Data Mining A.A. 2013/14

Final projects

List of projects

- Market basket context
  1. Entropy and sequentiality
  2. Promotions
- Mobility context
  3. Taxi cabs in San Francisco
  4. Big events analysis with Twitter data

Project assignment

- Form groups of 1-3 students and send names to the instructors
- One of the 4 projects that follow will be assigned to each group
  - Detailed description of the project and the needed data will be sent back to you shortly after
- 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
  - 15min total for each group/project

Entropy and sequentiality

Dataset

- Real data describing customers and transactions
  - Single department store belonging to the category “Supermarket”
  - Purchases performed over 6 months
  - Includes product details, customer ID

- SSET_ART_CORSODM
  - textual description of the products (in Italian)
- SSET_CLIENTE_CORSODM
  - basic information about customers (in Italian)
- SSET_DATA_CORSODM
  - translation table for date coding
- SSET_MKT_CORSODM
  - marketing hierarchy of products (in Italian)
- SSET_VEN_CORSODM
  - transactions, a line for each product sold

Key table

Entropy and sequentiality

Objective 1

- Data Exploration:
  - Examine data values and distributions
  - Understand what data can be useful
  - Identify significant issues or anomalies.

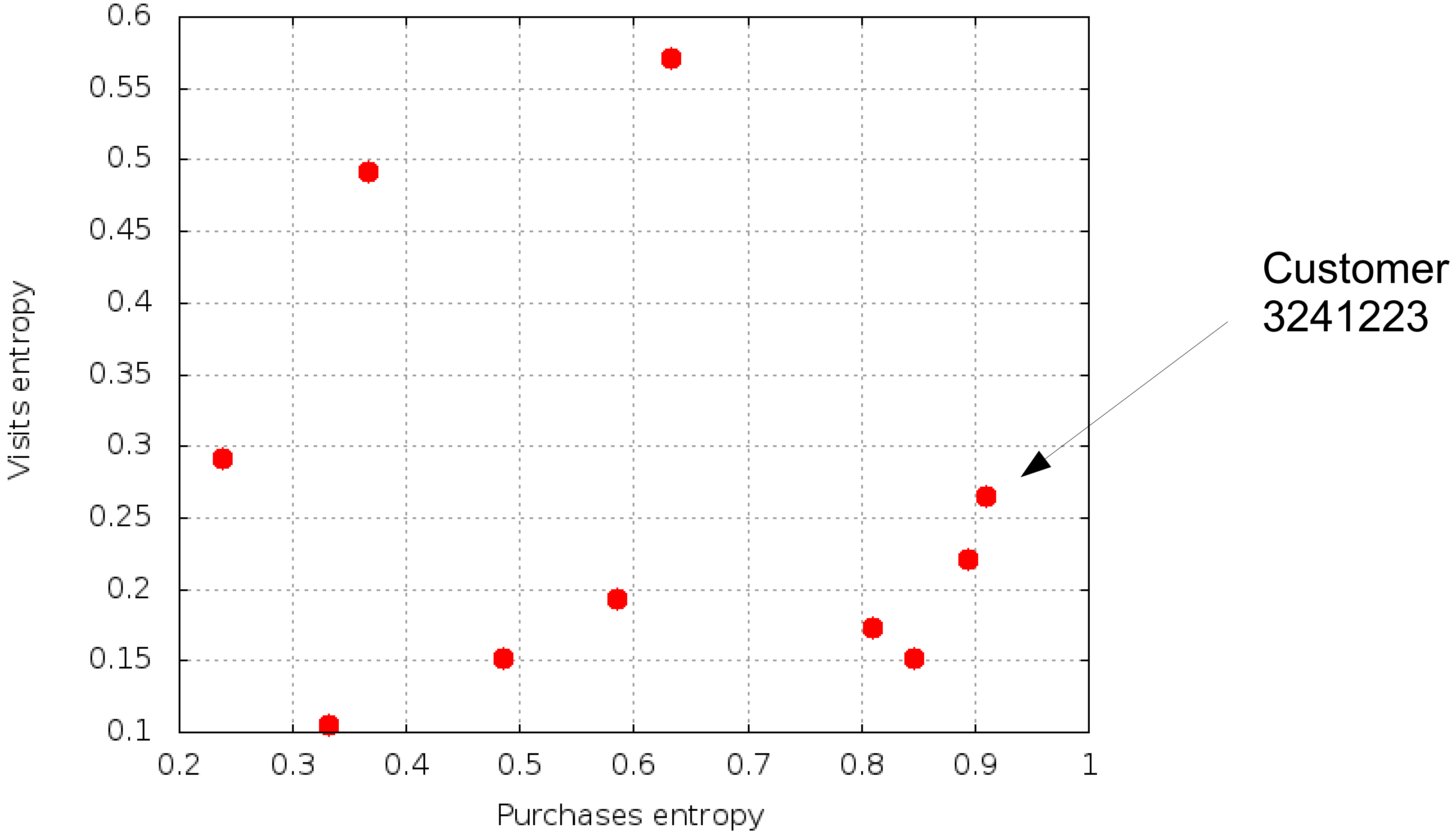
Entropy and sequentiality

Objective 2

- Purchases entropy vs. visits entropy
  - Purchases entropy: based on frequency of purchases of all products / product categories (you choose category level)
  - Visits entropy: based on frequency of visits to the store in days of week and/or hours of day (you choose time partitioning)
  - Study relations between the two entropies, through:
    - Visual inspection
    - Clustering

Entropy and sequentiality
