Towards a Data Literate Citizenry

Michael B. Twidale
twidale@illinois.edu

Catherine Blake
clblake@illinois.edu

Jon Gant
jongant@illinois.edu

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Abstract

We believe that data literacy should be a skill not just for scientists, but for all citizens. We make the case by considering literatures on various kinds of literacy, and using a number of examples to explore some of the challenges that emerge from trying to move to a more data literate citizenry. We consider the opportunities arising from developing technologies to help individual communities and intermediaries in understanding how data should inform personal and societal decision-making.

Keywords: data literacy, citizen participation, open data, information reuse

Introduction

Continuing growth of hardware, software and networking makes it possible to collect and use ever greater amounts of data. Along with opportunities, concerns arise about what we as a society want to allow, encourage or forbid, particularly with respect to privacy, security and ownership of different kinds of data. Debates on these societal and policy issues are themselves informed by analyses of data, as are other debates on priorities, finance, resource allocation, healthcare, climate change, etc.

This current data environment creates a set of concerns over the skills necessary to make intelligent use of available data. What needs to be known and who needs to know it? We see the growth of various professions of data analysts with specialist skills in creating, maintaining, using and interpreting data. However we claim that in a democracy it is important that we try to enable all people to have some level of data literacy in order to be able to fully participate in a discussion around important decisions that a society has to make – informed by ever more data. This is an argument for data literacy based on the importance of traditional literacy (what we might call 'book-literacy') in conventional societal participation. As the interest in Big Data continues, we believe that the question will recur of who gets to participate in the discussion and who has to simply surrender to trusting the judgment of experts.

Our goal in this paper is to explore what it means to have a data literate citizenry, and why it matters. This vision leads to a research agenda that builds on earlier ideas of information literacy, but considers the complex information ecosystem where people, information, data, methods to process data, and mechanisms to share all interact. As such, data literacy needs to be considered from multiple perspectives and involve researchers from a range of backgrounds – not just domain experts and curriculum designers, but also systems developers, designers of information visualizations, and researchers in online communities, communication, community informatics, the public understanding of science, informatics and library and information science. Information professionals will remain critical players as teachers, facilitators and intermediaries in data literacy interactions.

Acknowledgement: This paper is made possible in part by a grant from the U.S. Institute of Museum and Library Services (IMLS), Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12.

Twidale, M.B., Blake, C. & Gant, J. (2013). Towards a data literate citizenry. *iConference 2013 Proceedings* (pp. 247-257). doi:10.9776/13189

A Literature of Literacies

Building on established work on regular literacy (the ability to read and write text), researchers have explored other literacies. Even when talking about text, literacy researchers note that the basic ability to read and write is not enough – we should also think about abilities to understand and use what is read and written.

As a result there are now literatures on numerous literacies, each with multiple, often somewhat contradictory and overlapping definitions. For example, there are literatures on literacy, digital literacy, information literacy, scientific literacy, and data literacy. For our focus on data use we also need to consider statistical literacy and indeed basic numeracy. In many contexts, a degree of comfort with using computational technologies is also necessary – something addressed in the computer literacy literature.

In his review article on digital literacy Bawden (2001) notes that in the context of basic literacy (reading and writing), the concept of levels of literacy has proven useful. In highly developed countries few people are completely illiterate, but various studies have shown worrying proportions of the population to be functionally illiterate. Categories such as very low literacy and low literacy are useful in understanding the level of skill needed to make use of textual information in various ways. For example a person with low literacy may be able to read simple text, but be unable to look something up in an alphabetized directory. Bawden clarifies the relationships between various literacies, including a useful history of the development of the idea of information literacy. A later work (Bawden, 2008) provides additional contextualization of the concept of digital literacy and various meanings that different authors apply to different literacies.

Of these literacy terms, information literacy is the one that seems to generate the most interest in the field of Library and Information Science. The Association of College and Research Libraries (ACRL) (2000) defines information literacy as: a set of abilities requiring individuals to "recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information." This is then elaborated into six elements:

- Determine the extent of information needed
- Access the needed information effectively and efficiently
- Evaluate information and its sources critically
- Incorporate selected information into one's knowledge base
- Use information effectively to accomplish a specific purpose
- Understand the economic, legal, and social issues surrounding the use of information, and access and use information ethically and legally

Substantial work has explored ways of teaching information literacy skills to all ages; from elementary school, through college and to adults in public libraries. The Big6 approach developed by Eisenberg & Berkowitz (2011) is a particularly well-regarded example.

The importance of information literacy for civic participation has been emphasized for a long time, for example by Owens (1976): "Information literacy is needed to guarantee the survival of democratic institutions. All men are created equal, but voters with information resources are in a position to make more intelligent decisions than citizens who are information illiterates. The application of information resources to the process of decision-making to fulfill civic responsibilities is a vital necessity." Work in statistical literacy also emphasizes the importance for citizens to understand the issues underlying society and the economy (Podehl, 2003).

