

Public Data Exploration with Splunk

Phil Meyerson
Developer | EFSI



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.

Splunk, Splunk>, Turn Data Into Doing, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2019 Splunk Inc. All rights reserved.



Public Data Exploration with Splunk

Phil Meyerson
Developer | EFSI

Agenda

What this talk is

Where can we find data?

Data Ingestion, Normalization, and Quality

Let's go to the whiteboard

Tips Tricks Caveats

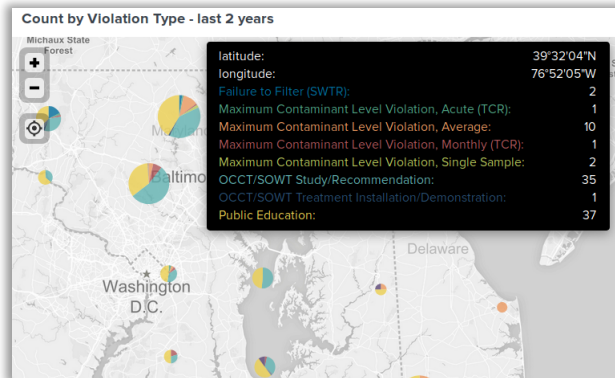
References

Q&A

What This Talk Is

Public data exploration

Water Quality Reports



Election Participation

PRECINCT REPORT Presidential General Election OFFICIAL RESULTS

Run Date: 12/12/16 Montgomery County, MD November 8, 2016

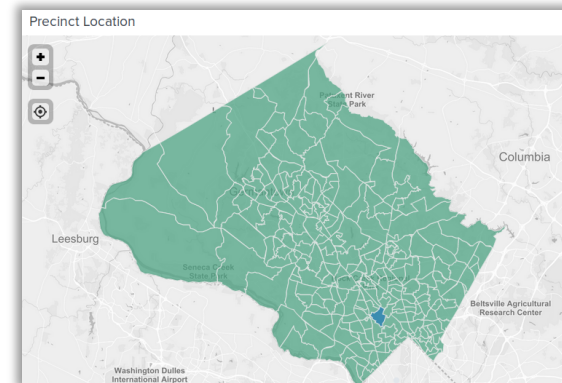
RUN TIME: 11:47 AM

0101 001-001	VOTES	PERCENT
REGISTERED VOTERS - TOTAL	2163	
BALLOTS CAST - TOTAL	1037	
BALLOTS CAST - BLANK	49	4.73
VOTER TURNOUT - TOTAL	47.94	
VOTER TURNOUT - BLANK	2.27	

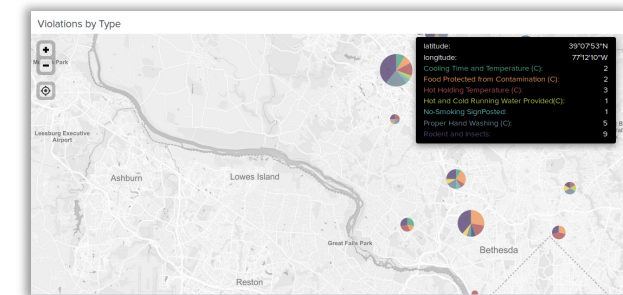
President - Vice Pres (VOTE FOR) 1		
Trump-Pence (REP)	550	53.40
Clinton-Kaine (DEM)	404	39.22
Johnson-Weld (LTB)	29	2.82
Stein-Baraka (GRN)	11	1.07
WRITE-IN	36	3.50
Total	1030	

U.S. Senator (VOTE FOR) 1		
Kathy Szeliga (REP)	572	57.03
Chris Van Hollen (DEM)	393	39.10
Hargett Jones (GRN)	34	3.39
WRITE-IN	4	.40
Total	1003	

Rep in Congress Congressional District 8 (VOTE FOR) 1		
Dan Cox (REP)	600	60.48
Janie Raskin (DEM)	332	33.47



