# Correlation analysis on spatiotemporal data

Heejong Kim(hk2451), Jun Yuan(jy50), Jiin Nam(jn1664)

*Cool Name Pending* Team

Big Data (DS-GA 1004) 18' Spring

NEW YORK UNIVERSITY

# Pipeline

## Data Preprocessing

- Data Cleaning
- Data Aggregation

## Data Analysis

- Mutual Information Calculation
- Analyzing correlation result

# Data cleaning


Number of columns for each issue in each dataset

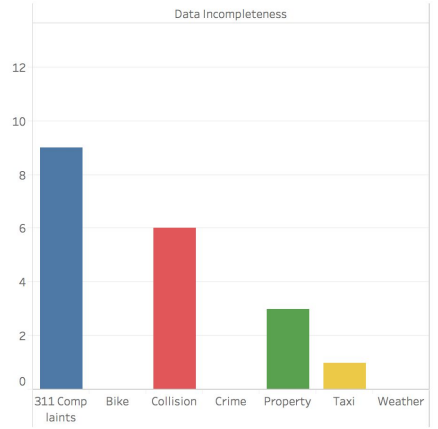
**Data Quality Issues**

1. Incompleteness
2. Inconsistency
3. Incorrectness
4. Redundancy

# Data Cleaning

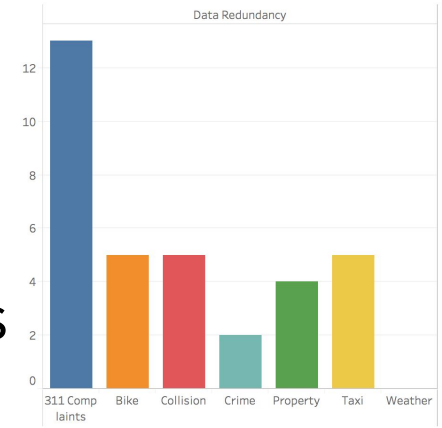
## Data Incompleteness

- Performance: A lot of “empty” values (e.g.: null, N/A, Unspecified, ect.)
- Solution: Delete columns with over 90% empty values.




## Data Redundancy

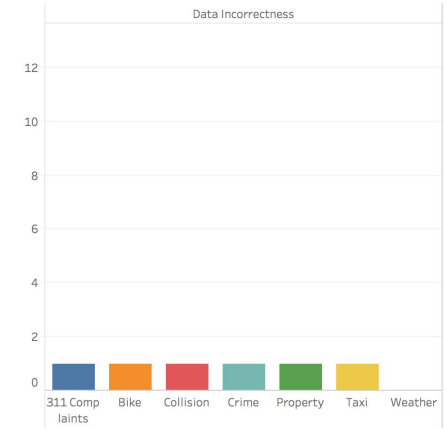
- Performance: Multi-columns describe the same / similar information
- Solution: Keep one column among the similar columns

