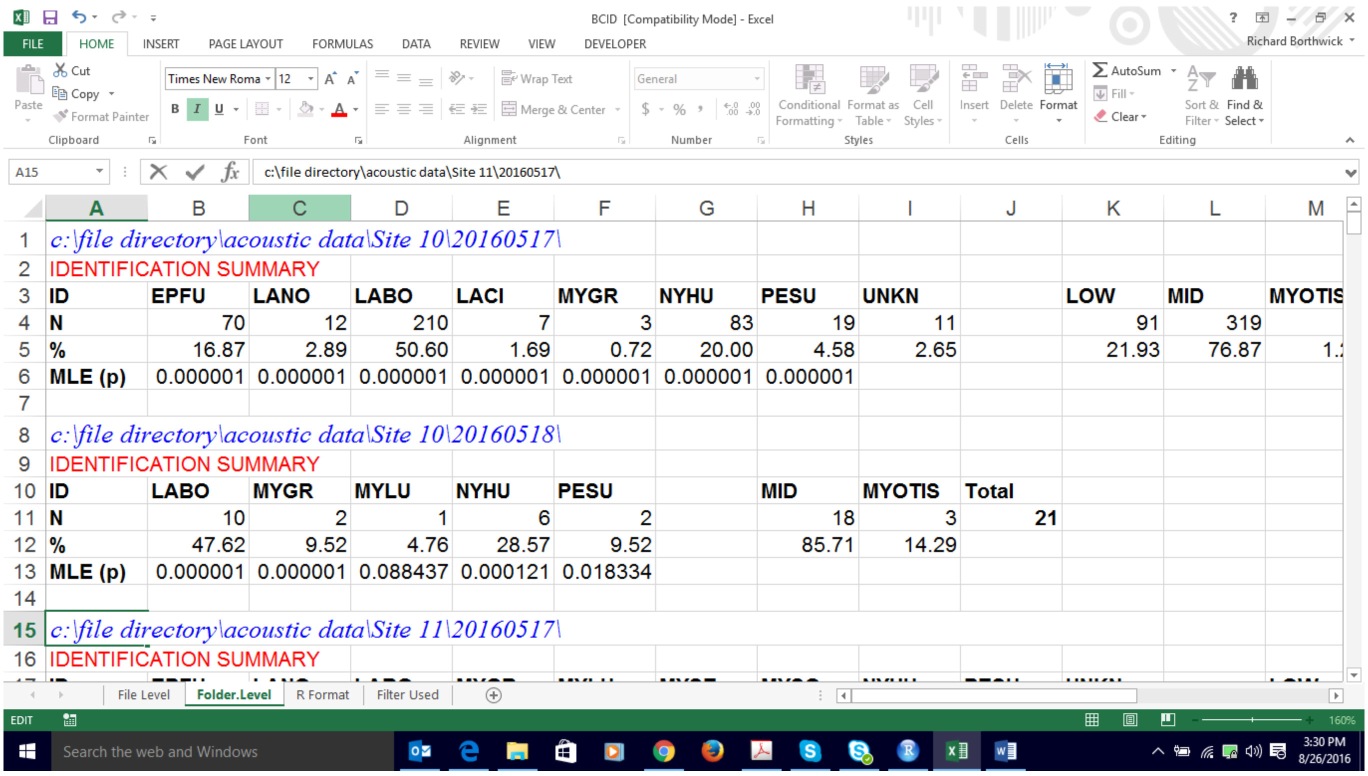
Mystery Data


BCID [Compatibility Mode] - Excel
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER
Richard Borthwick
File Level Folder.Level R Format Filter Used
EDIT
Search the web and Windows
3:30 PM 8/26/2016

A15 c:\file directory\acoustic data\Site 11\20160517\

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	c:\file directory\acoustic data\Site 10\20160517\												
2	IDENTIFICATION SUMMARY												
3	ID	EPFU	LANO	LABO	LACI	MYGR	NYHU	PESU	UNKN		LOW	MID	MYOTIS
4	N	70	12	210	7	3	83	19	11		91	319	
5	%	16.87	2.89	50.60	1.69	0.72	20.00	4.58	2.65		21.93	76.87	1.
6	MLE (p)	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001					
7													
8	c:\file directory\acoustic data\Site 10\20160518\												
9	IDENTIFICATION SUMMARY												
10	ID	LABO	MYGR	MYLU	NYHU	PESU		MID	MYOTIS	Total			
11	N	10	2	1	6	2		18	3	21			
12	%	47.62	9.52	4.76	28.57	9.52		85.71	14.29				
13	MLE (p)	0.000001	0.000001	0.088437	0.000121	0.018334							
14													
15	c:\file directory\acoustic data\Site 11\20160517\												
16	IDENTIFICATION SUMMARY												

Premise

When working with data, you'll commonly have to sift through stuff that you just don't understand. There will be columns that don't make any sense, rows that lack context, missing values... the list goes on.

The image above is lifted directly from a 2016 StackOverflow post, titled *Proprietary software produces ugly excel tables...can I get them into R?*, facing exactly this issue.

Gesticulating wildly towards their bad data, the author asks:

> "How do I take dozens of files from excel, that look like this, and get the import to add the rows, so that every set of four rows spaced by 2 rows (i.e. rows 3-6, 9-12, 15-18, etc--it's the same spacing for every project) are imported until empty space is reached (it'll be a different number of repetitions for each project)? I also want to `endCol` at empty space for each four-row segment..."

Colloquially, this process is known as *data wrangling*. Notice that not only does the author need this data to look better *on principle*, but also that the solution needs to be *dynamic*; that is, it will need to fit the parameters of their specific project on a case-by-case basis.

When you encounter this kind of data – and it's not a question of *if*, but *when* – what will you do? What strategies will you use to make sense of it, tidy it, and eventually, map it?

Activity (30 minutes)

Below, you'll find three mystery datasets.

Some of the datasets are clean, but in familiar formats; some are in familiar formats, but have been seriously gunked up. Some of the data might appear in a format you don't recognize at all.

To begin, download each dataset with `Right-click` ➡️ `Save as...`, and open it in the appropriate application. Using the "GIS friendly" principles that we discussed during lecture, group up with your peers – groups of 2-4 are fine – and answer each of the following questions. You don't need to submit this, but you'll probably need to write stuff down in a Word or Google Doc.

> **HINT:** if you aren't sure how to view one of the files, try opening the file in Visual Studio Code. It's a text editor that should be installed on all the Data Lab computers.

1. **Open all the files and determine whether the data is GIS friendly.** If not, make a list of problems that you notice. Then fix each one.

   I recommend using Microsoft Excel, and promise that the data can be wrangled using just two functions: `Find and replace` and `Text to columns`. Work together (and with the internet) to figure out how to locate and use these functions in Excel.

2. **What is the data trying to tell you?** Examine the field names and the information they depict.

3. **What kinds of fields are common across all the datasets?** Write down an example from each dataset. Why are these fields important? Note that the field might not have the same name in each dataset.

4. **Does the data contain information about where it came from?** As discussed in class, this "data about data" is simply called *metadata*. What metadata columns, if any, can you find?

5. **Does the data contain geographic information?** Again, if so, which fields contain it? Do you think this geographic information can be mapped? Why or why not?

Once you feel you have satisfactory answers to these questions for each dataset – or you have at least made an honest attepmt at discovering them – see if you can figure out what the data is. Can you snoop out each one?

GET DATA A

GET DATA B

GET DATA C