Crimes in San Francisco Neighborhood
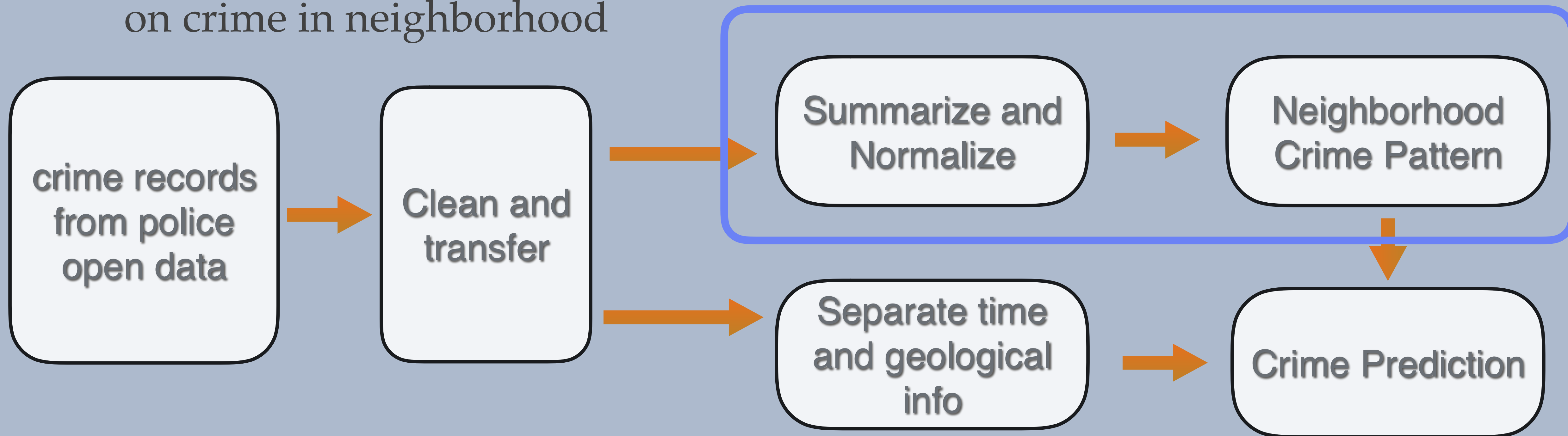
Zhenyu Chen
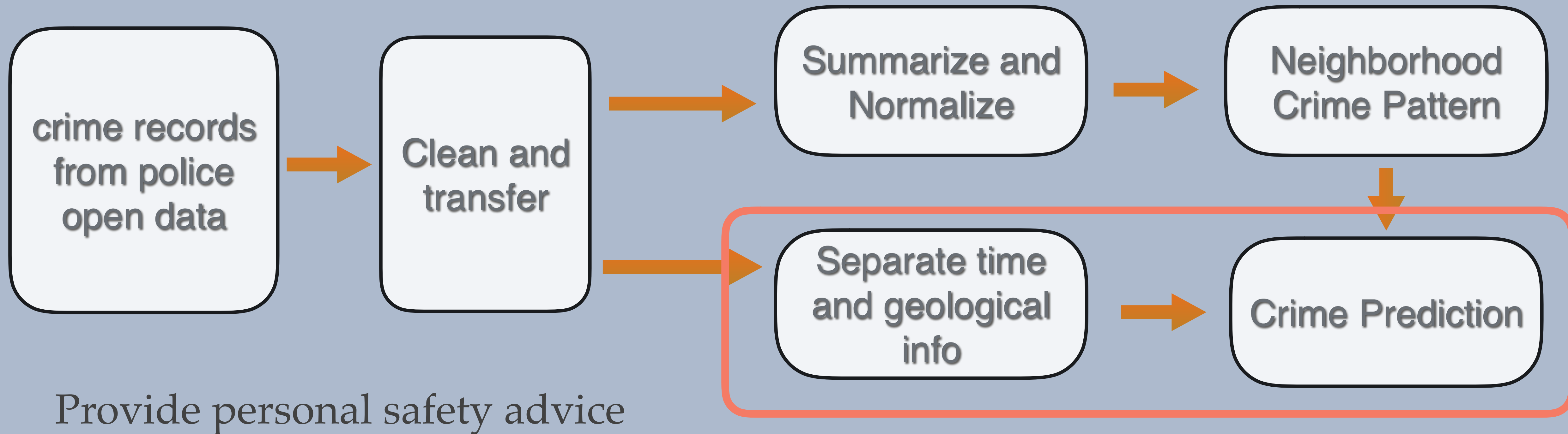
Data Flow

Easier and better understanding
on crime in neighborhood

crime records from police open data

→

Clean and transfer

→

Summarize and Normalize

→

Neighborhood Crime Pattern

→

Separate time and geological info

→

Crime Prediction

1. https://publicsafetydataportal.org/
2. Google Maps API
3. SF Opendata

Data Flow

```
crime records          Clean and          Summarize and          Neighborhood
from police     →      transfer      →      Normalize       →      Crime Pattern
open data                                                               ↓

                                          Separate time
                                          and geological    →    Crime Prediction
                                               info
```

Provide personal safety advice

1. https://publicsafetydataportal.org/
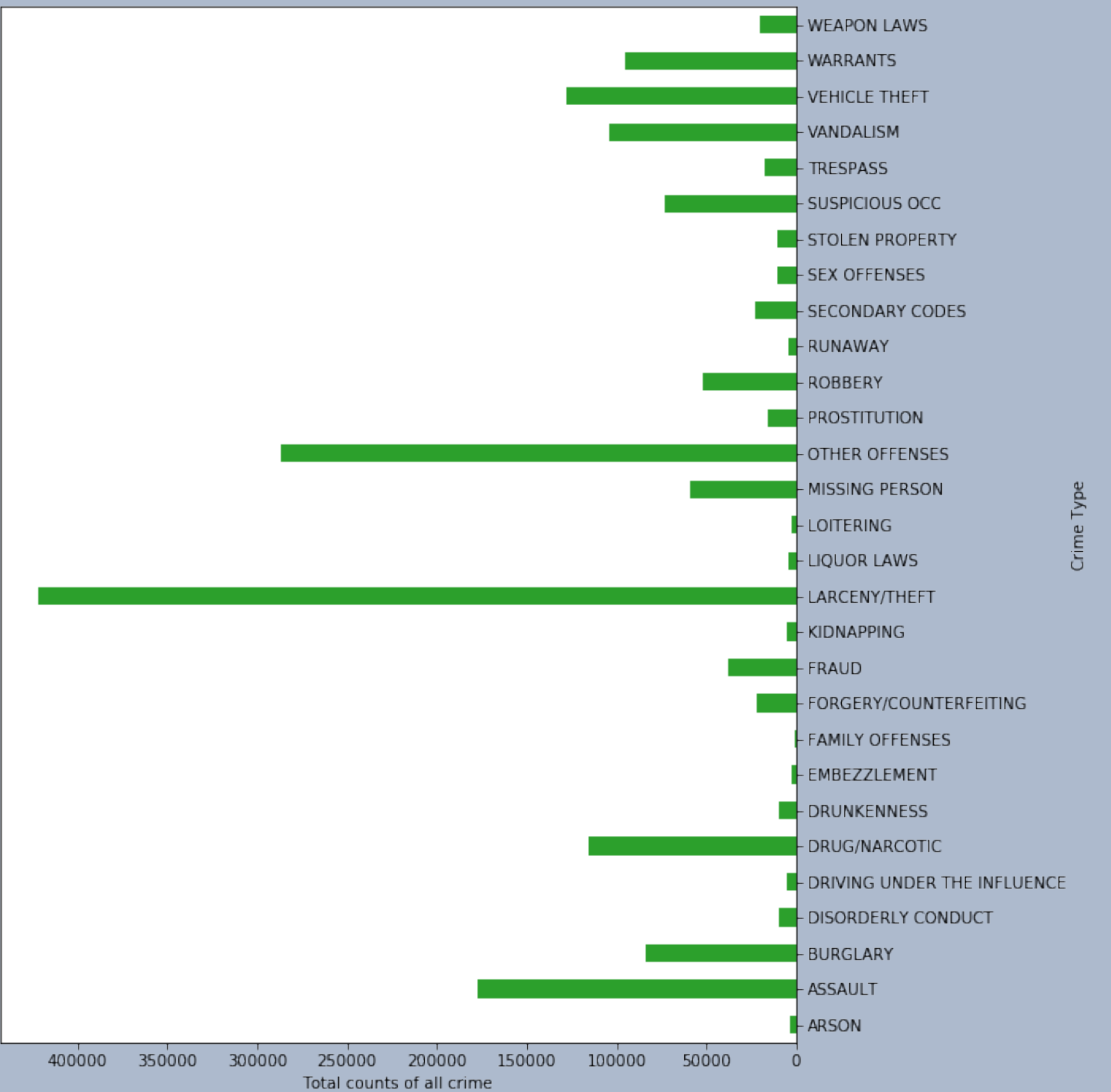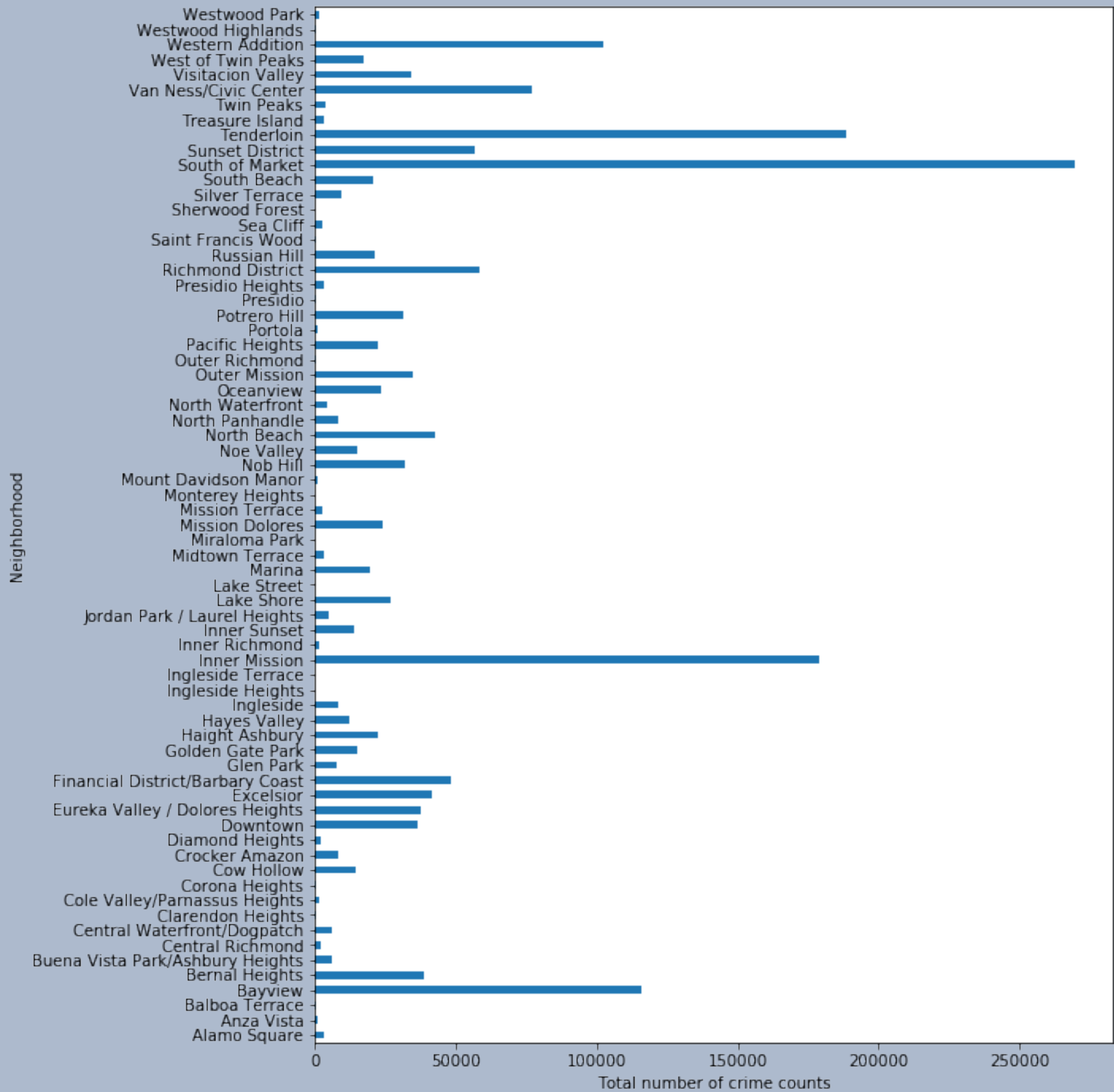2. Google Maps API
3. SF Opendata