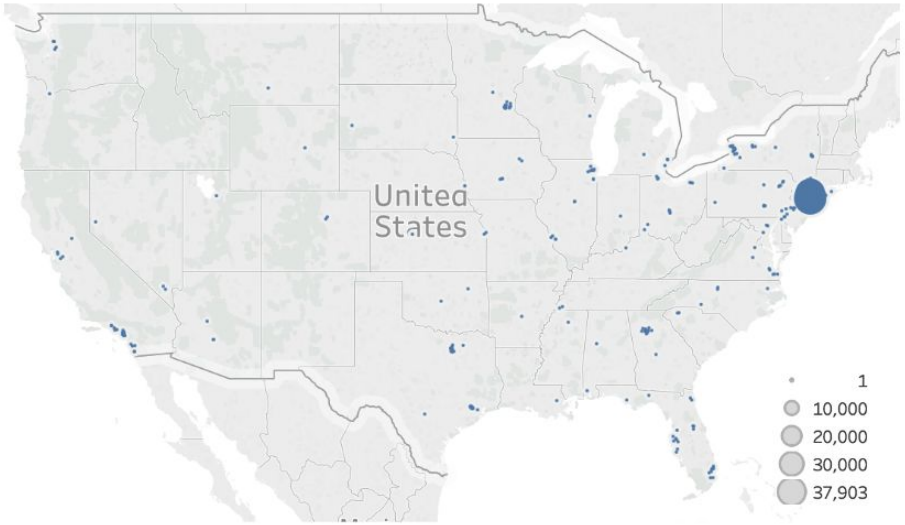


# Data Cleaning

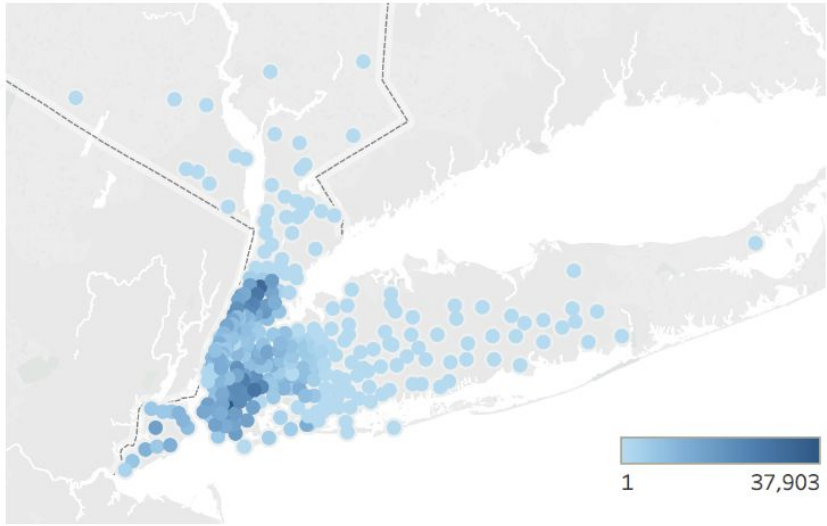
## Data Incorrectness

- Performance: location not in New York City
- Solution: NYC zip code validation




Locations **Before** Cleaning

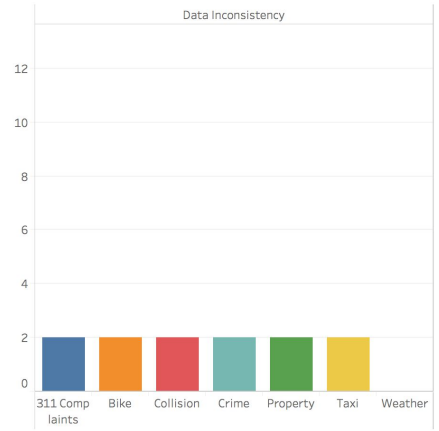


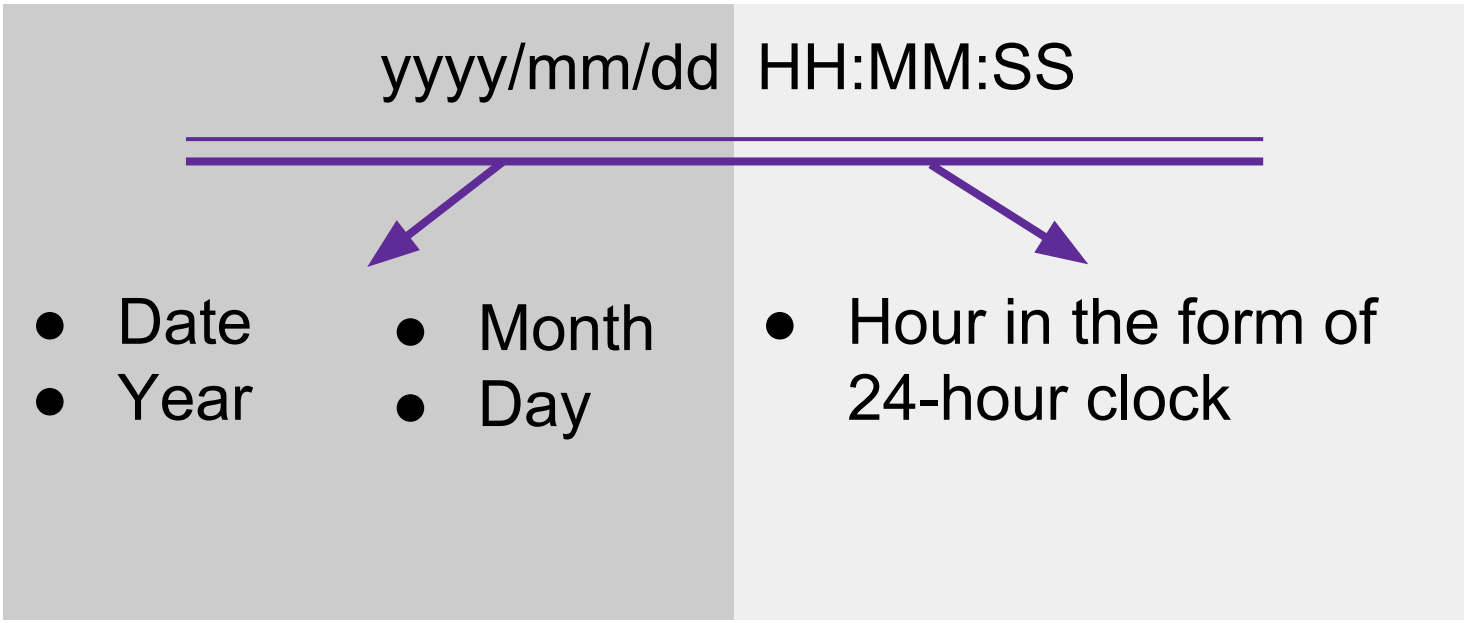

Locations **After** Cleaning

# Data Cleaning

## Data Inconsistency

- Performance: Different format and granularity
- Solution
  - For zipcode: keep first 5 digits
  - For time: Formatter and new attrs mapping



