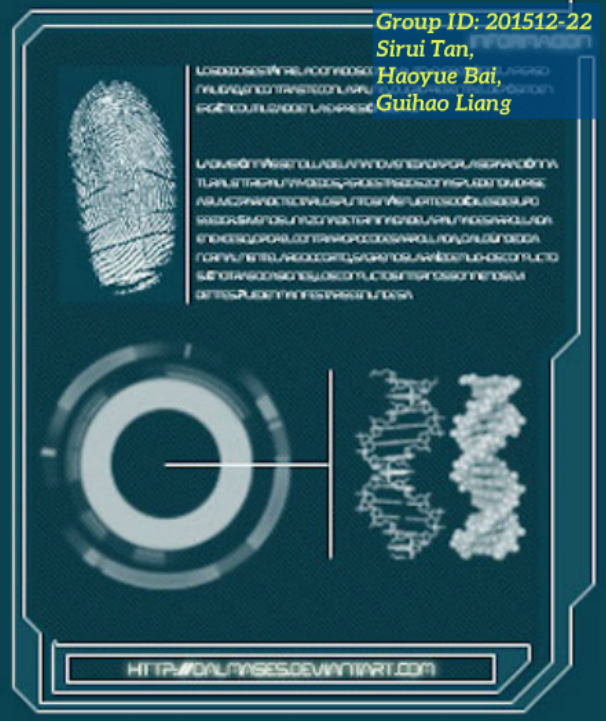
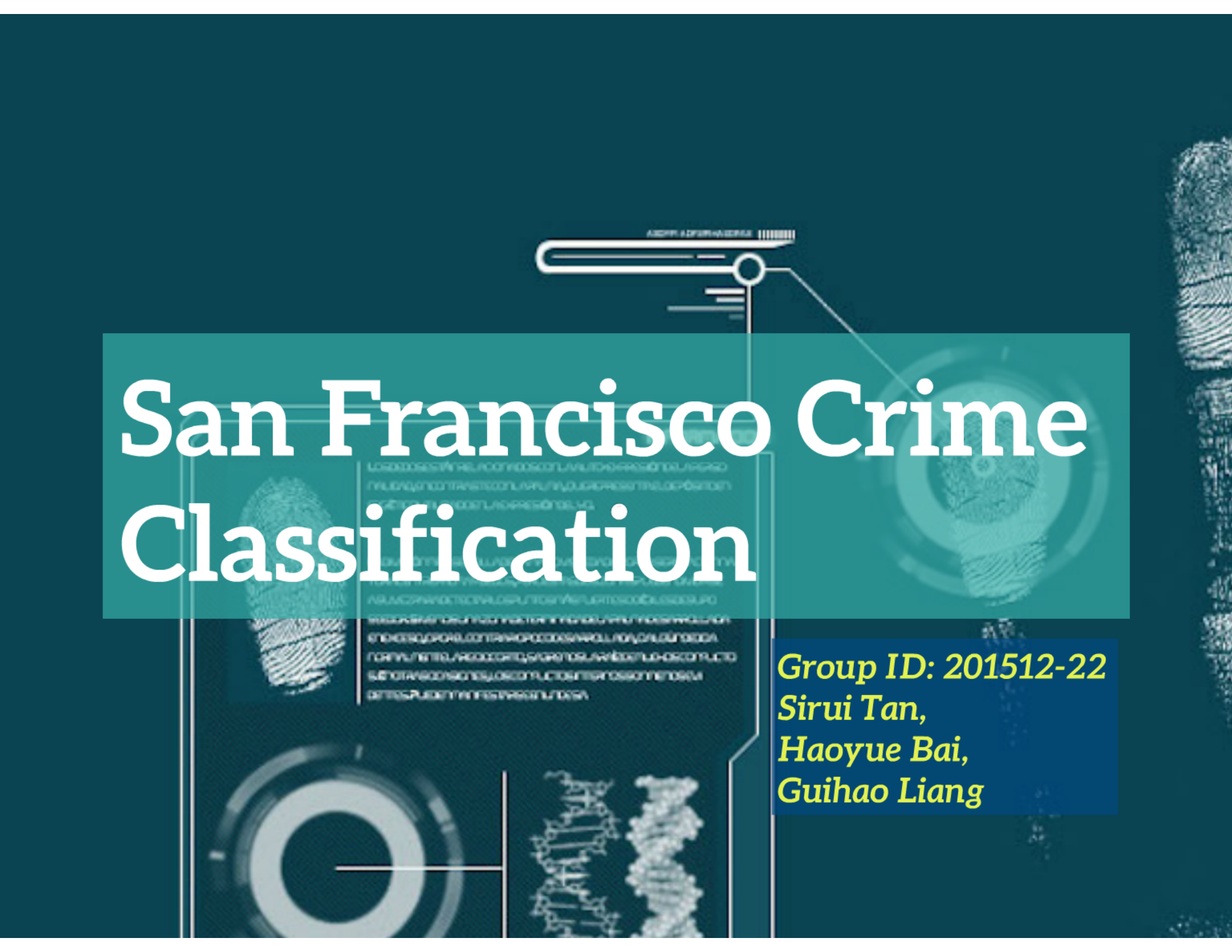
San Francisco Crime Classification

