4 Real world data fluency: How to use raw data

Wendy Stephens

The implementation of the Common Core State Standards warmed the hearts of many librarians in demanding evidence as part of argumentation. Many high schoolers (as well as their teachers) sought to integrate statistical evidence – hard numbers – in their products. But too often those percentages lose context in student work, as they neglect to mention the methods of sampling or specify the populations under study.

A command of datasets will enable teachers to help students understand where data is coming from and how it is presented, enabling more sophisticated inferences about the potential value of that information. By looking specifically at datasets and examining a range of tools for reading, integrating, representing and thinking critically about data, this very transferable skill will help students as they go forward, both academically and in the workplace.

As we peer over students' shoulders to help them make sense of what they find and what they create, we recognize that we need to nurture students' ability to do just-in-time analysis with freshly-encountered data. Whether students are using a table, a graph or a chart to help make determinations and evaluations, there are some general rules of thumb:

- **To use data, you should be able to articulate how data was generated or collected.** Who generated the data? Was it self-reported? Measured in a laboratory setting?
- **Why was the dataset generated?** This includes context around who or what organization is funding a study. Some

» associations are obvious, while others demand more research, but it's important to question research agendas.

» **Look at origins of data and what those relationships are between the variables.**

» **Use the Goldilocks principle to look for the just-right dataset** — one that's not too small or not too large.

» **Did you find a provocative statistic in your research?** Follow breadcrumbs back to the source for the most thorough understanding of the data point.

» **Always look for keys, practice axis awareness, and watch the units.**

» **Watch for language shifting and unit shifting,** matching things with the bigger picture.

» **Understand external factors when considering change over time.** Society and language changes as well, populations are collapsed or differentiated, making comparisons over time difficult.

I think of real world data fluency as a facility in thinking about numbers and values — what we can learn from different amounts of things and what those relationships suggest. To foster that sort of thinking, I suggest a collection of tools and associated strategies for integrating reading, representing, and thinking critically about different data sources. Those sorts of understandings will help our students become information-literate consumers and citizens that push past headlines.

Data fluency has much in common with data literacy, but like media literacy, it seems to include more transformative elements. Students should gain experience traveling backward from headlines, soundbites, and news accounts to find data points the headlines are based upon. We want them to become comfortable understanding where data comes from and the underlying relationships between measurements. This includes

how the study is diffused in the mass media and the way news accounts may or may not acknowledge the particulars of the investigation. Sometimes attribution may be simply an institutional name, but tracing back to particular research study is important.

To get started working with data in your class or library:

- **Provide students with examples of raw data with variables of interest.**
- **Discuss whether the parameters were made explicit or not.**
- **When students repurpose the data to create new products, stress attribution, citation, and responsible use of data.**
- **Practice extracting relevant data from charts, tables, and graphs. This facility is most germane in high stakes testing, too.**
- **Students might find messy data on their own, so teachers and librarians can help to structure assignments directing students to particular datasets.**

Young people who have mastered data fluency skills can make use of the myriad statistical information available today to make better informed decisions as a consumer, citizens, and social activists.

One way to start the conversation is to discuss the range of information-gathering strategies. Students should be able to identify opt-in participation like web polling or more rigorous population sampling, and discuss the way participation might affect results.

Thinking computationally

For teachers who aren't comfortable talking statistics, it is important that librarians model that teaching with data can be a cross-curricular proposition. Computational datasets and metrics exist in the humanities and across the subject areas like never before. Using datasets to explore bibliographic information or to map trends over time can be described as cultivating a computational mindset in ways that are not strictly arithmetic. There are a range of resources to help teachers in all disciplines talk about relationships in this way.

Tyler Vigen's Spurious Correlations

http://tylervigen.com

This fun, high-impact website allows for menu-based manipulation to explore sometimes humorous relationships between unrelated datasets.

The selection of the first dataset from the drop-down menu (for example, commercial space launches) produces a second menu of datasets, ranked in order of strength of correlation. Including the original numbers underpinning the relationship, the near-perfect arcs mapping honey produced by bee colonies in the U.S., and three years of visitor volume to Sea World in Florida (with a 0.9948 strength) is typical of the interesting but unrelated parallels.

With access points including morbidity and consumption of a range of dairy products, Vigen's site does drive home how two measurements can be correlated, without any representation or suggestion of a causal link between the two.

Social Explorer

http://www.socialexplorer.com

Getting hyperlocal or providing a national comparison are two potent ways to lead students to explore relationships. Social Explorer is a geographical tool including many paid datasets, but a wealth of census information is available in the free online version.

