BU Spark!
Fuqing Wang, Tian Chen, Zixiang Wei and Xiao Lu

Cambridge Eviction Study Project Proposal

Background:
Over the past two years, the City of Cambridge has been collecting and analyzing data from the state of Massachusetts online court dock system in order to achieve a better understanding of the factors that cause eviction both demographically and economically.

Goals:
Our project aims to explore research questions on what causes evictions under market-rate housing stock and where they might occur in the future. Strategically, our goals for this project are twofold. First, we are going to update and optimize the existing data collection and analysis procedures based on current data scraping and collecting tools. Second, we will develop an supervised-learning analytic model that demonstrates the relation between eviction and different social aspects(demographics, housing conditions, spatial characteristics, and personal reasons) both at present and in the future.

Non-goals, extensions:
As an extension of our proposed goals, one of our non-goals is to implement visualization tools that show the density and distribution of evictions and how these might change over time. On top of that, we would like to apply unsupervised-learning techniques to our existing model in order to achieve a deeper understanding of what causes evictions and possibly ways to prevent them, if time and datasets permitted.

Product:
By the end of this semester, we will propose an automated data analytical model as well as integrated visualization tools which could present analytical results right on the screen. Ideally, our analytical model should be scalable subject to adding additional data source and tools for data manipulation in the future.

Uncertainties:
Regarding uncertainties, one big question mark we have is about the scraping tool the city is currently using - our current documentation walks through how to set up the tool, but does not specify how the tool is built. Thus at present, there is some concern on how much flexibility it opens in terms of technical works such as modifying the scraping code and calibrating the workflow, etc. Second, based on current observation, the dataset is “noisy”. On the one hand, it is beneficial that it provides us endless possibilities for research topics. But on the other hand, is also risky whether this high-dimensional, yet low volume dataset would be robust enough to support our analysis.