Group ID: 201512-22
Sirui Tan,
Haoyue Bai,
Guihao Liang





San Francisco Crime Classification

Group ID: 201512-22
Sirui Tan,
Haoyue Bai,
Guihao Liang

Dataset

2003-01-07 07:52:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	5TH ST / SHIPLEY ST	-122.402843	37.779829
2003-01-07 04:49:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Tuesday	TENDERLOIN	ARREST, BOOKED	CYRIL MAGNIN STORTH ST / EDDY ST	-122.408495	37.784452
2003-01-07 03:52:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	OFARRELL ST / LARKIN ST	-122.417904	37.785167
2003-01-07 03:34:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	DIVISADERO ST / LOMBARD ST	-122.442650	37.798999
2003-01-07 01:22:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	900 Block of MARKET ST	-122.409537	37.782691
2003-01-06 23:30:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	BAYVIEW	ARREST, BOOKED	REVERE AV / INGALLS ST	-122.384557	37.728487
2003-01-06 23:14:00	WARRANTS	WARRANT ARREST	Monday	CENTRAL	ARREST, BOOKED	BUSH ST / HYDE ST	-122.417019	37.789110
2003-01-06 22:45:00	WARRANTS	WARRANT ARREST	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:45:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:19:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	GEARY ST / POLK ST	-122.419740	37.785893
2003-01-06 21:54:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	SUTTER ST / POLK ST	-122.420120	37.787757

Visualization of Dataset

- Topographic Maps : 39
- Year Distribution of Crimes (topographic maps) 12 x 39
- Month Distribution of Crimes (histogram) 12 x 39
- Hour Distribution of Crimes (histogram) 12 x 39

Visualization of Dataset

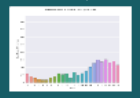
Topographic Map

Visualization of Dataset

Month Distribution of Crimes

Visualization of Dataset

Hour Distribution of Crimes

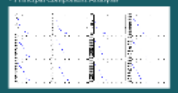


Feature Extraction Strategy

- Group by Year
- Temporal Features - Month, Date, Hour
- Geographical Features - Longitude, Latitude, Police Department
- Address Features - Log Odds Ratio

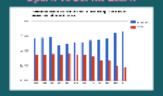
Feature Selection

- Principal Component Analysis



- The First 7 Features Selected

Spark vs Sci-kit Learn



- Limited Maximum Depth of Decision Trees
- Not Dedicated to Designed for Machine Learning

Evolution of Classification Strategy

Summary

- Empirical Interactions Between the Crime Incidence

- Decision Tree Outperforms SVM - Crimes related to both geographically and temporally
- Insights from PCA - Address Features are most significant than time correlated features