Objective 2

- Purchases entropy vs. visits entropy



Customer 3241223

Entropy and sequentiality
Objective 3

- Sequentiality index
  - Step 1: consider sequence of purchases of each customer and compute sequential patterns
  - Step 2: consider *flattened input sequences*, and *flattened sequential patterns*, e.g.:

    {p1, p2} → {p1, p3} → {p1, p2, p4}   =>   {p1,p2,p3,p4}

  - Step 3:

    Seq_index = support(SP) / support(flat SP)

Entropy and sequentiality

Objective 3

- Sequentiality index
  - Step 4: Compute top 20 sequential patterns with highest Seq. Index value
  - Step 5: give **interpretation** of the index, and discuss some possible **business exploitation**

Promotions
Dataset

- Real data describing customers and transactions
  - Single department store belonging to the category “Supermarket”
  - Purchases performed over 6 months
  - Includes product details, customer ID **and active promotions**

- SSET_ART_CORSODM
  - textual description of the products (in Italian)
- SSET_CLIENTE_CORSODM
  - basic information about customers (in Italian)
- … (as before)
- PROMOZIONI
  - description of promotions, linked by single purchases (see table SSET_VEN_CORSODM)

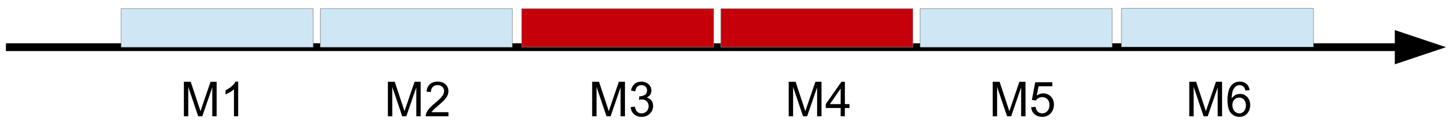
Promotions
Objective 1

- Data Exploration:
  - Examine data values and distributions
  - Understand what data can be useful
  - Identify significant issues or anomalies.

Promotions

Objective 2

- Sales forecasting under promotion
  - Focus on promotions that began in 3rd and 4th month of our dataset



  - Build a model to predict the amount of sales that will take place in the first 4 weeks of promotion:
    - **bad** (sales decrease >10% w.r.t. previous 4 weeks)
    - **neutral** (sales between -10% and +10%)
    - **weak** (sales between +10% and +30%)
    - **strong** (sales larger than +30%)

