Book Review

Uncharted: Big Data as a Lens on Human Culture.
Erez Aiden and Jean–Baptiste Michel. New York, NY:
Riverhead Books, 2013. 288pp.
ISBN–13: 978–1594487453. Hardcover, $19.16.

REVIEWED BY:
Carolyn Penstein Rose, Ph.D.
Carnegie Mellon University

As pressures to scale up education and assessment mount higher and higher, attention has turned towards techniques from the field of big data analytics to provide the needed boon. At first blush, Aiden and Michel's book *Uncharted: Big Data as a Lens on Human Culture* would not seem to speak to this issue directly, yet it does provide the opportunity for some needed reflection.

As pressures to scale up education and assessment mount higher and higher, attention has turned towards techniques from the field of big data analytics to provide the needed boon.

The vision of the idealized data science of the future has recently been characterized as something akin to archeology and geology (Knight et al., 2014), two fields where scientists conduct painstaking, careful, and reflective work to reconstruct the past from the fragments that remain. This characterization of our work challenges us to take greater care as we piece together evidence of psychological and social processes from the digital remains of cognitive and social activity taking place within the online world. In particular, it challenges us to take a step beyond just counting what can be easily counted, and push for greater theoretical depth and validity in our attempts at quantification and operationalization as we seek to make sense of the signals we can uncover using the growing number of powerful modeling technologies that have been developed in recent decades.

Within this sphere, Aiden and Michel's book is a popular press treatise designed to introduce a nontechnical readership to the capabilities of the Google Ngram Viewer.[1] It presents a fascinating new look at history through the lens of "robots," which are automated lexicographers that index arbitrary lengthed word sequences, referred to as ngrams, as they occur within the expanding Google Book collection.[2] The ngram viewer makes its debut in the book by producing a graph that challenges a claim about the historical event that triggered a shift in how the "United States" is treated grammatically, i.e., whether we treat it as a plural reference to a multiplicity of states or a singular reference to a collective whole. The shift in grammatical status is purported to reflect a shift in conception, and therefore has great historical significance, especially to Americans. The evidence of such a shift in usage is a graph of relative frequency of occurrence of "The United States are" and "The United States is" over time in the Google Book collection. The shape of the displayed trend is different from what one might think if it did indeed reflect the change in conceptual status and was indeed triggered by a historical event in that, it occurred gradually rather than suddenly, and it was not until fifteen years after the event that was believed to have triggered it when the dramatic difference in preference emerged. The reader is challenged to consider the extent to which previous conceptions of history might be challenged by viewing it through the eyes of these robot lexicographers.

The Google Ngram Viewer is a text visualization tool (Siirtola, Saily, Nevalainen, & Railha, 2014). One can consider its representation of text as something of a cross between world clouds, which give a cross–sectional view of word distributions from a corpus in graphical form, and graphs of topic trends, which use dimensionality reduction techniques like Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) or Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998) to identify themes and then plot the relative prevalence of those themes over time within a corpus. Word clouds are often used to suggest the values communicated through a text or text collection by displaying words with a relative size that indicates their relative frequency, with the implication that relative frequency says something about relative value. Topic trends present a more digested view, in that they collapse together sets of words that co–occur, and therefore might function together as elements that together communicate a theme. The representation of these automatically identified themes as a graph of their relative frequency over time is displayed through line graphs arguably provides a much coarser grained perspective on what is in the text, and yet it offers the possibility of comparing topic focus between different periods of time. And its coarser grained representation better leverages the richness in stylistic variation that language affords. Like a word cloud, the Google Ngram Viewer's representation displays relative frequency of ngrams as a representation of relative value. But unlike the cross–sectional nature of a word cloud, its representation allows us to see trends over time. Similarly unlike word clouds, it is extremely selective in which relative frequencies it displays. Thus, unlike topic trend representations, it does not consider the great variation that language affords in referring to an idea, or even in realization of a specific lexical construction. A rigorous interpretation of the significance of the graphs would take these contrasts into account.

The first chapter of the book recounts the history of the development of the Google Ngram Viewer and illustrates its use with some key examples. After that, with each of the next five chapters, a new and fascinating question that might be investigated using this tool is introduced and explored. The Google Ngram Viewer is posed as the data analyst's correlate of Galileo's telescope. While the richness of the signal provided by such a viewer is admittedly impoverished, it is compared to the remnants of monetary systems of old left behind for anthropologists to use to piece together an

[1] https://books.google.com/ngrams
[2] http://books.google.com/

image of cultural practices of old. The authors pose questions about the status of theory in light of the great multitude of hypotheses that can be imagined and quickly tested with such a resource.