Visualization of Dataset

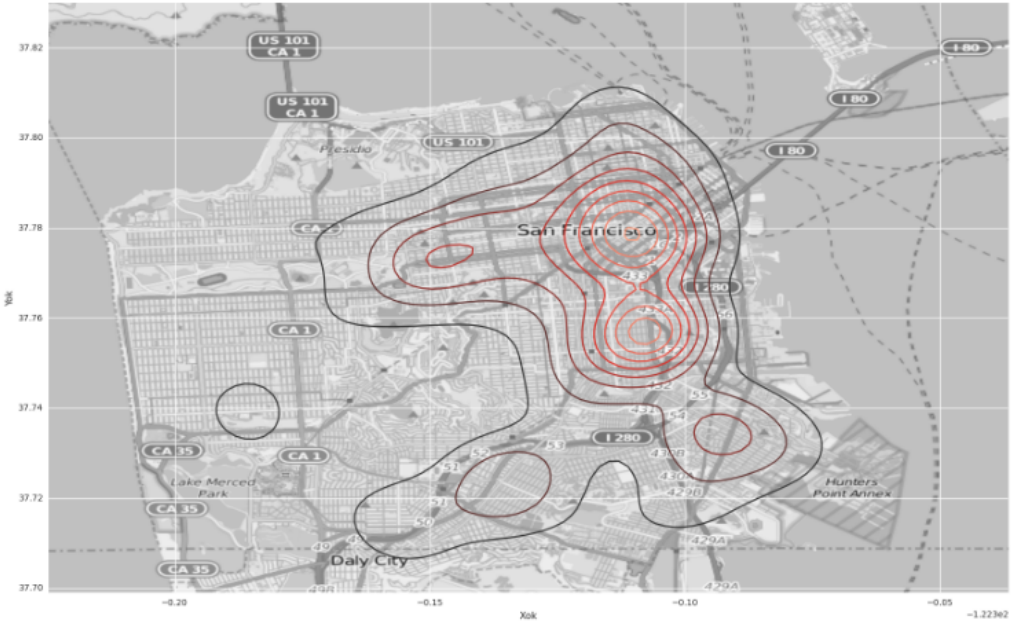
- Topographic Maps : 39
- Year Distribution of Crimes (topographic maps): 12 x 39
- Month Distribution of Crimes (histogram): 12 x 39
