Data Mining A.A. 2014/15

Final projects

List of projects

- Market basket context: Customer segmentation based on
  1. Entropy
  2. User purchase profiling
- Mobility context
  3. Taxi cabs 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

Market basket projects

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

Entropy
Objective 1

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

Entropy
Objective 2

- Entropy measures
  - Purchases entropy: based on frequency of purchases of all products / product categories (you choose category level)
  - Temporal entropy: based on frequency of visits to the stores (i) in **months**,(ii) **days of week** plus other optional time slots
  - Spatial entropy: based on the frequency of visits in each store

Entropy

Objective 3

- Analysis of entropy measures
  - Study correlations among different measures
  - Perform customer segmentation based on entropy measures
  - Evaluate and explain the clusters obtained

User purchase profiling

Objective 1

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

User purchase profiling

Objective 2

- Build User Purchase Profiles: for each user identify the set of products that are frequently bought in the same shop in the same time
- E.g. user 7999
  - purchase profile: { (milk, store_23, Monday), (bread, store_23, Monday), (fish, store_30, Friday)}
  - Interpretation: milk and bread are ferquently bought on Monday in store #23, while fish is frequently bought on Friday in store #30
- Product and time detail level should be chosen from the available hierarchies (marketing.csv and usual temporal hierarchy of days, hours, etc.)

User purchase profiling

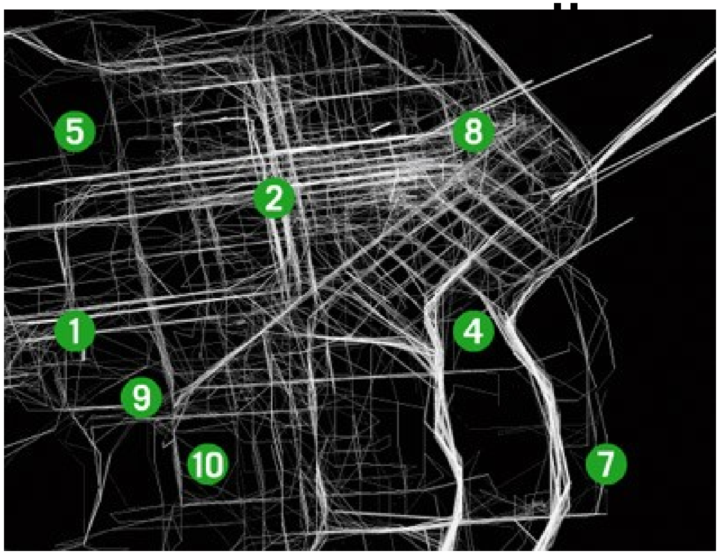
Objective 3

- Store Analysis:
  - select 2-3 significant stores
  - which are the typical user purchase profiles that occur in each of them?
- Suggested approach:
  - find two or more features to represent each user profile
  - perform a customer segmentation
  - give an interpretation of the clusters found and compare the results of the different stores

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

- Exploration of the data
  - general characteristics (distribution of key variables, spatial coverage, etc.)
  - possible issues (noise, strange behaviours, etc.)
  - origin-destination matrix to explore the distribution of flows across areas

Taxi cabs in S.F.

Objective 2

- Use the O/D matrix to select a significant area
- Compare the behaviour of “loaded trips” (i.e. taxi trips with passengers onboard) vs. the “unloaded” ones, adopting **three** different approaches

Taxi cabs in S.F.

Objective 2.1

- Approach 1
  - for each origin, compute the percentage of loaded trips that go towards the selected area "A", and analyze the distribution of the values obtained

Taxi cabs in S.F.

Objective 2.2

- Approach 2
  - divide the set of trips directed to the area “A” into
    - “AL” = loaded trips
    - “AU” = unloaded trips
  - Compute the main access patterns within the two sets, and compare the results

Taxi cabs in S.F.

