**Project Scope**: Understand the drivers of misprediction in Zillow estimates

* To limit scope, confine the evaluation area to Washington, DC
* Also to limit scope, leverage a set timeframe for data collection. This can be TBD for now until we clarify how much data we’re able to collect on a go-forward basis
* While original scope was to focus on distressed properties, opening scope to all properties will provide more data for model training and validation

**Data Sources**: Zillow property level data + Zillow neighborhood data + Opendata.dc zip and neighborhood info

* Zillow property level data will be directly accessed through the Zillow API.
  + **Challenge**: API only allows 1000 calls per day unless they approve call increase. Additionally, Zillow API requires property ID when making calls, so population of data will require creative solution.
    - One approach is to leverage the GetComps API to return data on additional properties. Another is to randomly select from property IDs in DS (commonly found in the 4xxxxx – 5xxxxx range). Not efficient, but a potential approach.
  + **Challenge**: In order to correctly define the target, we must identify the point in time relative to the sale date at which we want to capture the Zestimate. The more restrictive the window, the few target events are captured.
* Zillow neighborhood data will be accessed by the neighborhood / Zip code level information accessible in .csv format
* Opendata.dc has data on demographics, business and economic development, zoning etc. Suggest leveraging datasets that allow for the proxy of neighborhood growth and investment, growth in business investment and zoning changes, and demographic changes
* Primary key for combining all data will likely be at the zipcode level. For some Zillow data, there is the opportunity to drill down to the neighborhood level

**Methodology**: Drivers of misprediction can be predicted leveraging a number of statistical methods, essentially following the practice of residual modeling

* Linear regression, with a target of %relative error in Zestimate prediction
* GBM, with a target of % relative error in Zestimate prediction
* Unsupervised approach, clustering factors that are associated with differing levels of misprediction