Urban Data Mining:

Social Media Data Analysis as a Complementary Tool for Urban Design

by
Nai Chun Chen

M.S. Architecture
National Cheng Kung University, 2012

SUBMITTED TO THE DEPARTMENT OF ARCHITECTURE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ARCHITECTURE STUDIES AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2016

Signature redacted

Signature of Author: _____
Department of Architecture
May,18,2016

Signature redacted

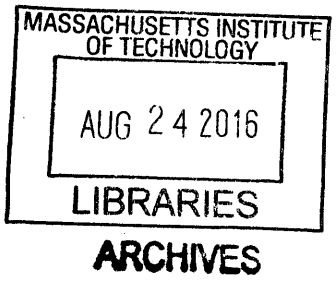
Certified by: _____
Kent Larson
Principal Research Scientist, Media Laboratory
Thesis Supervisor

Signature redacted

Certified by: _____
Takehiko Nagakura
Associate Professor of Design and Computation
Thesis Supervisor

Signature redacted

Accepted by: _____
Takehiko Nagakura
Associate Professor of Design and Computation
Chair of Department Committee on Graduate Students



Urban Data Mining:

Social Media Data Analysis as a Complementary Tool for Urban Design

by
Nai Chun Chen

SUBMITTED TO THE DEPARTMENT OF ARCHITECTURE ON MAY 18, 2016 IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN ARCHITECTURE STUDIES

Abstract

The emergence of "big data" has resulted in a large amount of information documenting daily events, perceptions, thoughts, and emotions of citizens, all annotated with the location and time that they were recorded. This data presents an unprecedented opportunity to help identify and solve urban problems. This thesis aimed to explore the potential of machine learning and data mining in finding patterns in "big" urban data. We explored several different types of user generated urban data, including Call Detail Records (CDR) data and social media (Crunch Base, Yelp, Twitter, and Flickr, and Trip Advisor) data on two primary urban issues. First, we aimed to explore an important 21st century urban problem: how to make successful "Innovative district". Using data mining, we discovered several important characteristics of "innovative districts". Second, we aimed to see if big data is able to help diagnose and alleviate existing problems in cities. For this, we focused on the city of Andorra, and discovered potential reasons for recent declines in tourism in the city. We also discovered that we can learn the travel patterns of tourists to Andorra from their past behavior. In this way, we can predict their future travel plans and help their travels, showing the power of data mining urban data in helping to solve future urban problems as well as diagnose and improve existing problems.

Thesis Supervisor: Kent Larson

Title: Principal Research Scientist in MIT Media Lab

Thesis Supervisor: Takehiko Nagakura

Title: Associate Professor of Design and Computation in Architecture Department

Reader: Michael Dennis

Title: Professor of Architecture Department

Acknowledgements

Thanks god, it is finished!

Appreciation for my thesis advisors, Professor Nagakura and Professor Larson for their profound instruction, encouragement, and patience, and for giving me this amazing opportunity to pursue research interests that I love.

Many thanks to my readers, Professor Dennis, and Professor Williams for their insightful advice, and knowledge that will guided me my whole life.

Many thanks to Professor Barzilay, Professor Mazereeuw, Professor Winston, Professor Jaakkola, Professor Jacobi, Professor Segal, Professor Welch, whose amazing courses I had the opportunity to take, and who took precious time to discuss with me and help me take my projects to a higher level.

Many thanks to my dear friends, Chin-Yi, Helen, Claire, Cindy, Yi-Min, Chin-Wen, Yu-Liang, Phil, Jean, Carson, Jason, Ira, RyanZ, Cheawen, Wenji, Difei, Agustina, George, Cynthia, and Michael Lin.

Many thanks to my UROPs, Marissa Stephens, Laura Peng, Sabina Chen, and Margaret Yu.

Many thanks to my friends in Smarts Urbanism group, ROCSA, and Media Lab Changing Place group

Many thanks to the Andorra Government for support. This work was part of the "Applied Machine Learning in Tourism of Andorra" project, which was supported by the Andorra Government.

Many thanks to Chen Sun for proof reading and idea discussions.

爸爸媽媽，姊姊弟弟，謝謝你們這三年來的支持，每次的越洋電話都讓我好想你們好想家。
謝謝你們，我愛你們！

Content

Chapter 1: Introduction

1.1 Research Questions and Goals

Swarm-scape: User-generated data for urban planning and design

The presence of web2.0 and traceable mobile devices create new opportunities for urban designers to understand cities by these user-generated data, including social media data and CDR data. The emerging of this "big data" has resulted in a burgeoning amount of information documenting daily events, perceptions, thoughts, and emotions of citizens, all annotated with the location and time they were recorded. This data presents an unprecedented opportunity for the purposes of gauging public opinion on the topic of interest.

In the past, urban designers seldom made use of (or had access to) such a large source of public data. In architecture of the 60s, modernist planners themselves pursued projects based on their own tastes and ideologies, and seldom consulted the public about the receptions of their designs.

For those who did make use of data to drive their designs, most made use only of government-collected data which are, in general, limited in scope.

However, emerging during this period of time was some critical urban theory that began to emphasize the relevance of social opinions as drivers of urban development. For instance, "The image of the city" by Kevin Lynch, "Death and Life of Great American Cities" by Jean Jacobs, and " City is not a Tree" by Christopher Alexander make compelling arguments for the use of bottom up social opinions. However, such work appears mostly ignored by planners at the time.

Even still, the methods advocated by these pioneering authors raise some practical problems. For example, in "The image of city", written by Kevin Lynch, the research method is to interview residents in the cities for learning what kinds of image have strong impression to residents. However, such traditional data collection methods, such as surveys and interviews, though insightful and personal, are slow and low-throughput in the data they produce.

The crux of the matter is that nowadays, such bottom-up techniques can be drastically

accelerated and supplemented by using geo-located data mining techniques of social data. These non-conventional data sources derived from activity registered on digital network offers extensive amount of information about human interaction. Furthermore, various statistical and machine learning techniques can, for the first time, enable analysis of this vast data, to deliver relevant conclusions. This is why I propose to make use of these user-generated data as an important complementary data source for urban designers in their analysis of cities, and this is why I propose to make use of powerful data mining and machine learning techniques, which will be explored and explained in this paper, to find patterns in such data.

1.2 Thesis Outline

Chapter two introduces a literature review, prior to discussing the Data mining and data visualization techniques we explored. In this chapter, the related literature discussion is divided into three subjects, including social media data analysis in related urban studies, CDR (Call Detail Records) data analysis in related urban studies, and analysis of innovation districts. Within the field of social media data analysis, researchers have tried to use novel ways of analyzing social media data in order to solve urban problems or to find interesting perspectives to read city life. Within the field of CDR (Call Detail Records) data analysis, papers in the field focus on using math methodologies and machine learning algorithms to find patterns of human activities. Finally, in studies of innovation districts, urban designers typically compare different types of innovation districts and categorized them into common types.

Chapter three explains the types of data we collected as well as the data visualization methods and different types of machine learning algorithms that we used. We describe our novel approaches for extracting social media (non-governmental) data and CDR data, which we believe is an emerging rich complementary source of urban data that is worth analyzing. We next show how to use statistical exploratory data analysis and geo-location based data visualization to represent the data. We then introduce different types of machine learning algorithms, including supervised learning methods and unsupervised learning methods, and finally, we describe how to apply them to find patterns by analyzing the data we collected.

Chapter Four titled "Swarm-scape 1.0: Data Mining Innovation Cities and Districts" explores the possibilities of using user-generated information, including geo-location data, text, and

images from different social media, such as CrunchBase, Yelp, and Twitter, to discover the characteristics of innovation districts.

Chapter five titled "Swarm-scape 2.0: Data Mining Andorran Tourism Patterns" collects data from social media, including Trip-advisor, Andorra Go! and Twitter, and overlays these POI (point of interest) data with CDR (call detailed records) data. Combing theses different datasets, we utilize statistical exploratory data analysis and machine learning algorithms in order to find the travel pattern of tourists from different countries.

Chapter six explores how these new data types and analysis methods help urban designers understand the city in a novel way. Conclusion are finally presented, and future directions described.

1.3 Work Distribution

The main ideas and concepts of this project were proposed by the author (NC).

Chapter 4 and chapter 5 introduce two projects which integrate different types of user-generated data with various statistical EDA and machine learning methods. The coding part was done by all participating members of this project's team, as follows:

The coding part of "Swarm-scape 1.0: Data Mining Innovation Cities and Districts" (chapter 4), is divided into three parts: data collection, data mining analysis and data visualization. NC collected all of the data from different platforms of social media, and did most of the data-mining analysis and data visualization. Rachael, a PHD student in the computation group, wrote the sentiment analysis code.

The coding part of "Swarm-scape 2.0: Data Mining Andorran Tourism Patterns" (chapter 5), is divided into three parts: data collection, data mining analysis and data visualization. The CDR data, Trip-advisor data in city-scale, and civic data from "Andorra Go!", were collected by the author. The Trip-advisor data in country-scale is collected by Marissa, a UROP in Media Lab "Changing Places" group. Marissa also helped the author to visualize the data in "Association Rule" method and "Origin-Destination" matrix. Another UROP, Laura Pang, helped the author develop the daily timeline of CDR spatial distribution application. The completion of this projects was the result of mutual cooperation and synergies within this team, and especially, these two fantastic and enthusiastic UROPs.

Chapter 2: Literature Review

2.1 Social Media Data in related Urban Study

In the thesis "Power Centrality as a Relational measure of Urban Hierarchy. Testing the Splintering Urbanism Theory with Social Media data from Santiago de Chile", Humeres M uses Power Centrality algorithm applied to different twitter statuses datasets generated by Metro users. In this work, the author analyzes data from social media data to identify spatial concentration, for proving the theory of Power Centrality, which is a relation measure of urban hierarchy, showing the importance of nodes within a network. The result evidenced how Metro could act as a mass public bypass that connects emerging centralities. (Humeres M, 2014)

In the thesis "Data Visualization and Optimization Methods for Placing Entities with Urban Area," Pranav Ramkrishnan use data-driven maps to reveals different urban environment ideas in seven projects, including Street Greenery, Bicycle Crashes, Footfall Density, Public-Private Transportation Efficiency, sky Prints Best Modes of Travel. In each project, the data-visualization map intends to provide unique perspectives about how cities work. The main contribution in this thesis is how optimization algorithms facilitate data visualization for analyzing urban environment, and, for identify possible areas of change in the city. Because this is a research paper from the computer science field as opposed to urban studies, this thesis is concerned with more technical issues than most other papers I have reviewed.

In the thesis " Seeing differently: cartography for subjective maps based on dynamic urban data," Xiaoji Chen creates a dynamic subjective map with animation and interaction, showing how representing cities through novel data visualization might be done. The dynamic data-driven map represents urban data, including transportation time, population density, and social connection in geographic space. This project alos embedded human activities within th GIS data, providing a new perspective in reading a city. (Chen,2011). Finally, Liu et al, 2014 utilized image analysis to deeply study the layouts of urban environments, and successfully used such techniques in a project called "C-IMAGE" to classify different types of urban changes, and correlate these different types of changes with indicators. Frias-Martinez et al 2014 proposes the use of geo-located tweets, coupled with the proposed use of clustering algorithms, to answer questions about land use.

2.2 CDR (Call Detail Records) data analysis in related urban studies

The CDR data analysis literature can be broadly understood to be of three different sub-topics. All of them have a common theme in their attempt to utilize novel machine learning and data mining/visualization techniques. Though this field remains young, due to the relative recency of big data collection, and privileged, due to the scarce availability of CDR data to the general public, researchers are increasingly interested in what kinds of conclusions can be made through the analysis of CDR data. The three sub-topics are as follows:

First, there is a literature trying to introduce and develop new methods of data acquisition, mining, and analysis into this field. For instance, Calabrese et al, 2013, presents some techniques for extracting data from mobile phone traces, to try to answer questions about human mobility. Arun J et al attempts to demonstrate the promise of using CDR data to make socio-economic inferences. In general, although such papers provide an important service in proposing new methods for others to use, they often, unfortunately, fall short in convincingly demonstrating the statistical and inferential power of their proposed techniques. This probably reflecting a mathematical tepidity, and instead, these authors choose to focus on demonstrating beautiful visualizations. Ultimately, they often fail to extract serious conclusions from their data.

Second, there is a literature focused on seriously testing and validating the differential use of machine learning and data mining techniques in analyzing CDR data. For instance, Ravichandran et al, 2012 compared different decision tree techniques, and random forest techniques in analyzing the behavior of mobile operator customers, and determining the respective advantages of each technique. Liu et al, 2013 attempts to build optimal classification models based on different known machine learning algorithms, to analyze human travel patterns. Hoteit et al, 2014 compares different types of interpolation methods (linear, cubic, etc) in the analysis of mobile phone data in inferring human trajectory, and concluded differential circumstances in which each interpolation method is most useful. This literature provides an important service in looking deeper at the various machine learning techniques that are proposed, and analyzing the pros and cons of each method to gauge its utility to the field.

Finally, there is a literature focusing on seriously applying tried and tested machine learning techniques to CDR data analysis to gauge new types of conclusions from CDR data. For instance, Doyle et al, 2011 analyzed mobile phone billing records in the Republic of Ireland using inference methods to discover distinct "modes" of travel of citizens. Tatem et al, 2014 applied validated data visualization techniques to help identify and evaluate at risk communities for malaria. Cavia, 2010 utilized spatial and network analysis to deeply interpret the social life of cities. The author specifically focused on contrasting the known social outlooks of three major European cities and analyzed them in terms of illustrative metrics from network theory, such as polycentricity, fragmentation, and centralization. Finally, Hager, 1999 was able to utilize statistical analysis of CDR from a live ATM network for some very practical applications: to plan network capacity and traffic. This final subtype of the literature is perhaps the pinnacle of this field, in trying to demonstrate the true utility of CDR data and associated machine learning algorithms, in extracting powerful and meaningful conclusions and predictions from the data.

2.3 Innovation Cities and Districts

In "Innovative Cities", James Simmie focuses on international comparisons of innovation and attempts to encourage high-technology industries to develop in particular European cities. He uses both urban economic and social perspectives. For the purpose of this work, innovation is defined as " new commodities, new technologies, new sources of supply and new types of organization"(Simmie, 2001) The author compares and contrast the contributions made in several European Cities, including Stuttgart, Milan, Amsterdam, Paris, and London. He analyzes the progressive changes of several urban economic variables, including employment rate, industrial types, and the number of firm establishment. He uses Agglomeration theory and modern evolutionary theory to examine this new industrial geographies. After comparing different cities, this research shows the shifts of agglomeration advantages from traditional large cities based on input and market scale economics to diversity, information exchange and knowledge intensive markets. The different attributes in each innovation city can be understood in its particular history and international trading connection. (Simmie, 2001)

In "The Rise of innovation Districts: A New Geography of Innovation In America," Bruce Katz and Julie Wagner argues that a new innovation urban model is emerging, and that this model is different from traditional innovation districts such as Silicon Valley which has

inefficient transportation systems, inefficient land use, extensive sprawl and continued environmental degradation (Katz and Wagner,2010). The new Innovation Districts present a denser pattern of local amenities, residential and working spaces, a well-developed mass transit, and re-growth in city centers. The research divides the innovation districts into three types: Anchor plus model, re-imagined urban area, and urbanized science park. Anchor plus models usually appear in downtown and mid-town of cities; re-imagined urban areas are often in the older industrial areas, close to downtown or waterfront areas, and urbanized science parks are usually located in suburban area. (Katz and Wagner,2010)

In the thesis " Spatial qualities of Innovation Districts: How Third Places are Changing the Innovation Ecosystem of Kendall Square" Kim uses Kendall square as a case study area, revealing the workplace in high tech clusters expanded from conventional office space to retail and public space. Subsequently, the work provides some recommendation in stimulating social interaction and collaboration from policy and design perspectives, in order to establish better ecosystems in innovation districts.

Chapter 3

We will first discuss the technical aspects of our data collection methodologies, as well as the key importance of the data types collected:

3.1 Data Collection

Data types used in this thesis include data from social media, WIFI data, and CDR data. All these different types of data required different and specialized methodologies to retrieve them.

CDR data was provided by the "Andorran Telecom" company. All the other types of data (the social media types) had to be retrieved algorithmically. In this section, we introduce different pipelines to gather data from these various platforms of social media. This is a very new method of gathering data in the urban design field. Typically, researchers obtain data directly from people and sources (e.g. via surveys, interviews, or direct measurements). In this thesis, we utilize the novel and emerging techniques of web-scraping and API, to gather social media data.

Web 2.0 and opportunities for data collection

Although traditional data collection methods such as surveys, interviews, and, more recently, data harvesting and analysis techniques have provided interesting insights into the social life of urban space, nowadays, they can be complemented by data from geo-located social media.

Web 2.0 describes a very exciting new model web platform that is dominated by "user-generated content", that is, content that is generated by individual web users. The emergence of Web2.0 provides a golden opportunity not only for users to generated data, but also for researchers to collect data. This is what we aim to do.

Web2.0 is characterized by users sharing their opinions and exchanging information by video, images and texts, giving a potentially very large and rich source of data that we may mine.

As of May 2015, Twitter has 500 million users. These hundreds of millions of users, each with many individual relations, form a complicated graph of hundreds of millions of nodes

and edges that define the giant internet community. The key important feature of these online networks is that we may exploit the API (Application Programming Interface, API) that the network provides, which allows us to obtain information in a very easy way.

Social Media APIs

"API"s, also known as application programming interfaces, are software systems that link smaller components of a software together into a bigger whole. The intuition is that due to the increasingly large scales that computer algorithms have become in recent years, it is often necessary to divide a complex system into smaller, more manageable parts. The APIs allow researchers to retrieve information in manageable chunks.

