Analyzing **Public Data** with the **Elastic Stack**

Jay Miller

elastic

HackSC 2022

Content Warning

Conversation contains:

- dealings with Police
- gender profiling
- systemic bias

The phrase "perceived" is going to be used.

This is profiling data

We'll address this with the upmost **respect** we're not in any way making statements around gender or LGBTQIA+ belonging

The Plan

Get Public Data

into Elasticsearch

and visualize it with Kibana

Get Public Data

Data Collected by a **Public Entity** (aka Governmental Body)

- Public Schools*
- Cities
- States
- Federal Agencies

We're focusing on **Open Public Data**

- **Accessible**
- **Open Data Portal**
- **Legally Usable**

Data
from a portal
designed for people to read...

This should be easy...**right?**

● ● ●

"***Readability*** *Counts*"

- Zen of Python

What data is this?

```
"2443","CA0371100","SD","10","2018-07-01","00:01:37","30","0","1", \
"Patrol, traffic enforcement, field operations","", \
"700.0","","Grand Avenue","","0","","SAN DIEGO","122","Pacific Beach 122", \
"1","0","0","25","Male","0","1","","No"
```

How about **this**?

```
stop_id, ori, agency, exp_years, date_stop, time_stop, stopduration, \
stop_in_response_to_cfs, officer_assignment_key, assignment, intersection, \
address_block, land_mark, address_street, highway_exit, isschool, school_name, \
address_city, beat, beat_name, pid, isstudent, perceived_limited_english, \
perceived_age, perceived_gender, gender_nonconforming, gend, gend_nc, perceived_lgbt

"2443","CA0371100","SD","10","2018-07-01","00:01:37","30","0","1", \
"Patrol, traffic enforcement, field operations","","700.0","", \
"Grand Avenue","","0","","SAN DIEGO","122","Pacific Beach 122", \
"1","0","0","25","Male","0","1","","No"
```

San Diego Police Stop Racial Identification and Profiling Act Data*

*1 of 11 spreadsheets associated with CA RIPA act of 2015 (AB 953)

Issues

Data Inconsitencies for Boolean values (1/0, true/false, yes/no)

`time_stop` - (obviously incorrect) **date entries not matching** "date_stop"

`intersection, address_block, land_mark, address_street, highway_exit` - **inconsistent**

More Issues

`perceived_gender` vs `gend` - Not CA gender assignment (Male, Female, Non-Binary)

`gender_nonconforming` vs `gend_nc` - does this include non-binary and trans identities

`perceived_lgbt` - could be more inclusive with language

`isstudent` - interactions not always reported[1]

[1] Reporting requirements regarding *students* only apply to interactions between officers and students that take place in a K-12 Public School

Untrustable Data

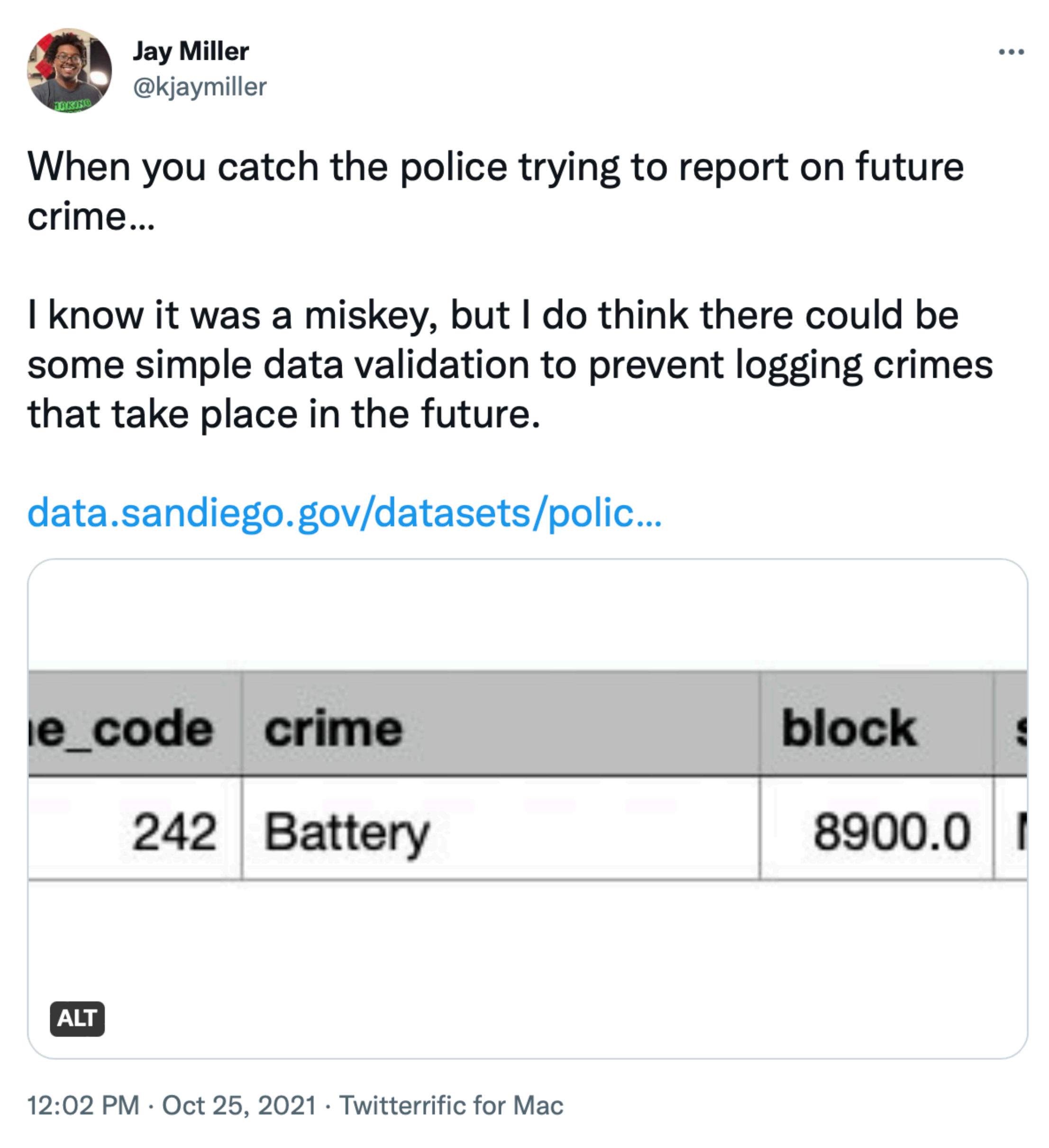
Leads to:

- incorrect **assumptions**
- positive action being **delayed**

data **doesn't get used**
& risks **losing funding**

What Can **WE** Do About it?

Bring **Awareness** to Problems



Jay Miller
@kjaymiller

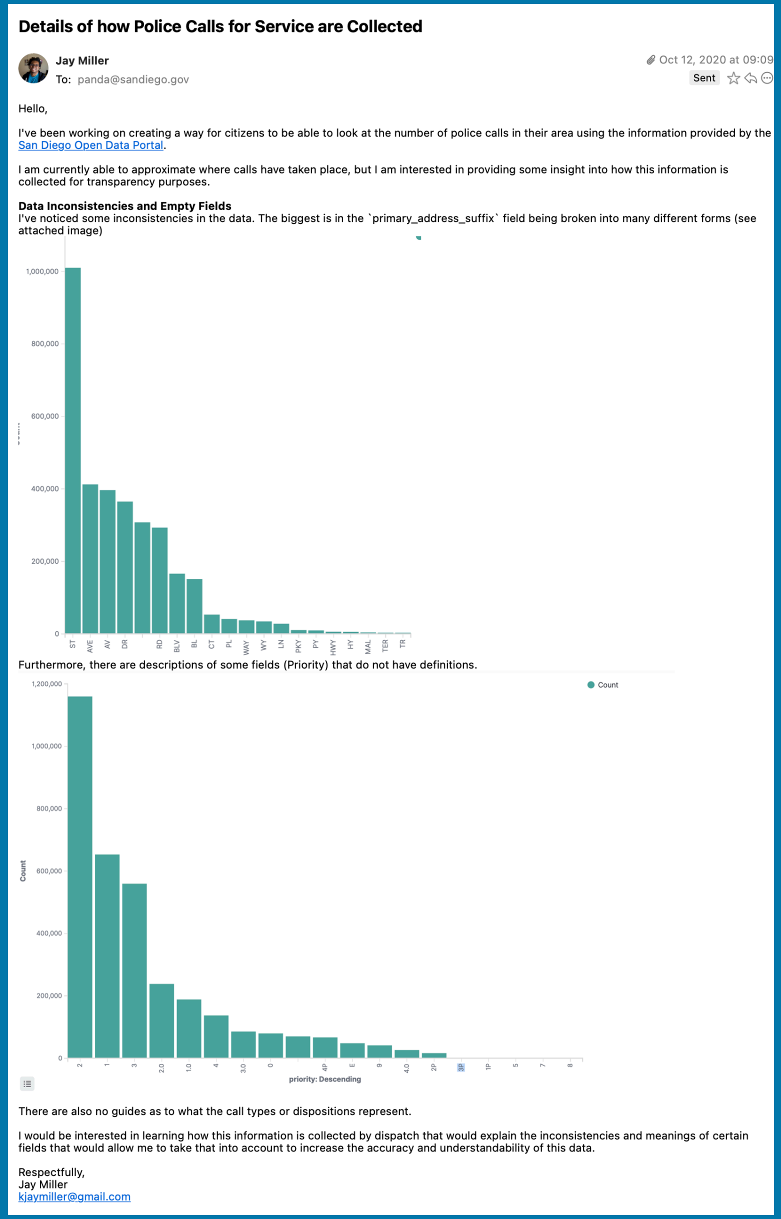
When you catch the police trying to report on future crime...

