# 11.205 – Intro to Spatial Analysis – Fall 2019

## Week 5 – Geocoding: Objectives

**Part 1. The Geocoding Process**

1. Obtain Data
2. Format Data
3. Geocoding with Online Geocoder
4. Visualize Using GIS
5. Use a Spatial Join to Summarize our Data
6. Format Data for QGIS Geocoder
7. Use QGIS Geocoder
8. Use Open Street Maps Geocoder Through QGIS.
9. Merge geocoded file to have one complete file.

**Part 2. PARSING ADDRESSES AND GEOCODING WITH DESKTOP GIS**

VI. Parsing addresses for geocoding.
VII. Geocoding using a desktop GIS

In lab this week we are going to focus on geocoding. The lab will be split into two parts; the first will introduce the geocoding process using a free, web-based geocoder called OpenCage Geocoder. In the second part we will use ArcGIS or QGIS to explore the backend of geo-coding. How does it work? Why does it work?

The Cambridge Public Health Department is seeking a list and map of all grocery and convenience stores in the city in order to inform policy around the alleviation of food deserts. They have asked you to find a list of grocery stores and convenience stores in the city of Cambridge, create a map of their locations, and see how many grocery stores are in each neighborhood.

## PART 1: THE GEOCODING PROCESS

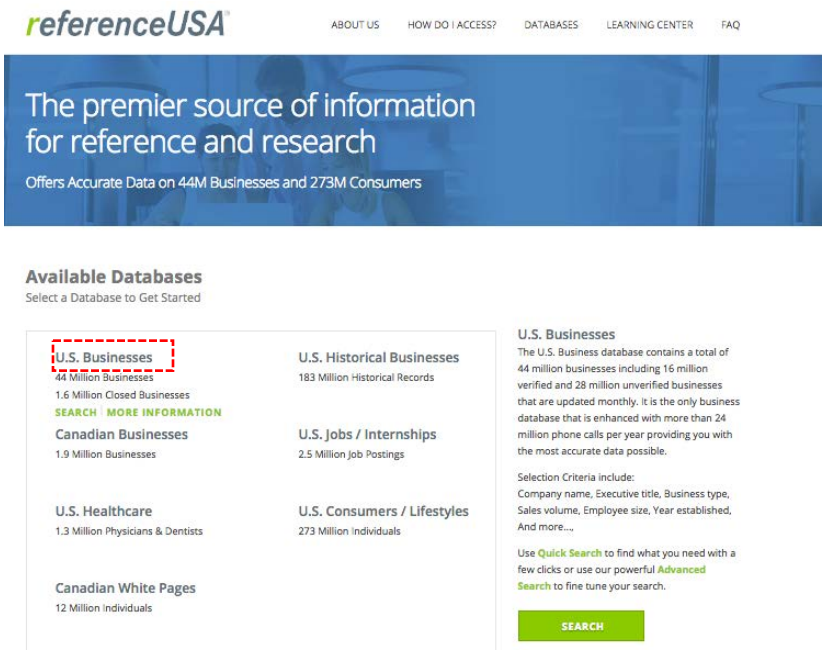
Locating and plotting addresses and point locations is a fundamental function of GIS software. The process of spatially locating tabular address data is called **geocoding**. In lab today, we are going to geocode a tabular dataset containing information on grocery stores in Cambridge. We will download the data, detail a web-based geocoding workflow, then look in-depth at how to clean, parse, and ultimately geocode addresses using a Desktop GIS.

### OBJECTIVE 1: OBTAIN DATA

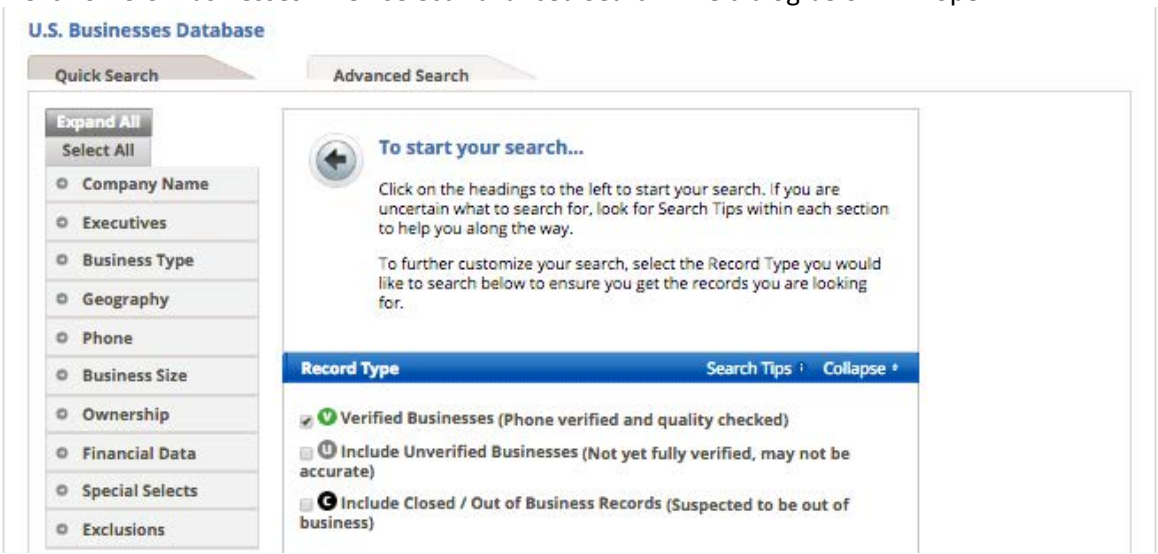
Let's get started! The first step is to acquire a tabular address dataset. This dataset can be any tabular data that contains addresses for each feature. In this exercise, we are going to download a table of grocery and convenience store addresses from a business database called Reference USA. Reference USA contains information on business locations and characteristics, and is a subscription service. MIT

has a subscription, and so do many major public libraries across the country. It is a valuable source of business-related information.

