DATA SOURCES AND TECHNIQUES TO MINE HUMAN MOBILITY PATTERNS

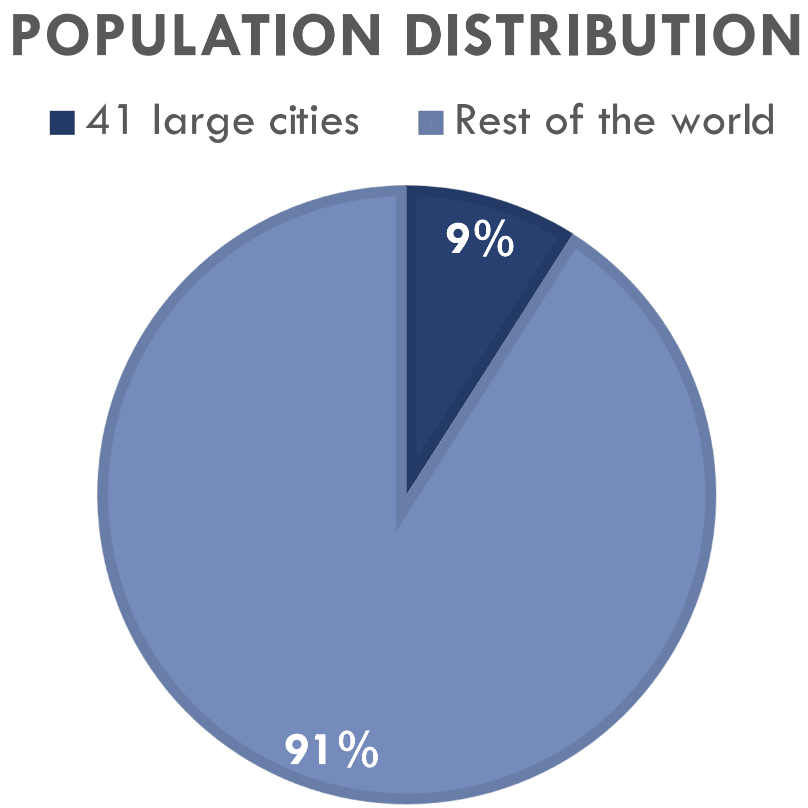
Ludovico Boratto (ludovico.boratto@acm.org)
Carmen Herrero (carmen.herrero@eurecat.org)
Andreas Kaltenbrunner (kaltenbrunner@gmail.com)
Matteo Manca (matteo.manca@zurich.com)
Giovanni Stilo (stilo@di.uniroma1.it)

The 5th IEEE International Conference on Data Science and Advanced Analytics
Turin, Italy – October 2, 2018

Motivation and outline

Motivation

- People is moving more and more from rural to urban areas



More than 50% lives in cities

Motivation

- Urbanization is changing people's lives
- Cities are becoming more and more complex
- New challenges arise:
  - air pollution
  - traffic congestion
  - resource allocation
  - mass tourism
  - …

Motivation

- Administrators and city planners need to know the dynamics of the city and how they interact
- Several actors play a role in these dynamics:
  - user habits
  - mobility patterns
  - most visited POIs (Points Of Interests)
  - Infrastructures management

Data sources

- To understand the trends of a city, several sources of data can be analyzed:

1. **User surveys**

Positive aspects	Negative aspects
• Very accurate (user residence, mobility patterns, and habits)	• High costs • Applied to a small sample of the population • Data is limited in space and time • Updated with low frequency

Data sources

- To understand the trends of a city, several sources of data can be analyzed:

2. **Wireless sensors**

Positive aspects	Negative aspects
• Very accurate • High-frequency data • Data can be collected for a long time	• High costs • Installation and management overhead • Spatial limitation

Data sources

- To understand the trends of a city, several sources of data can be analyzed:

3. **Mobile records – Call Detail Records (CDRs)**

Positive aspects	Negative aspects
• The localization of actions allows reconstructing human mobility • Support large-scale studies of aggregated behaviors • Rich data (not only positions, but also demo-graphic data: gender, nationality..)	• Designed for different purposes (i.e., billing) • Variable space granularity (location accuracy, depending on cell towers) • (Historically) low time granularity (number, frequency and uniformity of samples) • Not free and publicly available • User privacy

Data sources

- To understand the trends of a city, several sources of data can be analyzed:

4. **Apps and GPS**

Positive aspects	Negative aspects
• Very accurate • Allow to perform large-scale studies • High-frequency/real-time update of data	• Not free and publicly available • Users willingness to provide data about their attitude towards mobility (ratings, comments, surveys)

Data sources

- To understand the trends of a city, several sources of data can be employed:

5. **Social media**

Positive aspects	Negative aspects
• Easily collected through mobile devices • No temporal or spatial limitations • Allows large-scale studies • Accessible (almost) in real time	• Data collection and storing (GDPR) • Lower frequency collection (we rely on users' posts)

Tutorial structure

- This tutorial will focus on three data sources:
  - **Mobile records**
  - **Social media**
  - **Mobile apps**

- Objective is to answer the following question:
  - *To what extent each data source can be exploited to gain knowledge about human dynamics and mobility patterns?*

Tutorial structure

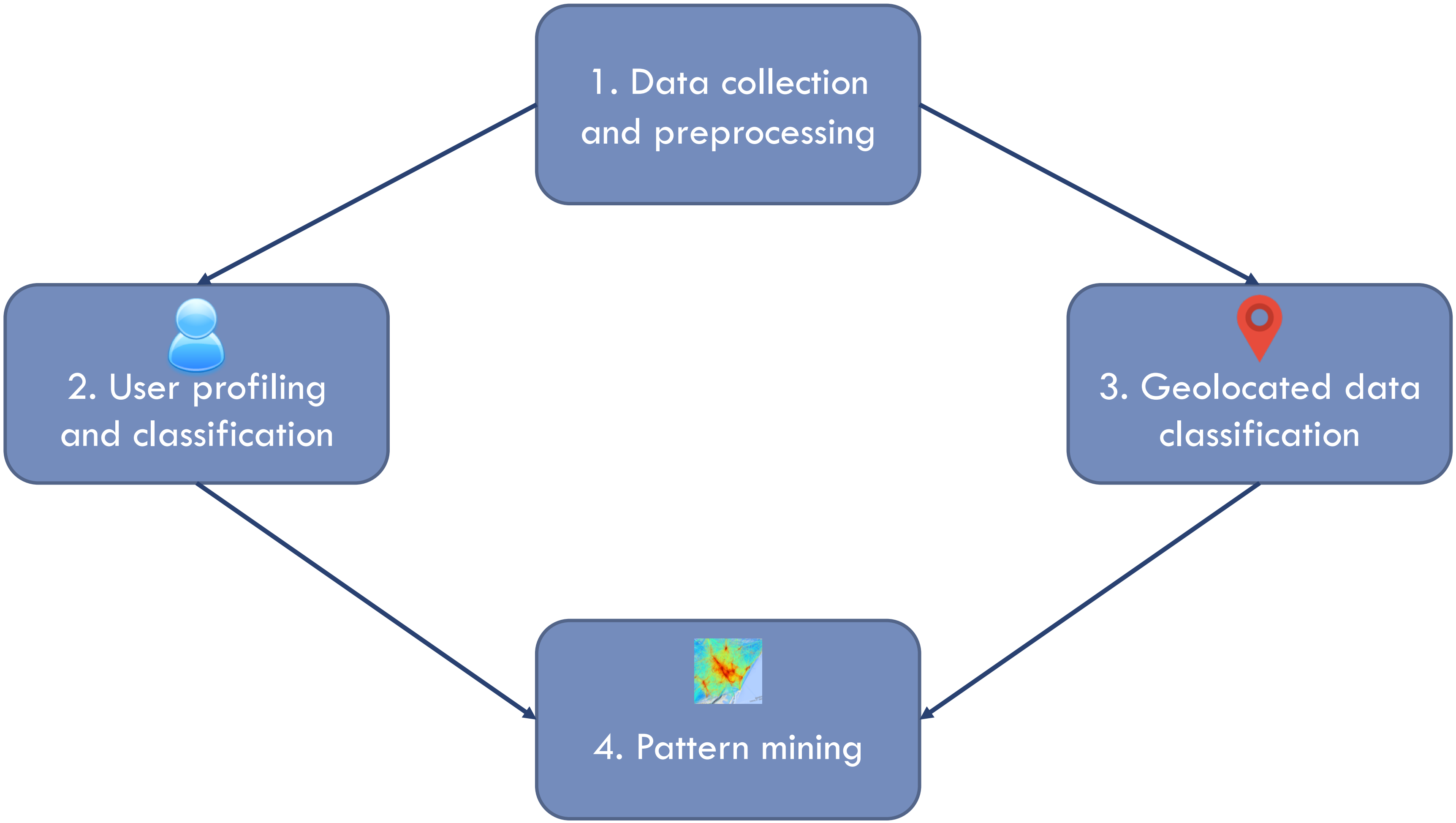
- We will define a workflow to mine human mobility patterns