Introduction

- ❖ 1.8 Million Crime records from Jan 2003 to Jan 2017

- ❖ ~ 130k / year, ~ 350 / day

- ❖ 860k Population on 2015[1]
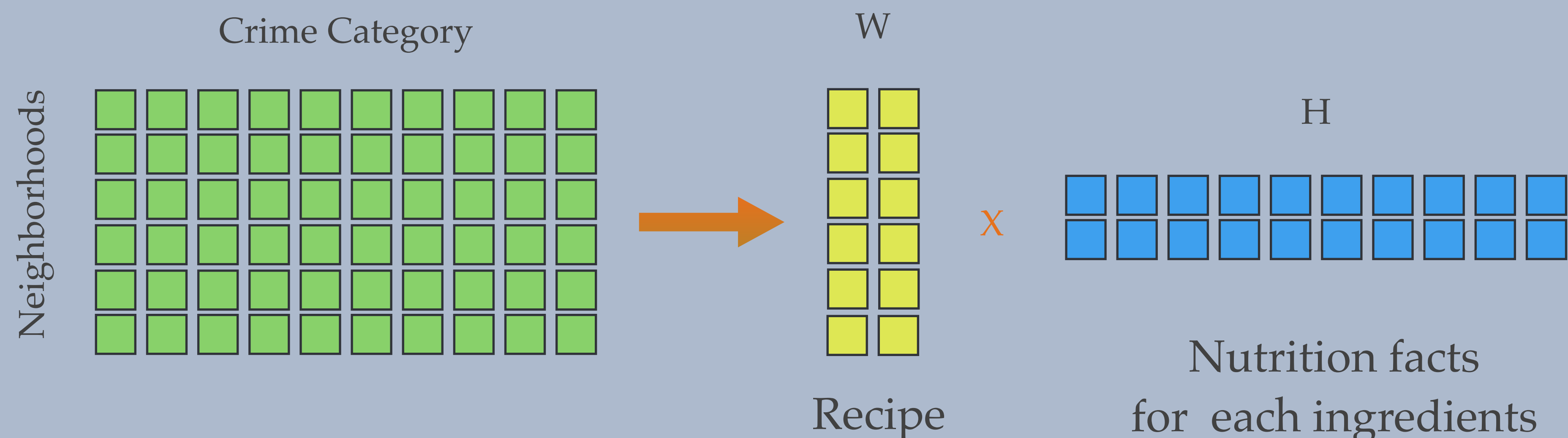
- ❖ 87 neighborhoods, 29 types Crime

⬇

Lots of Data!
Lots of Information!

1. wikipedia

Crime and Neighborhood

Non Negative Matrix Factorization

Crime Category

Neighborhoods

W

X

H

Recipe

Nutrition facts
for each ingredients

Nutrition Facts

Serving Size 6 pieces (30g)
Servings per container about 7
Calories 180
   Calories from Fat 120

Amount Per Serving	% Daily Value*	Amount Per Serving	% Daily Value*
Total Fat 13g	20%	**Total Carbohydrate** 14g	5%
Saturated Fat 8g	40%	Dietary Fiber 0g	0%
Trans Fat 0g		Sugars 12g	
Cholesterol 5mg	2%	**Protein** 2g	
Sodium 35mg	1%		