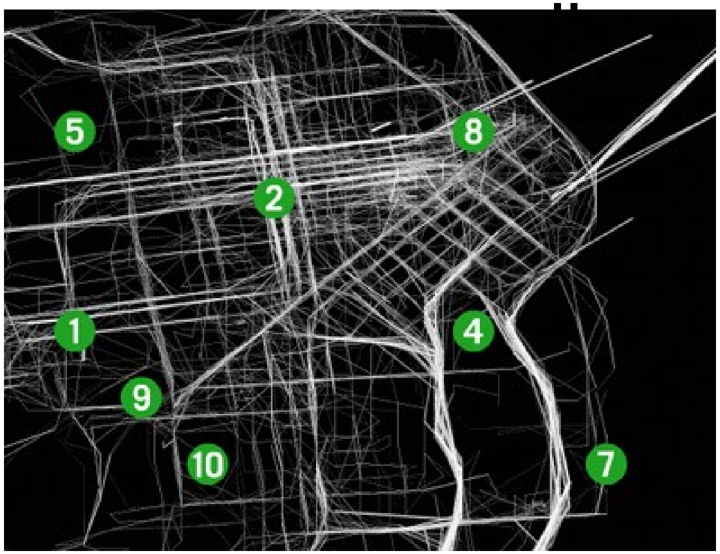
Promotions
Objective 3

- Behaviour analysis of sales under promotion
  - Build **time series** of weekly sells for the first 8 weeks of promotion
  - Discover **clusters** of promotions that show similar temporal evolutions of sells
  - Provide **interpretation**/characterization of clusters

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, containing
  - <Latitude, Longitude, Passenger?, Unix Timestamp>
  - E.g.:
    - 37.80246 -122.40186 0 1213034473
    - 37.8024 -122.40185 0 1213034409
    - 37.80245 -122.40166 0 1213034351
    - 37.80243 -122.40189 0 1213034287
    - ....
- Processed dataset:
  - Reconstructed trajectories (trips)
  - Separate trips with passengers from those without

Taxi cabs in S.F.
Objective 1

- Density of pick-ups
  - Evaluate the density of locations in S.F. Where taxis take passengers on-board
  - Which are the main pick-up areas?

Taxi cabs in S.F.
Objective 2

- Wandering patterns
  - How do taxis move when they have no passengers on-board?
  - Exploit (at least) T-Patterns

