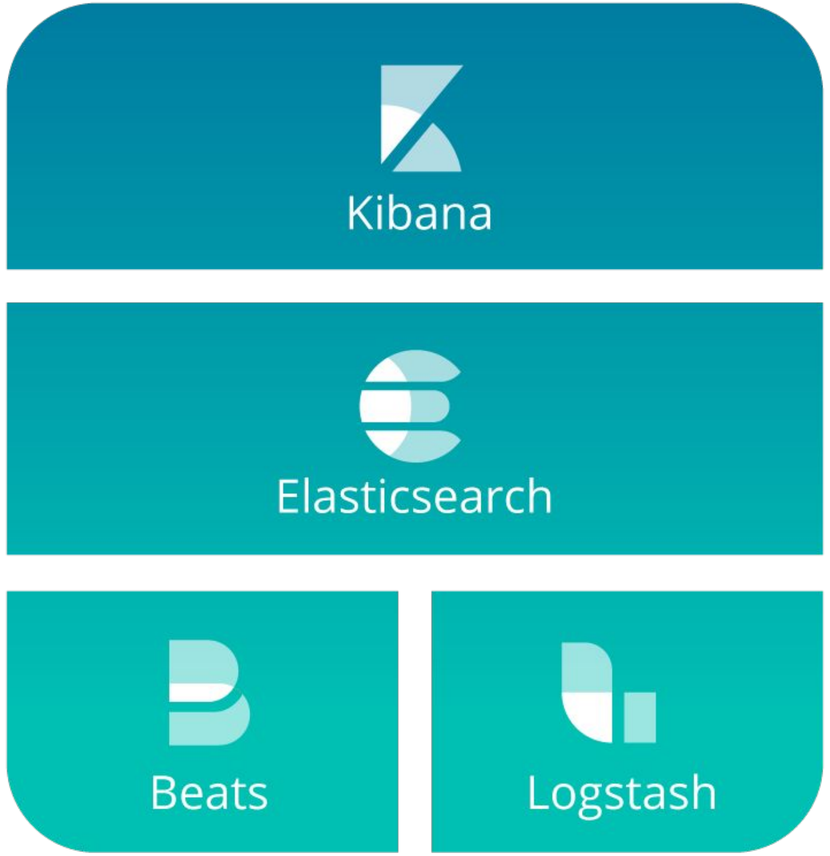
From data to dashboard

Rashmi Kulkarni, Tim Roes, Thomas Neyrick

Elastic

What is the Elastic Stack?

- Store and search data with **Elasticsearch**
- Move data into Elasticsearch with
  - **`FileBeat`**
  - **`Logstash`**
- Visualize data and administer the stack with **Kibana**
- **`In this demo:`**
- **`Elasticsearch – V6.2.0`**
- **`Kibana  - V 6.2.0`**
- **`Filebeat – V 6.2.0`**



What will we do in this presentation?

Full round-trip:
Build application to analyze traffic incident data in NYC

Ingest data into Elasticsearch with Filebeat
Build Kibana application to generate insights

… and enrich analytical experience with machine learning

The data source

NYC traffic incident data
https://opendata.cityofnewyork.us/
+1,000,000 traffic incidents, since July 2012

Tabular format
Fields indicate where and when, number of injuries and fatalities, type of vehicles involved

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	DATE	TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME	OFF STREET NAME	NUMBER OF PERSONS INJURED	NUMBER OF PERSONS KILLED	NUMBER OF PEDESTRIANS INJURED	NUMBER OF PEDESTRIANS KILLED	NUMBER OF CYCLIST INJURED	NUMBER OF CY
2	05/15/2017	0:00	BRONX	10452	40.8382	-73.92371	(40.8382, -73.92371)	WEST 168 STREET	WOODYCREST AVENUE		0	0	0	0	0	
3	05/15/2017	0:00	BRONX	10453	40.85592	-73.9172	(40.85592, -73.9172)	CEDAR AVENUE	WEST 179 STREET		0	0	0	0	0	
4	05/15/2017	0:00	BROOKLYN	11213				UNION STREET	SCHENECTADY AVENUE		0	0	0	0	0	
5	05/15/2017	0:00	BROOKLYN	11225	40.658897	-73.95851	(40.658897, -73.95851)			64 RUTLAND ROAD	0	0	0	0	0	
6	05/15/2017	0:00			40.689537	-73.75813	(40.689537, -73.75813)	119 AVENUE			0	0	0	0	0	
7	05/15/2017	0:00			40.767418	-73.7906	(40.767418, -73.7906)	33 AVENUE	FRANCIS LEWIS BOULEVARD		0	0	0	0	0	
8	05/15/2017	0:00			40.686287	-73.9695	(40.686287, -73.9695)	GREENE AVENUE			0	0	0	0	0	
9	05/15/2017	0:00			40.61137	-74.176315	(40.61137, -74.176315)			1130 SOUTH AVENUE	0	0	0	0	0	

