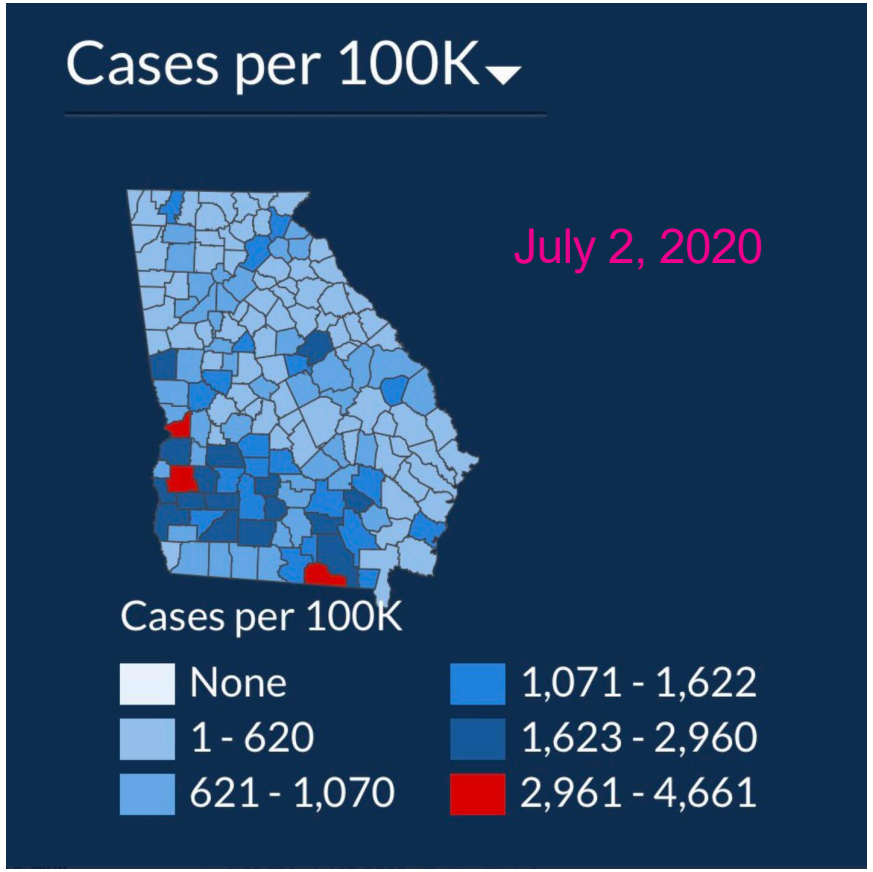
.conf20

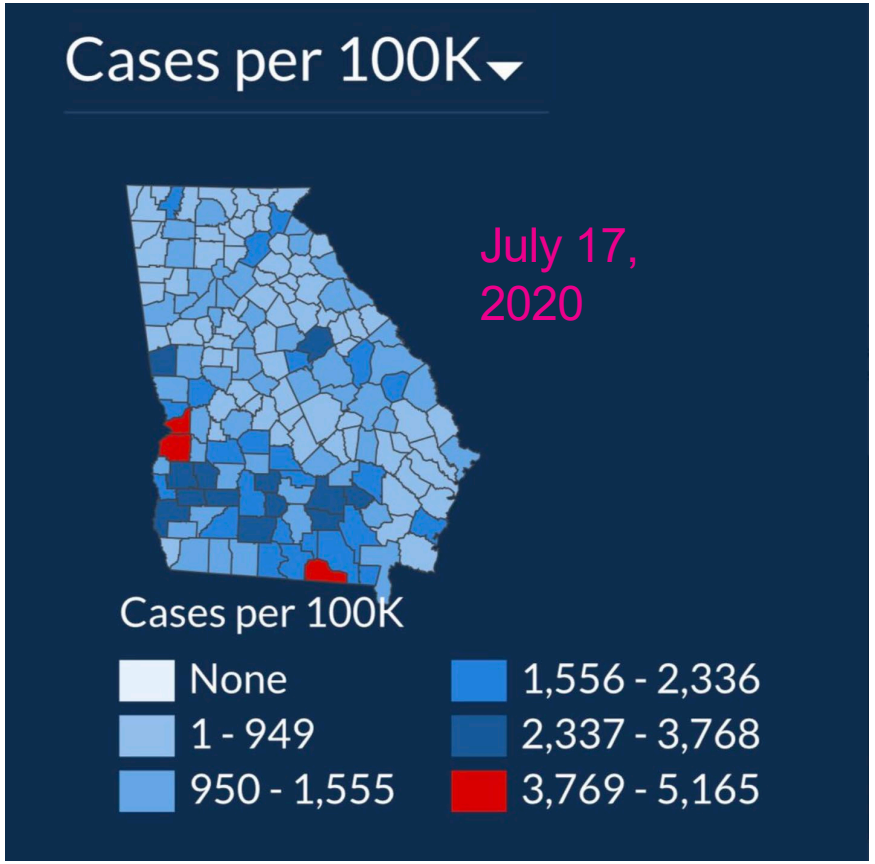
splunk>

Visualize the Data

Georgia Coronavirus Case Rates

Visualizing missing data

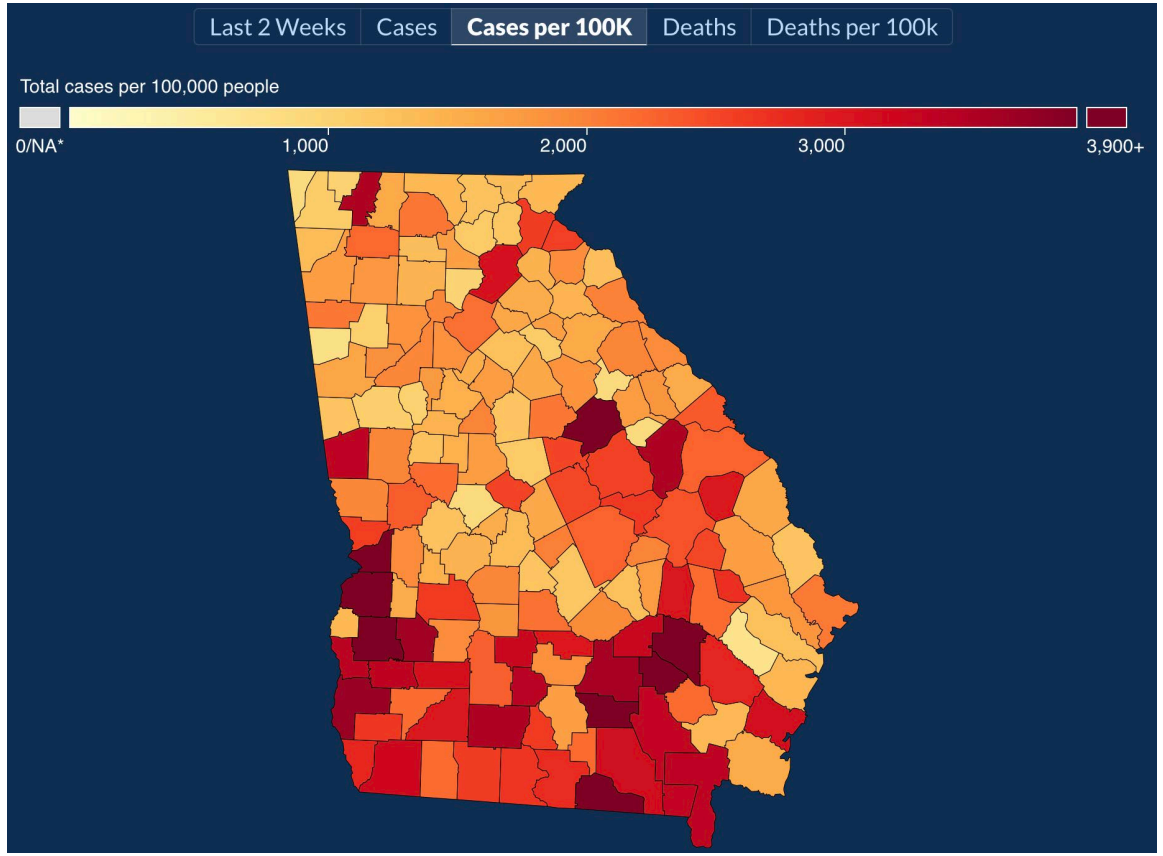