Objective 2.3

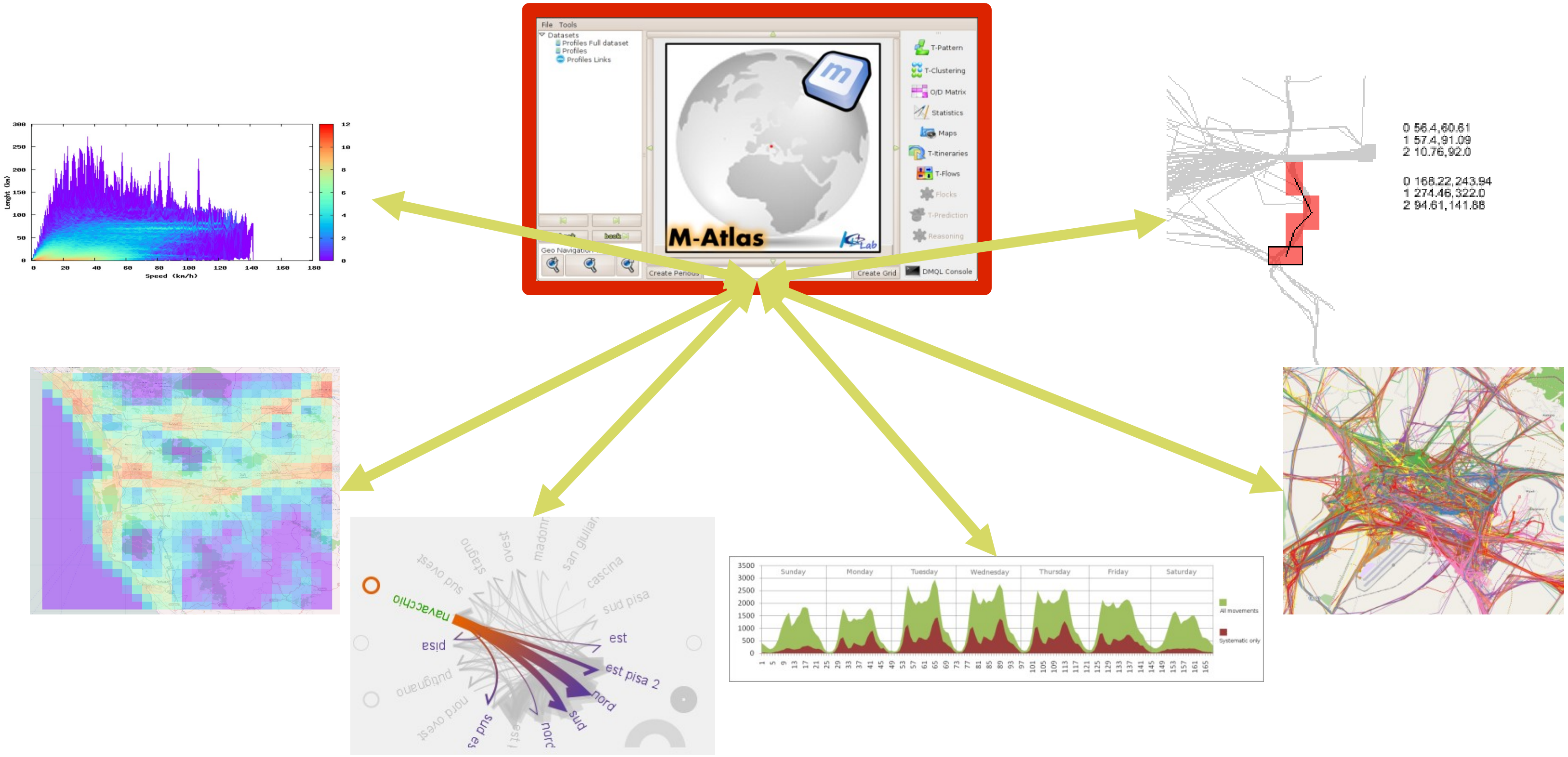
- Approach 3
  - characterize each dataset (“AL” and “AU”) through some set of indicators, such as travel duration, travel length, average speed, etc.

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!

Questions?