… but Elasticsearch requires JSON documents


```
{
  "date":"05/15/2017",
  "number_of_motorist_injured":1,
  "number_of_cyclist_killed":0,
  "on_street_name":"CHERRY AVENUE",
  "borough":"QUEENS",
  "number_of_persons_killed":0,
  "zip_code":"11355",
  "contributing_factor_vehicle":"Failure to Yield Right-of-Way",
  "number_persons_impacted":1,
  "intersection":"CHERRY AVENUE                    --COLDEN STREET",
  "@version":"1",
  "host":"thomas-XPS-13-9350",
```


- This document must conform to a `mapping`.
  - Field-values have to correspond to a datatype *(date, numbers, text, ...)*
  - Mapping informs how values are indexed at ingest-time *(and this impacts if/how they can be searched for at query-time)*

Ingesting Data via Filebeat (filebeat.yml)

```
filebeat.prospectors:
- type: log
  paths:
    - ./nyc_collision/nyc_collision_data.csv

output.elasticsearch:
  hosts: ["localhost:9200"]
  index: nyc_visionzero
  pipeline: nyc_collision

setup.template.enabled: false
```

Kibana is a window into the Elastic Stack

- *Index Patterns*
  - Points Kibana to one or more indices in Elasticsearch that share the same mappings
  - Manage
    - Time-based values
    - Formatting of values for display
    - Scripted fields for calculating values at query-time (<> *filebeat transformation at ingest-time*)

Kibana Visualizations

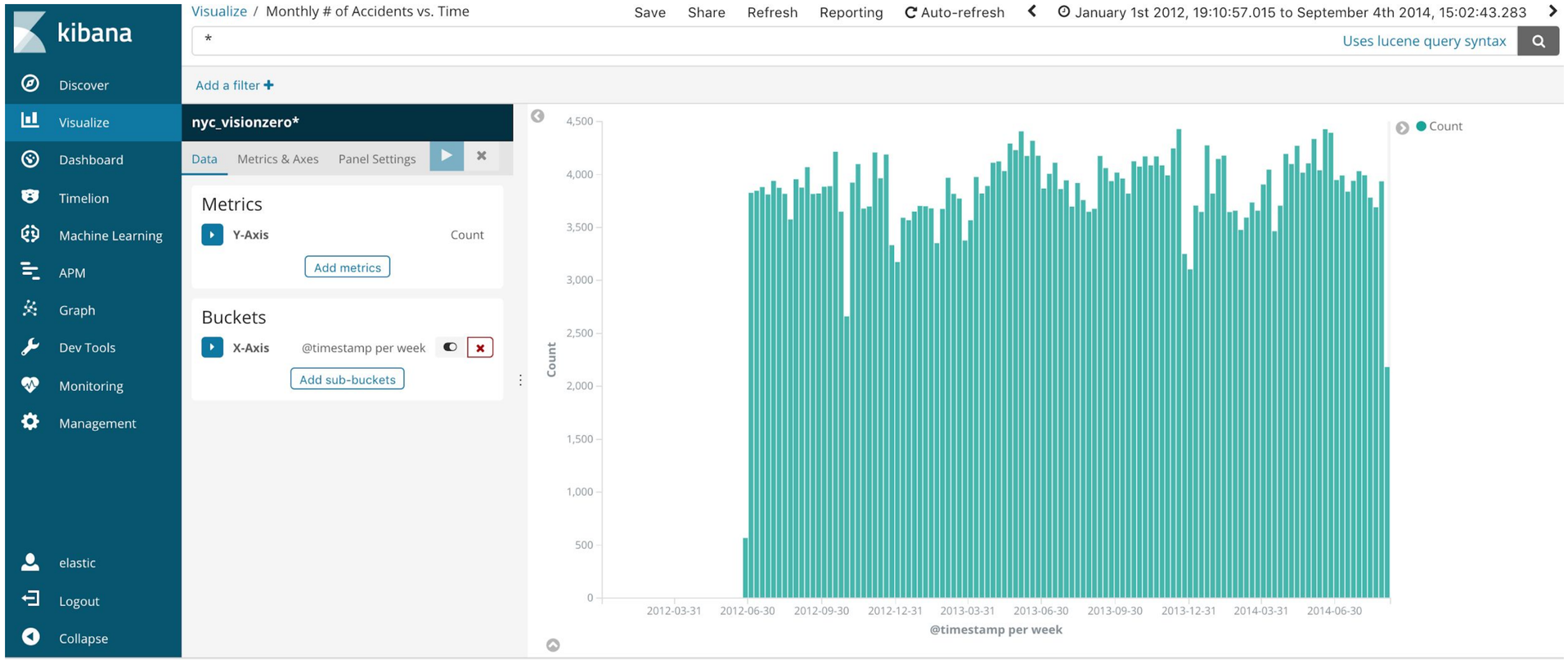
- Visualizations use the Elasticsearch _*search* API
  - REST-API with JSON-base query language
  - Can aggregate results (*similar to "group by" in SQL*)