Social Explorer makes it easy to present side-by-side comparison for examining change over time or contrasting geographical distribution of different types of data. The interface seems particularly suited for historical comparisons over time – vividly illustrating population shifts brought on by Westward Expansion, industrialization, and the Great Migration.

Hovering over visualizations provides access to the original data points, and the authentic datasets are explicitly described. Intuitive to manipulate in terms of both data and mapping, Social Explorer is an excellent way for students to build fluency.

Google Ngram

https://books.google.com/ngrams

Google Ngram reflects the preoccupations of English-speaking society as they appear in print, searching the scanned volumes from Google's groundbreaking book digitization project. The default results will span 1800 – 2000, and that built-in range is important to know, especially when researching current events.

Ngram allows you to search for a single word or a series of comma-separated terms across decades of published literature. Searching [socialist, terrorist, anarchist], for example, reveals evolving trends in public consciousness as well as shifts in vocabulary usage.

Ngram is a wonderful tool for thinking in a data-centric way, without necessarily operating in a number-centric content area. It is particularly useful in the humanities to show a shift in language over time. The site attempts to index all words that have appeared in published books. The y-axis shows the percentage of times the word appears based on overall documents from a given year, rather than computing the number of mere occurrences.

Tableau Gallery

https://public.tableau.com/s/gallery

Tableau is an embeddable graphic visualization product that is increasingly being used by news organizations to develop interactive graphics for online editions. Users can hover over a visualization to view the data used in creating the visual.

Tableau Public is the free online portal to basic visualization tools; for more sophisticated usage, you may wish to download a paid version of Tableau software, which may require that you have administrative rights for your school computer. But there are interesting instructional uses for Tableau even outside construction of your own visualization. On the Gallery tab, Visualization of the Day provides an often topical, sometimes innovative, depiction to spark conversations about data relationships with geographic aspects. A favorite of reporters, the Gallery recently reflected area distributions of oil derricks in Texas, taxi ridership in New York City, or graffiti in Chicago.

For those who find relationships they want to explore in more depth, Tableau creators can offer the ability to download datasets for re-use. Under the Resources tab, these are organized under headings like Technology, Lifestyle, Government, Sports, and Entertainment. If you have a student population that tends toward ideological homogeneity you wish to challenge, the Tableau Gallery is a great place to come and look at other visualizations that a range of media outlets have built.

538 P-hacking

http://fivethirtyeight.com/features/science-isnt-broken/

Five Thirty Eight's p-hacking interactive examines relationships between two bodies of information, using a graphical interactive to quickly transforming raw data into visualizations, exploding the 0.05 p-value level typically used for predictable publishable statistical significance all the while.

The interface encourages you to "Hack Your Way to Scientific Glory." The first step is choosing one of the two major political parties, and then choosing or excluding various levels of government representation through a series of checkboxes. As those changes alter the dataset under consideration, the p-value shifts. For example, by focusing on the terms of Democratic Presidents and Governors, including Employment, GDP and Stock Prices but excluding Inflation, one can produce a publishable result, as it achieves a p-value of less than 0.01, demonstrating Democrats have a positive effect on the economy. Adding in Inflation negates the significance of the finding.

The p-hacking site is designed to demonstrate that that threshold is so low that researchers can easily manipulate their findings, particularly when given a large range of data points, to claim significance where it might not exist. While researchers should establish those thresholds and variables from the outset as a part of research design, the p-hacking site allows for quick and tangible demonstrations of how multiple regression analysis can be problematic.

Google Public Data

https://www.google.com/publicdata/directory

Using data from the U.S. Census Bureau, Bureau of Economic Analysis, Bureau of Labor Statistics and other government sourc-

es, Google's public data site features quality sets with the ability to create some simple visualizations, all within an intuitive and easy-to-use interface.

Access to Census data can be found through a number of sites, but the menus Google provided through this tool allow for near-instant comparison and adjustment of graphs through easy access to the x-axis. Examining population by gender, for example, will allow you to consider the U.S. as a whole or drill down more locally. Perhaps students are interested in forecasting populations in particular countries. When accessing the Census Bureau's International Database, you can use checkboxes to include or exclude nations based on your area of interest. You can switch between cluster representations, bar graphs, and live graphs. Any attempt to measure with inadequate data will trigger a warning.

Additionally, you can upload your own datasets and use their visualization tools for plug-and-play displays. The site drives home the portability of data accessible by a variety of interfaces.

Finding existing datasets

When getting started with data, it's best to stick to the Goldilocks principle and look for a *just right* amount of data. Sometimes it's better for teachers to structure the assignments rather than let students haphazardly look for data. Student novices often run into roadblocks, not anticipating that some data simply cannot be collected and in other cases, people cannot be monitored all the time or might have reasons to not be candid. Using existing datasets helps students who might otherwise alight on a narrow data need that isn't generalizable. *Quick reads* of existing datasets also reinforce the process of drawing conclusions.

