Creative Cities

Contributors:
Moe Hadhrawi
Yu-Ann Wang

I. Overview and Motivation

It's now the 21st century and new cities are being built and old cities are being renovated to accommodate global population growth and urbanization. By 2009, more people in the world lived in cities than rural settings. This number will double by 2050 to 3 Bn (UN World Settlement Program).

Historically, people have long felt ambivalence towards cities. Cities have been the wellspring of ideas, the cornerstone of the industrial revolution, and centers of economic improvement. It's also been the source of socioeconomic equality, congestion, and crime. Inspired by several pieces of literature, we wanted to explore cities in a positive light and see what elements in a city could be related with higher levels of innovation.

Both our backgrounds tie into cities and creativity:

Moe is working in the Changing Places group, MIT Media Lab. He's investigating what makes cities creative/innovative using a tangible interactive decision support system. Moe is interested to connect tangible with graphical data representation to reduce cognitive load in understanding complex systems.

Yu-Ann has lived in multiple cities for work, including NYC, Los Angeles, Hong Kong, and Jakarta. In each place, she was struck by how urban planning has facilitated or disrupted people's ability to get together and work effectively. She is currently taking two courses on digital innovation and sustainable cities. She hopes to tie together these two courses and study why certain cities are effective in breeding entrepreneurship and innovation.

II. Related Work

Several pieces of work have attracted our attention:

Y combinator founder, Paul Graham, breaks down cities (NY, SF, LA, Cambridge) by its collection of people and its unique ability to generate innovative companies:

http://paulgraham.com/cities.html

The Economist's *A Cambrian Moment* discusses the decentralization of start-ups and how because there is widespread cloud computing power, it may be easier to start a company anywhere. This is less about public utilities, but does include emphasize the importance of buildings and wi-fi.

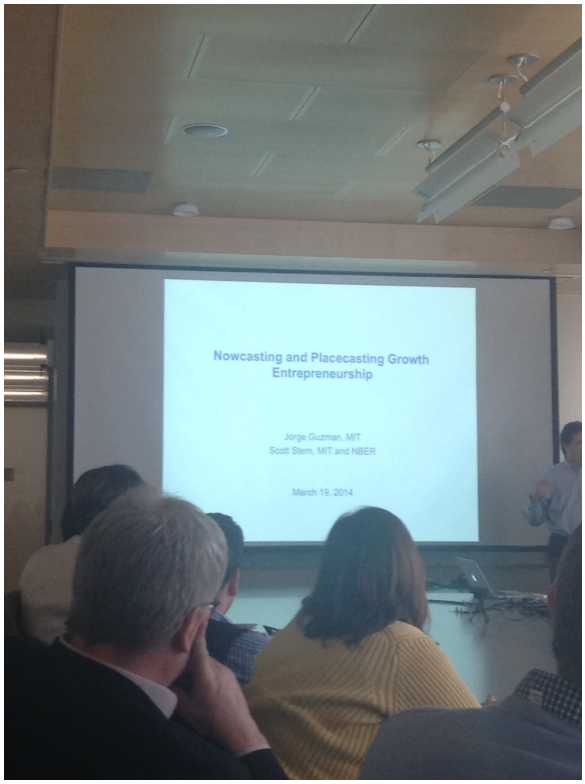
http://www.economist.com/sites/default/files/20140118_tech_startups.pdf