https://twitter.com/andishehnouraee/status/1284237474831761408

Georgia Coronavirus Case Rates

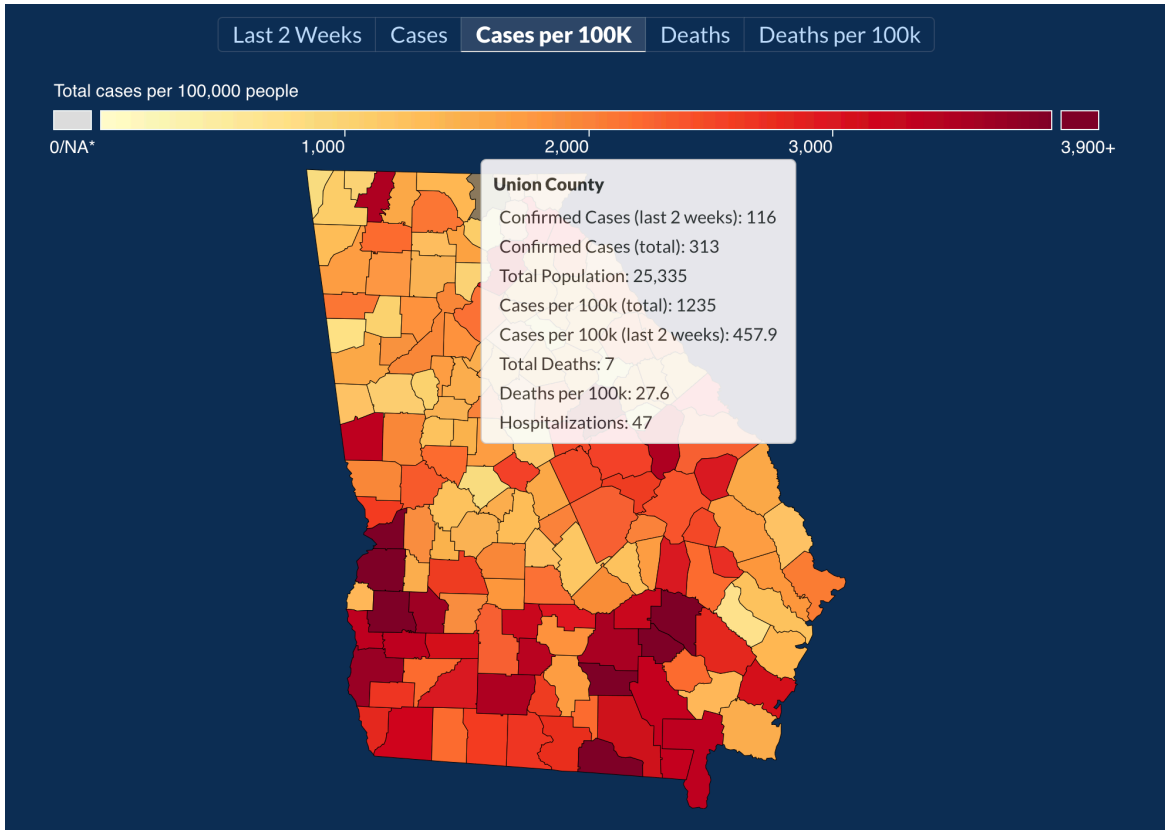
Visualizing missing data



https://dph.georgia.gov/covid-19-daily-status-report, screenshot taken August 13, 2020