Twitter API, for instance, includes two RESTful APIs, including REST API and Search API. The Twitter REST API methods allow developers to access core Twitter data. This includes update timelines, status data, and user information. The Search API methods give developers methods to interact with Twitter Search and trends data. The API presently supports the following data formats: XML, JSON, and the RSS and Atom syndication formats, with some methods only accepting a subset of these formats.

Web Scraping

"Web Scraping" is the use of algorithmic code in order to extract useful information from websites. The importance of this alternative pipeline is because some social networking platforms do not provide an API, or if it is difficult to obtain the desired information from their API. For instance, CrunchBase did not provide geolocation data in their API, so we had to crawl pages in their websites, in order to achieve our analysis goals.

Data held within websites tend to be written in a unified language such as HTML. "Web crawling" takes the URL (Universal Resource Identifier) address of websites, and extracts the HTML code behind that webpage in order to mine it for useful information.

The mining of webpage HTML data requires extensive methodologies. Luckily, several different types have been developed and accumulated into "libraries" that are available online for programmer use. The most popular methods library for web crawling is a library called "Beautiful Soup." In this thesis, "Beautiful Soup" is the library that we mainly use.

Beautiful Soup can extract data from HTML or XML files via Python library. Beautiful Soup parses complex HTML document into a tree structure: each node is a Python object, and all objects can be grouped into four types: Tag, NavigableString, BeautifulSoup, and Comments. Hence, using Beautiful Soup to do web crawling allows us to gather more information we need from HTML, which provides data to that which we gain from API.

3.2 Data Mining Methods and Data Visualization

After gathering data from various resources, the most important question we need to ask is why this data is interesting to urban planning and how can we find patterns within the data.

Data Visualization

A crucial reason why we have collected from (user generated) social media is that they contain accurate geo-location information of where that data was generated. This is crucial for the urban planner in order to understand not simply the events that occur in the city, but more importantly, *where* they occur, that is, the spatial distribution of events. We applied several pipelines of data visualization techniques to our accumulated data. First, we simply plotted our data using exploratory data analysis, such as histograms, pie-charts, and time series. Second, we created an interactive map, using a web-based interactive methodology, in order to more deeply visualize and understand the temporal dynamics of the data. In this project, we make use of Cartodb, D3.js, and google map API. Combining statistical analysis with geo-spatial analysis allow us to understand a city in both abstract and geographical ways.

Data Mining Method: Natural Language Processing

"Natural languages" refer to languages that people use in daily communication, such as: English, Chinese, and French. In contrast to artificial languages such as programming languages and mathematical languages, natural languages constantly evolve over the generations. This is the reason why it is so difficult to determine the clear rules that govern the construction of a natural language. "Natural language processing" is the group of algorithms that try to patterns within natural languages, and subsequently, try to interpret the text. For example, people have used techniques in natural language processing to find the most popular topic by analyzing the frequencies of words. More complex phenomena have also been analyzed, for instance, trying to decipher the sentiments of a sentence.

Chapter 4 in this project, titled "**Swarm-scape 1.0: Data Mining Innovation Cities and Districts",** utilizes a python library, called natural language toolkit (NLTK), to analyze tweets

from Twitter API. NTKL is an open source library, which contains software, data and documents. All of the information could be download from http://www.nltk.org for free.

Data Mining Method: Association Rules

"Associations" are cooperative links between two entities, and are the most basic type of relationships. This project make extensive use of "association rule" analysis to look for these basic relationships within our data.

A well-known story regarding association rules that have been discovered in business is the "Beer and diaper" association. In 1990, a manager in Walmart found an interesting phenomenon in analyzing sales' data: when young fathers went to the supermarket, they often bought beer and diaper at the same time. Inspired by this pattern, Walmart subsequently placed beer and diaper merchandise together. The relationship between beer and diapers is an association.

Association rule analysis is one of the core techniques of data mining. The use of association rules in data mining was first proposed in 1993 by R. Agrawal in IBM's Almaden research center. A classical association rule algorithm often implemented is called the Apriori algorithm, which has had a great influence in the field. Many online shopping websites, such as Amazon and eBay use this algorithm to recommend items for their customers (this is the algorithm behind the function "buy this product also buy that product at the same time" on their websites).

The intensity of an association is decided by three core concepts: support, confidence, and lift. For example, there are 10000 rows of dataset, and 1000 of them bought diaper, 500 of them bought bread, and 2000 of them bought beer. In common dataset part, 800 out of 10000 bought diaper and beer at same time; 100 out of 10000 bought bread and diaper at the same time.

"Support" means the probability of {x,y} appear at the same time in all of the datasets.
$Support(X \rightarrow Y)=P(X,Y)$

"Confidence" implies probability of {x,y} appear given x appears.
$Confidence(X \rightarrow Y)=P(Y|X)=P(X,Y)/P(Y)$

"lift" is a complementary measure to confidence. "lift" compares the two situations, the first one is people buy A and B at the same time, the second one is people did not buy A but buy B at the same time.

Lift(X→Y)=P(Y|X)/P(Y)=Confidence(X→Y)/P(Y)

Data Mining Method: Clustering Analysis

Clustering analysis is a method which divide dataset into various groups based on similarity and dissimilarity. It simultaneously is a data visualization technique, and also brings forth potential patterns (clusters) within the data. Clustering analysis is an unsupervised learning method. There are five most popular clustering algorithms: K-means, K-medoids, Density-based spatial clustering of application with noise, Hierarchical Clustering, and Expectation Maximization (EM). The three main methods we used are as follows:

K-means: "k" points are randomly picked from the dataset as starting centroid points, and the rest of the points are assigned into the clusters based on highest similarities. Next, the mean of the points' within each cluster in Euclidean space is defined as the new centroid points, and the process is iterated until convergence (no more changes in points' membership into the clusters).

K-medoid: is a clustering method very similar to k-means. The only difference is that instead of assigning the mean value of the points' in Euclidean space as new centroid points, we assign the minimum distance to the rest of the sample as new centroid points. The procedure is iterated until convergence.

Hierarchical Clustering: is a clustering method in which one does not need to set k clusters *a priori*. The algorithm is iterated, with each iteration combining pairs of closely related clusters into a single cluster, so that a tree-like diagram (dendrogram) of clusters is made to illustrate the relations found within the data at every scale.

Data Mining Method: Random Forest

Comparing to previous two machine learning algorithms (association rules and clustering), Random Forest is a relatively new algorithm. Random forest also avoids problems associated with another emerging technique: neural networks. Neural networks, though highly accurate and predictive once trained, are computationally intensive; random forest does not

suffer from this. The Random forest algorithm was developed by Breiman, 2001, who modified a (classical) Classification Tree algorithm into the Random Forest. Classification trees suffer the problem of inaccuracy, due to outliers during training. Random forest makes numerous classification trees based on (randomly chosen) subsamples of the given data, and then collects the results of each of these individual classification trees, and outputs a result based on majority vote. This thus produces a result that is simultaneously highly accurate (like the neural network and unlike the original classification trees), while also being computationally efficient (like the classification trees and unlike the neural network

Chapter 4: Swarm-scape 1.0: Data Mining Innovation Cities and Districts

4.1 High Tech Cities and Innovation Districts

Importance of Innovation cities and districts

A remarkable shift is occurring in the spatial geography of new and existing cities: the emergence of "new high tech" cities or so called "innovation cities". The development of such cities reflects emerging trends in the global society: the increasing importance of high technology industries and products. These trends are creating profound changes in society. To catch wind of these untapped economic opportunities, city planners and urban designers are beginning to incorporate the unique demands by innovative cities into city designs. However, such efforts have had mixed success. For instance, despite the economic success of Silicon Valley, suburban corridors of spatially isolated corporate campuses make for very inefficient transportation, and urban designs there do not place emphasis on the quality of civic life or on integrating housing and amenities.

Traditional High Tech City----San Jose

New High Tech City----New York

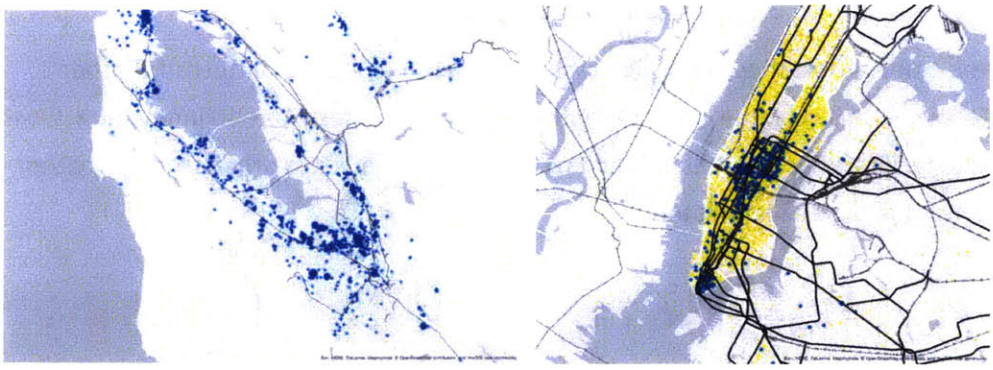


***Figure 4.1.1**: Comparing Spatial Distribution of Companies in Traditional and New High Cities*

A new urban form is emerging, give rise to what we and others are calling "innovation districts." Innovation districts are placed within the new high tech cities, and have the unique potential to spur productive, inclusive and sustainable economic development. They provide a strong foundation for the creation and expansion of firms and jobs by helping companies, entrepreneurs, universities, researchers and investors. An investigation into the attributes that facilitate the success of innovation districts could be very beneficial as these innovation districts become more and more numerous.

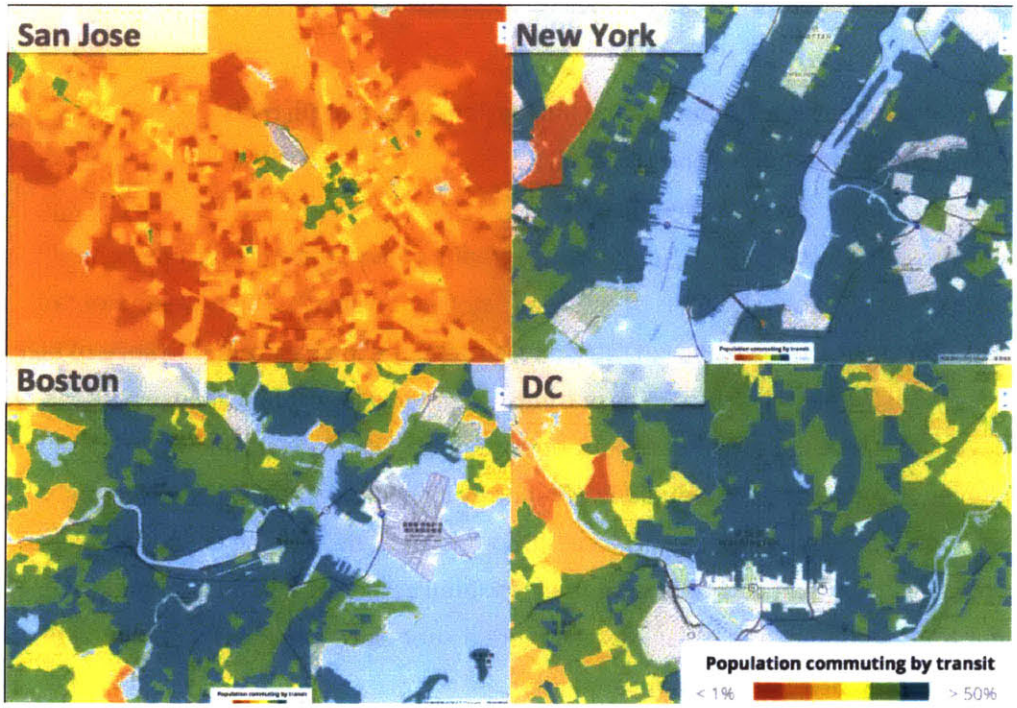


***Figure 4.1.2**: Comparing Population Commuting by Transit in different cities*

While some cities shrink, these new high tech cities become popular patterns for other cities to follow. In this way, the question "what are the attributes of urban innovation districts that facilitate their success?" becomes a critical issue in this urban transformation. It is the topic of this present chapter.

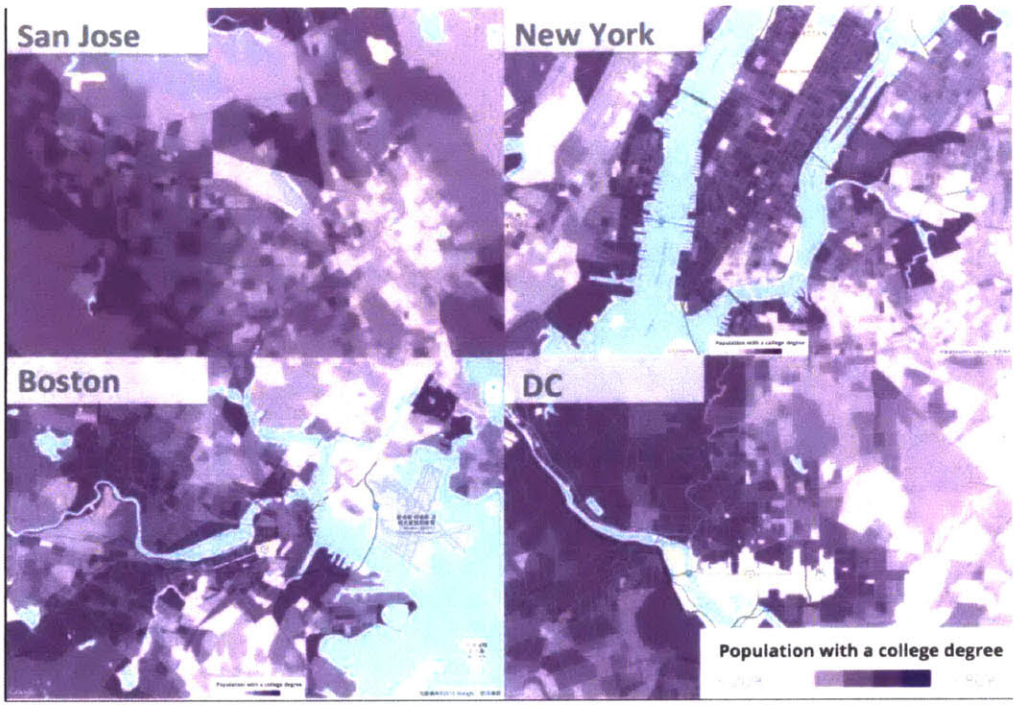


***Figure 4.1.2**: Comparing spatial density with college degree in different cities*

Comparing to traditional high tech city, such as San Jose, new high tech cities, such as Boston, New York, and D.C, have much more high ratio of population commuting by transit. In addition, these cities also have a common phenomenon, high educated people tend to live in the city center area, which has high density.

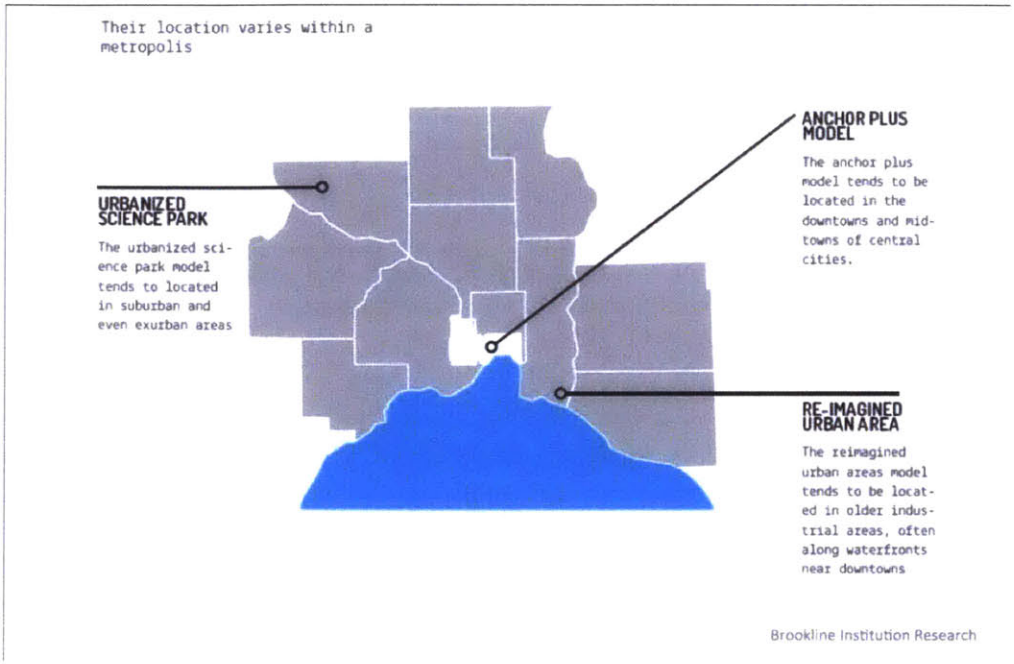


***Figure 4.1.3**: Different types of new innovation districts (Image from: Brookline Institution Research)*

Definition: Innovation district

Before discussing further, we must agree on definitions in our discussion of innovation, to have consistency. Here, the primary innovative entity that I will discuss will be the "innovation district", a zone of a certain scale, with certain attributes that define it:

I utilize an "innovation district", according to Bruce and Julie (2010), as a "geographic area where leading –edge anchor institutions and companies cluster and connect with start-ups, business incubators and accelera
tors. They are also physically compact, transit-accessible, and technically-wired and offer mixed-use housing, office and retail". (Bruce and Julie,2010).