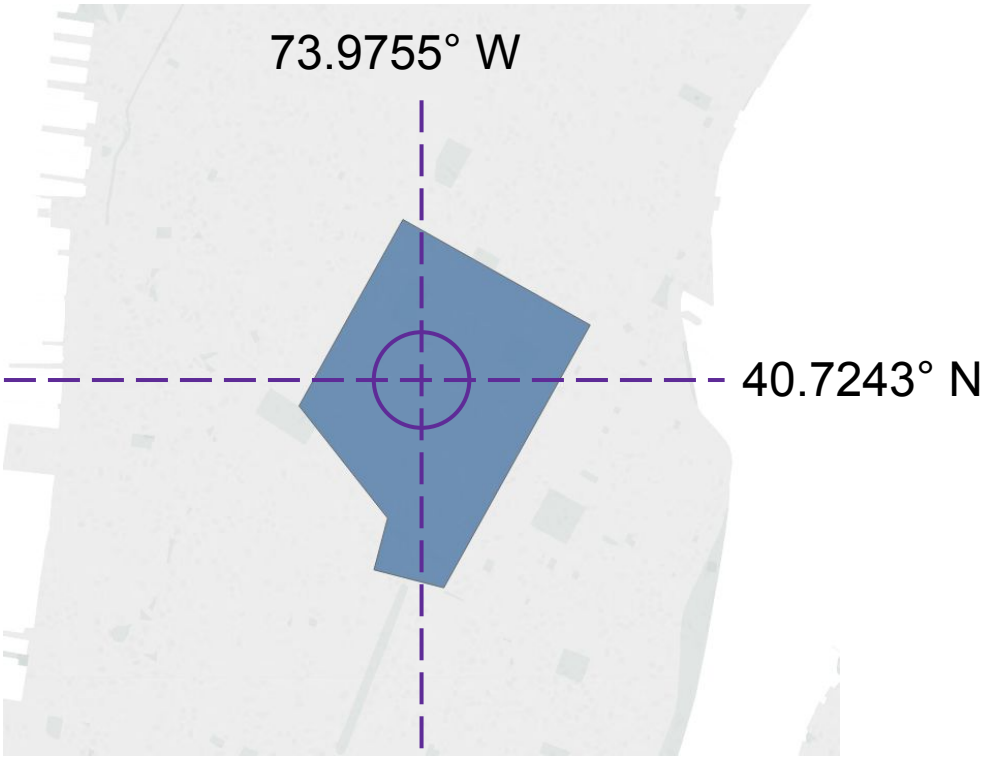



# Data Aggregation

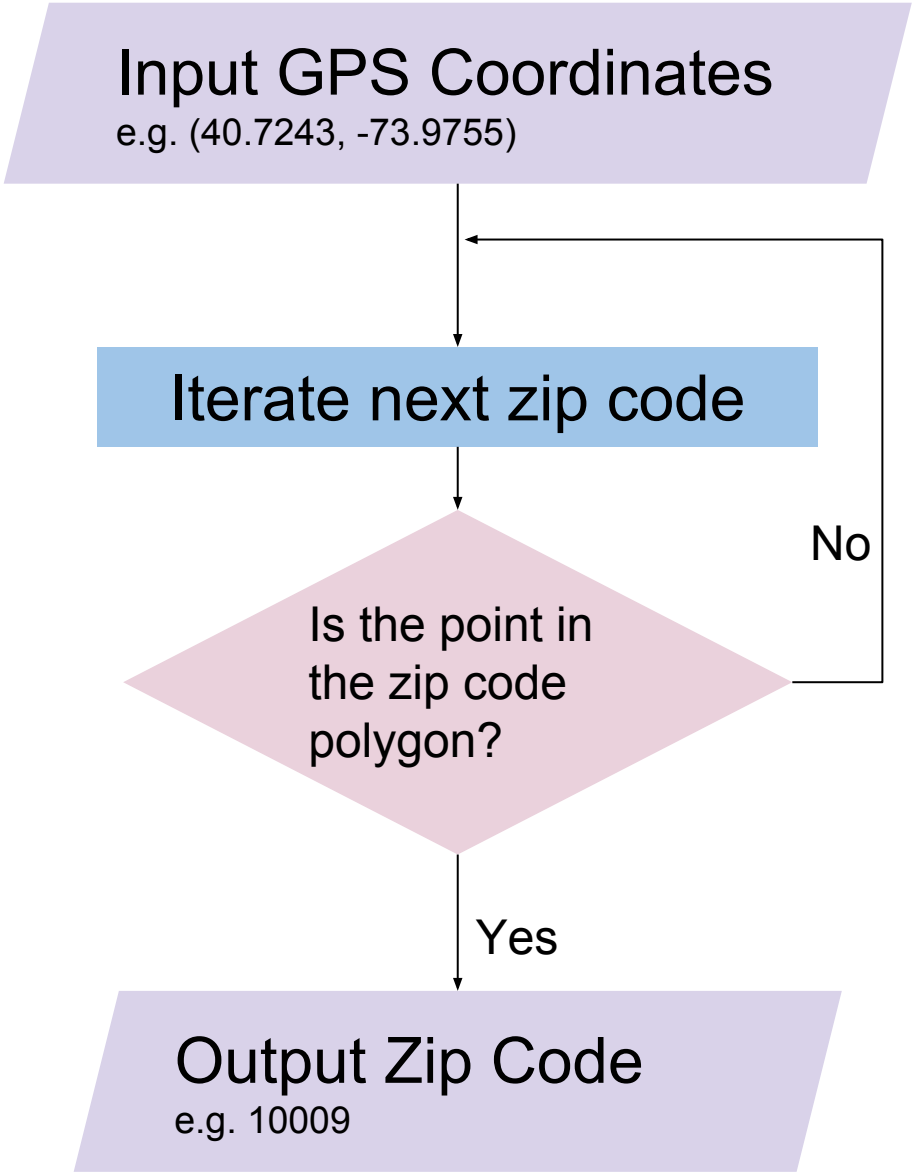
- Zip Code Mapping
- Attribute Transformation
- Attribute Categorization