Vitamin A 0% • Vitamin C 0% • Calcium 6% • Iron 0%

*Percent Daily Values are based on a 2,000 calorie diet.
Your Daily Values may be higher or lower depending on your calorie needs.
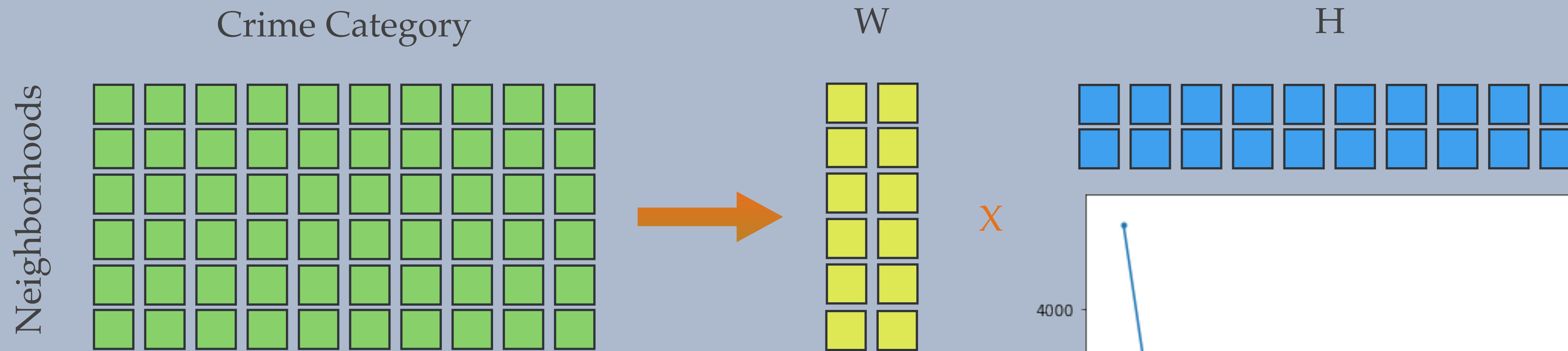
Pure Chocolate "Creamy Milk & White"
INGREDIENTS: SUGAR, COCOA BUTTER, WHOLE MILK POWDER, COCOA MASS, CONDENSED MILK
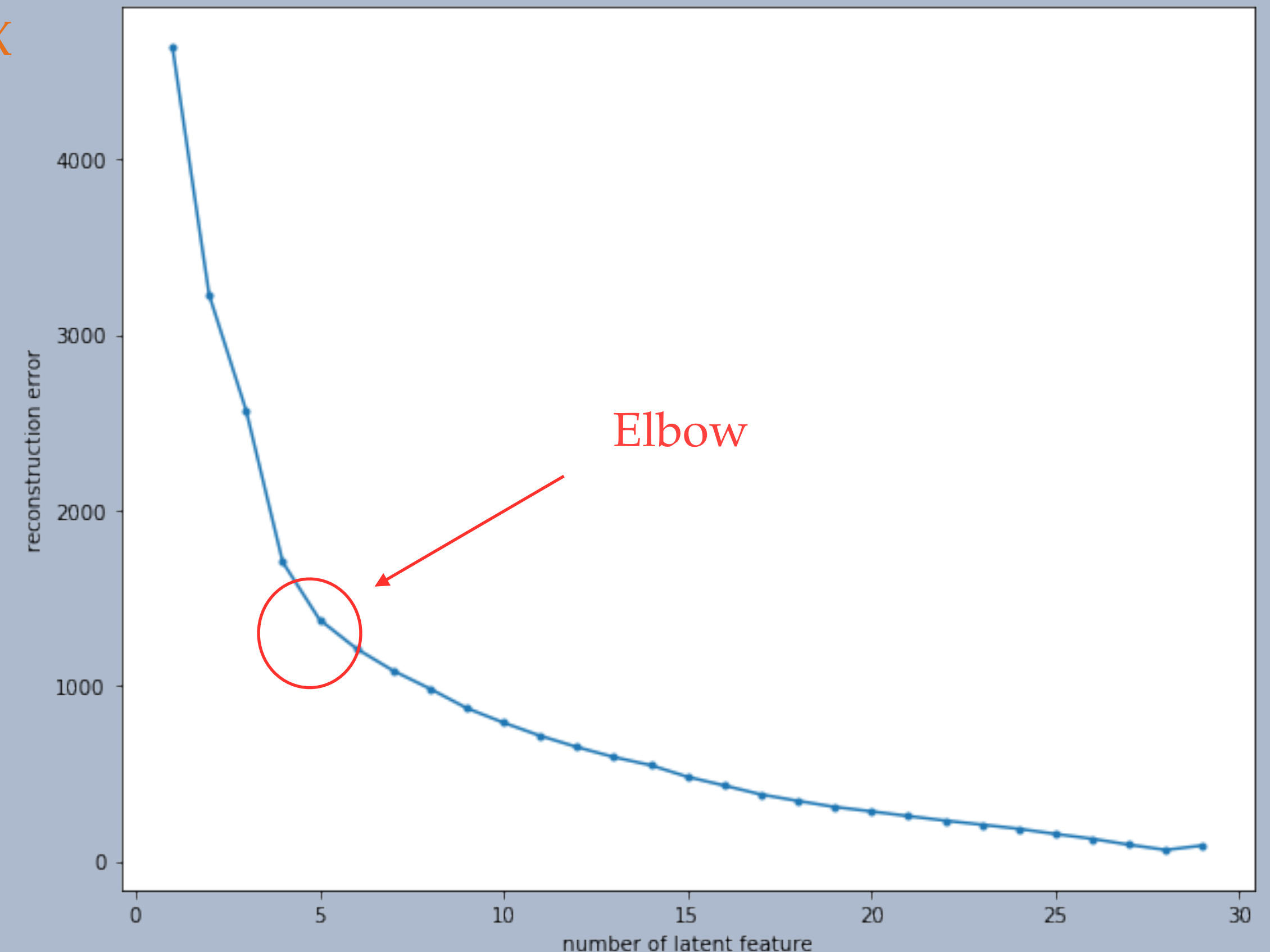POWDER, SKIM MILK POWDER, SOY LECITHIN, NATURAL AND ARTIFICIAL FLAVOR.
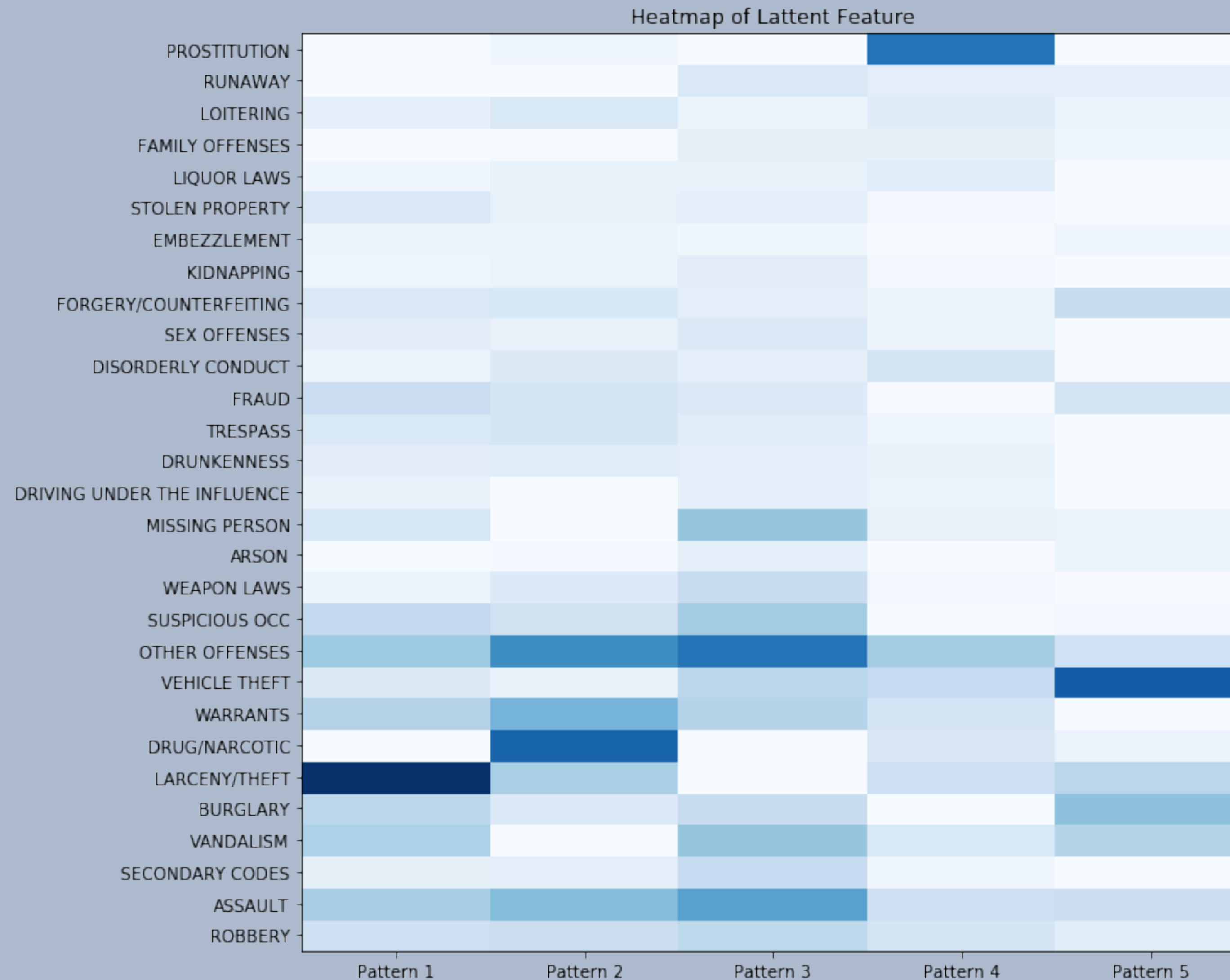**CONTAINS: MILK, SOY BEANS**

141222-D

Non Negative Matrix Factorization

Crime Category   W   H

Neighborhoods

X

❖ Find the ingredient by decompose to two matrix

❖ Optimize cost function $|X - WH|^2$

❖ Weight by inverse document frequency

$(0.5 + idf)*X$

❖ Elbow method

Elbow

reconstruction error

number of latent feature

Crime patterns

Heatmap of Lattent Feature

* Each crime appear in all class of pattern

* Each pattern emphasize on several type of crime

* Each neighborhood is linear combination of all patterns
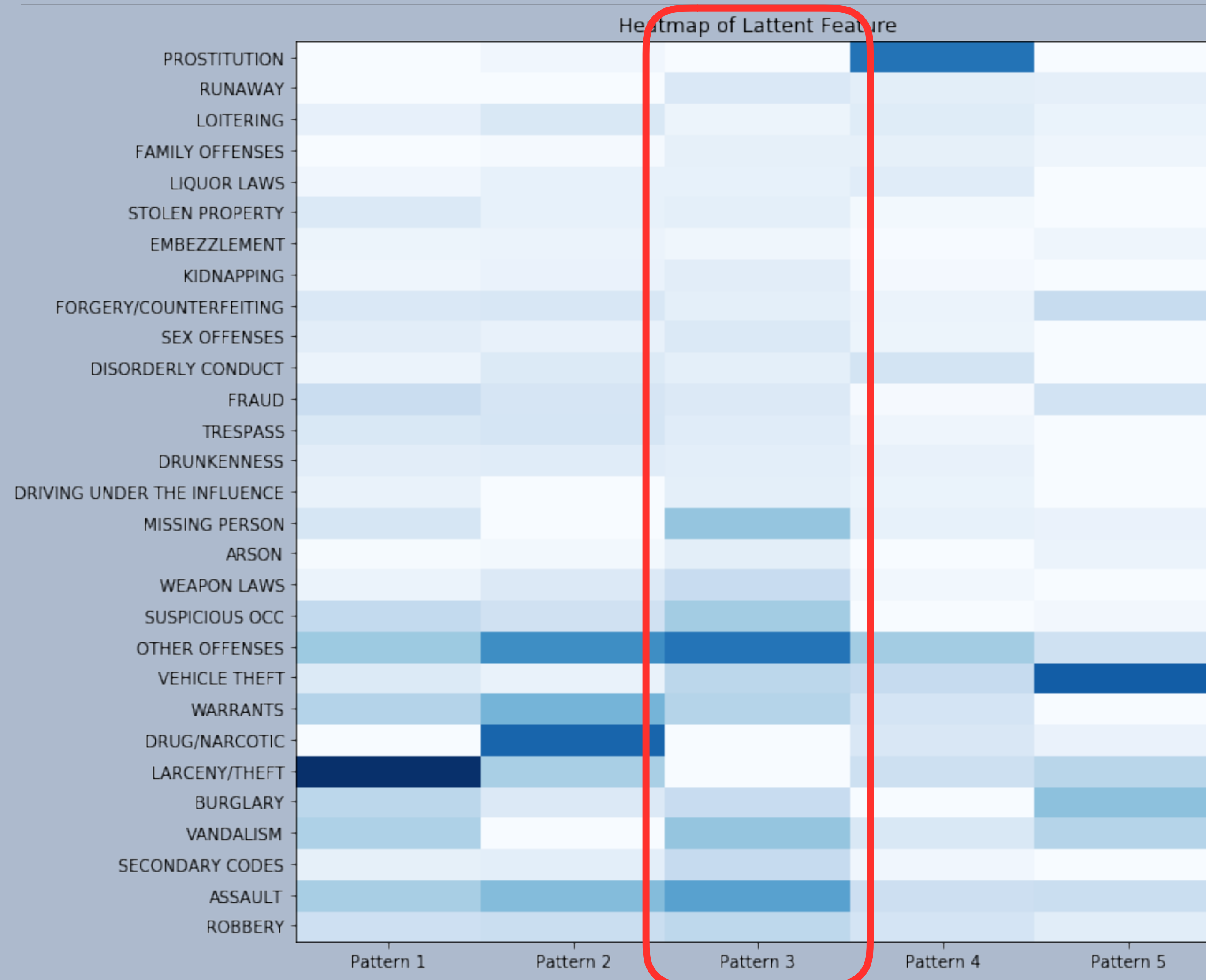
* Crime types happen on similar area