Similarly, Shapiro & Hughes (1996) make the case for information literacy as a prerequisite for participation in an information society. They build on the 18th century vision of Condorcet, who advocated for access to information (and the skills necessary to read and use it) so that educated citizens "will no longer depend for every trivial piece of business, every insignificant matter of instruction on clever men who rule over them in virtue of their necessary superiority." We aim to extend this argument to ideas around data literacy.

In a similar vein, Zuccala (2009) shows how open access to scientific literature can and should benefit not just fellow scientists but also the layperson. Healthcare information is a primary area of interest for many people (particularly that relating to an individual's or family member's medical condition), but it is certainly not the only area. Zuccala's use of the resonant term "laity" reminds us of the risks of delegating all power of use and interpretation to its antonym: priesthood. It is possible that a priestly caste of scientists are the only ones able to collect, analyze, and interpret data, and then remain the only conduit to the laity about what the findings mean for making important societal decisions in lawcourts and

legislative assemblies. With open access to research papers (Zuccala's point) and underlying data (our extension) the laity can have at least an opportunity to try and verify certain claims and even maybe make their own discoveries. Such freedom entails numerous risks (as do all freedoms). Knowing how to search, select and interpret findings are critical skills to avoid misinterpretation. Access is necessary, but not sufficient. Zuccala's analysis shows how we can understand current and future needs for civic scientific information literacy by weaving together work on information seeking, science communication in society, and the public understanding of science. This leads to a consideration of the potential contributions of intermediaries including science journalists, museums, scientist-popularizers, and (we claim) librarians and other information professionals.

Schield (2004) connects three literacies (data, statistical and information literacy) all necessary to enable critical thinking skills including analysis, interpretation and evaluation – concepts that we further explore in the next section.

Researchers using the term data literacy have looked at its use in various contexts including enabling educators to use student performance data to improve teaching (Love 2004), and for undergraduates studying geography (Hunt, 2004). Qin & D'ignazio (2010) consider the requirements for a new course addressing data-related literacy for science students. Their investigations revealed the central role that metadata should play in their course. Carlson et al. (2011) use the compound term data information literacy to refer to the skills needed to prepare students to participate in e-research. They used the ACRL six element framework as a way to examine the topics they had derived from an analysis of faculty or student practices and needs, finding considerable conceptual overlap.

Surveying multiple overlapping fields can be rather confusing. There is something of a terminological mess as different authors can mean slightly different things when they use the same x-literacy term, and very similar subconcepts can be part of more than one x-literacy. One way to help make sense of the richness and variety of literacies literature can be to consider issues of scale, genre and use. Different skills are needed at different scales. To cope with, or make use of (that is, read) one book you need basic literacy (book-literacy). To cope with or make use of 100 or 1,000 books (a small library) you also need certain kinds of information literacy. Making the transition from a small school or public library to that of a large university research library (1,000,000 or 10,000,000 books) requires additional information literacy skills. Having access to online journals and indeed the entire web (1 to 100 billion documents) requires yet more information literacy skills. Each of these levels of skills supplement those of the smaller scales – they certainly do not replace them.

As we switch genre from text documents (books, articles, web pages) to data elements (in millions of datasets, each potentially containing billions of values, plus associated metadata of varying form, format, quality and completeness), effective use also requires data literacy. Similarly, reading a single book on science may require a certain amount of scientific literacy (Hazen & Trefil, 2009), basic numeracy, and perhaps statistical literacy. Making use of larger datasets and findings derived from multiple, perhaps heterogeneous datasets requires further skills relating to statistics, estimation, error, and as Qin & D'ignazio (2010) reveal, metadata.

In data literacy just as other literacies we are concerned with not just consumption (reading) but also production (writing). Levels of sophistication matter; much can be improved without everyone becoming an expert. Fluency means an ability to gain an overview as well as follow details. In the context of very large amounts of resources, different access and evaluation skills emerge and grow in importance. Finally, literacies are not just individualistic. They are about participating in a community, and sociotechnical systems can be designed to help that community nurture ongoing learning and growth..

Motivating Example: Data in the Courts

As a very simple, constrained example consider the case of statistical evidence in court proceedings. Expert witnesses are brought in, but their testimony needs to be understood by the lawyers, the judge and of course the jury. Problems arise when the level of understanding and (it must be admitted) the clarity of the evidence is lacking. A recent UK case highlights the problem. A 2010 UK Court of Appeal case (Regina v. T.) involved the use of Bayesian methods to assess the rarity of the print from the shoes of the defendant that matched those found at the scene of the crime (Saini, 2011). Under imperfect information about the exact number of particular brands sold in the UK, the expert witness had had to make some informed estimates. The process was perhaps poorly explained and the judge consequently quashed the conviction. Additionally the judge also ruled against using similar statistical

analysis in the courts in future. This has been misreported as an exaggeration claiming that the judge banned all use of Bayesian statistics in UK courts. Rather the judge said that Bayesian methods were an inadmissible way to present expert evidence — except for DNA and "possibly other areas where there is a firm statistical base". The sales estimates used as a component of the shoe match calculation were considered not "firm", but of course this opens up a mess of what is to be determined to count as firm.

