BOOK REVIEW / CRITIQUE DE LIVRE

The Accidental Data Scientist: Big Data Applications and Opportunities for Librarians and Information Professionals. Amy Affelt 2015. 240 pp. Softbound. ISBN 978-1-57387-511-0. Regular Price USD$39.50. Available from: http://books.infotoday.com/books/Accidental-Data-Scientist.shtml#ixzz3x3c0RH00.

Data has become a hot topic in librarianship; conference meeting schedules are packed with presentations about data, articles and books are published about data at a rapidly increasing rate, and career opportunities for data librarians are becoming more common. As data becomes more popular, many librarians have been quick to respond to prove they have the skills necessary to be relevant in this area. One example of this is Amy Affelt's book *The Accidental Data Scientist.* Affelt is the Director of Database Research Worldwide at Compass Lexecon, a global economic consultancy. She writes and speaks frequently about "big data" in corporate contexts. This book takes a corporate approach to librarianship and data, explains the wide variety of tools that can be used in data science, and provides examples of working with data in industry settings to establish the librarian skill set as one well suited to this area.

As an academic health sciences librarian who provides data management education, develops data discovery tools, and collaborates on data projects with basic science and clinical researchers, I can attest that although *The Accidental Data Scientist* may be applicable to corporate librarianship, it does not translate well to an academic health sciences research setting.

Throughout the book Affelt outlines scenarios where librarians locate data for big companies and corporate firms to target advertising markets. For example, to establish librarians as "21st Century Librarians" working with data, Affelt highlights the skills of librarians as search experts, stating it is the librarian's responsibility to find external data for corporate stakeholders. For reference, external data can be defined as any public dataset that is available for free, via license, or for purchase (e.g., Canadian Census, Canadian Community Health Survey). Affelt claims it is the librarian's job to "determine which data is the best fit for a project" (p. 146), and to assess the "accuracy, consistency, reliability, completeness, timeliness, reason and validity" (p. 130) of this external data. This approach is hugely problematic in a biomedical environment. First, Affelt assumes that data analysis is a traditional skill of librarians, but these skills are really more aligned with the work of data analysts and statisticians. Second, librarians traditionally do not have the expertise to determine which datasets are most appropriate for analysis or use. This task should be left to the researchers, who are the experts in their field. To give an example from the health sciences, population health researchers use external datasets such as the census or national health surveys to identify populations and evaluate health outcomes. In this scenario, a population health researcher is far more knowledgeable about their research than a librarian would be and, therefore, more capable of making decisions about selecting relevant external data.

My concern is reaffirmed in a feature interview from the book with Kimberley Silk, a Data Librarian at the University of Toronto. Silk agrees that in her work it is best to leave the analysis and decision-making about what data to use up to the experts on her team. For health sciences librarians, we are more than capable of supporting the discovery of external data by pointing researchers to data sources such as Canada's Open Data Portal or building discovery tools for specific communities to find external datasets [1], but it is not our place to analyze data or decide what data is valuable, unless we have the subject expertise and authority to do so.

Due to the corporate nature of the book, Affelt also omits the topic of research data—data that is created, collected, or observed to produce original research results [2]. In biomedical research, research data can take the form of biospecimens, video recordings, images, software programs, algorithms, and even paper lab notebooks. It is during the collection of these types of data that health sciences librarians can play a significant role in managing that data. The omission of data management in *The Accidental Data Scientist* is surprising considering Affelt's goal is to prove the librarian's skill set is suited to data-specific professions. Librarians play a key role in information management on a regular basis; we make information discoverable, accessible, and understandable. This role is no different when it comes to managing research data and making it available to others. Our expertise in organizing information, assigning metadata, and providing access to information can all be applied to research data. To learn more about the role of libraries and research data management in the context of the health sciences, I recommend two valuable resources: *Research data management* [3] and *Research data management and the health sciences librarian* [4]; they provide a high-level overview of the role librarians can play in data management and highlight common issues biomedical researchers face when collecting data.

Finally, Affelt's liberal use of the buzzword "big data" is something I take issue with. The world of data is plagued with jargon that serves to obscure the discipline, rather than clarify it. Affelt acknowledges this fact, yet perpetuates the confusion by discussing big data's "many definitions" using examples from a historical search of the phrase in mainstream media. Examples of mainstream media's interpretations of big data include everything from social media data to smartphone data to financial market data. These examples are problematic because they do not explain why these types of data are considered big. Social media and smartphone data, for example, are not defined by their

"bigness"—this data can also be collected and analyzed in small quantities. My problem with Affelt's approach is that she never defines big data on her own terms, instead settling on a definition supplied by the Oxford English Dictionary even after devoting an entire chapter to its many interpretations. Without a clear definition of big data, Affelt does not lay the necessary groundwork to help the reader fully understand its meaning in the book's subsequent chapters.

The point that could have been made in the *The Accidental Data Scientist* is that data can come in many sizes: big or small. In the health sciences, a researcher could collect and analyze data from thousands of patients across multiple institutions or, conversely, produce a few spreadsheets of demographic data from a small cohort of study participants. Regardless of the size and type of data in the research environment, the term big data does not add value to the conversation. Speaking from personal experience, eliminating jargon when speaking with physicians and researchers about their data is the first step toward gaining credibility with them. Health sciences librarians should treat all the data they encounter as part of their work as simply that—data—and focus on their skills as information organizers, managers, and sharers when working in a data-driven environment.

References

1. Read K, Surkis A, Lamb I, et al. Promoting data reuse and collaboration at an academic medical center. *IJDC*. 2015; 10(1):260–267. doi: 10.2218/ijdc.v10i1.366.

2. University of Edinburgh Information Services. *Research data management programme: research data management home* [Internet]. Edinburgh, UK: The University; 2014. [cited 13 Feb 2015]. Available from: http://www.ed.ac.uk/schools-departments/information-services/research-support/data-management/data-management-home

3. Surkis A, Read K. Research data management. *JMLA*. 2015; 103(3):154–6. doi: 10.3163/1536-5050.103.3.011.

4. Creamer A, Martin E, Kafel D. Research data management and the health sciences librarian. In *Health sciences librarianship*. Ed. S. Wood. Chicago, IL: Rowman & Littlefield and Medical Library Association. 2014; p. 252–274.

Kevin Read MLIS, MAS
NYU School of Medicine
577 First Ave
New York, NY 10016, USA
Email: kevin.read@med.nyu.edu