© 2013, *Global Media Journal -- Canadian Edition* Volume 6, Issue 1, pp. 67-72
ISSN: 1918-5901 (English) -- ISSN: 1918-591X (Français)

Making Journalism better by Understanding Data

A Review Article by

Fred Vallance-Jones
University of King's College, Canada

The Data Journalism Handbook
Edited by Jonathan Gray, Lucy Chambers, and Liliana Bounegru
Sebastopol: O'Reilly Media, 2012. 242 pp.
ISBN: 9781449330064.

The Filter Bubble
By Eli Pariser
New York: Penguin Press, 2011. 304 pp.
ISBN: 9780143121237.

PDQ Statistics
By Geoffrey R. Norman and David L. Streiner
Hamilton: BC Decker, 2003. 218 pp.
ISBN: 9781550092073.

There was a time when a notebook, pencil and telephone were pretty much all the tools a reporter needed for a successful career in journalism. A good rolodex helped too, but that was about the only database a reporter would likely ever access. And most reporters would never delve into something as seemingly arcane as an academic research paper. But just as the flask of whiskey hidden in the bottom desk drawer has given way to endless cups of coffee, so has the practice of journalism shifted from a nearly blue collar occupation to nearly a profession, with formal codes of ethics and much greater demands for careful research. Whereas schools of journalism once focused on the fundamental skills of storytelling, advanced research courses are now a common element in the curriculum. All of this is a necessary development in that the world reporters must chronicle has become increasingly more complex, particularly with the arrival of computer technologies that have revolutionized journalistic research as much as they have that in other fields. Suddenly, some of the reporters who once got by on a box of index cards imprinted with sources' names and numbers have had to learn how to navigate online databases, make sense of databases, and decipher the methods and findings of researchers using sophisticated statistical

analysis. Even tasks that seem simple, such as "Googling" a search term on deadline, are actually more complicated than they first appear, just because of everything that is going on behind the scenes when you click that search button and send your request to Google's vast servers.

Google has become the world's most used search engine partly because its results seem so uncannily useful. Results that seem most relevant pop to the top of the search results, almost like magic. Of course, there is no magic involved, just a lot of very sophisticated computer programming. This is true not only with Google, but with social networking sites such as Facebook that also use sophisticated algorithms to determine what to show the user and what to leave out. This means that any presumption on the part of journalists or others that these services are providing a neutral view of the world, with the best and most useful sources coming to the top of the search results page, is illusory. Google has always ranked results based on factors that may be irrelevant to a researcher, such as the relative popularity of a site as measured by the number of other sites that link to it.

However, as Eli Pariser points out in his book *The Filter Bubble* (2011), the technology is becoming much more sophisticated, to the point where two searches by two people, using exactly the same search terms, may yield subtly or substantially different results. Based on your location, your previous search history, and a lot of other clues, Google decides what you might be most interested in and serves that up first. Facebook, similarly, shows you more content in your news feed from people you have interacted with most often or about things you have clicked on. Pariser calls it "the age of personalization", and sees in it profound implications. For example, Pariser explains that stories on Apple computers get more prominence on the algorithm-driven Google News aggregation service than do stories on the war in Afghanistan, even though the latter is arguably more significant. Pariser suggests that lost in the thickets of the search algorithms will be stories about matters that are "important but complicated". That has always been true to some extent, Pariser concedes, but the ability to tailor content for specific audience members makes it far more likely that some people will see little or nothing about important but disquieting things.

> Consuming information that conforms to our ideas of the world is easy and pleasurable; consuming information that challenges us to think in new ways or question our assumptions is frustrating and difficult. . . . An information environment built on click signals will favour content that supports our existing notions about the world over content that challenges them.
>
> (Pariser, 2011: 88)

The makers of these tools see themselves as giving users more of what they want (and in that, probably making them better commodities to sell to advertisers) but Pariser's central argument is that users end up in self-reinforcing loop, cut off from things that might shake up their view of the world, and feed them more of the things for which they have already shown an interest—an "antiseptically friendly world" (Pariser, 2011: 150). Worse, Pariser argues, it is all happening without most users really being aware of it.

