

Sample Questions for the [Drinking Water Dataset](#)

Table of contents (linked to questions below):

[What is the average arsenic of filtered samples vs. non-filtered?](#)

[How much arsenic is in unfiltered water samples \(by school\)? \(3 ways to graph it:\)](#)

[What proportion of the data are from each type of well?](#)

[What are the arsenic levels within each well type?](#)

[Of the total concentration of arsenic found in all wells, what proportion of it is accounted for by drilled, dug or driven \(etc\) wells?](#)

[Are some filters more effective than others at filtering out arsenic?](#)

[What is the prevalence of arsenic in drinking water in participating towns in Maine?](#)

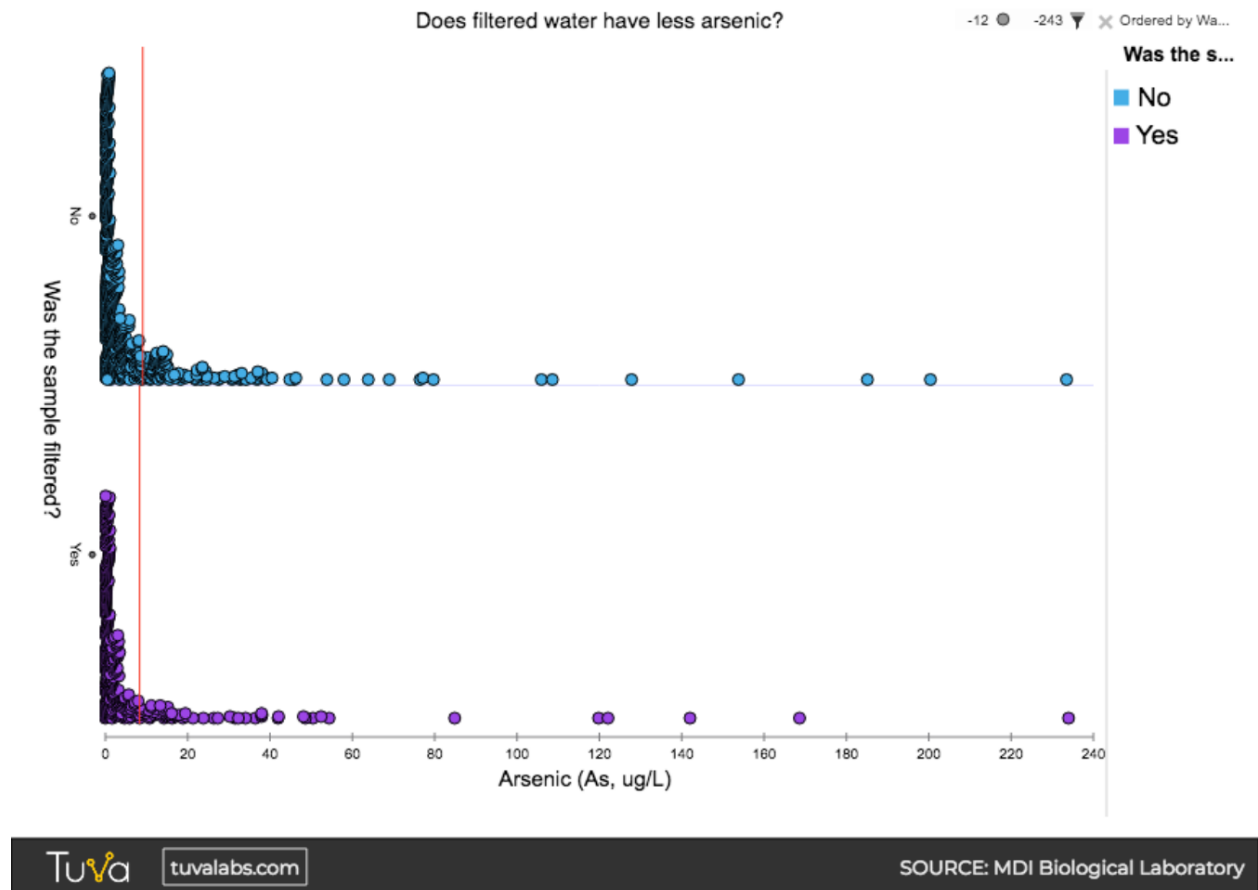
[Where are the study areas for the All About Arsenic Project?](#)

[What proportion of unfiltered samples from New Hampshire were above the 10ug/L threshold?](#)

