*Pre-11/23/2013*

* Revised geocoding scripts for new data
* “Hard-coded” expected data types in SQLite for better data management later on (this created a bit of a mess in the earlier effort to scrape data from the assessment role website)

*11/23/2013*

* Geocoded approximately 10,000 addresses via Google.
* Attempted to find boundaries of municipalities used in assessment role data
  + Try 1
    - Create raster of municipalities from point attributes
    - Raster -> Polygon
  + Try 2
    - Create polygons from point attributes using convex hull (slow) [QGIS]
  + Both methods affected by geographically misplaced data (however still workable; could be fixed manually)

*11/24/2013*

* Found municipality shapefile from geocommons (WGS84), boundaries match results of convex hull (minus discrepancies) providing some validity to both
  + Used convex hull around outlying points to create regions not included with the geocommons map
  + Manually clipped to outlying regions to prevent overlap (this will inevitably mean there will be a small amount of unmatched addresses from here)
* Selected points on assessment role in areas not covered by geocommons map. Perform convex hull on just these points to generate the remaining polygons not included from geocommons
  + The reason for using an existing map is that convex hull can generate some overlap from interlocking geographical features since it generates a footprint from the extents the furthest in every direction from the center (elastic band closing in around pegs analogy). This means irregular boundaries such as those determined rivers can be misrepresented.
* Finished geocoding of 2009 sales (for 2011 assessment role)
  + 17014 addresses total
  + 15085 successfully geocoded (Google API)
  + 407 possible duplicates (most seem unlikely to be wrong – often resulting from two different names for the same locale)

*11/25/2013*

* assigned municipality labels to geocommons shapefile from consensus of assessment role points within it
  + Sampled 5% of data (56979 points – 1.3 million becomes a bit unwieldy)
  + Spatially joined polygon (municipality) name to sample points
  + Wrote python script: Iterate through shapefile points, assign most common (will be an overwhelming majority) municipality code used by assessment role to polygon
    - Created lookup shapefile for reverse comparison of geocoded mls sales
* Plotted geocoded sales, spatially join municipality code to each sale depending on which municipality polygon it falls in
  + Used this data as criteria for SQL matching query for mls sale to assessment role (street\_number, street\_name, municipality\_code)
    - (arrondissement shouldn’t be necessary? Ask if there’s any instances of multiple streets having the name like in Boston)
  + 12666 of the geocoded mls sales are located within the region covered by the assessment roll data
  + 7885 addresses matched to the assessment roll

Notes:

* Is there an orientation field in assessment data?
* If there the street number listed in the MLS sales contains a letter and the Google geocoded response does not, it will fail when it should not (e.g, 3848A rue Drolet if 3848 rue Drolet is returned)
  + This will require another script function run against failed\_lookup.sqlite
* PCI should be trimmed from geocoded databases