anywaze

: analytics on Waze data at scale

(or how to attack google’s servers for fun and {hopefully} profit)

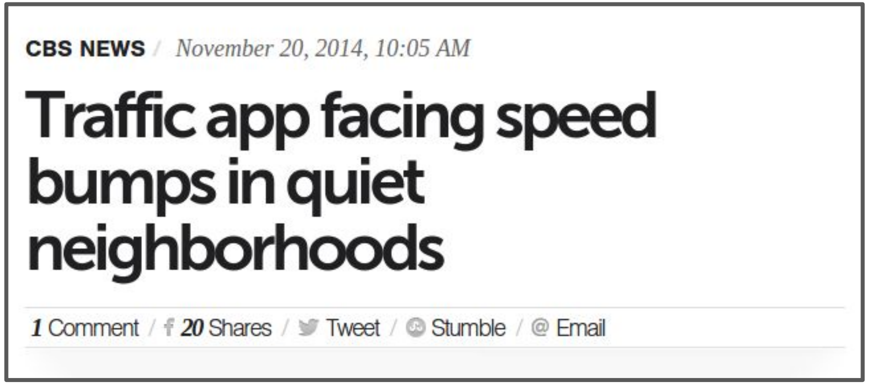
Jason Gors
Insight Fellow
Data Engineering 2016

anywaze.xyz

Motivation:

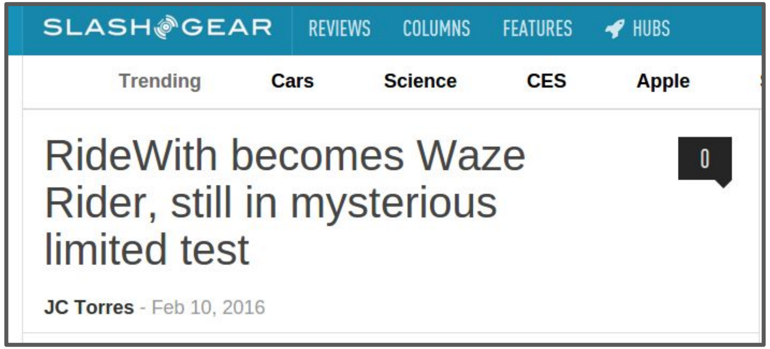
crowdsourced data == controversial:

- Dynamically reroutes users through neighborhoods for quicker driving directions.
- Users report real-time police locations.



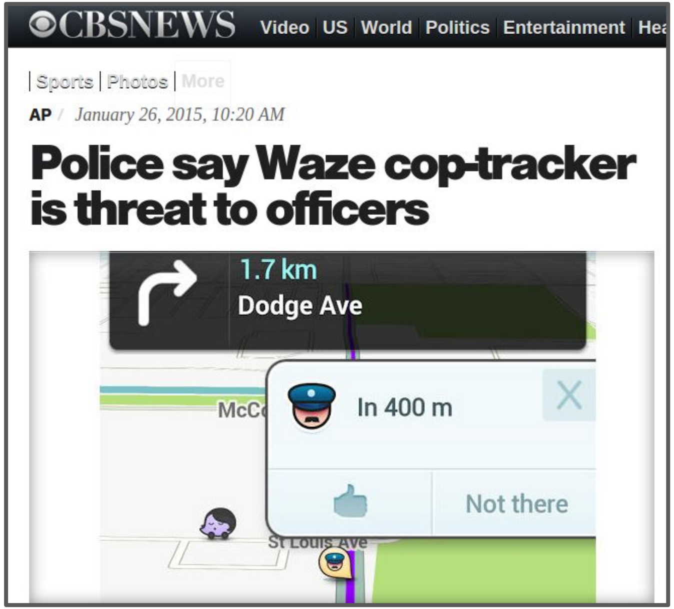
CBS NEWS / *November 20, 2014, 10:05 AM*

Traffic app facing speed bumps in quiet neighborhoods



RideWith becomes Waze Rider, still in mysterious limited test

JC Torres - Feb 10, 2016



AP / *January 26, 2015, 10:20 AM*

Police say Waze cop-tracker is threat to officers

Data:

- Real-time data from 25 largest U.S. cities:
    - 20 x 20 mile square over city centroid (eg. next slide).
- No official API for Waze:
    - had to be “creative” in getting the data.
    - thus data is very messy!
- Currently ~1.5TB and growing.