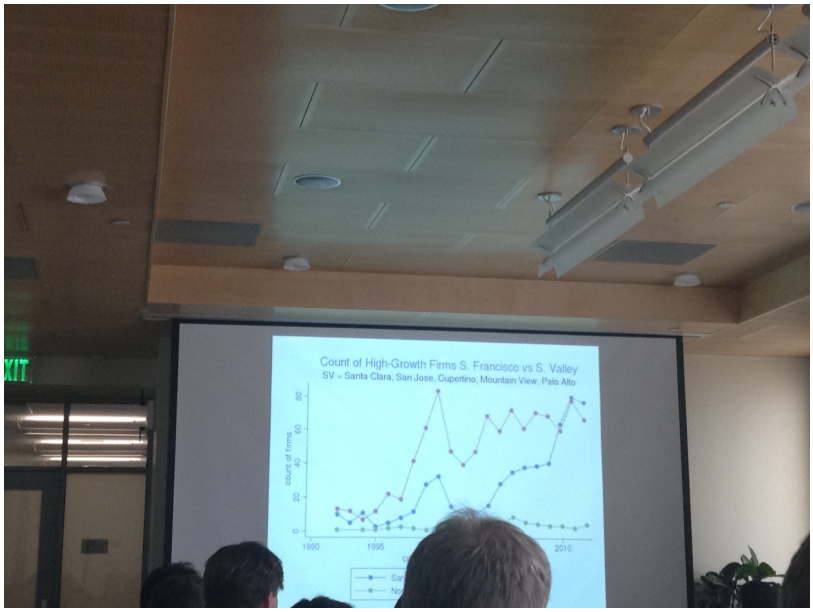
Tony Hsieh and the Las Vegas Downtown Project. Tony's idea is to facilitate creative collisions between people. He believes that innovative results come from unplanned interactions between people.

http://www.wired.com/2014/01/zappos-tony-hsieh-las-vegas/

Scott Stern and Jorge Guzman: Placecasting and Nowcasting

This paper has not been published yet, but two professors from NBER and Sloan used incorporation data to develop a predictive model on growth entreprenuership. The professors tried to extract data and compare the shift in company growth from Route 128 to Cambridge, and Silicon Valley to San Francisco. Several images from Scott and Jorge's talk are below:





Last of all, this report from the Urban Land Institute discusses increasing a city's livability factor (i.e., parks, playgrounds) and partnerships with academic and private institutions to increase innovation.

http://uli.org/wp-content/uploads/2012/06/Building-Innovation-Murphy.ashx_2.pdf

Questions

The main question we're trying to answer is in a city where there are neighborhoods for innovation, is there a correlation with smart city planning (i.e. availability of public transportation, common areas to meet and congregate, and public education availability)

Data (source, scraping method, cleanup)

In looking for data, we conducted a web search in addition to interviewing several people from Code for America, Code for Boston, an HBS Real Estate Professor, and two previous employees in the NYC government (for details on their inputs, please see here: http://bit.ly/Q7cOno

For looking at proxies of innovation, we pulled the following data
- Paper from the US Patent Office breaking down the number of utility patents in major cities
- Total number and relative percentage of creative professions in the city (data pulled from the American Community Survey using arts, entertainment, and recreation as a proxy for a creative profession)
- Pulling information on articles of incorporation (accessing this from Open Data NYC, SF) and manually pulling the information from the corporations database for Boston (the current database has session cookies which prevented us from scraping the webpage)

For looking at independent variables which could be related with innovation, we are currently surveying the following:
- Public transportation data (supplied by the MBTA, MTA, and BART) providing average ridership in the subway by station during weekdays
- Cafes (NYC, SF, Boston Open Data) as a proxy for places to informally congregate and discuss ideas
- Public school data (NYC, SF, Boston Open Data) to gauge the prevalence of education access in a city

Initially after our first milestone, we were fairly optimistic of retrieving more city data sets. On suggestion of our product TF, we pushed to introduce one more city (San Francisco). However, upon graphing and representing more of our data, we encountered the following barriers:

- Data, even the items found in OpenData tended to be inconsistent and require standardization

    - For example, depending on the dataset and our choice of visualization, we ended up requiring addresses, zip code, and lat long for each data point
    - Inconsistent data resulted in us using Google Developer API, Geocoding service to convert street address into lat/longitude values
- We found ourselves being unexpectedly bogged down by legal definitions, especially in our proxy for innovation
    - After consulting the Harvard Law School, we discovered that DBA's ("Doing Business As") certificates and articles of incorporation don't always overlap in different cities. We were pushed to reevaluate our datasets

Exploratory data analysis

We didn't focus as much on making mock visualizations as we did on outlining a set of go/no decisions based on the availability of the data. We were fairly confident that we could get innovation data on the state level and maybe the city level, but were unsure how we would go about collecting data on the more local level.