Kibana Visualizations display the result of aggregations, not the values of individual documents.

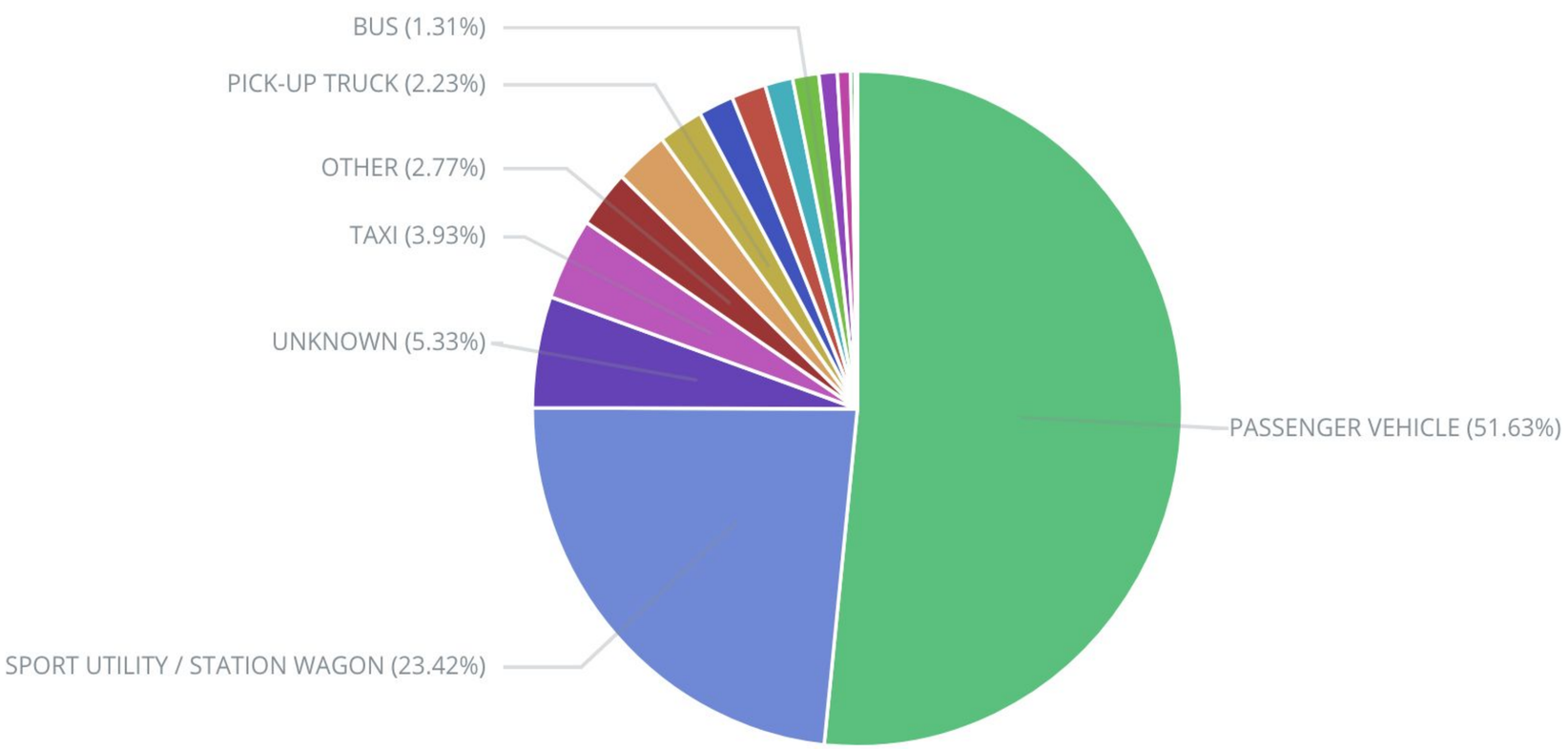
  - This scales better
  - Different data-types have different type of roll-up
    - e.g.
      - seconds, minutes, hours for date values
      - ranges for number values