The case has raised considerable concern. An article in Nature (Fenton, 2011) notes that fallacies of statistical reasoning have influenced verdicts in dozens of widely documented cases. The article reports the creation of: "an international consortium of statisticians, forensic scientists and academic and practicing lawyers (80 people signed up in the first 2 months) to develop guidelines for when and how Bayesian reasoning should be used to present evidence".

The point of the example is not to point out innumeracy in the British Judiciary (British judges already have a reputation with being out of touch with much popular culture). Rather we want to make the case that as data becomes more widespread, it can and will be used in evidence. This evidence will need to be weighed and will no doubt be contested by the other side in the case. If judges, lawyers and juries don't understand how to handle evidence derived from data, this is as problematic as a legal system where textual evidence is used but some of those involved cannot read.

Data Literacy for Greater Civic Participation

Data produced and collected by government are the basic ingredients for governments to provide services, make policy, and be held accountable for their performance (Heeks, 1999). "In democratic societies citizens have a basic right to know, to speak out, and to be informed about what the government is doing and why and to debate it." (Stiglitz, 1999 p.29). Strategies to promote more open and participatory approaches for government center on "using information and communication technologies to operate transparently, facilitate easy and low cost access to public records, and to make civic and social data available in standardized formats that support productive public use of data" (Knight Commission, 2009).

Openness and transparency are being enhanced as governments implement open government initiatives to increase transparency, participation, and collaboration. Open government refers to "government that co-innovates with everyone, especially citizens, shares resources that were previously closely guarded; harnesses the power of mass collaboration, drives transparency throughout its operations, and behaves not as isolated department of jurisdiction, but as something new, a truly integrated and networked organization" (Lathrop & Ruma, 2010). A commitment to access is laudable, but without the appropriate data literacy skills, co-innovation with citizens is unlikely.

A Vision and How to Get There

We envision a sociotechnical ecology where data, information, people and technology co-evolve. That means that we are not advocating for a particular curriculum, but rather for how we might combine resources in different ways to increase levels of data literacy. There are likely to be multiple solutions and indeed different kinds of research that can contribute to these goals. We illustrate those below.

.

Extending Data Literacy to Citizens

Most of the existing literature and discussion focusses on scientists, trainee scientists, information professionals, and other experts. We want to extend this to considering the general public as a whole, as a data literate citizenry. Other points on this dimension may be non-experts or people with different amounts of expertise but with a particular need. For example how do you get a jury up to speed with data-based evidence, and how is it best presented and discussed?

Similarly, although we increasingly talk about "citizen scientists", the nature and depth of participation in the scientific process can be very variable. While it is entirely appropriate to have certain activities that are very basic but nonetheless useful to science and meaningful to the participants, it would be good to also provide structures for ever greater participation at greater depth. A useful analogy would be how Wikipedia provides ways for more sophisticated participation not just in editing web pages but adjudicating disputes and managing the entire process. In a similar way we might explore ways that we

can help people migrate from “citizen lab assistants” to something more like true “citizen scientists”. Figure 1 best characterizes our vision to include citizens with little or no statistical training into the fold of data literacy.

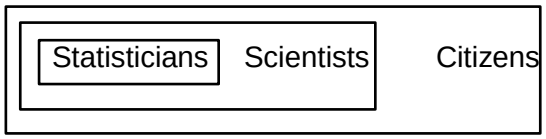

Figure 1. Work on cyberinfrastructure pushes data literacy from the realm of statisticians to scientists. We envision citizen users with little or no scientific training.

Using Existing Information Infrastructures for Learning

The work of Qin & D’ignazio (2010) and Carlson et al. (2011) are examples of approaches geared to teaching data information literacy in formal learning contexts (courses offered at universities). There are alternatives of more informal learning, including science museums, book popularizations of science, TV shows, podcasts, etc. There are also approaches that are somewhere inbetween or happen in an overlapping context. For example, information literacy may be taught in a formal lecture, or may arise in a help-giving interaction with a reference librarian that can involve teaching by stealth. In an online health community there may be peer support and activity that is more about incidental or co-learning - where the help giver(s) did not know the approach at the outset, but discovered it along with the help seeker(s). Variation and innovation can occur along this spectrum. In addition to peer-to-peer learning, the growth of MOOCs shows one possible way of enabling very large groups of people to learn certain topics at low cost (Hyman, 2012).

Empowering New Teachers

Online communities around topics such as personal healthcare (Preece 2000) and tech support (Singh & Twidale, 2008) often involve teaching skills of data use and data manipulation. This kind of teaching or help-giving may be done by those formally designated with that task, or as a more spontaneous emergent activity.