- For each data source, we will:
  - Survey the existing literature
  - Present a case study based on the city of Barcelona

- We will conclude with open issues and future research directions

Workflow to mine human mobility patterns

Workflow to mine human mobility patterns



1. Data collection and preprocessing

- A crucial and not negligible aspect of the data mining process
  - 80% of the whole data mining process consists of data preparation [Zhang et al. 2003]

- Data is usually collected **through APIs**, that allow to perform several activities
  - download a stream of data in **real time**, specify a time window, a set of **keywords**, specify a **bounding box**, …

- **Problem.** How can we collect a set of geolocated objects? How can a preprocessing task **transform** raw data into **consistent** data that can be analyzed?
  - Common issues: presence of **errors** and **outliers**, **missing** values, and **inconsistencies** in the data

2. User profiling and classification

- It is the problem of identifying different **classes** of users in a specific area
- City planners and administrators are interested in studying **different aspects** of urban areas
- There is need to identify different classes or types of users (e.g., locals/tourists, active/passive)

- **Problem.** Given a set of users, how can we identify a **discrete** number of **categories**, based on **specific criteria**, and profile each user by **assigning** her to a given category?

3. Geolocated data classification

- **Profiling** of each **geolocated** data object, obtaining a segmentation of the initial dataset (e.g., considering **temporal** segments, or **geographic** areas)

- **Problem.** Given a set of geolocated data objects, how can we find a set of categories and identify to which category each object belongs?
  The objective is to segment the initial set of geolocated objects in to multiple groups.

4. Pattern mining

- In order to mine the mobility patterns, it is necessary to form **a *path* or *trajectory***
  - It orders all the places visited by a user
- Then, mobility patterns can be extracted considering:
  - The paths
  - The classification of the places visited by a user
  - The class to which the user belongs

- **Problem.** Given a **dataset of users** and their related **geolocated** data objects (posts), how can we **extract** the user **paths**, **avoiding** those that are too **short**, or that span over too **long** time periods?

Mobile records

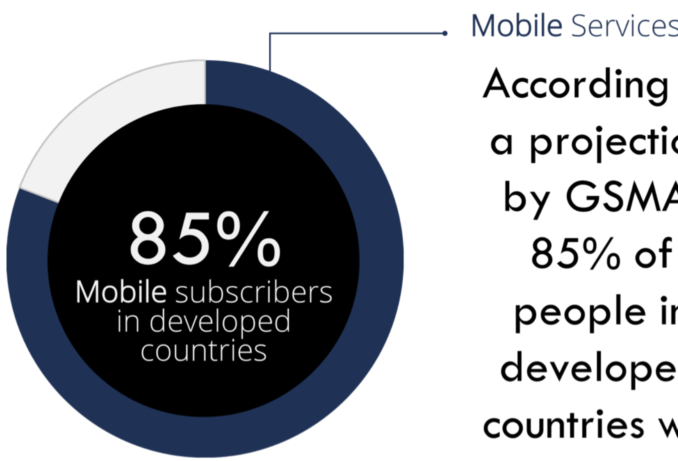
Mobile Records

Call Detail Records (CDR)

A CDR is a **summary ticket** of a telephone transaction, including the type of activity (voice call, SMS, 2G/3G/4G data connection), the user(s) involved, a time-stamp, technical details such as routing information, and the identifer of the cell ofering connectivity to the hand-terminal during the transaction.

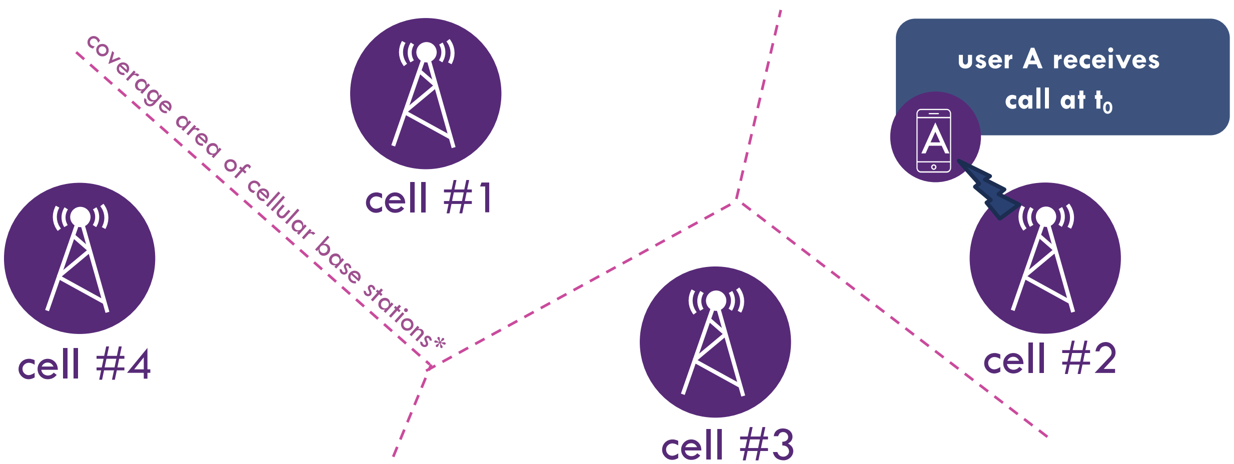
Mobile Records

- Call detail records are good set of data due to high - availability and penetration.
- It is not a surprise that both network operators and the research community look at mobile technologies as an unprecedented information source.
- Every terminal produces an enormous amount of meta-data that can be exploited to study aggregated behaviours and trends.



According to a projection by GSMA, 85% of people in developed countries will be mobile subscribers by the end of 2017

Mobile Records



example of a transaction

user_id	timestamp	action_type	cell_id	…
A	t_0	inbound_call	2	…

user_id	age	gender	rate	address
A	30	M	MyRate™	Rambla, BCN

dimension (user-related metadata)

cell_id	type	latitude	longitude	…
2	micro	40.9220	1.7612	…

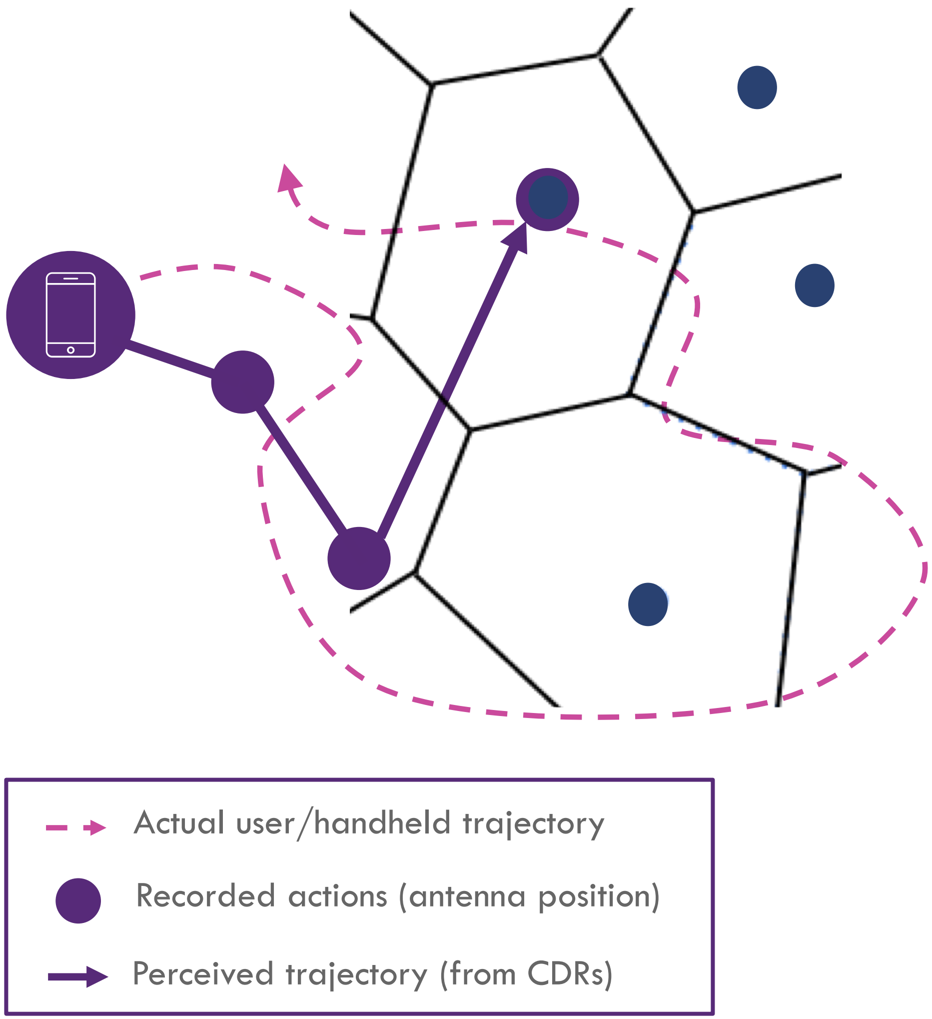
dimension (cell-related metadata)