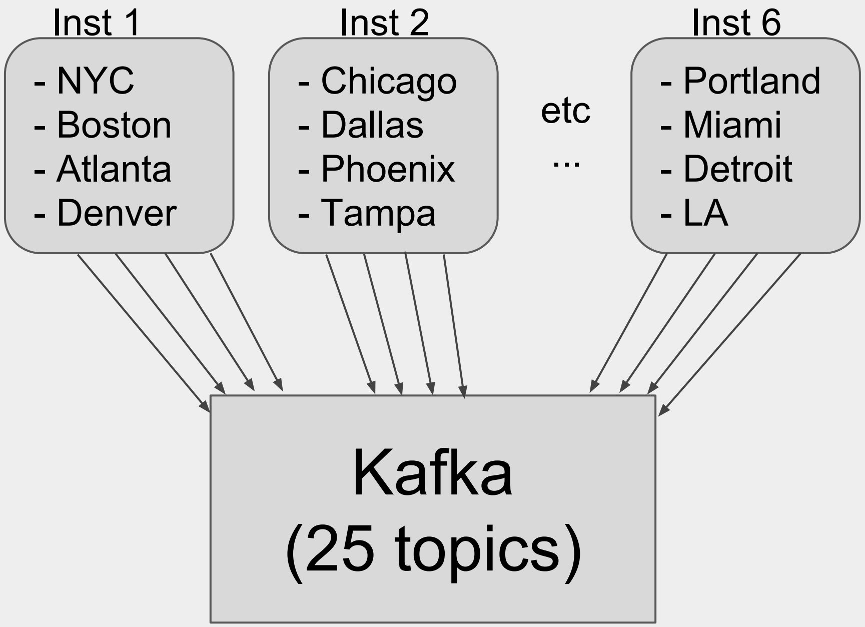
Data:

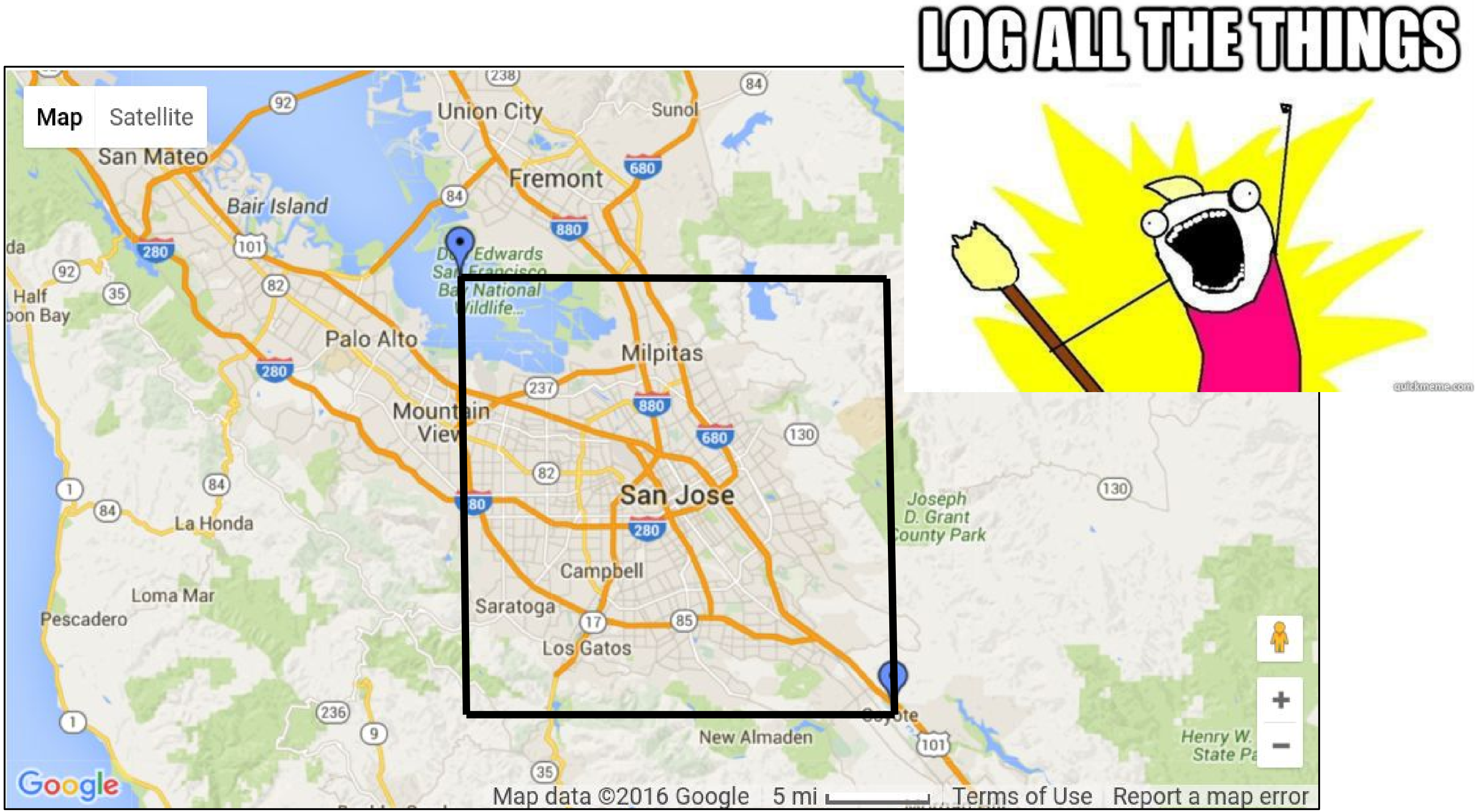
- Real-time data from 25 largest U.S. cities:
  - 20 x 20 mile square over city centroid (eg. next slide).
- No official API for Waze:
  - had to be “creative” in getting the data.
  - thus data is very messy!
- Currently ~1.5TB and growing.