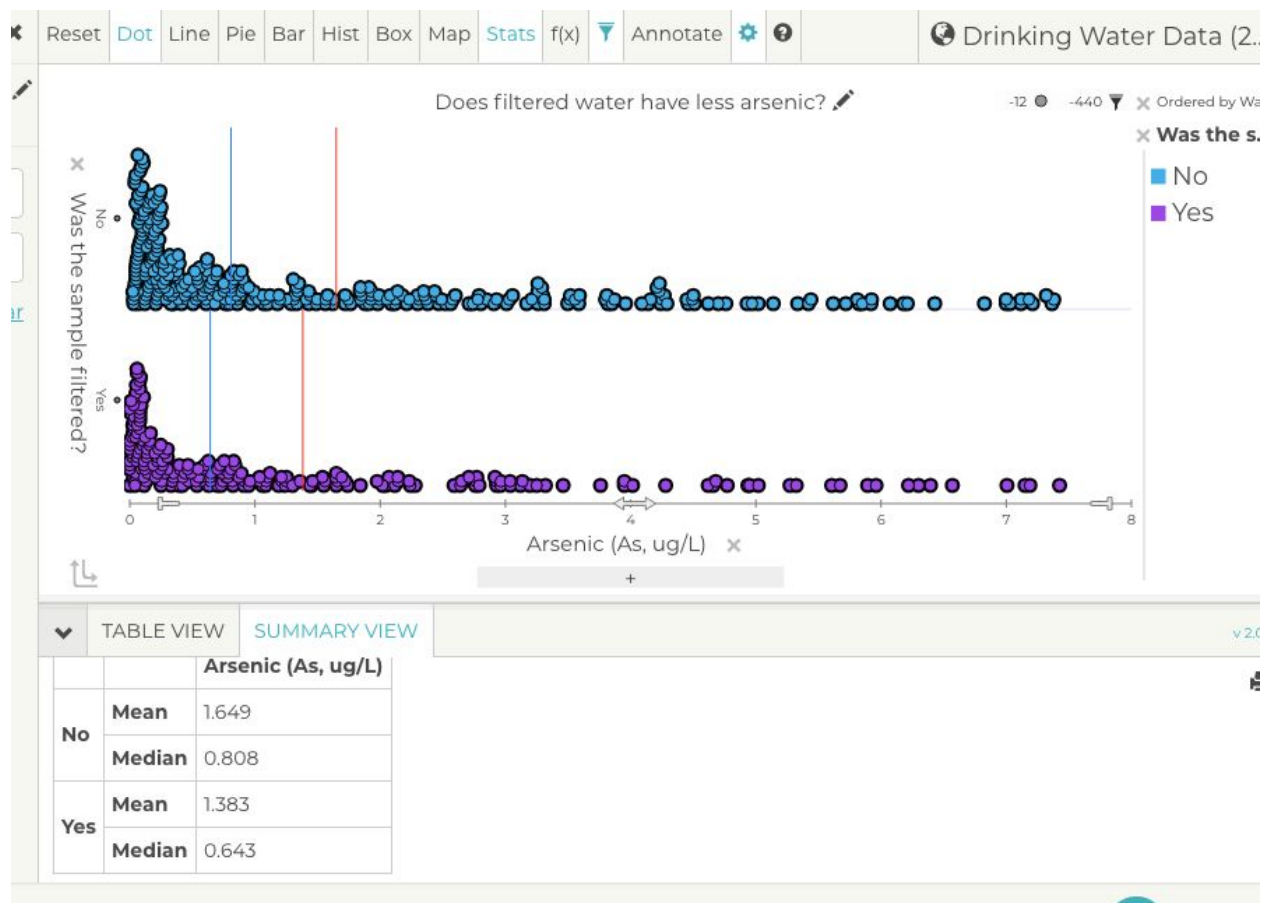
Is there evidence of any drinking water problems in our community?

What is the average arsenic in filtered samples vs. non-filtered?

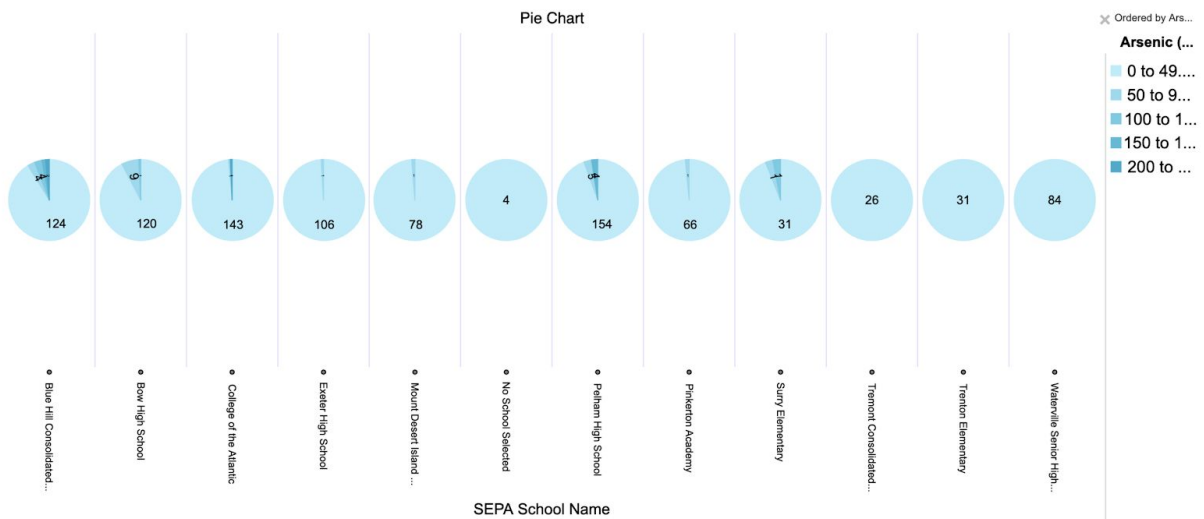
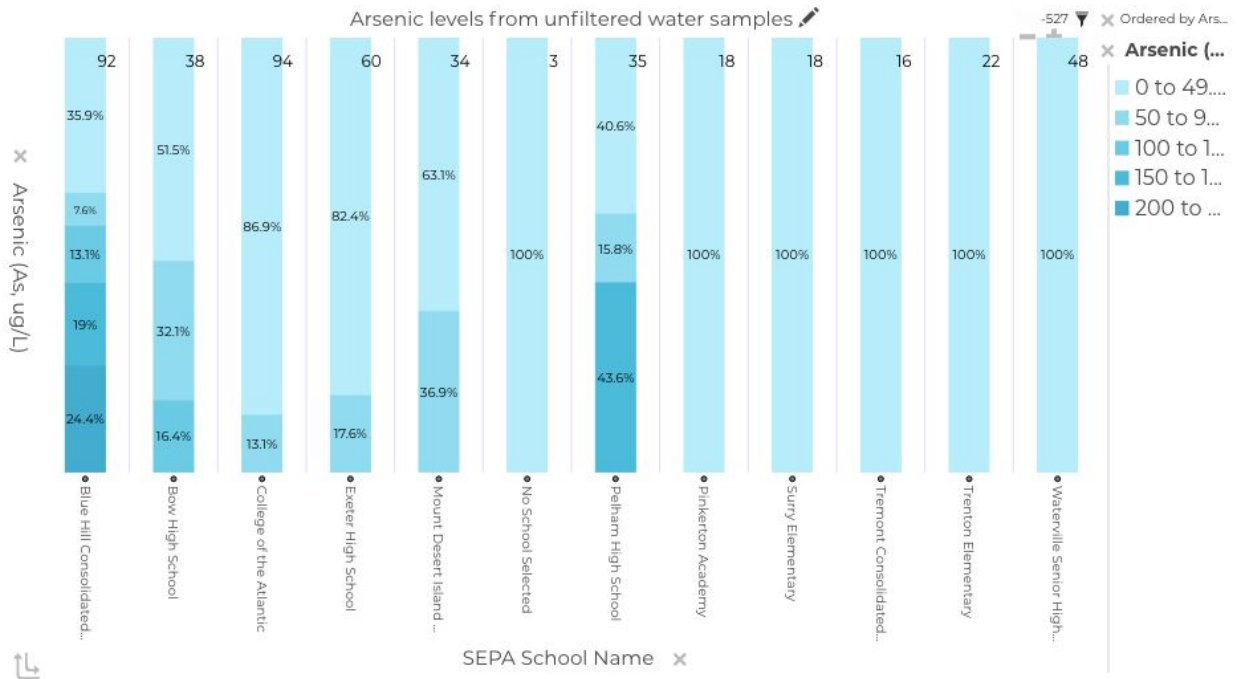
Alternate wording: Does filtered water have less arsenic than non-filtered water does? (Does filtering make a difference?) (For HS: is the difference significant?)