If we wish to extend this, how do we design/build/engineer an online community that helps people explore and understand a particular kind of data (say, personal health information)? Or do we just hope that it emerges? What can be built so that more established members can help newcomers to understand a dataset, what to do with it, and how to avoid getting confused or making incorrect inferences? Most online communities are text forums, for obvious reasons of convenience and familiarity. But when talking about data and its creation, interpretation and use, is text really the only and best medium for discussing and explaining? How can visualizations and software use be better folded into these discussions?

Empowering Old intermediaries

There are many kinds of intermediary who may help in supporting greater data literacy; just as different information professionals have been involved with basic literacy and with information literacy. For data literacy these intermediaries include librarians, data curators, journalists, expert witnesses, and scientists acting as popularizers via books, TV shows or online. An example of the latter in the context of data analytics is Hans Rosling (2007), who uses Gapminder data visualizations to help make complex (and often counter-intuitive) points about social development in an accessible way. There is always the risk that intermediaries end up doing the work for the end users, rather than helping the end users do it for themselves. This would be the difference between say a professional letter writer providing a vital service for an illiterate person, and a teacher helping that same person learn to write their own letter. As the analogy indicates, learning data literacy is likely to be time consuming and difficult. It is understandable why people may be daunted, give up or delegate the understanding to experts. As

advocates of a data literate citizenry, we will need to explore how to lower the learning barriers of cost, time and effort.

Building New Technology Support

Data literacy discussions often focus on who is teaching whom and what is being taught. This is entirely appropriate. But it is also useful to consider the ways that different technologies can help that process by making it easier to understand and use data. Typically these were initially developed for the benefit of experts, but with some extra work they can be made more accessible to a wider population. This is what happened with library catalogues and many bibliographic databases. Initially they were developed for use by highly skilled librarians who learned complex query languages in order to maximize recall and precision in a single (extremely expensive) query. These became online public access catalogues to libraries with easier to understand and use interfaces intended for patrons. Similar routes were followed by various databases, and of course by search engines. We can aim for similar approaches to data analytic tools.

Examples of tools that could be developed or refined to be more accessible to a general audience include bibliometric search, data mining, text mining, data visualization, claims analysis, and tools to help with understanding of terminology or methods (which may be as simple as Wikipedia lookup and YouTube videos). In addition, a less data literate population is likely to encounter difficulties and need help in how to use tools for different purposes. Consequently there is also a need for technologies to support the discussion of, learning and sharing of skills in online and face-to-face communities

Some tools may be so powerful as to work in a standalone context for many users (as OPACs often do). An example may be a tool that aggregates and summarizes a literature on a topic in medical research, as well as providing explanations of what it did, and what various technical terms mean. Although it is highly desirable to have standalone applications that can explain something without needing an intermediary, much can still be gained by developing applications that partially help the end user, but still need an intermediary to explain the subtleties.

Tools can be useful for facilitating data literacy even if they are not complete standalone infrastructures that enable all kinds of data analysis or create a “citizen data analytics environment”. It can be useful to have a diversity of tools of varying sophistication, which individuals and groups can assemble to meet particular needs. For citizens this may include many more simple, basic and low cost tools, while scientists require more complex tools that are harder to learn and are more expensive. That is, we need to support ways of combining tools like Many Eyes and Google Refine as well as similar more powerful applications. It remains an ongoing challenge to design tools to optimize various ad hoc assembly and repurposing. Equally, we need to design these tools to enable the learning of data analysis skills and not to be so daunting to learn and use that they serve to further alienate people already somewhat intimidated by datasets and lacking confidence in basic mathematical skills.

Levels of Data Literacy Sophistication

Inspired by research in basic literacy highlighting the importance of considering levels of literacy, we think it will be fruitful to develop a better sense of levels of data literacy that permit particular kinds of understanding, interpretation and data use.

For example, a basic level of understanding of a straightforward dataset might involve understanding simple measures (such as mean, median and variance) and how they can be used to assess patterns and trends over time. More advanced concepts may be around how large datasets can be used to predict likely future outcomes, such as the probabilities of different events. Yet more advanced would be ideas like Type I and Type II errors, and simple Bayesian statistics, as in our court case example.

Other cases might be more about a degree of guided literacy – where citizens may not be able to operate independently, but can with help from others; who may not necessarily be experts. For example, we might assess whether citizens can follow a data-driven argument and question things that seem odd. In the context of data driven findings from experts, can we get to the state of the Reagan quote of “trust but verify”? That is, where individuals or groups can (with help) actually check some of what the experts are saying.