Data Collection:

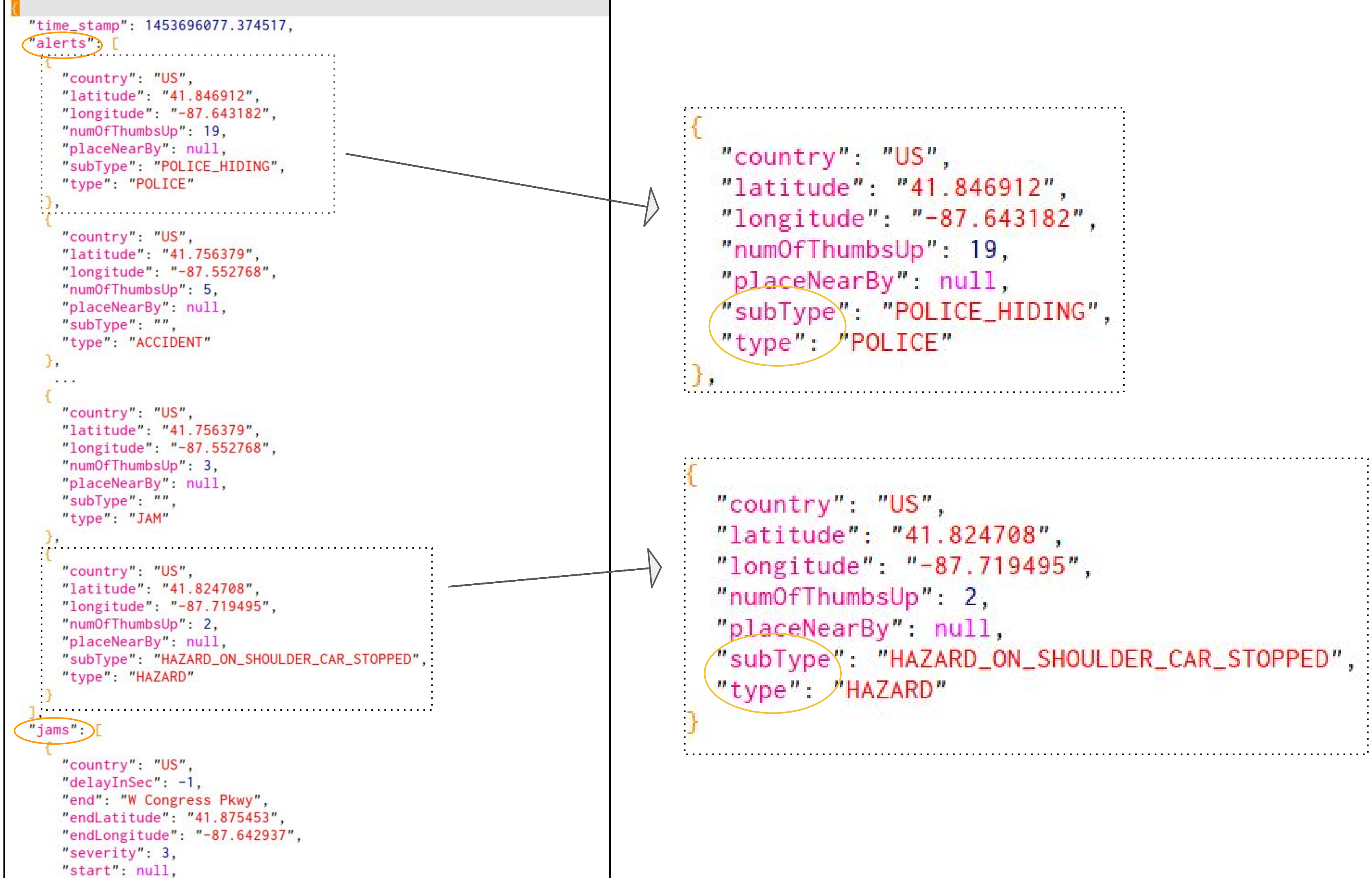
(read: my **“waze botnet”**)