We initially hit some snags when we discovered that patent data was hard to download and didn't provide zipcodes for their authors. Another issue was working on articles of incorporation, and discovering that many cities still had their documentation in paper form (Boston offered to pull their articles of incorporation data for us if we paid them money and submitted a written letter, but we found that cumbersome, nontransparent, and time consuming).

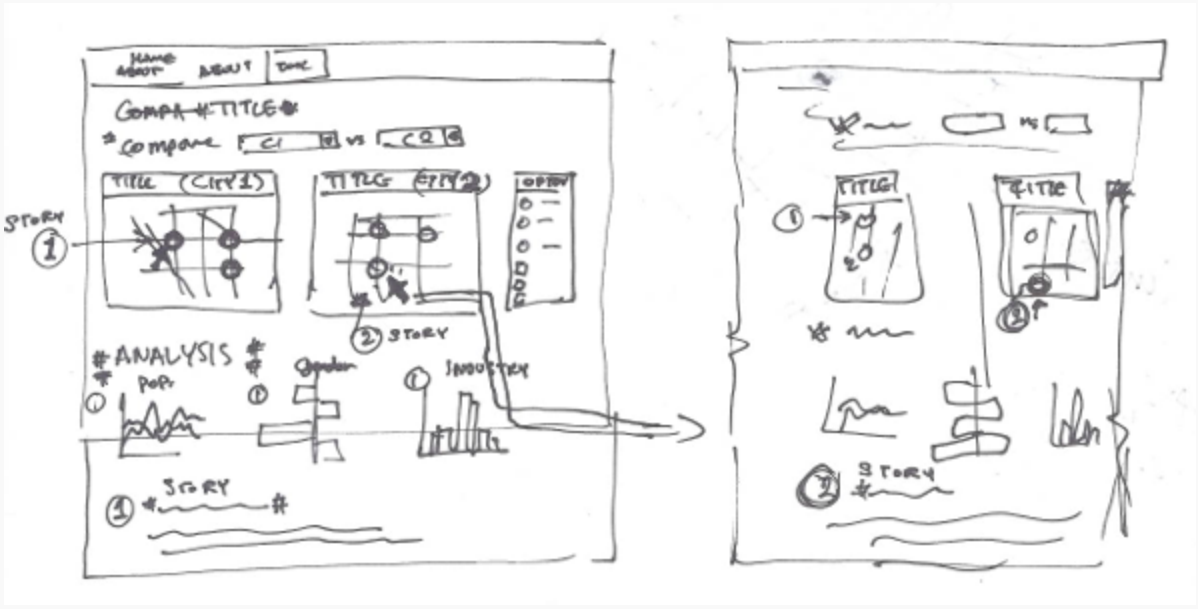
Ultimately, because our focus was on innovation in neighborhoods, we decided to go with a geovisualization with supplemental time series line graphs. We were fortunate enough to be able to retrieve data on ridership in public transportation (pulling raw data and APIs from the MTA, MBTA, and BART websites), public schools, and cafes/restaurants on the neighborhood level with details on latitude/longitude, zip codes, and addresses.

Design Evolution

We were initially inspired by the states level data we pulled for HW4 and also the geo visualizations found in the MIT Media Lab Changing Places (http://www.media.mit.edu/research/groups/changing-places) and Senseable City Lab (http://senseable.mit.edu). We were also inspired by Chris Whong's data visualization on New York city turnstiles : http://chriswhong.com/open-data/visualizing-the-mtas-turnstile-data/.

From our early discussions and design studio feedback, we realized that a geo visualization may not be helpful enough for comparing two different cities. Depending on how many layers the users examine, it's very easy to get lost in the details. We opted to also include bar charts, time series graphs on innovations and their correlated variables in the bottom of our visualization so it would be easier for our user to compare metrics in different cities.