***Figure 4.1.4**: Four Examples of different types of new innovation districts (Image from: Brookline Institution Research)*

Previous work

Studying the attributes for the success of the Innovation district is an important topic, and some work has progressed to understand these attributes. For example, "Innovating the City: Challenges and Opportunities in establishing incubators and Districts in Paris and Boston" by Karen Johnson is a Masters Thesis in Urban Planning at MIT that is very much related to this topic. This thesis offers insights on the unique political, economic and cultural systems that shape innovation strategies in innovation cities. The study utilizes in-depth interviews of various Innovative zone design projects to gain insight into these factors. But, in order to give translatable and design-usable advice for Urban designers in Innovative districts, one must dig deeper in understanding the city attributes that facilitate Innovation. But for this, the contemporary studies done so far, in my view, are not enough.

Particularly lacking in the study of the Innovative district is an understanding of how the Citizen who lives there perceives the changes. Ultimately, positive changes must come to the Citizen. But, whether citizens have benefitted directly or not, and a deeper understanding of how they have benefitted, is unknown. This is due to the inherent limitations of the types of data used so far. In the Johnson study above, interviews, though in-depth, are still limited in scope as they occur only to the top companies participating in a project. Thus, they project only one point of view. Other studies appear to be restricted to

government sourced data (e.g. census data) which, while informative, remains restricted by the types of data that the government has collected.

4.2 Social media data and Innovation districts

This chapter makes use of web-based API, particularly social media, which we hypothesize can shed enormous insight into the consequences of urban designs in the innovative city, from the perspective of the citizen. We will then mine it using techniques from machine learning and inference in order to more deeply study the features of districts that are most critical for Innovation.

Methods:

For a reasonable sample number, four new high tech cities will be considered: Boston, New York, the District of Columbia, and Los Angeles. This way, conclusions obtained can be tested for generality, and will not be due to the special circumstances of a specific region. At a lower urban scale, three Innovation districts will be considered: Kendall Square innovation district in Cambridge MA, South Boston Innovation district, and Silicon Alley in New York.

Within these zones, both government GIS data as well as web API data were collected, and analyzed, taking into special account temporal relations and location information.

GIS data was mined for the diversity and the density of offices, jobs, people, income, rental prices from census to examine relations among these metrics and their relation to the success of the Innovative zone.

Besides the government GIS data, several social media sources were mined, for several different purposes. **Twitter** content and the associated geo-locations was mined to identify real time activities of residents and where they are occurring. Data from **Google maps API** was mined for attributes (e.g. density) of local amenities, including cafes, restaurants, shops, and bars. **Flickr** images as well as the captions and comment contents was mined for identification of the popular use open space and city perceptions by residents.

Mining and gathering data from web-based API in each district will indeed yield an enormous amount of data. Thus, the next difficult technical challenge is how to treat all of

this data to make meaningful conclusions. We will apply data mining techniques:

RESULTS AND DISCUSSION

Data mining the text, image, and geographical location with WEB API

1. Geo Location Mining

- ***Activities***
- ***Amenities***
- ***Startup location***

2. Text Mining

- ***Trend***
- ***Feeling of Area***

3. Image Mining

- ***Usage of Urban Space***
- ***Life of Residents***

The goal in this Chapter is to discover the features of districts that support innovation. More specifically, the aim is to investigate the use of user generated data as a source of knowledge for designing innovation districts, and make recommendations for the public policy, design, and infrastructure that cities should put in place to increase the potential for "innovation."

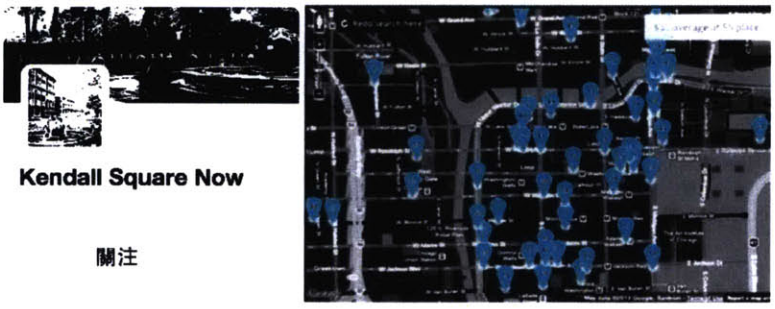


***Figure 4.2.1**: Three main types of user generated data could be gathered from social media: Geolocation data, text, and Image.*

4.3 Start-scape: Geo-location Mining with Crunch Base and Yelp

We makes use of data with geo-location information, in order to understand the spatial distribution of events. For example, we could investigate a startup company's growing pattern by observing geolocated changes and events through the years. We can find restaurant clustering areas by geo-location data, and we can search for particular daily events by twitter geo-location data.

Crunch-Base as Innovation Trend Detector

Crunch-Base is a Web 2.0 platform that has become increasingly popular with new businesses. It allows startup companies to register, and update their company data, location, categories, founded time, employment data, etc. It also became a key resource for us, owing to its wealth of publically available information. Using the Crunch-Base Web API, we compiled information regarding startup companies in the Boston region.

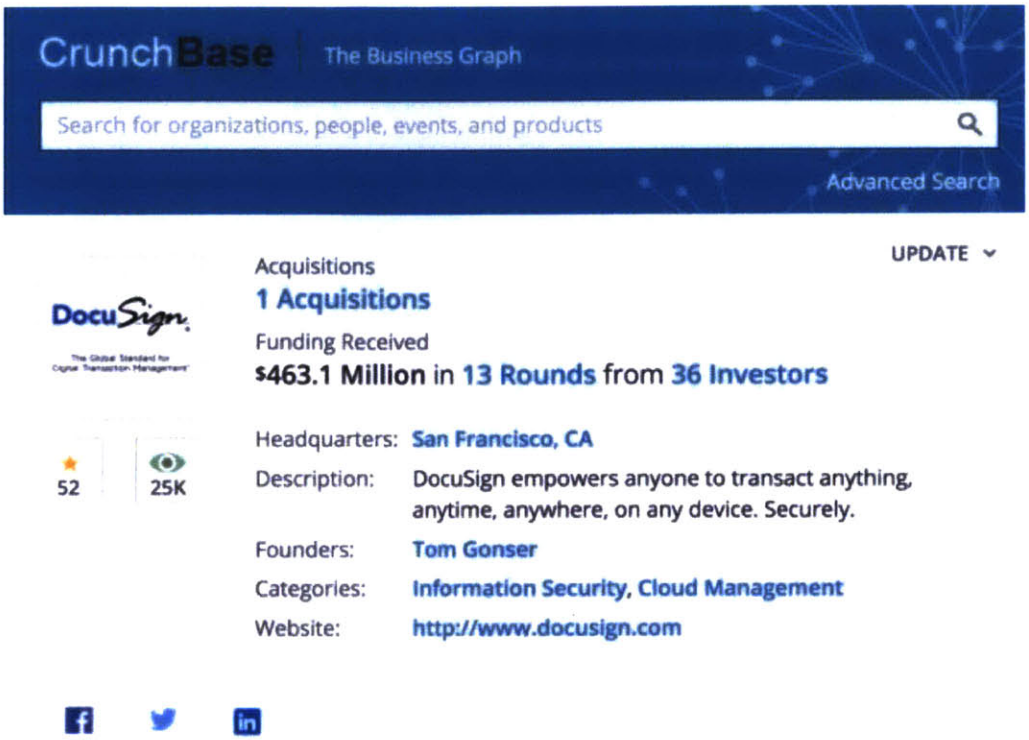


***Figure 4.3.1**: Three main types of user generated data could be gathered from startup website" CrunchBase": Geolocation data, Company Category, and Founded time.*

From the data collected from Crunch-Base, we were then able to apply visualization techniques to examine the spatial patterns of high tech startup tech companies in the Boston area. Several findings can be seen from our visualizations.

In general:

Traditional tech companies in the 20th century were located in suburban areas, to take advantage of cheaper land and cheaper labor wage. It can be seen, however, that new high tech companies are more often located in the city center in the Greater Boston area, rather than the suburbs. We can speculate that this is in order to attract young and talented people graduated from universities around the city. This observation also suggests that the transformation of local industries through the generations have also changed the preferences of new company locations. The other mainly factor for where these new high tech companies choose to localize themselves is proximity of MBTA, presumably because the company is then able to attract labor forces that commute by public transportation system.

***Figure 4.3.2**: Based on the geolocation data and company category from CrunchBase, the map shows the pattern of distribution of high tech company.*

If we zoom into the Boston and Cambridge area, we may see several of these patterns in greater detail.

1. Proximity to MBTA

Data visualization from CrunchBase reveals the relationship between startup companies, their Head Quarters, and the MBTA. These startup companies tend to be located close to the MBTA station, such as Kendall Square, Central Square, Financial District, Beck Bay, and the South Station area. This patterns occurs presumably because the company is then able to attract labor forces that commute by public transportation system.

2. Locations of the "big" high tech companies

The "big" tech companies, such as Facebook, Google, and Microsoft, seem also to have become critical attractors for smaller startups to choose their locations. There are two possible reasons: First, there could be a "flow of labor" from one campus to the smaller ones. Many of the members of these new startup companies might have been former employees of the large tech companies. Second, there could be a "flow of ventures." The big tech companies tend to invest in the startups that they better understand; local startups thus have an advantage in these critical social relations.

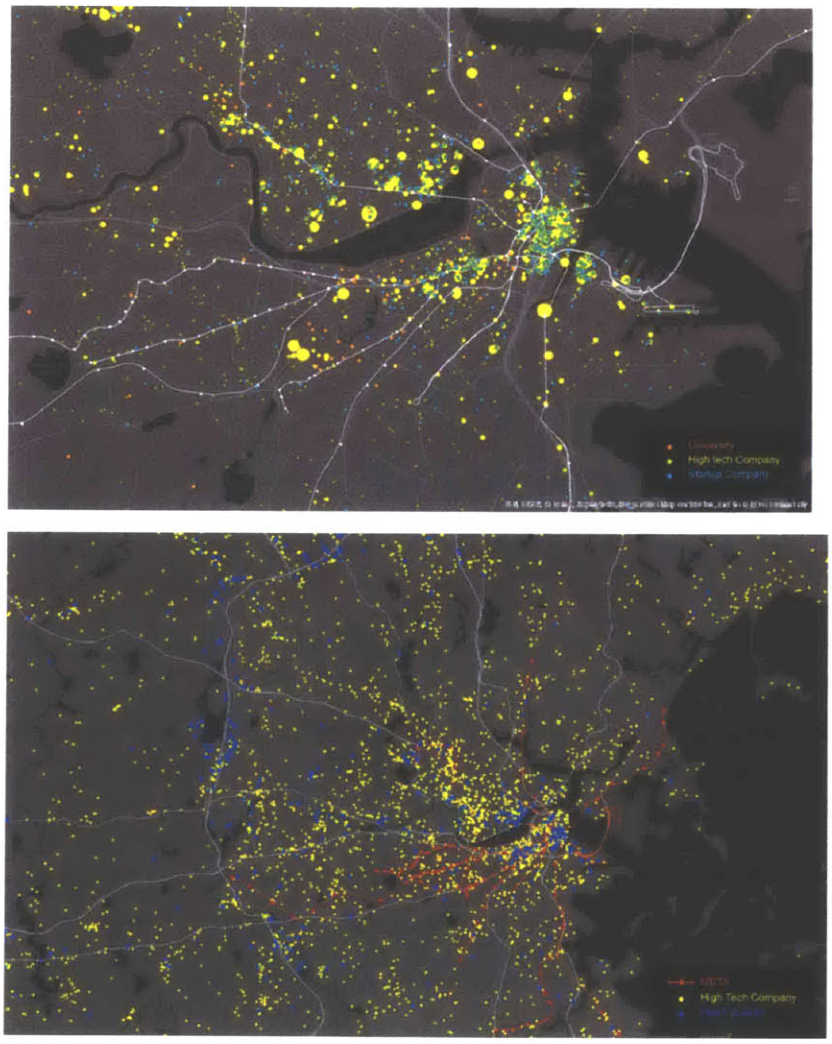


***Figure 4.3.3**: Overlaying with transportation facilities, the map shows start up company tends to located close to T station and Head Quarter.*

3. Locations of the institutions

In the Boston and Cambridge area, it can be seen that research Institutions also seem to serve as anchors for startup clusters. We can see that Harvard, MIT, BU, and many of the medical institutions have startup clusters nearby. For instance, when talented graduate students and professors from local institutions (such as MIT) invent novel technologies, they often establish a startup company close to their base operations (their labs, which are located locally). In this way, they are able to manage both their company and their research lab.

4. Distribution of startup company types

Crunchbase data also provide us insight about the spatial distribution of different types of startup. Within the Boston area, we can see, for instance, that there are many biotechnology companies located in Kendall Square. We may reasonably suspect that big pharma companies like Novartis, and Pfizer provide a hotbed for this type of startup company to grow.

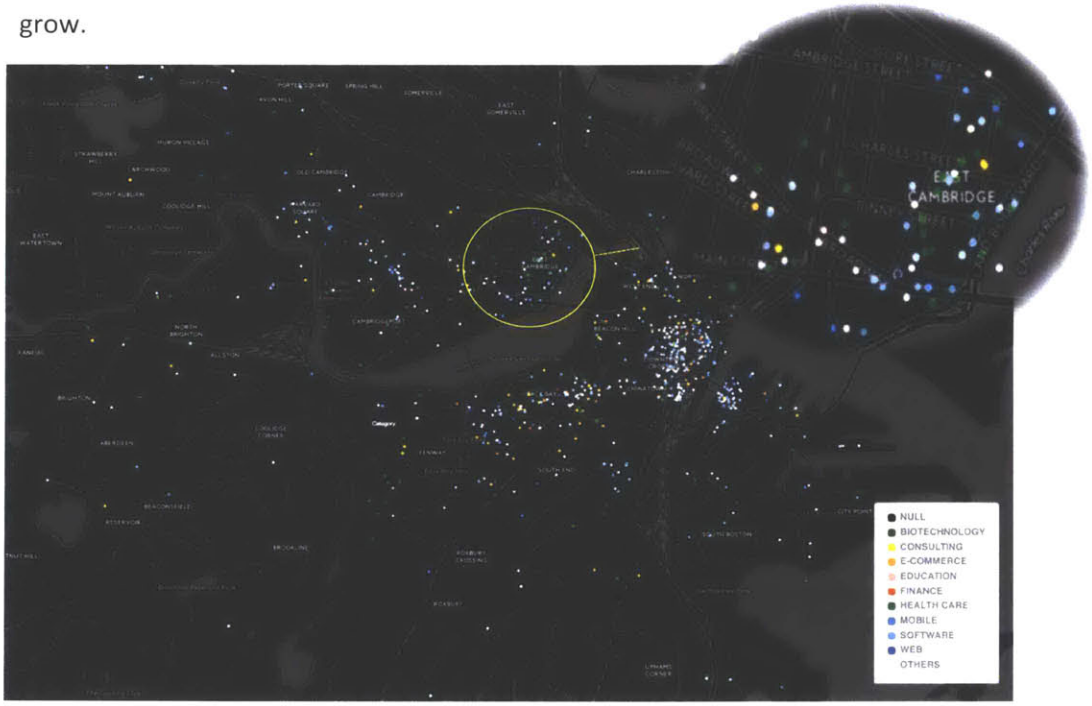


***Figure 4.3.4**: Biotech Company tend to cluster around MIT.*

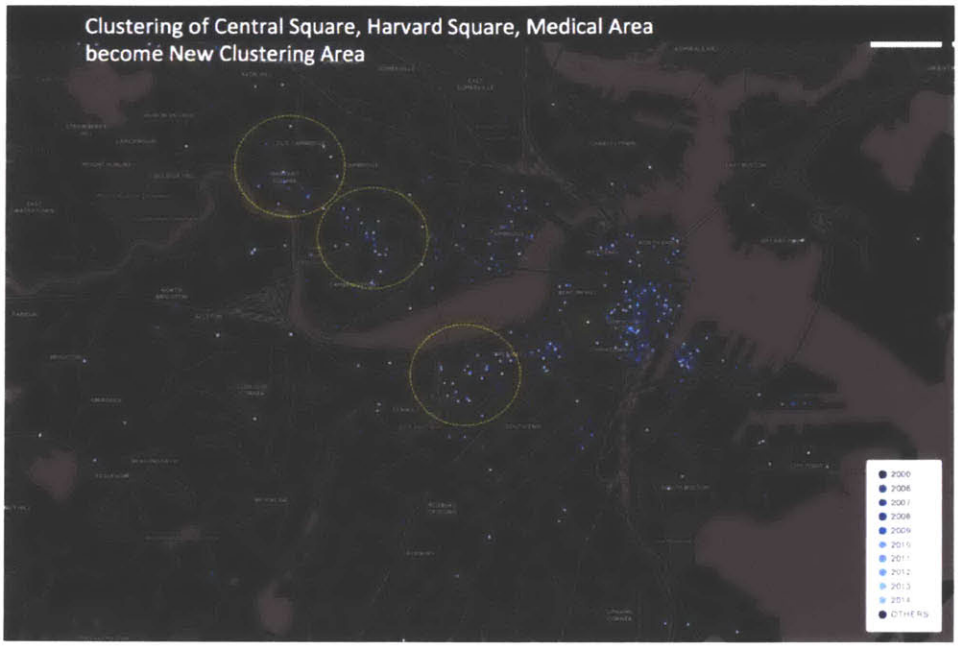


***Figure 4.3.5**: Based on timestamp, new startup company will locate in cheaper rental price with public transportation system area*

Established temporal and spatial distribution

From the timestamp and geo-location data provided by startup companies in Crunchbase, we can also trace the dynamics of startup developments in the Boston and Cambridge area. In this visualization, the brighter the point represents the later of the time a company established. The timeline range is from 2000 to 2014. There are two types of districts popular among startup companies:

First, in the innovation districts established by government, such as Kendall square and the Boston Seaport innovation district area, the earliest tech companies established here soon rapidly attracted new startups into the area. Second, there are "self-emerging innovation districts", i.e. districts that emerged spontaneously without government intervention, such as Harvard Square, Central Square, and the Financial District. These have become popular districts recently for startup companies. Why they spontaneously emerged is an interesting question. The next section of this chapter deals with this question.