- 6 AWS instances, each running a web server, each hitting 4 to 5 cities every second:




LOG ALL THE THINGS
quickmeme.com
Map
Satellite
Union City
Sunol
San Mateo
Fremont
Bair Island
Don Edwards San Francisco Bay National Wildlife...
Half Moon Bay
Palo Alto
Milpitas
Mountain View
San Jose
Joseph D. Grant County Park
La Honda
Campbell
Loma Mar
Saratoga
Pescadero
Los Gatos
New Almaden
Coyote
Henry W. Coe State Park
Map data ©2016 Google
5 mi
Terms of Use
Report a map error

demo

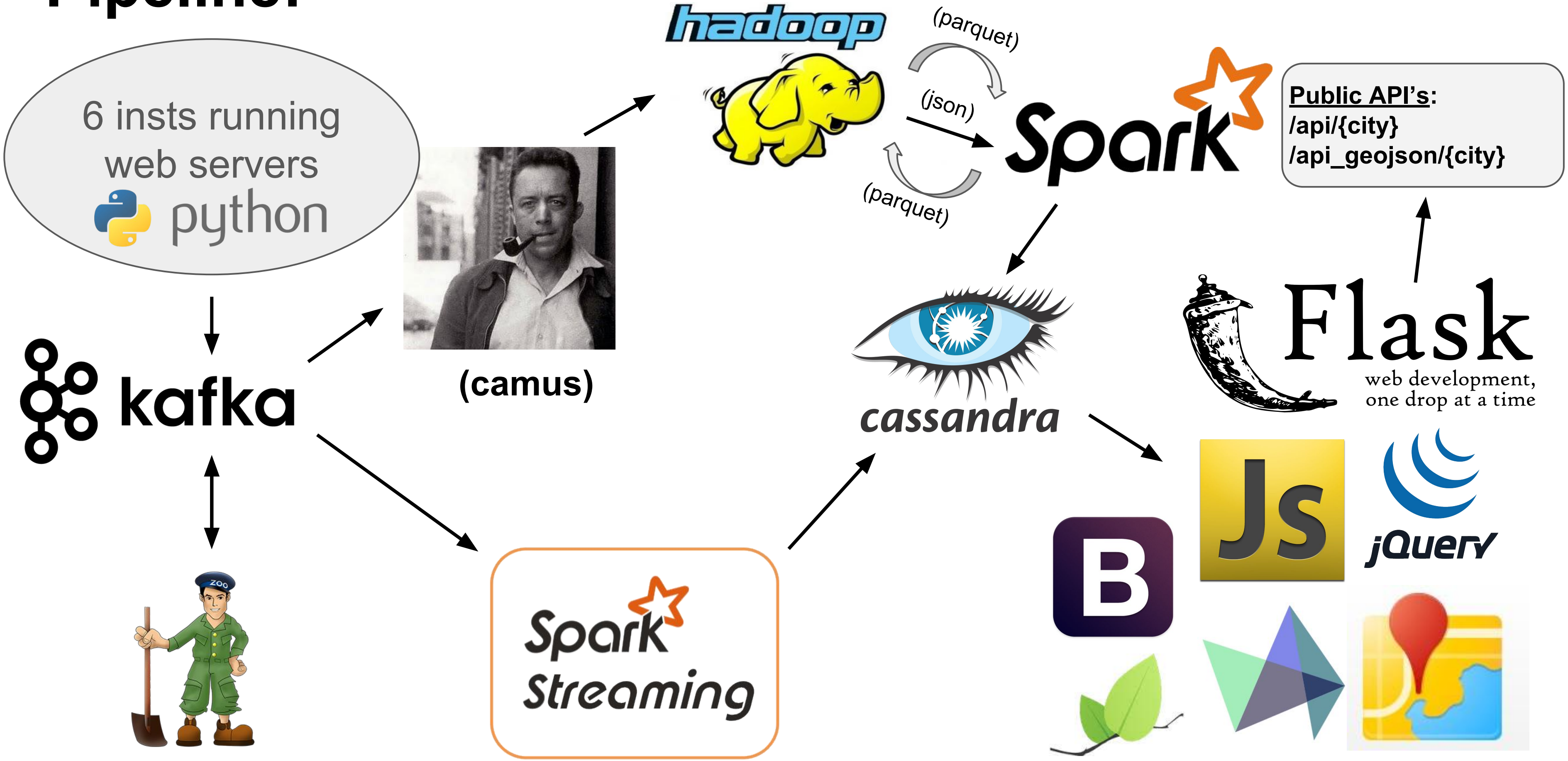


```
{
  "time_stamp": 1453696077.374517,
  "alerts": [
    {
      "country": "US",
      "latitude": "41.846912",
      "longitude": "-87.643182",
      "numOfThumbsUp": 19,
      "placeNearBy": null,
      "subType": "POLICE_HIDING",
      "type": "POLICE"
    },
    {
      "country": "US",
      "latitude": "41.756379",
      "longitude": "-87.552768",
      "numOfThumbsUp": 5,
      "placeNearBy": null,
      "subType": "",
      "type": "ACCIDENT"
    },
     ...
    {
      "country": "US",
      "latitude": "41.756379",
      "longitude": "-87.552768",
      "numOfThumbsUp": 3,
      "placeNearBy": null,
      "subType": "",
      "type": "JAM"
    },
    {
      "country": "US",
      "latitude": "41.824708",
      "longitude": "-87.719495",
      "numOfThumbsUp": 2,
      "placeNearBy": null,
      "subType": "HAZARD_ON_SHOULDER_CAR_STOPPED",
      "type": "HAZARD"
    }
  ],
  "jams":[
    {
      "country": "US",
      "delayInSec": -1,
      "end": "W Congress Pkwy",
      "endLatitude": "41.875453",