While the authors compare the Google Ngram Viewer to the telescope of Galileo, the book does not come across to my academic ears as designed as a serious foray into data science, nor meant to make serious contributions to the fields of humanities and social sciences. To its credit, it raises some methodological concerns even in the first chapter where the authors affirm the need to validate interpretations from quantifications and acknowledge the difficulty of doing so in a corpus as large as the Google Books archive. Thus, it would not be fair to critique it based on methodological standards of the fields of data science. Nevertheless, it is useful in the context of a special issue on learning analytics, and assessment specifically, to consider what message this book might have for us as a community as we reflect on our own practices of scientific inquiry.

Nevertheless, it is useful in the context of a special issue on learning analytics, and assessment specifically, to consider what message this book might have for us as a community as we reflect on our own practices of scientific inquiry.

Consider the following anecdote. A recent New York Magazine article reported that personnel at Pinterest had noticed a strong trend for numerous women to collect substantial numbers of pins related to weddings. The interpretation of this strong focus on weddings was that these women were most likely preparing for their respective weddings. Thus, the organization then proceeded to send an email to them with text that implied they were indeed preparing to get married. It turned out, however, that most of them were single and were collecting the pins for other reasons. Some responded in a way that suggested they were dismayed at the mistake. This anecdote illustrates well how easy it is to misinterpret what a pattern might be telling us, even when the pattern appears strong and clear. The problem is that Pinterest was not designed to provide others with insight into the reasons why people are interested in or collect the items that they do, and thus it is not valid to assume that upon viewing ones pins the viewer would get insight into these reasons.

Similarly, in the case of the Google Ngram Viewer, it is easy to imagine that while the view provided by the robots has some advantages over our own human perspective on history (e.g., perfect memory, long time view, ability to consider every word in the entire book collection, etc.), we must consider the important ways in which the view it provides might be obscured by what its missing. For example, the contrast between "The United States is" and "The United States are" neglects the fact that the great majority of mentions of the phrase do not place it as the subject of the copula, and therefore will be skipped in this analysis.

Furthermore, the contexts in which it is positioned this way are not a random sampling of mentions since this form is indicative of a definitional statement, although the grammatical treatment of the phrase in other contexts is equally a reflection of the conception of its status as an entity. It is equally important to note that books included in Google Books might not be a random sampling of published books, and the language of book publications might not be a random sampling of language produced. Furthermore, the analysis fails to take into consideration that many genres of writing include language that reflects not the style or perspective of the author, but perhaps the style or perspective of a synthetic culture created as a fictional character or culture, or the author's potentially mistaken conception of how some other would present him or herself. All of these issues and more threaten the validity of the conclusions one might draw from the graphs, no matter how compelling they might appear.

Coming back to the focus of this special issue, what does this tell us about the use of big data analytics for assessment? The book is well worth a thoughtful read by all who look to big data analytics to play a growing role in large scale assessment. It is not to say that the book should either encourage or discourage such a movement. It should simply provide the opportunity to reflect on issues related to validation of interpretation. And specifically with respect to assessment based on analysis of textual data, issues related to the incredible richness and variability of language usage should be appreciated and allowed to raise an appropriate level of skepticism.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Knight, S., Wise, A., Arastoopour, G., Shaffer, D., Shum, S. B., Kirschner, P., & Collier, W. (2014). *Learning analytics: Process vs practice*. Presentation at the at Learning Analytics for Learning and Becoming in Practice Workshop, International Conference of the Learning Sciences, Boulder, CO.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis *Discourse Processes, Special Issue: Quantitative Approaches to Semantic Knowledge Representations, 25*(2–3), 259–284.

Roy, J. (2014, September 4). Pinterest accidentally congratulates single women on getting married. *New York Magazine*. Retrieved from http://nymag.com/daily/intelligencer/2014/09/pinterest–congratulates–single–women–on–marriage.html

Siirtola, H., Saily, T., Nevalainen, T., & Railha, K.–J. (2014). Text Variation Explorer: Towards interactive visualization tools for corpus linguistics. *International Journal of Corpus Linguistics, 19*(3), 417–429.