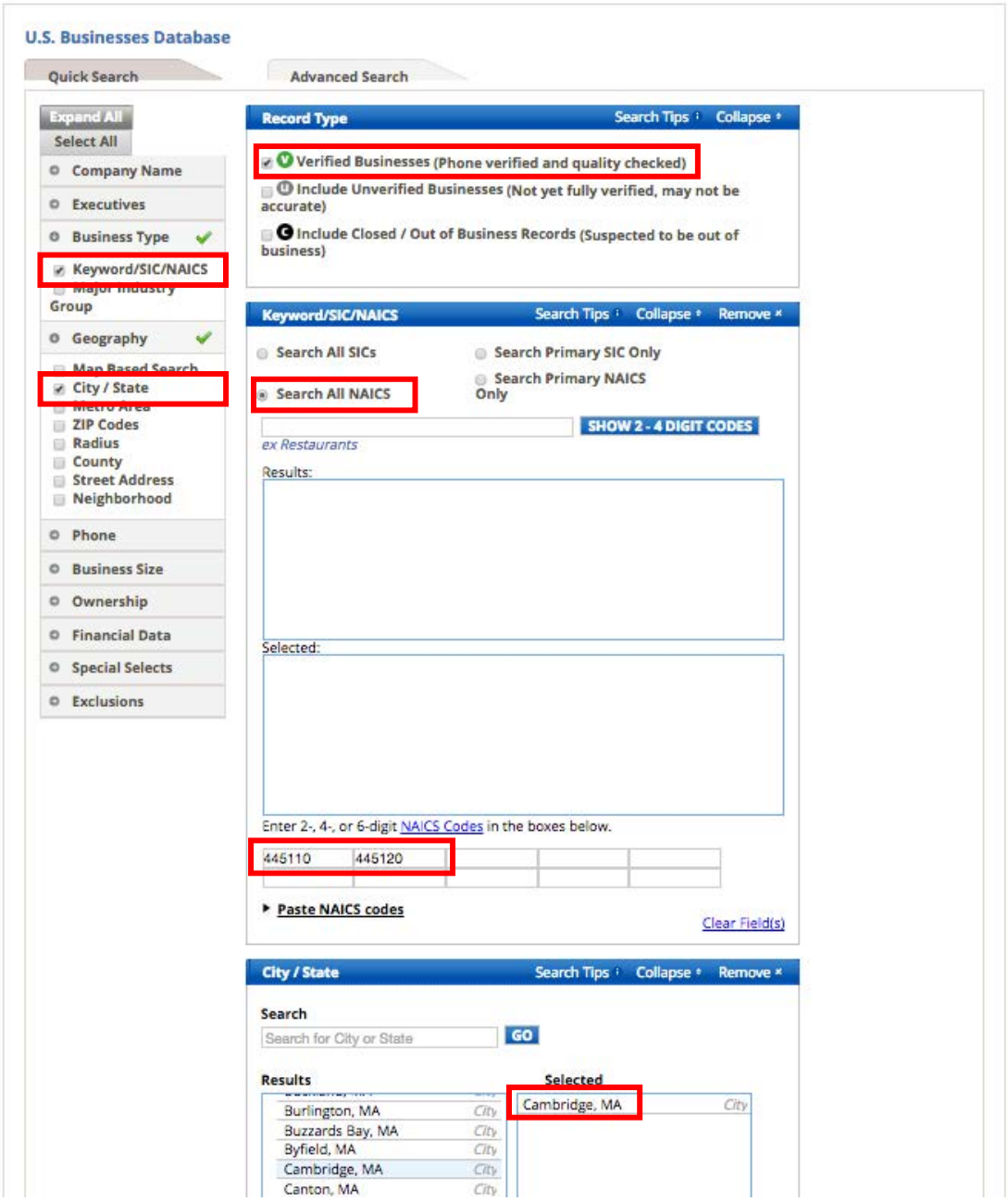
1. Navigate to Reference USA (http://libraries.mit.edu/get/ref-usa). You must have MIT certificates on your computer or be on the MIT network.




2. Click on U.S. Businesses. Then select Advanced Search. The dialog below will open.




In the Custom Search dialog, we want to specify the data we are looking for.






Fill out the form with the following:

- **Record Type:** Include only Verified Businesses

- **Business Type:** Keyword/SIC/NAICS - Search All NAICS -> 445110 (Grocery Stores) and 445120 (Convenience Stores) See the entire listing of NAICS here: http://www.naics.com/search
- **City/State:** Cambridge, MA

When done click View Results.

> *Side Note: What is NAICS?*
> From the Census website: "The North American Industry Classification System (NAICS, pronounced Nakes) was developed under the direction and guidance of the Office of Management and Budget (OMB) as the standard for use by Federal statistical agencies in classifying business establishments for the collection, tabulation, presentation, and analysis of statistical data describing the U.S. economy. […] NAICS is based on a production-oriented concept, meaning that it groups establishments into industries according to similarity in the processes used to produce goods or services."

3. You will now see your resulting table from our query; we see 70 results, but as this database is regularly updated, you may get slightly more or fewer results. It will contain listings for all the verified grocery stores and convenience stores in Cambridge. Click Download to download a tabular table of the data and save it to your working directory.

After clicking download, select:

1. Excel
2. Summary
3. All Records
4. Download Records and save as **cambridge_grocery.xls.**

## OBJECTIVE 2: FORMAT DATA

The first step towards working with tabular data is cleaning and preparing the dataset. To do this, open the file in Microsoft Excel. You may see a warning message reading 'The file format and extension of 'Summary[timestamp].xls don't match. The file could be corrupted or unsafe. Unless you trust its source, don't open it. Do you want to open it anyway?' This is fine – go ahead and open it.

When you open the data, notice that there is an **'Address'** field that contains the street number and name. Also note there are **City, State, and Zip Code** fields. These fields are necessary for our geocoding, as they contain the information we need to create an Address Locator and locate points. Other fields might be useful as well, though we won't necessarily use them this time around.

1. Delete the Fax Number Combined, IUSA Number, Primary SIC Code, Primary SIC Description, SIC Code 1, and Record Type field.

2. Remove spaces from field headers and reduce their length. These should **make sense to you**. ArcGIS cannot handle headers longer than 8 characters, which is a sensible enough threshold, even if you are a Q user. I used these names:

| | |
|---|---|
| Company Name | name |
| Executive First Name | execFrst |
| Executive Last Name | execLast |
| Address | address |
| City | city |
| State | state |
| ZIP Code | Zip |
| Credit Score Alpha | credit |
| Executive Gender | execGdr |
| Executive Title | execTtl |
| Location Employee Size Range | empSize |
| Location Sales Volume Range | salesVol |
| Phone Number Combined | phone |
| SIC Code 1 Description | type |

3. Finally, we also need to create **a unique identifying column**. Create a new column called 'ID' at the left of your sheet, place a number 1 in the first row. In the second row, use the formula **=A2+1** to increment by 1. Then fill this to the bottom of the table. Click the column, copy it, right click the same column, and select Paste Special > Values.
4. We need the address to be in one field so we need to format it.
5. Create a new field called **Address2**.
6. Click on the cell and copy the following formula so that address is in one line.

   =CONCATENATE(F2, ", ",G2, ", ", H2, ", ",I2)

Notice that formula has added spaces and comma so the program know the differences between the town, state, and zip.

