High tech industries are playing an important role in the economic development in the United States. While some cities are shrinking, the "innovation" cities are growing. The attributes that cause some cities to successfully become innovative is a very relevant 21st century topic and will be investigated here. Previous work conduct city analysis through conventional government GIS or census data but such analyses do not answer questions about the perception of citizens inhabiting the city, and the activities they conduct. The novelty of this current project is to make use of large-scale bottom-up data available from social media. Several social media sources—CrunchBase, Twitter, Yelp, and Flickr—were data mined pertaining to four innovation districts in Boston. We found that the success of innovation districts in Boston were correlated with several important variables: the most successful districts tended to occur near research institutions, in very "mixed use" areas, and were unexpectedly not correlated with land and labor prices, unlike technology districts in the past. Based on our study, we make recommendations for the urban design that cities should put in place to increase the potential for "innovation".

**Keywords:** Smart Cities, Social Media, Innovation District, Spatial Analysis, Data Mining, Natural Language Processing
uses of urban spaces, for finding novel attributes in innovation districts and evaluating the districts from the perspective of residents, which helps urban planners design a better one.

**Previous Work**

Studying the attributes contributing to the success of the Innovation districts is an important topic, and some work has already been done to understand these attributes. For example, "Innovating the City: Challenges and Opportunities in establishing incubators and Districts in Paris and Boston" is a Master's Thesis in Urban Planning at MIT that is related to this topic (Johnson 2014). This thesis offers insights into the unique political, economic and cultural systems that shape innovation strategies in innovation cities. But, in order to give translatable and readily deployable advice for urban designers in Innovation districts, one must delve further into understanding the city attributes that facilitate innovation.

Particularly lacking in the study of the Innovation district is an understanding of how residents perceive the changes. Whether citizens feel benefited directly, and a deeper understanding of how they have benefited, 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 by the types of data that the government has collected (Ciuccarelli et al. 2014). The current project attempts to make important use of large scale bottom-up data available from social media.

**DATA AND METHODS**

In this project, we utilized geo-located tweets from Twitter, pictures from Flickr, and the density of amenities from Yelp. We developed and used algorithms to extract data from these social media (Kumar et al. 2014).

**CrunchBase**

CrunchBase 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, launch date, employment data, etc. Using the CrunchBase Web API, we compiled information regarding startups in the Boston region, and applied visualization techniques to examine the spatial patterns of high tech startups in the Boston area.

**Yelp**

Yelp is a company that publishes crowd-sourced reviews about local businesses, and was thus, an important source of information for us regarding the spatial distribution of diverse types of businesses and amenities.

**Twitter and Flickr**

Twitter and Flickr are both social networking services that allow users to send and receive messages and pictures, respectively. The key attribute is that all of these public information are geo-located (Russell 2011). We also recorded timestamps of "tweets" on Twitter, allowing us to understand the different local temporal distributions of activities in the innovation districts. Finally, we applied natural language processing and sentiment analyses to the tweets. Analyses of image tags from Flickr revealed that the pictures uploaded in different innovation districts tended to vary according to localized urban focus.

**Application**

These different social media sources were mined to reveal urban issues such as the locations of improperly used or underused land areas. The richness of the data enables us to make inference about the reasons behind such unpopular spaces. This social media data thus shows the image of the city from the point of view of residents and allows urban designers and government officials to solve these urban problems (Lynch 1960).
ANALYSIS
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.

*Crunch-Base as Innovation Trend Detector*
Based on the data collected from CrunchBase, we were able to apply visualization techniques to examine the spatial patterns of high tech startups in the Boston area. Several findings can be seen from our visualizations. In general, it can be seen that traditional tech companies in the 20th century were located in suburban areas to take advantage of cheaper land and wages. However, new high tech companies are more often located in the city center in the Greater Boston area, rather than the suburbs. We speculate that this is in order to attract young talent graduating from universities in the city. This observation also suggests that the transformation of local industries through the generations have also changed the preferences of company locations from sub-urban to the city.