This is a representation of our design in Milestone 1:



Since then, we've made several decisions which have impacted the look and feel of the visualization:

(1) The shift from using TopoJSON maps to OpenStreetMap

Because we used TopoJSON maps in HW4, this was naturally our default choice when we started the final project. However, we soon realized that these maps only go to the granularity of county level:

(TopoJSON map with one county selected)

Ideally, we wanted to demonstrate data on the neighborhood and street level, so the visualization was switched over to OpenStreetMap, which was more visually arresting. Implementing OpenStreetMap proved challenging at times, especially because the spatial data model is only points and lines, and the data elements are not always consistent across geographies.

(2) Using Voronoi Tessellations instead of Tiger/Line shapefiles

The US Census Bureau provides Tiger/Line shapefiles which break up the maps by zipcodes. However, when we selected the shapefiles, we often found the boundaries very uneven and difficult to compare across different cities:

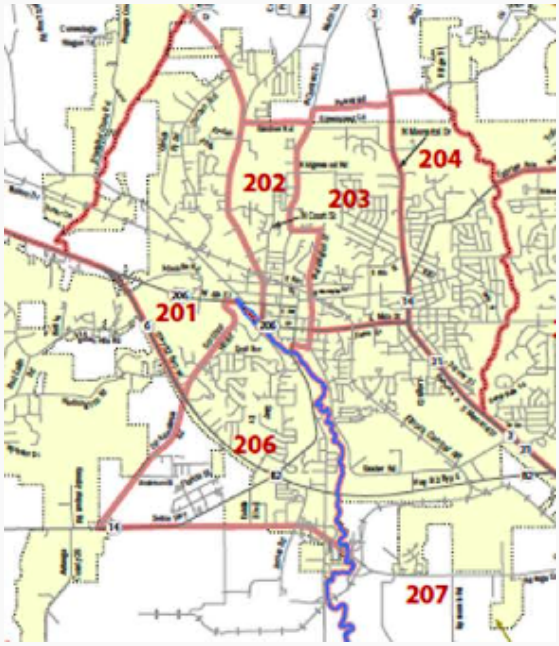


Image of zip code shape file from TIGER

Inspired by several papers on Voronoi diagrams (including a discussion on gerrymandering and Voronoi and also state capitals and Voronoi tessellation) we decided to test out a visual representation of our neighborhoods using a Voronoi diagram. A Voronoi diagram is a technique to divide up space so each "cluster" is closer to a dictated point than a set of other points provided. In our case, the points we provided were based on latitudes and longitudes provided by the Google API when querying zip codes. We also thought using a Voronoi tessellation would be easier if we decided to use neighbor queries for point location data in the future.