We now test whether datasets from other social media platforms may shed insight to how these spontaneously emerging innovation districts occurred.

***Figure 4.3.6**: Spatial distribution of business type from yelp. The map shows some areas are lack of civic facilities, such as Kendell Square and South Boston Innovation District*

Yelp as an amenity indicator

One hypothesis of why these locations (Harvard Square, Central Square, etc.) were preferred startup locations is that they differ in some of the local amenities that are offered, from other area of Boston. We mined Yelp data to examine the diversity and density of amenity in Boston. From the figure, we can examine the local spatial distribution of amenities. We compare the "government established innovation districts" with the "self-emerging innovation districts". The first category of Innovation districts (including Kendall Square, and Boston Seaport) lack, to this day, a diversity of amenities in the local area, including restaurants, bars, and grocery stores. In contrast, "new emerging innovation districts", such as Harvard Square, Central Square, and Financial District, not only have higher density of civic amenities, but also have higher diversity in civic types.

***Figure 4.3.7**: Map shows no grocery store in Kendell Square and South Boston Innovation District*

Focusing further, we can examine the amenity pie charts of four key areas only: Kendall Square, Seaport Innovation District, Harvard Square, and financial district. If we only focus on "grocery stores" in data visualization analysis, it can be seen that the "self-emerging innovation districts" also tend to have more groceries stores, which plays critical role in providing daily supplies of life. This information suggests to urban planners that mixed used in land use is very important in the future for injecting humanity to a new innovation district, and is something deficient within the "government established innovation districts"; a consideration for further improvement.

***Figure 4.3.8**: Map shows civic facility and ratio in different districts.*

***Figure 4.3.9**: Kendall Square and South Innovation District have high ratio of high tech company but have low ratio of civic facilities*

4.4 Twitting-scape: Twitter as Activity and Land Use Sensor

Our hypothesis regarding data mining geo-located data from Twitter was that we could answer questions pertaining to when and where activities are occurring. In contrast to the civic data that can be provided from Yelp, geo-location data from Twitter can provide information about whether a building or open space is being used or not. One particularly interesting investigation we did was to look at the difference between daytime and night time use of a piece of land or building; more uniform use of the buildings during the day and night mean that the urban design goal for more efficient building use was achieved.

We can see, for example, that geo-located Twitter data in Harvard square is active both during the daytime as well as at night. In contract, Kendall square has a lot of data activity during the daytime, which attenuates at night. This implies that human activities in Kendall square are not as popular as Harvard square during the night time. This information informs the urban designer that the mono-functional land use in Kendall square does not provide as many activities as the mix-used land use in Harvard Square, a consideration for further improvement.

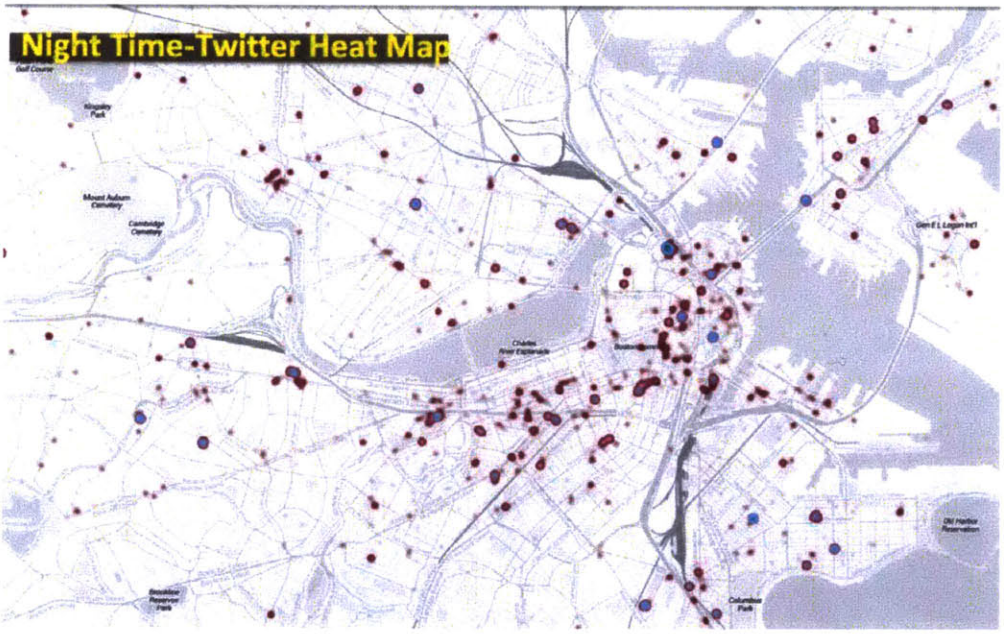


***Figure 4.4.1**: Twitter Geolocation data identifies the night activities in the city. Cambridge and South Boston Innovation District do not have night events.*

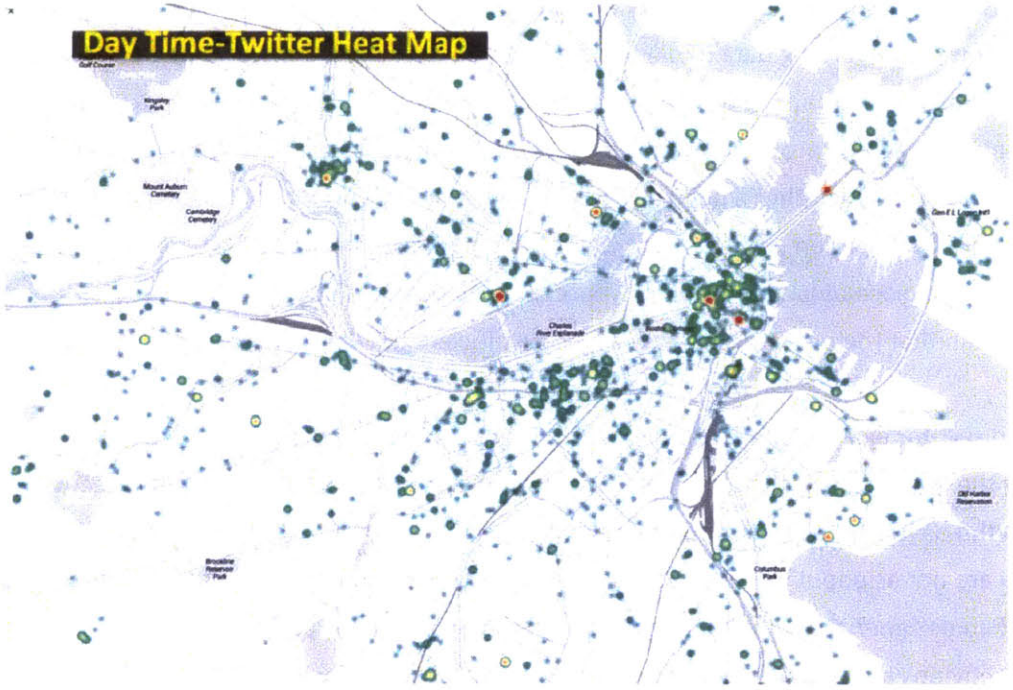


***Figure 4.4.2**: Twitter Geolocation data identifies the daytime activities in the city*

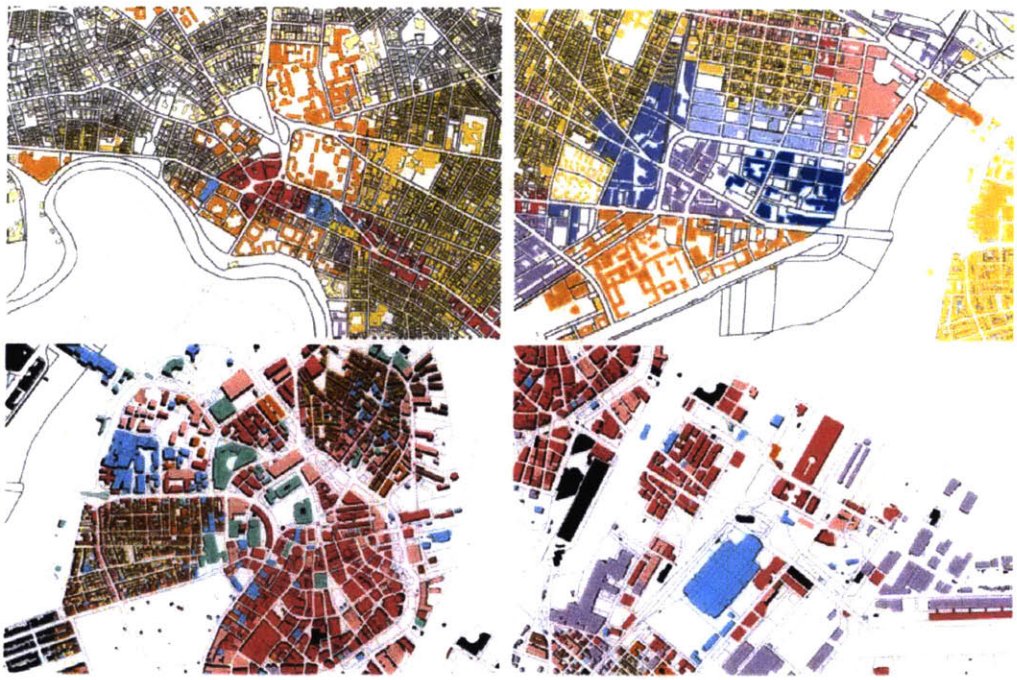


***Figure 4.4.3**: Overlay with land use map, the mono function land use in Cambridge and South Boston Innovation District may cause lack of civic facilities and night life in the area.*

A final example of this principle can be seen, using twitter data, to test whether the public

spaces constructed in Boston are as popular as desired. If the twitter geo-location data never shows up in the open space, it would imply the open space is not being used.

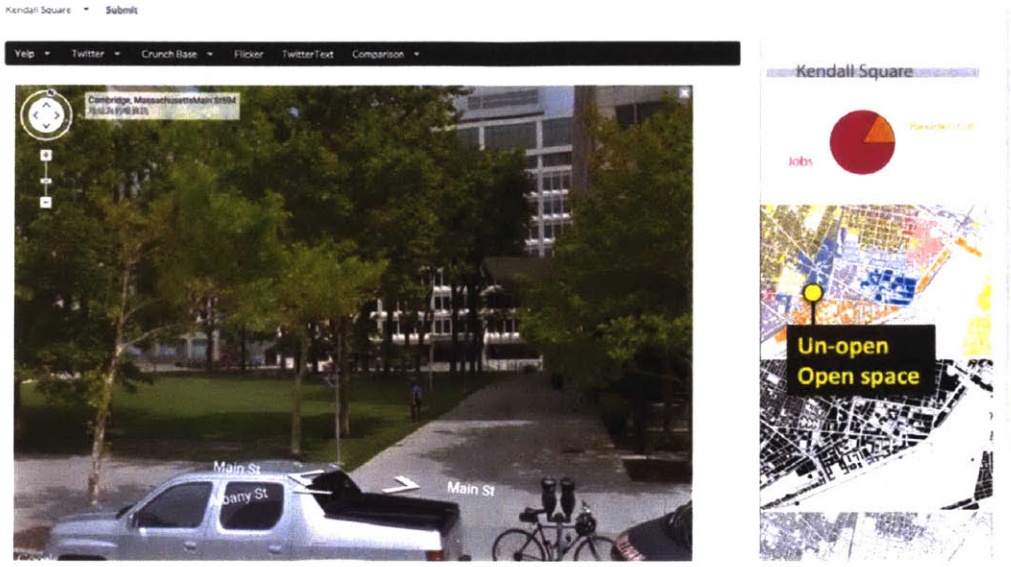


***Figure 4.4.4**: Twitter data could help urban designers identify some un-use open space*

For example, the open space "Tech Square" in Kendall Square innovation district, is seldom used on weekdays at night and weekend's all day. Thus, we may advise the Urban designer to find the opportunity to "re-innovate" the open space which people are not frequently using.

Tweets as a Local Trend Speaker

We hypothesize that using Nature Language Processing to mine tweets (from Twitter) could help the urban designer understand important social trends and priorities of local people. Comparing to traditional methods such as interviews, surveys, and fieldwork, utilizing social media text data is a novel perspective to understand the city. **Sentiment analysis** (also known as **opinion mining**) refers to the use of ***natural language processing, text analysis*** and ***computational linguistics*** to identify and extract subjective information from source materials.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. In other words, it aims to extract data about social opinions in a very high throughput way.

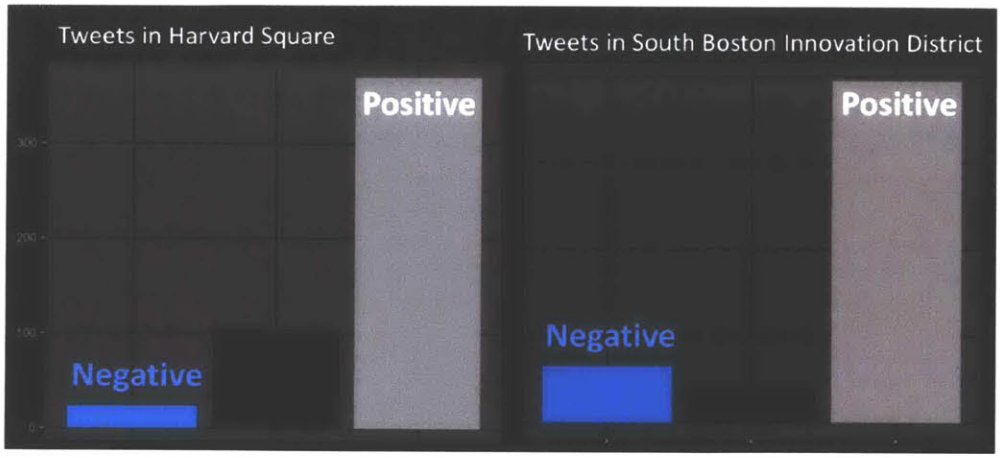


***Figure 4.4.5**: Tweets with Natural Language Processing could help urban designers understand the people describe the district is positive or negative.*

From sentiment analysis, we can understand if people's emotions are positive or negative. One observation we made was that tweets related to the South Boston innovation district contains more negative words than tweets which is related to Harvard square, implying that the perceived quality of daily life in South Boston does not in many ways meet peoples' expectations. Urban designers have a role to play in this, and should take this information into consideration. Urban designers can investigate what kinds of urban spaces make people feel more positive or negative, and go back to analyze the city form in a data driven way.

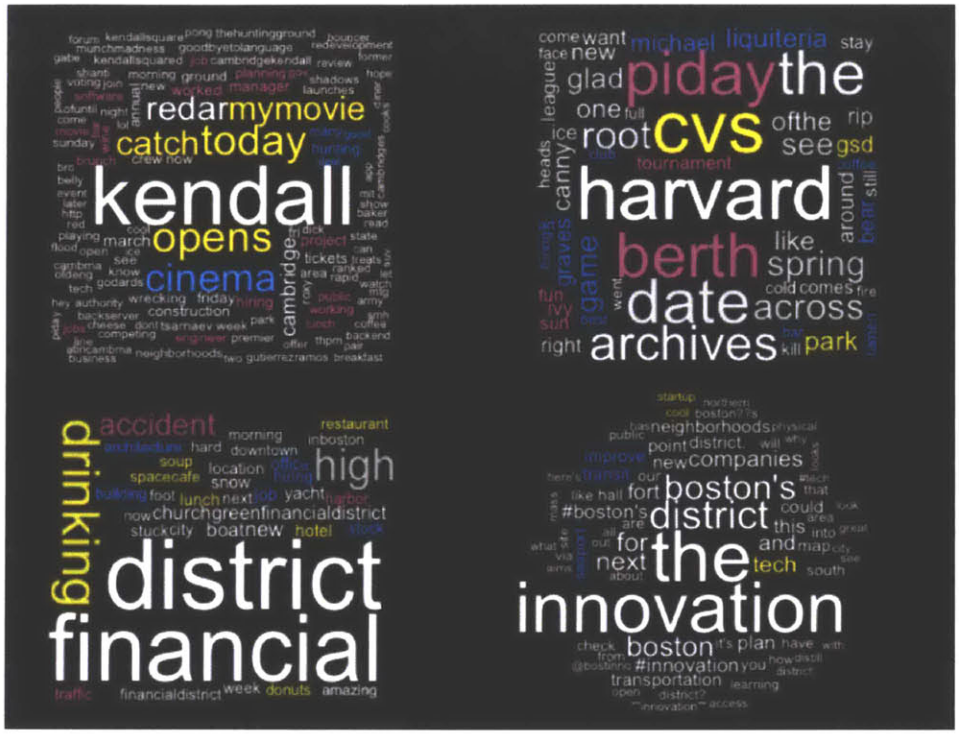


***Figure 4.4.6**: Tweets with Natural Language Processing could help urban designers understand the popular topics in the districts.*

Analyzing the frequencies of words and topics (not just sentiments now), from tweets, provide us with information about what people care about and the image of the local district within the citizen's daily life. For example, some local amenities, such as CVS, Park, and Drinking, frequently appear as the topic of Tweets in Harvard Square and Financial District. This implies that these places seem to play important roles in local life. Within Kendall Square, the cinema is the only amenity topic that is significantly discussed; the rest of the topics that show up in Tweets in that area regard "jobs" and "high tech". This problem is even worse in the South Boston innovation district: no civic topics show up at all. Only word related to industrial, such as "tech" and "startup" show up, suggesting an almost monolithic focus within this area; not a sign of a vibrant district.

Finally, Tweets, in addition to analyzing spaces, can also be used to analyze events. For example, "pi day" in Harvard square could be noticed, from the Twitter data of March 14, 2015. Thus, when and where major activities happen can be mined from Twitter data.