Mobile Records

- The localization of actions allows reconstructing human mobility
- Support large-scale studies of aggregated behaviours
- Wide range of applications (socio-economical studies, transport optimization)
- Rich data (not only positions, but also demo-graphic data: gender, nationality..)

- Designed for different purposes (i.e., billing)
- Variable space granularity (location accuracy, depending on cell towers)
- (Historically) low time granularity (number, frequency and uniformity of samples)

Mobile Records



- Location accuracy depends on cell position and radio planning
- Main CDR limitation: limited time granularity. This is currently changing due to data connections

1. Data collection and preprocessing

State-of-the-art

1. Data collection and preprocessing

State-of-the-art

Different sources and data collection/processing strategies

- [Gonzalez et al. 2008] analyzed six-month of mobile phone dataset finding that human mobility is characterised by a high degree of temporal and spatial regularity.
- [Song et al. 2010] exploit mobile phone data to highlight the lack of variability in mobility predictions.
- [Pappalardo et al. 2015] use mobile phone and GPS data to study user mobility, discovering two main classes of users: returners, who focus their mobility to a few locations and explorers, whose mobility is not limited to few locations.
- [Fiadino et al. 2017] show the evolution in terms of data volume and quality on the CDRs.

1. Data collection and preprocessing

State-of-the-art

Different sources and data collection/processing strategies

- [Berlingerio et al. 2013] implemented a system that uses the location of mobile phone data to identify travel patterns in a city with the aim to help decision makers to improve the public transport systems.
- [Gabrielli et al. 2014] use mobile phone data to study the mobility behavior of visitors in a urban area.
- [Jiang et al. 2016] extract individual mobility networks comparable to the activity-based approach on Singapore.
- [Barbosa et al. 2018] review that can be used both as an introduction to the funda-mental modeling principles of human mobility, and as a collection of technical methods applicable to specific mobility-related problems.

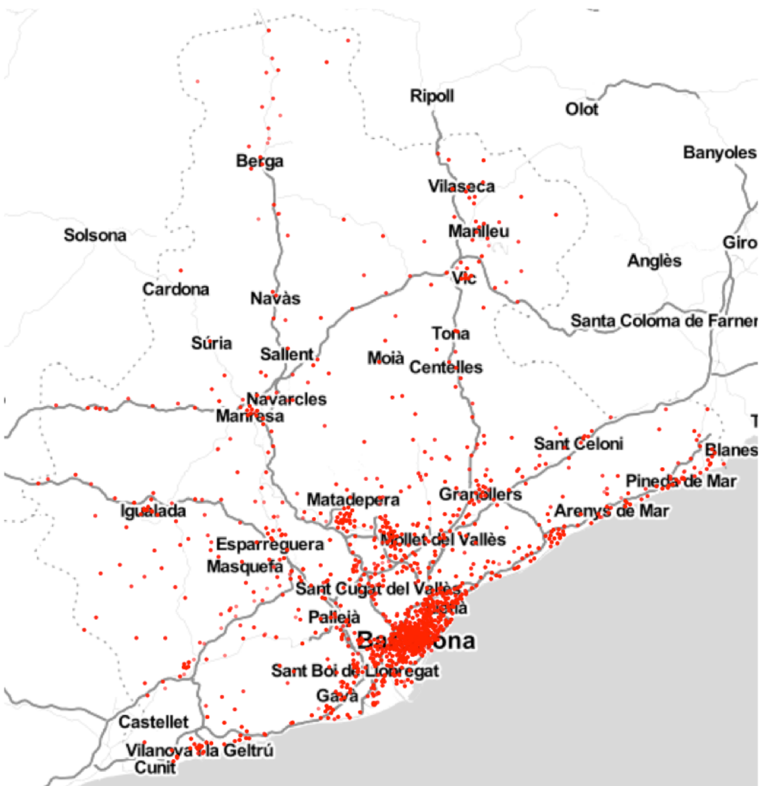
1. Data collection and preprocessing

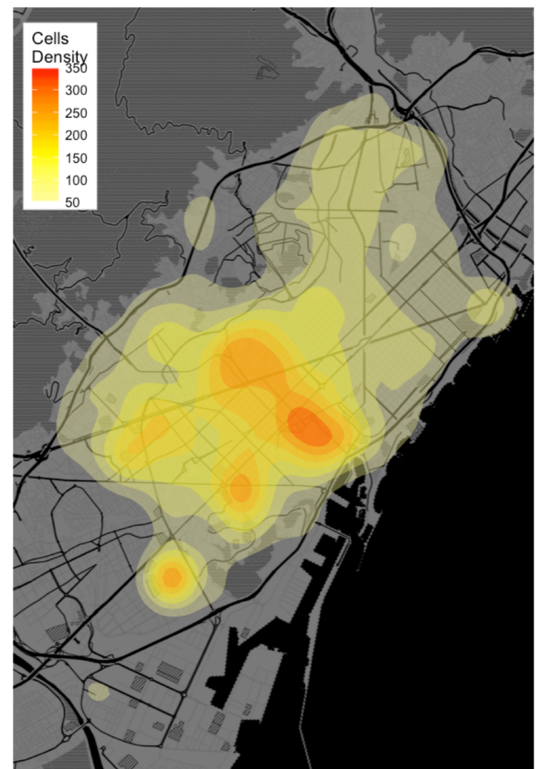
Case study

1. Data collection and preprocessing

Case study

- In the Barcelona case-study, we used data from a National mobile services provider.





1. Data collection and preprocessing

Case study

- Data range and records comparison

	Collection period	Region	Length	Records/day	Users/day
2018 Data	Q2 - 2018	Catalonia Region	15 days	1.3 billion	14 million
2016 Data	Q2 - 2016	Spain	31 days	1.1 billion	11 million
2014 Data	Q3 - 2014	Spain	31 days	350 million	9 million