Double-click on the right-hand side of the cell to make the formula go all the way down the addresses. The result should look like the following.

7. Now we need to make this concatenated address permanent. Same as before, select the column, copy it, right click the same column, and select Paste Special > Values. You will know you did it correctly if you don't see a formula when you click in any of the Address2 cells.

8. Save the prepared file as '**cambridge_grocery_cleaned.xls'**. If you get a message box saying that 'some features in your workbook might be lost if you save as Excel 97-2004 Workbook (.xls).' Click 'yes', and save the file. This will be our starting point for additional cleaning work!

## OBJECTIVE 3: GEOCODING WITH THE US CENSUS BATCH GEOCODER

There are a large number of geocoding services available and they change all the time – one free option if you are geocoding something that is under 2,500 records is OpenCage. What's great about OpenCage is that it uses a number of different opensource geocoders in the background. This means that your data is still owned by you. If you use Google's Geocoding API technically they own your data. OpenCage prides itself in following the guidelines of GDPR.

1. Go to the OpenCage Website (https://opencagedata.com/) and sign-up for an account. Once you have your account set-up you should see on your dashboard an API key. You will need that to geocode your addresses. Copy your API key somewhere as you will need it later.

2. OpenCage makes it easy to geocode using a lot of different format – everything from Java to R. We are going to use the option that uses the least amount of coding – Google Sheet. Yes, you can geocode directly from Google Sheets.

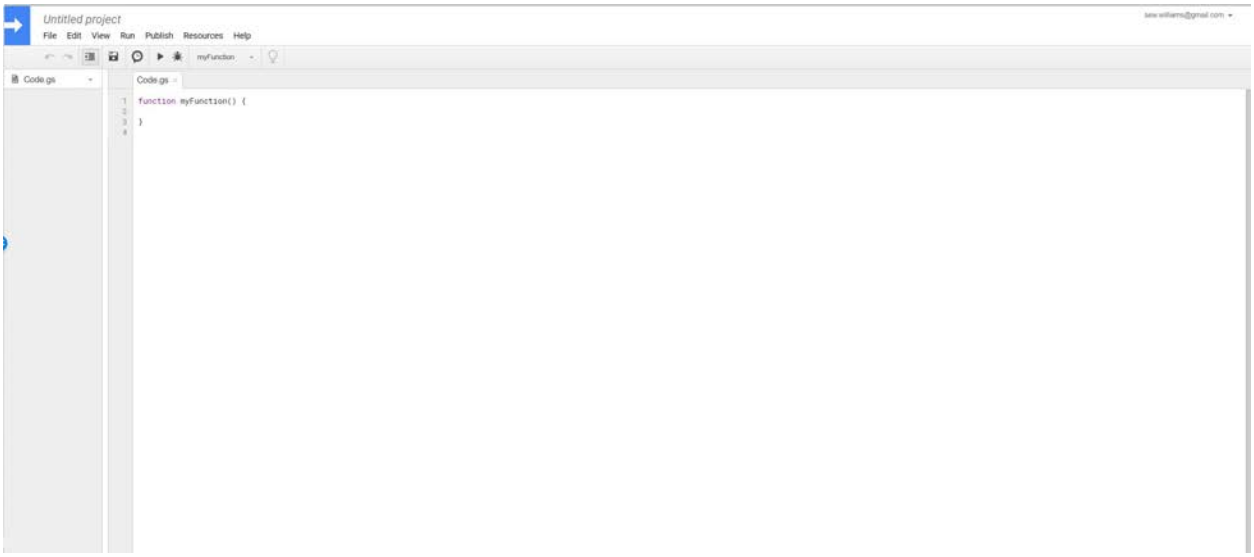
**Load the Excel file you just created into Google Drive.**

1. Create a new **Google Sheet.** Select File > Import > Upload, and select **cambridge_grocery_cleaned.xls. In the import file box that appears, select 'replace Spreadsheet' and import the data.** Do not simply upload the **.xls** file as we need the file to be natively in Google Sheets**.** Name the file **Cambridge Grocery**

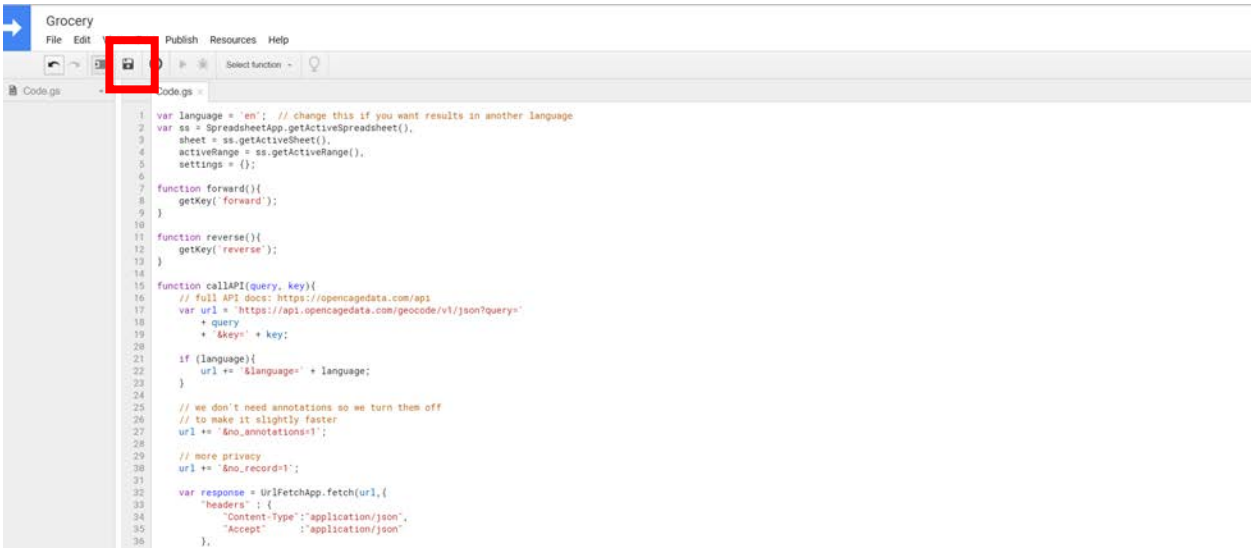
2. Once Opened in Google Sheets **add 2 new fields** - latitude and longitude.

**Add the Open Cage Script to Google Sheets**

1. **In your Google Sheet click Tools> Script Editor and the following dialog will appear. We need to add our OpenCage script into this dialog.**




2. Go to OpenCage and click on **SDKs/Tutorials** from the main menu. Scroll down and pick the Google Docs Tutorial.
3. The tutorial asks you to go to [opencage.js](opencage.js).
4. Now copy all the code in the code space and go back your Google sheets script editor and completely replace with this script. Call the script project Grocery and save the script using the save icon. It should look like below.