DEMO TIME !!!

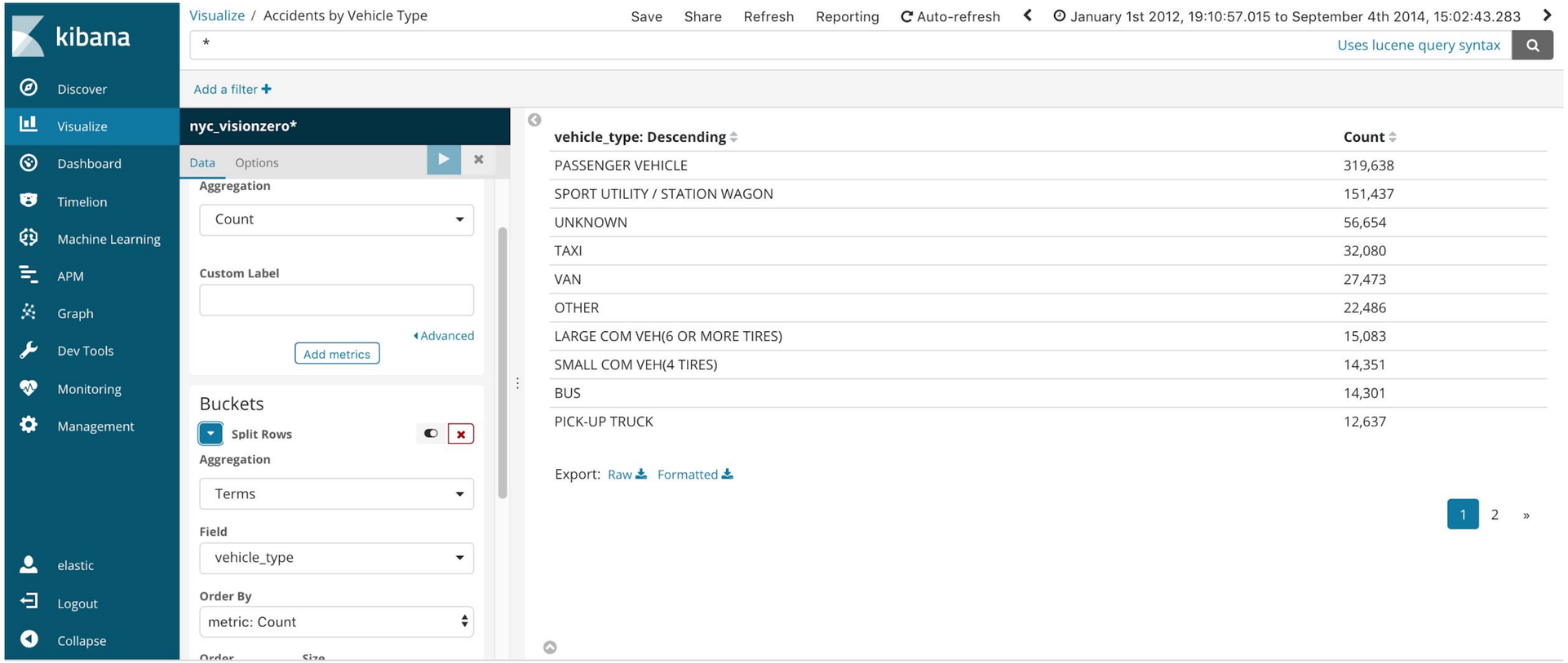
Vertical Bar Chart Monthly # of Accidents vs Time



