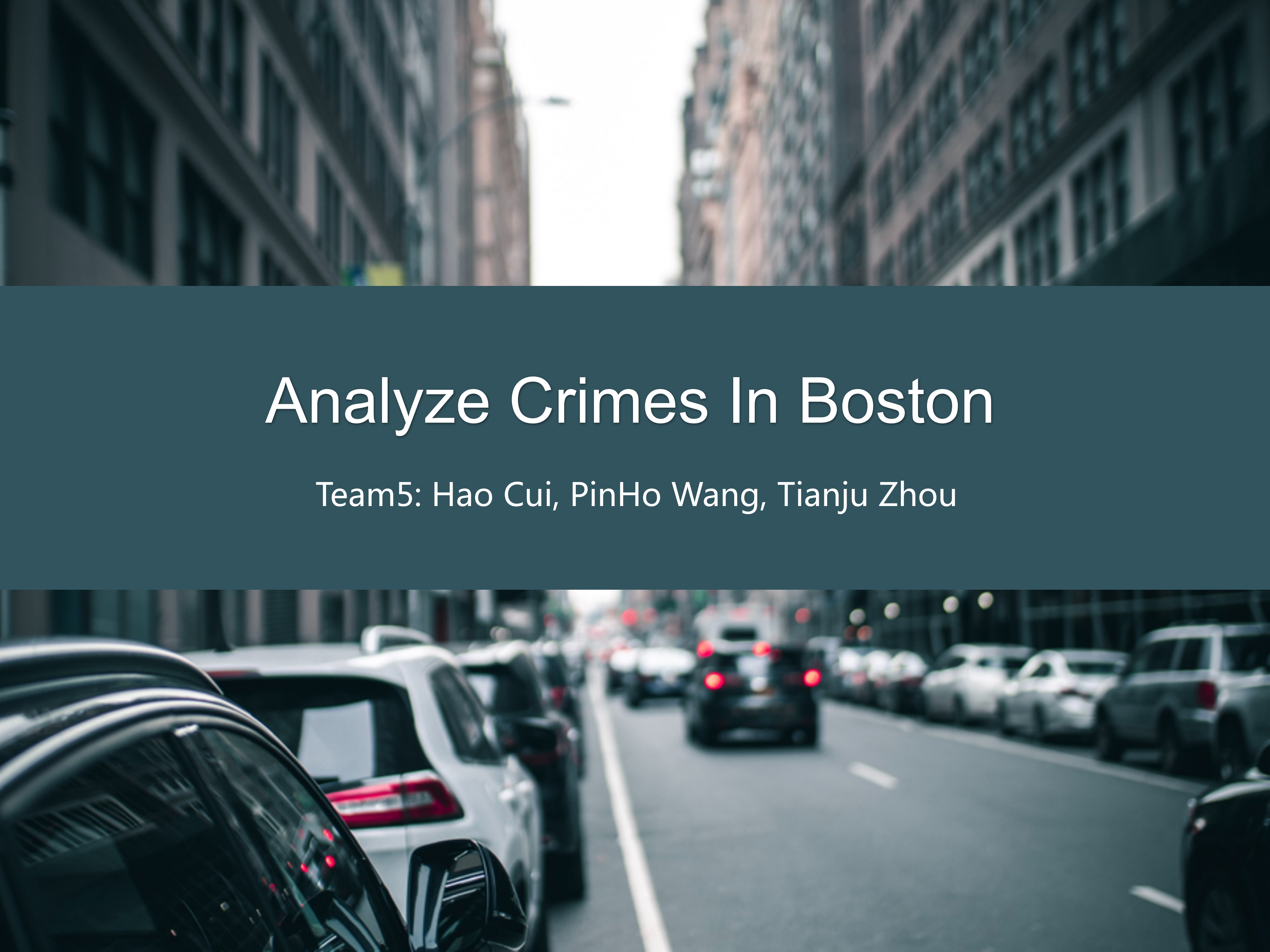

# Analyze Crimes In Boston

Team5: Hao Cui, PinHo Wang, Tianju Zhou

# Use Cases

1. Users input queries (e.g. date) and receive lists of matching records

2. The system will cluster criminal locations according to giving crime longitude and latitude

3. The system will predict the number of crimes in the next few days

4. The system will make clusters to filter crimes and select crimes which happens around holidays

5. The system will analyze the dangerous place for each holiday, then create the related hot-spot chart

# Methodology

1. Ingest data from csv files

2. Use Spark to process data (data cleaning and missing data processing)

3. Implement K-Means algorithm to cluster crime locations according to longitude and latitude

4. Use Holt Winter model to predict crime numbers in the next 365 days based on the records from 2015 to 2018

5. Utilize TextRank algorithm to make clusters for crimes related to personal safety

6. Visualize data through Zeppelin notebook and Play Framework

# Data Sources

## Crimes in Boston

https://www.kaggle.com/AnalyzeBoston/crimes-in-boston

1. Provided by Analyze Boston

2. 17 columns (INCIDENT_NUMBER, OFFENSE_CODE, OFFENSE_CODE_GROUP, OFFENSE_DESCRIPTION, DISTRICT, REPORTING_AREA, SHOOTING, OCCURRED_ON_DATE, YEAR, MONTH, DAY_OF_WEEK, HOUR, UCR_PART, STREET, Lat, Long, Location)

3. 319073 pieces of data in the dataset.

4. The records begin in June 14, 2015 and continue to September 3, 2018.

# Milestones

11.15

## Data Processing:

- Deal with missing data
- Remove unnecessary data
- Build the program frame- work to analyze data

11.22

## Analyze:

- Analyze data with Spark SQL
- Implement K-Means Algorithm
- Use Holt Winter Model to predict crime numbers in next few days

11.29

## Data Visualization:

- Visualize data on Zeppelin with DataFrame and SparkSQL
- Build front-end web pages using Play Framework

12.6

## Optimization:

- Optimize the program
- Prepare for the final presentation

# Code

- The program will be coded in Scala and Spark.

- Repository: https://github.com/CSYE7200/Analyze-Crimes-Boston

# Acceptance Criteria

1. All Spark SQLs should be executed within 5 seconds
2. The Alpha, Beta and Gamma value of Holt Winter Model should be within 1.0 +- 0.1
3. The Sum of Squared Errors (SSE) of K-Means should be within 0.01 +- 0.001
4. The accuracy of predicated model should higher than 70%

# Goals

01

- Learn to use Zeppelin, Spark and Spark SQL to analyze big data
- Learn to utilize Play Framework to build simple front end pages
- Understand the basic usage of Holt Winter Model
- Gain more knowledge in algorithms like K-Means and Text Rank

02

Provide information:

- The most frequent crime type
- The most dangerous place in Boston
- The most dangerous month in a year
- The most dangerous day in a week
- The most dangerous hour in a day
- High crime rate locations during holidays or weekends in Boston
- Other factors that may relate to crime rate

# Demo

*Demo*

*Thank You*