1. Data collection and preprocessing

Case study

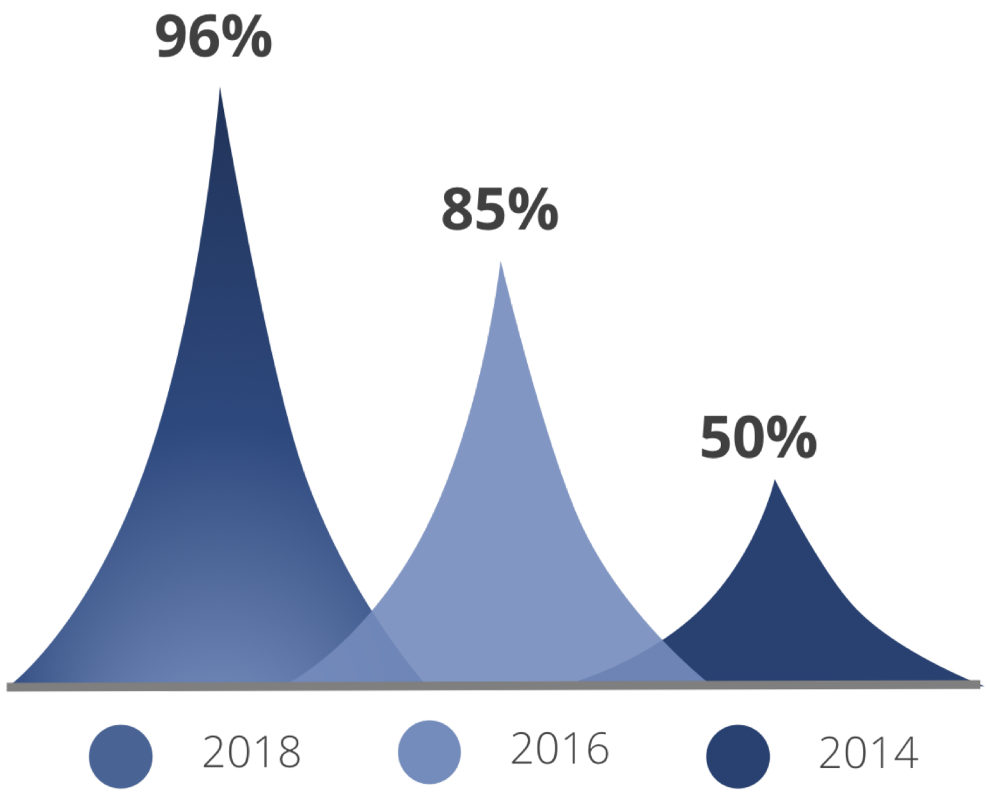
- 7-nodes EMR cluster (1 master, 6 workers)
- 28 CPUs, 213 Gb memory, 10TB S3 storage
- Apache Spark (for processing and data analytics)

Apache Zeppelin

1. Data collection and preprocessing

Case study

- Radical change in usage patters: more data connections.
- We have more actions and with more temporal granularity



Number of data connection records

2. User profiling and classification

Case study

2. User profiling and classification

Case study

- Identify users from which we can extract mobility.

Higher activity degree (number and *stability* of action pattern) → Higher number of position samples (i.e., temporal granularity) → Perceived mobility is closer to actual mobility (less bias)

To reduce bias in mobility results, we only consider Highly Active Users (HAU)

Define HAU samples according to requirements
e.g. urban mobility studies requires higher time granularity than Nation-scale studies

Even with strict requirements, the HAU sample will be large and statistical relevant
users are in general more active and the usage of data connections is widespread

Thresholds from an operational study	
Days Of Visibility (DOV)	>=75%
Hourly Action Rate (HAR)	>=1
Average Lag Time (ALT)	<=30m
Total Inactive Time (TIT)	<=75%
Entropy (H)	0.7<>0.9

ALL conditions →

Highly Active Users samples	
DS2014	7.5% (25% day only)
DS2016	12.5% (38% day only)

Sample size rather large: up to 38% of entire user population in DS2016

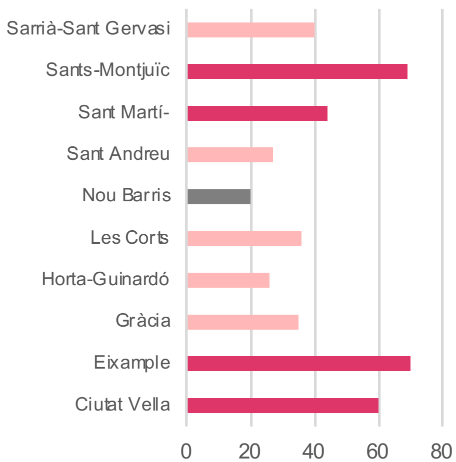
Statistical relevant samples: demographic (e.g., gender, age) statistical characteristics are preserved

2. User profiling and classification

Case study

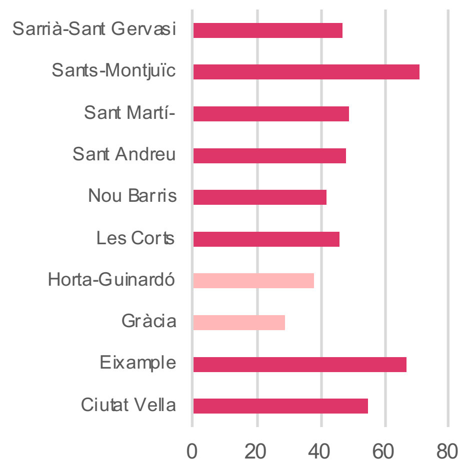
Classification of users

Tourist



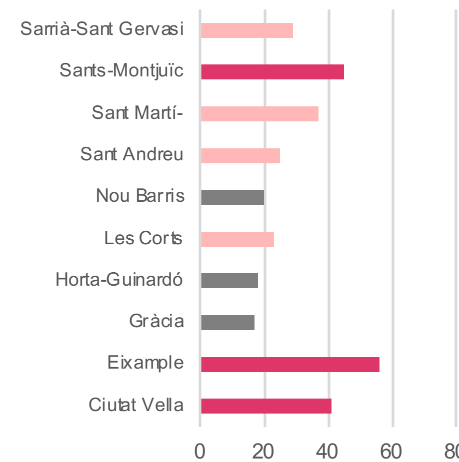
A visit starting and ending in two different days

Excursionist



A visit starting and ending the same day, with a maximum duration of 23:59 h.

Nightlife visitor



A visit to begins to register activity from 18 h. and stops recording it until 6 h.

3. Geolocated data classification

Case study

3. Geolocated data classification

Case-study

- We have a dataframe with all the actions and the corresponding user and location information

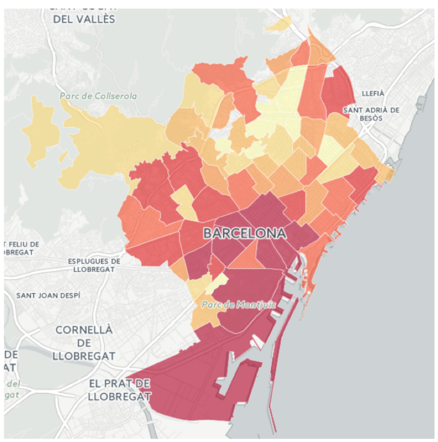
user_id	cell_id	time	action_type	cell_id_fromcells	latitude	longitude	azimuth	location	postcode	province	hour	minute	day
0366b440ed755add9...	214030050050000	2018-05-09 16:55:46	data_connection	214030050050000	41.3962	2.1756	70	BARCELONA	8009	BARCELONA	16	55	9
255b56382ba9cccee...	214030051643010	2018-05-09 11:21:22	data_connection	214030051643010	41.9321	2.2575	60	VIC	8500	BARCELONA	11	21	9
2b3da355b675faecd...	214030212630736	2018-05-09 20:26:44	data_connection	214030212630736	41.3817	2.1335	80	BARCELONA	8028	BARCELONA	20	26	9
3e4326cfe3249b4de...	214030050613010	2018-05-09 09:46:39	data_connection	214030050613010	41.3797	2.1468	50	BARCELONA	8015	BARCELONA	9	46	9
6e8f258b4ef4aa19b...	214030212618072	2018-05-09 06:12:47	data_connection	214030212618072	41.4004	2.1441	70	BARCELONA	8006	BARCELONA	6	12	9

4. Pattern mining

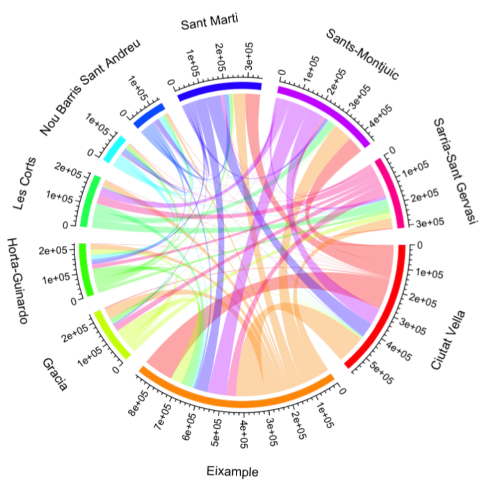
Case study

4. Pattern mining
Case study

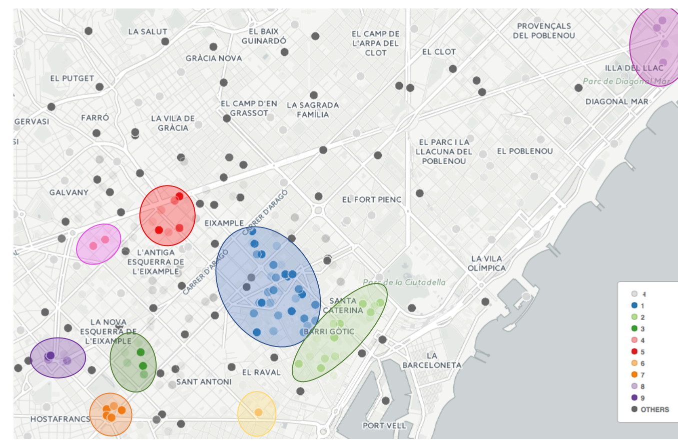
- We have used the data to study the concentrations of people on the area, the trajectory between neighborhoods and the detection of Points of Interest.