Georgia Coronavirus Case Rates

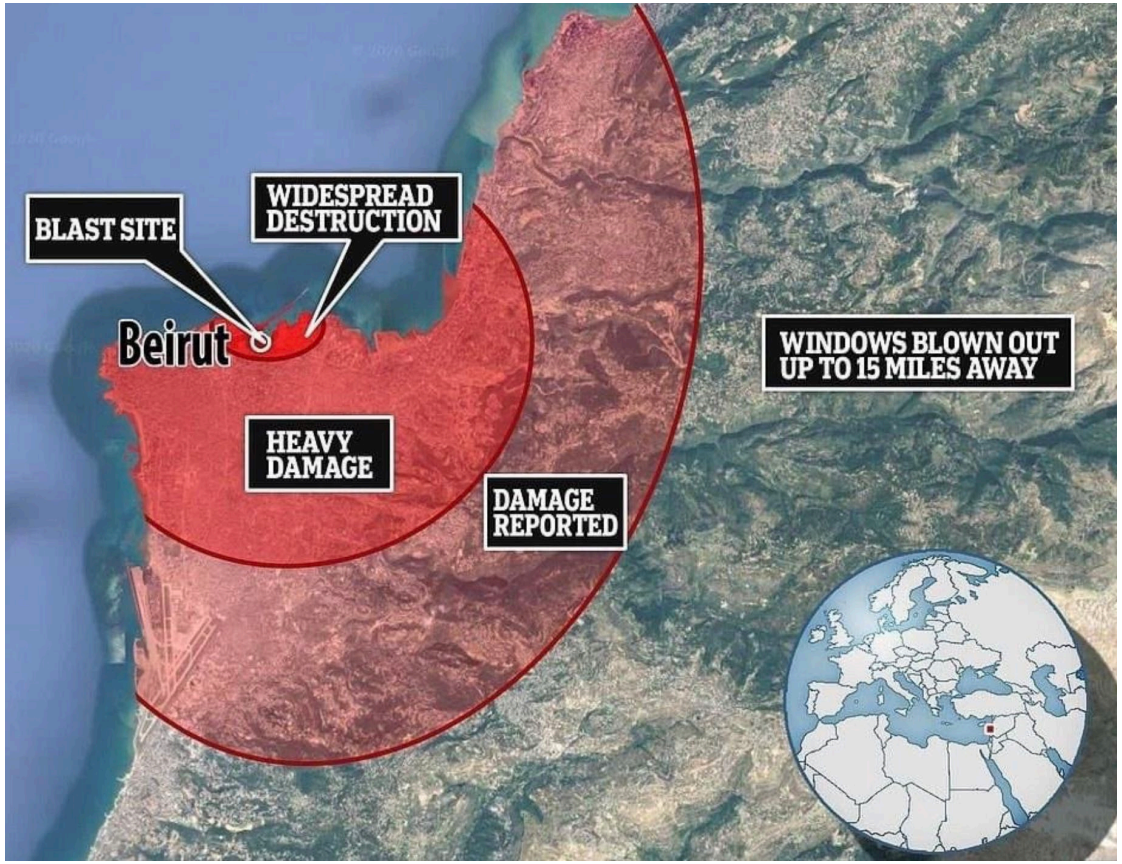
Visualizing missing data



https://dph.georgia.gov/covid-19-daily-status-report, screenshot taken August 13, 2020