- Hour Distribution of Crimes (histogram): 12 x 39

Visualization of Dataset

Topographic Map

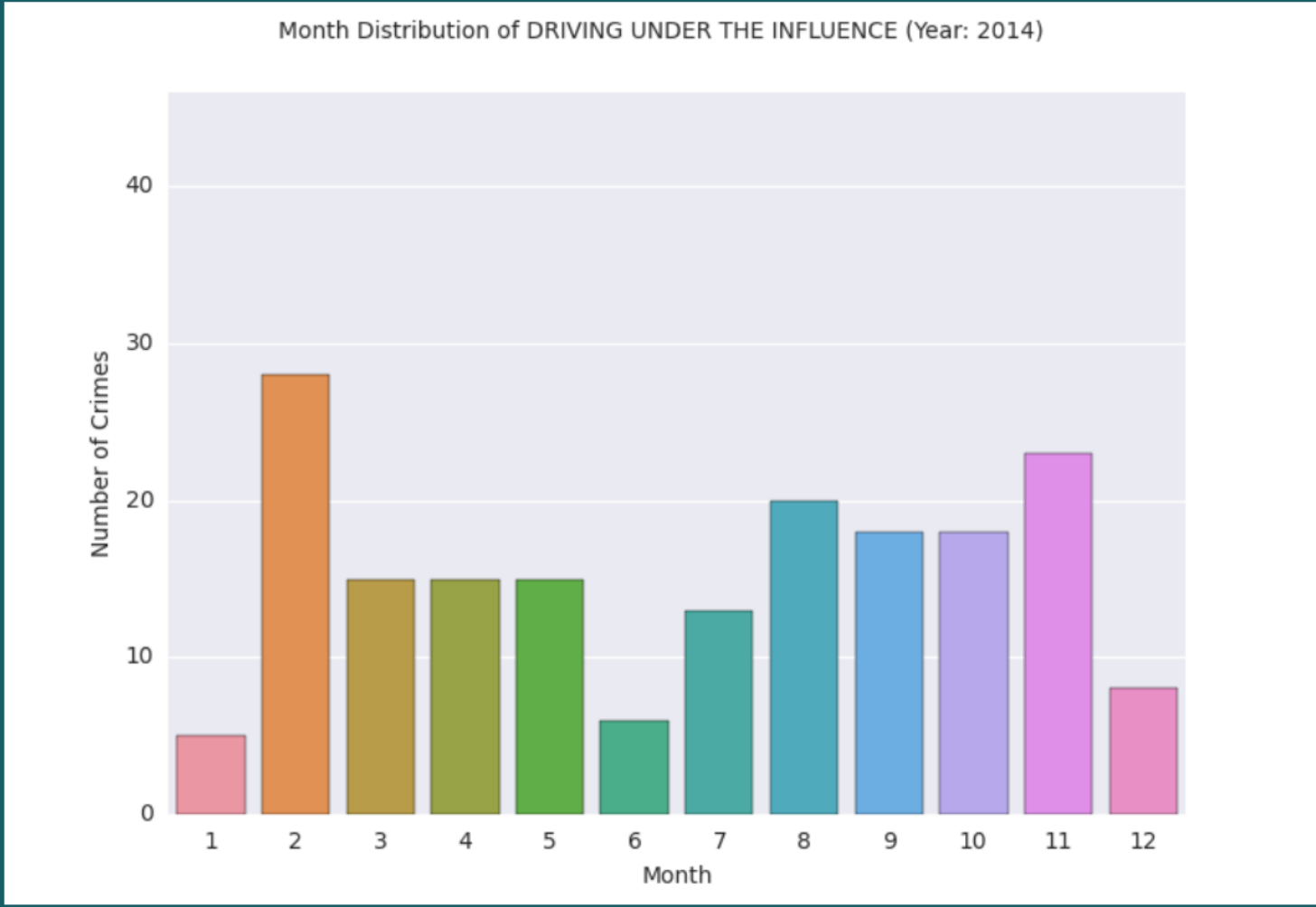
Topographic Distribution of DRUG/NARCOTIC (2003 - 2014)