# Data Aggregation

## Zip Code Mapping

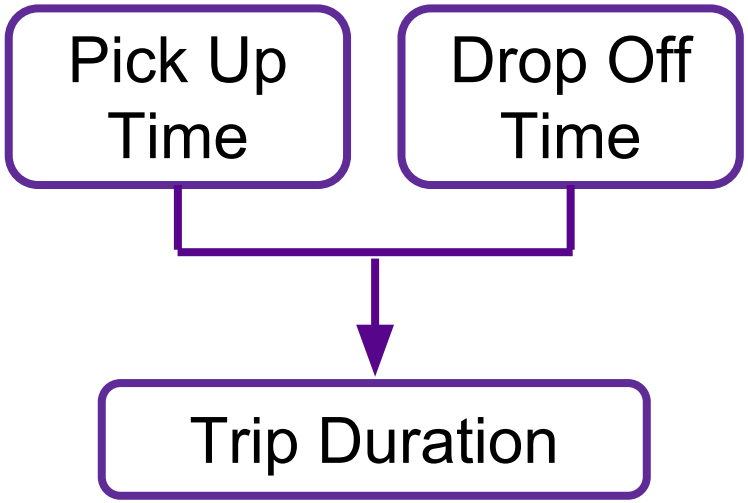


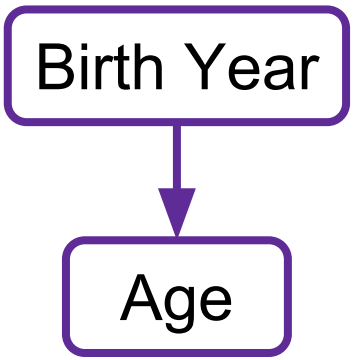

Zip code: 10009



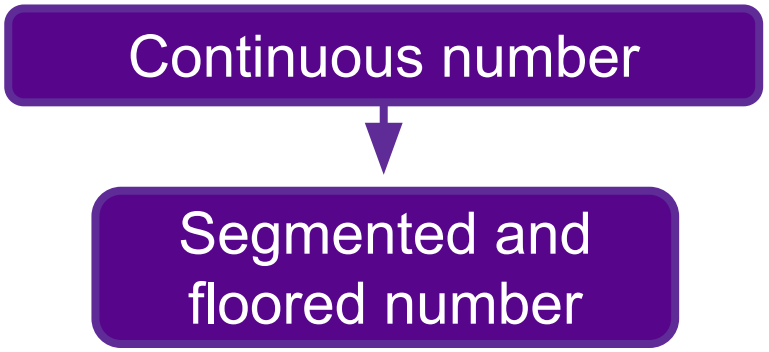
# Data Aggregation

## Transformation

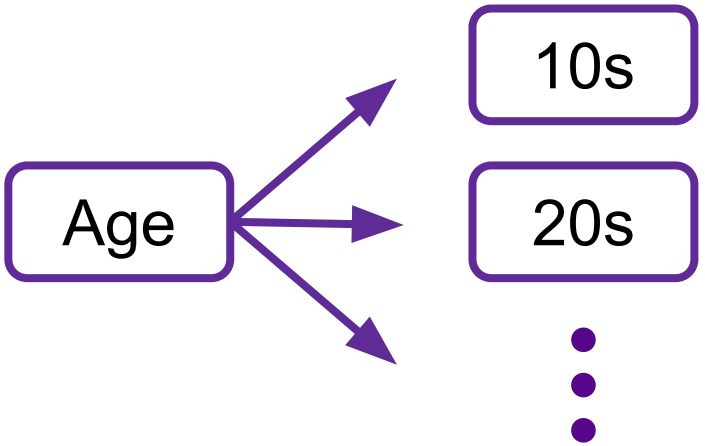



## Categorization

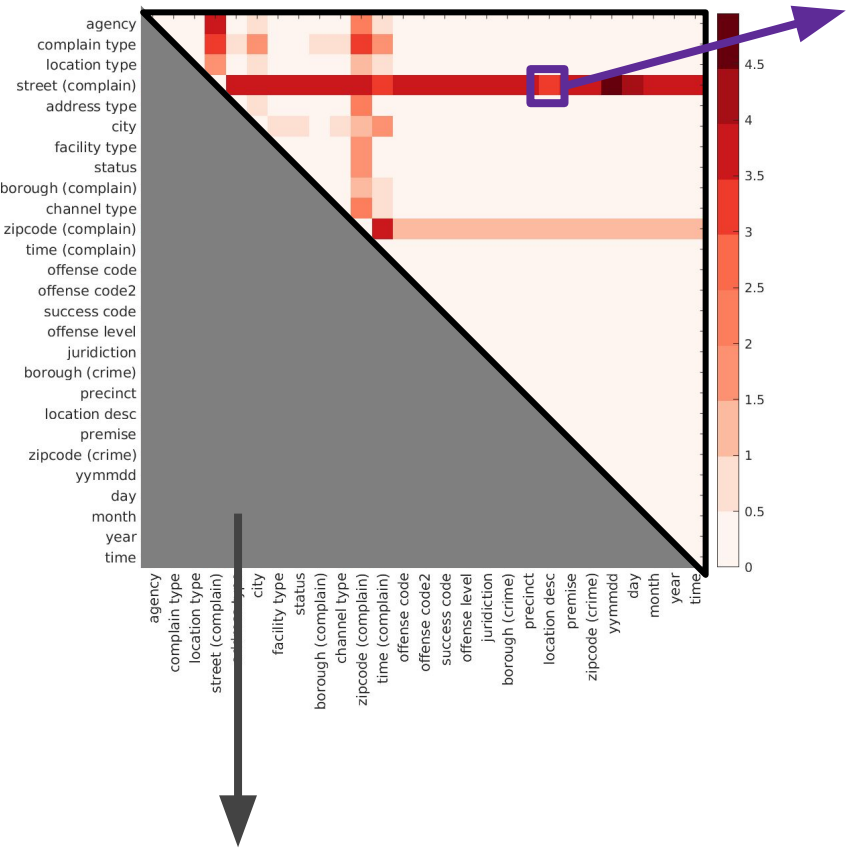



Example: speed, temperature, etc



# Mutual Information

## Calculating “Mutual Information” with example




Mutual information between
“Street name (X) from complain dataset”
& “Location description (Y) from crime dataset”

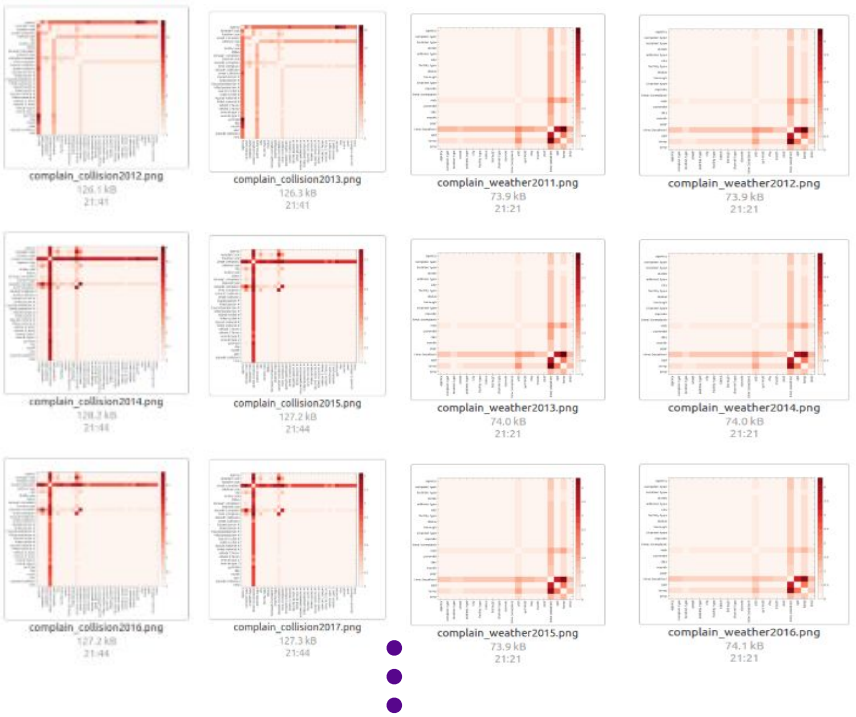
1) Join two dataset with dates (“yymmdd”)
2) Calculate
   a) P(X) and P(Y)
   b) P(XY)
   c) Mutual Information =
      P(XY) * log( P(XY) / (P(X)*P(Y)) )

