Università di Pisa	A.A. 2015-2016
Data Mining II	

Project assignment

Taxi cabs & crimes in S.F.

General information

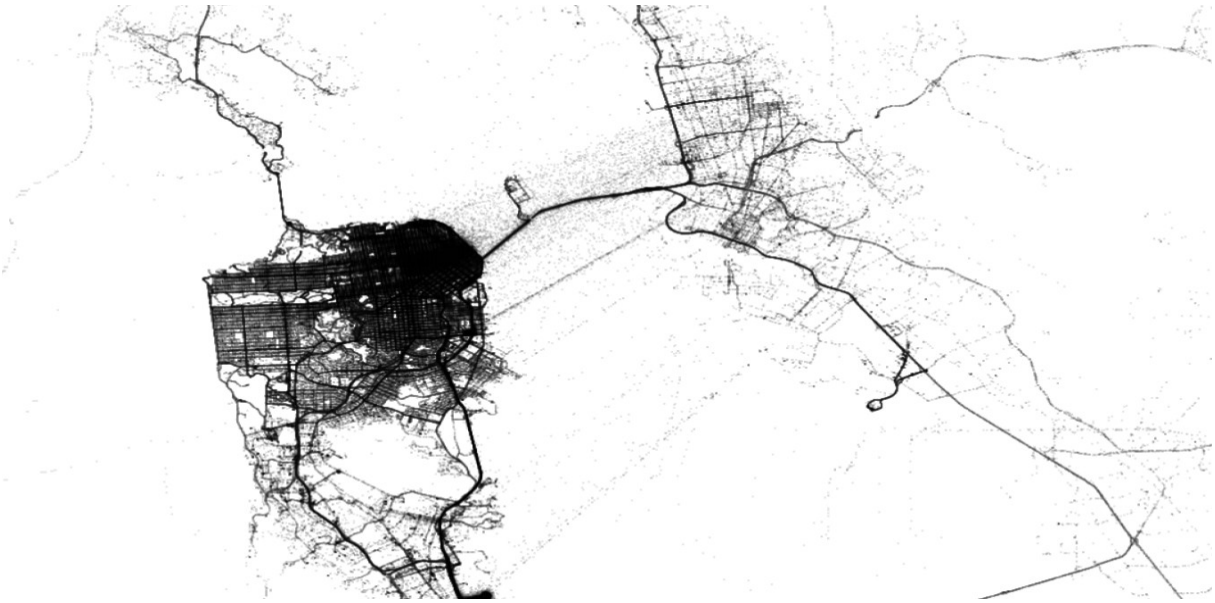
Objective of this project is to perform a few analyses on a dataset of mobility data involving taxis in San Francisco. The general guidelines for this assignment are the following:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the processing and analyses indicated in the text, trying to answer to each request. Any spontaneous addition to that is welcome yet optional, and cannot replace the original TODO list;
3. each group should summarize the work done in a short report (indicatively 5-15 pages), loosely following the guidelines of the CRISP model;
4. each group is totally free to choose the tools and software it prefers;
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni (mirco.nanni@isti.cnr.it).

The dataset

The data provided include two main sources:

- **GPS traces** of ~500 taxis over 30 days. Each San Francisco based Yellow Cab vehicle is currently outfitted with a GPS tracking device. The data is transmitted from each cab to a central receiving station, and then delivered in real-time to dispatch computers via a central server. This system broadcasts the cab call number, location and whether the cab currently has a fare. The following picture shows (a zoom of) the density distribution of points in S.F. downtown.