Food Inspections



What This Talk Is

Tool exploration

TA-data-scrape

- My custom data scraper and ingest tool

SA-Socrata

Splunk Add-On Builder



Finding Data and Ingest

Many Sources

Sometimes In unexpected places

Federal, state, local data portals

Infrastructure/utility operators

Local government and NGO websites

Scientific or research institutions

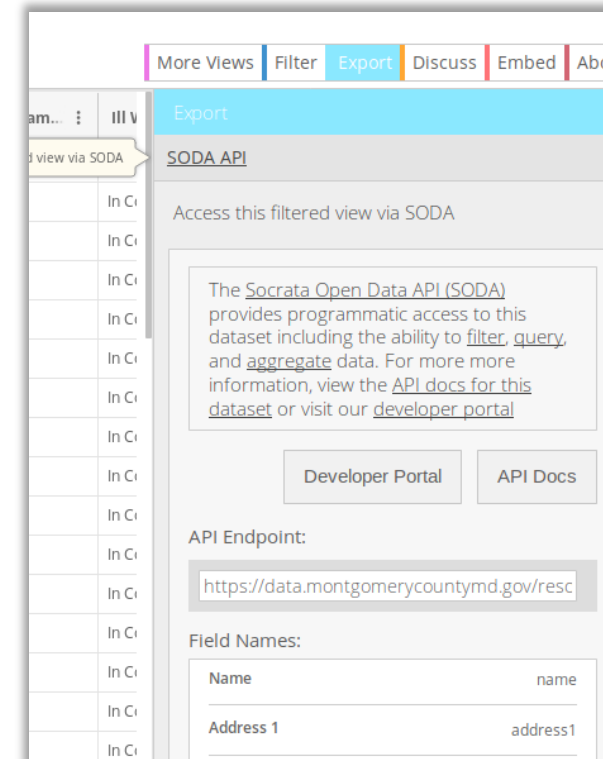
Data collated by a 3rd party



Getting Structured Data

Structured Data Capabilities

- SODA API: SA-Socrata
- REST API: Splunk Add-on-Builder
- CSV: Lookup Import, Lookup Editor App, or “Getting Data In” Local Ingest



Getting Unstructured Data

Board of Elections
Montgomery County, Maryland

2016 General Election Results by Polling Place

01-01	01-02	01-03	01-04	01-05	01-06	01-07
02-01	02-02	02-03	02-04	02-05	02-06	02-07
02-08	02-09	02-10	02-11			
03-01	03-02					
04-01	04-02	04-03	04-04	04-05	04-06	04-07
04-08	04-09	04-10	04-11	04-12	04-13	04-14
04-15	04-16	04-17	04-18	04-19	04-20	04-21
04-23	04-24	04-25	04-26	04-27	04-28	04-29
04-30	04-31	04-32	04-33	04-34	04-35	04-37
04-38	04-40		04-42	04-43		
05-01	05-02	05-03	05-04	05-05	05-06	05-08
05-09	05-10	05-11	05-12	05-13	05-14	05-15
05-16	05-17	05-18	05-19	05-21	05-22	05-23
06-01	06-02	06-03	06-04	06-05	06-06	06-07
06-08	06-09	06-10	06-11	06-13	06-14	06-15
07-01	07-02	07-03	07-04	07-05	07-06	07-07

PRECINCT REPORT Presidential General Election OFFICIAL RESULTS
Run Date:12/12/16 Montgomery County, MD
RUN TIME:11:47 AM November 8, 2016

0101 001-001

	VOTES	PERCENT
REGISTERED VOTERS - TOTAL	2163	
BALLOTS CAST - TOTAL	1037	
BALLOTS CAST - BLANK	49	4.73
VOTER TURNOUT - TOTAL		47.94
VOTER TURNOUT - BLANK		2.27

President - Vice Pres
(VOTE FOR) 1

Trump-Pence (REP)	550	53.40
Clinton-Kaine (DEM)	404	39.22
Johnson-Weld (LIB)	29	2.82
Stein-Baraka (GRN)	11	1.07
WRITE-IN.	36	3.50
Total	1030	

U.S. Senator
(VOTE FOR) 1

Kathy Szeliga (REP)	572	57.03
Chris Van Hollen (DEM)	393	39.18
Margaret Flowers (GRN)	34	3.39
WRITE-IN.	4	.40
Total	1003	

Rep in Congress Congressional District 8
(VOTE FOR) 1

Dan Cox (REP)	600	60.48
Jamie Raskin (DEM)	332	33.47

Python Event Scraper

Python web scraper for my local board of elections data

- Traverse a tree of web links
- Download, parse minimally, and ingest via HEC
- Remainder of parsing at search time





Demo



Data Quality & Normalization

Field Drift

The more things change...