In the legal example of the problems with Bayesian reasoning we can see how that might play out. It would certainly be very nice if in the future all school leavers had a rudimentary understanding of these issues, thereby being able to serve effectively as jury members; assessing statistical evidence and able to participate in political debates where data is marshaled and contested. But lobbying to add yet more content to a curriculum is a long term endeavor. In many countries various unintended consequences of high stakes testing make it difficult to focus on more subtle analytic skills as opposed to easily memorizable facts and exam-passing tricks. It would also be nice and somewhat easier to lobby for all people graduating from university to have a basic set of data literacy skills, perhaps as part of a General Education requirement. We would definitely support that, and with greater hope of change happening faster than at the school level. But in the meantime there are ameliorations that are well worth doing.

It is incumbent on experts to acquire not just domain skills but communication skills. For example, what are the best ways of communicating Bayesian statistics so that judges, lawyers and juries can understand them? What are some classic recurrent misconceptions? How might intermediaries develop repertoires of detecting and addressing these misconceptions? How might information professionals help in not just accessing data but co-interpreting it? We believe there is a vital, indeed radical role for librarians, amongst others, in this space.

Integrating Different Approaches to Data Literacy

Experts in curriculum design can clearly contribute by identifying core skills and prerequisites. We have seen substantial work in this area in various x-literacies. But we want to emphasize that it is not the only kind of research contribution. More analytic work can help to understand the barriers to data literacy, ranging from better understandings of math and computer phobia, to identifying and understanding commonly recurring misconceptions about data, probability and statistics. Work on misconceptions in physics (Brown, 1992) and statistical reasoning ((Tversky & Kahneman, 1974) indicate the potential of misconception analysis.

Builders of technology can explore ways to make data analysis easier to do – or easier to explain to less expert individuals. Improved visualization of results is highly desirable, but there is also a need for better visualizations of other analytic work: how the results were obtained, errors in the results, the claims and counter claims in the literature and step-by-step how-to guides of using tools. Similarly, carefully crafted data visualizations can help in illustrating concepts that we know that people find difficult like Type I and Type II errors.

In the past the barrier was just getting access to the data. This is becoming easier. It is necessary but not sufficient for a citizenry to be in a position to be able to verify claims. As well as understanding the claims of scientists as they relate to policy, data literate citizens can also make use of that data in their lives and in participating in science. We have advocated for data literacy as desirable for civic participation in general, but this can seem a rather abstract idea. A more immediate application of data literacy skills through activities such participating in citizen science, or in personal health management can be a powerful motivator for developing those skills. We now give two scenarios that illustrate what this might look like in the near future.

Situating Text in the Data Literacy Conversation

Scientists already use published peer-reviewed literature to inform public health policy. For example, systematic reviews conducted by the Cochrane Collaboration and Health Technology Assessment play an important role in evidence based medicine which in turn can influence government policies on standards of health care. In epidemiology meta-analytic results (a quantitative form of a systematic review) can influence legislation, product label requirements, and public services.

The systematic review process is typically a group activity whereby scientists identify a comprehensive collection of articles, extract information from those articles, verify the accuracy of those extracted facts, and analyze the extracted facts using either qualitative or quantitative techniques (Blake & Pratt, 2006). Although systematic reviews accurately capture evidence, the process is time-consuming, taking 28 months from the original conception through to publication (Petrosino, 1999) and 1139 hours (Allen & Olkin, 1999). With more than 21 million citations in MEDLINE and an additional 1900 new citations added every week, the manual techniques currently used are becoming increasingly difficult to

apply. Consider a breast cancer expert. It would be difficult, but necessary for her to consider the 33,883 articles published on breast cancer during the 28 months required to conduct a systematic review. Faced with the daunting task of sifting through currently available and recently added articles, our breast cancer expert may turn to other strategies to reduce the number of articles, such as constraining her hypothesis or her selection criterion. However, both of these constraints introduce undesirable biases, and thus reduce the validity of her review to inform public policy. Citizen scientists could play an important role in this process by participating in the information extraction activities, for example you don't have to be expert in medicine to identify the number of people in a study.

The second key challenge of a systematic review is that articles considered are drawn from published literature and thus may suffer from publication bias; where articles that find statistically significant findings are more likely to be published than articles that do not show statistical significance, even though the methodology of both studies are the same. "For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results" (Rosenthal, 1979).

The Multi-User Extraction for Information Synthesis (METIS) system is an example of using text for analytics in that the system automatically identifies information required in a meta-analysis from full text scientific articles. Such an approach can be used within the existing systematic review process to reduce the time between when findings are published and when public policy is updated. The automatically extracted information, after being verified by citizens, could be used to create a synthetic control group estimate so that information from articles that would not be considered in a traditional meta- analysis could be incorporated into the analysis. Although this is a controversial suggestion from a meta- analytic perspective, the goal here is not to replace meta-analysis, but rather provide alternative ways to leverage textual "big data".