Beirut Explosion Maps

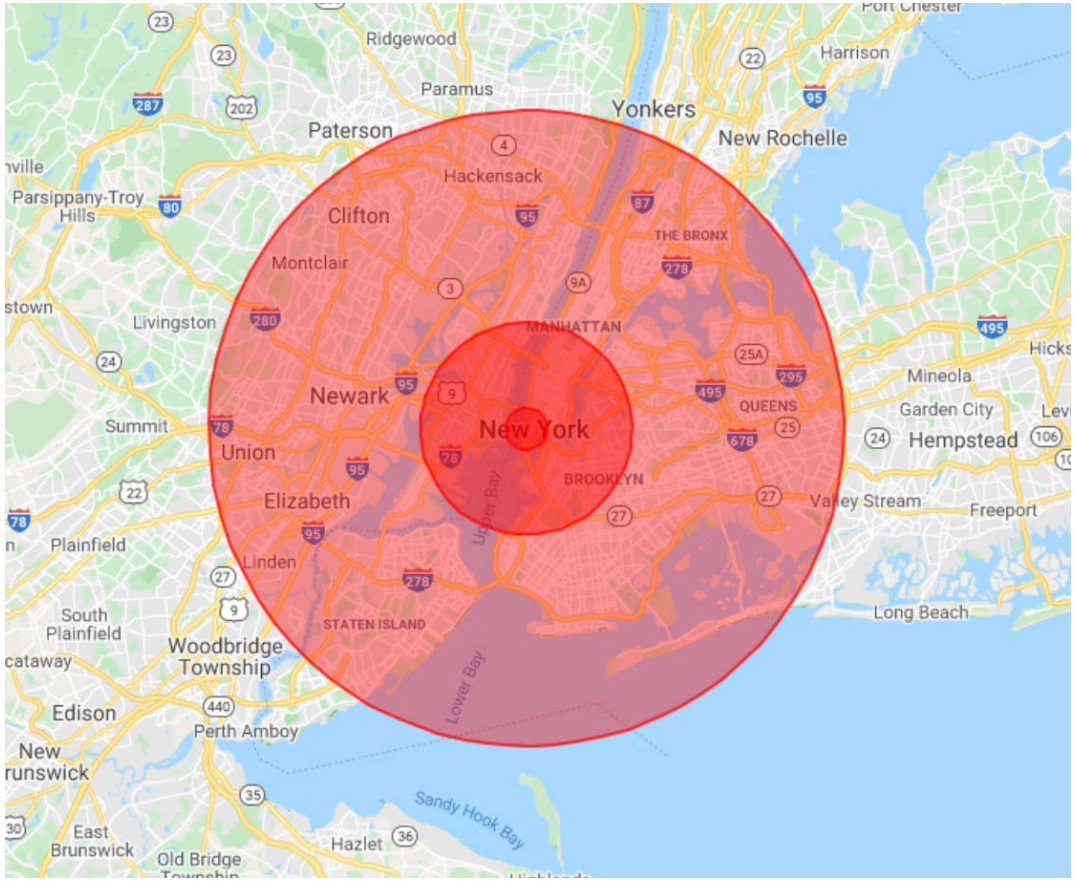
Visualizing missing data



https://twitter.com/JoannaMerson/status/1291095463119056896?s=20

Beirut Explosion Maps

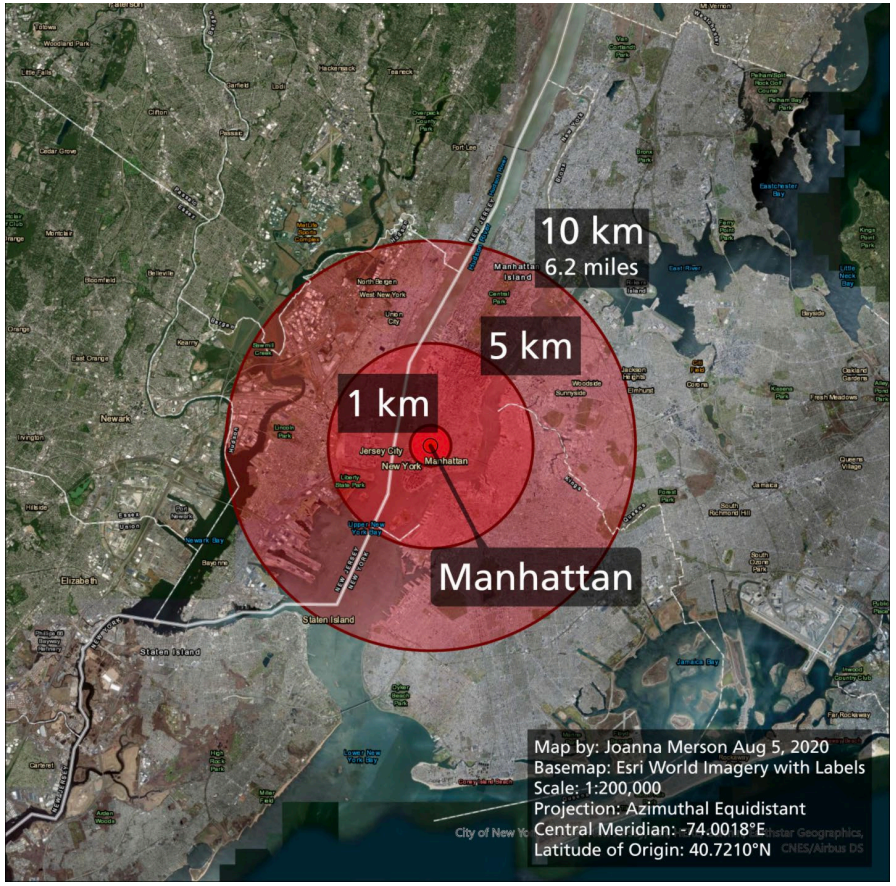
Visualizing missing data



https://twitter.com/JoannaMerson/status/1291095463119056896?s=20

Beirut Explosion Maps

Visualizing missing data



https://twitter.com/JoannaMerson/status/1291095463119056896?s=20

