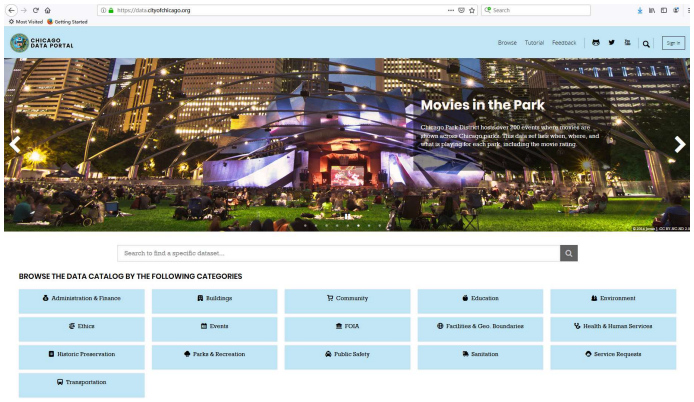
## DATA

There are many factors will be put into consideration, and I think there are a few areas of particular importance, which include safety, education, and neighborhood.  There are publicly available data in this regard. From Chicago Data Portal(http://data.cityofchicago.org/), for example, we can get census data, school data and crime data etc.

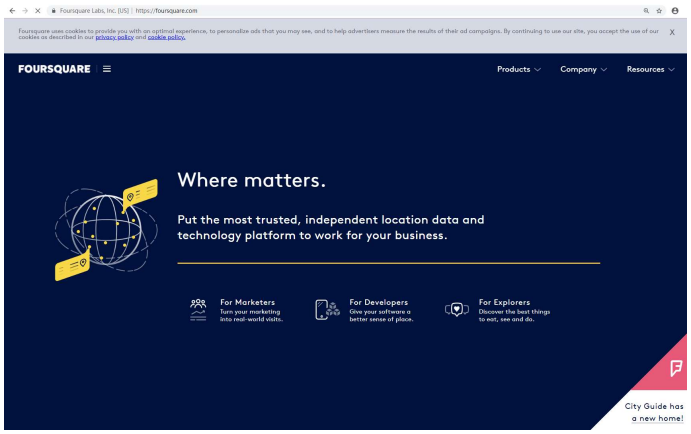



For census data, this dataset contains a selection of six socioeconomic indicators of public health significance and a “hardship index,” by Chicago community area, for the years 2008 – 2012. The indicators are the percent of occupied housing units with more than one person per room (i.e., crowded housing); the percent of households living below the federal poverty level; the percent of persons aged16 years or older in the labor force that are unemployed; the percent of persons aged 25 years or older without a high school diploma; the percent of the population under 18 or over 64 years of age (i.e., dependency); and per capita income.

The school data is from Chicago Public Schools - Progress Report Cards (2011-2012). This dataset shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year. There are many type of data in the report, and I’m focusing more on SAFETY SCORE: Student Perception/Safety score from 5 Essentials survey, and COLLEGE ENROLLMENT (NUMBER OF STUDENTS): Total school enrollment.

The crime data is from Crimes - 2001 to present report. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present. It mainly included report time, type of location, exact position, type of crime and whether arrested or not. Since it is a huge dataset, so only a very small sample of this dataset is used in this research.

Furthermore, leveraging Foursquare location data can help explore or compare based on their first-hand information. Food / restaurants available is an important fact for consideration, and I think many will agree on this.




## METHODOLOTY

A. Census data

With census data, we can have a basic understanding of the different community areas in Chicago. Because there are 77 community areas in total in Chicago, each has different attributes, it's better to put them into different groups. And for this purpose, **K-Means clustering** would be an ideal and fair model, as it does not need any preset on the inter-relationship between different attributes.

I used this model to put all 77 community areas into 4 groups. After the clustering, I plotted the grouping results with Axes3D, with axis set to 'Hardship', 'Housing Crowded', and 'Household Income'.