Taxi cabs in S.F.
Objective 3

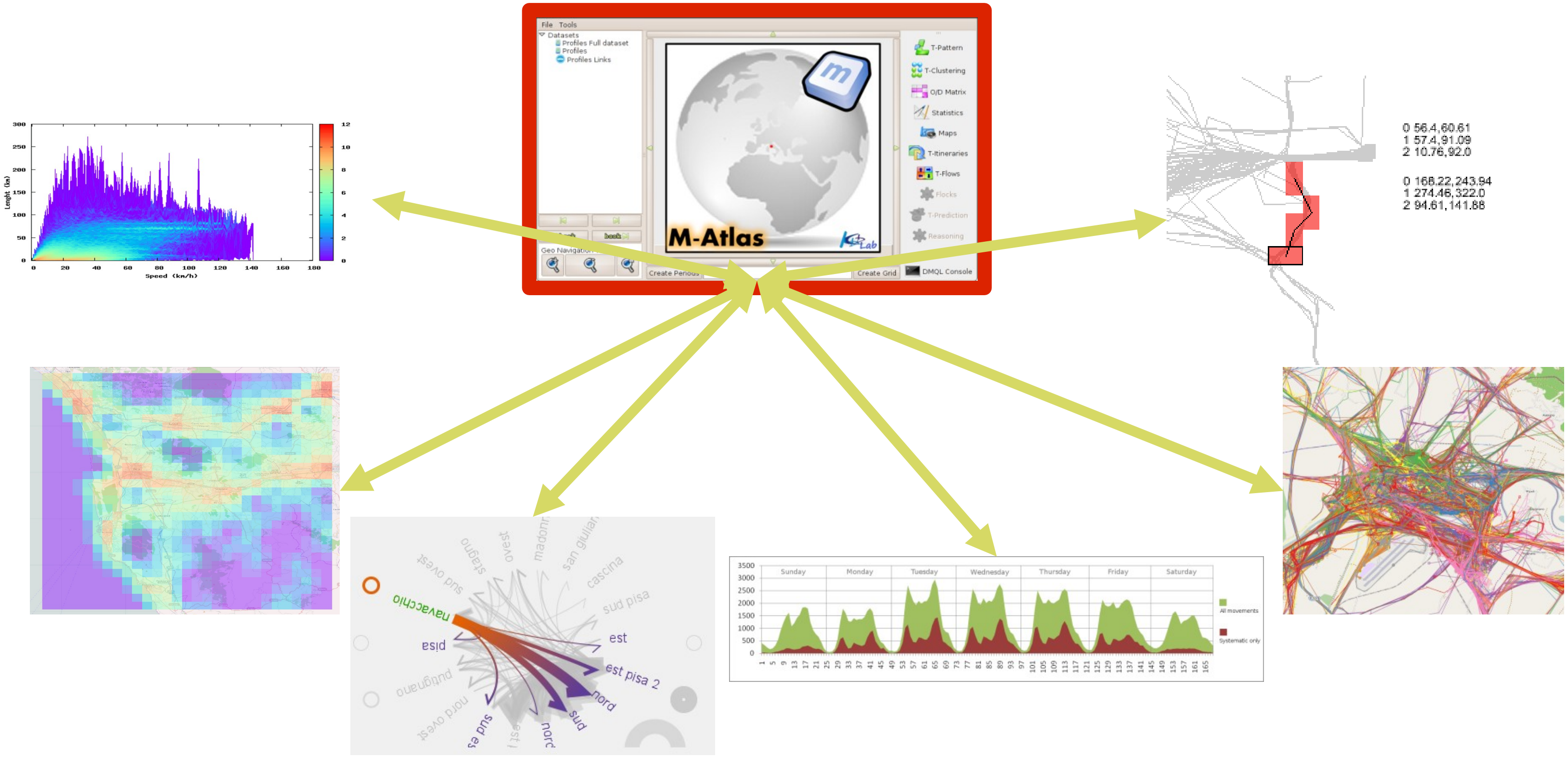
- Comparative O/D
  - Compute a grid over the area
  - Compute two Origin-Destination matrices:
    - That relative to trips with **passengers on-board**
    - That relative to trips of **empty taxis**
  - Provide qualitative comparison

Taxi cabs in S.F.

Analysis tool

M-Atlas platform

- A tool kit to extract, store, combine different kinds of models to build mobility knowledge discovery processes.