5. Go back to your Cambridge Grocery Google Sheet (not the script editor) and reload the page. After about 30 seconds you should see a new tab in your sheet called Geocode.

6. Select on three columns: "Address2, Latitude, Longitude

7. Then go to the Geocode tab and select Geocode > Address to Latitude, Longitude

8. You will be prompted to give the script permission to run. You will need to select "Continue".
9. A popup will appear asking you which Google account you want to authorize. Select one.
10. The popup will display a warning that our app isn't verified. Click "Advanced"
11. Scroll down and then click on the link to "Go to Grocery (unsafe)"
12. Now you will need to give the project access to your spreadsheets. Click "Allow"
13. A popup will appear in your spreadsheet asking you for your OpenCage API key. Enter it and click OK. If you forgot where you got this, you can open the emailOpenCage sent you when you signed up or you can find it in your Opencage dashboard.
14. You will see a “Running script” message on. Now we wait. According to the Opencage’s website, it takes roughly 7 seconds for each address to be found through their API. 7*77 = 539 seconds when you divide that by 60 you get roughly **9 minutes** (it’s usually a little quicker). So, take a break to ask a question about the process so far, organize your files or anything else you want to do.
15. Below is what the results should look like.

You will notice that it tried to geocode the Header field. Return that to “Latitude” and “Longitude”.

You will also notice that there are a number of errors. What do you think the reason for this might be? Ok I can’t help myself... it is because of the qualifying numbers such as 468 Broadway #2. The #2 is confusing the geocoder service. Delete the #2 in the Address2 column. Now, highlight the Address2, Latitude and Longitude fields for only that row that you corrected and then run the geocoder again on just that record. It should fix the situation.

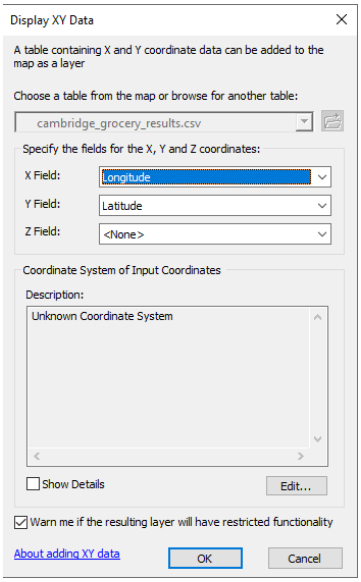
Now go ahead and fix all the Addresses that have errors. Can you think of systemic ways to fix this rather than going through them one by one?

Once you have corrected the addressed and geocoded all 77 address let’s download them so we can visualize in GIS. Say File>Download and download as a CSV. I called mine **cambridge_grocery_results.csv**

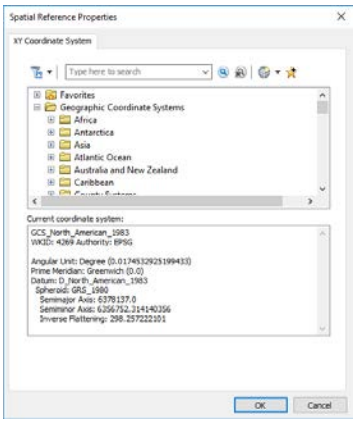
**Objective 4: Visualize Using GIS**

Now, open up ArcMap, and add this sheet to your TOC. Next, let’s display XY data by right clicking on the cambridge_grocery_results.csv file → Display XY data.

A display data dialog will appear. Here we need to set our settings for properly showing the data.




Set Longitude to be the X field, and Latitude to be the Y field. Before clicking OK, we need to tell the GIS what coordinate system our data is in. The data is in NAD 1983, which is a geographic coordinate system for latitude and longitude points. Under 'Coordinate System of Input Coordinates', click Edit…



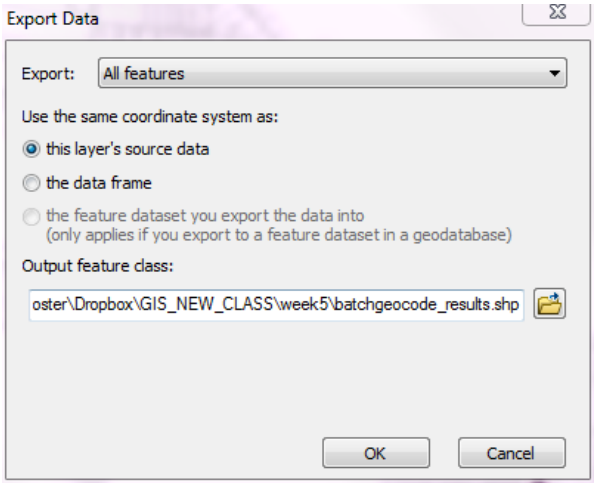

**In the Spatial Reference Properties window, go to Geographic Coordinate Systems -> North America -> USA and Territories -> NAD 1983.** Click OK. This sets the geographic coordinate system. If we don't do this, our points will show up in a different location. We will go into depth on coordinate systems in a couple of weeks in lecture and lab.

How did we know it is in NAD 1983? This can be found in the Metadata (FAQs) on the census site.

https://geocoding.geo.census.gov/

In the Display XY Data window, make sure your settings are good, and click OK to add points to your map. You will see a warning message pop up, this simply tells us we are getting an events layer, **not an actual shapefile**. This means we need to export the data to work with it.

Points will populate on your map. It will look like the following.



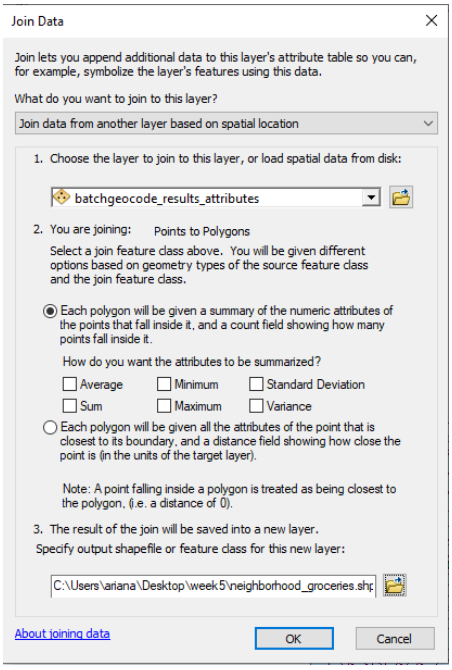

To work with the dataset, right click and export the data to a new shapefile. Call it **'cambridge_grocery_results.shp'** and save it to our weekly workspace.