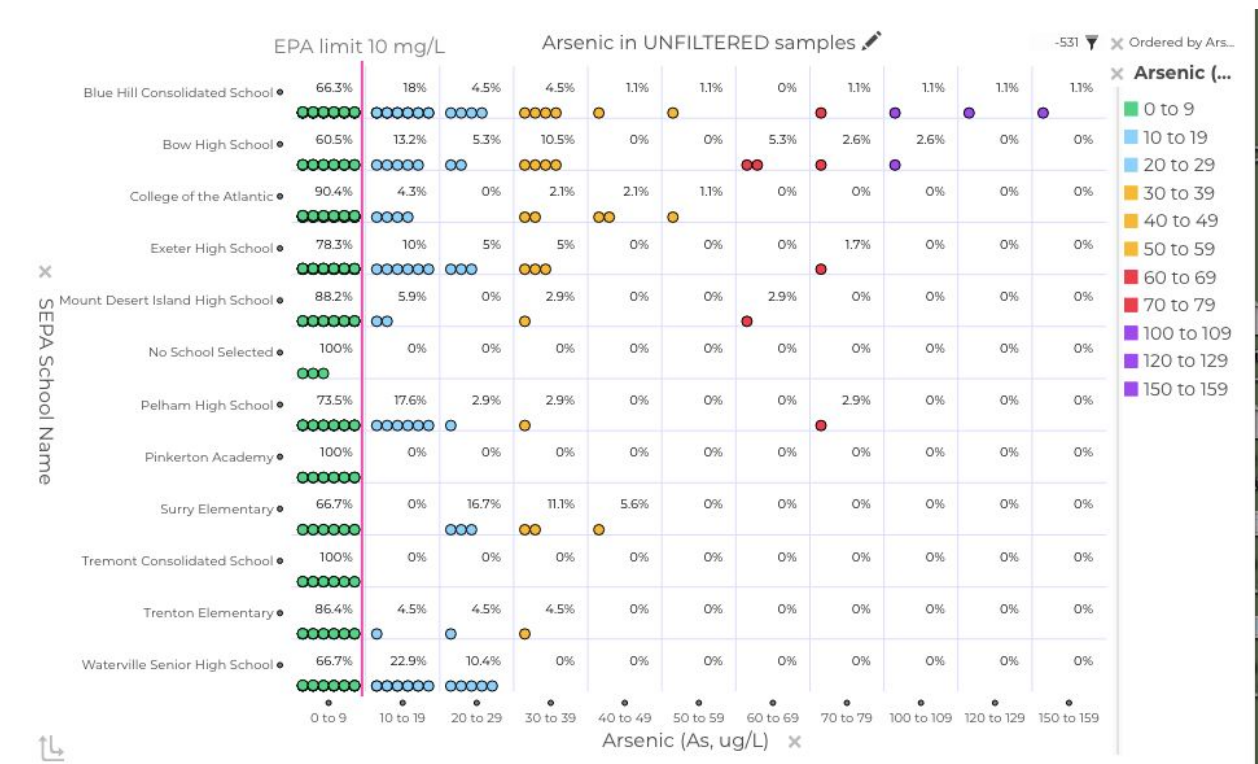


It's hard to compare the **means** because the distribution is so strongly skewed. Extreme values pull the average in one direction, so it is less representative than the median, thus the **median** would be better to use for the center: (Mean = red line, median = blue line. The next image is filtered to zoom in to just the lowest levels (8 ug/L or less).