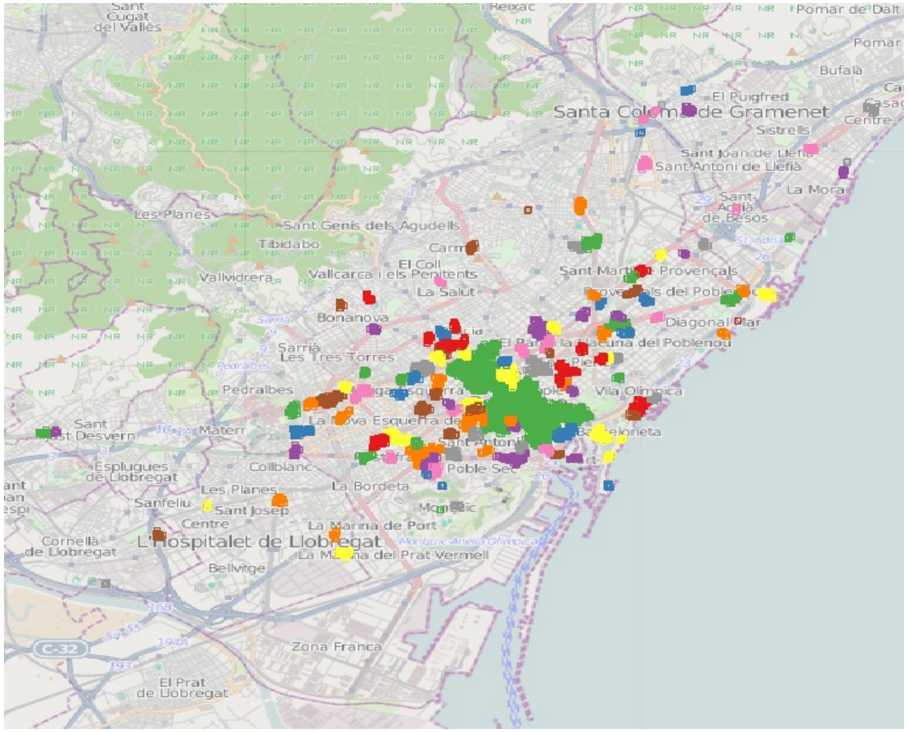


... detailed description in next lessons!

Big events analysis with Twitter data

Dataset

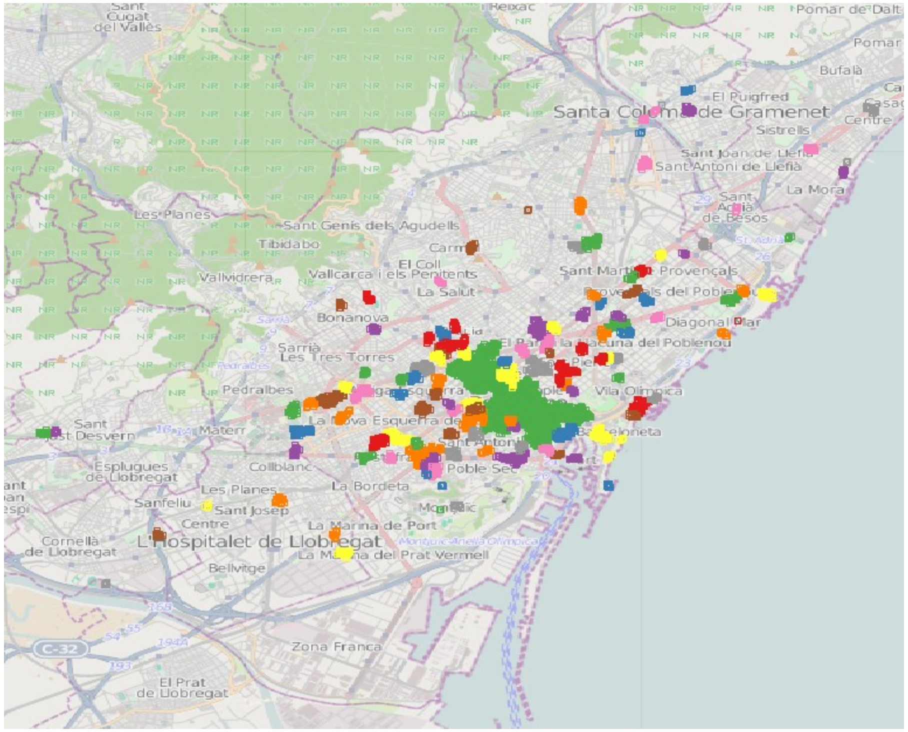
- Tweets generated in Barcelona along 3 weeks
- Contains the “Mobile Week Congress 2012”
- Tweets of the same user linked by same ID
- Geo-referenced