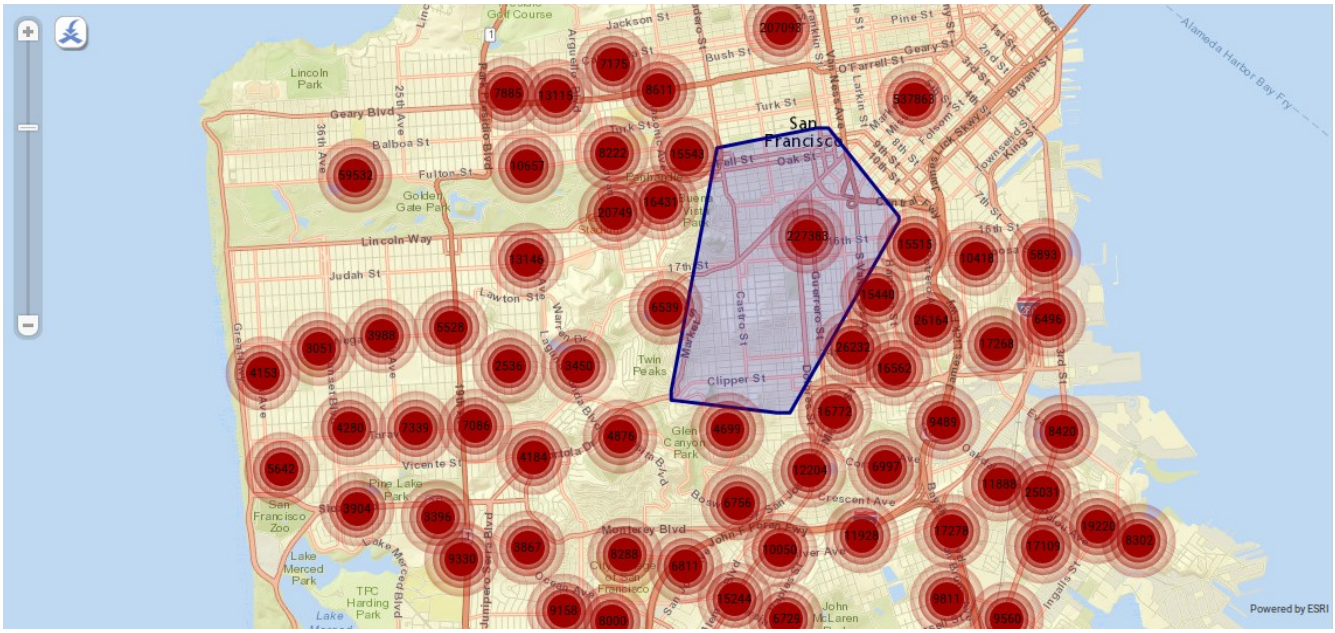


The raw dataset includes ~500 files, one per cab, containing <Latitude, Longitude, Passenger?, Unix Timestamp>. E.g.: 37.80246 -122.40186 0 1213034473.
The processed dataset consists of two sets of trajectories: one for taxi trips with passengers on-board, and one for trips without passengers.

- **Crime event records** for S.F. over several years – including the period covered by the GPS traces – equipped with geographical position. The basic source for this data is the Kaggle data challenge (https://www.kaggle.com/c/sf-crime), which provides the files train.csv and test.csv. This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set. They have the following fields:
    - Dates - timestamp of the crime incident
    - Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.
    - Descript - detailed description of the crime incident (only in train.csv)
    - DayOfWeek - the day of the week
    - PdDistrict - name of the Police Department District
    - Resolution - how the crime incident was resolved (only in train.csv)
    - Address - the approximate street address of the crime incident
    - X - Longitude
    - Y - Latitude

  As alternate source for this information, the file SFPD_Incidents_-_from_1_January_2003.csv is provided, directly obtained from SF government web site: https://data.sfgov.org/, where also additional data are available. A map

of the crimes can be seen here: https://data.sfgov.org/Public-Safety/Map-Crime-Incidents-from-1-Jan-2003/gxxq-x39z (see sample screenshot below)



Objectives

The following activities should be performed and reported:

1. **Exploration** of the two data sources, aimed to describe their general characteristics (distribution of key variables, spatial coverage, etc.) and detect possible issues (noise, strange behaviours, etc.).
2. **Study** the relation between crimes in the area and the taxi drivers' activity, trying to answer to the following questions:
   - Do taxi drivers avoid the areas with highest crime rates when driving?
   - What is the relation between crime rates and number of taxi pick-ups / drop-offs? E.g. do people in high-crime areas prefer taxi to other public transport?
   - Are there specific cases of crimes or crime bursts that apparently affected the taxi activity – globally or in the area of interest of the crimes?
   - Any other question you deem interesting.

Some discussions about Uber (carpooling-based service) and criminality in S.F. can be found here: https://newsroom.uber.com/crime-knowledge-demand-proxy/