## OBJECTIVE 5: USE A SPATIAL JOIN TO SUMMARIZE OUR DATA

We now have a point shapefile of all verified grocery store and convenience store locations in Cambridge. Using a Spatial Join, we are going to find out how many grocery stores and convenience stores are in each Cambridge neighborhood.

1. Add the geocoded addresses point shapefile we just created to a new ArcGIS project alongside **Cambridge_StreetCenterlines.shp**, and **Cambridge_neighborhoods.shp**.

2. We are going to join the grocery points to the neighborhoods. Right-click on the **neighborhood layer** and select Joins and Relates -> Join.




3. In the dialog that appears, select the following.

a. 'Join data from another layer based on spatial location'

b. 'Each polygon will be given a summary of the numeric attributes of points that fall inside it, and a count field showing how many points inside'

c. Set the output to save into your workspace, and name it 'neighborhood_groceries.shp'.

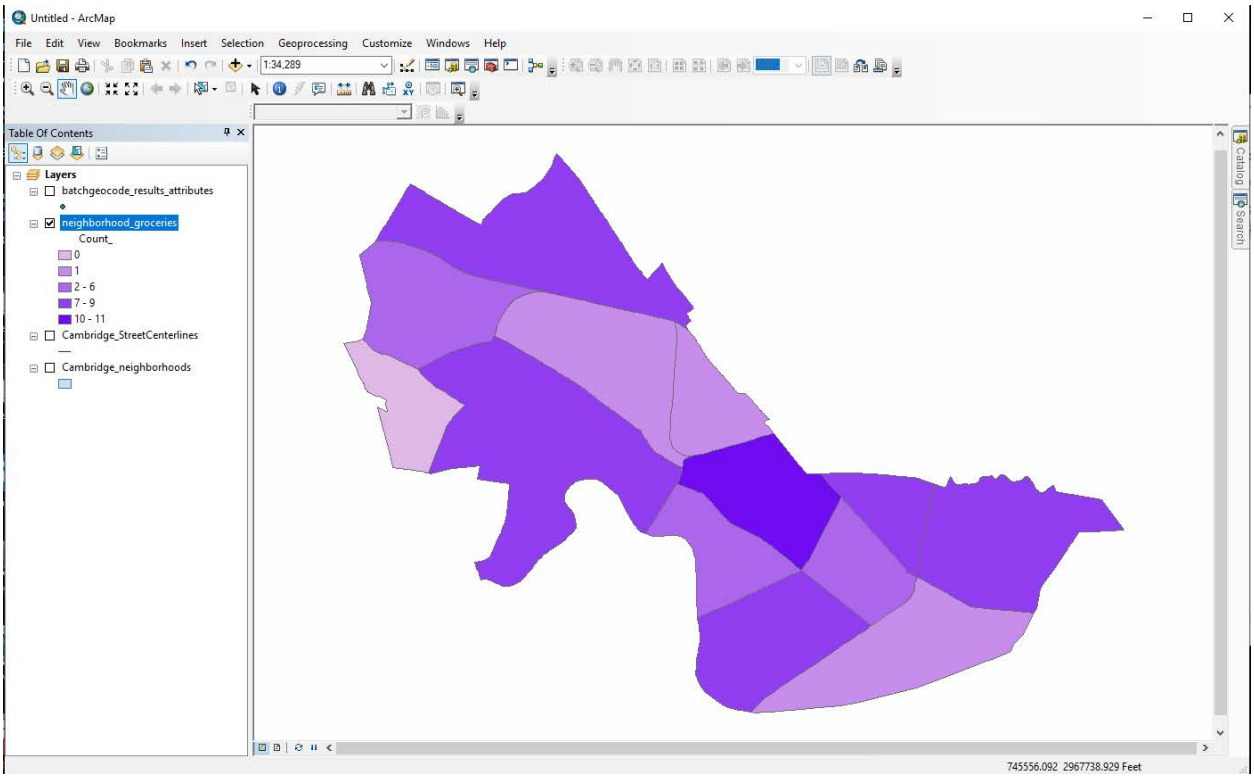
4. Click OK. The Spatial Join will create a new polygon shapefile that has the summary counts of our grocery stores for each neighborhood. Note that the spatial join creates a new shapefile, where tabular joins do not. Open the attribute table and see the summary information.

Table

neighborhood_groceries

| Webpage | N_HOOD | NAME | GlobalID | Count_ |
|---|---|---|---|---|
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/2.aspx | 2 | Area 2/MIT | {7E7AA8AC-B184-4114-8848-CA4EA161AFB9} | 1 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/5.aspx | 5 | Cambridgeport | {BA795F74-0E4E-4544-A5AB-97309EE9ADE1} | 8 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/7.aspx | 7 | Riverside | {BC1F44A9-60BA-4DC6-8B71-1FCA98C190A4} | 4 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/6.aspx | 6 | Mid-Cambridge | {54BA721E-5EC2-48D7-8740-CEE618B77F88} | 11 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/3.aspx | 3 | Wellington-Harrington | {BCF123EA-5006-4D6B-8591-3318C200CC9D} | 9 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/1.aspx | 1 | East Cambridge | {1533C532-2655-4F3E-A274-AB150DC1B985} | 7 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/12.aspx | 12 | Cambridge Highlands | {846719F3-3C41-4A17-85AC-3031318E4C36} | 6 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/13.aspx | 13 | Strawberry Hill | {E9ABF249-C82D-4719-8B99-062ED890B4B8} | 0 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/10.aspx | 10 | West Cambridge | {FC9B6813-1DAC-4354-900A-B157984DA2F7} | 8 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/11.aspx | 11 | North Cambridge | {CB9A6B5D-9F49-40BE-8872-966360EC4ADF} | 8 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/4.aspx | 4 | The Port | {93279187-CD9E-428B-A49B-265B9C552D1D} | 6 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/8.aspx | 8 | Agassiz | {63D00AF5-E679-40F6-94F1-34D3E6B04E33} | 1 |
| http://www.cambridgema.gov/CDD/planud/neighplan/neighs/9.aspx | 9 | Neighborhood Nine | {8ABEBD64-DB5A-443C-88F9-EB0ACCF41C44} | 1 |

5. Symbolize the map by the number of grocery stores (Count_). Your map document should look similar to this.




**Now continue on to part 2, where we'll learn to geocode using Desktop GIS!**

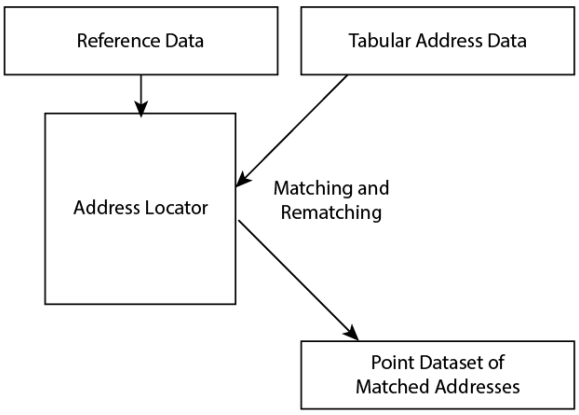
## PART 2: PARSING ADDRESSES AND GEOCODING WITH A DESKTOP GIS

In Part 1 of this exercise, we used an online service to geocode data downloaded from Reference USA. This is convenient, but it hides much of the inner workings from us; as a service, it's great, but it leaves us with much to learn!