Crime patterns

Heatmap of Lattent Feature
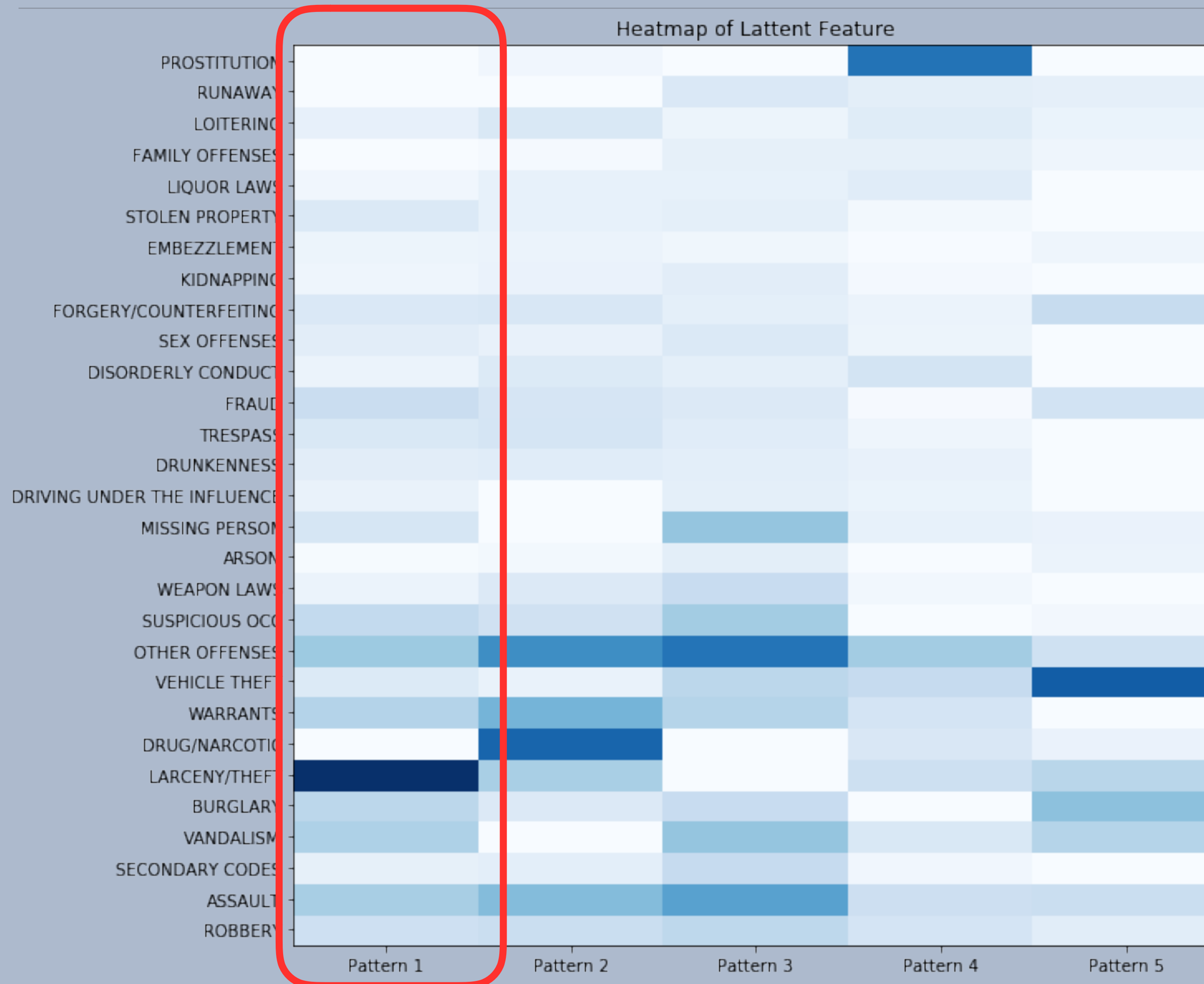
Crime pattern 1:
  Larceny / Theft
  Assault
  Other offenses
  Warrants
  Vandalism

Crime patterns

Heatmap of Lattent Feature
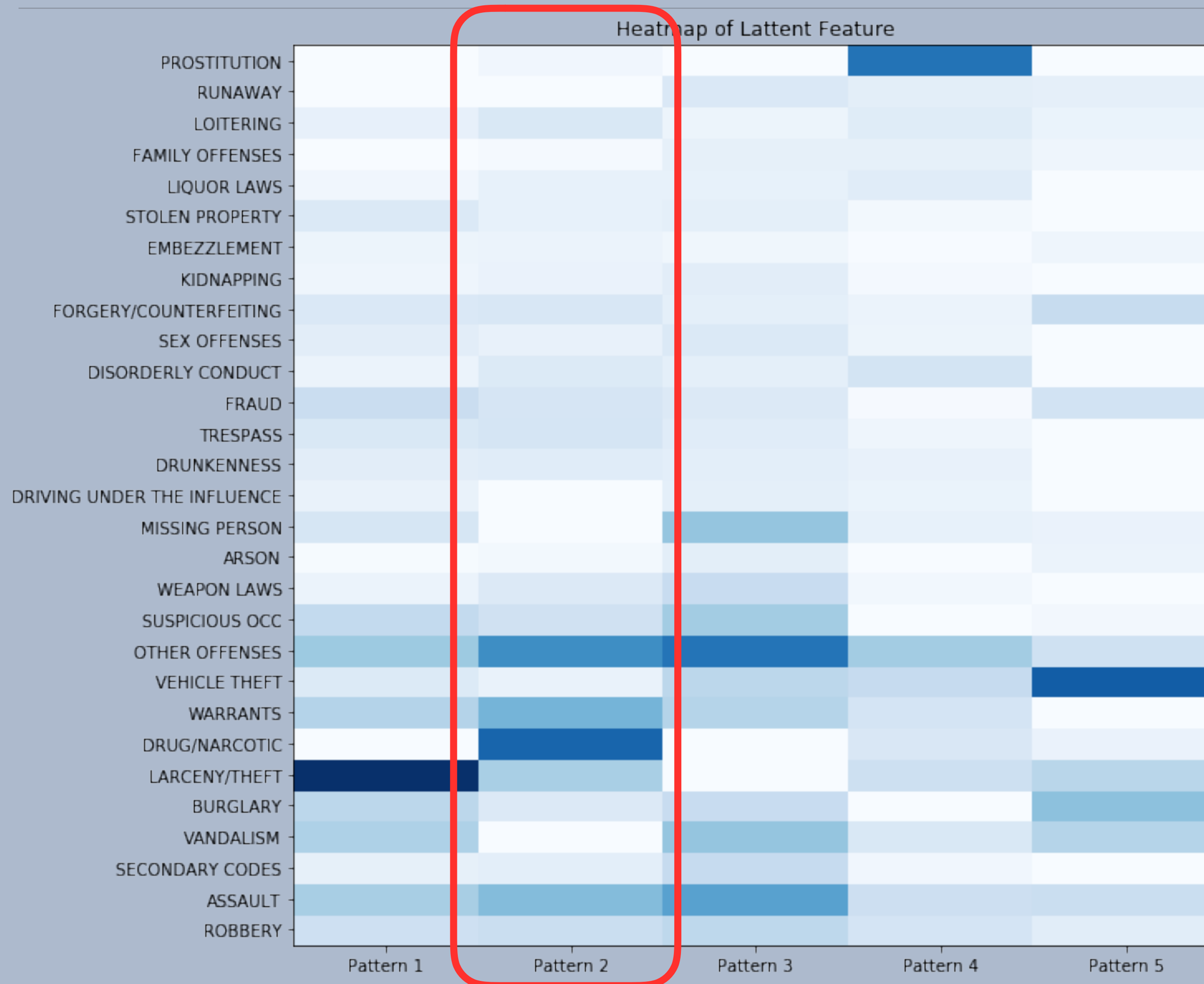
Crime pattern 2:
Drug / Narcotic
Other offenses
Warrants
Assault

Crime patterns

Heatmap of Lattent Feature

Crime pattern 3:
Other offenses
Missing person
Assault
Vandalism

Crime patterns

Heatmap of Lattent Feature
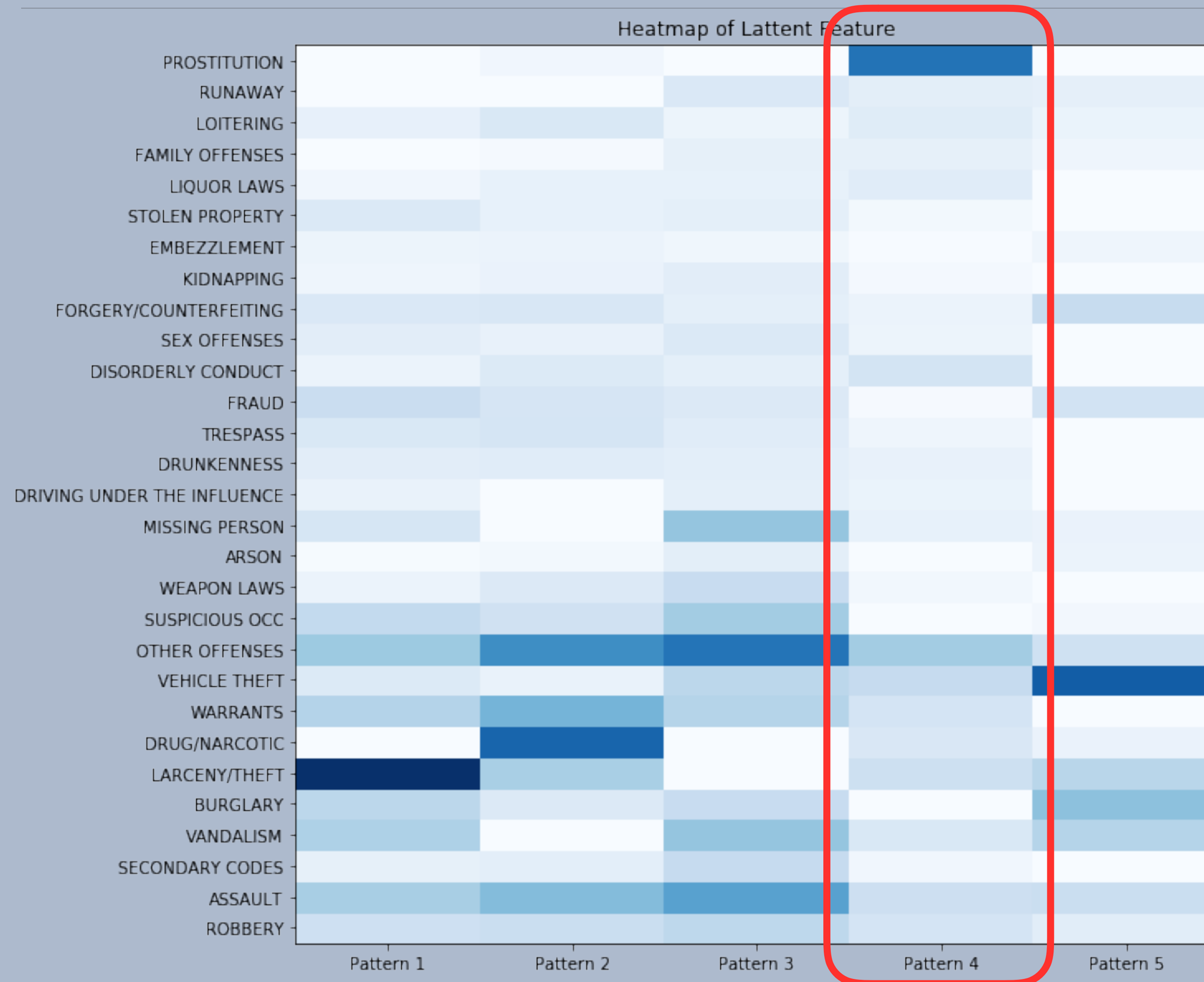
Crime pattern 4:
  Prostitution
  Disorderly conduct
  Other offenses