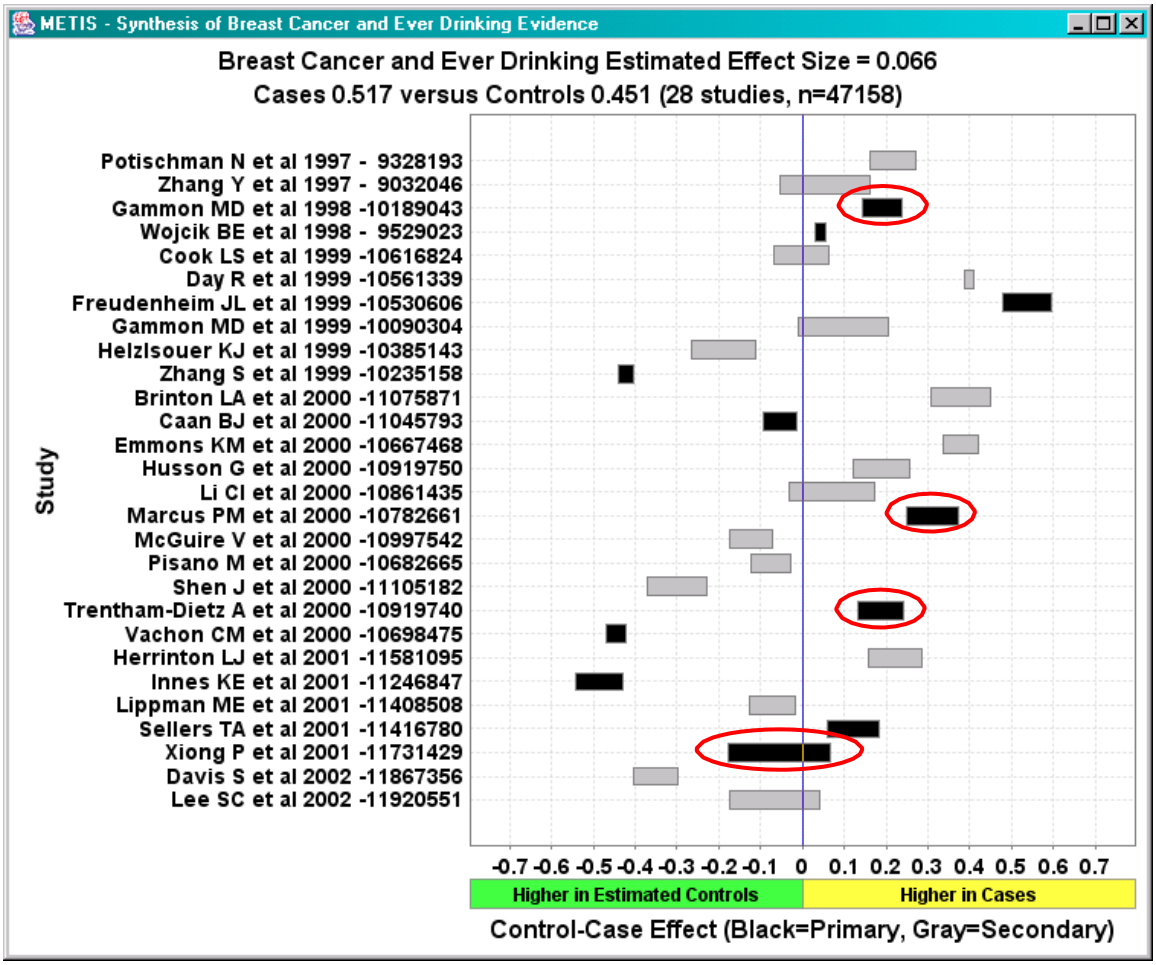

Figure 2. The METIS Summary of breast cancer articles published between 1997- 2002.

Figure 2 shows this new type of analysis using breast cancer articles in key epidemiology journals, which were published between 1997 and 2002. Articles that report alcohol consumption as primary information (in the title, keywords or abstract text) are shown as black and articles that report consumption only within the full text (secondary information) are shown in grey. Although scientists who conduct systematic reviews place a high priority on obtaining the "file drawer" articles, studies of the manual systematic review process (Blake & Pratt, 2006) show that users review the abstracts before retrieving the full text, so it is unlikely that these studies would have been found. In one traditional meta-analysis of breast cancer and alcohol consumption (Ellison et al., 2001) 71 of the 72 articles included have "alcohol" or a synonym in the title, keywords or abstract. Some would argue that methodologically studies that do not report the disease and the risk factor as primary information should not be included in a meta-analysis.

The results of this study show that more than 60% of the breast cancer articles that report alcohol consumption (17 out of the 28 studies) include the data as secondary information (only in the full text) and thus would not be included in a traditional meta-analysis. More important than the number of primary versus secondary articles is the degree to which the findings reported in those articles differ. Of the four articles that would be included in a traditional analysis (circled), three suggest that ever drinking is higher in subjects with breast cancer (the cases) and only one study suggests that alcohol consumption is not a breast cancer risk factor. In contrast to the 17 studies that report alcohol consumption as secondary information, 6 show positive effect, 6 show a negative affect and 5 show no effect. Despite including both primary and secondary information, the METIS results are consistent with the earlier cited traditional meta-analysis, which suggests a small positive effect size between ever drinking and breast cancer.