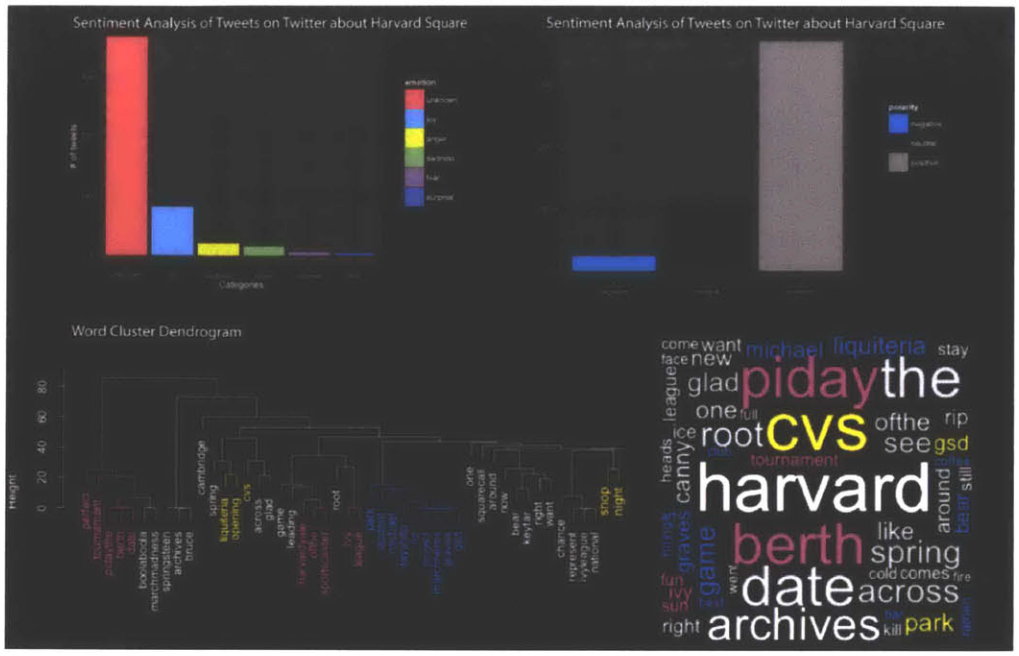


***Figure 4.4.7**: Via Word Cluster, Sentiment Analysis, and Topic Word Cloud, urban designers could better understand trend and how people feel in urban districts.*

4.5 Flicker-scape: Flickr as Urban perceptron

We hypothesized that mining image data with Flickr can help understand how people perceive the urban environment. Images collected from social media not only provide image data, but also simultaneously provide related information such as tag-texts, geo-locations, and timestamps. In this way, we possess not only the image content, but also data to understand how people utilize urban space and how people feel, within the space.

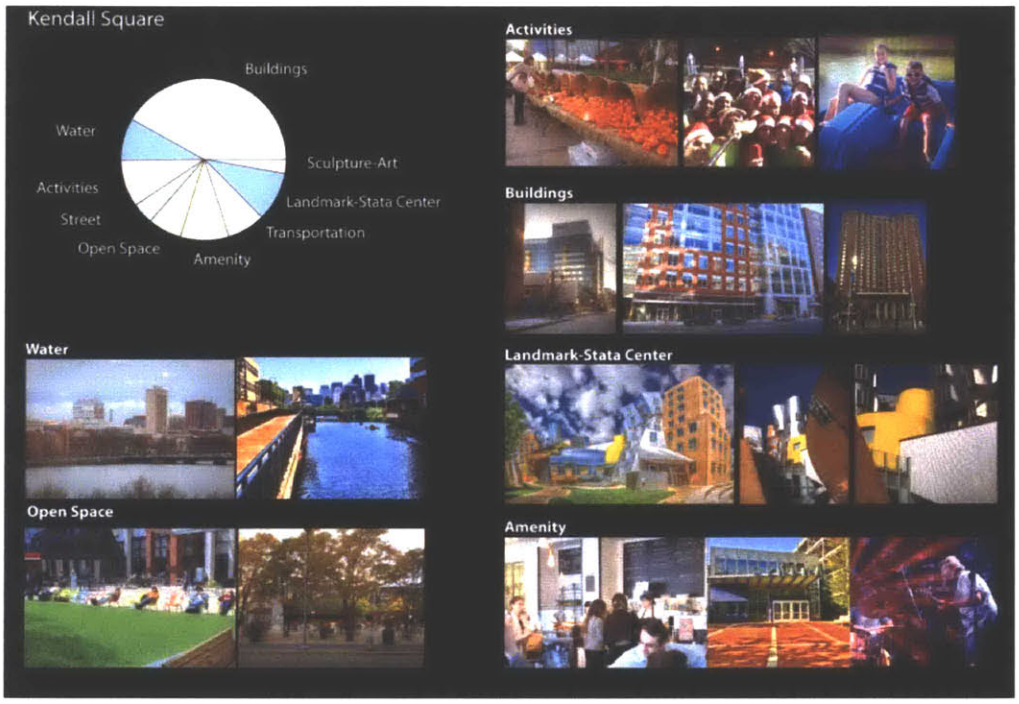


***Figure 4.5.1**: Identifying images from Flickr could help urban designers understand the "image" and the activities in the districts.*

For example, image data from Kendall square innovation district most often pertained to the urban space. Popular topics included the local buildings, sculptures art, landmarks, transportation, amenities, open space, street, activities, and the Charles river. Thus, the most popular and important "image of the city" within the Kendall Square region is the urban space itself. Of particular note is that images of the Stata Building in MIT account for about one fifth of all images taken. Thus, in fact, Stata Building could be considered a landmark within the Kendall square area!

***Figure 4.5.2.1**: Comparing the number of images related to activities implies Kendall Square area is not as popular as Harvard square for activities*

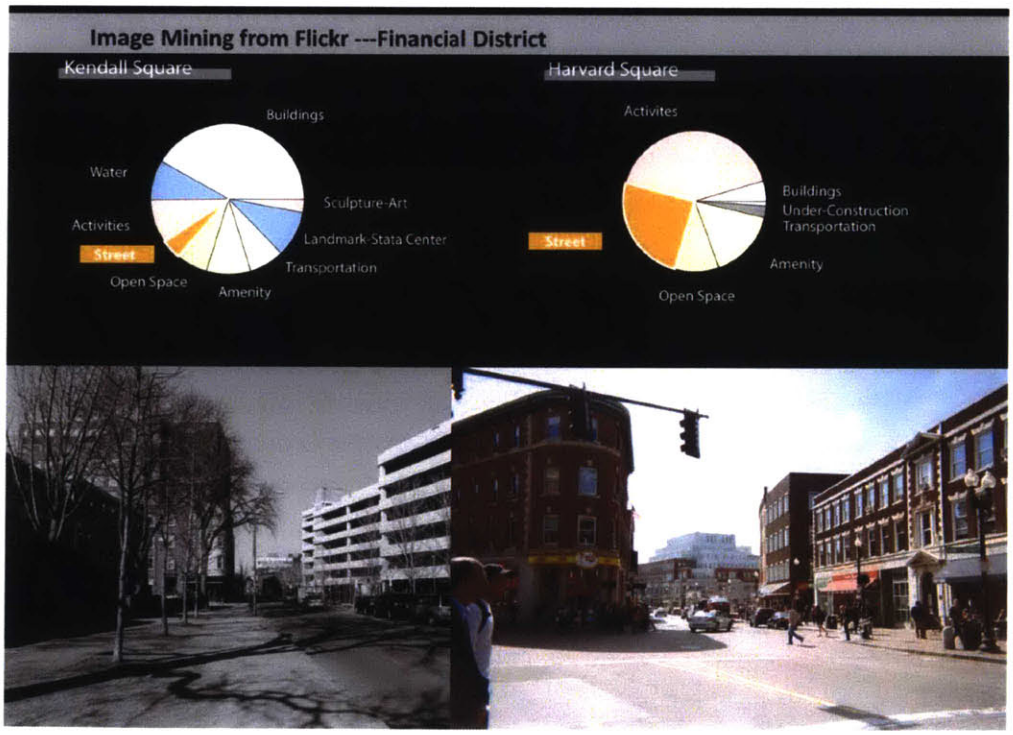


***Figure 4.5.2.2**: Images related to Street life in Kendall Square is less than in Harvard square.*

***Figure 4.5.3**: Images related to Open Space in Kendall Square is less than in Harvard square.*

In contrast, Harvard square Flickr data contain more images related to activities, street life, and open space. This observation, combined with some of the other observations above about Harvard square, suggest a plausible reason why Harvard square spontaneously developed into an Innovation district, despite it not being a government designated one. More and more startup companies tend to come this "mixed used" area, where there are so many convenient civic amenities, as well as a world class institution, Harvard University.

This focus about Harvard Square and the "self-emerging" innovation district brings up an interesting and important question: if there was ever a particular piece of land that did not meet the original purpose (and of course, there are plenty of examples of these!) how could we improve the space? In this chapter, we drew lessons from the successful districts and the less successful ones that we data mined. We drew lessons from the peculiar districts that spontaneously emerged as Innovation districts, and asked why they did so. In this chapter, we compared the urban design in both areas, and made numerous such suggestions for improvements.

Urban Implementation:

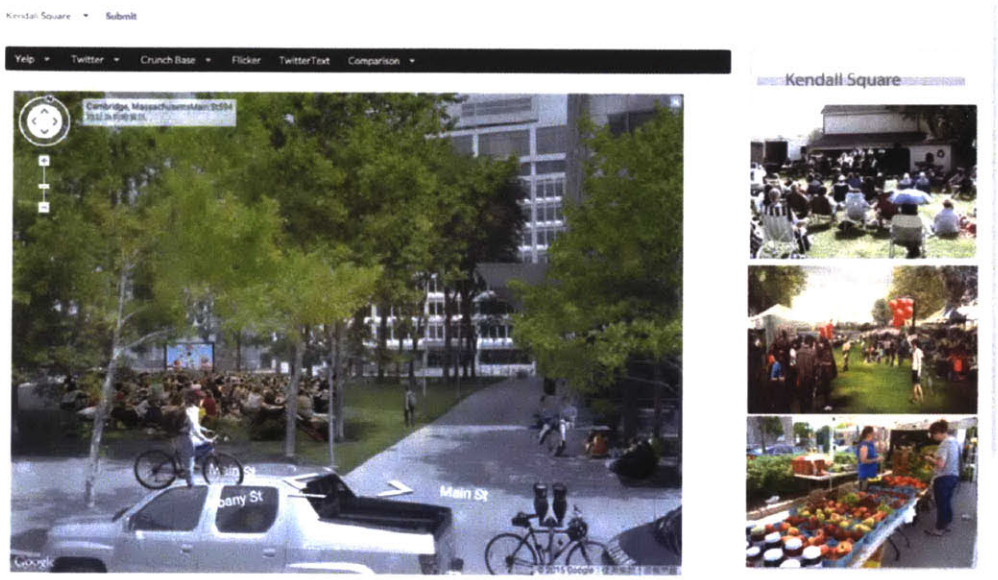


Figure 4.5.4: Urban Planner could implement some festival in the un-use open space, based on the activity sensor, twitter.

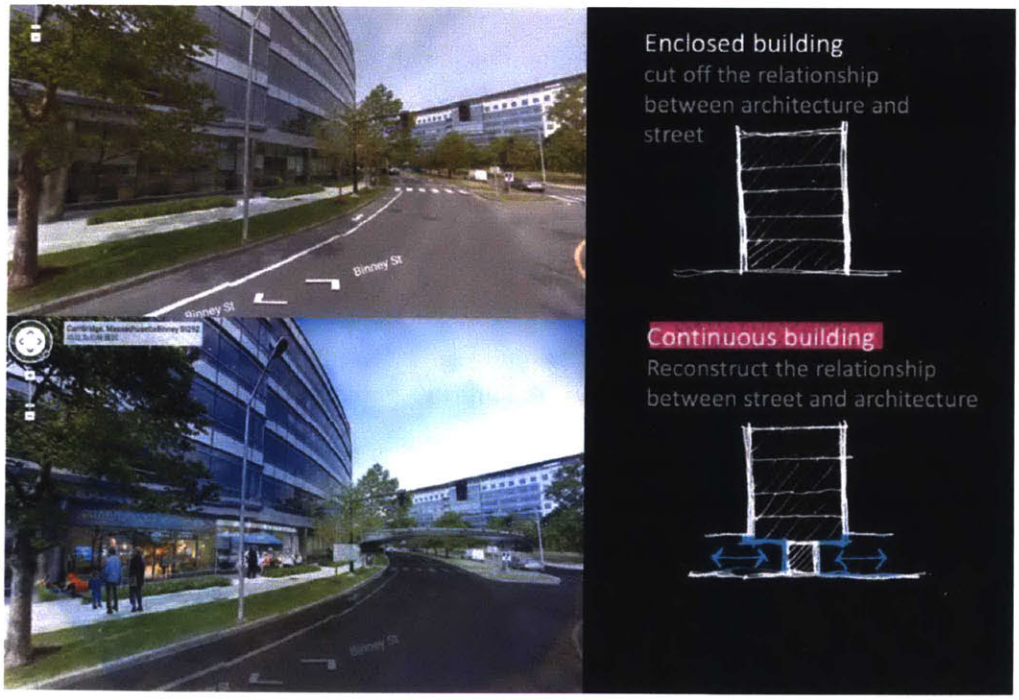


***Figure 4.5.5**: Urban designer could re-plan the usage of building, open retail store in first floor.*

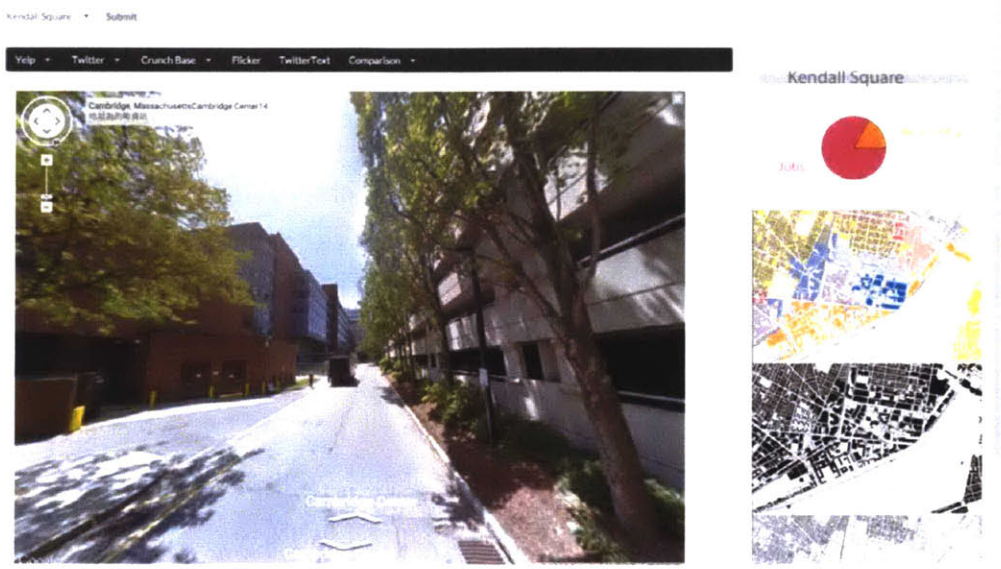


***Figure 4.5.6**: From mono function zoning to mix use zoning could increase the activities in the area without nightlife.*

***Figure 4.5.7**: Simulation of re-design the area based on the issues discovered by social media.*

Chapter 5: Swarm-scape 2.0: Data Mining Andorran Tourism Patterns

5.1 Andorra as a Tourism Country and Trip-Advisor

The presence of web2.0 and traceable mobile devices create new opportunities for urban designers to understand cities by these user-generated data, including social media data and call detail record (CDR) data. The emerging of this "big data" has resulted in a burgeoning amount of information documenting daily events, perceptions, thoughts, and emotions of citizens, all annotated with the location and time they were recorded. This data presents an unprecedented opportunity for the purposes of gauging public opinion on the topic of interest.

In the past, urban designers seldom made use of (or had access to) such a large source of public data. For those who did make use of data to drive their designs, most made use only of government-collected data which are, in general, limited in scope. But nowadays, such bottom-up techniques can be drastically accelerated and supplemented by using geo-located data mining techniques of social data. These non-conventional data sources derived from activity from digital network offers extensive amount of information about human interactions. In this project, we extract and collect data both from call detail records (CDR) as well as from TripAdvisor social media, for the purpose of analyzing tourist patterns. We do so, with data from the European country of Andorra, whose tourism industry occupies 80% of its economy, but which has, recently, experienced a decline in its tourism numbers. Our goal is to mine our collected urban and social data for patterns that may lead to useful recommendations to Andorran tourism authorities. This sheer amount of available data, however, brings its own, difficult problem: how to extract useful patterns and information from this enormously "big data".

Machine learning has become a powerful tool in other data driven disciplines but has rarely been applied in urban design. We thus ask whether machine learning is capable of extracting urban patterns from the tourist data in Andorra, and deliver relevant conclusions. We applied, (to the first of our knowledge to urban data) the k-means clustering technique, and were able to separate Andorra into different regions based on distinct land use types. We next applied, (to the first of our knowledge to urban data), natural language processing (NLP) and sentiment analysis to tourist attraction and restaurant reviews on TripAdvisor, and found distinct clusters of local Andorran amenities that either had uniformly positive and uniformly negative reviews. We applied, again to the first of our knowledge to urban data, association rule algorithms and discovered several distinct "types" of tourist travel patterns,

and with this knowledge, showed, using the random forest algorithm, that we could predict, to a very high degree, future travel plans of individual tourists from their past travel patterns. From this, we made recommendations to Andorran authorities about which regions in the country had the best and worst tourist attractions, and cater their tourism packages to the particular "types" of tourists that we discovered. In an illustrative way, we propose that this powerful approach of combining user generated "big data" with state of the art machine learning techniques can give unique and commanding data-driven insights to the urban designer and planner that would be hard to discover, by any other method.