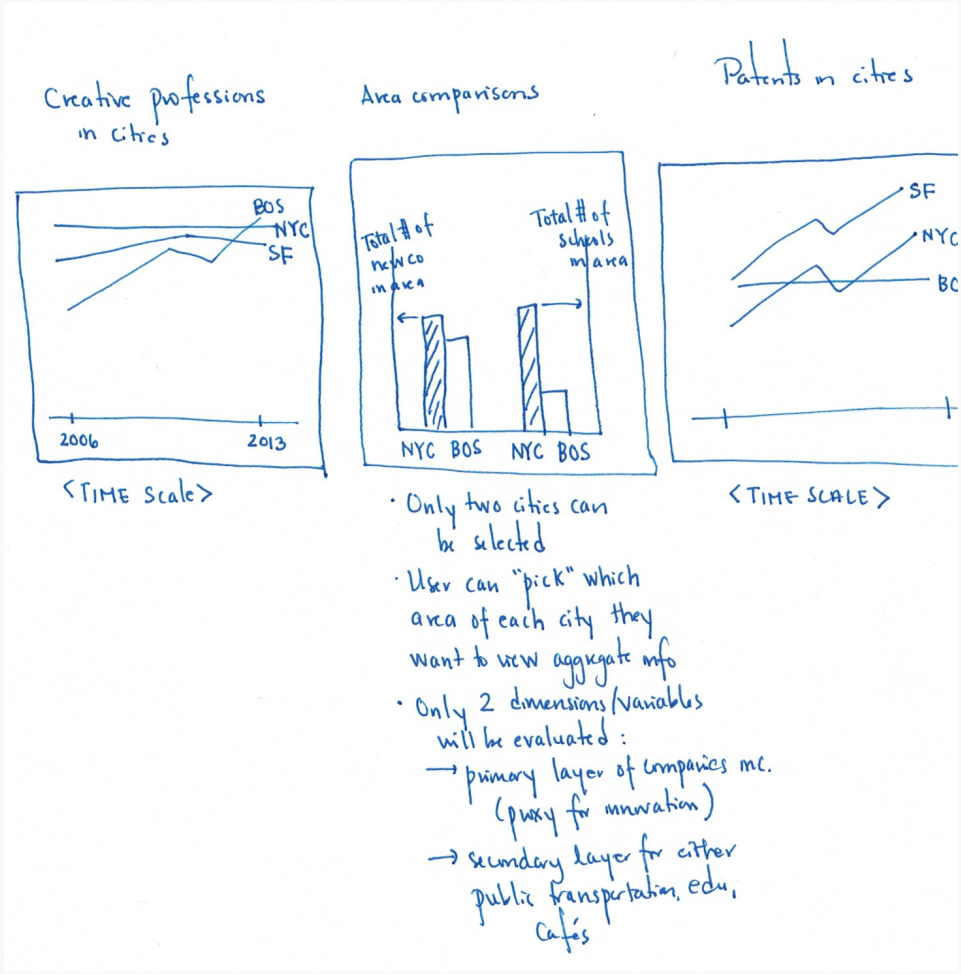
(3) Selecting which independent variables would be represented by areas and which by points

We definitely wanted our "innovation" measure to be represented by area and opacity (the more opaque an area, the more companies that have been incorporated in the area for the last five years). However, we deliberated about representing public schools, transportation, and cafes as points versus areas.

Ultimately, public schools and cafes were represented as areas (as we plotted cafes for locales like Manhattan, we noticed significant overlap in points and thought a lot of the visual data would be lost). Public transportation was represented in points form because we wanted to introduce a third dimension: ridership for the size of the dots on top the location of the stations.

(4) Fixed graphs versus brushed graphs

Based on the data collected, we were able to compare creative professions and patents issued across the three cities. We wanted another interactive graph where we could compare aggregate numbers across cities. Right now, it displays aggregate information, but in future implementations, we will allow the user the option to click on different areas of each city map and have them compare the "on"-layers variables:



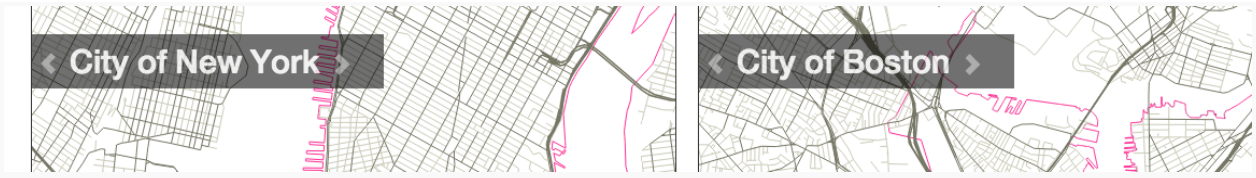
(5) Adding "stories"

Just as important as the data visualization story is the "why." Originally, we wanted to "zoom" in neighborhood area by neighborhood area, but discovered that we didn't have that much neighborhood data available. Instead, in our next implementation, we will create a stories section which will paint broad strokes on (1) how the specific variable has