The implications of these kinds of technologies for journalists and other researchers are profound and hardly need to be stated. If the world that is presented to us by apparently "neutral" search engines and social networking sites is actually a world that has been carefully computer-manicured for us, it puts that much more onus on us to proactively search out that which we might be missing. If we just take the first few choices offered to us by a personalized search

engine, we might miss that which is really important or significant. A good researcher ought to know this anyway, but it is somewhat disquieting to think that today's most ubiquitous research tool actually changes like the shape shifters of *Harry Potter* fame. It would be as if the card catalogue I once knew and loved magically changed its contents depending on who was pulling open the drawer. I know from teaching in the King's journalism school that many young researchers rely on Google almost exclusively. While we have always taught them to go beyond it, it has the potent attraction of ease of use.

Pariser also warns of a big data industry that knows more and more about every one of us, with large data repositories such as the awkwardly named Axciom possessing billions of pieces of data on just about every American and hundreds of millions of people beyond the U.S., ready to be sold and resold to help companies market goods with almost eerie precision. Pariser gives the example of a traveller who goes to an online travel site to shop for airline tickets. As the search is made, the site deposits a cookie (a small data file with information that can be retrieved later by a remote server) on the traveller's hard drive, allowing the computer to be recognized next time it goes online. The fact that the computer searched for the ticket is then sold to a big data storehouse such as Axciom, which can then resell that information to an airline, all in a few fractions of a second. Now, as you visit seemingly unrelated websites, ads for airline fares on the same route start appearing. This "retargeting" increases the chance we will eventually buy. It is as insidious as it is almost invisible, and mildly off putting for those who might imagine their travels on the Internet are somehow anonymous or private.

All of this gives great power to those who happen to control these critical information gateways and repositories. Pariser notes that young, smart programmers such as Facebook founder Mark Zuckerberg go from nothing to having "great authority over the doings of 500 million human beings" (2011: 88) and may not give a lot of consideration to the social and ethical questions raised by their inventions. "Like pop stars who are vaulted onto the global stage, world-building engineers are not always ready or willing to accept the enormous responsibility they come to hold when their creations start to teem with life" (Ibid: 181-182).

Getting beyond the personalized world created by these technologies becomes the challenge for those who wish to retain some degree of control over their online activities. For researchers it becomes more important than ever to use tools such as advanced search to tailor results to what the searcher is looking for rather than on what Google thinks the searcher wants to see, use other search engines, and, shocking as it may seem to a younger generation, to visit a library. In our private lives, Pariser suggests taking actions such as deleting cookies from our hard drives and being more diverse in our search activities, to make it harder for the algorithms to pigeonhole us or for the data miners to track our online activities.

The Filter Bubble is not a very challenging or very long read, and at times Pariser's prognostications seem apocalyptic, seeing a world in which even a person's DNA becomes a commodity to be digitized and marketed to the highest bidder. But its central point is one that anyone trying to navigate the digital universe ought to keep in mind.

Of course, journalists and other researchers do not need to be passive consumers of data dished out by large search engines and social networks. The explosion in data collection and the growth of Internet have created opportunities for research and storytelling that would have been unheard of not long ago. A subfield, first dubbed "computer-assisted reporting" and now more commonly called "data journalism", is taking advantage of the new opportunities for both research and presentation.

Today, data have literally swept over the news business, driven partly by the emergence of a new generation of web-based technologies that have made the presentation and visualization of data-driven stories easy even for those with no database or web development experience. With little more than a basic familiarity with a spreadsheet program, journalists can use online tools to make interactive maps and charts that once required significant programming skills.

Recently, the folks at the European Journalism Centre and the Open Knowledge Foundation teamed up with O'Reilly Publishing to create *The Data Journalism Handbook* (2012), edited by Jonathan Gray, Lucy Chambers, and Liliana Bounegru. It is an introduction to various data techniques written by a diverse group of practicing journalists in Europe and North America. It is not the first book on the subject, or even the most exhaustive, but because it was crowd sourced, it has a great many contributors, and covers a lot of ground at a very basic level. So someone who wants to gain a glancing understanding of what data are, where they can be found, how they have been used and how they can be presented, can do so by reading this volume. Given that working with data is going to be something new for most journalists, such a light and easy volume may be just what is needed to tease people into looking a little more deeply. This is helped by the fact that the book is available as a free online edition, as well as for purchase from O'Reilly.

The down side of this is that the book has the character of a collected series of executive summaries. Topics such as obtaining and cleaning data, writing scraper scripts and building news applications are dealt within a few paragraphs or pages, when some of these could easily have an entire book devoted to them. The book also includes almost 20 case studies, brief descriptions of stories that saw their genesis in data, though again, some so short that they are more like case teasers. The publishers may have overreached when they titled this book a "handbook". The reader will surely be disappointed if he or she has taken this title seriously. All this said, the book provides a useful introduction to a rapidly growing field, and enables that deeper exploration if the reader so desires.