The project, titled "**Swarm-scape 2.0: Data Mining Andorran Tourism Patterns**", focuses primarily on the social media data and CDR (Call Detail Record), which is telecommunications data, consisting of every call or SMS message that routes through an Andorran cell tower. From this data, the location of the user can be approximated to a city level, and thus their position can be tracked throughout the country of Andorra.

From social media data and CDR data, we can observe how the tourists move across the country, and predict where they will go next for different interest. The first portion of this project involved determining the local attractions of Andorra via a web scraper that extracted ratings, locations, and other information from Tripadvisor. Next, the movement of tourists throughout Andorra for a single day was obtained via an OD matrix and displayed. After getting a sense of how tourists moved through the country, the Association Algorithm was used over a year's worth of data to determine how connected cities were to one another for various months and holidays. Lastly, various machine learning algorithms were used to predict where an individual would visit next based on their previous locations and the nationality of the user.

Overview and Motivation

Understanding customer mentality is essential for a successful business. In fact, lots of IT companies have launched Location recommendation services based on a person's past behavioral patterns to predict future location. For example, Google Now and Yelp Nearby can both provide personalized recommendations. Similarly, understanding travel patterns is key to design personalized travel recommendation and to improve local business to attract and serve tourists. In this way, I am interested in learning the travel patterns of tourists to predict their next locations, and provide recommendations based on their past behavior.

Andorra has becoming a popular tourist destination in Europe. It is a small country bordered

by Spain and France. At present, there are 54,619 reviews in Tripadvisor relevant to Andorra. I would like to explore our ideas with such manageable dataset. Related Work Previous work has been done on tracking and predicting location based on information collected through smartphone. For example, Nokia's Mobile Data Challenge, a paper by Do & Gatica-Perez used the smartphone data (where, when and what apps people used in their smartphone) behavioral patterns to study how generic behavior models can improve the predictive performance of personalized models. However, the behavioral pattern during vacation might differ a lot from daily life. For example, instead of going to the same places over and over again, tourists might be an "exploration" mode to visit different places.

Trip-advisor as POI and Speaker

This part of project focuses primarily on how Andorra is perceived on social media. Specifically, this project works with Trip-Advisor data, but the process may be used to analyze other social media sites. Trip-Advisor is an online review platform that focuses specifically on travel. It has over 290 million user reviews and opinions and 350 million monthly unique visitors, making it the world's largest travel site . More than ever, people are using online sites such as Trip-Advisor to help them plan their next travel destination. A 2013 study showed that 75% of travelers use social networking sites to look for shopping related deals and 30% specifically look out for travel deals.

More importantly, looking at reviews has become a prerequisite for booking travel 3 for many people. 77 percent of Trip-Advisor users usually or always reference Trip-Advisor reviews before selecting a hotel, 50 percent usually or always reference Trip-Advisor reviews before selecting a restaurant and 44 percent usually or always reference Trip-Advisor reviews before selecting an attraction. Furthermore, 80 percent of respondents read at least 612 reviews before making their decision. It is clear that reviews are a powerful and significant part of the 4 travel planning experience.

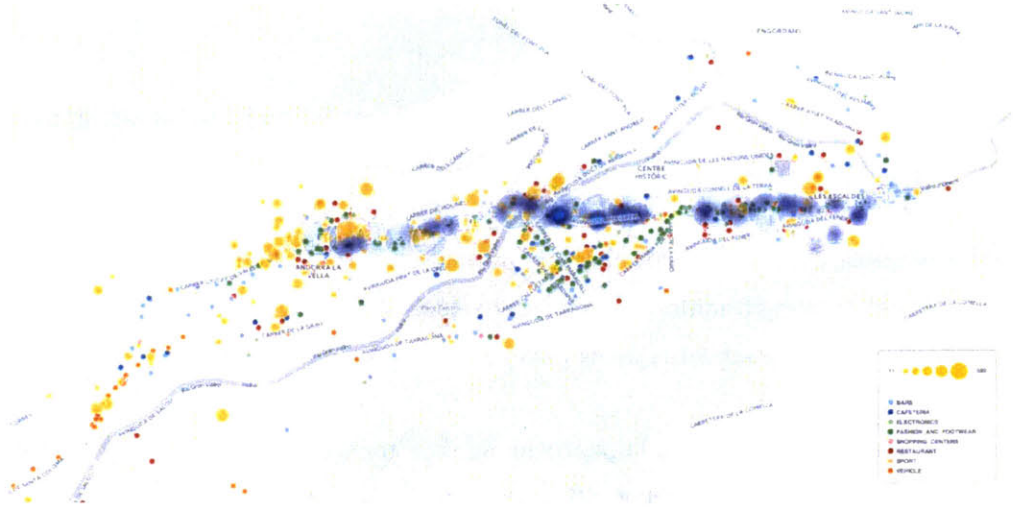


***Figure 5.1.1**: Overlying Trip-advisor POI data and Cell Phone wifi data could help urban designers discover the relationship between activities and civic facilities*

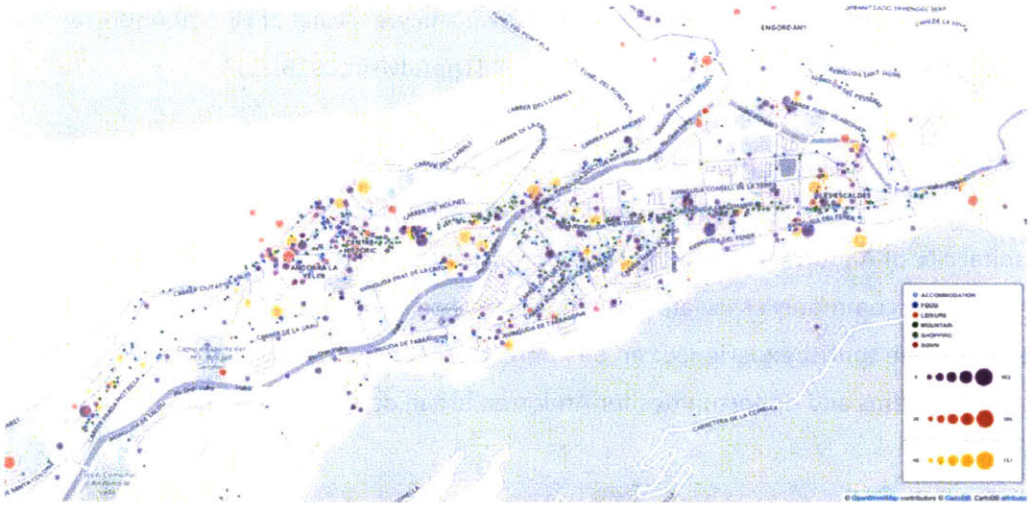


***Figure 5.1.2**: Different social media provide different POI information, such as purple dots show popular area from Instagram; red dots show popular area from AirB&B, and yellow dots show the popular attraction from Trip Advisor*

Therefore, the focus of "NLP with Trip-advisor" is on the sentiment of the users towards the hotels, restaurants, and attractions in Andorra, specifically focusing on the city of Andorra La Vella, the capital city of Andorra. By analyzing the words used in the review, this project aims to extract the topics most commonly associated with good and bad reviews, thus giving a greater sense of how each location can be improved.

5.2 Trip-scape: NLP as a Urban Design Tool

Introduction

Compelling arguments for the use of bottom-up social opinions to inspire urban designs can be found in influential books such as "The Image of the City" (Lynch, 1960), "Death and Life of Great American Cities" (Jacobs, 1964), and "City is not a Tree" (Alexander, 1966). A large scale survey of public opinions for this purpose, however, was difficult in the 1960s because it relied on time-consuming traditional ethnographic tools such as surveys and interviews. Presently, modern geo-located data mining techniques can be deployed.

This project aims to complement traditional ethnographic tools by mining social media data for the purpose of better understanding cities. These techniques are applied to analyze tourism data from the country of Andorra, a small country between Spain and France. Our goal is to show how mining social media data for urban patterns may lead to useful recommendations to Andorran tourism authorities. We investigate how Andorra is perceived by tourists by analyzing social media. Specifically, we analyze a total of 68,500 Andorra-related tourist reviews obtained from Trip-Advisor. (TripAdvisor,2016)

We used "Natural language processing" (NLP) with Trip-Advisor to determine the sentiment of the users towards the hotels, restaurants, and attractions in the city of Andorra La Vella, the capital city of Andorra. By analyzing the words used in the review, this project extracts the topics most commonly associated with good and bad reviews, thus giving a greater sense of how the tourist experience can be improved for each Andorran attraction. This pinpoints problems and opportunities for Andorran urban designers and planners to focus on.

Previous work

The origins of natural language processing sentiment analysis has come from previous work such as "Sentiment Analysis and Opinion Mining" (Liu et al., 2012) that have used sentiment analysis to examine the text of online movie reviews to automatically detect opinions about the various movies. Following this paper, others have begun to use sentiment analysis to examine other product reviews. Our approach is to apply it to urban design decision making.

Data and Methods

Our data analysis workflow involves 6 steps, summarized in Figure 1. The steps are described below.

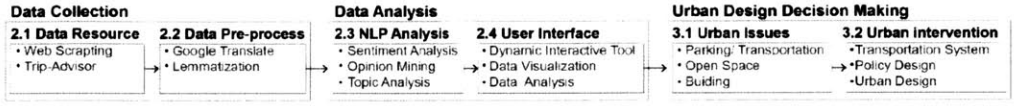


Figure 5.2.1. *Data analysis workflow*

Data Resource: Web-Scripting Trip-advisor

The first step of our analysis pipeline was to extract the relevant review information from each review for the Andorran destinations of interest. On Trip-Advisor we focused on three primary categories of tourist destinations: restaurants, hotels, and attractions. A scraper was employed to extract the name, address, rating, and review attached to each of these destinations and record them in table format. From these reviews, we would identify sentences describe urban issues, such as parking. We used a script to parse the data from website based on the structured web format.

Reviews were in multiple languages. Each review was translated into English, using Google translate API, in order to enable subsequent language analysis.

Data Pre-process: Translation and Lemmatization

Having standardized each TripAdvisor review page into English, the next step in our analysis pipeline was to simplify the page enough to enable NLP analysis by 1) extracting root words, and 2) removing superfluous words. This was done in several steps.

First, reviews had to be "tokenized", that is, broken into single words using the Natural Language Processing Toolkit (NLTK) tokenizer (http://www.nltk.org). Sentences such as "the parking is awful" were broken into its individual words ("the", "parking", "is", and "awful").

Second, each word detected was "lemmatized", that is, their root word was extracted. The purpose of this step was to decrease the number of words that have to be analyzed by equating words like "run", "runs", "ran", and "running" into their respective single root word "run".

Next, the 200 most common words in the English language such as "the", "is", "are", "a", etc. were detected and deleted from the review pages, thus preventing these simple and frequent words from obscuring the more important and relevant words in the text that are actually important for understanding the topic matter. Noun phrases were next detected in

the text using NLTK libraries.

NLP Analysis: Sentiment and Topic Analysis

We classified the reviews based on "sentiment". Based on NLTK libraries, we detected words that have a "positive" sentiment, and words that have a "negative" sentiment, within the reviews of each tourist destination. Each sentimental word was attached to its neighboring noun (which serves as the particular subject/topic), and the collection of sentiment-noun pairs were collected. For example, the sentiment-noun pair "beautiful lake" was collected.

Based on the sentiment analysis done at each Andorran destination, an interactive visualization summary of all Trip-Advisor destinations for La Vella was created. It allows for a heat map break down of reviews, sentiment, and relevance for all specific topics examined, as well as a breakdown by language of reviews to estimate the demographics of visitors to each destination.

User Interface: Dynamic Interactive Data Visualization and Analysis

The key technological development in this project was the production of a searchable visualization summary of all the Trip-Advisor locations to be used in the Andorra City Scape model. It includes a heat map of reviews, sentiment, and relevance to any searchable topic definable by a key word (e.g. "street", "parking", "shopping", etc.). It can also provide a breakdown by review language (Spanish, French, Russian, etc.), to further analyze the demographics of visitors to each specific attraction in Andorra. In addition to being searchable, each location has a popup card that displays information such as rating, summary of review, sentiment, and most popular keywords.

Figure 2 shows an example of a search conducted in the user interface, searching for "street", in Spanish reviews. The degree of popularity of locations is indicated by regions colored from green (less popular) to red (most popular) with restaurants (red dot) and hotels (blue dot).

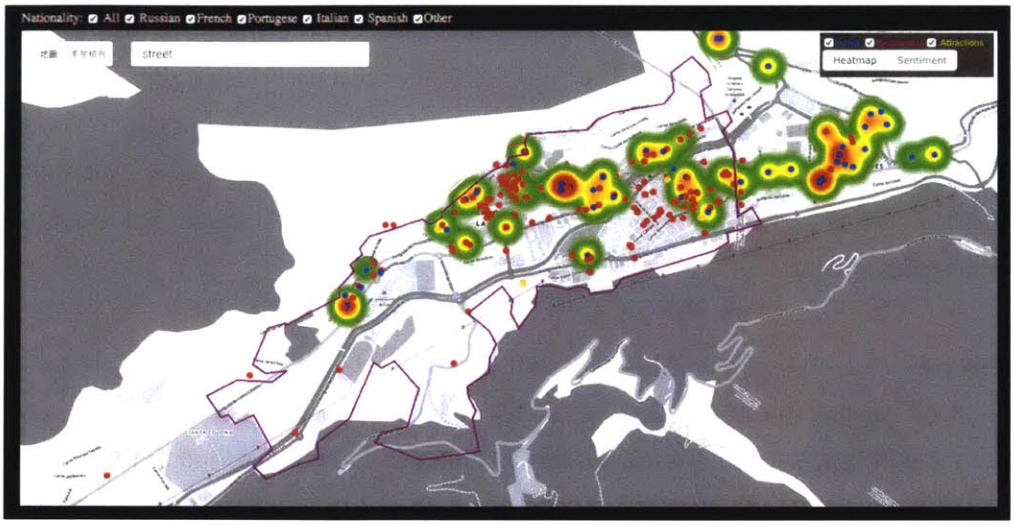


Figure 5.2.2. *HEAT map for search of "street" in the Spanish reviews*

Figure 3 shows the same search, but with the added condition to display the *sentiment* of the reviews, to show which locations were positively (green) and negatively(red) experienced by Spanish visitors.

Figure 5.2.3. *Map for Sentiment analysis search of "street" in the Spanish reviews*

Result: Data driven Urban Design based on Sentiment Analysis

Urban Issues

We now demonstrate how our User Interface is a valuable tool for urban design decision making.

We compare the Trip-advisor sentiment map to a land use map. There are two districts, the old city center area and the new pedestrian district, that have a particularly high concentration of negative reviews (Figure 4 below, red circled areas). We immediately note a strong correlation between the areas of the city that are most negatively reviewed, and areas that have a lack of parking (Figure 5 below). This is a novel observation that is a direct result of our data mining.

It is in this way that sentiment analysis can point out issues of importance to the urban designer. Most tourists in Andorra visit by car, and if the city supply of parking facilities does not meet the demand, it invariably has a negative effect on city tourism, and reduces the likelihood that people will want to visit or stay in Andorra.

Figure 5.3.1. *Sentiment map for positively (green) and negatively (red) reviewed streets, with areas that are concentrated with negative reviews circled in red*

Figure 5.3.2. *Availability of parking (blue), plotted with negative review-concentrated areas of the city circled in red*

Urban intervention

Our social media data mining provides bottom up information from users for urban designers to make decisions about where their future designs should focus on improvement. In this case, urban designers may need to alter the parking facilities around the old shopping centers or renovate the area.

In parallel with the parking problem, there is another more general problem regarding transportation systems: with increasing car usage, the traditional road system in Andorra Le Vella does not meet road usage demand. One can address both of these problems simultaneously. One possible solution is to increase the public transportation system for intra-city network connections, which would help alleviate both traffic congestion as well as parking issues. The other possible solution is to improve accessibility in the city by implementing urban systems amenable to shuttles, bicycles, and taxis.

Conclusion

Our analysis of social media data revealed the most positively and negatively reviewed tourist locations. By comparing the regions to the land use map, we identified a prevalent issue of parking in the city. We suggest that our approach of combining social media "big data" with natural language processing to detect patterns of sentiment is a useful new methodology for the urban designer and planner, and can give data-driven insights that would have been hard to collect otherwise.