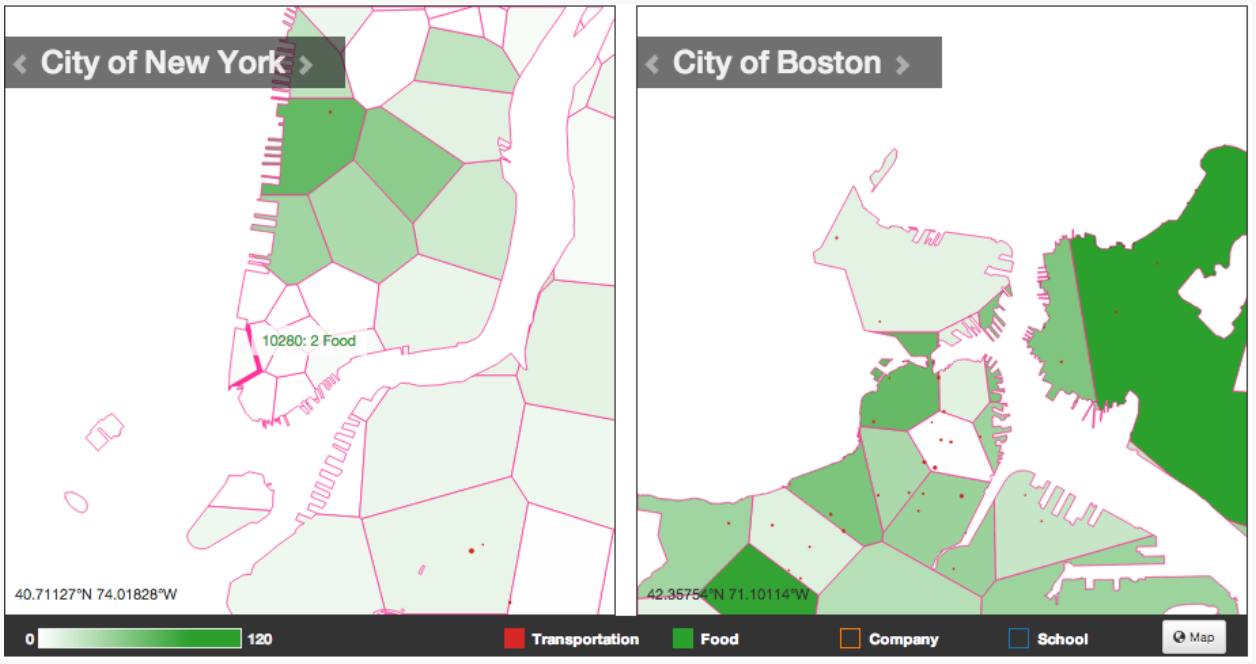
changed through recent years, (2) how it might potentially affect innovation based on local headlines we found.

Implementation

In this interface, the user selects two different cities from a menu for comparison. The choices are : NYC, Boston, and San Francisco.