100 People: A World Portrait

http://www.100people.org/statistics_100stats.php?section=statistics

Concerned with global measures, this site makes it easy to compare one group to another by converting each statistic to being a certain percentage of 100. This makes the site a good match for younger or beginning learners.

Curious about the distribution of overall literacy globally? Students will find that 86% can read and write and 14% cannot, then find a pdf from UNESCO used to inform those percentages.

When accessing Detailed Statistics through the toolbar, scrolling to the bottom provides a link back to datasets – including the CIA World Fact Book, the World Health Organization, and World Bank – providing citations for each statistic.

100 People provides lesson plans for teachers, and using the information contained here, data becomes accessible for even the youngest students and those with learning differences. Students can make charts and graphs using data in 100-unit breakdowns, or they can take the data and work with it in analog capacities.

U.S. Census Data — People and Households

https://www.census.gov/people/

From tracing ancestry to counting the number of same-sex couples, the range of data collected by one-time and periodic supplemental surveys conducted by the U.S. Census Bureau offers a vivid portrait of evolving American life.

This menu offers access to a range of datasets, each with specified parameters, explicit terminology, and clear date of collection. There are variations between how data is conveyed. Some datasets – for example, Computer and Internet Use – offer access to tables and visualizations.

For American students, U.S. Census data is among some of the most comprehensive information available, but students should be aware the government involvement in collection presents its own issues, especially for individuals with anti-authoritarian perspectives. Any use of Census data can involve discussion about how the data is collected from the populace at regular intervals and supplemented with interview data.

Examining the distinctions between survey approaches can illuminate the way that methodologies inform outcomes. In the midst of the negotiations between the FBI and Apple over access to a locked iPhone belonging to suspected San Bernadino terrorists, the website The Verge featured two surveys about consumer privacy protections only days apart. One survey found the populace sided with the FBI in terms of mandating access, while the other survey found the majority siding with Apple on the side of privacy. The language of the surveys varied from the abstract to the concrete. Should Apple be forced, through this precedent, to unencrypt phones? Most respondents felt not. But should Apple cooperate with the terror investigation? In that case, the majority felt they should. Asking the question two different ways got two different but valid responses.

Registry of Data and Search Repositories

http://www.re3data.org/

This registry collocates a range of dataset repositories. Most subject areas focus on biological and physical sciences, but also include research on digital humanities projects, linguistics, archaeology and big data from hard science.

For example, typing [student athlete NCAA] into the home page search box retrieves the NCAA Student-Athlete Experiences Data Archive. Clicking on the archive's link (http://service.re3data.org/repository/r3d100010824) provides the url (http://www.icpsr.umich.edu/icpsrweb/NCAA/) for the datasets from a topic area with potential to interest student researchers.

The repository is in the public domain and licensed under Creative Commons CC-BY, meaning you can copy and redistribute the material in any medium or format and remix, transform, and build upon the material, as long as you give attribution.

Dryad

http://datadryad.org/

Dryad takes some 15,000 datasets and associates them with the research articles they generated. This service is provided for long-term preservation, assigning Digital Object Identifiers for proper attribution through data citation.

Dryad's curated Twitter feed demonstrates its range of data-based projects represented there and models the possibilities and technique of research using existing datasets. A recent article tackled "Disparities in influenza mortality and transmission related to sociodemographic factors within Chicago in the pandemic of 1918" (http://datadryad.org/resource/doi:10.5061/dryad.48nv3; Grantz et al., 2016).

The idea behind Dryad is allowing researchers to move beyond reading others' results to allowing readers to replicate the experimental design and confirm earlier findings. This site offers tools for the highest aims of replicability, something that should be of concern even to student researchers.

Data Portals

http://dataportals.org/search

Data Portals offers a clearinghouse for data uploaded from a range of government and nonprofit projects and agencies. Datasets are tagged and organized around subjects, allowing for searching as well as browsing their catalogued collection of links.

The attributes of each dataset – including usage rights, current activity, and the availability of application programming interfaces (APIs) – are part of the catalog's record for each set.

Want to know the geographic distribution of public toilets in Bath, U.K.? How about the books "exiled" by the Third Reich (including Dos Passos, Hemingway, and Upton Sinclair)? This treasure-trove of information offers thousands of potential jumping-off points for student researchers. It is interesting to explore the many governments and organizations that have made statistics and research available to citizenry through the open data access movement. Users can propose links to their own or other existing datasets, too.

