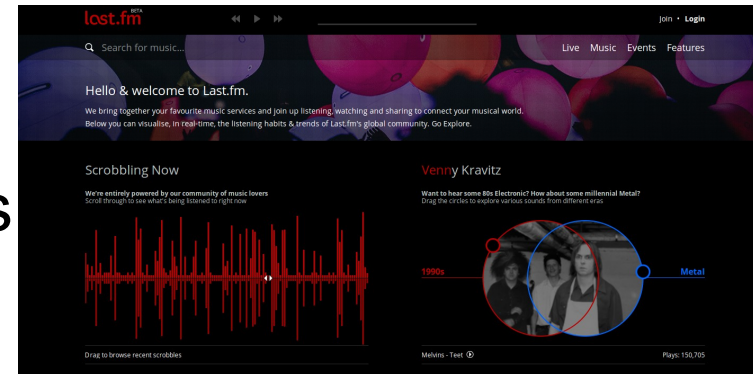


Data Mining A.A. 2015/16

Final projects

# List of projects

- Market basket context
  - Individual vs collective purchase behaviours
- Online services
  - Churn analysis on LastFM listenings
- Mobility context
  - Taxi cabs & criminality in San Francisco



# Project assignment

- Form groups of 1-3 students
- Send names and project chosen to the instructors
  - Detailed descriptions of the projects will be put online now
  - The datasets will be sent upon receiving your email
- Write a report on the analyses performed and the results obtained and send it before the final exam
  - Final exam will include a presentation with slides
  - 10-15min total for each group/project

# Project assignment

- Each project includes
  - A preliminary data exploration phase
  - Data analysis: central phase, driven by the general objectives assigned to you
  - Conclusions, where a summary of the key results, limitations and issues met is provided

# Individual vs collective supermarket purchase activity

- General idea: provide the customer a self-awareness of what he does w.r.t. the others



# Market basket project

## Dataset

- Real data describing customers and transactions
  - Several department stores
  - Purchases performed over 12 months
  - Includes product details, customer ID
- articolo.csv
  - textual description of the products (in Italian)
- cliente.csv
  - basic information about customers (in Italian)
- data.csv
  - translation table for date coding
- marketing.csv
  - marketing hierarchy of products (in Italian)
- venduto.csv
  - transactions, a line for each product sold

**Key table**

# Market basket project

## Top 10 most purchased products

- Choose the proper product category level to adopt
  - 70cl Whole Milk Brand X? Whole Milk? Milk?
- Identify an interesting period of day
  - 17-18? Mornings? Thursdays 16-19? Weekends?
- Discover top-10 products in the period for each customer
- Compute purchases distribution on them, for each customer
  - Milk: 50%, Bread: 30%, Wine: 10%, ...

# Market basket project

## Customer segmentation

- Segment customers into homogeneous groups
- Characterize each group
  - Purchase distributions
  - Other info derived from original data



# Market basket project

Individual vs. collective

- Select (small) sample of customers
- Compare the customer to the segment he belongs to
  - Highlight similarities and deviations
  - Sketch a self-awareness-style service

# LastFM & Churn

- General idea: who and why does stop listening to some music artist or genre?

The screenshot displays the Last.fm website interface. At the top left is the 'last.fm' logo with 'BETA' above it. To the right are navigation icons and links for 'Join' and 'Login'. Below the logo is a search bar with the placeholder text 'Search for music...'. To the right of the search bar are links for 'Live', 'Music', 'Events', and 'Features'. The main content area features a large banner with the text 'Hello & welcome to Last.fm.' and a sub-header 'Scrobbling Now'. Below this is a red waveform visualization. To the right of the waveform is a section for 'Venny Kravitz' with a circular image of the band Melvins and a play button. The text 'Melvins - Teet' and 'Plays: 150,705' is visible at the bottom right.

last.fm<sup>BETA</sup> Join • Login

Search for music... Live Music Events Features

Hello & welcome to Last.fm.

We bring together your favourite music services and join up listening, watching and sharing to connect your musical world. Below you can visualise, in real-time, the listening habits & trends of Last.fm's global community. Go Explore.