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

**Proximity to MBTA.** Data visualization from CrunchBase reveals the relationship between startups, their headquarters, and the MBTA. These startups tend to be located close to MBTA stations, such as Kendall Square, Central Square, Financial District, and the South Innovation District. This pattern occurs presumably because the company is then able to attract labor forces commuting by public transportation system (see Figure 1).

**Locations of the "Big" high Tech Companies.** The location of "big" tech companies, such as Facebook, Google, and Microsoft, seem to have also become a critical factor influencing smaller startups when they

---

**Figure 1**
Distribution pattern of high tech companies, based on geolocated company data from CrunchBase. (Startup companies in cyan and headquarters in dark blue.)
Figure 2
Distribution of start up companies, and biotech companies in green.

Figure 3
South Innovation District and Kendall Square have a smaller proportion of civic facilities such as restaurant (purple) and grocery (green) than Harvard Square and Financial District.
choose their locations. There are two possible reasons: First, there could be a “flow of labor” from one campus to the smaller ones. Many members of these new startups 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 startups that they better understand; local startups thus have an advantage in these critical social relations.

**Locations of the Institutions.** In the Boston and Cambridge area, 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 (see Figure 2). For instance, when talented graduate students and professors from local institutions (such as MIT) invent novel technologies, they often establish a startup close to their base operations (their labs, which are located locally).

**Established Temporal and Spatial Distribution.**
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 to grow.

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. There are two types of districts popular among startup companies:

First, the innovation districts established by government, such as Kendall square and the South Innovation District. The earliest tech companies established here soon rapidly attracted new startups into the area. Second, "self-emerging innovation districts" that emerged spontaneously without government intervention, such as Harvard Square, Central Square, and the Financial District. Recently, these have become popular districts for startups. Why they spontaneously emerged is an interesting question. The next section deals with this question.

**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, compared to other areas of Boston. We mined Yelp data to examine the diversity and density of amenities in Boston. We compared the "government established innovation districts" with the "self-emerging innovation districts". The first category of Innovation districts (including Kendall Square, and South Innovation District) lack, to this day, a diversity of local amenities, such as 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 a greater variety (see Figure 3).

**Land Use and Amenity Distribution.** We examined in more detail the amenity pie charts of four key areas: Kendall Square, South Innovation District, Harvard Square, and financial district. If we only focus on "grocery stores" in the 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 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." This is a consideration for further improvement (see Figure 4).

**Twitting-scape: Twitter as Activity and Land Use Sensor**
We suggest that Twitter can 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. We investigated the difference between daytime and night-time usage of a piece of land or building; more uniform use of the buildings during the day and night means that the urban de-
sign goal for more efficient building use was achieved (see Figure 5).

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 contrast, Kendall square has a lot of data activity during the daytime, which decreases at night. This implies that human activities in Kendall square are not as popular as Harvard square during the night. This informs the urban designer that the mono-functional land use in Kendall square does not provide as many activities as the mixed-used land in Harvard Square, a consideration for further improvement (see Figure 6).

**Tweets as a Local Trend Speaker.** Using Natural Language Processing to mine tweets could help the urban designer understand important social trends and priorities of local people. We use sentiment analysis or opinion mining to identify and extract subjective information from source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker/writer with respect to some topic or the overall contextual polarity of a document.
Red indicates mixed use buildings, while purple and blue indicate mono functional use. The high prevalence of mono function buildings in Cambridge and South Innovation District contribute to the scarcity of civic facilities and night social activities in these areas.

For instance, we can understand if people's emotions are positive or negative (see Figure 7). One observation we made was that tweets related to the South Innovation District innovation district contain more negative words than tweets which is related to Harvard square, implying that the perceived quality of daily life in South Innovation District does not in many ways meet people's expectations.

Analyzing further the frequencies of words and topics in tweets provides 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, parks, and bars, frequently appear as the topics 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 this area are
related to "jobs" and "high tech". This problem is even worse in the South Innovation District: no civic topics show up at all. Only industrial-related words such as "tech" and "startup" show up, suggesting an almost monolithic focus within this area; not a sign of a vibrant district (see Figure 8).