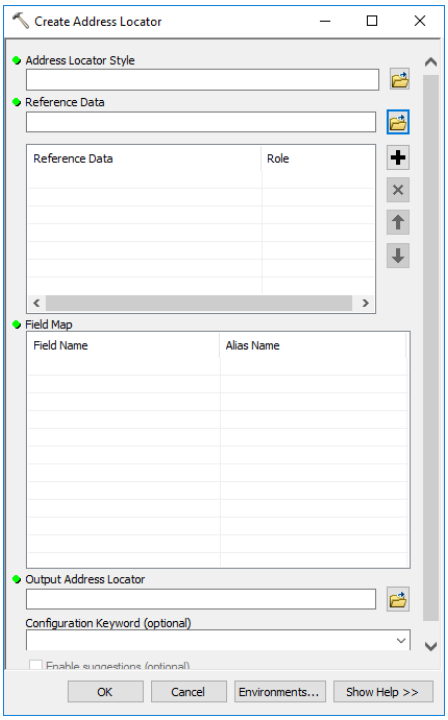
In order to geocode using a Desktop GIS, you need to have two sources of information: **a table containing addresses**, and a **reference file or database that contains streets and information about their name, address number ranges, and zip codes**. With these two files, you can create what is called an **address locator**, which is a functional tool in ArcGIS that runs geocoding based on criteria you set. The address locator will match tabular address data to a reference file. The address locator will not

match everything, so we must match unmatched addresses manually using a troubleshooting process. Finally, we can export our geocoded points to a shapefile.

The workflow for Address Locators is based on the following model.



## OBJECTIVE 2: CREATE AN ARCGIS ADDRESS LOCATOR




To begin geocoding the data we must now create an Address Locator in ArcGIS. **Address Locators** tell the software how to convert tabular data into spatial points using a reference file. The reference file contains line work with an address range and street name attributes, and interpolates where our address is located along the line work in the network. Open ArcMap.

1. Open the ArcCatalog window. In your working space for the week, right click the folder and select New -> Address Locator. The Create Address Locator dialog will appear.

2. Click on Address Locator Style. Here we set the style of address that we have in our spreadsheet of address data. Set the style to be **U.S. Addresses – Dual Ranges**. This is because our data set has street ranges on both sides of the streets.

3. Click on **Reference Data**. This is where the reference dataset for our address locator is identified. For this first dataset, we are going to use the City of Cambridge Street Centerline dataset. The City provides this free on its GIS website, and it contains attributes in the file that allow for geocoding in the City of Cambridge. This street dataset is provided for you in your weekly materials. Navigate to and select Cambridge_StreetCenterlines.shp as our reference file. Set the role for it to be the Primary Table.

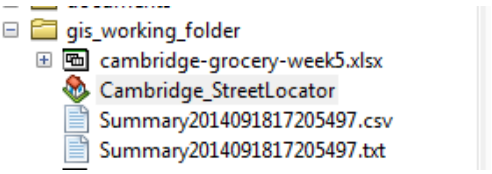
4. Click on the **Field Map**. This is where we match fields in our reference dataset to components of the Address Locator. For example, match the field that contains the Zip Code on the right side of the line to the Right Zip Code field name. Match the fields from the Cambridge Streets dataset as follows:

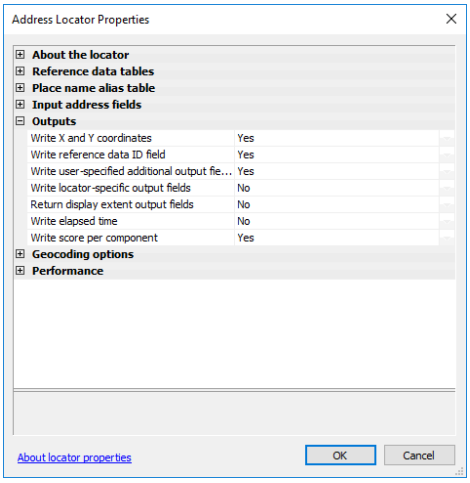
| **Field Name** | **Alias Name** |
|---|---|
| Feature ID | FID |
| *From Left | L_From |
| *To Left | L_To |
| *From Right | R_From |
| *To Right | R_To |
| … | |
| *Street Name | Street_Nam |
| Suffix Type | Street_Typ |
| … | |
| Left ZIP Code | ZIP_Left |
| Right ZIP Code | ZIP_Right |
| … | … |

***Leave the rest as <none>***

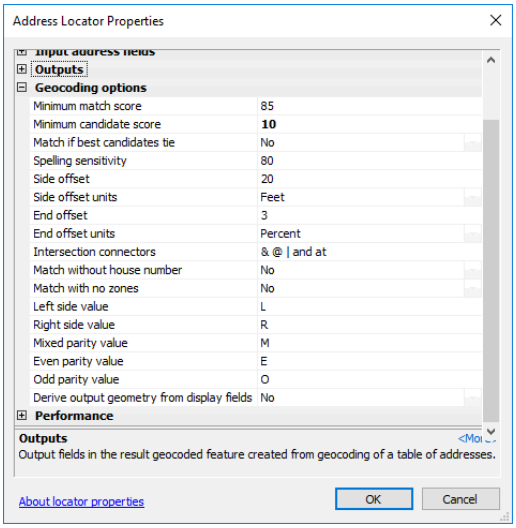
5. Save the Address Location Service in your weekly workspace as Cambridge_StreetLocator. Click OK. You may need to wait a second or two as the process runs in the background and the results appear in ArcCatalog. Leave Configuration Keyword blank (this is optional).

6. Next we need to configure the properties of the Address Locator to give us the desired results. In the ArcCatalog tab, right click on your newly created Cambridge_StreetLocator Address Locator and select **Properties**. We need to set the Outputs, Geocoding Options, and Input Address Fields (in no particular order).




7. In properties, expand the "Outputs" tab to set what is included in our output file. Make sure that **"Write X and Y Coordinates" is set to 'yes'** (so that we will have latitude and longitude information).