Topographic Distribution of MISSING PERSON (2003 - 2014)



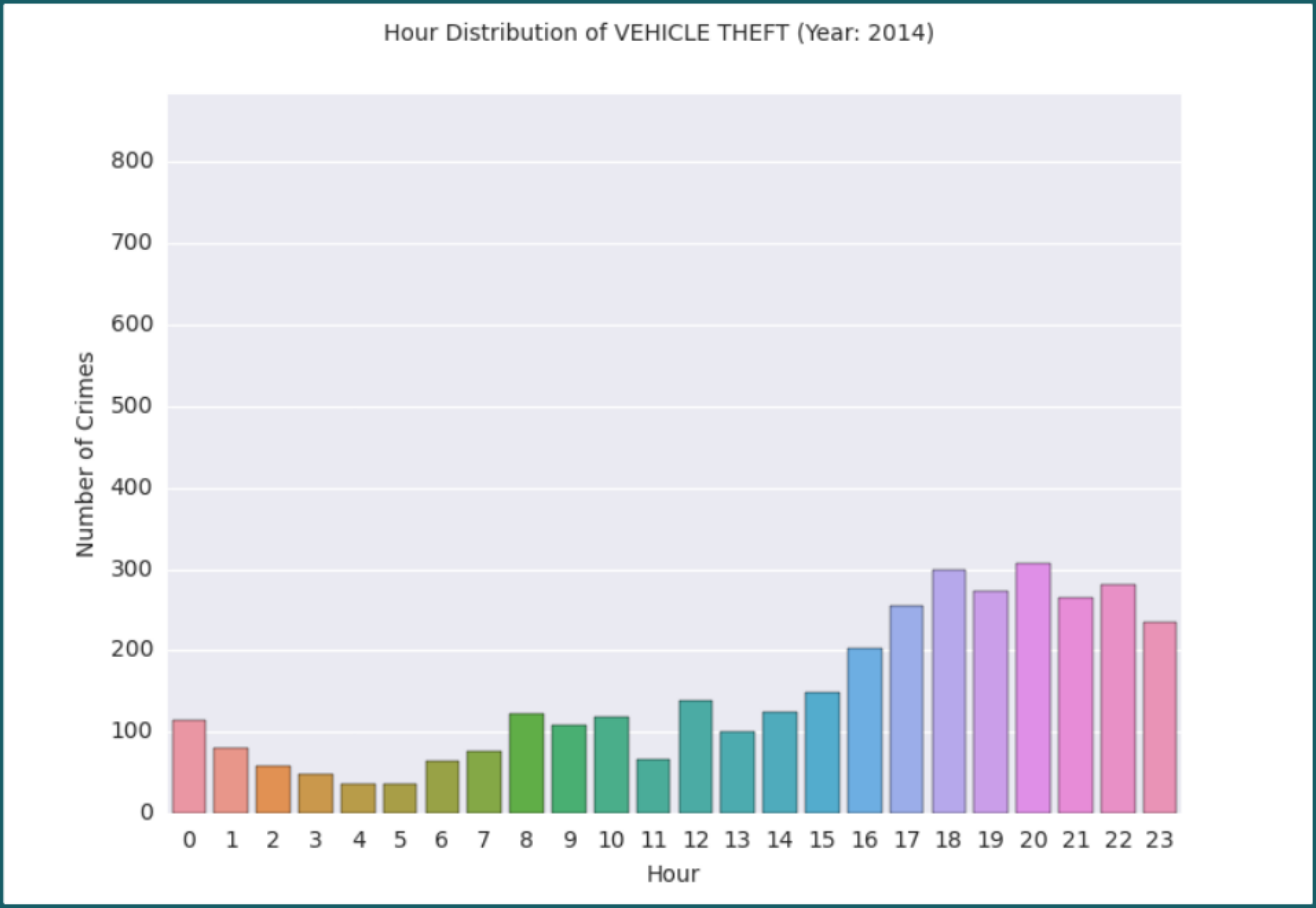
Visualization of Dataset

Month Distribution of Crimes



Visualization of Dataset

Hour Distribution of Crimes

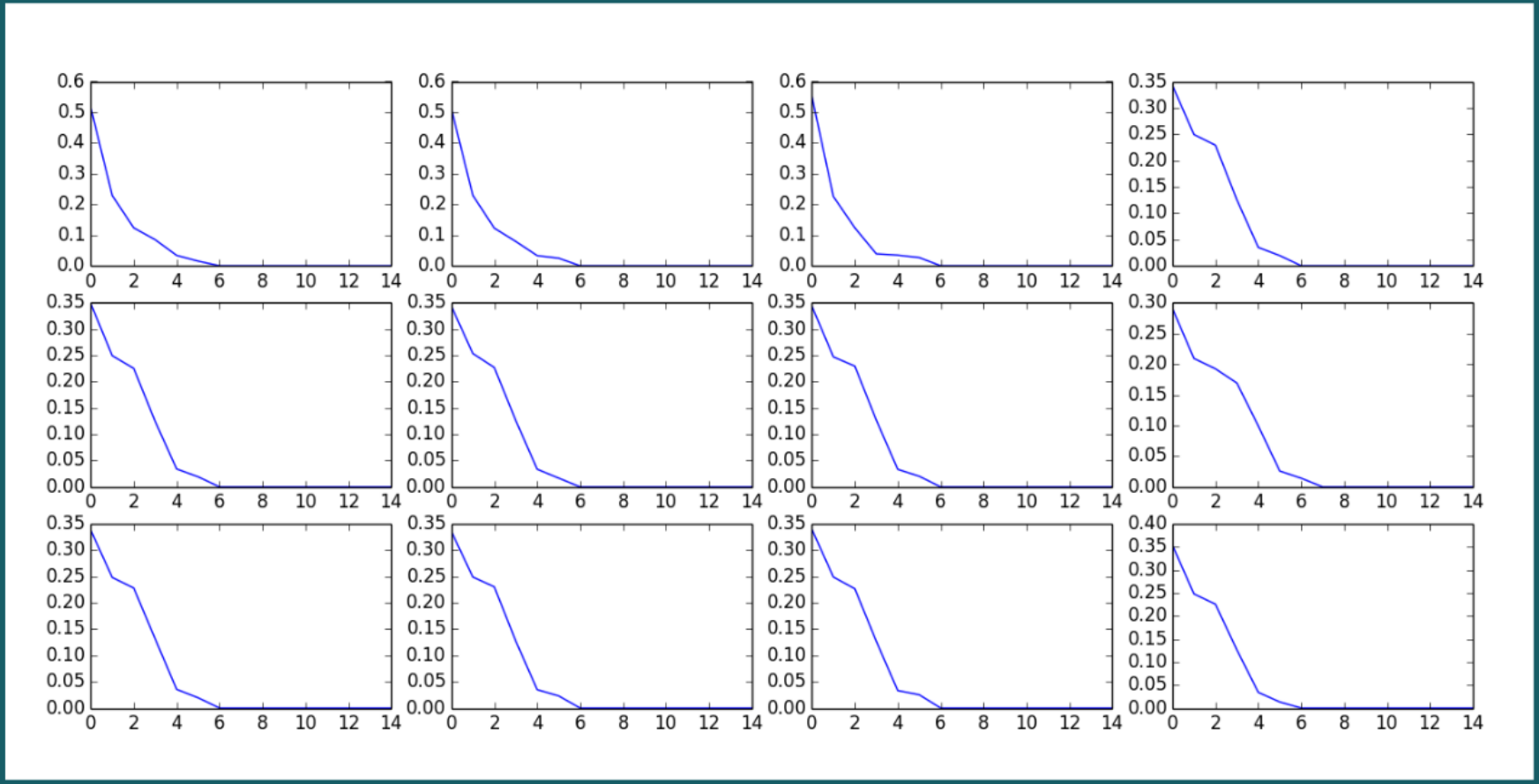


Feature Extraction Strategy

- Group by Year
- Temporal Features - Month, Day, Hour
- Geographical Features - Longitude/Latitude, Police Department
- Address Features - Log Odds Ratio

Feature Selection

- Principal Component Analysis



- The First 7 Features Selected

ANN

- Feature Space Relatively Small
- Non-trivial Computation Complexity

SVM

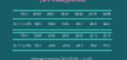
- Inappropriate for Tangled Features
- Hard to Converge
- Time Consuming

Random Forest

- Perfect for Tangled Features
- Much More Efficient
- Stable Classification Result
- Avoid Over-fitting

Gradient Boosting Decision Tree

- Update of Random Forest



Evolution of Classification Strategy

ANN

- Feature Space Relatively Small
- Non-trivial Computation Complexity

SVM

- Inappropriate for Tangled Features
- Hard to Converge
- Time Consuming

Random Forest

- Perfect for Tangled Features
- Much More Efficient
- Stable Classification Result
- Avoid Over-fitting

Gradient Boosting Decision Tree

- Update of Random Forest

Classification Accuracy (24 Categories)