I know it was a miskey, but I do think there could be some simple data validation to prevent logging crimes that take place in the future.

data.sandiego.gov/datasets/polic...

e_code	crime	block	s
242	Battery	8900.0	I

12:02 PM · Oct 25, 2021 · Twitterrific for Mac



Details of how Police Calls for Service are Collected

Jay Miller
To: panda@sandiego.gov
Oct 12, 2020 at 09:09

Hello,

I've been working on creating a way for citizens to be able to look at the number of police calls in their area using the information provided by the San Diego Open Data Portal.

I am currently able to approximate where calls have taken place, but I am interested in providing some insight into how this information is collected for transparency purposes.

Data Inconsistencies and Empty Fields
I've noticed some inconsistencies in the data. The biggest is in the `primary_address_suffix` field being broken into many different forms (see attached image)

Furthermore, there are descriptions of some fields (Priority) that do not have definitions.

There are also no guides as to what the call types or dispositions represent.

I would be interested in learning how this information is collected by dispatch that would explain the inconsistencies and meanings of certain fields that would allow me to take that into account to increase the accuracy and understandability of this data.

Respectfully,
Jay Miller
kjaymiller@gmail.com

Contribute to Open Source

Sorry...No

Make it **Readable**

- Data **Validation**
- **Remove inconsistencies**
- Make Data **Segmentable**

into Elasticsearch

Elasticsearch

- **Readable** Document-based storage (JSON)
- **Search** Prioritized (You Know, for Search)
- powerful **visualization** (Kibana)
- **Scale Discovery** (APIs)

Prep SD RIPA **Data**

Things we want to do:

- **clean up** keys
- **remove** confusing keys
- assign **mappings**
- **correct formatting**
- create **easier to understand** values
- **upload** data into elasticsearch

Changes from CSV

field changes

("date_stop","time_stop")

date_time_of_stop

"~~stopduration~~"

stop_duration_minutes

"~~stop_in_response_to_cfs~~"

response_to_service_call

("~~officer_assignment_key~~")

officer_assignment

~~percieved_lgbt~~ **percieved_lgbtqia**

Deletions

~~address_block~~

~~address_street~~

~~address_city~~

~~isschool~~

~~beat_name~~

~~intersection~~

~~land_mark~~

~~ori~~

~~agency~~

~~assignment~~

~~isstudent~~

~~gender_nonconforming~~

~~gend~~

~~gend_nc~~

Additions

driver

perceived_transgender

address_description

Now Data looks like **this**

```
POST /sd-ripa/

{
   "stop_id": "478466",
   "exp_years": "10",
   "beat": "243",
   "pid": "1",
   "perceived_age": "63",
   "perceived_gender": "Male",
   "address_description": "The intersection of MIRAMAR and KEARNEY VILLA",
   "stop_datetime": "2021-09-30T11:07:00-07:53",
   "stop_duration_minutes": 6,
   "response_to_service_call": false,
   "officer_assignment": "Patrol, traffic enforcement, field operations",
   "city": "SAN DIEGO",
   "driver": false,
   "perceived_limited_english": false,
   "perceived_transgender": false,
   "perceived_lgbtqia": false
}
```

Mappings

```
PUT /sd-ripa-<agency>
{
    mappings: {
    "stop_id": {"type": "keyword"},
    "pid": {"type": "integer"},
    "exp_years": {"type": "integer"},
    "stop_datetime": {"type": "date"},
    "stop_duration_minutes": {"type": "integer"},
    "officer_assignment": {"type": "keyword"},
    "address_description": {"type": "text"},
    "perceived_age": {"type": "integer"},
    "perceived_gender": {"type": "keyword"},
    "driver": {"type": "boolean"},
    "response_to_service_call": {"type": "boolean"},
    "perceived_lgbtqia": {"type": "boolean"},
    "perceived_transgender": {"type": "boolean"},
    "perceived_limited_english": {"type": "boolean"},
    "beat": {"type": "keyword"},
    "city": {"type": "keyword"},
}
```

NOTE: Mappings are not always required but great for:

- Documentation in Code
- Exact definition of fields

Do you need to **format** your data?

- `mod_csv.sh`
- `parse_address()`
- `parse_gender_lgbtqia()`
- `parse_driver()`

Bringing Data In

- Automatically using our Agents
- Create your own ways
  - (clients in most major (programming) languages)
  - Logstash allows for flexible control the inbond, mutation, and outflow of data
- Bring in (almost) everything keep what you need as long as you need
  - Searchable Snapshots

This isn't a consultation, if you're dealing with others data do it responsibly.

and visualize it with Kibana

Kibana

- A way to interact with data visually
- Accessible (with A11y in mind)
- Extensible (make it what you need it to be)

What's Next?

- EUI
- Runtime fields (create new data based on data)
- Machine Learning

More Information on this data:

- RIPA (AB 953)
- SD RIPA Data - SD Open Data Portal
- CALIFORNIA CODE OF REGULATIONS

Let's Connect

Jay

@kjaymiller -

https://kjaymiller.com/contact

elastic Community

https://elastic.co/community

@elastic

You can find similar projects at https://github.com/stars/kjaymiller/lists/es-public-data-projects