Vehicles involved in accidents – Pie chart



Data Table Visualization – Accident by Vehicle Type



Region Map Visualization

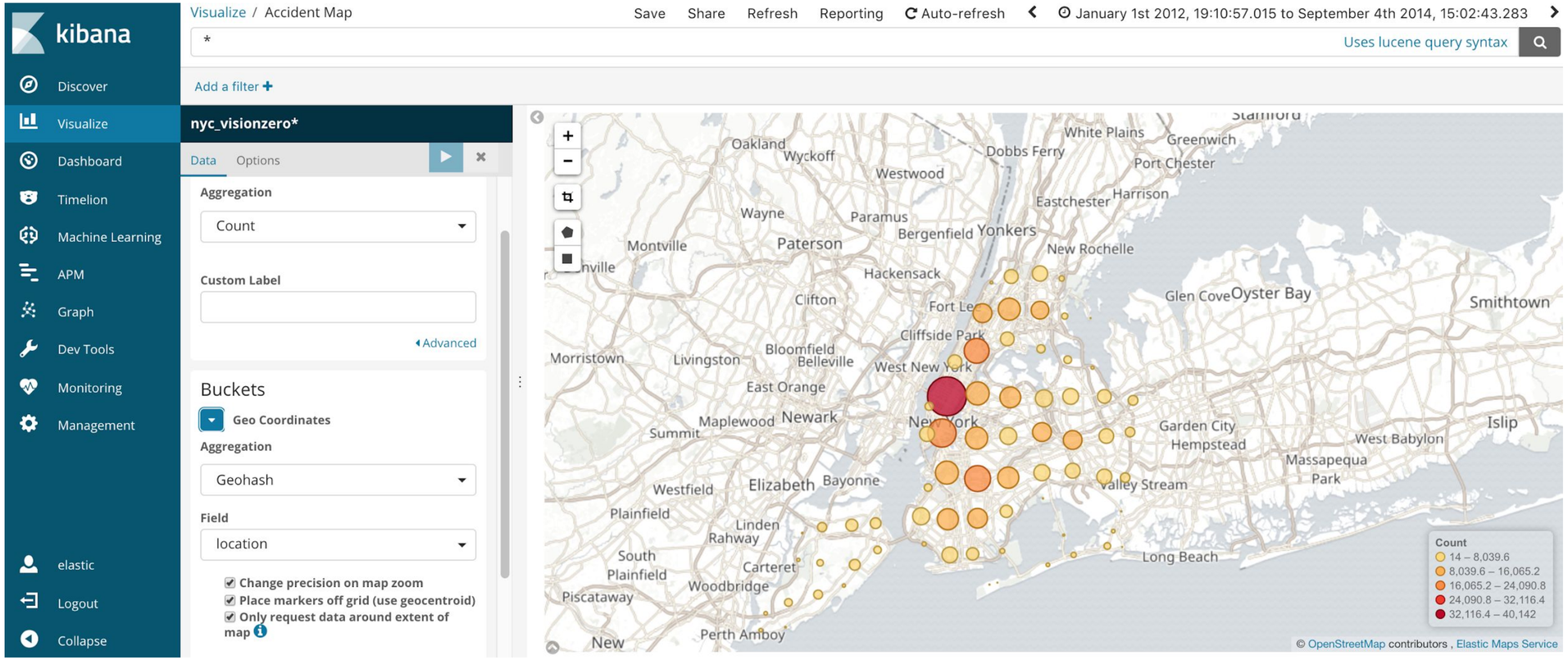
Create choropleth maps

Inner join of results of “terms” aggregation with reference shape data

Link custom data-service in config/kibana.yml (requires CORS support)

```
regionmap:
 layers:
    - name: "NYC Boroughs (self-hosted)"
      url: "http://localhost/region_map/data/nyc_boroughs.json"
      fields:
        - name: "name"
          description: "Borough Name"
    - name: "NYC Council districts (self-hosted)"
      url: "http://localhost/region_map/data/nyc_councildistricts.json"
      fields:
        - name: "CounsilDist"
          description: "District #"
```