Analyze the Data

Declining Bird Populations

Analyzing missing data

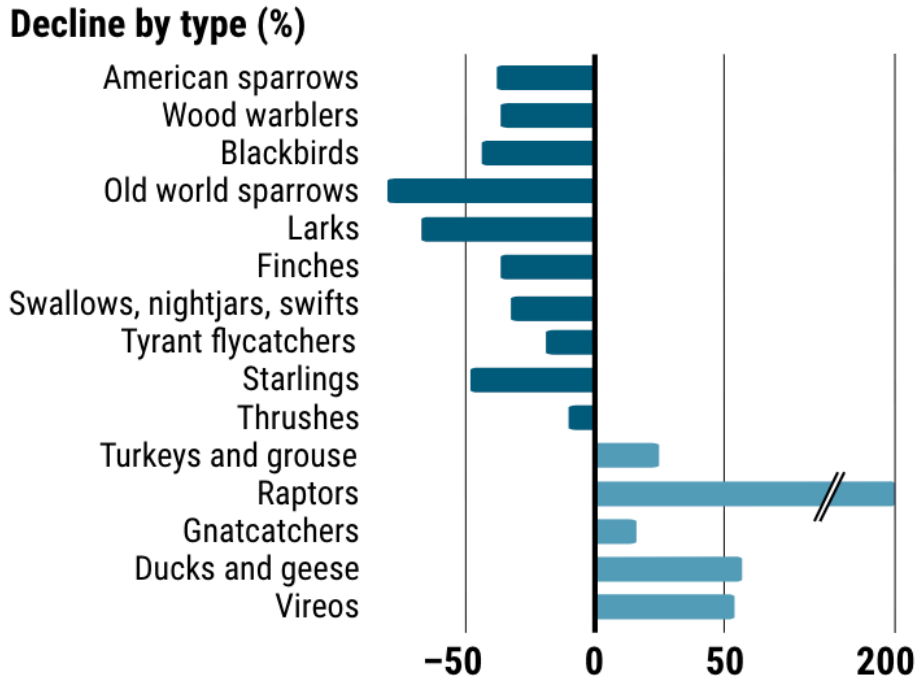
Three billion North American birds have vanished since 1970, surveys show

By Elizabeth Pennisi | Sep. 19, 2019 , 2:00 PM

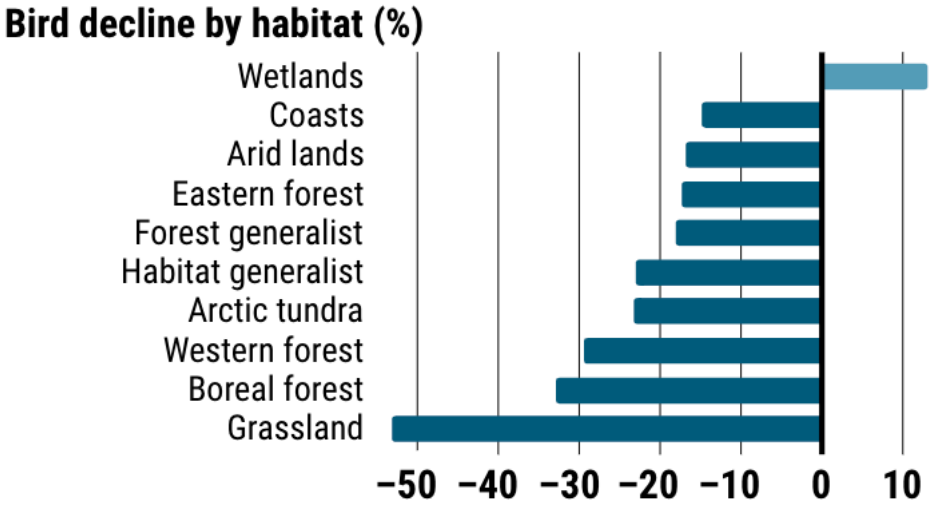
https://www.sciencemag.org/news/2019/09/three-billion-north-american-birds-have-vanished-1970-surveys-show