Examine data closely after ingest

Correct changes in value or field formatting across data sets

| fieldsummary is your friend

CANDIDATE	Candidate
CONTEST_ID	ContestName
CONTEST_NAME	ContestNumber
ELECTION_DATE	ElectionDate
ELECTION_NAME	ElectionName
PARTY	Party
PRECINCT_NUMBER	PrecinctNumber
VOTES	Votes
WARD	WardNumber



Let's See What We've Got

Voter Participation

Computer... Enhance

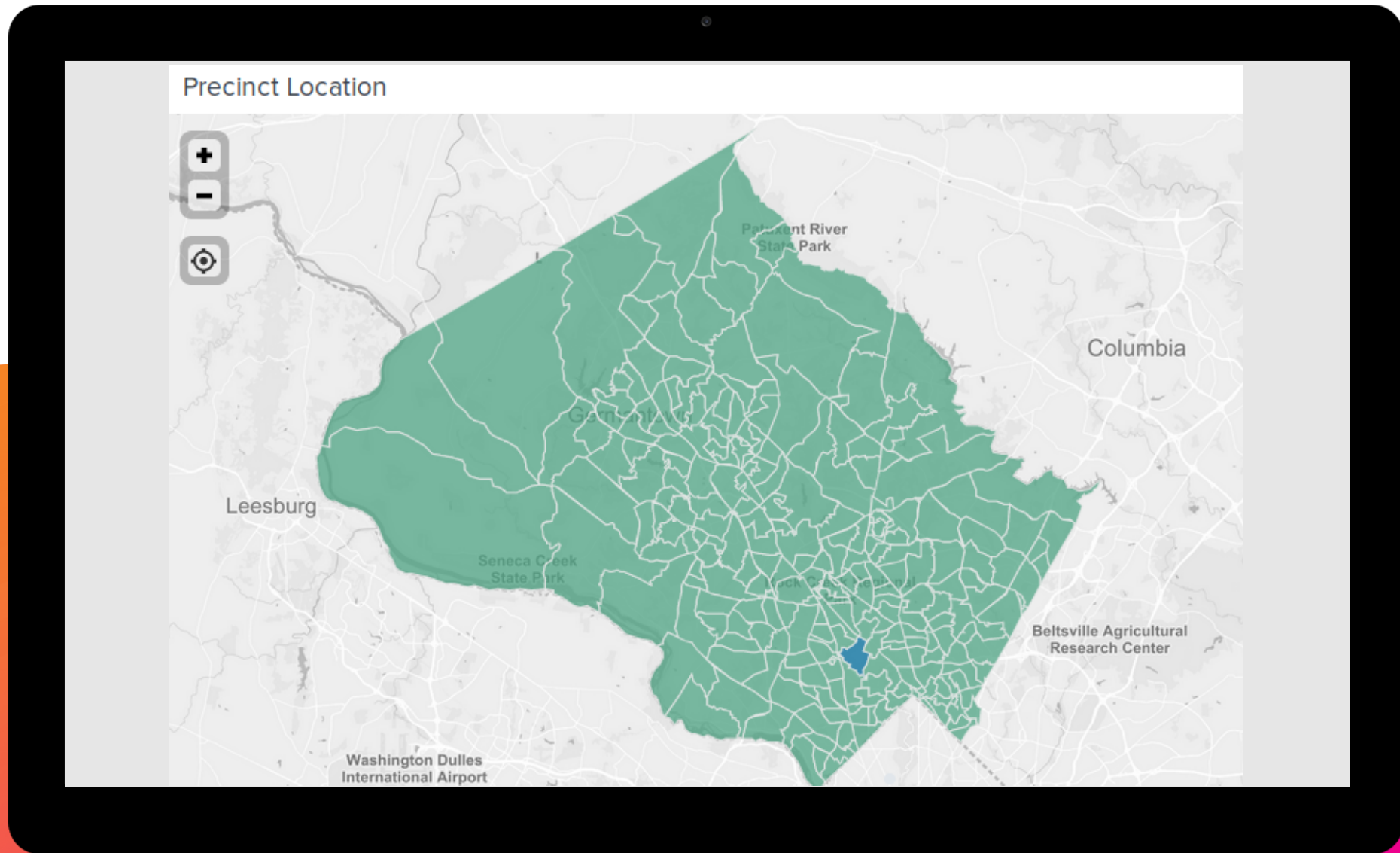
The screenshot shows a configuration window with a sidebar on the left containing three tabs: 'General', 'Markers', and 'Tiles'. The 'Tiles' tab is selected and highlighted with a blue border. The main content area of the 'Tiles' tab contains the following settings:

- Show Tiles:** A toggle switch set to 'Yes'.
- Tile Opacity:** A text input field containing '100' followed by a '%' symbol.
- URL:** A text input field containing the URL 'http://{s}.tile.openstreetmap.org/{z}/{x}/{y}.png'. Below the field is a small text note: 'The URL to use for requesting tiles, ex: http://{s}.tile.openstreetmap.org/{z}/{x}/{y}.png'.
- Min Zoom:** A text input field containing '0'.
- Max Zoom:** A text input field containing '12'.
- Populate from preset configuration:** A dropdown menu currently showing a downward arrow.

The screenshot shows a configuration window with a sidebar on the left containing three tabs: 'General', 'Markers', and 'Tiles'. The 'General' tab is selected and highlighted with a blue border. The main content area of the 'General' tab contains the following settings:

- Zoom on Scroll:** A toggle switch set to 'Yes'.
- Latitude:** A text input field containing '38.99' with a location pin icon to its right.
- Longitude:** A text input field containing '-76.91'.
- Zoom:** A text input field containing '12'.

A large green arrow points from the bottom of the 'General' tab area towards the text 'Populate with current map settings'.





Demo

EPA Water Violations in My Area?

EPA.GOV SDWIS Federal Reports Search

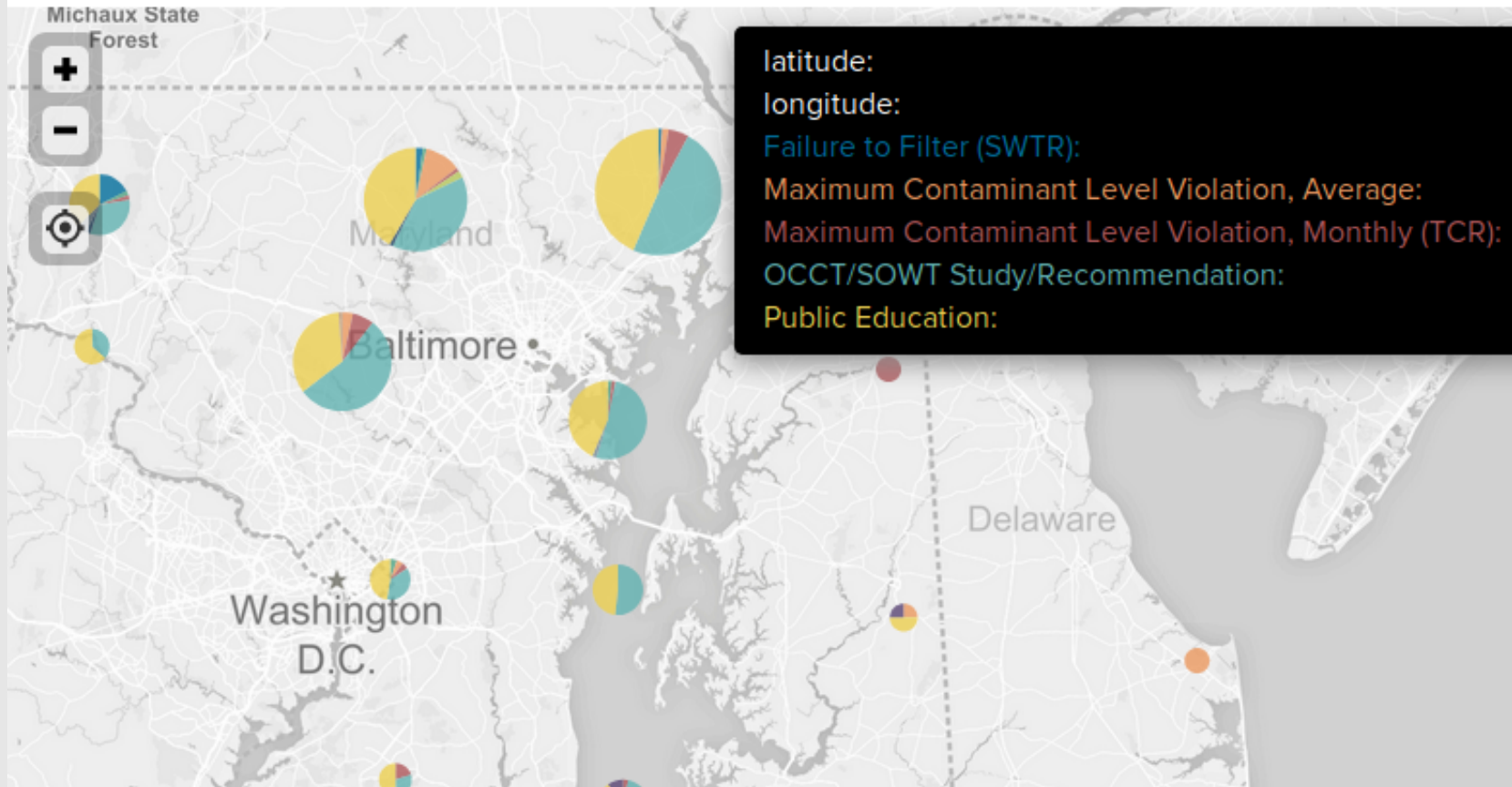
- Is Health Based = 'Y'
- State Code = 'MD'
- Compliance Period End Date < 2 Years ago