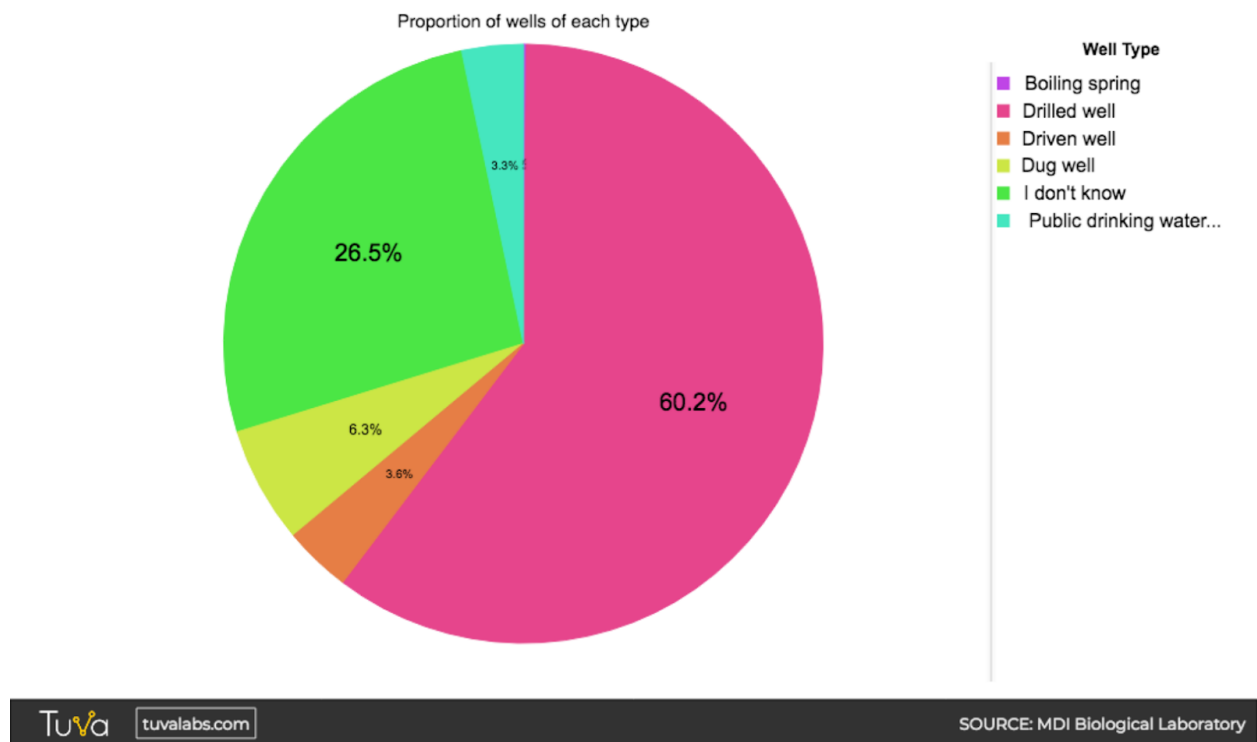


How much arsenic is in unfiltered water samples (by school)? (3 ways to graph it:)



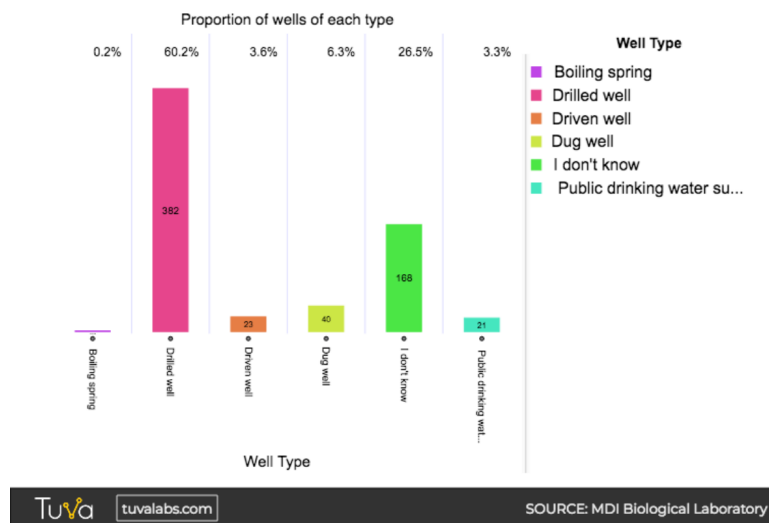
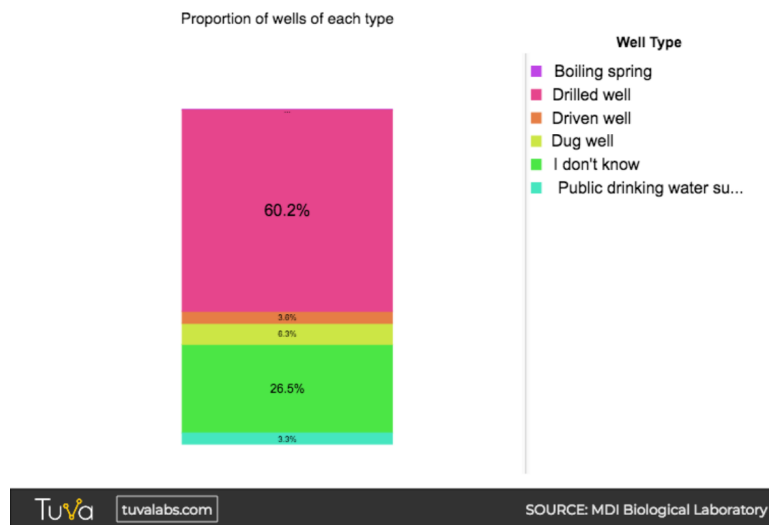


What proportion of the data are from each type of well?