Text is often missing from existing conversations about data analytics, but a manual version of the proposed strategy which took approximately six weeks to complete on 1008 articles [Personal communication T.Tengs, 2002], has been used to quantify the association between smoking and impotence (Tengs & Osgood, 2001). Assuming that a similar amount of time would be required for the 240,000 breast cancer articles such an analysis would take 27 years. Although issues such as access to full text need to be resolved before this approach can be applied to all breast cancer articles, METIS can be used on the subset of articles that are available electronically. Moreover, METIS can reduce the time required for a traditional analysis by automatically identifying information from the articles and thus reduce the time to integrate new scientific findings into public policy.

Making it Personal

People will have more and more access to data about themselves – recording their own activities and with a move to personalized medical care. Looking for health information online is already widespread. As healthcare costs continue to soar a data literate citizenry could help control the expense of managing chronic diseases by accessing the latest scientific information and by incorporating those findings into their daily lives. We begin by describing Morgan, our data literate citizen of the future, as a data consumer.

Morgan remembers a radio story that suggested certain mushrooms could help reduce risk of breast cancer, but she doesn't know which type of mushroom would be best for her. She uses her computer to identify mushroom study results that are most relevant her stored genetic data. Unfortunately none of her genomic characteristics match the study profiles in the mushroom study, so the system accesses the latest literature on mushrooms. As with any disease or treatment the scientific literature reports many different studies from different countries that each report slightly different results. The system weights each study result based on the quality of the study design and the similarity between Morgan and the subjects used in the study. Finally the system converts the outcomes to more accessible language and visualizations that Morgan can understand and suggests three different types of mushrooms would be beneficial.

Greater access to scientific data in an understandable form is a wonderful first step, but the real benefit of the envisioned data ecology is that Morgan can also actively participate. Let us consider also Morgan as a data producer and analyst.

Based on the system recommendation, Morgan decided to include shiitake mushrooms into her diet. But rather than exploring this alone, she participates in an ongoing community dietary study. Like her neighbors she has also noticed that her cholesterol levels have been lowered, but there seems to be a difference depending on where the mushrooms were purchased. She tells the system to record the date,

time and location of each purchase and the system establishes that particular farms appear to have bigger impacts. Combining these results from other growers around the world reveals that soils with certain trace elements are particularly beneficial.

Conclusion

As well as being essential for doing data driven science, some level of data literacy is becoming increasingly important for lay people to participate fully in society in various ways including democratic decision-making. This is likely to come about through a mixture of technological resources and people: teachers, experts, helpers and peers in communities both online and face to face.

This vision for a data literate citizenry will certainly have some skeptics. Some will argue that a basic understanding of numeracy is critical to use the quantitative data that is available. We don't dispute this claim, but we do suggest that citizens have a range of online learning resources that were previously unavailable. Moreover, there are people in society (including librarians and amateur help-givers) who want to help others understand. In the current online world, this desire emerges in forums and blogs. Despite evidence that many people have poor numeracy skills, there are still domains where many people have a great interest in statistical data, such for sporting statistics (in various cultures, baseball and cricket seem to be particularly data heavy). Participatory activities such as citizen science and personalized health data collection and analysis as part of a health oriented online community can simultaneously serve as both an impetus to begin acquiring data literacy skills and a means for learning them in a more participatory manner.

Regardless of variable data literacy, we as citizens and as a society already make decisions based on data. The motivating legal example hinges on the necessary uncertainties and the degree to which imperfect data quantified as a range of uncertainties can be factored in with other evidence to determine innocence or guilt. Similarly, our decisions about what to eat, how much to exercise and which medications to take directly affect our very existence. We posit that a data literate citizenry would enable us all to have a higher quality of life. Condorcet's advocacy for access to information and the means to make use of it as an 18th century Enlightenment value also applies to the skills to interpret 21st century data.

We need to think carefully about exactly which skills and to what level of sophistication people may need them for various purposes. But as with basic literacy, leaving the laity out of the process of access and interpretation creates a degree of dependency on a priesthood of experts. Equally we must acknowledge that there is a very real risk of lay people misunderstanding and misapplying data. But there is no point trying to prevent greater access – it will happen inexorably anyway. What is needed is a way to help far more people make the transitions from data, through information and knowledge, and so to wisdom.

References

Association of College and Research Libraries (2000). Information Literacy Competency Standards for Higher Education. In *American Library Association.* Retrieved from http://www.ala.org/acrl/standards/informationliteracycompetency

Allen, E., & Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. *Journal of the American Medical Association*, 282(7), 634-635.

Bawden, D. (2001). Information and digital literacies; a review of concepts. *Journal of Documentation*, 57(2), 218-259.