Big events analysis with Twitter data

Objective 1

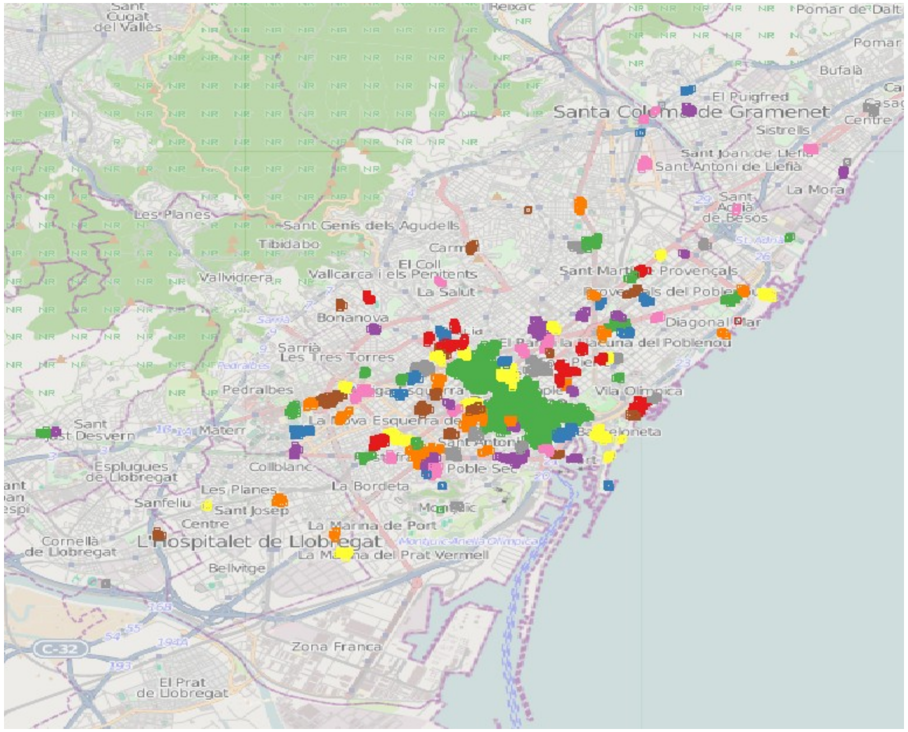
- Analysis of presence:
  - Which are the most active areas?
  - Does presence vary during the conference?
  - Which are the spots with higher peaks & variations?



Big events analysis with Twitter data

Objective 2

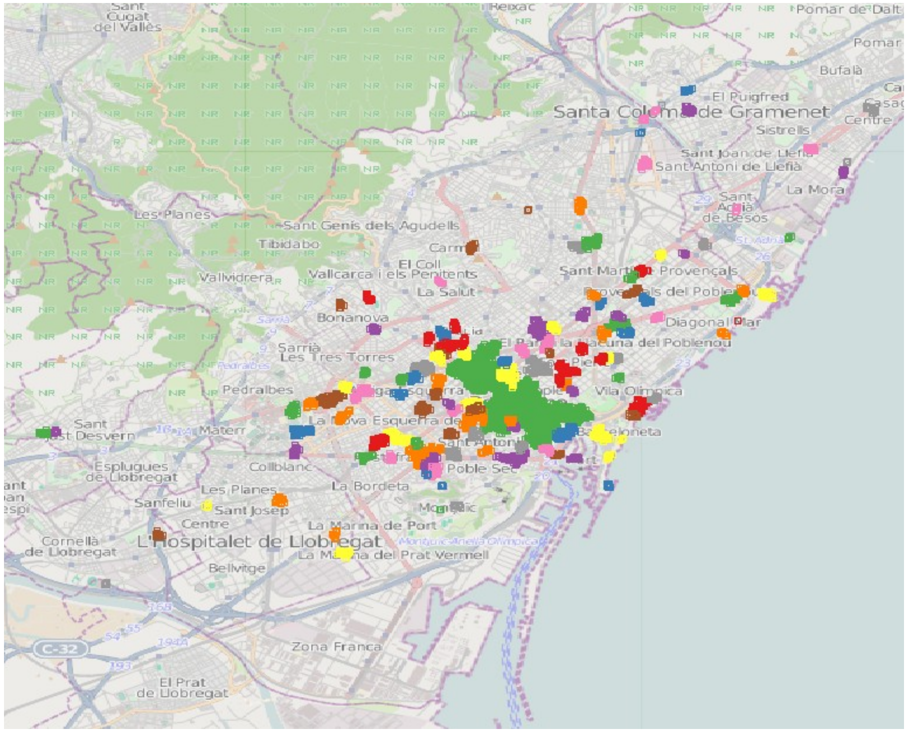
- Access path analysis
  - Detect the most likely area of the conference
  - Which are the main origin locations and access paths of people who joined the conference?



Big events analysis with Twitter data

Objective 3

- O/D analysis
  - Divide the territory into cells (e.g. a grid)
  - Compute Origin-Destination matrix
  - Which are the main flows?



Questions?