Note: for this question, Count can be left checked, if desired -- it just gives the number of wells in each group, in addition to %, which is meaningful (unlike summary values for arsenic concentrations). Here are three ways to graph this question:

- Pie chart (Above; Click Stats/Percent by Pie Section) (and uncheck Count, if desired)
- Stacked bar chart (Click Stats/Percent by Bar Section)
- Stacked bar chart with Well type on X (Click Stats/Percent by Category)

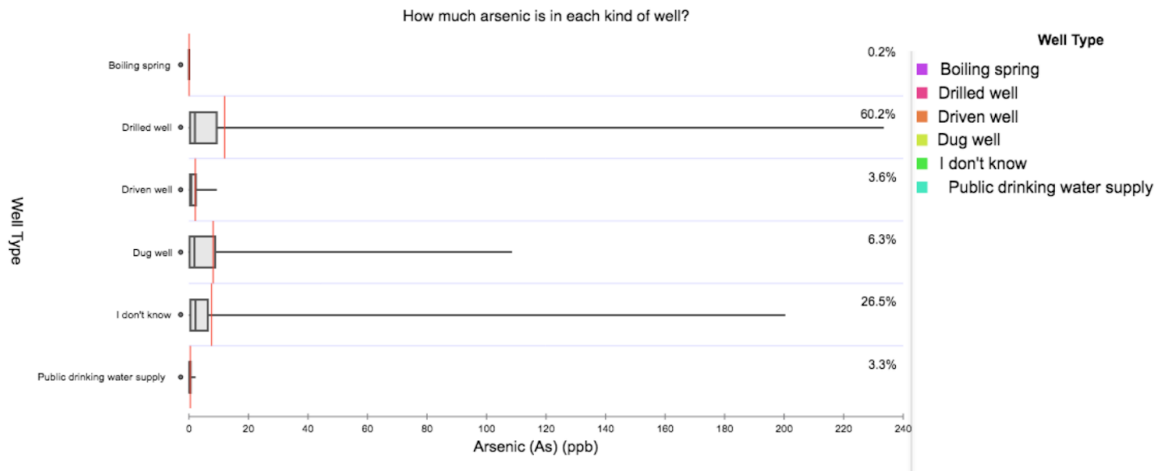
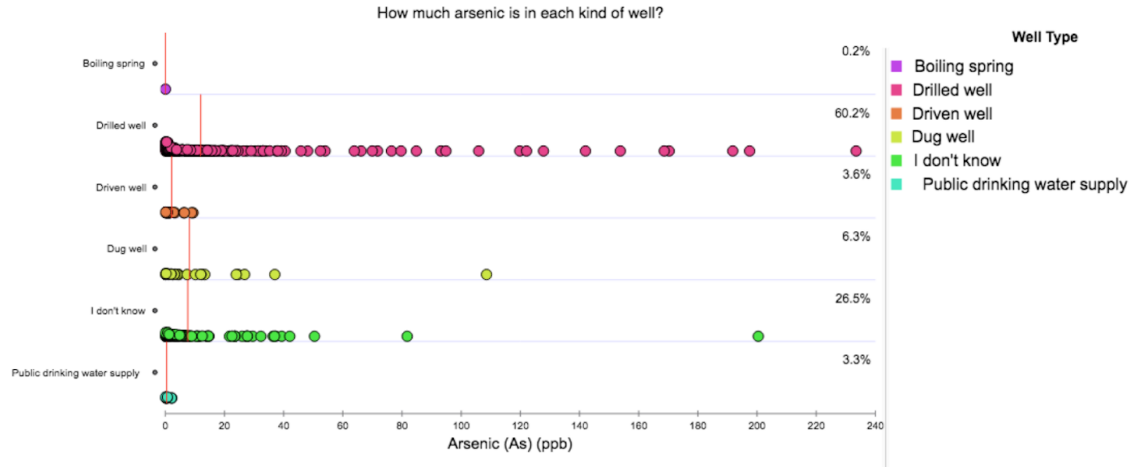


“...and what the arsenic levels were within that type...”

What are the arsenic levels within each well type?

This is really a separate question from the first one, as it is about comparing the well types (a categorical attribute) in the amounts of arsenic they have (a quantitative measure). To compare, you could compute the average As concentration for each group (but then you don't see how variable each well type is), or you could compare the distributions (which allows you to see the mode(s), and the mean, and how variable they are).

- Put Arsenic on X and Well Type on Y.
- Click Dot (or Box, or Histogram).



"The program was creating a SUM of the total arsenic found in all the wells of that type. Can we change this to a Mean, and maybe mode?"

I think students might be trying to do too much here. Pie charts are good for showing proportions of a sum total, but they can't show a distribution -- they are by nature summarizing each section as a count or a percent. The averages or modes for each well type wouldn't relate

to a meaningful total (ie. the total of all the averages for each pie section? The averages would not necessarily add up to 100.

This could be a good opportunity to ask students to sketch what we called in the Data Literacy Project a “hypothetical graph” -- what sort of output are they picturing? Sketching a “hypothetical graph” forces students to think through what they want their evidence to look like. (Just sketch it loosely, don’t worry about scale, or accuracy, just label axes and roughly sketch how you think it should look.)

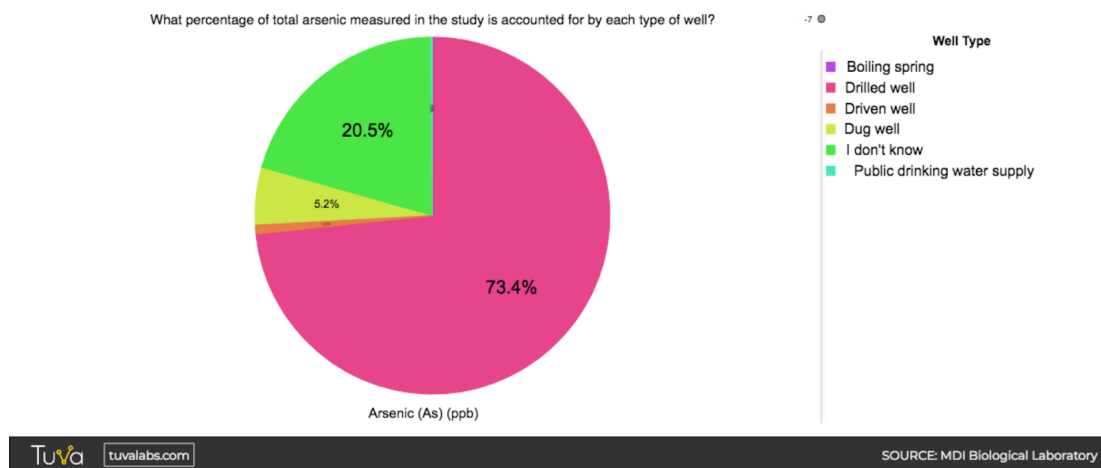
It would be way more useful information than the total sum of all those wells?”