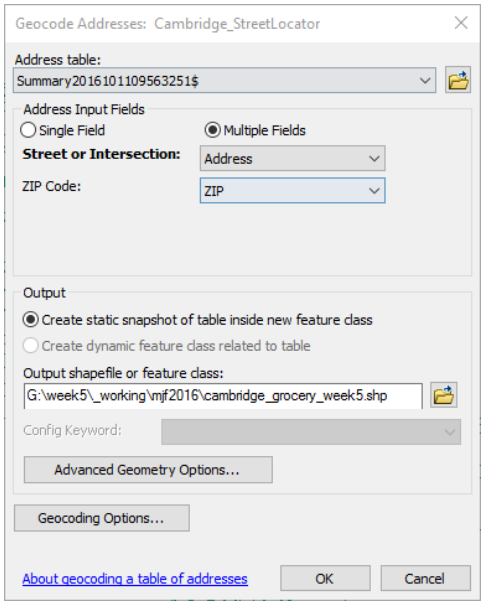

8. Select Geocoding options. Set **minimum candidate score to 10** (we don't want to miss any possible matches). Change **Match if best candidates tie to No** (we don't want to match up candidates if they tie). Set **side offset to 20 feet**, **end offset to 3 percent** (to represent typical block/street patterns). This will offset points to 20 feet to the side of the street.

9. Now, let's click **on Input Address Fields**. This option sets names of fields in our dataset for the address locator to recognize. If we have custom or different names, this is where you set this. **Make sure ZIP is included under the ZIP option**. It will likely already be there, but we want to double-check. **We want to do the same with the 'Street' field, make sure 'Address' is listed next to it**. Click OK.

## OBJECTIVE 4: GEOCODE ADDRESSES

1. In ArcMap, add the street file, Cambridge_StreetCenterlines.shp, and the tabular address data, you downloaded in the OpenCage part of the exercise -called **cambridge_grocery_results.csv**

2. In the **List by Source** View  in the Table of Contents, right click on the tabular grocery address data, select **Geocode Addresses**…

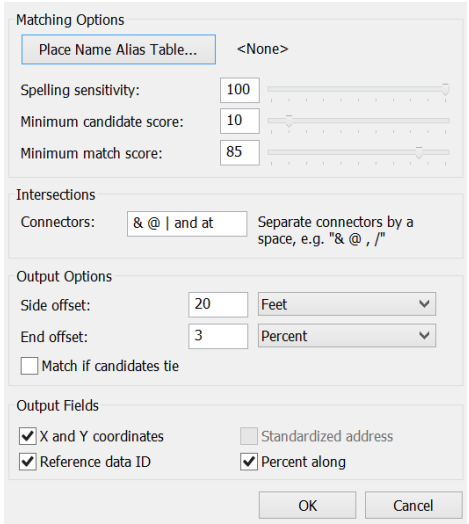



3. A dialog will appear allowing you to choose a geocoding service. A list will appear as well. If you don't see the locator you just created - Click Add - and browse to the Address Locator we just created called Cambridge_StreetLocator. Add it, highlight it in the dialog, and click OK to select it as our Address Locator service for the geocode.

4. The Geocode Address dialog will appear.

Street or Intersections = "Address"
Zip Code = "ZIP"

5. Change the Output file name to be 'cambridge_grocery_week5.shp' and set it to save to your weekly workspace, but don't click OK yet. We need to check the Geocoding Options.

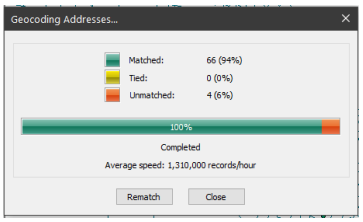



6. Click on the **Geocoding Options** button. Match your options to the screen to the left. Set the settings as follows:

- Set the **Spelling sensitivity** to 100.
- Set the **Minimum candidate score** to 10.
- Set the **Minimum match score** to 85.
- Select **Side offset** of **20 feet** (make sure to check the unit and change if necessary) to make sure that your points will fall inside polygons on their correct side of the street.
- Also, make sure that in the **Output Fields** section you specify X and Y Coordinates. This field will be useful after you have geocoded your address locations because it will allow you to automatically re-map addresses using the ADD X & Y Event function.

- Also, check **Percent Along and Reference Data ID (if not checked by default)**. This allows you to determine the percent along a given segment that an address falls.

7. Click OK. This will return you to the previous, main geocoding dialog. Click OK again and the geocoding process will start. This may take some time to run. The Address Locator is locating your addresses.

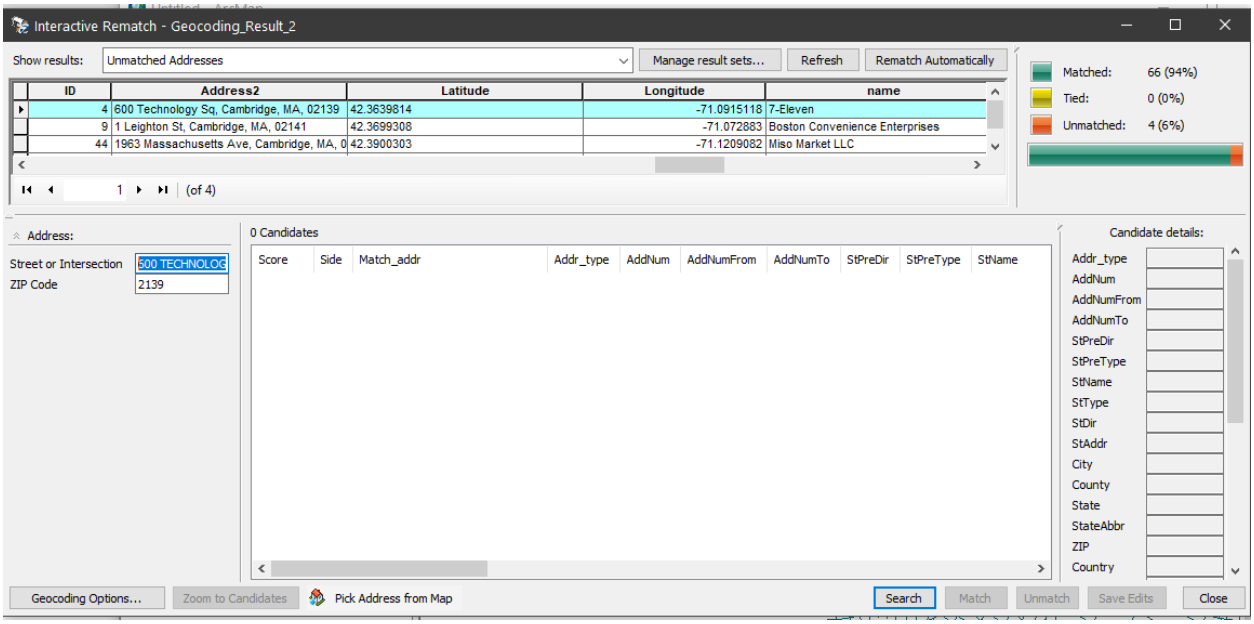



When complete, addresses should be tied to points. A dialog will appear that will show how many addresses were matched, how many were tied, and how many were not matched. Click **Rematch** to proceed and match all unmatched addresses.