Accident Map



Simple Machine Learning job

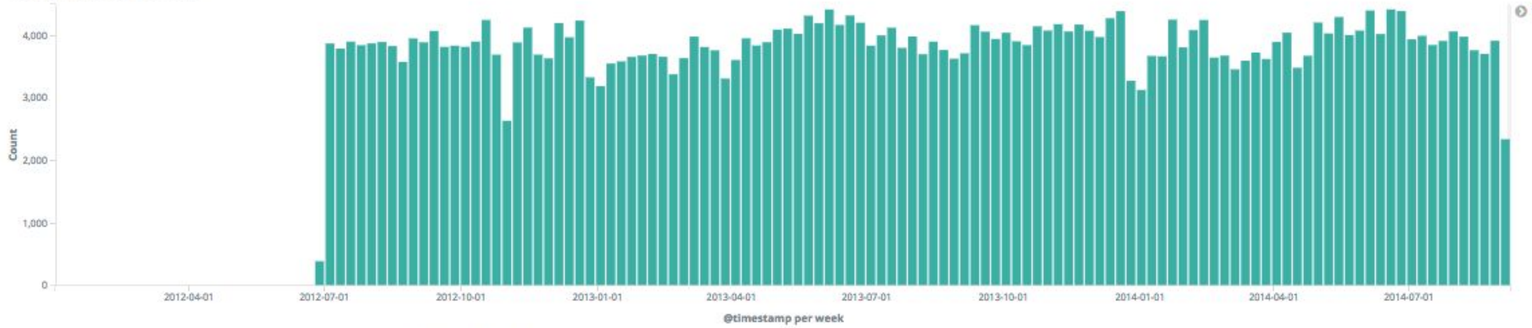
Accidents

Count

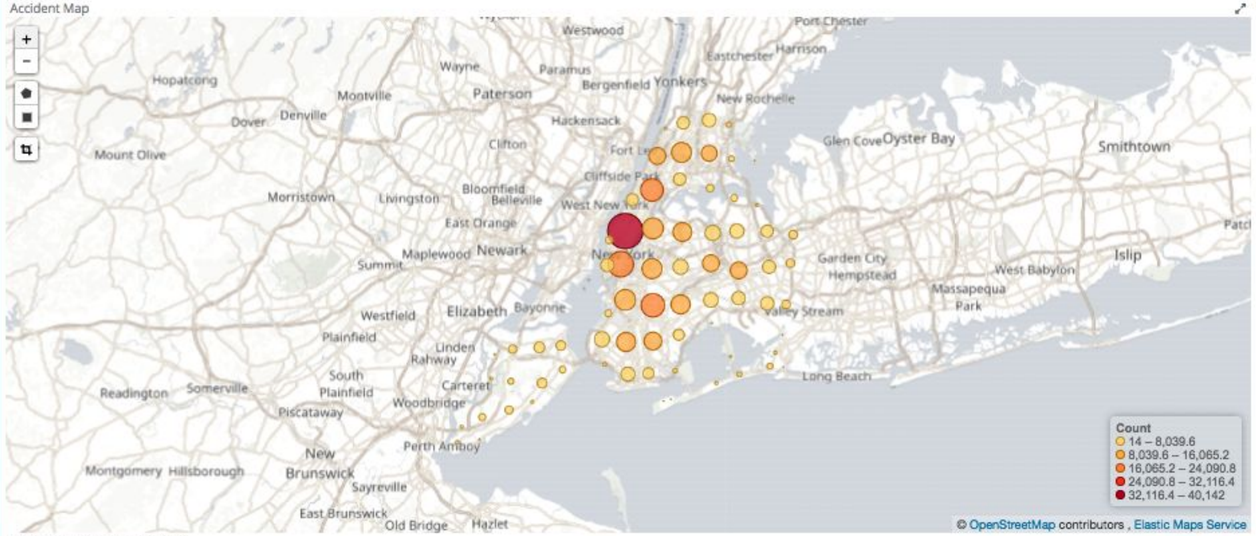
442,319

Count

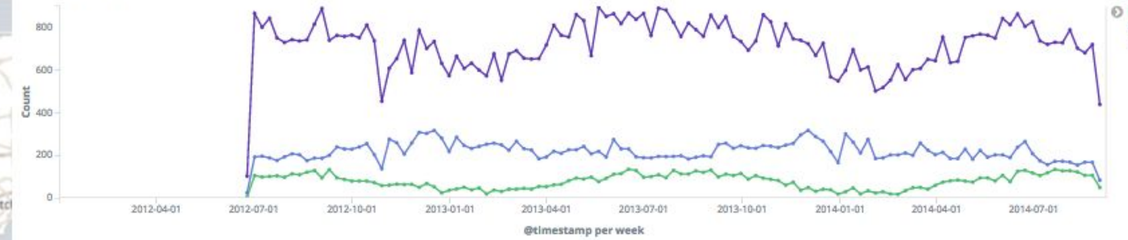
Monthly # of Accidents vs. Time