Yes! To show percents instead of counts (totals), click on Stats in the top menu, click Count, and uncheck Sum by Pie Section. Then click on Percent, and Percent by Pie section.

Here’s one more question you might have been trying to answer:

Of the total concentration of arsenic found in all wells, what proportion of it is accounted for by drilled, dug or driven (etc) wells?

- Reset the graph to start over.
- Click once on Well-type
- Click Pie (top menu)
- Drag Arsenic to X
- Click Stats, then uncheck Sum by Pie Section and check Percent by Pie section



You could say that 73.4% of the total arsenic concentrations measured in the study so far came from drilled wells. It’s an odd way of putting it because as you pointed out, totalling concentrations is somewhat arbitrary, and it could be higher just because there are more drilled wells. (See the last question below as a way to test that hypothesis).

The box plots suggest that while drilled wells have many wells with higher arsenic, and the range is highly variable, they are **typically** not that different from others (because their boxes -- interquartile ranges, or the middle 50% of the data points -- are somewhat aligned).

A key take-home here is for students to clarify the question they are trying to answer. I like to see it written out. It helps to analyze for one question at a time.

Finally, here's another question your students might have been trying to ask:

Do drilled wells have a higher *proportion* of wells with high arsenic (ie. >10ug/L) than other types of wells do?

The question is about proportions, so a pie or stacked bar is a good choice.

The question is about wells with arsenic either above the EPA criterion, or below, so we need to change the Arsenic attribute from Numeric to Categorical with Numeric Intervals

- Click the icon beside Arsenic
- Under Type, select Categorical with Numerical Intervals
- Change the Interval size to 10

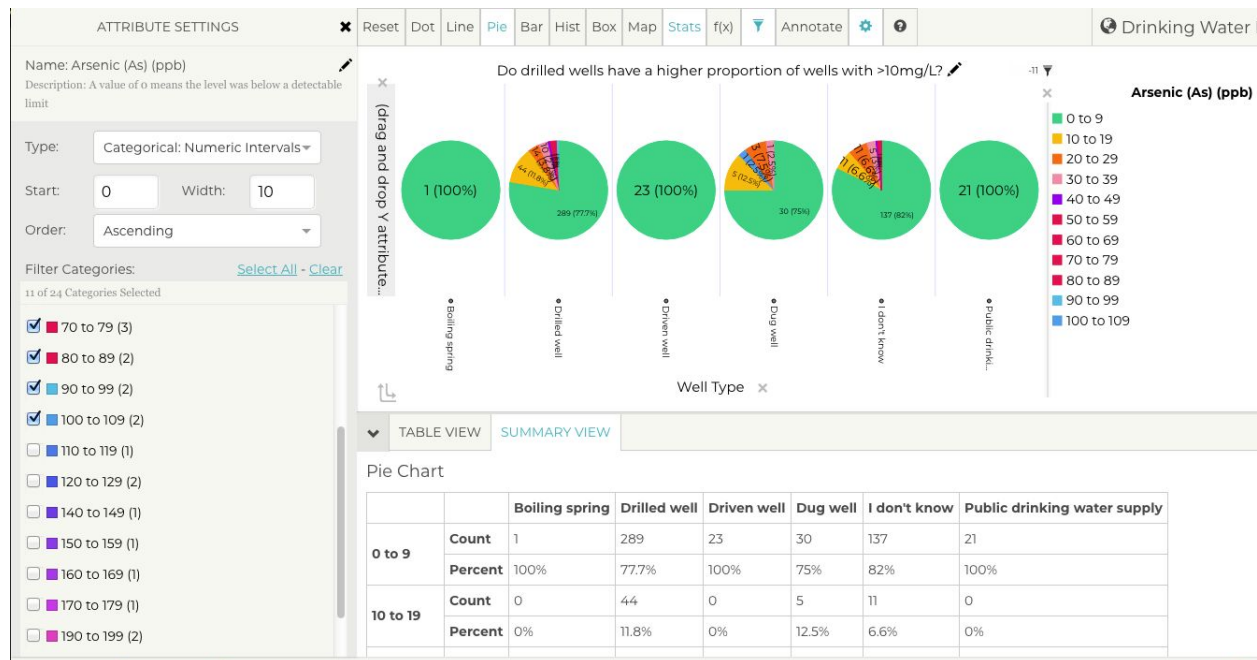
(Optional: To visually distinguish between the <10 interval and the interval groups >10, you can change the colors for each group so <10 is one color (ie. green, to mean ok), and make all of the others another color (ie. red, to mean over the safe threshold, or orange and red if you want to separate out those that are just over 10). (Or you can keep the groups each a different color).

- With the Arsenic attribute settings still open, click on the colored boxes beside each interval group and select the desired color.)

Close the Attribute Settings (two clicks to get back to the attribute list) and make the pie.

- Put Well Type on X (so we get a pie for each well type)
- Click once on Arsenic, so it knows to show the proportion of wells in each As interval group
- Click Pie (Or Stacked Bar, either one works well)
- Click Stats / then Percent / Percent by Bar Section
- (You can either keep or remove the Counts -- that gives the number of wells in each interval).
- To see the actual numbers for the thinner sections, hover over the section with the cursor, or (better) open up the Summary view at the bottom, and you can read the values for each section for each pie.