Bawden, D. (2008). Origins and concepts of digital literacy. In C. Lankshear & M. Knobel (Eds.), *Digital literacies: concepts, policies and paradoxes* (p. 17-32), New York: Peter Lang.

Blake, C., & Pratt, W. (2006). Collaborative Information Synthesis I: A Model of Information Behaviors of Scientists in Medicine and Public Health. *Journal of the American Society for Information Science*, 57(13), 1740-1749.

Bloom, B. S. (1984). *Taxonomy of educational objectives*. Boston, MA: Allyn and Bacon, Pearson Education.

Brossard, D., & Shanahan, J. (2006). Do they know what they read? Building a scientific literacy measurement instrument based on science media coverage. *Science Communication*, 28(1), 47–63.

Brown, D. E. (1992). Using examples and analogies to remediate misconceptions in physics: Factors influencing conceptual change. *Journal of Research in Science Teaching*, 29, 17-34.

Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *Sciences New York*, 11(2), 629-657. Czarnocki, S. & Khouri, A. (2004). Filling the Gap: Doing Stats in the Library. Presented at *IASSIST 2004 Conference*, Madison, WI. Retrieved from http://dpls.dacc.wisc.edu/iassist2004/program.html

Eisenberg, M. and Berkowitz, R. (2011). The Big6 Workshop Handbook: Implementation & Impact (4th ed.). Linworth Publishing.

Ellison, R. C., Zhang, Y., McLennan, C. E., & Rothman, K.J. (2001). Exploring the Relation of Alcohol Consumption to Risk of Breast Cancer. *American Journal of Epidemiology*, 154(8), 740-747.

Fenton, N. (2011). Science and law: Improve statistics in court. *Nature* 479, 36–37. doi:10.1038/479036a

Hazen, R. M. & Trefil, J. (2009). Science Matters: *Achieving Scientific Literacy* (Reprint ed.). Anchor.

Heeks, R. (1999). *Reinventing Government in the Information Age*. Routledge.

Hunt, K. (2004). The Challenges of Integrating Data Literacy into the Curriculum in an Undergraduate Institution. Presented at *IASSIST 2004 Conference*, Madison, WI. Retrieved from http://scholar.uwinnipeg.ca/khunt/iassist2004/index.cfm

Hyman, P. (2012). Stanford Schooling—Gratis!. *Communications of the ACM*, 55(3), 22.

The Knight Commission (2009). *Informing communities Sustaining Democracy in the Digital age*. The Aspen Institute.

Lathrop, D. & Ruma, L. (2010). *Open Government: Transparency, Collaboration and Participation in Practice*. O'Reilly Media.

Love, N. (2004). Taking Data to New Depths*. Journal of Staff Development* 25 (4): 22–26.

Owens, M. R. (1976). State, government and libraries. *Library Journal*, 101(1), 27

Petrosino, A. (1999). Lead Authors of Cochrane Reviews: Survey Results*. Report to the Campbell Collaboration*. Cambridge, MA: University of Pennsylvania.

Podehl, M. (2003), Statistics in the classroom: learning to understand societal issues. Presented at the *IASE Satellite Conference on Statistics Education and the Internet*, Berlin.

Preece J. (2000). *Online Communities: Designing Usability and Supporting Sociability*. John Wiley & Sons, Inc.

Qin J. & D'ignazio J. (2010). The Central Role of Metadata in a Science Data Literacy Course*, Journal of Library Metadata*, 10(2-3), 188-204.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results*. Psychological Bulletin,* 86(3), 638-641.

Rosling, H. (2007). Visual technology unveils the beauty of statistics and swaps policy from dissemination to access. *Journal of the International Association for Official Statistics,* 24(1-2), 103-104.

Saini, A. (2011). A formula for justice. *The Guardian Newspaper*. Retrieved from http://www.guardian.co.uk/law/2011/oct/02/formula-justice-bayes-theorem-miscarriage

Schield, M. (2004). Information literacy, statistical literacy and data Literacy. *IASSIST Quarterly* 28(2/3): 6-11.

Shapiro, J. J. & Hughes, S.K. (1996). Information Literacy as a Liberal Art, *Educom Review*, 31(2).

Singh, V. & Twidale, M. B. (2008). The Confusion of Crowds: non-dyadic help interactions. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 699-702.

Stiglitz, J.E. (1999). Transparency in Government. In Islam, R. (Ed.) *The Right to Tell: The Role of Mass Media in Economic Development*. World Bank

Tengs, T., & Osgood, N. D. (2001). The link between smoking and impotence: two decades of evidence. *Preventive Medicine*, 32(6), 447-452.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Twidale, M.B. (2005). Over the shoulder learning: supporting brief informal learning. *Computer Supported Cooperative Work* 14(6) 505-547.

Zuccala, A. (2009), The lay person and Open Access. *Annual Review of Information Science and Technology*, 43, 1–62. doi:10.1002/aris.2009.1440430115