Local Events and Activities Speaker. 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 observed from Twitter data mined.

Flickr-scape: Flickr as Indicator of Urban Life
Data mining images from Flickr can help understand how people perceive the urban environment. Image files collected from social media provide graphic data, together with 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 (see Figure 9).

Ratio in Types of Image. For example, image data from Kendall square innovation district most often pertained to the urban space. Popular topics included the local buildings, sculptures and other art, landmarks, transportation, amenities, open space, street, activities, and the Charles River. Thus, Those targets are considered to be the popular and important "image of the city" within the Kendall Square. (Lynch 1960) Of particular note is that images of the nearby Stata Building designed by Frank Gehry on MIT campus account for a landmark within the Kendall square area (see Figure 10).

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, suggests a plausible reason why Harvard square spontaneously developed into an Innovation district, despite not being a government-designated one. More and more startups tend to come to this "mixed used" area, where there are so many convenient civic amenities, as well as a world class institution, Harvard University.

APPLICATION TO URBAN DESIGN
Our social media analysis allowed the discovery of urban issues that were overlooked when the innovation districts were first designed by the government in a top-down fashion. For example, Kendall Square revealed some land misuse, even though it is a popular location for startup companies (due to its proximity to MIT and big companies already present). From Twitter data, it can be seen that there is very little activity in this Square at night and on non-working days. Also, many of its open spaces, encircled by companies, are not being used.

We have discovered a possible reason for this: the prevalence of mono-functional land use and enclosed open space in Kendall Square. Analysis through Natural Language Processing supports this. Tweets in Kendall Square primarily concern job and employment-related topics, in contrast to Harvard Square which features mixed-use land, and generates tweets including references to more events (like Pi day) and leisurely locations (such as parks and restaurants), at all times of the day.
Thus, based on this land use principle discovered by social media, several suggestions for improving this district can be made. First, government officials can change building and zoning codes to encourage and reward buildings for putting civic facilities in the ground floors of buildings in the area, and even encourage these civic facilities to extend onto the streets, so that they do not have such an "enclosed" feeling to them (see Figure 12,13). Second, street landscapes can be altered by urban designers to have more street furniture and amenities in order to facilitate a more mixed-use landscape for residents, and encourage more street activity. Third, government officials can facilitate more social events such as festivals in this area to stimulate activity in these unused open areas. In this way, appropriate urban modifications are suggested from the social media data mining study that we conducted, to further improve the innovation districts (see Figure 11).

CONCLUSIONS

All of these findings are in contrast to traditional tech companies of the 20th century that were generally located in suburban areas to take advantage of cheaper land and lower wages, and may reflect new social characteristics and needs of the 21st century. From our results, we made recommendations for the urban plan and design that these cities should put in place to increase the potential for "innovation". For the city of Boston, for instance, it became very clear to us that Kendall Square is a district lacking in diversified amenities, and based on our findings, a plan to increase local "mixed" land use will increase the ultimate success of the Square. We suggest that our approach using social media analyses is a powerful
methodology to study the emerging "innovation districts", an important topic of the 21st century.

ACKNOWLEDGMENTS
We thank C. Sun, M. Dennis, and S. Williams for discussions and paper preparation, C. Haung for help in getting the user interface to run, J. Chen, C. Summit, M. Lin and J. Nawyn for helpful discussions and comments, members of Media Lab Changing Place group. In memory of W. J. Mitchell.

REFERENCES
Ciuccarelli, P, Lupi, G and Simeone, L 2014, Visualizing the Data City Social Media as a Source of Knowledge for Urban Planning and Management, Springer

Johnson, K 2014, Innovating the City: Challenges and Opportunities in Establishing Incubators and Districts in Paris and Boston, Master’s Thesis, MIT
Kumar, S, Morstatter, F and Liu, H 2014, Twitter Data Analytics., Springer
Russell, M 2011, Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, O