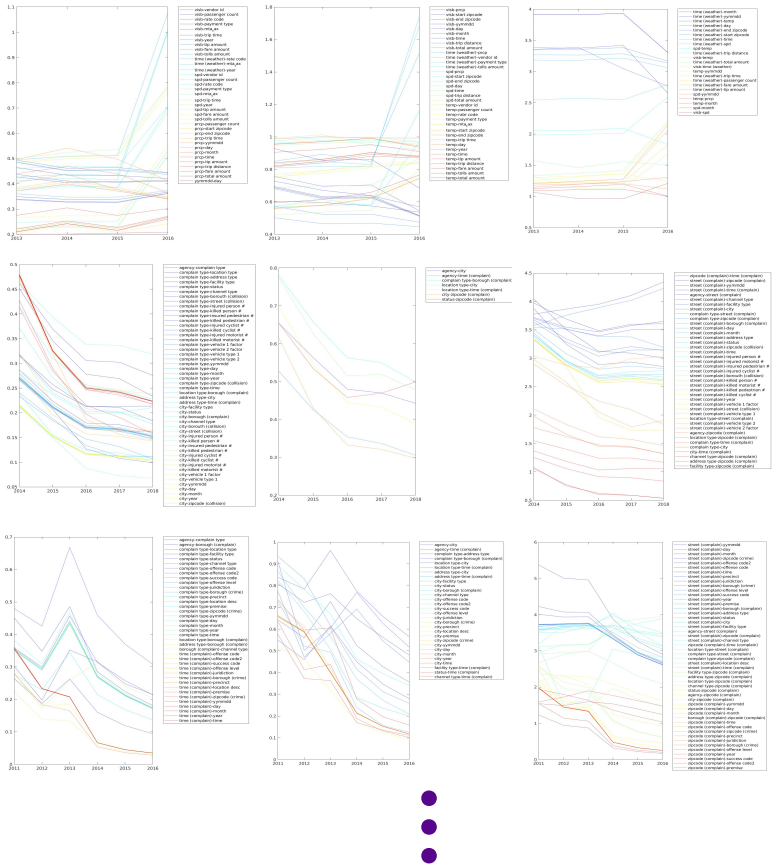
Correlation matrix is symmetric

⇨ We calculated upper triangle only for efficiency

# Correlation results

## Overview

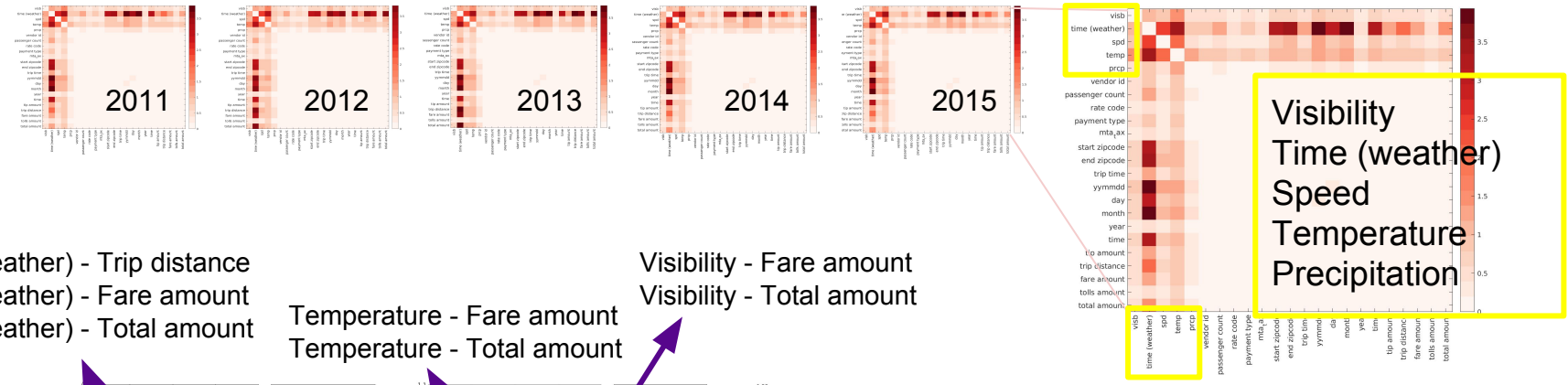


| | Collision | Taxi | Complain | Bike | Crime | Property | Weather |
|---|---|---|---|---|---|---|---|
| Collision | . | | | | | | |
| Taxi | . | . | | | | | |
| Complain | . | . | . | | | | |
| Bike | . | . | . | . | | | |
| Crime | . | . | . | . | . | | |
| Property | . | . | . | . | . | . | |
| Weather | . | . | . | . | . | . | . |

Tables Join By Year

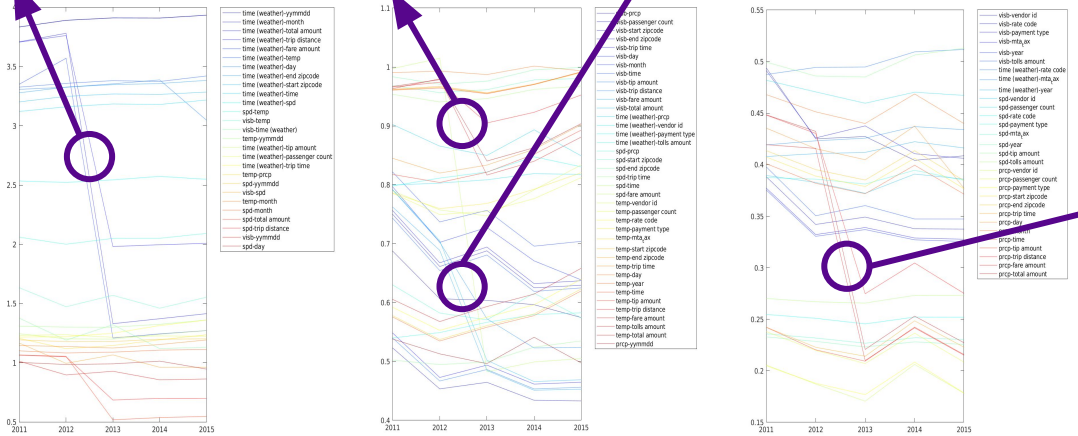
# Case study 1: Weather and Taxi




Time (weather) - Trip distance
Time (weather) - Fare amount
Time (weather) - Total amount

Temperature - Fare amount
Temperature - Total amount

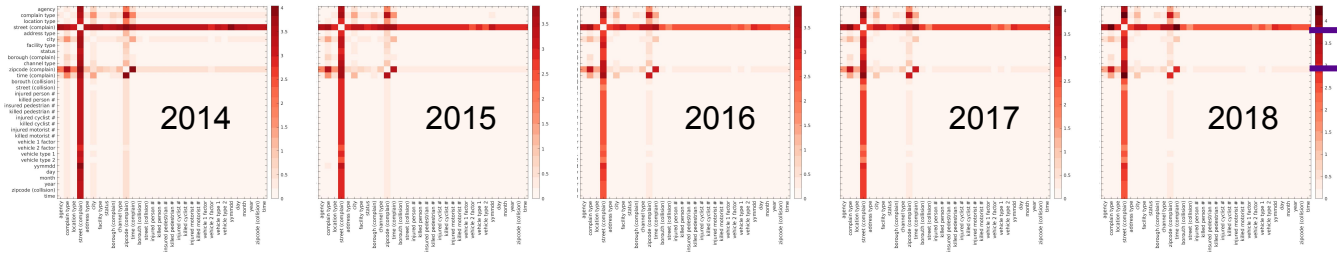
Visibility - Fare amount
Visibility - Total amount



Time (weather) - Trip distance
Time (weather) - Fare amount
Time (weather) - Total amount