Heat maps: study concentrations of people per area



Transitions: between pairs of PoI, cities or neighborhoods



Human activity clustering: DBSCAN on weighted (by action count) tower locations

Social media

Social media

- We will analyze how **social media** can be employed **to mine urban mobility patterns**
- For each task in the workflow, we will illustrate how the **approaches** in the **literature** perform it
- **Then** we present a **case-study** based on the city of **Barcelona** and **Twitter** data, to compare the mobility patterns of local citizens with respect to those of tourists.

1. Data collection and preprocessing

State-of-the-art survey

1. Data collection and preprocessing

State-of-the-art survey

Different sources and data collection/processing strategies

- [Fuchs et al. 2013] considered all the tweets of users who **stayed in the Seattle area** for at least 10 days during a two-month period, **and outside for less than 10 days**; the data was preprocessed to remove tweets that contained Foursquare logins
  - They only considered geolocated tweets
- [Preotiuc-Pietro and Cohn 2013] considers only users who used **Foursquare** at least three times per day over one month
- [Ferrari et al. 2011] collected all the tweets in **Manhattan** for a **1-year period**

1. Data collection and preprocessing

State-of-the-art survey

- [Shelton et al. 2015] aims to verify if the city of Louisville (Kentucky) is actually **divided** into **east** and **west**, according to the notion of the '9th Street Divide'
  - They collect two datasets from Twitter of 703 users from the east and 662 from the west

1. Data collection and preprocessing

Case study

1. Data collection and preprocessing

Case study

- In the **Barcelona** case-study, we considered Twitter data with the aim to **extract mobility patterns**
- We used the **Twitter Streaming API** to collect the data
- We filtered data by location specifying a **bounding box** for Catalonia
  - Two comma-separated pairs of longitude and latitude representing the coordinates of the bottom-left point and of the top-right point: [**0.1592, 40.523, 3.3326, 42.8615**]

1. Data collection and preprocessing

Case study

- Initial dataset: 12,873,348 tweets posted in 2015
- Preprocessing to **filter out:**
  - the tweets that were **not geolocated**
  - the tweets that **were not published in Barcelona**, using a shape file
  - the tweets published **by bot accounts**, published by the same user on the exact same latitude and longitude
- Final dataset: **1,120,216 tweets**

2. User profiling and classification

State-of-the-art survey

2. User profiling and classification

State-of-the-art survey

- Given the geolocated data of a user, a user profile in the form of a vector that characterizes her preferences can be formed
- [Jin et al. 2016] builds a vector whose **elements** are the **points** that **Foursquare** awarded to the user in the considered week
- [Fuchs et al. 2013] mines mobility patterns associated to the lifestyle of the users and defines **22 categories** represented by **keywords** (such as food, family, etc.)
  - Each user is profiled based on the relevance of each category for her (i.e., the relative frequency with which the keyword occurred)

2. User profiling and classification

State-of-the-art survey

- Some approaches separate users between **locals** and **tourists**
  - to analyze the **spread** of an **illness** [Cao et al. 2015]
  - to study **global mobility** patterns [Hawelka et al. 2014] Others are interested in analyzing either the locals or the tourists
  - to analyze where the Seattle **locals tweet** [Andrienko et al. 2013]
  - to mine the mobility patterns **of tourists in Florence** [Girardin et al. 2007]

2. User profiling and classification

Case study

2. User profiling and classification

Case study

- Goal: separate the initial set of users into two subsets, ***locals*** and ***tourists***
- To do so, we considered:
  - The *userLocation* field of the tweets of a user. We considered a **set S of locations** that make the user local (*S = ["bcn", "barcelona", "badalona", "hospitalet"]*)
  - The number of consecutive days in which the user **tweeted inside Barcelona** (if more than 20 she is local, otherwise a tourist)

2. User profiling and classification

Case study

- After the profiling task, the dataset is structured as follows:

Number of geolocated tweets in Barcelona	*1,120,216*
Proportion of tourists’ tweets	19%
Proportion of locals’ tweets	81%
Number of unique users	*93,946*
Proportion of tourists	*57.5%*
Proportion of locals	*42.5%*

3. Geolocated data classification

State-of-the-art survey

3. Geolocated data classification

State-of-the-art survey

- Most approaches classify a social media data object as belonging to a geographic area
- In [Lee et al. 2010], each geolocated tweet is given as input to the k-means clustering algorithm, which defines "**Regions of interest**" (RoIs)
  - Close places with the same tweeting activity
- [Hasan et al. 2013] assigns each Foursquare check-in to a **200 meters × 200 meters** square into which a city is divided
  - Squares are then **ranked by popularity** for the subsequent pattern mining step

3. Geolocated data classification

State-of-the-art survey

- Other approaches present **a topic-based** classification of geolocated tweets
- [Fuchs et al. 2013] **classifies** tweets based on their **content**, considering **22 lifestyle-related keywords**
- [Frank et al. 2013] measures the **degree of happiness** with respect to the covered distance in a travel
  - Each word is given a score from 1 (sad) to 9 (happy)
- in [Cao et al. 2015] a tweet is "**flu-flagged**" if it contains a set of keywords, such as *flu*, *cough*, *sneeze*, and *fever*

3. Geolocated data classification

Case study

3. Geolocated data classification

Case-study

- We classify each tweet either as “weekend” or “working day” tweet
  - 370,942 “weekend” tweets
  - 854,257 “working day” tweets
- Moreover, we add a label to each tweet indicating the district name it was posted from

Ciutat Vella	333,183
Eixample	245,517
Gràcia	63,775
Horta-Guinardó	55,490
Les Corts	74,238

Nou Barris	49,626
Sant Andreu	55,936
San Martí	55,490
Sants-Montjuïc	137,328
Sarrià-Sant Gervasi	73,905

4. Pattern mining

State-of-the-art survey

4. Pattern mining

State-of-the-art survey

- Each point in a path might include just the place visited by the user, or take the form of a **tuple** with other information, like:
  - **time**
  - category of the **venue** (in case of check-in data)
  - **content** of the tweet
- Most of the approaches represent a path as a sequence of ***<location, timestamp>*** pairs

4. Pattern mining

State-of-the-art survey

- **Clustering-based approaches** cluster the individual user paths, to discover how an area has been used by the users
- K-means is used in [Frías-Martínez et al. 2012] to detect 4 clusters and **characterize** the tweeting **behavior** in Manhattan
- [Cranshaw et al. 2012] clusters venues with a spectral **clustering** approach to find which areas of a city are **characterized** by the same **dynamics**
- Non-Negative Matrix Factorization is used in [Jin et al. 2016] to capture temporal and spatial characteristics of users' **Foursquare check-ins**

4. Pattern mining
State-of-the-art survey

- **Model-based approaches** build models that consider a set of observed geolocated data points in a path and assign a category to which this set of points belongs (i.e., a pattern)
- Most approaches are based on Latent Dirichlet Allocation (LDA)
  - [Long et al. 2012] discovers local **geographic topics** from **Foursquare** check-ins
  - [Ferrari et al. 2011] discovered 30 **geographical topics** that characterize **Manhattan**
  - [Liu et al. 2014] found out that the observed **spatial interactions** in migration flows are governed by a **power law** distance decay effect

4. Pattern mining
State-of-the-art survey

- **Path-distribution-based approaches** study the distribution of the data points in a path, in order to analyze the mobility patterns
- [Noulas et al. 2011] studies the complementary cumulative distribution function of **Foursquare** check-ins
  - 20% cover a distance of 1 km
  - 60% are between 1 and 10 km
  - 20% take place at distances over 10 km
  - ~5% go beyond 100 km