Crime patterns

Heatmap of Lattent Feature
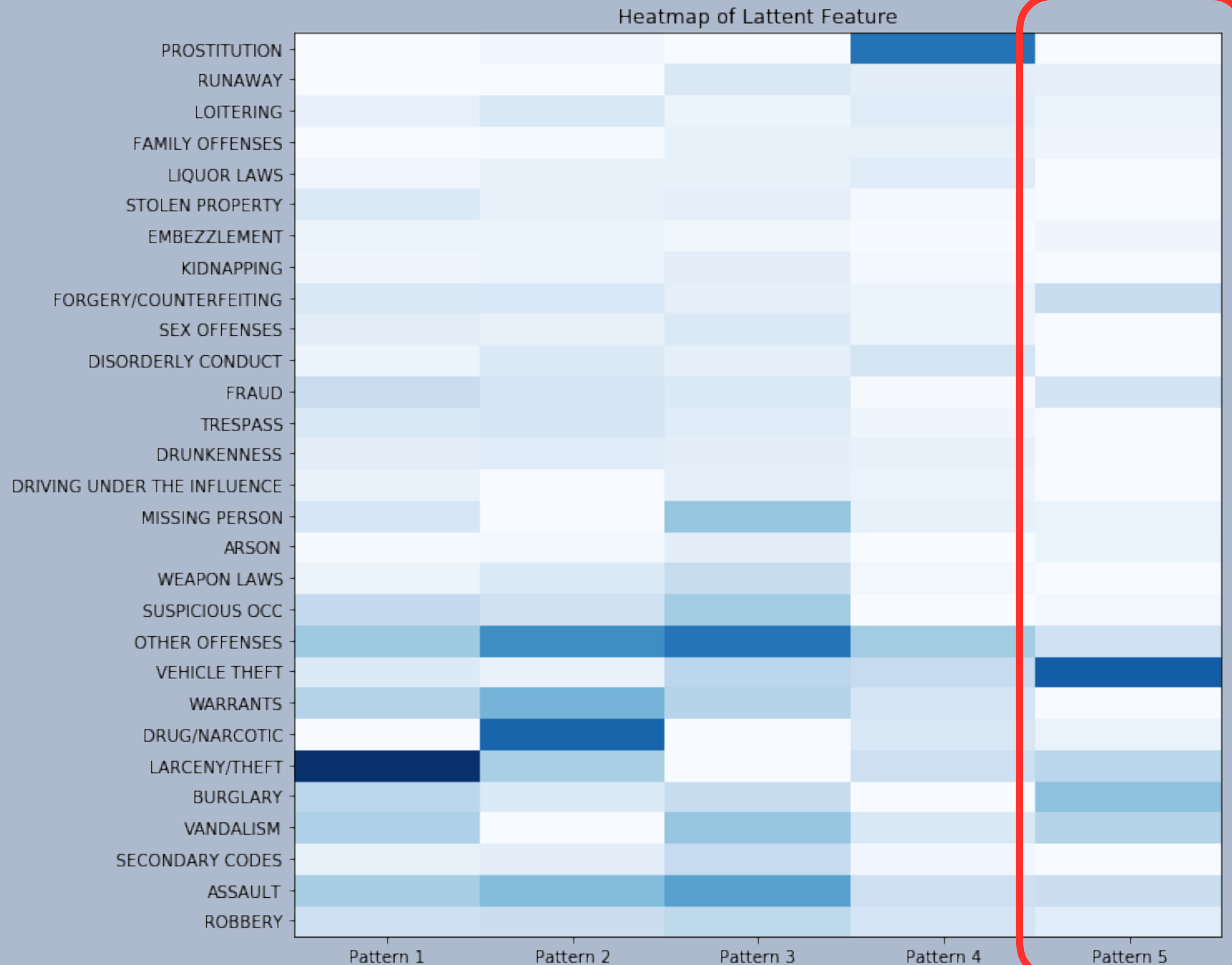
Crime pattern 5:
Vehicle theft
Burglary
Larceny / Theft
Vandalism

Crime Maps

San Franscisco Commuties Categoried by Crime -- 2003

pattern 1: Theft & Burglary
pattern 2: Drug related
pattern 3: Minor crime & Assaut
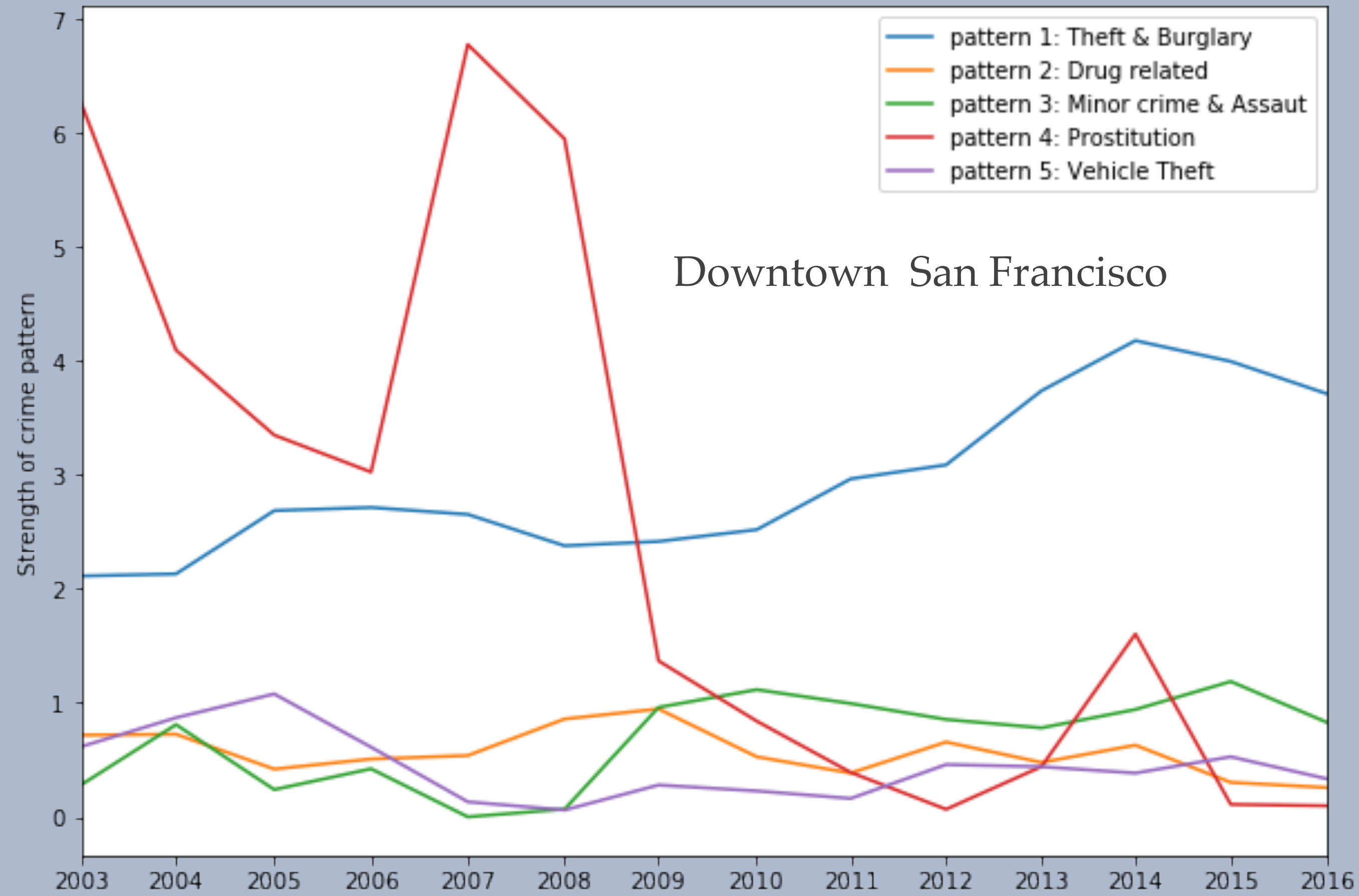pattern 4: Prostitution
pattern 5: Vehicle Theft
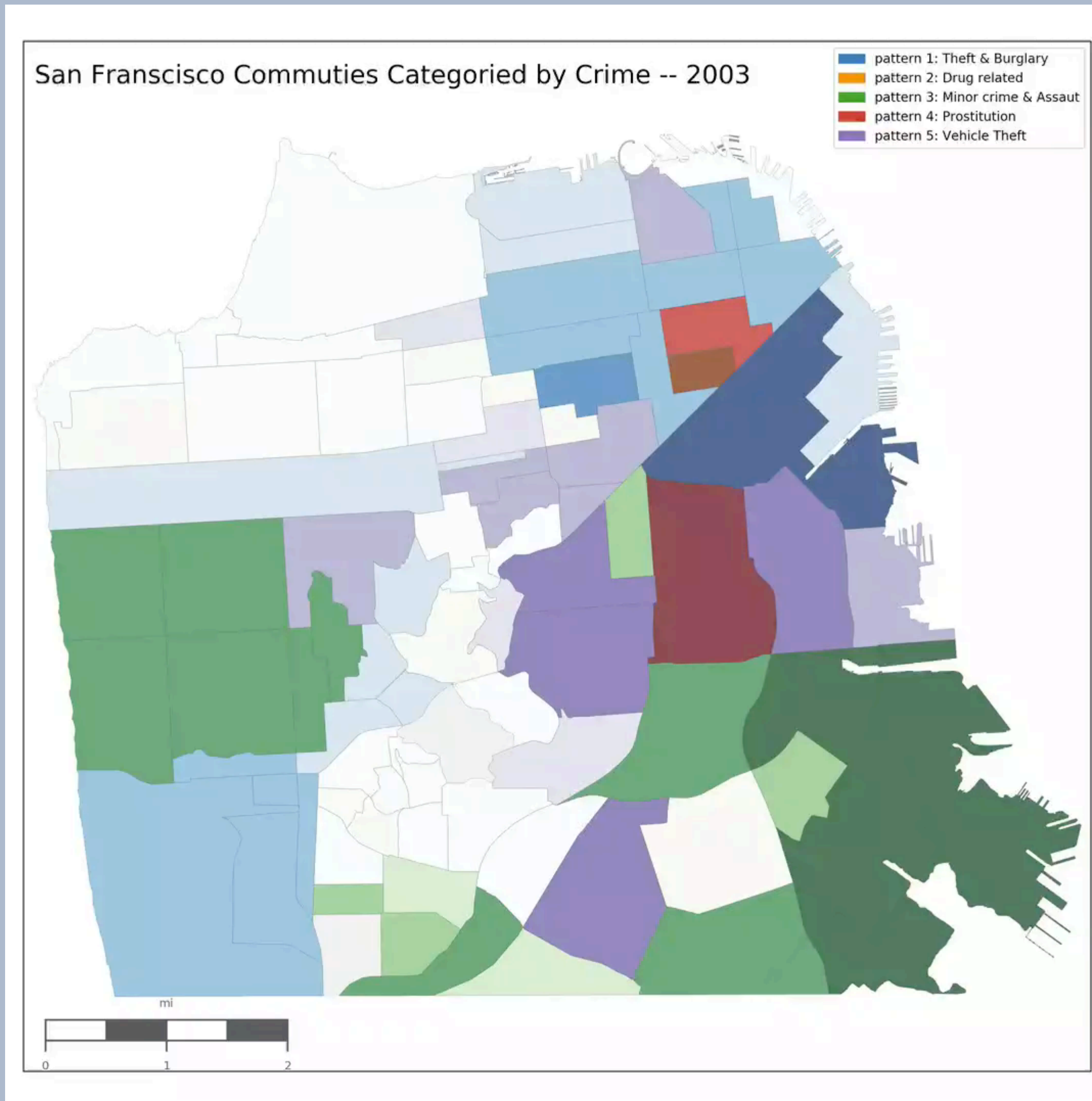
❖ Cluster to most representing pattern