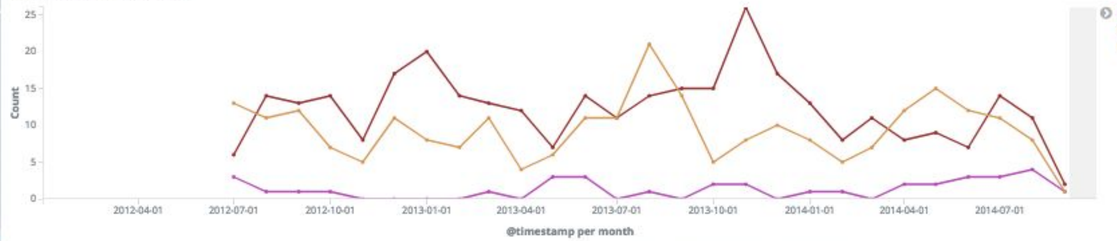
Accident Map



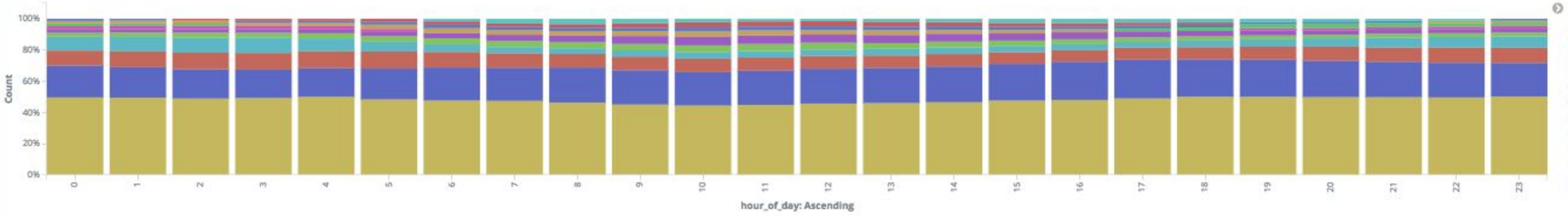
Weekly # of Injuries vs. Time



Monthly # of Fatalities vs. Time



Hourly Trend (by Vehicle Type)



Top Contributing Factors by Borough

Hourly Trend

Thank you!!