4. Pattern mining

State-of-the-art survey

- When analyzing the '9th Street Divide', [Shelton et al. 2015] found out that the **two neighborhoods** can be considered as fluid
  - Users freely more from one to the other
- [Girardin et al. 2007] considers the **paths built with Flickr** photos and build inbound and outbound maps that show how **tourists move in Florence**, and analyzed the **most frequent** flows
  - **Americans follow** a **specific** graph constituted by the nodes of Florence, Siena, Pisa, Genova and Perugia
  - **Italians** are more **adventurous**

4. Pattern mining

Case study

4. Pattern mining

Case study

- We propose to extract patterns using a notion of *path* considering **one hop**, i.e., only two points (t_1, t_2)
- In order to form a path, the following must hold:
  - Two following tweets t_1, t_2 must be published in the same day
  - The distance between two subsequent points has to be **higher than 150 meters**
    - It is the distance between two parallel consecutive streets in Barcelona
  - The difference in hours between two **subsequent** tweets must **be lower than 10 hours**

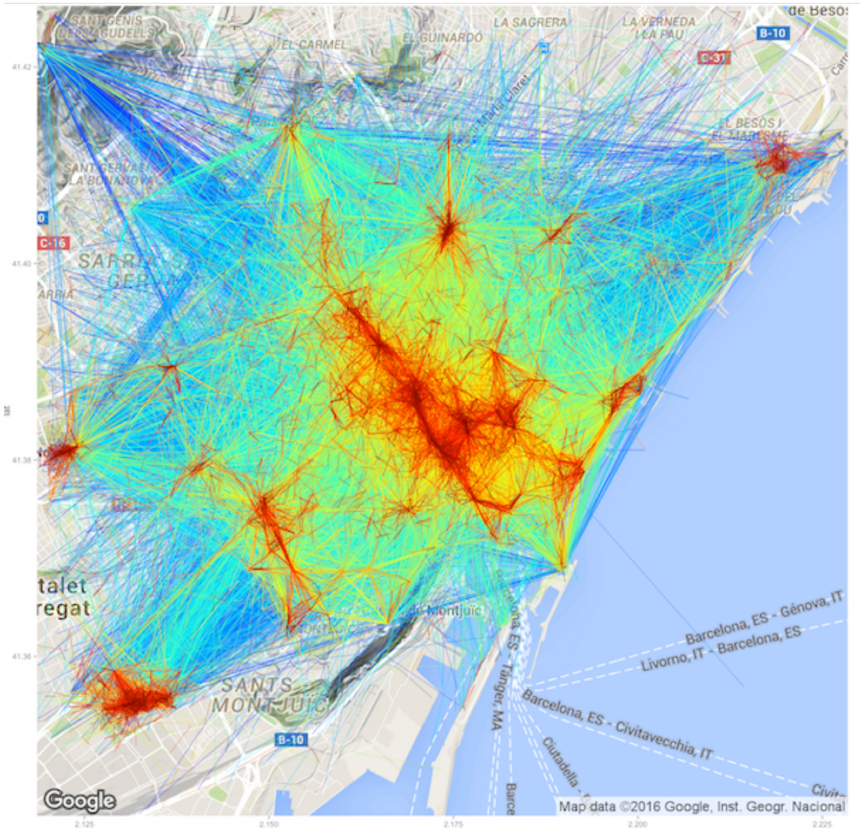
4. Pattern mining
Case study

- Result: **165,998** user paths
  - **41,626** performed by **tourists**
  - **124,372** by **local** citizens
- We plot the paths using *ggmaps*
  - An R library for the visualization of spatial data
- The **longer paths** have been represented with **cool colors** and the **shorter** ones with **warmer** colors

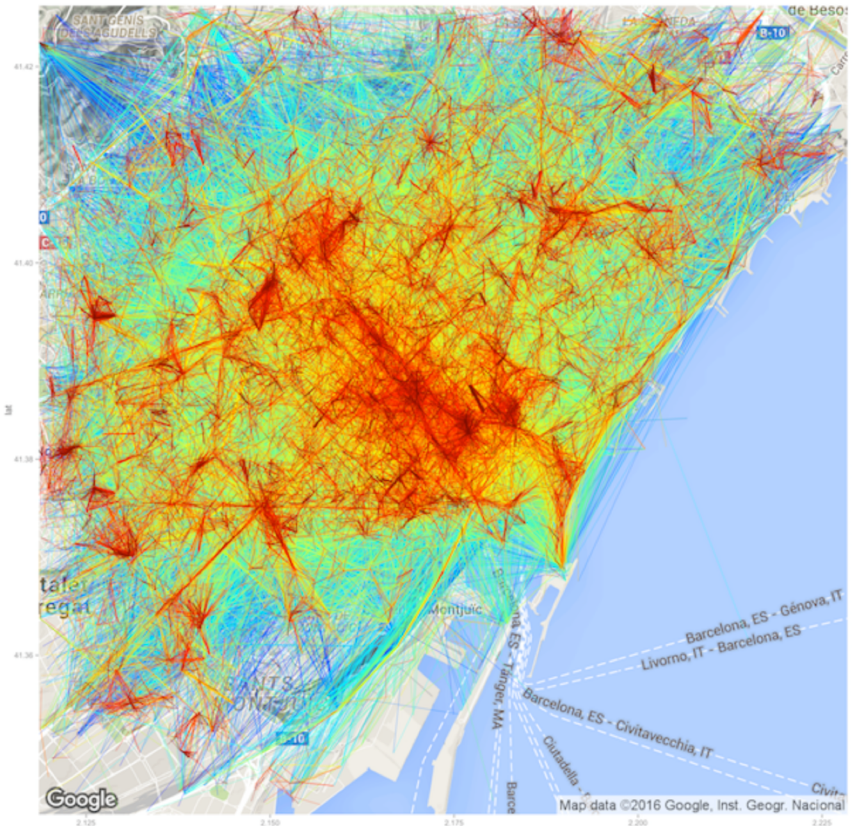
4. Pattern mining

Case study

- Tourists' mobility behavior based on one-hop paths



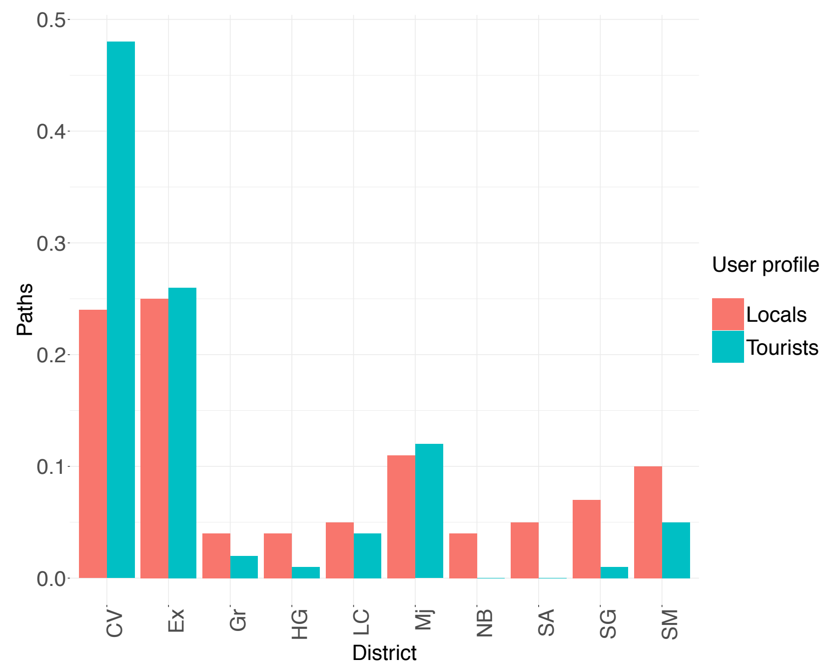
- Locals' mobility behavior based on one-hop paths



4. Pattern mining

Case study

- We also studied the distribution of tourists and locals in all districts of Barcelona
  - CV: Ciutat Vella
  - Ex: Eixample
  - Gr: Gràcia
  - HG: Horta-Guinardó
  - LC: Les Corts
  - NB: Nou Barris
  - SA: Sant Andreu
  - SM: Sant Martí
  - Mj: Sants-Montjuïc
  - SG: Sarrià-Sant Gervasi



4. Pattern mining

Case study

- We also analyzed how users move inside and throughout the districts of the city during working days and during the weekend

	Inside	Across
Locals weekends	0.36	0.64
Locals working days	0.38	0.62
Tourists weekend	0.39	0.61
Tourists working day	0.4	0.6