❖ Color represent the class and severity of crime

General situation in SF

Details in Neighborhood
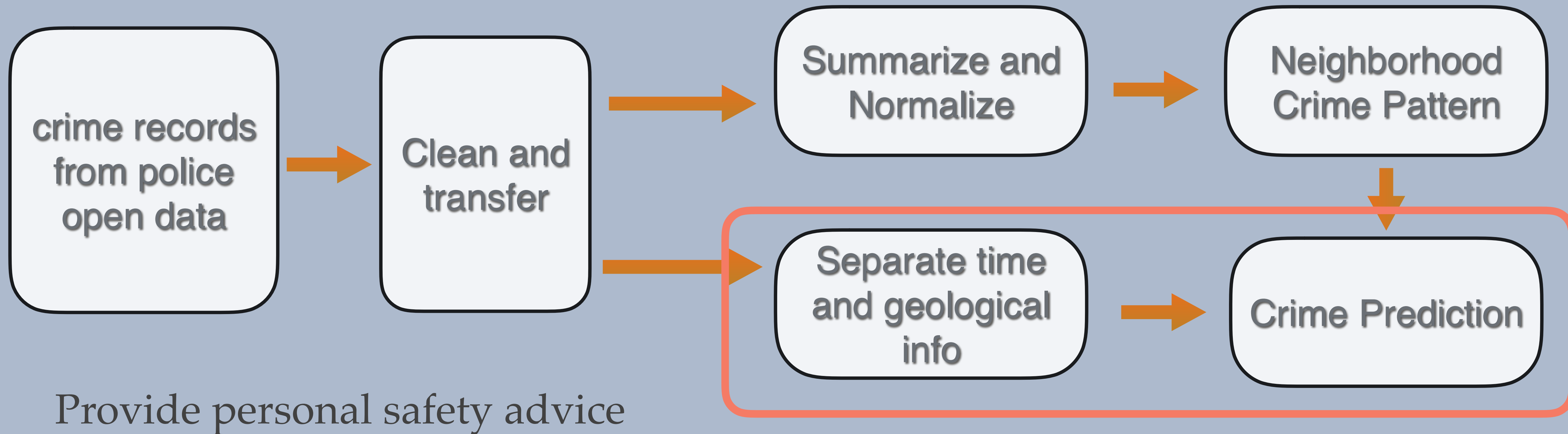

Downtown San Francisco
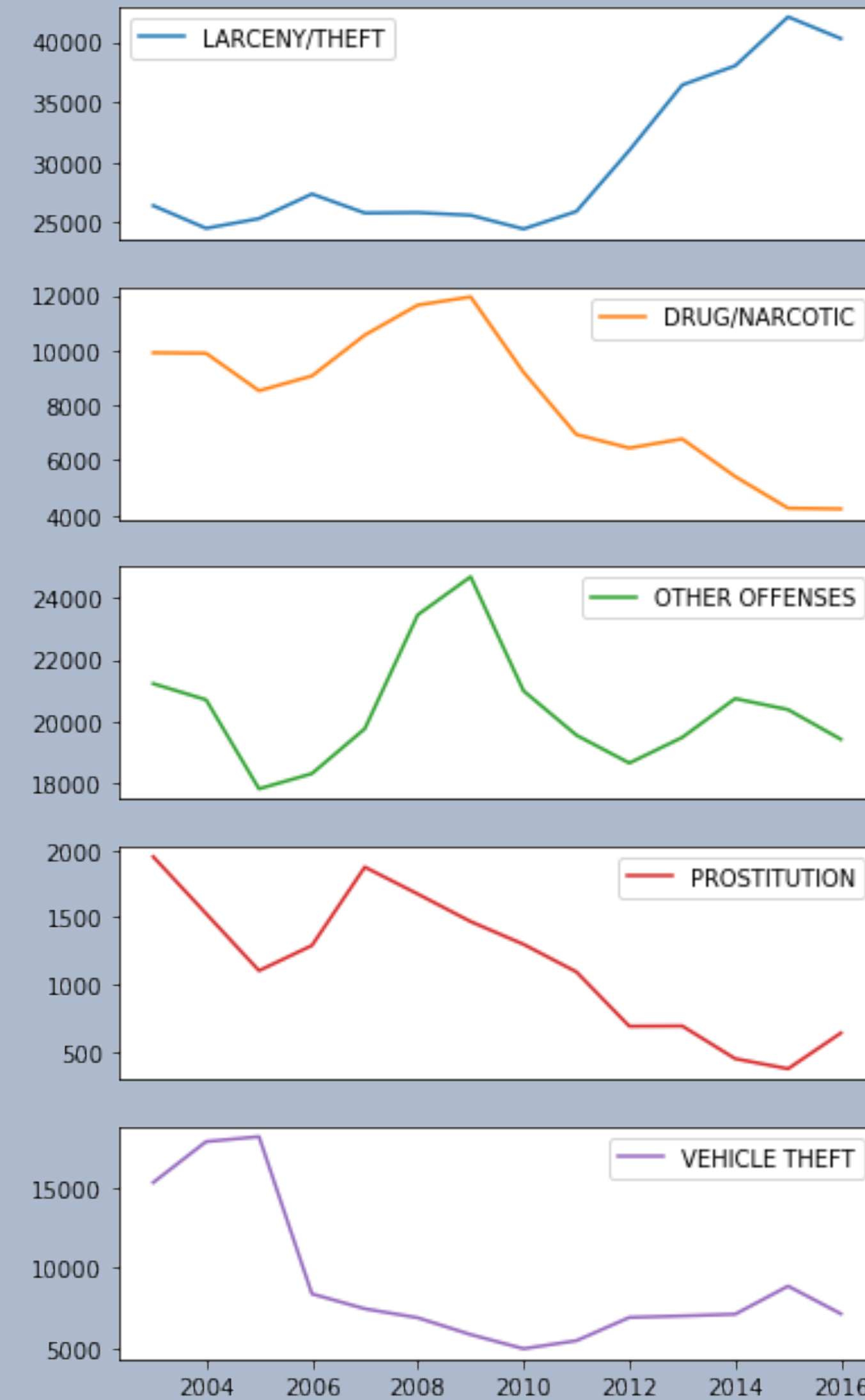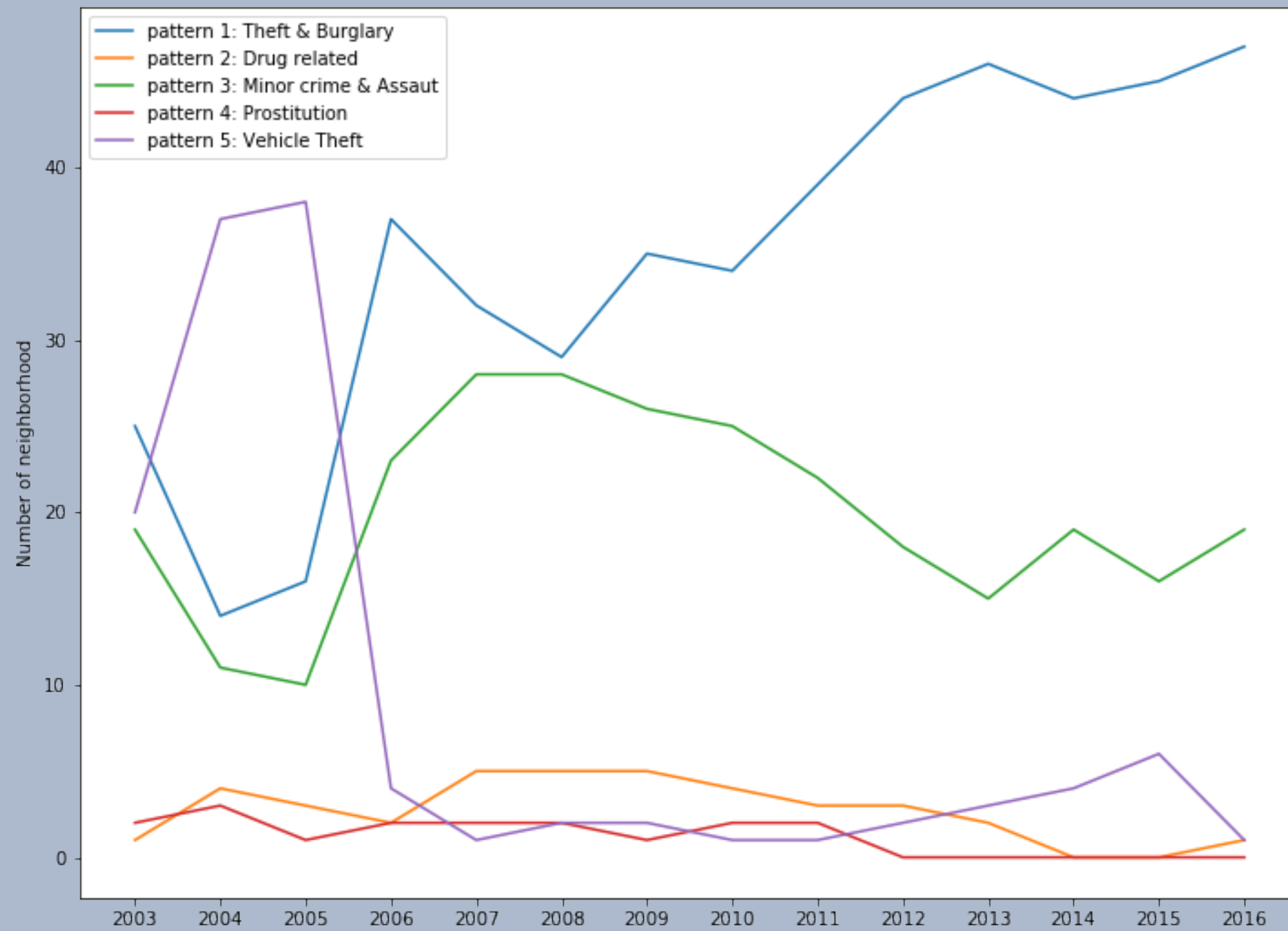
Each neighborhood is linear combination of all crime pattern