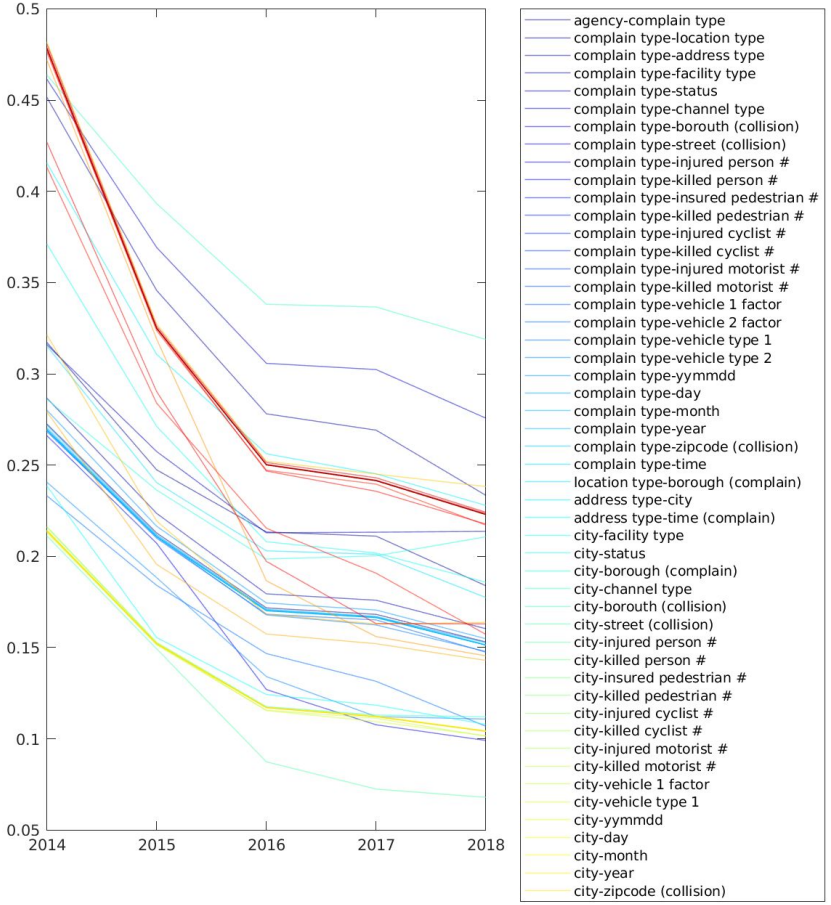
- Weather attributes are highly correlated to the taxi tips, passenger count, trip time.
- Correlations between the weather and some taxi trip attrs drop a lot in 2013. Maybe because Uber became popular and got approved in NYC in 2013 [1][2].

# Case study 2: Complaint and Collision




Street attribute in Complaint
Zip code in Complaint

Difference between business area and residential area

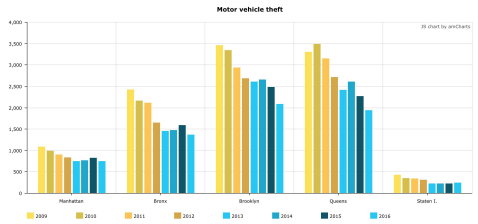



zipcode (complain) - attributes

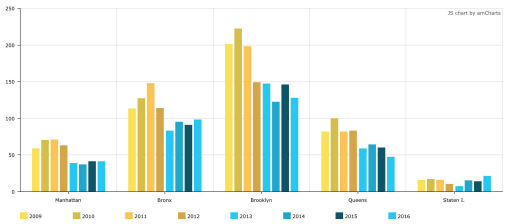
- Complain type is correlated to locations, streets

- The correlations are decreasing

- More expensive traffic tickets
- Less Crime in Brooklyn/Bronx => less correlated




Motor vehicle theft [3]




Murder [3]

# References

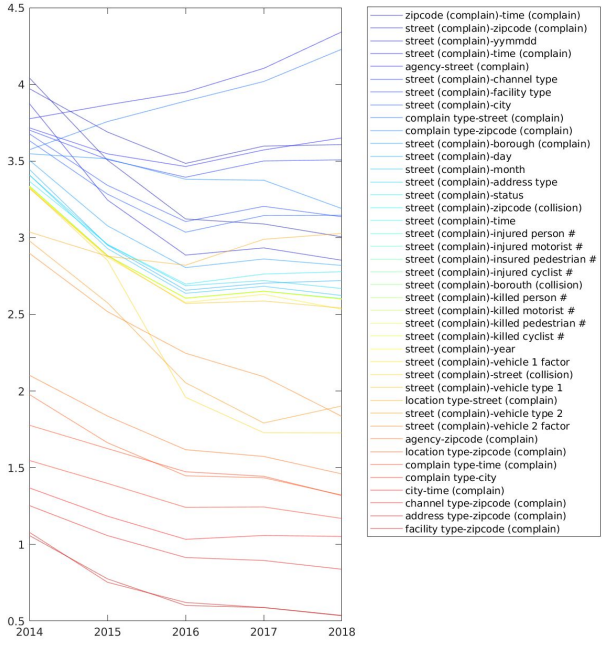
[1] Uber Closes Yellow Taxi Cab Service In New York City: https://www.forbes.com/sites/tomiogeron/2012/10/16/uber-closes-yellow-taxi-cab-service-in-new-york-city/#fe9787051e27
[2] After long battle, Uber becomes first taxi app to get approved in New York City: https://www.theverge.com/2013/4/26/4271490/uber-becomes-first-taxi-app-to-get-approved-in-new-york-city
[3] Major crime in New York City, 2009-2016: https://projects.newsday.com/databases/long-island/new-york-city-crime-rate/

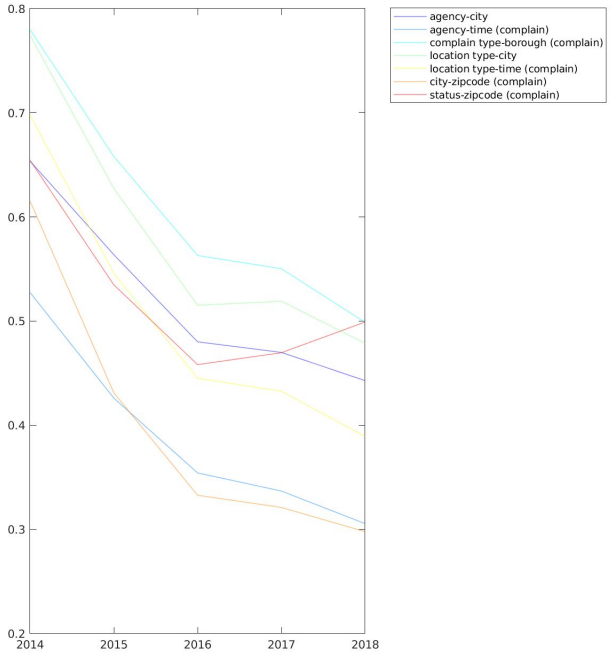
# Thanks :)