Declining Bird Populations

Analyzing missing data



K. ROSENBERG *ET AL.*, *SCIENCE*, ADAPTED BY N. DESAI/*SCIENCE*



K. ROSENBERG *ET AL.*, *SCIENCE*, ADAPTED BY N. DESAI/*SCIENCE*

https://www.sciencemag.org/news/2019/09/three-billion-north-american-birds-have-vanished-1970-surveys-show

Song FRatings

Analyzing missing data

splunk> .conf20

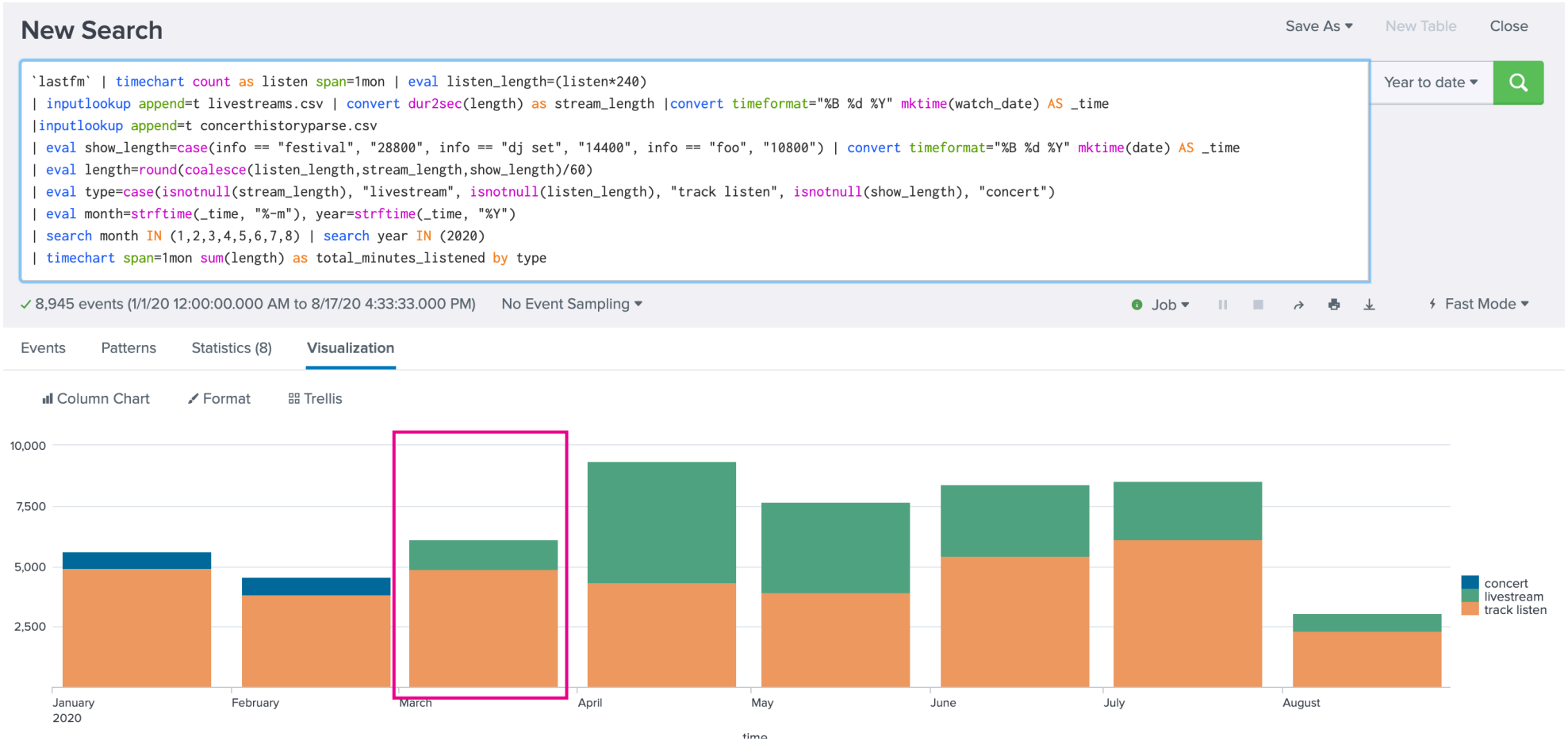
Music Event Duration

Analyzing missing data

```
|inputlookup append=t concerthistoryparse.csv
| eval show_length=case(info == "festival", "28800", info == "dj set", "14400", info == "concert", "10800")
```