Open Access Directory

http://oad.simmons.edu/oadwiki/Data_repositories

Another repository linking to the growing number of open data initiatives, Open Access Directory (OAD) is a crowd-sourced, wiki-based initiative to sort open data sources into disciplinary buckets as diverse as linguistics and archeology.

Of particular interest are the Interdisciplinary datasets the OAD collects, including FigShare (https://figshare.com), a site which allows access to unpublished data and studies which produced negative results, potentially revolutionizing literature reviews by allowing the inclusion of research not otherwise accessible.

Gapminder

https://www.gapminder.org/data

Gapminder is perhaps the easiest way to create visualizations based on their integrated datasets. The research tends to come from international, nongovernmental sources like the World Health Organization and the International Labor Organization.

The sets are available for download as spreadsheets or can be viewed through their web interface in both one-time versions and those which reflect change over time. Many of the data points in Gapminder will be of interest to students, for example: *how many cell phones per 100 people exist in the world?*

Tracking tends to be over time, like income per capita in different countries. Gapminder enables students to easily grasp the effects of globalization and population shifts.

Visualize Free

https://visualizefree.com/datasets.jsp

Visualize Free offers integrated datasets, as well as a mechanism for you to upload your own.

Easy-to-operate filters allow you to target particular data points on topics including Walmart locations across the country and date the store opened.

The site also offers ways to understand data gleaned from *Forbes' America's Top Colleges*, including cost, the percent of students receiving financial aid, average test scores, and faculty-to-student ratios, something particularly meaningful for high school students.

Knoema

https://knoema.com/atlas

With a definitive social science slant, Knoema offers interesting international data. One example reflects expenditures spent on food displayed as a percentage of household budgets.

The site auto-generates charts and graphs and allows for export of tables as comma delimited .csv files, so students can use spreadsheet plug-ins to customize their own visualizations. For

teachers looking to exploit some of the more nuanced versions of Excel, this site could prove valuable using raw data to explore properties which inform the function of spreadsheet cells.

Representing data responsibly

One of the most exciting learning opportunities comes when students collect their own data. In Geoff Herbach's novel *Fat Boy vs. the Cheerleaders* (2014), his protagonist charts the consumption of products from the high school soda machines to support his argument that band kids, not the dance team, deserve the proceeds.

But if students collected information themselves, they'll need some help with formatting that data. The first step is usually removing duplicate or incomplete data and looking for multiple values in the same field. Students should develop some protocol for working with data and systematic methods grouping together different versions of the same reality. They will also want to consider the distribution and whether there are outliers you may want to remove. OpenRefine and Tableau Public are intuitive tools to help students clean up their own datasets.

OpenRefine

http://openrefine.org/

OpenRefine is an interactive data transformation tool, like a spreadsheet, but both easier and more powerful in specifying the attributes of a particular aspect of data. OpenRefine is dynamic in that all the data changes immediately and is reversible, taking away much of the anxiety of working with large datasets.

OpenRefine is particularly useful for consolidating different spellings or related concepts — our state library association used it for one project to bring together a population who variously described themselves as school librarians or media specialists, for example. OpenRefine is dependent upon the ability to install software on a local computer.

Tableau Public

https://public.tableau.com/s/

Tableau can use your geographically based data to create easy visualizations about the density of populations through heat mapping.

Using its ZIP code option, one of my professional organizations used it to discern where members lived to inform our targeting of certain areas in membership drives.

These are just some of the possibilities for using data with students. As a school librarian, one of the first data projects I worked on with a whole class involved a stock market contest. Each economics student chose one stock and tracked three months of closing prices before graphing the gains and losses to determine an overall profit leader. The main takeaway was that the scale of units matter in visual depictions. Excel generated a scale based on the prices in the spreadsheet, but adjusting the axes from the default let students know they had choices which would add or remove inherent drama from the graphic visualization. When students contrasted identical stock prices on 20-unit scale vs 120-unit scale, the two graphs look very different. Consciousness of units is critical in building awareness about the importance of attending to axes.

Conclusion

Teachers can encourage students to think computationally about relationships in the data with which they are working. Those are important information literacy skills. Working with data does not necessarily mean performing mathematical operations.

To master data fluency concepts within limited time frames, educators can locate existing datasets, and design assignments where students manipulating data and teasing out the relationships is the goal and then making comprehensible statements about the relationship between the data.

Presenting data in a responsible way involves students getting their point across using the language, the variables, and even the types of charts and graphs that will portray their data accurately and not necessarily overstate the data relationships. Constructing charts and graphs will make students subsequently better readers of information in these formats.

Resources

Herbach, Geoff. 2014. *Fat Boy vs. the Cheerleaders.* Naperville, IL: Sourcebooks.