EPA Water Violations in My Area?

Detail

PWS Name ⇅	PWS Type ⇅	Customers ⇅	Violation Type ⇅	Contaminant ⇅	Violation Measure ⇅	Status ⇅	Compliance Period Begin ⇅	FirstReported ⇅	Action ⇅	Is Health Based ⇅
TOWN OF UNION BRIDGE	Community water system	997	Failure to Filter (SWTR)	Surface Water Treatment Rule	-	Returned to Compliance	01-FEB-1999	28-AUG-1999	State Compliance achieved	Y
TOWN OF UNION BRIDGE	Community water system	997	Maximum Contaminant Level Violation, Average	Nitrate	11.1	Returned to Compliance	15-MAY-1994	25-AUG-1994	State Compliance achieved	Y
TOWN OF UNION BRIDGE	Community water system	997	Maximum Contaminant Level Violation, Average	Nitrate	10.5	Returned to Compliance	06-DEC-1993	23-MAY-1994	State Compliance achieved	Y

Count by Violation Type - last 2 years





Demo

Local Food Safety

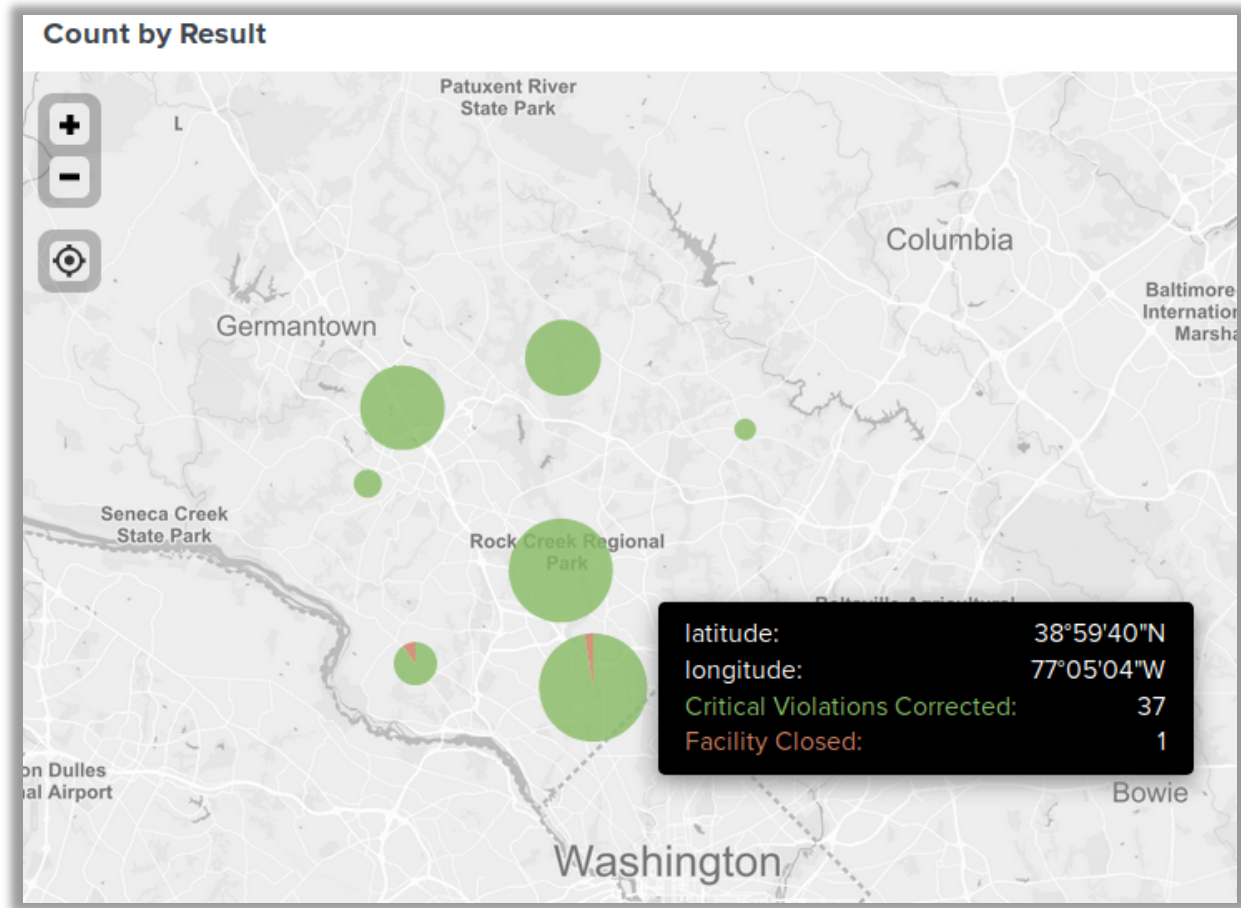
Inspection Data published using Socrata Open Data API (SODA)