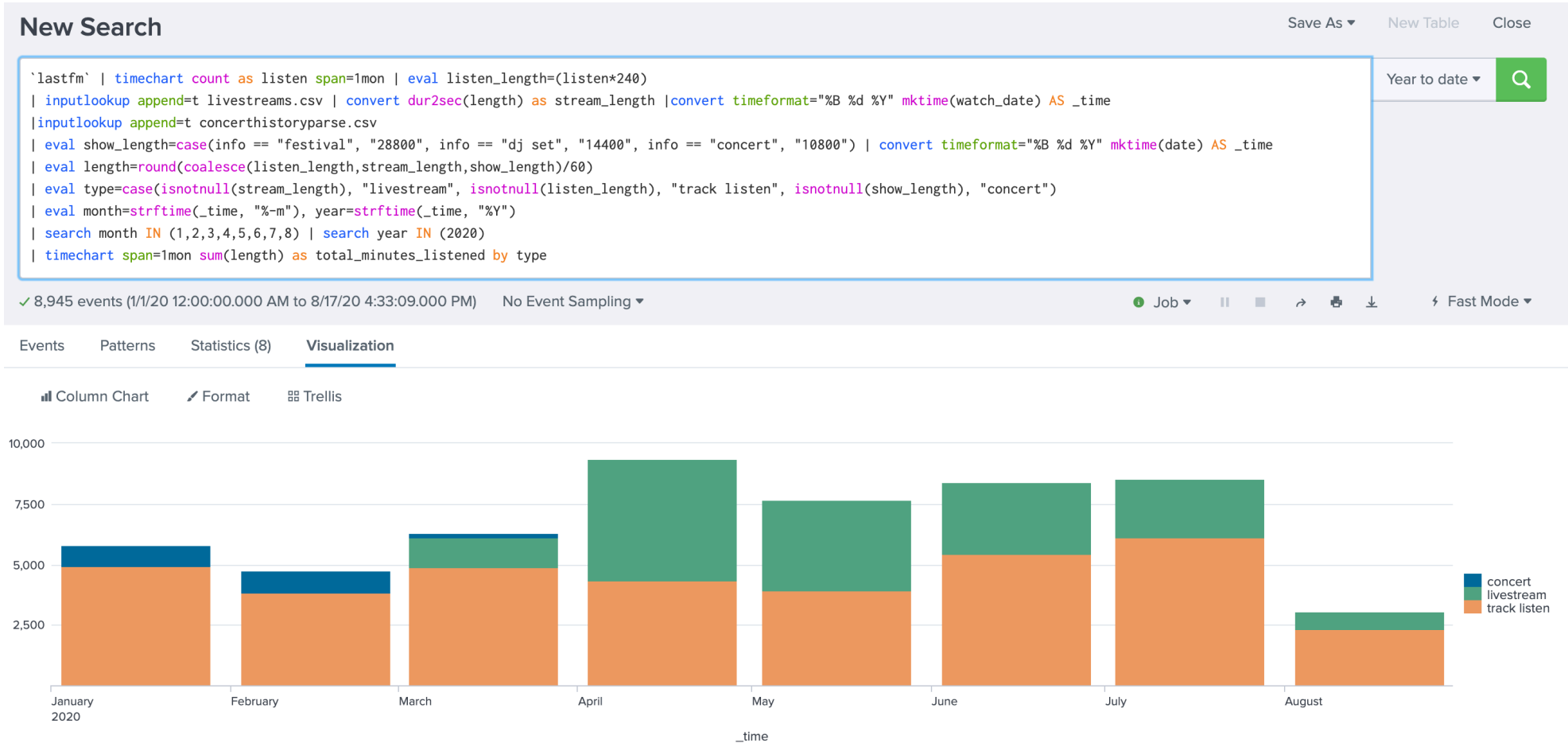
Music Event Duration

Analyzing missing data



Music Event Duration

Analyzing missing data



Manage the Data

Extract, transform, load, monitor, steward

Missing Website Metrics

Managing missing data

Events | Patterns | **Statistics (8)** | Visualization

20 Per Page ▾ | Format | Preview ▾

_time	count
2020-07-15	720328
2020-07-16	64714
2020-07-17	0
2020-07-18	0
2020-07-19	0
2020-07-20	0
2020-07-21	0
2020-07-22	0