Compared to the relatively light and breezy approaches of *The Filter Bubble* and *The Data Journalism Handbook*, *PDQ Statistics* (2003) is a book intended for those who would like to have a much deeper understanding of the subject at hand, in this case providing an overview, but not a how-to, of statistical methods commonly used by academic researchers.

Many journalists pursued their particular line of work because of an intense dislike of anything to do with math, and that distaste often manifests itself in a misunderstanding of such statistical concepts as statistical significance and variance. To the unwitting reporter, that a study says a result is significant sounds important. In fact, significant is often misinterpreted to mean just that, an important finding. It is not far from that to the headline about a "breakthrough". That significance means that an outcome is likely not due to chance alone is probably lost on those outside of the initiated. And the methods used to reach that conclusion will surely be Greek to most whose research is of the journalistic rather than academic variety.

Yet having some basic understanding of how statistics work would make for much more discerning journalists, especially those covering science, healthcare, and other fields in which statistical research is commonplace. But how to achieve this when so many journalists would not know a standard deviation from a regression analysis and probably has never even contemplated a university-level statistics course. The average statistical text is probably also a non-starter for those not comfortable with math, but the approach taken by *PDQ Statistics*, by Geoffrey R. Norman of McMaster University and David L. Streiner of the University of Toronto, has the

potential to give the journalist or other reader willing to take the time, a good understanding of the methods, if not the ability to actually employ them.

This is not a new book—the most recent edition dates from 2003—but unlike the algorithms of Facebook and Google, statistical principles and tests persist year to year, even if the software mostly used today to do statistical analysis is constantly being updated and improved. As Norman and Streiner point out, the intent of the book is not to train users on how to do statistics, but to provide them with enough understanding of the various tests, and how they should be used, to allow the reader to read research intelligently and insightfully, as well as identify when tests are being misused or misapplied.

The book starts off with basics, variable types, distributions and different measures of central tendency. Then it steps pretty systematically through various parametric, nonparametric, and multivariate statistical methods. Norman and Streiner also do a good job of explaining the circumstances under which different tests are used, and the basic design flaws that can make calculated statistics meaningless, such as a poorly selected sample; "C.R.A.P. [Convoluted Reasoning or Anti-intellectual Pomposity] detectors" point out ways the tests can be misused. Tables, graphs, and examples of formulae help illustrate what is otherwise a fairly text-heavy book. Touches of humour help ease the reader through what, even with the non-textbook approach, is not the sort of book you would take to the beach for a light afternoon read.

This would also be a good book to keep as a reference, so as one reads through a study that uses statistical methods, one can refer back to the book for a refresher on how the methods are supposed to be used. A sharp journalist might even be able to start spotting flaws in research, rather than skipping to the discussion section in search of findings that can quickly be converted into headlines. He or she might even be tempted to take it to the next level, and start doing some statistical journalism.

It is important here to highlight the big picture, taking these three very different books together and reasons behind recommending them to anyone who wants to take his or her journalistic practice, or learning, to a more sophisticated level. Understanding the sophisticated ways that today's search engines and social networks massage the results you receive is a key step toward doing more effective online research; having some basic understanding of data journalism and its methods may be necessary to even survive in a fast-changing journalistic environment; and knowing something of the statistical methods used in academic research has the potential to demystify what has often been impenetrable for journalists who got into their line of work precisely because of previous bad experiences with mathematics. The outcome would be a better understanding of data of different kinds, be that the unstructured data that come from Google searches, the structured data contained in relational databases, or sophisticated statistical results. All of this can only act to improve the quality of journalism at a time when traditional journalism is under great strain.

About the Reviewer

Fred Vallance-Jones is an Associate Professor in the School of Journalism at the University of King's College, with over 25 years of experience as a professional journalist. He specializes in investigative and data journalism. He has won and been nominated for many national and

international awards, as have his investigative reporting students. He has conducted several national studies on freedom of information for Newspapers Canada, which represents the country's daily newspapers. He has recently authored a chapter in *Brokering Access: Power, Politics, and Freedom of Information Process in Canada*.

Citing this review article:

Vallance-Jones, Fred. (2013). Making journalism better by understanding data [Review of the three books *The data journalism handbook*, *The filter bubble*, and *PDQ statistics*]. *Global Media Journal -- Canadian Edition*, *6*(1), 67-72.