Data Flow

crime records from police open data → Clean and transfer → Summarize and Normalize → Neighborhood Crime Pattern

Separate time and geological info → Crime Prediction

Provide personal safety advice
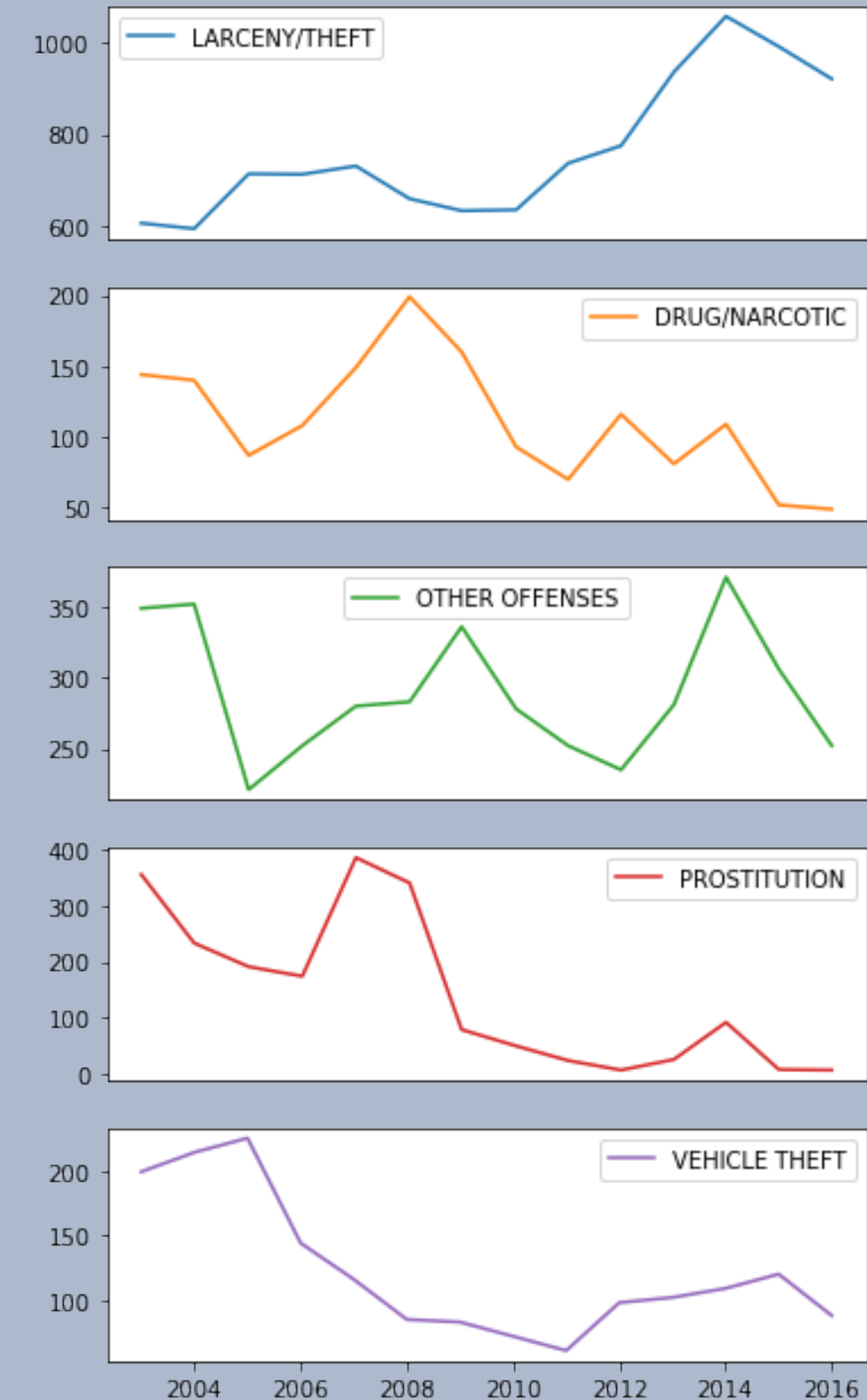
1. https://publicsafetydataportal.org/
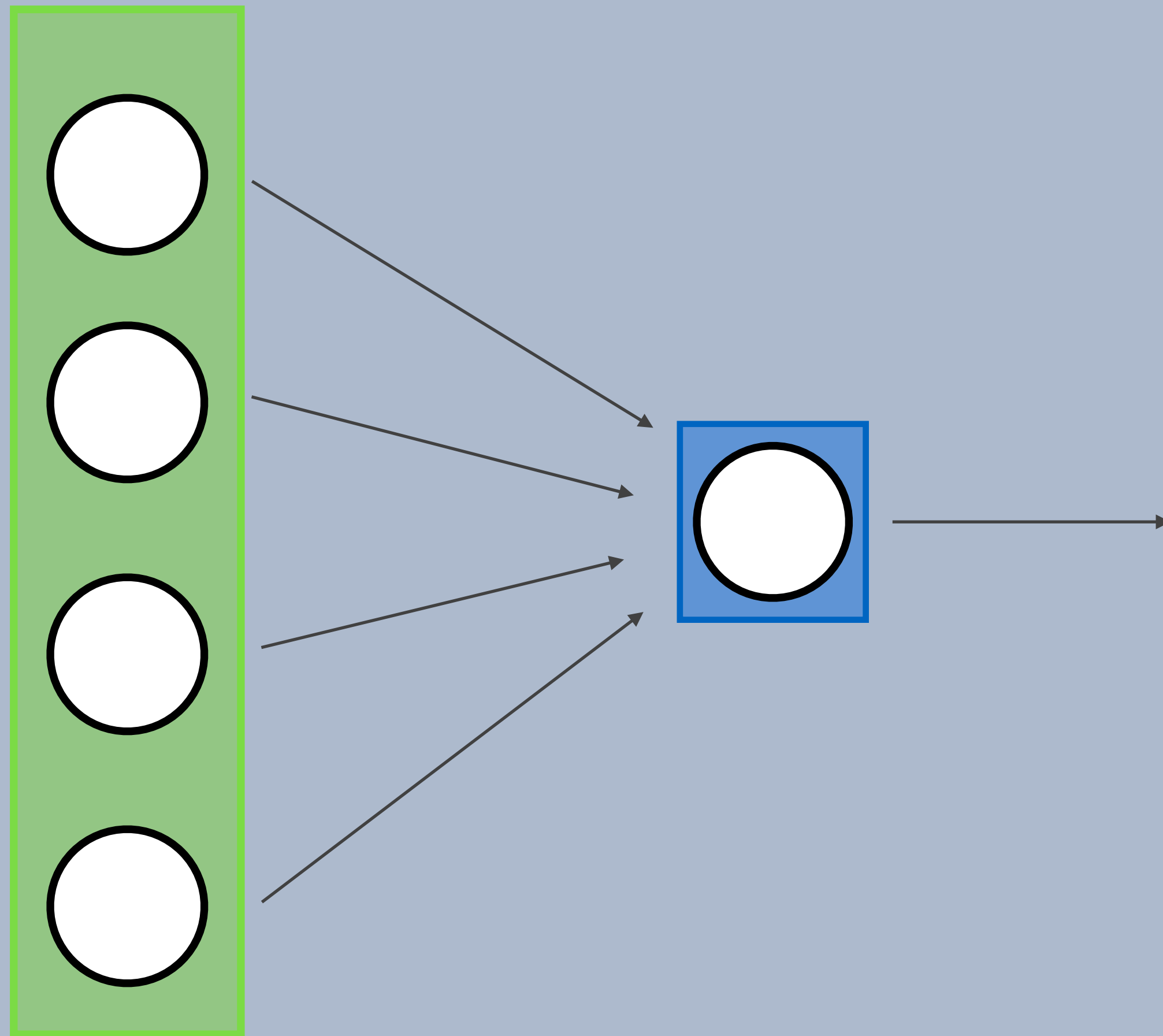2. Google Maps API
3. SF Opendata