      "endLongitude": "-87.642937",
      "severity": 3,
      "start": null,
```

```
{
  "country": "US",
  "latitude": "41.846912",
  "longitude": "-87.643182",
  "numOfThumbsUp": 19,
  "placeNearBy": null,
  "subType": "POLICE_HIDING",
  "type": "POLICE"
},
```

```
{
  "country": "US",
  "latitude": "41.824708",
  "longitude": "-87.719495",
  "numOfThumbsUp": 2,
  "placeNearBy": null,
  "subType": "HAZARD_ON_SHOULDER_CAR_STOPPED",
  "type": "HAZARD"
}
```

Pipeline:



6 insts running web servers

(camus)

(parquet)

(json)

(parquet)

Public API's:
/api/{city}
/api_geojson/{city}

Pipeline (cont’d):

6 instances:
Data collection web servers

c4.xlarge
(network io)

4 instances:
Kafka, Camus, HDFS, Cassandra

m4.large
(storage)

5 instances:
Spark

m4.xlarge
(compute)

1 instance:
Flask

m4.large
(general)

Challenges encountered:

- Data is very messy -- difficult to wrestle into a usable format.
    - Each GET Response from their servers has changing nested json fields.
- Bringing nodes back online after failure.
- Scaling instance storage in real-time (== much pain).
- Adding new nodes into active clusters in real-time.
- How best to represent 25 cities of data in Kafka:
    - 25 topics, 1 Partition each
- Unit & CI testing is hard with these distributed tools.

About me...

- Previously a Data Scientist
- ...and Software Developer
- Graduate work in Statistics
- A Bayesian
- Python & Linux evangelist

Anywaze, take a look:

@github: jgors/anywaze

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.

Questions?