**Wrestling with data:**

As you saw in your readings, there are many potential pitfalls you can run into when you’re working with data. You should always assume that you data is “dirty” or has flaws. But that doesn’t mean you can’t use it. You might just have limitations you have to work around, or you might have to clean it up.

You might also just have a format problem – that the data is not structured in quite the way you need to work with it. We’re going to look at some examples of that.

Once you have the format set, you might also need to:

1. Understand your data and get answers to whatever questions you encounter.
2. Physically alter the data to repair inconsistent names, problematic dates, etc
3. Add something to your data to make it more valuable

If you look at the [“Notes about DOD database”](https://gist.github.com/chrislkeller/9c89207cfcd26982893a) that I linked on Moodle, you’ll see a paper trail of problems reporters encountered when working with the “1033 program” data showing surplus defense equipment acquired by local law enforcement agencies. You can see that the reporters had to learn about what the data contained – tactical equipment, versus controlled equipment, for example .

Under the “data smells” section, you can see the flaws they found in the data – i.e. the data doesn’t indicate the condition of the equipment (is it fully functioning?) , the names of the recipient agencies are inconsistent, etc.

Let’s look at a common example of “dirty data” – FEC data showing contributions made to U.S. Senate campaigns (open Excel file – “wrestling with data\_practice”)

**Dirty Data:**

--variation in name spellings & occupation (even city field is problematic in remainder of dataset)

--“chardate” – this is how the date field originally came through

--“charamount” – FEC uses COBOL to store these amounts; there is a conversion you need to go through to convert these to the correct dollar amounts; negatives are particularly problematic

Later this semester we’re going to learn how to use a tool called Open Refine that makes it easy to fix the name inconsistencies. It’s also possible to do it in Excel or Access, although Open Refine is much quicker and simpler.

You might also need to add some value to your data. There are many times that the data is right for your analysis, but you can’t get the answer you need without making some changes. Typically this requires adding something.

Thinking about what your data DOESN’T have – but needs – is a crucial part of doing data analysis. Sometimes you can get a better story that way. For example, [this story by ProPublica](http://www.propublica.org/article/vying-for-market-share-companies-heavily-promote-me-too-drugs) analyzed how much money drug companies were paying doctors for consulting, giving speeches, etc. In order to find a pattern in the data, they had to look beyond the data. The data merely showed how much money each pharma company spent related to a particular drug. It’s easy enough to add in some value to the data by adding a field that indicates what the drug is primarily used for (i.e. diabetes, high blood pressure, etc). ProPublica went one step further and added a field that indicated whether the drugs were big medical breakthroughs or cures --- the analysis found that companies were mostly spending money on drugs that were not.

Here are a few common examples:

Again…use the Excel file “wrestling with data\_practice”

**Dates:**  
--getting the year, month or day out of a date

**Split names:**

--using string functions to split apart names, addresses, etc

**Lookups & lookups2**

--data doesn’t have the name of the county (only the FIPS code); Can use VLOOKUP to add that information

**Classify**

--this is crime data; I’d like to know which shift this call came in on – day, evening or night – so I want to add a field that indicates ; also can use VLOOKUP to do that; this adds value to your data

Now let’s go to some common problems you might encounter when you first get data. Sometimes, the format or structure of the data simply doesn’t work, for various reasons.

**CensusData:**

--this is a download from AmericanFactfinder. It’s supposed to be a format that is “ready” to go into database software. But it’s not. The two header rows will screw you up quickly. Other than that, though, this is a pretty good file.

--Solution: set up just one header row. Ideally, this would be better than either of the ones here. I prefer database-friendly column headers like “pop\_2010” or “geoID”

**CensusData\_PresentationReady:**

--that same data as before, but this time I downloaded the “presentation ready” Excel format. A bunch of problems here: the headers encompass more than one row; there are hidden columns; there are merged cells (A&B columns); and the data is in text format.

--Solution: Unmerge A & B columns; Get rid of blank hidden columns; Create a single header row; use the value() function to convert all the data to numeric format.

**Transpose:**

--This is a lovely little feature in Excel that can come in handy. Open the “Excel Magic” workbook and go to the “Reshaper1” sheet. Copy the data and put your cursor in an empty part of the worksheet, right-mouse click and choose “paste special” and then click the box that says “transpose”

**Tableau Reshaper:**

--this is an “add-in” for Excel that is really useful for reorganizing data, even if you never intend to use Tableau. This tool “normalizes” data so that there is only one value per line -- that’s how Tableau prefers data. And you’ll likely find that other data apps you might build may want this format as well. Use “reshaper2” worksheet

<http://kb.tableausoftware.com/articles/knowledgebase/addin-reshaping-data-excel>

**Graduation rates**

--this shows graduation rates (and the data used to compile those rates) for each school district in Minnesota, across several years. This is a tough spreadsheet to even look at and it’s impossible to do any analysis on it in this format. Problem: three rows for each district, but district name is on a separate row. (there are also data notes at the bottom)

--Solution: start by applying the appropriate district name to each line of data (using IF statement). Copy-Paste Special.  filter on column A for rows that start with “graduation”. Select-All and go to visible cells only. Paste into new worksheet. (repeat for the other rows, if you want). Put header rows back on new table.

**Firearm discharge:**

--these are police firearms discharge reports. You’ll see there are multiple lines for each incident, but really there is just one line of data that we’d likely want. This also has a header problem, so it’s not possible to even filter this spreadsheet.

--Solution: Fix the header problem. Turn on filters and choose only rows where there is a value in the first column (called “count”). Then do select-all and go to “visible cells only”. Get error message that says “can’t do that with a merged cell.” Go to Format Cells and turn off the merged cells option. Repeat select-all, go to visible cells, copy and paste into a new file. Save the file as older version of Excel to get rid of hyperlinks.

**Crime data:**

--this is crime data from the Minnesota Bureau of Criminal Apprehension. It shows the number of offenses, number of those offenses that were cleared and the clearance percentage, plus a crime rate for each police jurisdiction, by each crime type.

--solution: create a unique ID column to number each row. Use IF statement to pull out the jurisdiction name (i.e. “Aitkin Sheriff”) so that it’s on the same row as offenses; Use If statement to pull the population into a new column on the same row. Then filter to get just the offense rows. Since you have the population, you can then calculate the crime rates on your own (using the BCA’s numbers in original sheet to check your work)

**Key things to keep in mind when creating header rows in Excel:**

--keep it on one row.

--put it directly above your data (no blank space between)

--leave a blank row above it (or have it be the top row of your spreadsheet)

--make the labels as short as possible (especially if you want to use PivotTables or export this data to other software)

--make sure you don’t have extraneous info at the bottom

--use worksheets to keep data chunks separate

--goal is to get one record for each thing you’re analyzing

**Some Excel things you should know about:**

--Merged cells (get rid of them!)

--How to hide/unhide columns

--Wrapped text

--How to delete entire rows or columns

--copy paste special

--Transpose

--go to special (visible cells only)

**Getting data out of PDFs**

Some government agencies will play nasty and only give you data in a PDF. If you get one where the data you want is in there – and it looks like it’s set up in a table format – there are ways to get the data out.

See links on Moodle for articles on this topic and links to particular tools (most of them online; there are some that are software you load onto your machine)