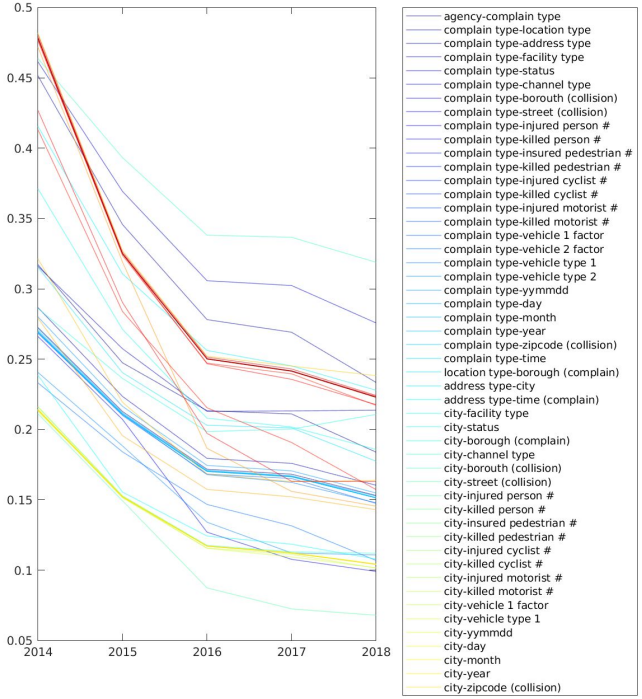
Cool Name Pending Team

# Supplementary

## Complain collision 2014-2018

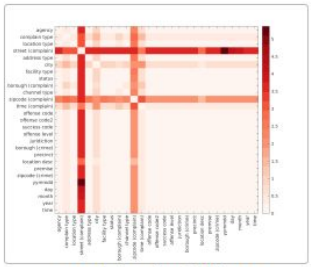


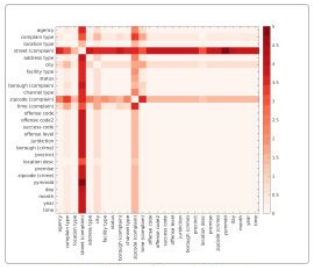



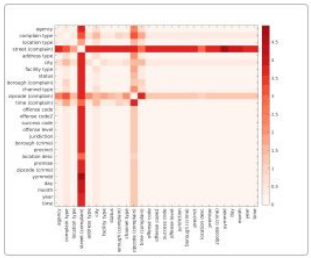



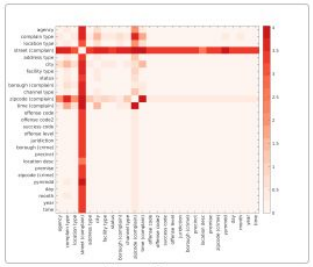
# Supplementary

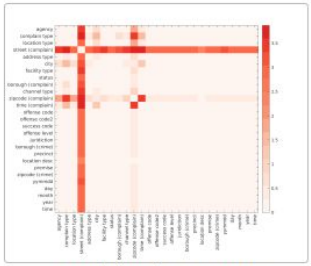
## Complain crime 2011-2016

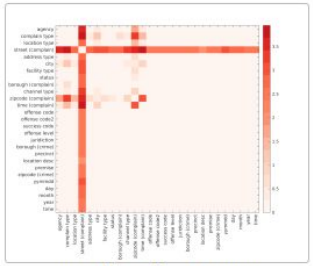

complain_crime2011.png

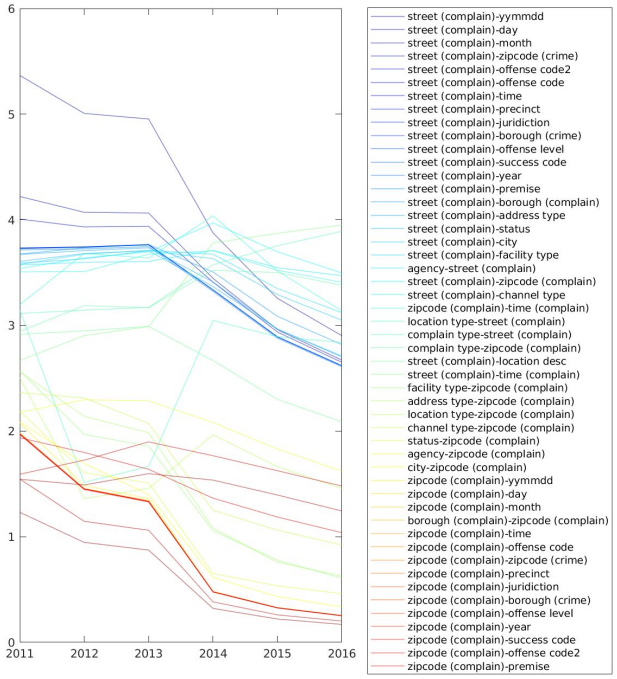

complain_crime2012.png

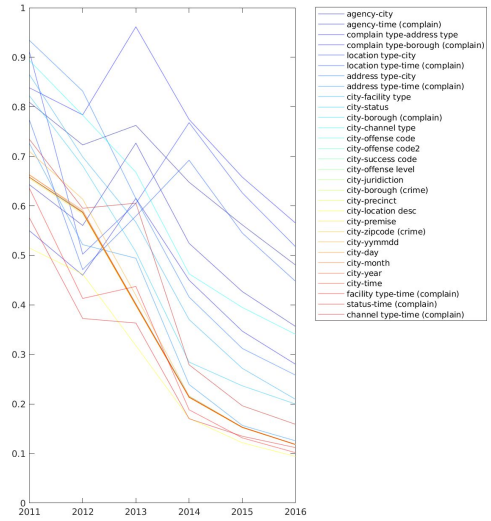

complain_crime2013.png

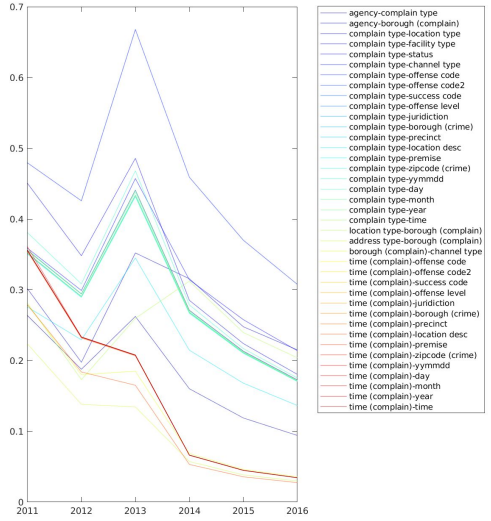

complain_crime2014.png


complain_crime2015.png


complain_crime2016.png

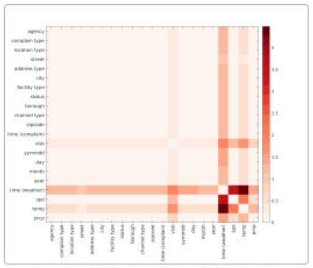


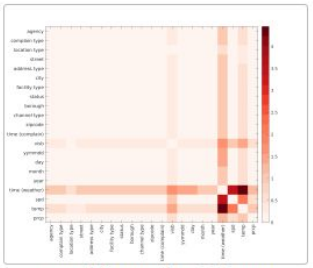



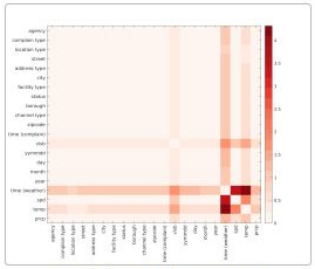



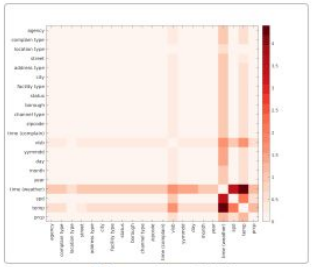