5.3 CDR-scape: Dynamic Urban Computing and Responsive Urban system

In this chapter, urban data analysis is conducted using call detail records (CDR) data, in order to find mobility pattern within cities and between cities. Based on different aggregation methods, including by time, area, and tourist nationality, and visualizing the results, we aim to discover relationships between preferred events and tourist types. We designed three types of visualization maps in web applications, in order to understand different spatial-temporal tourist patterns. The three dynamic data visualization maps were constructed with the help of three different UROPs. The Spatial Aggregation map was done with Laura Pang, the Association Rule Map was done with Marissa Stephen, and the OD map was done with Margaret Yu. Ultimately, we use the knowledge gained to suggest different urban design strategies. The possible urban design implications from each type of CDR data analysis are as follows:

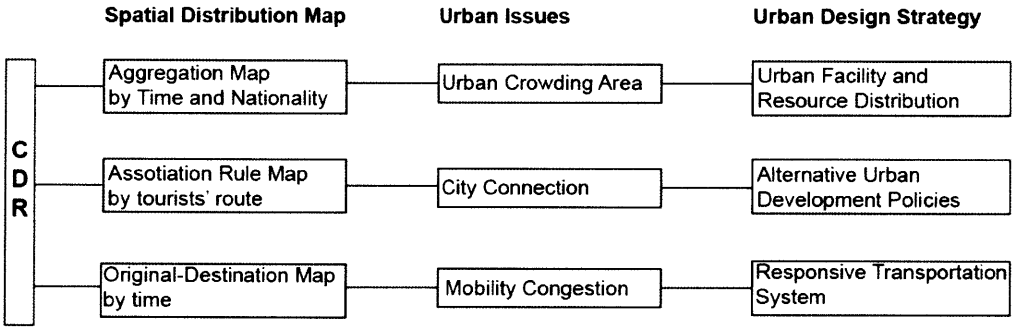


Figure5.3.1. *Framework of Spatial Data analysis and possible urban implement.*

Call Detain Record (CDR) data is originally of the following form:

- DS_CDNUMORIGEN: ciphered phone number
- DT_CDDATAINICI and DT_CDDATAFI: start time and end time
- NUM_DURADA: duration of phone call in second
- ID_CDOPERADORORIGEN: code of cell phone carrier
- ID_CELLA_INI and ID_CELLA_FI: start tower and end tower

	DS_CDNUMORIGEN	DT_CDDATAINICI	DT_CDDATAFI	NUM_DURADA	ID_CELLA_INI	ID_CELLA_FI	ID_CDOPERADORORIGEN
0	9aff9f9d53ebd77d2cc0edc6eddb761b0c1d5ae7166ea8...	2015.01.02 00:03:50	2015.01.02 00:03:50	0	2021	2021	20801
1	9aff9f9d53ebd77d2cc0edc6eddb761b0c1d5ae7166ea8...	2015.01.02 00:54:39	2015.01.02 00:54:39	0	2021	2021	20801
4	161e38db04f53740a00aaf1734d5b07a19381567261ebf...	2015.01.02 00:01:58	2015.01.02 00:01:58	0	19091	19091	20801
5	161e38db04f53740a00aaf1734d5b07a19381567261ebf...	2015.01.02 00:00:38	2015.01.02 00:00:38	0	19091	NaN	20801
11	a1497171f3fa9f67427840a9a2793445478f5dce51a717...	2015.01.02 01:24:00	2015.01.02 01:24:00	0	9162	9162	20801

Figure5.3.2. *Original Data Frame*

Based on this original data set, we aggregate the data with different features.
As an example, we aggregate the data based on timestamp, and apply time series analysis.

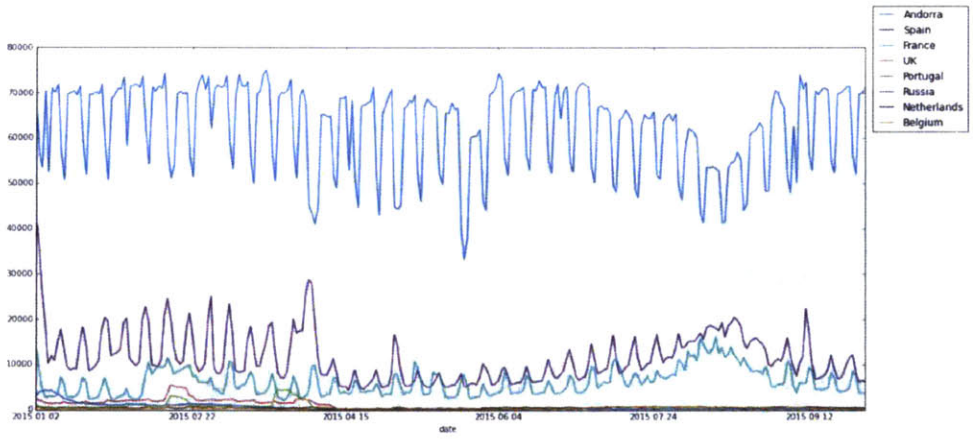


Figure5.3.3. *Time Series Data Analysis from CDR data*

From the Time Series Analysis, we have two small conclusions: i) Spanish tourists and French tourists make up the majority of tourists in Andorra, and ii) there are several peak times for tourism in the country, namely in January (new year), April (Easter), August(Summer peak), and September(Sport Event). Knowledge about these two primary tourist groups, and their relevant peak times, will focus the rest of the analysis conducted.

CDR Spatial Distribution Interactive Map

The first question we ask is how populations change in different cities during a particular tourist event. To answer this question, we aggregate the data not only with timestamp, but also with tower geolocation to provide spatial data specification. We use the K-Means algorithm to cluster the towers with similar location, in order to aggregate the data based on their geolocation.

After the data was aggregated by the k-means algorithm, we create a dynamic interactive map to visualize the population distributions during specific tourist events. For example, on October 26, a famous musical festival occurred in Andorra. We found city population increases not only in the capital city of Andorra La Vella where the event, was being held but also other cities. Particularly, the border cities PasDeLaCasa, and StJulia also increased in population. This suggests that tourists not only attend events in the main cities, but also do other complimentary activities before returning home. The economic benefit from a successful event thus has a ripple effect across the country.

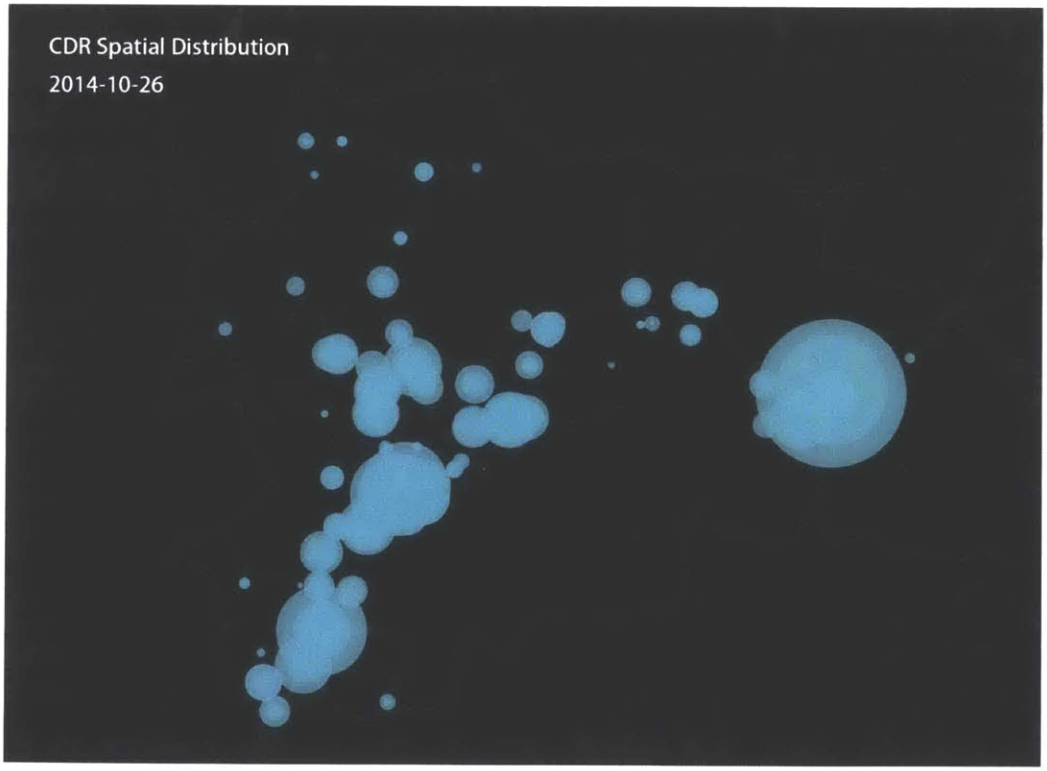


***Figure5.3.4**. Spatial Distribution in 2014 October 26, Musical Event*

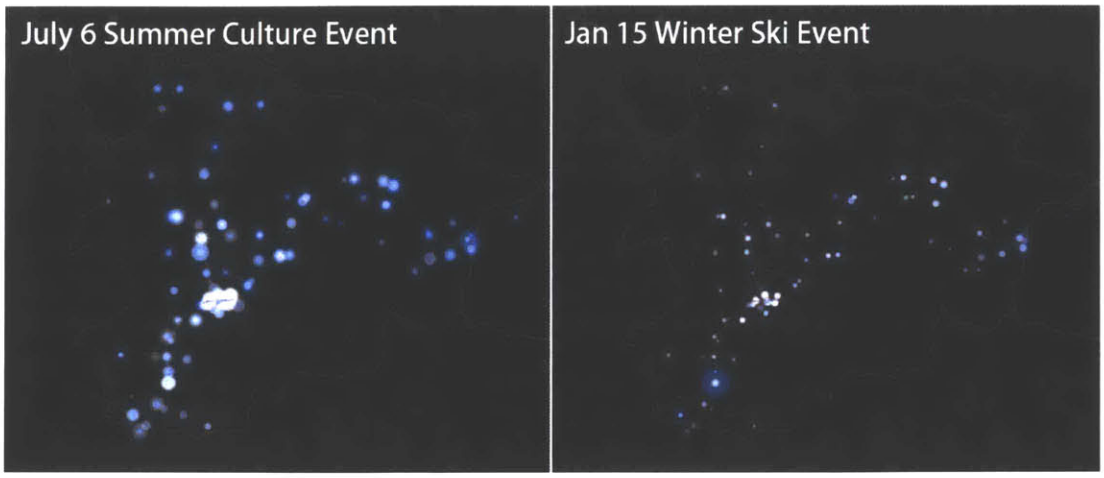


***Figure5.3.5**. Comparing different population spatial distribution in different events.*

We can conduct temporal analysis in addition to the spatial analysis already described. For instance, we may ask if different tourist events elicit differential population changes. Figure 5.2.3 compares the different urban population pattern during summer and winter events. During the summer peak season, Andorra La Vella held Cirque du Soleil (July 6th). It can be seen that tourists also visited other cities throughout the country, including nature attraction areas (La Massana and Aran), and border cities. In contrast, the winter ski event (Jan 15) also attracted amount of tourists to the country, but they only stay and visit the city (St Julia) which is next to the ski resort, revealing different tourist travel plans depending on the particular event.

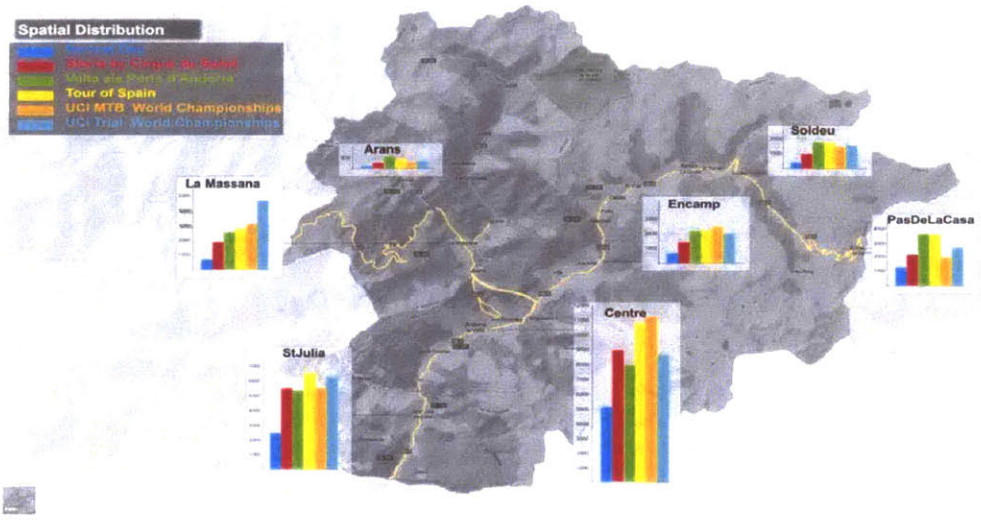


***Figure 5.3.6**: Spatial distribution of events: Method of combining Call Detail Record (CDR) and Open Data to inform the deployment of new mobility systems, adjustment of existing transit services, and optimiztion of the overall network.*

Thus, we have shown that the aggregation of CDR data reveals the spatial and temporal nature of the tourism relationships between cities, and the events they hold.

Original-Destination (OD) Interactive Map

Based on the CDR timestamps and geolocations that we collected, we next aggregated the data by tourists' initial towers' location and subsequent towers' location, in order to obtain the *mobility paths* of tourists. As an example, we found that at 10 am in the morning, the path (road) from the border city of PasDeLaCasa (neighboring with France) to Encamp has heavy movement congestion, whereas in contrast, there is high movement congestion from Soldeu to Encamp at 7pm.

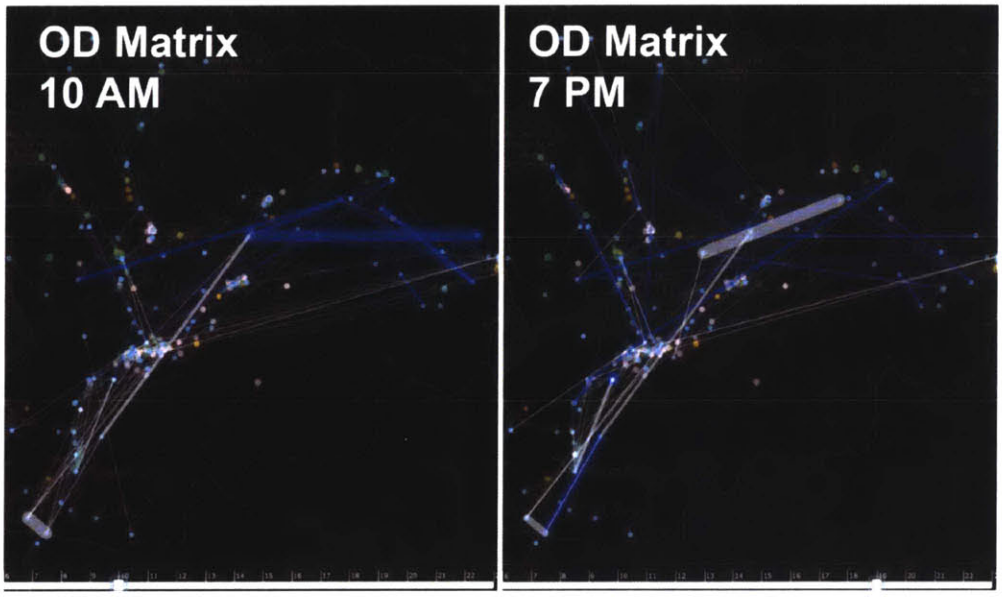


***Figure 5.3.6**: Origin-Destination Matrix showing changes in mobility pattern between morning and night time, with white lines representing Spanish tourist movements and blue lines representing French toursit movements. (Data Source: Andorra Telecom)*

Thus, this type of analysis allows us to find functional traffic problems the country, and allows us suggest to the government on the implementation of new transportation system in specific areas of the country that need it most.

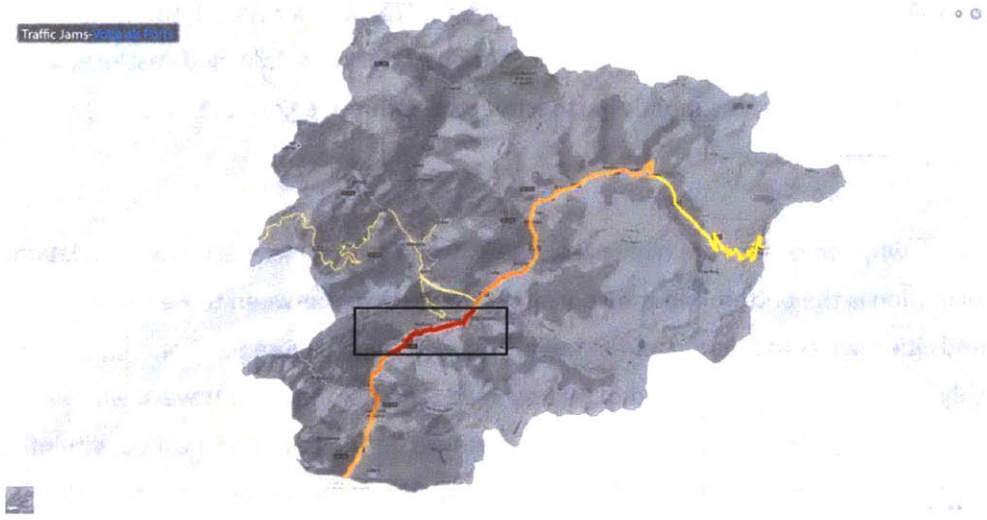


***Figure 5.3.7**: Traffic on road segments*

City Association Map

After finding the mobility pattern from the OD matrix, the next question we have is whether we may use the found associated travels between cities to prediction where tourist will travel next, from knowledge of their past travels. In this section, we use two machine learning algorithm to predict future tourist travels. As a first pass, we use the association rule algorithm to find out whether certain cities are traversed together systematically or if visitations to sequences of cities are random. We then used a more advanced algorithm, the "Random Forest" algorithm, to predict future visits based on past travel history.

Association rule

"Associations" are cooperative links between two entities, and are the most basic type of relationships. This project make extensive use of "association rule" analysis to look for these basic relationships within our data. Association rule analysis is one of the core techniques of data mining, and is applied here. We were able to discover several distinct "types" of tourist travel patterns which suggest that, if we are able to classify the behaviors of individual tourists, we may be able predict his or her future travel plans, and cater to their needs and desires, accordingly.

Applying association rule to the CDR data, we looked for cities that are highly "associated". We examined, for instance, city relations during the Easter holiday, and found that Soldeu, Andorra Le Vella, and PasDeLaCasa are highly associated (below) for French tourists. This implies that French tourists most likely to visit Soldeu, Andorra Le Vella, and PasDeLaCasa in a single trip. From an urban economic perspective, these three regions are thus economically tied to each other.

One may ask why some cities are highly associated, while other cities have low association. One explanation is the better road transportation connectivity between these highly associated cities, which facilitates tourist travel between them. Figure 5.3.7 and 5.3.8 show that highly connected cities also tend to be traversed together in tourist travels, whereas less well connected cities (such as those in the northwest) tend to less associated with other cities. These results suggest that improvement of road networks to these less associated cities (e.g. northwest) can enhance tourism in these cities, as "associated" travel to and from

these destinations is facilitated.