Crime prediction
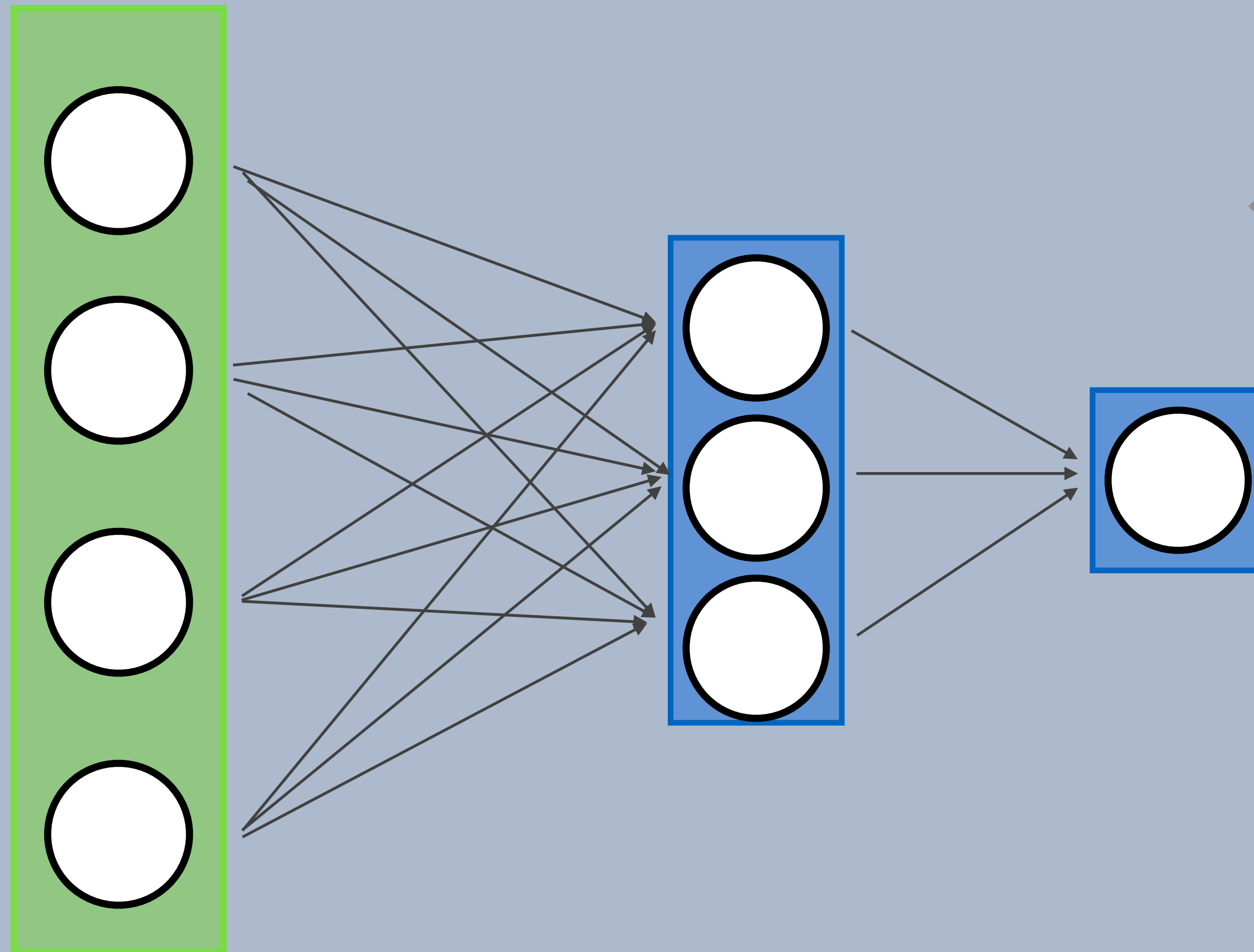
❖ Predict crime type based on the assumption of crime will happen

❖ Location, time and neighborhood crime pattern as input feature

❖ Advantage of neural networks:

    1. non-linearity

    2. easy for multiple classification

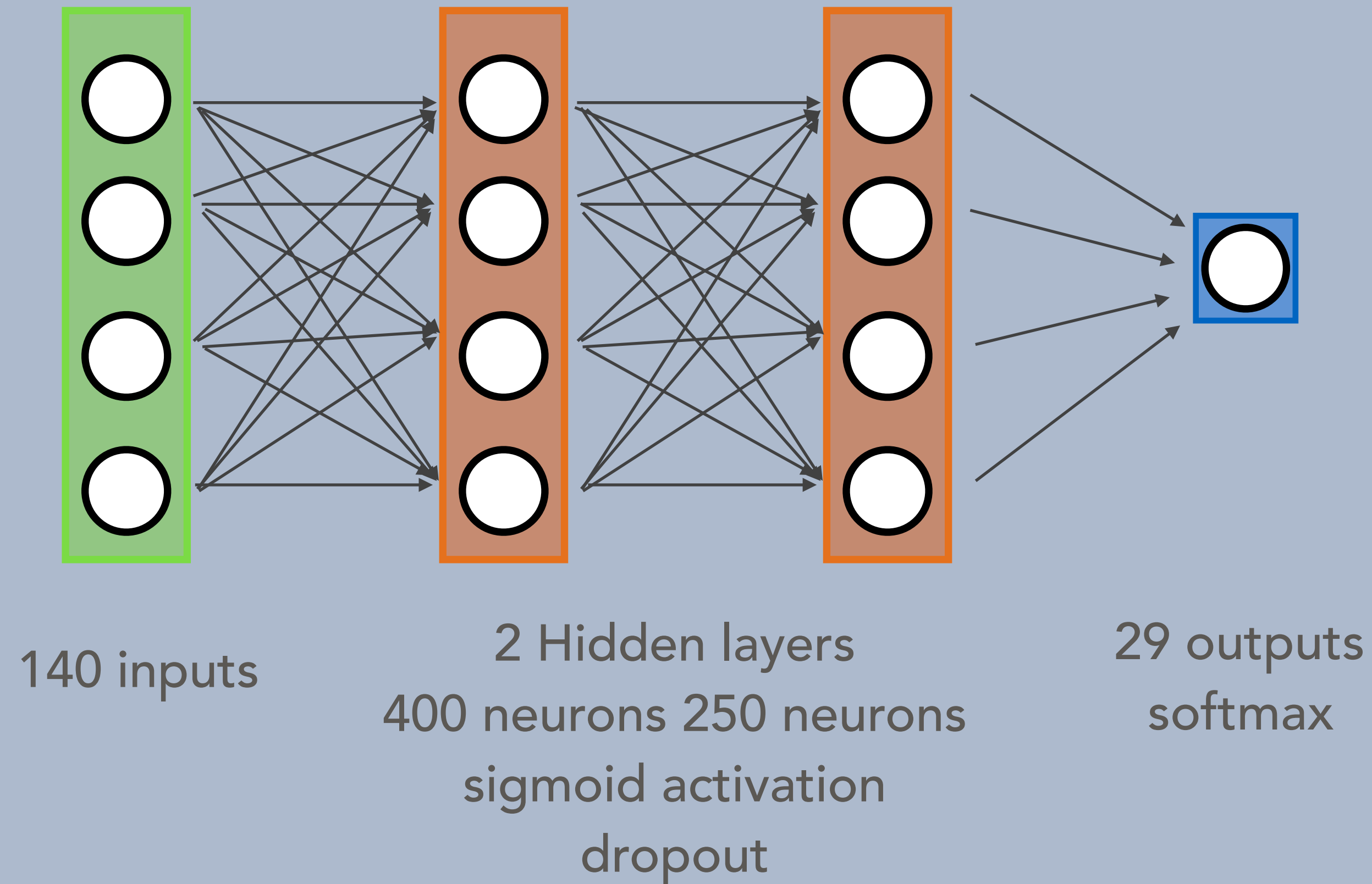   3. online learning ability and parallelization by GPU

Neural Networks

- ❖ Neuron: transfer input by activation function
- ❖ Logistic regression

Neural Networks

❖ Add neuron will add non-linearity

Neural Networks

140 inputs

2 Hidden layers
400 neurons 250 neurons
sigmoid activation
dropout

29 outputs
softmax

Time, Location, past clustering ➡ likelihood of crime type ➡ Top k crime type

~ 24% top 3 accuracy by 10 fold cross validation

Summary

❖ use non-negative matrix factorization

   find the crime pattern in neighborhood

   clustered neighborhoods

   neighborhoods crime evolution

❖ predict crime type by neural networks

Future works

❖ Find a better way to weight low count but high impact crimes

❖ Build map into interactive fashion, easily to find details in neighborhood

❖ Looking for more related information (e.g. House Price, Weather, Personal)

Thank You!
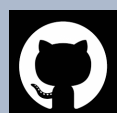
github.com/livenb

zclivenb@gmail.com

linkedin.com/in/james-zyc/