Mobile apps

Mobile apps

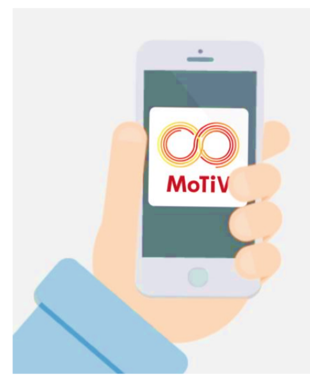
- Mining mobility patterns through mobile apps can offer additional insights
  - If the user interacts with it, we can get insights on her attitude towards mobility
    - Ratings, comments, answers to questions, …
- **MoTiV (Mobility and Time Value)**: ongoing Horizon 2020 project that will also try to extract patterns from data collected through a mobile app

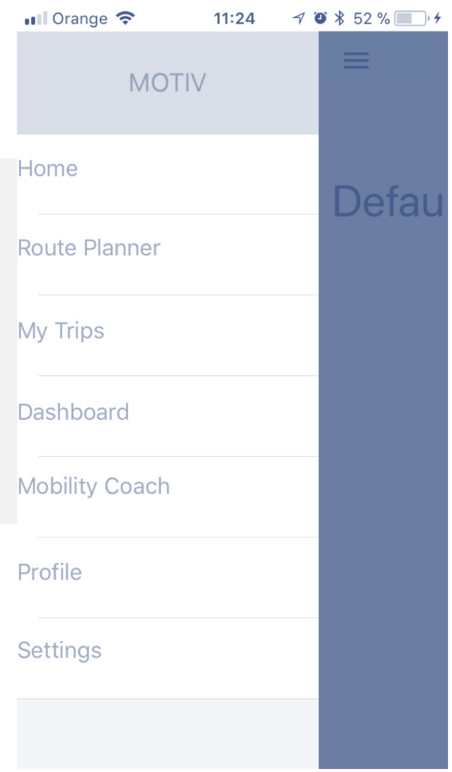
MoTiV Objectives

- To introduce and validate a conceptual framework for the estimation of value of travel time (VTT)
- Broaden definition of VTT beyond "time savings"
- Gain knowledge on traveler's reasons/purpose connected to the perceived value proposition of mobility
- Assess to what extent ICT connectivity and transport services/infrastructure affect VTT
- Provide specific actions and recommendations for all stakeholders (including end users) shaping the value proposition of mobility

MoTiV European-wide Data Collection

- Main types of data to be collected:
  - Mobility data
  - Experience (satisfaction) data
  - Activity data
  - Profile data
  - Attitudes data and
  - Insights data
- The final dataset will be open





Collecting the data from the users of MoTiV mobile App

>5.000 users from at least 10 European countries

Get involved

1 If you want to take part to the data collection and volunteer...

2 If you are interested in following our results...

Get in touch with us after the tutorial and visit

www.MoTiVproject.eu

Follow @MoTiV_Project

www.linkedin.com/groups/13568338

www.facebook.com/motivprojekt/

Open issues and future directions

Open issues and future directions

- Even though the number of advantages and possible solutions that can be developed to extract human mobility patterns is large, a set of open issues and research challenges still exists

Privacy issues

- Even though in case of social media people are aware of sharing personal information publicly, **privacy concerns** about collecting data without the users' consent exist
- To overcome this issue, approaches such as [Preotiuc-Pietro and Cohn 2013] **anonymize user** and **venues ids**
- However, aggregating the individual data points to extract information without the users' consent might still **violate privacy requirements**

Big data issues

- Both **social media** systems and **mobile** records generate data at a **very high rate,** leading to the widely-known ***big data problem***
- This might create **challenges** for real time storing, **processing**, and **indexing** of the data [Silva et al. 2014], which can have an impact on the mining of **up-to-date mobility patterns**

Data collection from third parties

- Mobile records are **not public**
- Also social media records can refer to **data that** is **not publicly available** (e.g., coming from apps like Waze)
- This might limit the **human mobility pattern** mining process

Low frequency of data sharing in social media

- Users share **geolocated** data with unequal **low rates**
  - On Twitter, **less than 1%** of published posts are geotagged
- This poses **challenges** on the analysis and **interpretation** of the data and on the ability to collect individual paths
- Some approaches tried to infer **geolocation** information exploiting **different features** of social media data [Backstrom et al. 2010, Kong et al. 2014, Han et al. 2014, Jurgens et al. 2015]
- **Incentive** mechanisms, such as **micro-payments**, are possible solutions being investigated for this issue [Silva et al. 2014]

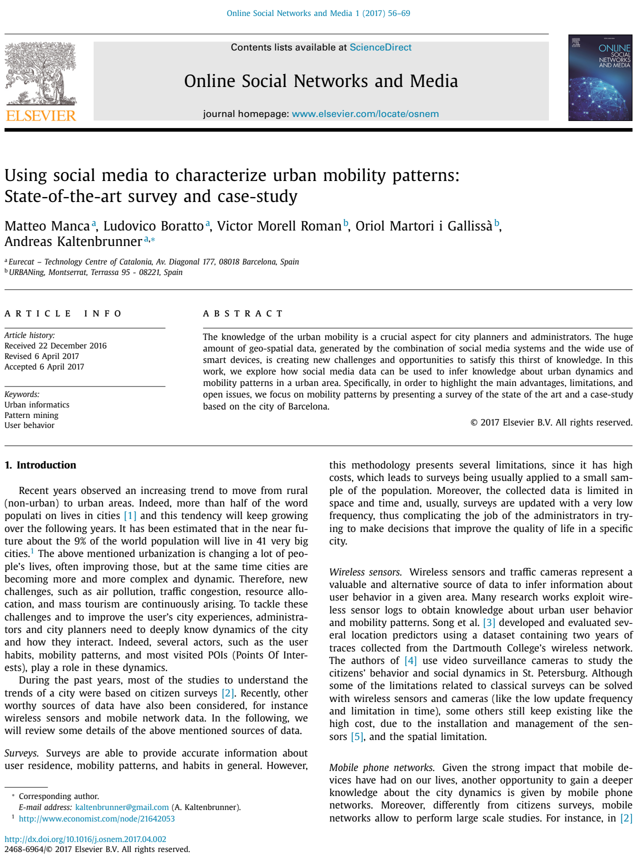
Conclusions

Conclusions

- The analysis of the state of the art and the conducted case-studies have **highlighted that social media and mobile** records may be valuable sources of data for **city planners, administrators, and urban scientists**
- At the same time, these relatively new sources are generating many challenges, which will be the **objective of future research**

Main reference

- The content of most of this tutorial (except the "Mobile records" part) is based on the paper: **"Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study"**, by Manca et al., published in *Online Social Networks and Media.*



Online Social Networks and Media 1 (2017) 56–69

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study

Matteo Manca[a], Ludovico Boratto[a], Victor Morell Roman[b], Oriol Martori i Gallissà[b], Andreas Kaltenbrunner[a,*]

[a] *Eurecat – Technology Centre of Catalonia, Av. Diagonal 177, 08018 Barcelona, Spain*
[b] *URBANing, Montserrat, Terrassa 95 - 08221, Spain*

A R T I C L E I N F O

Article history:
Received 22 December 2016
Revised 6 April 2017
Accepted 6 April 2017

Keywords:
Urban informatics
Pattern mining
User behavior

A B S T R A C T

The knowledge of the urban mobility is a crucial aspect for city planners and administrators. The huge amount of geo-spatial data, generated by the combination of social media systems and the wide use of smart devices, is creating new challenges and opportunities to satisfy this thirst of knowledge. In this work, we explore how social media data can be used to infer knowledge about urban dynamics and mobility patterns in a urban area. Specifically, in order to highlight the main advantages, limitations, and open issues, we focus on mobility patterns by presenting a survey of the state of the art and a case-study based on the city of Barcelona.

1. Introduction

Recent years observed an increasing trend to move from rural (non-urban) to urban areas. Indeed, more than half of the word populati on lives in cities [1] and this tendency will keep growing over the following years. It has been estimated that in the near future about the 9% of the world population will live in 41 very big cities.[1] The above mentioned urbanization is changing a lot of people's lives, often improving those, but at the same time cities are becoming more and more complex and dynamic. Therefore, new challenges, such as air pollution, traffic congestion, resource allocation, and mass tourism are continuously arising. To tackle these challenges and to improve the user's city experiences, administrators and city planners need to deeply know dynamics of the city and how they interact. Indeed, several actors, such as the user habits, mobility patterns, and most visited POIs (Points Of Interests), play a role in these dynamics.

