Innovation Hubs - A Story of Cities and Patents —

# **Overview**

This capstone project will investigate the factors that facilitate innovation growth for a company **using publicly available patent data, regulatory factors and census data.** And then, we will aggregate the company’s data to city level, in order to predict the innovative cities as well.  
- To understand how this process develops, we will analyze patent data in the United States from 2001-2020.   
- The regression analysis will help in exploring many features that influence the growth of innovation.  
- Upon running multiple analyses across the years we will find those few features that have higher influence on patent output amongst the top cities.   
- We will also come to know if any of these feature is missing among cities with less patent output.   
- What this experiment would recommend to cities desiring greater patent output is that they should invest in higher education, in earning Small Business Innovation Research (SBIR) grants, and looking into becoming empowerment zones.  
- This way, we will identify how cities can invest in their economic future. This is accomplished by developing an innovation score. This will provide an understanding of the urban structure that facilitates the growth of innovation in the form of patent production.

# **Different Data points used for measuring Innovation**

### **Patent Data:** The patent data obtained is from Patents View. It contains detailed information on every patent assigned to a United States based organization. These are as follows: - a list of assignees (those who own the rights to the patent) - and a list of inventors (those who contributed to the innovation itself — each with an associated company and location) - Additionally, every patent contains a number of citations( the amount of new innovations built off of this patent). By aggregating the number of patents assigned and patents invented with their associated citations, two scores were generated that indicate innovative development: Patents Assigned and Patents Invented.

### **Regulatory Factors:** The Regulatory Factor data includes three aspects.

These are as follows:

- A federal award data: This can be defined as money the federal government has promised to pay to companies, organizations, government entities or individuals. This is done by contracts, grants, loans or direct payments. **Federal awards data are available, with each year having millions of awards. Each award has 260 features** ranging from funding agency, federal obligation, to recipient, and performance center, and location. **So, each year’s data will be aggregated to average amount of federal obligation and total number of awards based on recipient city and primary place of performance.** The average amount will be used instead of the total amount of federal obligation to account for the large variation in the sizes of cities being analyzed.  
  
- Additionally, the Small Business Innovation Research (SBIR) program will be investigated as well. The SBIR program is a federal funding program that enables small businesses to get financial awards from federal agencies' R&D budgets helping thousands of small businesses with over $100 million awarded every year since 1982. **For the SBIR program, data is available from 1983 to 2019 with an average of six thousand awards each year and each award containing general information of each business including its location, the amount of award the business receives, its funding agency and topic/field for each awarded project. The average amount of funding per business received and number of businesses awarded were calculated for each city in the model. The goal is to measure if this program has encouraged innovation within cities.**  
-The Empowerment Zone Initiative is a tax incentive and public funding program initiated by the US Department of Housing and Urban Development in 1994 that intended to revitalize economically distressed areas**. Empowerment Zone data are available at city level and included in our model to see if these zones witnessed transformations into innovation hubs. For these empowerment zones, a binary variable was put in place to indicate if the city federal assistance.**

### **City Diversity Data:** Demographic and household data can be collected decennially going back to 1970 from the Census’ IPUMS National Historic GIS at the place levels for the entire U.S. Features of interest include total population, median household income, education, and nativity. In order to determine how many people fit Richard Florida’s creative class, we collected US Economic 5 Year Data from the Census API. Years available were 1997, 2002, 2007, and 2012. Data collected includes the number of employers and employees per each job title as described by North American Industry Classification System (NAICS) per Census designated place. Richard Florida describes the creative class as those in academia, arts, and other professions requiring an advanced degree. We mapped the job titles to create, or not, and summed the number of creative and non-creative employees per city in order to determine the size of each city’s creative class. Processing the US Economic Census data resulted in five features including the number of ‘creative’ and ‘regular’ employees and employers, and a unique city identifier.

# Data Aggregation

A unique code per each city was designed as such, city-state (ex. sanjose\_ca), among all datasets in order to join them. We then performed a left join on of our collected and processed data onto the patent data in order to keep as many of the original 1000 top patent producing cities.

# Model Selection

"Our analysis consists of creating a classification and using coefficients to understand each feature’s influence on the classes. To do this we applied a logistic regression to our given data, splitting our data into two classes. We defined the split at the 75th percentile of patent production and can tentatively say the top 25% of cities for each score constitute innovative centers. What this experiment hopes to do is to identify features, and subsequently policy decisions, the bottom 75% of cities might hope to enact for greater patent output. Multiple model scoring metrics were used to determine the model performance including log loss, area under curve (AUC), confusion matrix and precision-recall. After calculating the scores per feature per year, scores were formatted into a time series across our years of interest to highlight feature influence changes over time.

A random forest regression will be used to confirm and add a measure of robustness to the feature importance results from the logistic regression. Outside of the impurity scoring feature, importance given by sci-kit learn, we calculated the model score using a method that permuted the features and dropped a feature over multiple iterations.

As a result, we are able to determine each feature’s added value to the model."

# Limitations

A major limitation faced by this experiment was the decline in number of municipalities through various stages of this study. With the original top 1,000 patent producing cities, the retention rate after all the joins with other datasets was found to be between 55%-80%. One reason this occurred was due to Patents View irregular use of geographies within the same location feature. For instance, New York City, NY and Woodlawn, IL -- a neighborhood within Chicago -- were both listed within the top 1,000 patent producing cities. Finding neighborhood level data across various data sources was not possible and, as a result, neighborhoods, towns, and small  
municipalities were mostly dropped. This amount of data reduction allowed for limited model selection, with the random forest and logistic regressions as the best options.  
External data sets not only limited the number of cities, but also the years of analysis. Patents View data went back to the 1970s, however, between all these external data sources, the window of analysis was shorted to 2001 through 2012. This window allows for the experiment to determine what features are significant to city patent production, but not how these cities became patent producing 'hubs'.