Use SA-Socrata to pull data

Local Food Safety

Name	Address 1	Address 2	City	Category	InspectionResult	Inspection Type	raw_field	Cooking Time and Temperature (C)	Cooling Time and Temperature (C)	Establishment ID	Food Protected from Contamination (C)	Holding Temperature (C)
ABYSSINIA RESTAURANT	8221 GEORGIA AVE.	NULL	SILVER SPRING	Restaurant	Critical Violations Corrected	Monitoring	violation6a	In Compliance	In Compliance	20691	In Compliance	Out of Compliance
6-TWELVE CONVENIENT MART	14100 C DARNESTOWN RD.	NULL	GERMANTOWN	Market	Critical Violations Corrected	Monitoring	violation6a	In Compliance	In Compliance	37048	In Compliance	Out of Compliance
KENNY'S SUB SHOP	11210 GEORGIA AVE.	NULL	WHEATON	Restaurant	Critical Violations Corrected	Complaint Invest w/Comprehensive	violation6a	In Compliance	In Compliance	17453	In Compliance	Out of Compliance
ANTOJITOS RESTAURANT	12 E EAST DIAMOND AVE.	NULL	GAITHERSBURG	Restaurant	Critical Violations Corrected	Change of Ownership - Environmental	violation22 violation6a	In Compliance	In Compliance	37186	In Compliance	Out of Compliance

Local Food Safety



Tips and Caveats

Limits.conf – max_count default 500,000
Subsearches max 10,000

Engaged with experts

For data across multiple regions look to National sources

Data normalization is a key step!

References

Limits.conf – max_count default 500,000
Subsearches max 10,000

Engaged with experts

For data across multiple regions look to National sources

Data normalization is a key step!

Tips and Caveats

<https://www.splunk.com/blog/2017/06/01/enhancing-splunk-visualizations-with-mapbox.html>

<https://www.splunk.com/blog/2013/07/31/zip-code-mapping.html>

<https://sideviewapps.com/2871/custom-maps/>

<https://github.com/hire-vladimir/SA-socrata>

<https://mapbox.com>

<https://dev.socrata.com/docs/endpoints.html>

<https://docs.splunk.com>

Tips and Caveats

EPA SDWIS 2019 Violations Report for MD. Via

https://ofmpub.epa.gov/apex/sfdw/f?p=108:9:0::NO::P9_REPORT:VIO

Montgomery County, MD Board of Elections 2018 Gubernatorial Primary Election Official Results. Via

<https://www.montgomerycountymd.gov/Elections/Results/2018GPPrecinctResults.html>

Numerous sources from <https://opendata.maryland.gov> and <https://data.montgomerycountymd.gov/>



Thank You!

Go to the .conf19 mobile app to

RATE THIS SESSION