Year	2003	2004	2005	2006	2007	2008
Accuracy%	58.5	56.8	54.8	59.4	60.5	60.6

Year	2009	2010	2011	2012	2013	2014
Accuracy%	60.0	60.6	63.0	63.4	66.5	67.4

Average Accuracy (2003-2014) = 61.0%

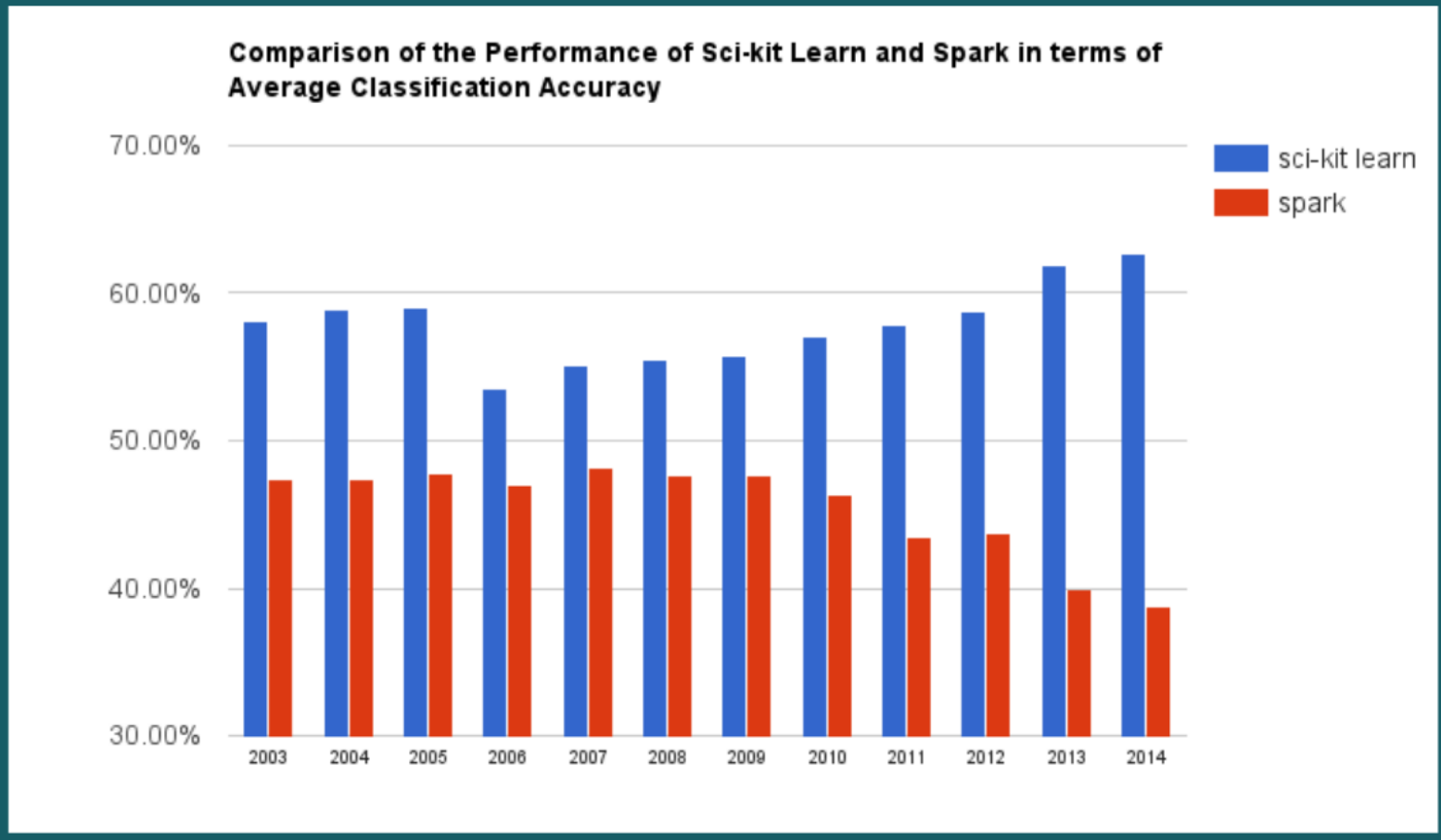
Classification Accuracy (24 Categories)