8. We have four unmatched addresses that we need to locate on our map.

The **Interactive Rematch** window will open. Rematching is an iterative process of troubleshooting. The **Interactive Rematch window** looks like the one below. We'll be using it to manually rematch records that didn't have an exact match during the automatic geocoding process.

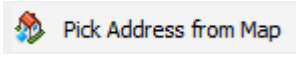
In the **Show results** drop-down menu, select **Unmatched Addresses** to only see the addresses with no match. Let's try to locate these addresses!




9. **Start with 600 Technology Square**. Click on the row for this record. (You'll have to scroll to the right on the list of unmatched addresses to find it.) The **Candidate Window** in the middle shows potential locations for our address point. Unfortunately, the candidate window is empty when we look at this one, meaning the Address Locator could not find any potential matches.

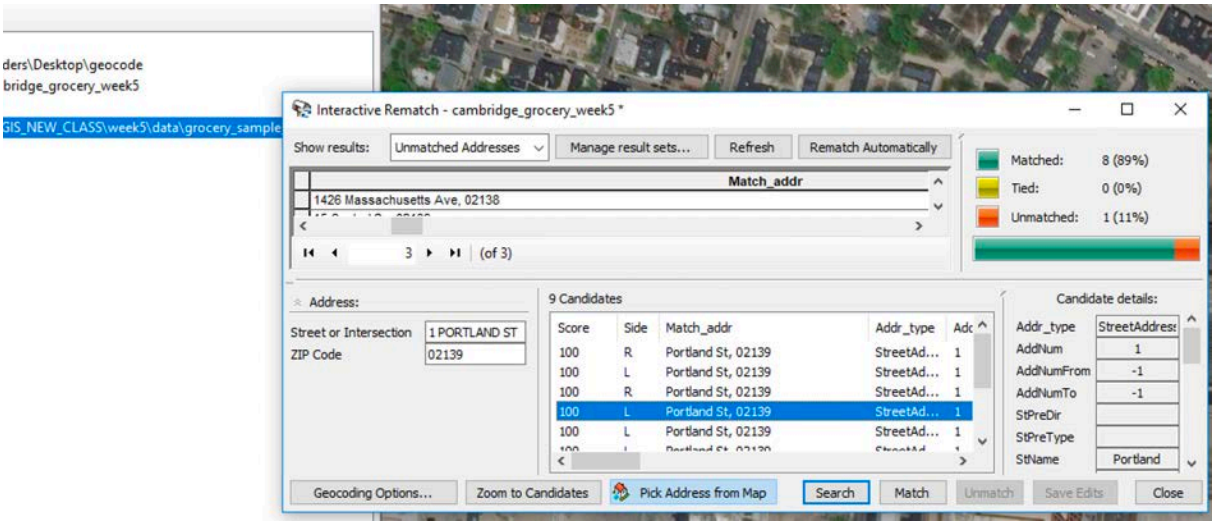
This is because while our locator performed well, certain reference files will miss things. In this circumstance, we can get a second opinion and use a secondary source. We'll use Google Maps here,

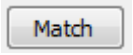
but you can also check Mapquest, Bing, or any other online mapping service. Go to Google Maps and search 7-Eleven, 600 Technology Square, Cambridge, MA. You should see the following.


10. We may have no candidates for this record, but we can locate where the address appears on Google Maps and add the point where we want it to go. Back in ArcMap, Click on 

in the Rematch dialog, and in ArcMap, click on the street in front of where the point should be and hold the mouse button down. A flash dialog will appear telling you the rough address. Mine said **1 Portland St.** It may be a little different depending on where you click the tool. This is a relatively common error to find while geocoding. Technology Square is the building complex, and our reference data represents streets. Because the numbers don't match, it might signify either a limitation or error in the reference data. We will match this manually.

Now that you know what the address should be, enter it in the Street or Intersection Dialog. See Below. After you enter in the **Street Intersection** part of the dialog and click 'Search', you will get a list of possible addresses on the right hand side of the screen.




You will see a number of candidates appear. Scan through and select the one that appears closest to the correct location. Highlight it and click the 

button in the lower right corner of the dialog. This will match that location with our address.

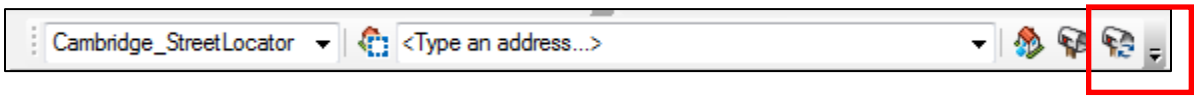
A point will appear on the map at 1 Main St location that represents an estimate of the location of 600 Technology Square. Click **Refresh** at the top of the Interactive Rematch dialog to remove this address from the unmatched address list.

Match your remaining unmatched records (if any) following any of these methods.

Then, click Close on the Interactive Rematch dialog, deselect your world imagery file, and zoom to the layer of geocoding results. Your map document window should look similar to the following.

11. Export your result to a new point shapefile that contains all the verified grocery stores and convenience stores in the City of Cambridge. You can do this by right-clicking on your 'Geocoding Result: Cambridge_grocery_week5' file in your TOC → Data → Export. Make sure to save the file as 'cambridge_grocery_geocoded_week5.shp' with Shapefile as the file type into your weekly workspace.

*Note: If at any point during the exercise you closed out of the **Interactive Rematch** dialog. You can get back to it by opening up the Geocoding Toolbar. You can find it by going to 'Customize >Toolbars' and selecting the 'Geocoding' box. The following tool bar should pop up on your window. You can then clickon the 'Review/ Rematch' icon to reopen the Interactive Match Dialog. You need to have the shape file you are using selected to do this.*



**Can you think of a reason that you would us the ArcGIS geocoder instead of the OpenCage? OpenCage has a limit of 2500 points a day. You can pay for something that has more than that, but if you have a huge dataset you might want to use ArcGIS and use OpenCage to find the files that ArcGIS couldn't find that means you have to find less files overall.**

# 11.205 Introduction to Spatial Analysis

Fall 2019