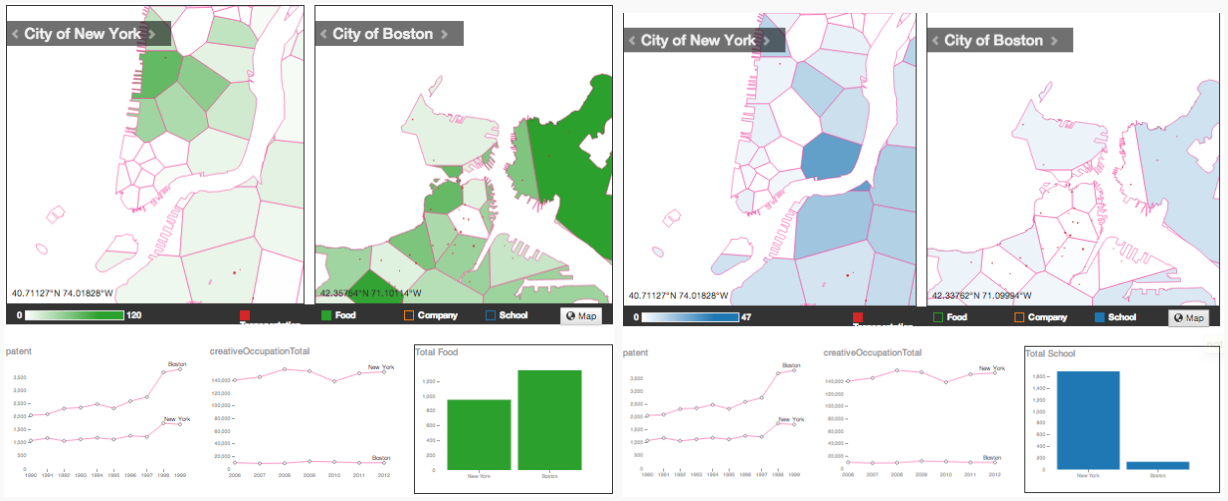
On the base of the two maps, there is a panel with checkbox inputs that allow users to turn on/off multiple spatial data layers. The data layers available for turning on/off include: public transportation, public schools, and cafes.



Below the maps, non spatial data is provided for comparison. The graphs that are fixed include aggregate city level data about patents and creative professions across a time scale (often times these two variables serve as proxies for creativity and innovation).

The last chart is dependent on (1) which cities are selected for comparison, (2) which "layers" are selected for comparison, and (3) which Voronoi areas are selected for

comparison. Here are two charts with the same cities (New York and San Francisco), but with different layers selected for comparison.



Finally, the story on the bottom is interchangeable based on which layer (public transportation, cafes, schools) is selected and which cities are selected.

Evaluation

Our major findings are:

(1) In terms of patents, Boston is the clear winner compared with NYC or SF. However, when we use other measures of creativity and innovation, we discover that Boston trailed behind the other cities in number of creative occupations and number of companies incorporated in the last five years. This shows that our proxies are not always correlated with each other (this may even be inversely correlated with older, larger companies able to sponsor expensive patents) and we may need to narrow our definition of innovation.

(2) Across all three cities, food and "innovation" (measured by new companies incorporated) are related. Because we only have current data, this poses a question of which came first, new restaurants or new companies.

(3) In terms of public transportation, SF was the only city that a clear relationship with innovation and public transportation. This also could be

because its subway system is so concentrated in the Mission District, Market Street, and Embarcadero.

(4) When analyzing schools and companies, we uncovered a clear relationship in Boston and SF. Public schools were concentrated in Jamaica Plains and Brookline, while in SF schools were located in Mission and SOMA. In New York, there was less of a relationship, with public schools being located downtown and new companies being incorporated further up town.

Overall, besides food, we've discovered each city has its unique footprint in how its public services are developed and where its pockets of innovation are located. What we thought to be very global observations ended up varying widely from city to city.

Going forward, we believe the following changes could improve our visualization:

(1) Including more independent variables (initially we had pitched internet connectivity and walkability scores as other items which could be conducive to innovation and creativity).

(2) Define innovation and creativity on a narrower band, or at least demonstrate what is the correlation for each proxy. Right now, we're using companies incorporated as our major proxy for innovation, but we also have thrown in creative professions and patents filed as two alternative proxies. We'd like to investigate how closely these items are correlated, if at all.

(3) Investigate the notion of lagging variables. For example, maybe well developed public transportation/higher ridership volumes or cafes opened up after an area became the innovation district/neighborhood. There's really no way to tell because we're investigating things on the same timeline. However, if we introduce time sequences into our geovisualization (i.e., public transportation started increasing first in this neighborhood, and then new companies started to form around this area), it would be a more insightful story.

(4) Introduce more creative ways of using Voronoi tessellations. We bounded our Voronoi tessellations using zip codes, but as we could see from the

gerrymandering and Voronoi article, there are many alternative ways to bound regions (such as population density). In the next iteration, we may diverge from zip codes and use another variable to bound our regions.