What conclusions can be drawn from the pies to answer the question? (Screenshot below).



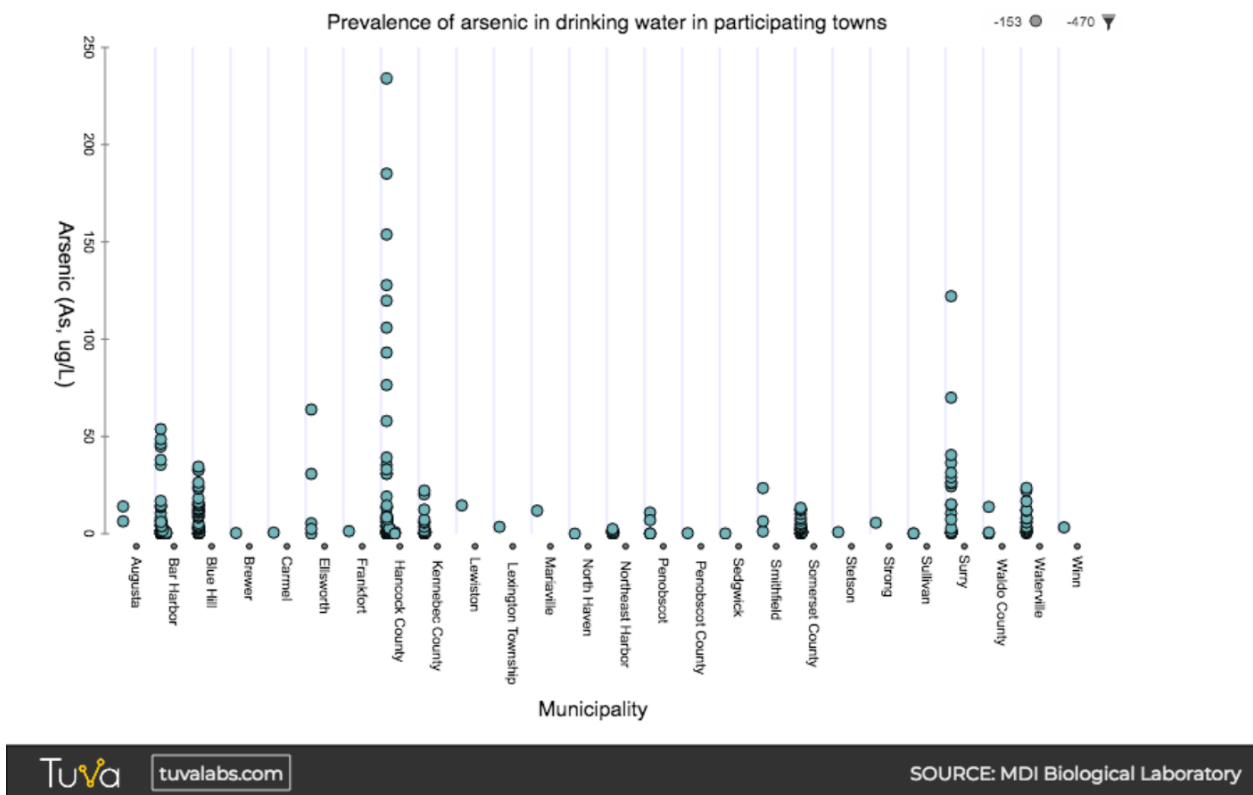
Are some filters more effective than others at filtering out arsenic?

Please note: On the sample registration, we asked people to report on their filtration system EVEN IF they bypassed it for the purposes of this test. Many people report that they do have a filter, but sent in a raw water sample. Without removing data from samples that bypassed the filtration system, the graph will show an artificially high number of records where it appears the filtration system does not remove arsenic to a satisfactory level.