Microsoft Excel and Genes

Managing missing data

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

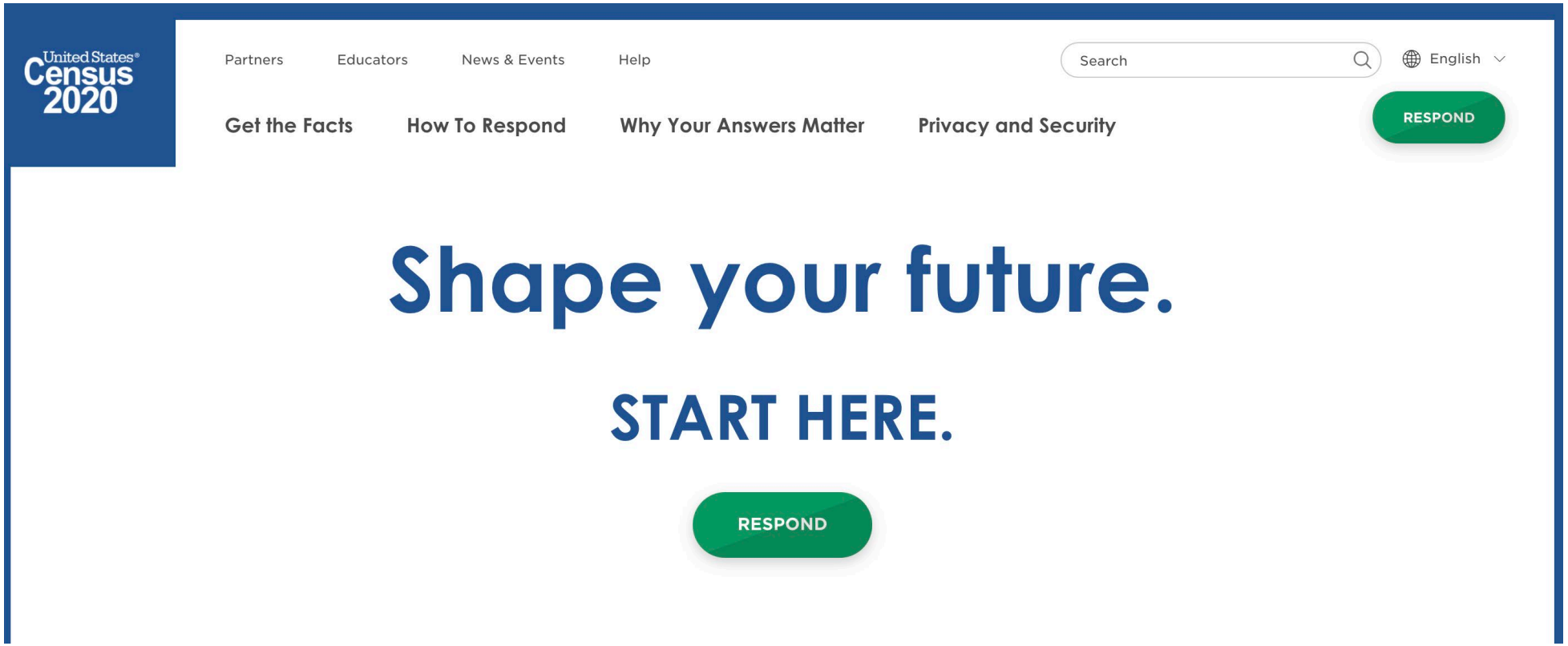
SHARE

https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

Collect the Data

U.S. Census 2020

Collecting missing data



https://2020census.gov/en.html

Illinois Coronavirus Data

Collecting missing data

Environment & Public Health

Illinois Has Holes In Its COVID-19 Data. Will That Hinder Planning For Future Outbreaks?

Data kept by the state show that in 80% of COVID-19 cases, the patient's job is unknown — vital to preventing potential future outbreaks.

By Kristen Schorsch

May 28, 6 a.m. CT

https://www.wbez.org/stories/illinois-has-holes-in-its-covid-19-data-will-that-hinder-planning-for-future-outbreaks/20bfea8f-c140-404d-bf2e-4c8795e2ce8c

Define the Question

Data analysis starts with a question

Define the Question

- Decide what questions you want to answer
- Use questions to guide your process
- Evaluate your questions for bias

Digital Film Archive

Missing data when answering a question

https://twitter.com/footage/status/1265723417132556288

Data can go missing at any stage of the data analysis process

Decision | Communication | Visualization | Analysis | Management | Collection | Question

Take action to reduce bias from missing data

Deciding with the data

1. Define the questions being answered with data
2. Identify missing data
3. Ask questions of the data analysis before making decisions

Take Action To Reduce Bias From Missing Data

Working with the data

1. Steward and normalize data
2. Analyze data at multiple levels of aggregation and time spans
3. Add context to reports and communicate missing data

Thank You

@smorewithface on Twitter
@smoir on splunk-usergroups.slack.com
smoir@splunk.com

Please provide feedback via the

SESSION SURVEY

ⓘ Helpful Resources

Books, articles, Splunk docs,
apps, and Twitter threads

What happens when data is missing?

- Read: https://eyeondesign.aiga.org/finding-the-blank-spots-in-big-data/
- Review: https://github.com/MimiOnuoha/missing-datasets
- Read: https://datasociety.net/library/data-voids/
- Read: https://www.aljazeera.com/indepth/opinion/data-collection-solution-europe-racism-problem-200728131435298.html
- Read: Invisible Women: Data Bias in a World Designed for Men by Caroline Criado Perez (but know also that trans women and nonbinary folx are missing)

Missing Data At The Management Stage

- Read: https://www.duanewaddle.com/proving-a-negative/
- Read: https://www.splunk.com/en_us/blog/tips-and-tricks/sourcetypes-whats-in-name.html
- Read: https://docs.splunk.com/Documentation/Splunk/latest/Data/Usepersistentqueues
- Read: https://community.splunk.com/t5/Splunk-Search/How-to-find-the-retention-period-of-an-index/td-p/299191#M90176
- Install: https://splunkbase.splunk.com/app/4621/
- Install: https://splunkbase.splunk.com/app/2949/

Missing Data At The Visualization Stage

- Read: https://flowingdata.com/2018/01/30/visualizing-incomplete-and-missing-data/
- Read: https://blog.datawrapper.de/colorblindness-part1/ and parts 2 and 3

Make Decisions With Data

- Are you using the right data?
- Read: https://medium.com/@gibsonbiddle/4-proxy-metrics-a82dd30ca810
- Are you starting with a question and using data consistently?
- Read: https://amplitude.com/blog/stop-data-snacking