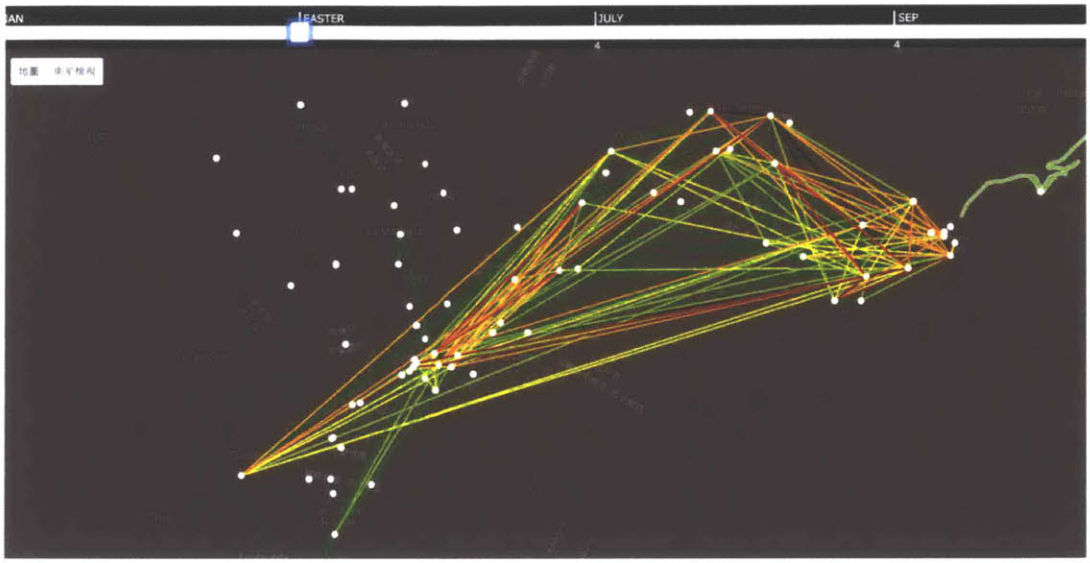


***Figure 5.3.8**: Association map show the relationship between different cities*

Random Forest

Knowing that city sequences are traversed by tourists in systematic ways, we next ask whether tourists' future travel plans may be predictable from their past travels. For this, we harness the Random forest algorithm. The Random forest algorithm was developed by Breiman, 2001, who modified a (classical) Classification Tree algorithm into the Random Forest, and created a predictive algorithm that is simultaneously highly accurate (like the neural network), while also being computationally efficient (like the classification trees); the best of both worlds. Applying the random forest algorithm, we were able to predict, to a very high degree (> 0.67), the future travel plans of individual tourists based on their past travel patterns

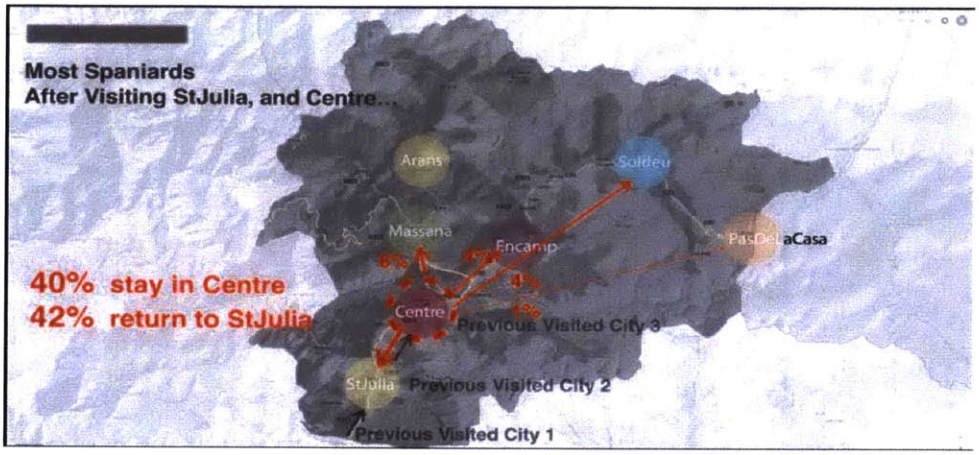


***Figure 5.3.9**: Spanish tourist tend visit other cities near the main central cities, which implies transportation system could help them to visit other cities.*

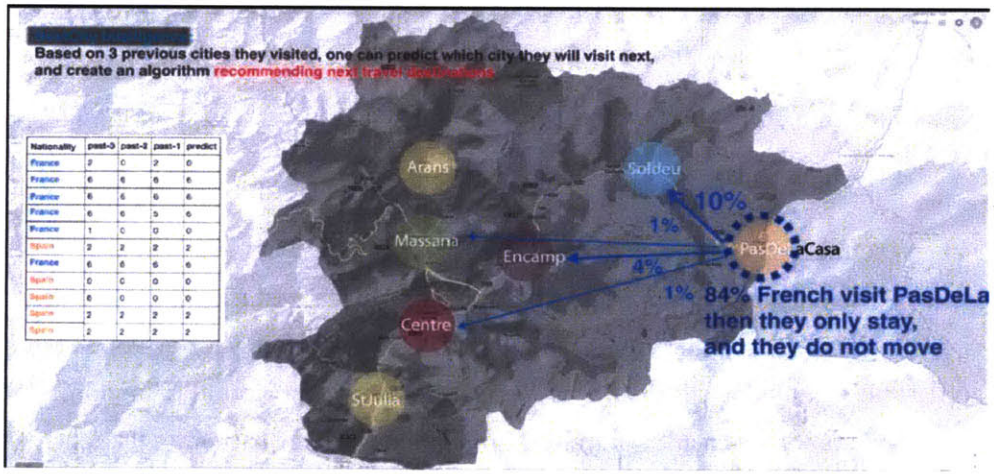


***Figure 5.3.8**: French tourist tend to stay in the same cities when they visit Andorra, which implies the urban facilities in border city meets French Tourists' need when they visit.*

To apply this work, I wish to, in my future work, create recommendation systems for tourist based on their travel histories, which could optimize the usage of urban transportation and allocate tourists equally, to avoid traffic congestion. A very simple example is the following: it seems that French tourists typically stopover in City 6. Thus, recommending restaurants as well hotels in City 6 will be beneficial to French tourists. In contrast, Spanish tourists stopover in City 1 and then immediately travel to City 0 to stay overnight. Thus, for Spanish tourists, recommending restaurants may be enough for City 1. But for City 0:CL1-center they

will need hotel recommendations. These predictive services help identify tourists' behavior.

Urban Implementation with CDR: Dynamic urban computing

The country of Andorra experiences a profound change in population during tourist events. This puts a large strain on resource allocation in the country. In fact, resource allocation is a general problem in the 21st century world, owing to resource scarcity. In general, within cities, a significant source of resource and energy waste stems from the rigid and centralized methods by which urban centers distribute resources. Schedules for distribution are typically static over time, blind to mismatch between supply and demand of resources, and so, generates waste.

A dynamical scheduling system can be suggested to solve this problem. We conceive that a dynamical system, updating resource allocation based on real time localized demand, is able to actively update areas of supply and demand mismatch in real time, and is thus able to optimize resource management at all times.

We suggest the employment of our social media data and call detail record (CDR) data mining tools as a source for updating the state of resource demand in our dynamical system. The reason is that both of these data sources satisfy characteristics of being i) geo-located with a high degree of spatial resolution, ii) user generated in real time, and iii) available over an extended period of time for reflecting influence of time and external events, all of which are necessary for our purpose.

In the future, we wish to explore this further. We plan to evaluate the use of social media and CDR data to infer resource demands by testing three uses in three major urban planning resource problems: i) how to optimally distribute public transportation services and dynamically allocate vehicles; ii) how to optimally distribute emergency services in event of man-made or natural disruptions, and iii) how to optimally distribute utilities (e.g. electricity) to prevent overloading the grid.

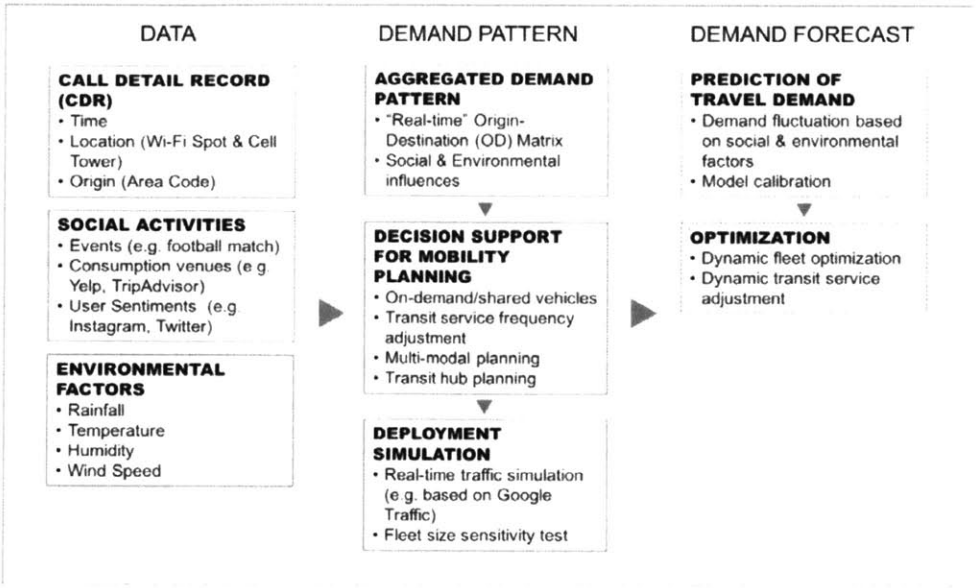


Figure 5.3.10 *Method of combining Call Detail Record (CDR) and Open Data to inform the deployment of new mobility systems, adjustment of existing transit services, and optimization of the overall network. (diagram cooperate with Phil Tinn)*

Conclusions:

In this project, we explored whether we could learn the travel patterns of tourists to Andorra from their past behavior, and in this way, predict their future travel plans. Through these analysis, we were able to identify that by far the two largest groups of tourists to Andorra were from France and Spain. We were then able to identify different types of tourist travel plans, among both the French and the Spaniards. We were also to predict future tourist travel locations from past locations. To apply this work, I wish to, in my future work, create recommendation systems for tourist based on their travel histories, which could optimize the usage of urban transportation and allocate tourists equally, to avoid traffic congestion.

Conclusion and Future Work

This thesis aimed to explore the potential of machine learning and data mining in finding patterns in "big" urban data. We explored several different types of government urban (CDR) data and social media data including Crunch Base, Yelp, Twitter, and Flickr, and Tripadvisor on two primary urban problems. First, we aimed to explore an important 21st century urban problem: how to make successful "Innovative district". Using data mining, we discovered several important characteristics of "innovative districts". Second, we aimed to see if big data is able to help diagnose and alleviate existing problems in cities. For this, we focused on the city of Andorra, and explored reasons for recent declines in tourism in the city. We also explored whether we could learn the travel patterns of tourists to Andorra from their past behavior, and in this way, predict their future travel plans and help their travels.

6.1 Conclusion

The presence of web2.0 and traceable mobile devices creates new opportunities for urban designers to understand cities through an analysis of user-generated data. The emergence of "big data" has resulted in a large amount of information documenting daily events, perceptions, thoughts, and emotions of citizens, all annotated with the location and time that they were recorded. This data presents an unprecedented opportunity to gauge public opinion about urban issues. This thesis aimed to explore the potential of machine learning and data mining in finding patterns in "big" urban data.

Chapter four explores an important 21st century urban problem that comes with a changing economy: how to make successful "innovative urban districts". To answer this question, we employed several different social media sources, including Crunch Base, Yelp, Twitter, and Flickr. We found CrunchBase to be a good "startup detector", revealing the information about relationship between the location of startups and the other physical or hidden factors. In future work, this can let urban designers understand how to make an area more attractive for enterprise and startups. We found Yelp to be a good "amenities indicator", showing the amount and quality of basic amenities in different geographic locations of cities. This helps identifies the districts still in need of civic facilities improvement. Twitter was found to be a good "activities sensor", helping visualize areas and times of day with high urban activities, which helps urban designers to rethink urban zoning, and correct land use properly. Tweets, as a local speaker, reflects not only the trend and the topic in the area, but also automatically senses public opinion (positive or negative) from the words they use. Flicker,

as an "urban perceptron", could identify "the image of the city" based on the most area for people to take pictures. Based on the information from different social media, we discovered important city indicators, and made recommendations to improve the city form for innovative urban use.

Chapter five explores alleviate an existing tourism issue in the city of Andorra. We employed Trip Advisor to help us understand the distribution of point of interest in the cities, and we applied natural language processing to this data to understand peoples' perceptions of the city. These user generated perspectives brought to light urban problems such as parking issues in the city. We also found we could successfully learn the travel patterns of tourists to Andorra from their past behavior, and in this way, predict their future travel plans and help their travels, which enables, in future work, the creation of recommender systems to help tourists in their travels. In this way, data mining social media data helps diagnose and improve an existing urban problem.

Our analysis of social media data revealed urban issues with different algorithms. By comparing the regions to the land use map, we identified a prevalent issues in the city. We suggest that our approach of combining social media "big data" with Data Mining algorithms to detect patterns of city is a useful new methodology for the urban designer and planner, and can give data-driven insights that would have been hard to collect otherwise.

6.2 Future Work

Combining Image mining with Computer Vision

Social media consists not only of user generated texts, but also user generated images and pictures. Thus an ability to automatically analyze the pictures that people post online would be a very powerful addition to our analysis toolkit. We aim to employ computer vision to detect the activities and the image of cities in social media images.

Creation of recommender systems

We have shown an ability to predict future tourist travel plans from their travel histories. An important application of this ability is to create recommendation systems for tourists based on their travel histories. In Andorra, review sites such as Yelp are unavailable, so predictive services based on our models may be very useful for travelers to that country.

Other Future work will be conducted as follows:

TripAdvisor data analysis will be compared to government GIS data to validate our spatial observations.

Analysis of other types of social media (Facebook, Instagram, etc.) will be conducted to reduce sampling bias in our data from analyzing only one type of social media.

TripAdvisor data analysis will be overlaid with Call Detail Record (CDR) data to understand the mobility patterns of tourists.

Bibliography

Agarwal, Basant. Prominent Feature Extraction for Sentiment Analysis. Place of Publication Not Identified: Springer, 2016.

Alexander, Christopher. A City Is Not a Tree. London, 1966.

Belussi, Alberto. Spatial Data on the Web Modeling and Management. Berlin: Springer, 2007. Print.

"CIA - The World Factbook -- Andorra." CIA - The World Factbook -- Andorra. Accessed April 22, 2016. http://people.uvawise.edu/pww8y/Supplement/-ConceptsSup/Maps/CountryFactBook/NationsFactbook04/print/an.html.

Ciuccarelli, Paolo, Giorgia Lupi, and Luca Simeone. Visualizing the Data City Social Media as a Source of Knowledge for Urban Planning and Management. Dordrecht: Springer, 2014. Print.

"Fact Sheet - TripAdvisor." Fact Sheet - TripAdvisor. Accessed January 22, 2016. https://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html.

"Goslate: Free Google Translate API" Goslate: Free Google Translate API — Goslate 1.5.0 Documentation. Accessed January 22, 2016. http://pythonhosted.org/goslate/.

Guzman, Emitza, and Walid Maalej. "How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews." 2014 IEEE 22nd International Requirements Engineering Conference (RE), 2014. doi:10.1109/re.2014.6912257.

Jacobs, Jane. The Death and Life of Great American Cities. Pelican Books, 1964.

Janert, Philipp K. Data Analysis with Open Source Tools. Sebastopol, CA: O'Reilly, 2011. Print.

Karen Johnson, Innovating The City: Challenges and Opportunities in Establishing Incubators and Districts in Paris and Boston, MIT 2014

Kumar, Shamanth, Huan Liu, and Fred Morstatter. Twitter Data Analytics. N.p Print.

Liu, Bing. Sentiment Analysis and Opinion Mining. San Rafael, CA: Morgan & Claypool, 2012.

Lynch, Kevin. The Image of the City. Cambridge, Mass.: MIT, 1960. Print.

Lynch, Kevin. A Theory of Good City Form. Cambridge, Mass.: MIT, 1981. Print.

"Natural Language Toolkit" Natural Language Toolkit — NLTK 3.0 Documentation. Accessed January 22, 2016. http://www.nltk.org/.

Innovation Ecosystem of Kendall Square, MIT 2013

Minjee Kim, Spatial qualities of innovation districts: How Third Places are Changing the City, MIT 2013

Pont, Meta, and Per Haupt. Spacematrix: Space, Density, and Urban Form. Rotterdam: NAI, 2010. Print.

Pranav Ramkrishnan, Data Visualization and Optimization Methods for Placing Entities Within Urban Areas, MIT 2014

Ryan, Brent D. Design after Decline: How America Rebuilds Shrinking Cities. Philadelphia: U of

Pennsylvania, 2012. Print.

Rogerson, Peter. Statistical Methods for Geography: A Student's Guide. 3rd ed. Los Angeles: Sage, 2010. Print.

Russell, Matthew A. Mining the Social Web. Sebastopol, CA: O'Reilly, 2011. Print.

Simmie, James. Innovative Cities. London: Spon, 2001. Print.

"Text Mining Online | Text Analysis Online | Text Processing Online." Text Mining Online Text Analysis Online Text Processing Online. Accessed January 22, 2016. http://textminingonline.com/about.

"TextBlob: Simplified Text Processing" TextBlob: Simplified Text Processing — TextBlob 0.11.1 Documentation. Accessed January 22, 2016. https://textblob.readthedocs.org/.

Xiaoji Chen, Seeing Differently: Cartography For Subjective Maps Based On Dynamic Urban Data, MIT 2011