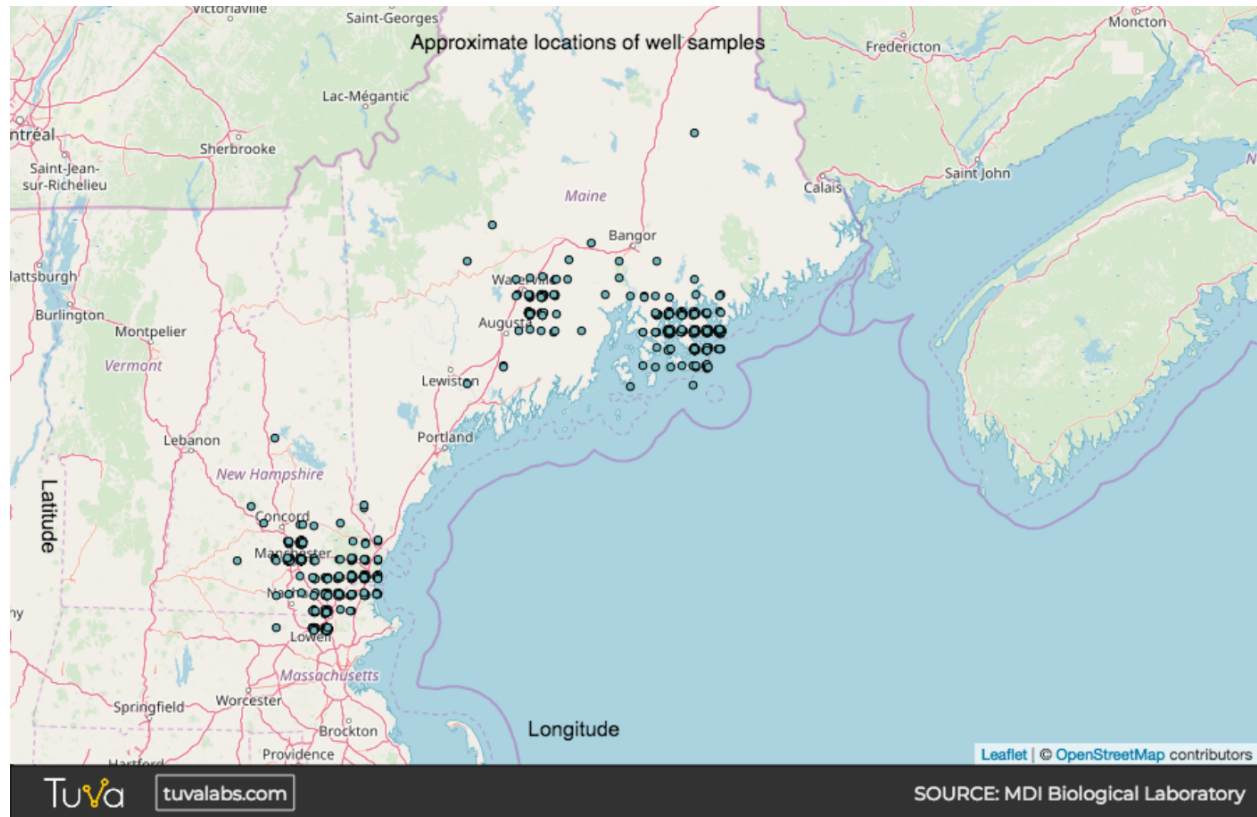
-73 ● -709 ▼



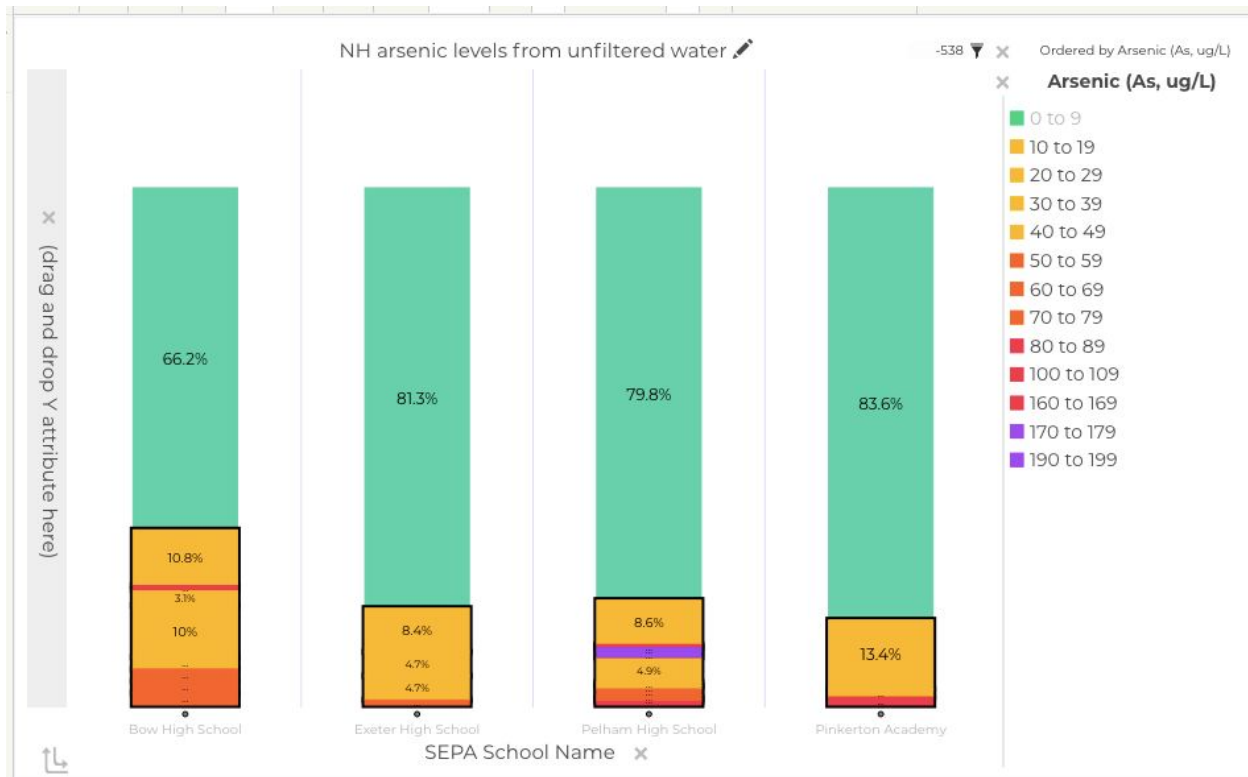
What is the prevalence of arsenic in drinking water in participating towns in Maine?



Where are the study areas for the All About Arsenic Project?



What proportion of unfiltered samples from New Hampshire were above the 10ug/L threshold?



To make the plot:

- Filter out all but the four NH schools (Click the Attribute settings Icon beside School, and uncheck the ones you don't want.)
- Put School Name on X
- Click Bar / Stretched Bar chart. (You could also pick stacked bar -- the heights of the bars will reflect the number of samples from each SEPA School.
- Change Arsenic from a 'Numeric' attribute to 'Categorical with Numeric Intervals' (Click the attribute settings icon beside Arsenic, then change Type to 'Categorical with Numeric Intervals')
- Change the interval size to 10, so the 0-10 safe group is shown separately.
- Adjust the colors of all the groups, ie. green for the 0-10 group, yellow for the next 3 or 4 intervals, orange for the next intervals, etc.
- To put the darker box around the over 10 intervals as a group, click on each interval in the legend that is over 10 while holding down the shift key.
- Under Stats, click Percent by Bar section. (You can uncheck count, or leave it if you want to see how many wells fall into each group.
- Add a title if desired