During the past years, most of the studies to understand the trends of a city were based on citizen surveys [2]. Recently, other worthy sources of data have also been considered, for instance wireless sensors and mobile network data. In the following, we will review some details of the above mentioned sources of data.

Surveys. Surveys are able to provide accurate information about user residence, mobility patterns, and habits in general. However, this methodology presents several limitations, since it has high costs, which leads to surveys being usually applied to a small sample of the population. Moreover, the collected data is limited in space and time and, usually, surveys are updated with a very low frequency, thus complicating the job of the administrators in trying to make decisions that improve the quality of life in a specific city.

Wireless sensors. Wireless sensors and traffic cameras represent a valuable and alternative source of data to infer information about user behavior in a given area. Many research works exploit wireless sensor logs to obtain knowledge about urban user behavior and mobility patterns. Song et al. [3] developed and evaluated several location predictors using a dataset containing two years of traces collected from the Dartmouth College's wireless network. The authors of [4] use video surveillance cameras to study the citizens' behavior and social dynamics in St. Petersburg. Although some of the limitations related to classical surveys can be solved with wireless sensors and cameras (like the low update frequency and limitation in time), some others still keep existing like the high cost, due to the installation and management of the sensors [5], and the spatial limitation.

Mobile phone networks. Given the strong impact that mobile devices have had on our lives, another opportunity to gain a deeper knowledge about the city dynamics is given by mobile phone networks. Moreover, differently from citizens surveys, mobile networks allow to perform large scale studies. For instance, in [2]

* Corresponding author.
E-mail address: kaltenbrunner@gmail.com (A. Kaltenbrunner).

[1] http://www.economist.com/node/21642053

http://dx.doi.org/10.1016/j.osnem.2017.04.002

Acknowledgments

- This tutorial is partially supported by project **MoTiV** (Mobility and Time Value), funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 770145.

References

[Andrienko et al. 2013] G. L. Andrienko, N. V. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, D. Thom, Thematic patterns in georeferenced tweets through space-time visual analytics, Computing in Science and Engineering 15 (3) (2013) 72–82

[Backstrom et al. 2010] L. Backstrom, E. Sun, C. Marlow, Find me if you can: Improving geographical prediction with social and spatial proximity, in: Proceedings WWW '10

[Berlingerio et al. 2013] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, M. L. Sbo- dio, AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 663–666

[Barbosa et al. 2018] H. Barbosa, M. Barthelemy, G. Ghoshal, C.R. James, M. Lenormand, T. Louail, R. Menezes, J.J. Ramasco, F. Simini, M. Tomasini, Human mobility: Models and applications, Physics Reports (2018)

[Cao et al. 2015] G. Cao, S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang, K. Soltani, A scalable framework for spatiotemporal analysis of location-based social media data, Computers, Environment and Urban Systems 51 (2015) 70–82

[Cranshaw et al. 2012] J. Cranshaw, R. Schwartz, J. I. Hong, N. M. Sadeh, The livehoods project: Utilizing social media to understand the dynamics of a city, in: Proceedings of the Sixth International Conference on Weblogs and Social Media, 2012

[Fuchs et al. 2013] G. Fuchs, G. Andrienko, N. Andrienko, P. Jankowski, Extracting personal behavioral patterns from geo-referenced tweets, in: 16th AGILE Conference on Geographic Information Science, 2013

[Ferrari et al. 2011] L. Ferrari, A. Rosi, M. Mamei, F. Zambonelli, Extracting urban patterns from location-based social networks, in: Proceedigs of LBSN 2011

References

[Fiadino et al. 2017] P. Fiadino, V. Ponce-Lopez, J.A. Torrero-Gonzalez, M. Torrent-Moreno, A. D'Alconzo, Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era". In *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks* (Big-DAMA '17). ACM, New York, NY, USA, 43-48

[Frank et al. 2013] M. R. Frank, L. Mitchell, P. S. Dodds, C. M. Danforth, Happiness and the patterns of life: a study of geolocated tweets, Sci Rep 3 (2013) 2625

[Frías-Martínez et al. 2012] V. Frías-Martínez, V. Soto, H. Hohwald, E. Frías-Martínez, Characterizing urban landscapes using geolocated tweets, in Proceedings of PASSAT 2012

[Gabrielli et al. 2014] L. Gabrielli, B. Furletti, F. Giannotti, M. Nanni, S. Rinzivillo, Use of mobile phone data to estimate visitors mobility flows, in: Proceedings of MoKMaSD, 2014

[Girardin et al. 2007] F. Girardin, F. Dal Fiore, J. Blat, C. Ratti, Understanding of tourist dynamics from explicitly disclosed location information, in: Symposium on LBS and Telecartography, Vol. 58, 2007

[Gonzalez et al. 2008] M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, Nature 453 (7196) (2008)

[Han et al. 2014] B. Han, P. Cook, T. Baldwin, Text-based twitter user geolocation predic- tion, J. Artif. Int. Res. 49 (1) (2014)

[Hasan et al. 2013] S. Hasan, X. Zhan, S. V. Ukkusuri, Understanding urban human activity and mobility patterns using large-scale location-based data from online social media, in: Proceedings of UrbComp@KDD 2013

[Hawelka et al. 2014] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, C. Ratti, Geo-located Twitter as proxy for global mobility patterns, Cartogr Geogr Inf Sci 41 (3) (2014) 260–271

References

[Jiang et al. 2016] S. Jiang, J. Ferreira, Jr., M. C. Gonzalez, Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore, IEEE Transactions on Big Data, TBD-2015-12-0163

[Jin et al. 2016] L. Jin, X. Long, K. Zhang, Y. Lin, J. B. D. Joshi, Characterizing users' check-in activities using their scores in a location-based social network, Multimedia Syst. 22 (1) (2016) 87–98

[Jurgens et al. 2015] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu, D. Ruths, Geolocation prediction in twitter using social networks: A critical analysis and review of current practice, in: Proceedings of ICWSM 2015

[Kong et al. 2014] L. Kong, Z. Liu, Y. Huang, Spot: Locating social media users based on social network context, Proc. VLDB Endow. 7 (13) (2014)

[Lee et al. 2010] R. Lee, K. Sumiya, Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection, in: X. Zhou, W. Lee, W. Peng, X. Xie (Eds.), Proceedings LBSN 2010

[Liu et al. 2014] Y. Liu, Z. Sui, C. Kang, Y. Gao, Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data, PLoS ONE 9 (1) (2014)

[Long et al. 2012] X. Long, L. Jin, J. Joshi, Exploring trajectory-driven local geographic topics in foursquare, in Proceedings of Ubicomp '12

References

[Noulas et al. 2011] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare, in: Proceedings of the Fifth International Conference on Weblogs and Social Media (2011)

[Pappalardo et al. 2015] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, A.-L. Barab´asi, Returners and explorers dichotomy in human mobility, Nature Communications 6 (2015)

[Preotiuc-Pietro and Cohn 2013] D. Preotiuc-Pietro, T. Cohn, Mining user behaviours: a study of check-in patterns in location based social networks, in: WebSci '13

[Shelton et al. 2015] T. Shelton, A. Poorthuis, M. Zook, Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information, Landscape and Urban Planning 142 (2015) special Issue: Critical Approaches to Landscape Visualization

[Silva et al. 2014] T. H. Silva, P. O. S. V. de Melo, J. M. Almeida, A. A. F. Loureiro, Large- scale study of city dynamics and urban social behavior using participatory sensing, IEEE Wireless Commun. 21 (1) (2014)

[Song et al. 2010] C. Song, Z. Qu, N. Blumm, A.-L. Barabsi, Limits of predictabil- ity in human mobility, Science 327 (5968) (2010)

[Zhang et al. 2003] S. Zhang, C. Zhang, Q. Yang, Data preparation for data mining, Applied Artificial Intelligence 17 (2003) 375–381

DATA SOURCES AND TECHNIQUES TO MINE HUMAN MOBILITY PATTERNS

Ludovico Boratto (ludovico.boratto@acm.org)
Carmen Herrero (carmen.herrero@eurecat.org)
Andreas Kaltenbrunner (kaltenbrunner@gmail.com)
Matteo Manca (matteo.manca@zurich.com)
Giovanni Stilo (stilo@di.uniroma1.it)

The 5th IEEE International Conference on Data Science and Advanced Analytics
Turin, Italy – October 2, 2018