Scrobbling Now

We're entirely powered by our community of music lovers  
Scroll through to see what's being listened to right now

Venny Kravitz

Want to hear some 80s Electronic? How about some millennial Metal?  
Drag the circles to explore various sounds from different eras

1990s Metal

Melvins - Teet

Plays: 150,705

Drag to browse recent scrobbles

# LastFM & Churn Data

- Data about listenings: last 200 listening performed by a set a users:
  - user\_id: identifies the user
  - date: timestamp of the listening
  - track: title of the song listened
  - artist: artist of the song
  - album: album of the song

# LastFM & Churn Data

- Music genres: association of the predominant / best fitting genre for a given artist, according to LastFM weights:
  - artist: artist/group's name
  - genre: genre of the artist

# LastFM & Churn Data

- Network of friendships of the users:
  - user\_id1: user\_id contained in listening file
  - user\_id2: user which is friend of user\_id1  
(Notice: he is not necessarily in listening file)

# LastFM & Churn

## Churn analysis

- Choose an artist, set of artist or a whole genre
- Study the churn phenomenon for that:
  - Identify the users that consistently listen to them
  - Identify those that, at some point, abandoned the artist/group/genre (churn)
  - Try to understand what determined the churn, and build a model able to predict it in advance.
    - Possible causes to consider: features of the user, of the artist/group/genre, friends' features, etc.

# LastFM & Churn

## Customer segmentation

- Build a customer segmentation of LastFM users based on as much information as you can infer:
  - what they listen to
  - when they do that
  - friendships
  - etc.

# Taxi cabs & crimes in S.F.

- General idea: does crime influence how taxis operate their service?





# Taxi cabs in S.F.

## Dataset

- 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 number, location and whether currently has a fare

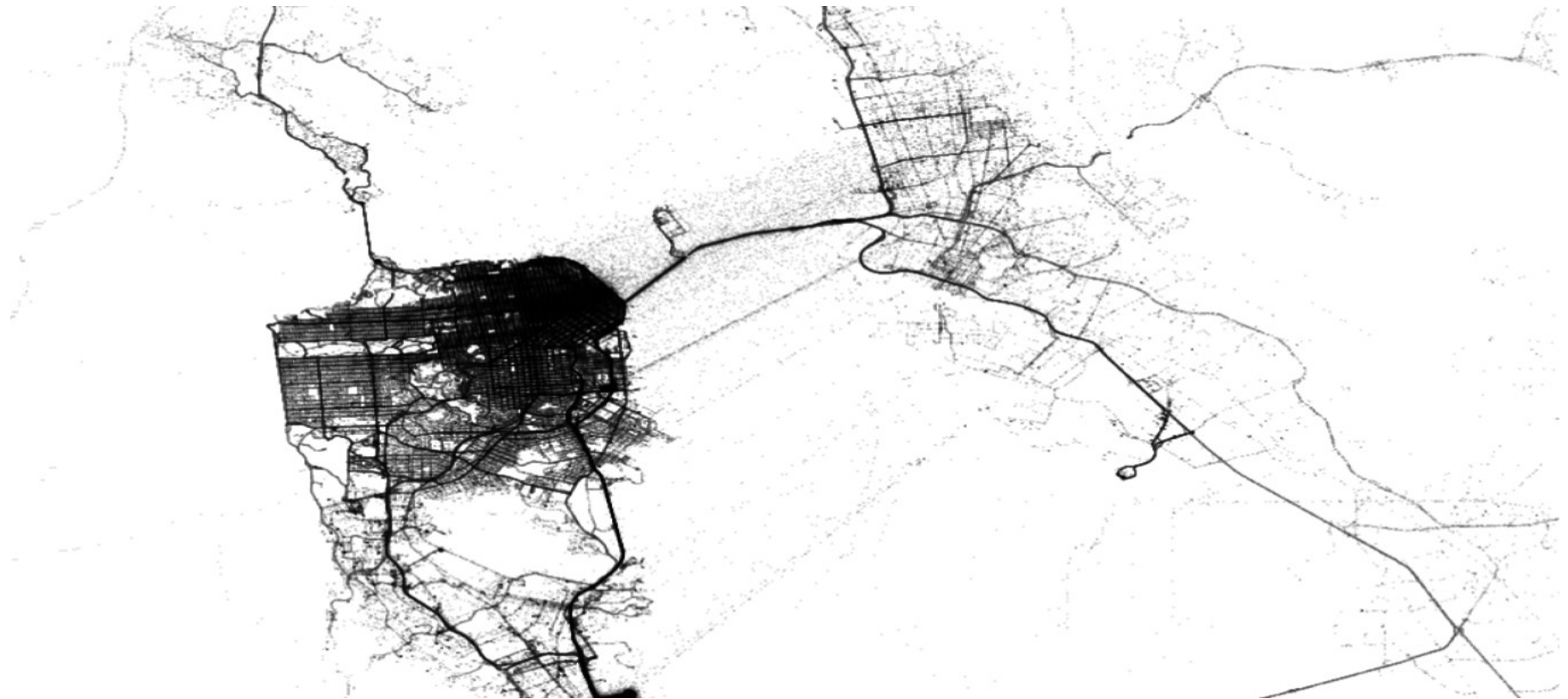


# Taxi cabs in S.F.

## Dataset

- Raw dataset: ~500 files, one per cab:

<Latitude, Longitude, Passenger?, Unix Timestamp>



# Crimes in S.F.

## Dataset

- Crime event records for S.F. over several years
  - Source: Kaggle data challenge  
<https://www.kaggle.com/c/sf-crime>
- Incidents derived from SFPD Crime Incident Reporting system.
- The data ranges from 1/1/2003 to 5/13/2015

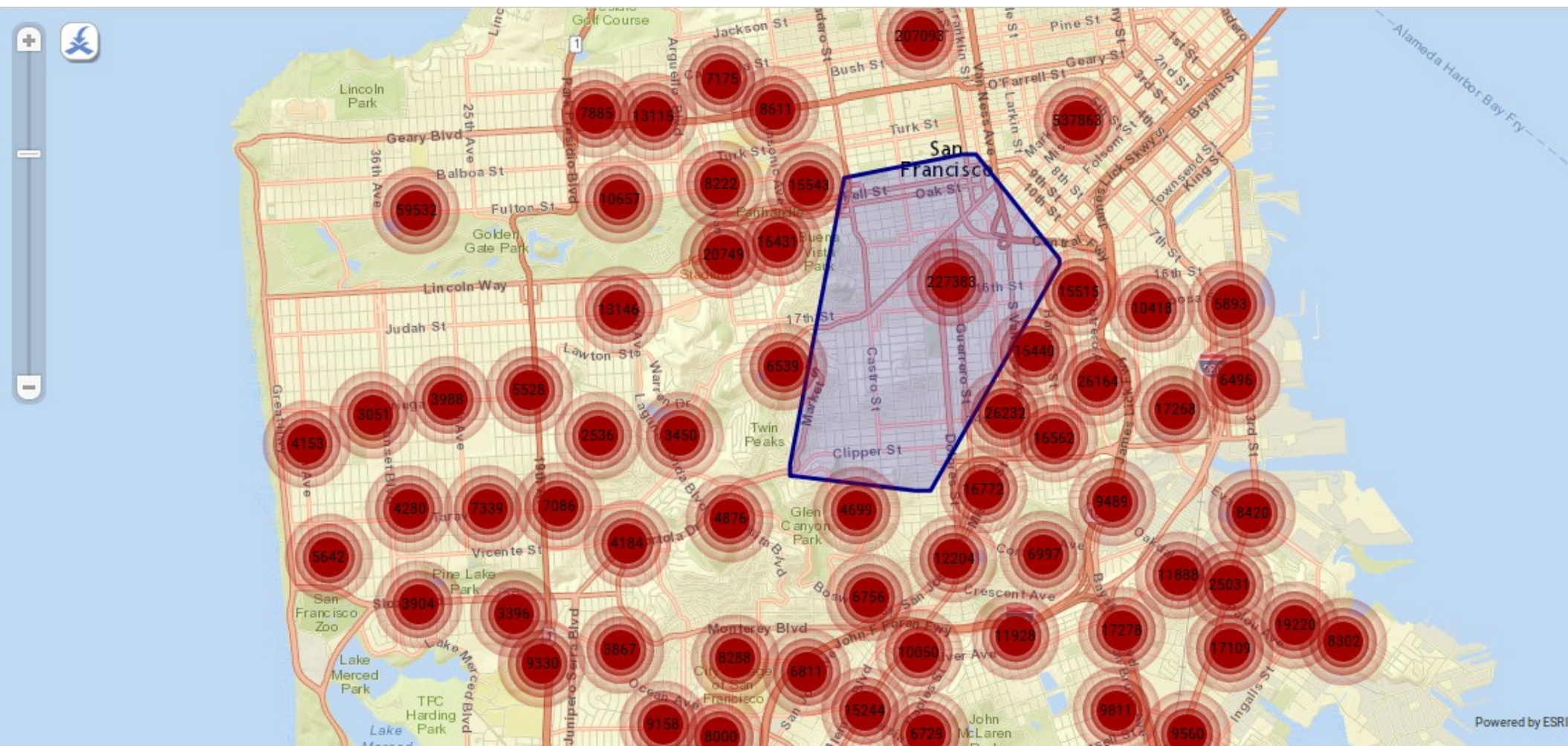
# Crimes in S.F. Dataset

- Data format:
  - **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.
  - **Description** - 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

# Crimes in S.F. Dataset

- Additional data available from

<https://data.sfgov.org/>



# Taxi cabs & crimes in S.F.

## Objectives

- Relation between crimes and the taxi drivers' activity
- Basic 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.

Questions?