# Supplementary

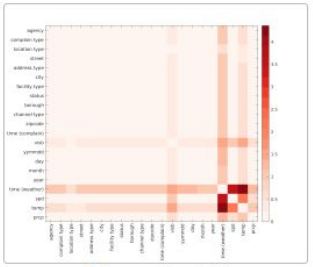
## Complain weather 2011-2017

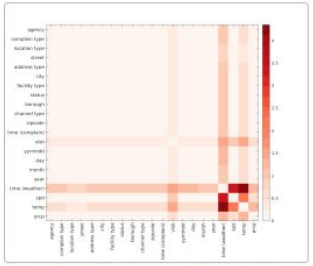

complain_weather2011.png

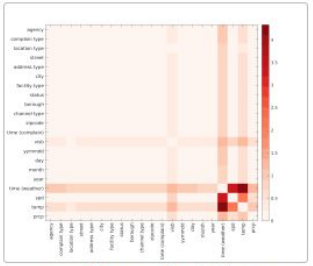

complain_weather2012.png

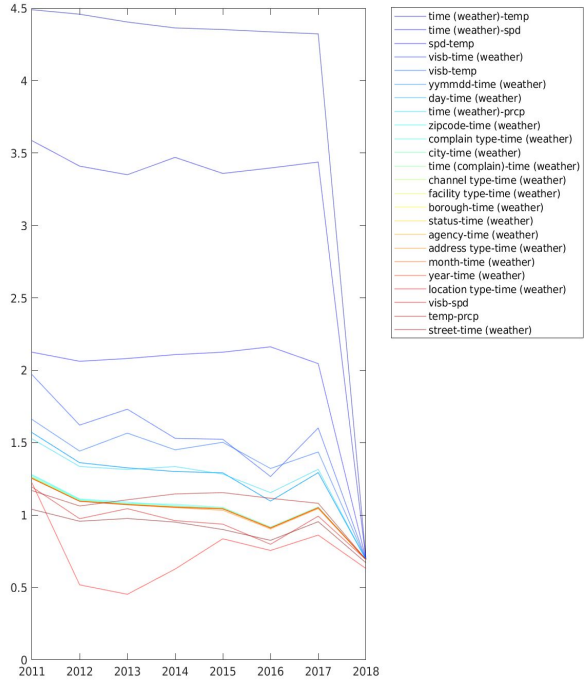

complain_weather2017.png

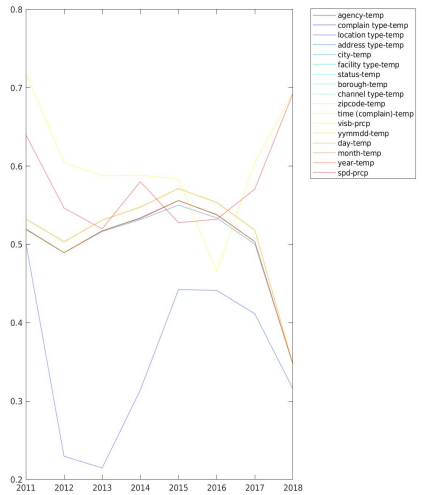

complain_weather2013.png

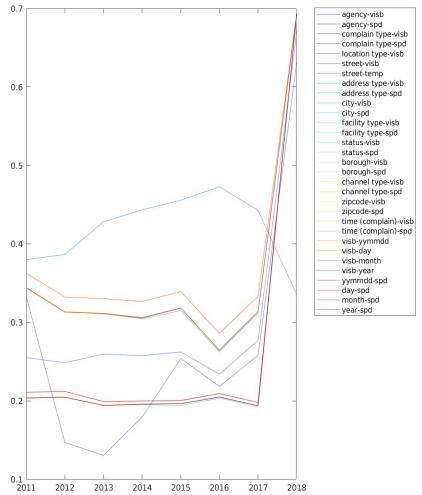

complain_weather2014.png


complain_weather2015.png


complain_weather2016.png

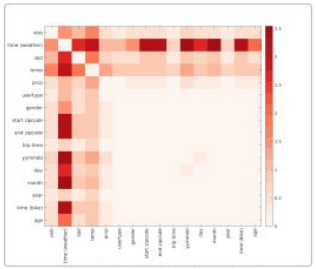


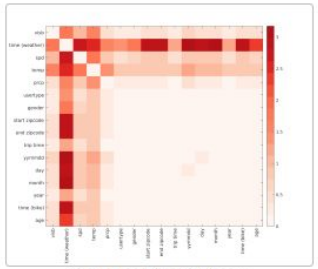



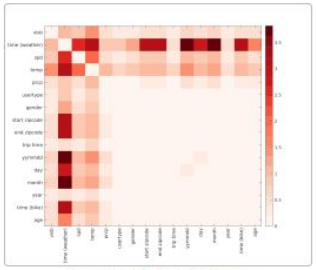



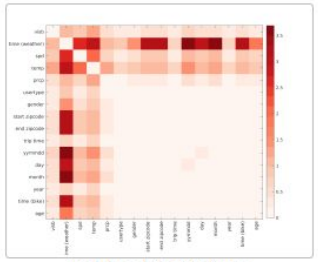
# Supplementary

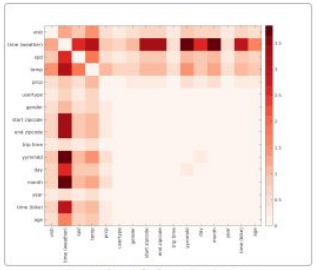
## Weather Bike 2013-2018

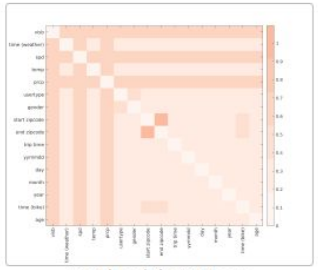

weather_bike2013.png


weather_bike2014.png


weather_bike2015.png


weather_bike2016.png


weather_bike2017.png


weather_bike2018.png