Year	2003	2004	2005	2006	2007	2008
Accuracy%	58.5	56.8	54.8	59.4	60.5	60.6

Year	2009	2010	2011	2012	2013	2014
Accuracy%	60.0	60.6	63.0	63.4	66.5	67.4

Average Accuracy (2003-2014) = 61.0%

Spark vs Sci-kit Learn

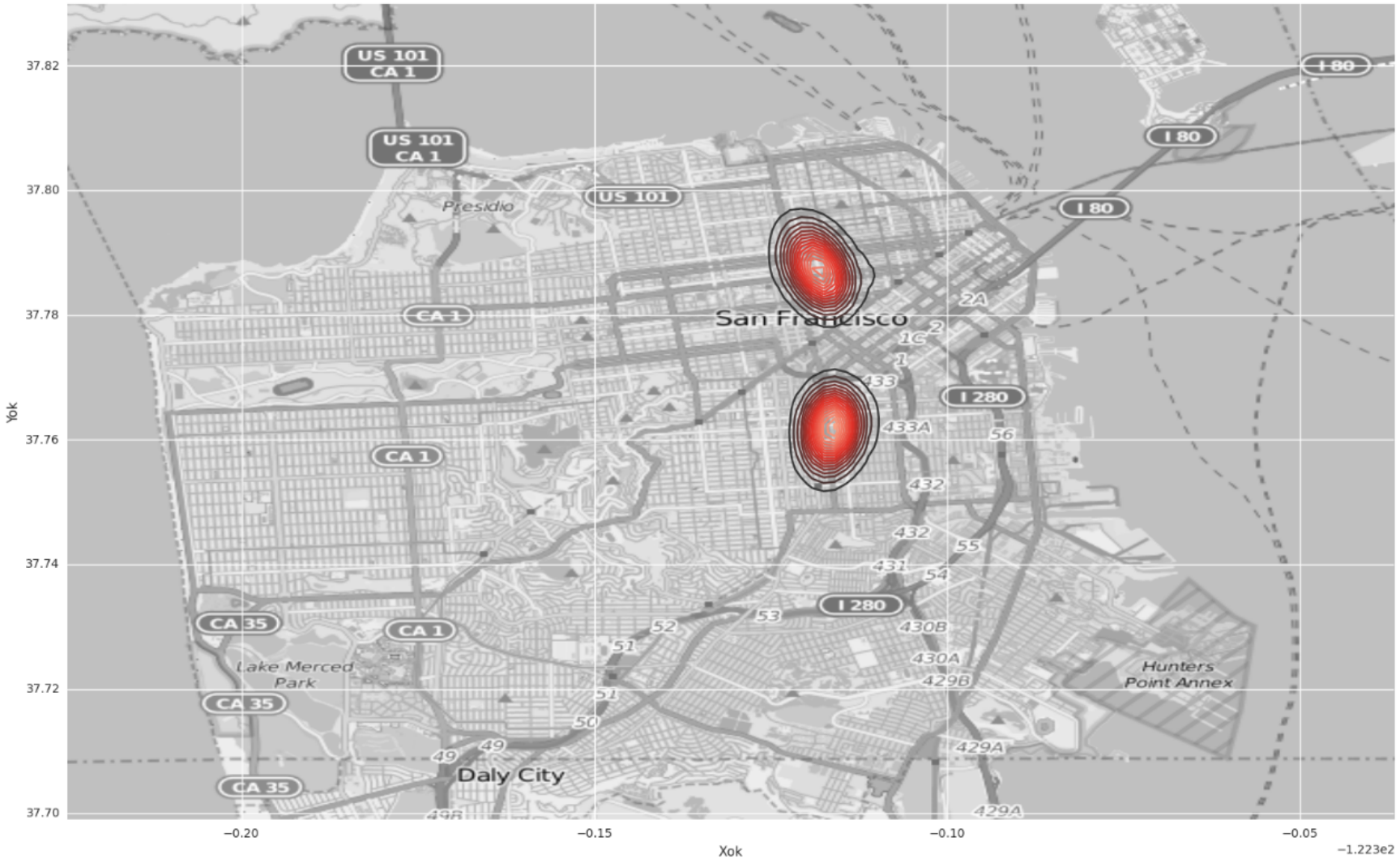


- Limited Maximum Depth of Decision Trees
- Not Dedicatedly Designed for Machine Learning

Summary

- Empirical Information Benifits for Crime Reduction

- Decision Tree Outperforms SVM - Crimes mixed up both geographically and temporally.
- Insight from PCA - Address Features are most significant than time-correlated feature.


US 101
CA 1
I 80
Presidio
San Francisco
CA 35
I 280
Lake Merced
Park
Hunters
Point Annex
Daly City
37.82
37.80
37.78
37.76
37.74
37.72
37.70
Yok
−0.20
−0.15
−0.10
−0.05
Xok
−1.223e2

Summary

- Empirical Information Benifits for Crime Reduction

- Decision Tree Outperforms SVM - Crimes mixed up both geographically and temporally.
- Insight from PCA - Address Features are most significant than time-correlated feature.

San Francisco Crime Classification

Group ID: 201512-22
Sirui Tan,
Haoyue Bai,
Guihao Liang