**Capstone Data Ingestion Process**

1. Determine business problem.
   1. Abandoned homes are a financial and safety burden on local governments and their constituents.
2. Determine scope.
   1. Indianapolis, IN has a documented vacant and abandoned housing problem. Initial research will be done on a subsection of Indianapolis, including the Arsenal Heights, Englewood, and Fountain Square neighborhoods. These are all residential areas.
3. Labeled or unlabeled data? – Supervised or unsupervised machine learning?
   1. I was able to find a labeled dataset from the city of Indianapolis (<https://data.indy.gov/datasets/abandoned-and-vacant-housing>) with 7215 instances of vacant or abandoned buildings. Furthermore, this data is updated regularly and is current up to 01/21/2021 as of this writing. Since I found a labeled dataset from a veritable source, I chose to follow a supervised machine learning process.
4. Source the data for the basemap.
   1. My initial intuition was to use satellite imagery data as the base image for the model training. I found a data source (<https://lidar.jinha.org/download.php?cname=marion&clon=-86.13305839196093&clat=39.779844384833936&years=2011,2016>) to download satellite imagery data from The U.S. Department of Agriculture, Natural Resource Conservation Service (NRCS), Indiana State Office, and the Indiana Geographic Information Council (IGIC), under separate BAA Agreements in partnership with the United States Geological Services (USGS) 3DEP Program. This data is publicly available and free to use without restriction.
   2. The data from LIDAR scans, which are similar to RADAR and stands for Light Detection and Ranging. This data is available for download in a compressed format, which had to be decompressed using either a paid software application or through an open-source software, LASzip, available to windows OS exclusively. Since I work on Mac OS, I created a Windows Virtual Machine and connected through a remote desktop client in order to decompress the .laz files into a format I could use.
   3. Determine the Coordinate Reference System (CRS) of the data. You are familiar with one CRS: latitude/longitude. Another is the US Survey Foot, which is the format this particular data came in.