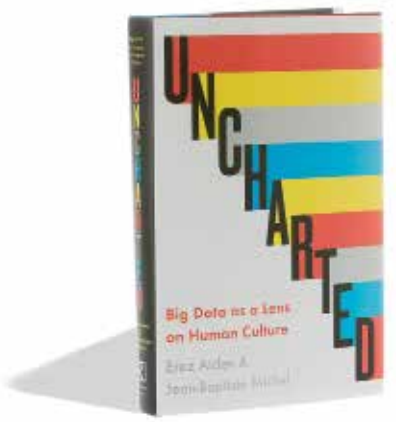
A Play on Words



Erez Aiden and Jean-Baptiste Michel

Uncharted

Big Data as a Lens on Human Culture

Riverhead Books, New York, 2013, 288 pp., $27.95 (cloth).

There are many things that are uncharted in this book. But Erez Aiden and Jean-Baptiste Michel don't mean by "uncharted" that things are left out—in fact, a more appropriate title might be "Charted."

The book tells the story of collecting the billions of words in all the world's books, words that were previously lost in the meaning of the text—"uncharted" as it were—but can now be charted to our heart's content. The authors hope that the charting process will lead us to discover interesting aspects of our culture, which they refer to as "culturomics." Aiden and Michel collaborated with Google to make a powerful Web tool, but their claims about its usefulness are perhaps extravagant.

This is not to say that the book isn't fun. That's what you'd expect from the acknowledgments, in which Aiden thanks his three children and includes the middle name of a daughter: Banana. (At least he's quirkily consistent; his son is Galileo.)

Now I'm all for fun. But Aiden and Michel are doing important scientific work, and they don't do themselves any favors by giving "fun" examples. It doesn't take big data to convince us that the word chupacabra (a blood-drinking creature reportedly sighted in Puerto Rico in 1995) is much rarer than Sasquatch or the Loch Ness Monster. It also seems silly to chart the changing usage of "argh" and "aargh" in books published sometime between the 1940s (it's hard to tell the starting date from the chart reproduced in the book) and 2000. There's a quote on the book jacket from *Mother Jones* that calls the Ngram Viewer "the greatest timewaster in the history of the Internet." It was bold of the publisher to include that.

To document a cultural history by getting robots to read every word of every book ever published is ambitious. So what do I mean by saying there's a lot left out of this effort? Aiden and Michel acknowledge that they are searching through a tiny sample of words, and although they say that Google has so far scanned some 30 million books (probably more by now), there are still some 100 million to go.

Further, if a word's usage is a clue to our cultural history, many sources are ignored in this book: newspaper and magazine articles, letters, movies, TV and radio interviews, transcripts, lectures—in fact, everything written or spoken, but not published in a book. Besides, after books are written, they are often edited and revised for grammar and spelling, not to mention translated into other languages. Every author knows that editors change the text according to their publisher's house style. I wonder if the language in books, even 30 million of them, is a reliable source of changing language usage.

I expect Aiden and Michel would argue that the books Google has scanned are all they have to work with, but given what's missing, their "lens on human culture" theory may be too bold a claim.

About the charts: there are many, and they are stripped down; printed in black, white, and gray; and generated directly from the data. There is nothing wrong with simple charts, but the relative lack of labels and grid lines, and the chart lines themselves (sometimes as many as six) in minimally differentiated shades of gray, make for difficult reading. The authors point us to the Web, where all these problems are taken care of: colors differentiate the lines, and clicking on them at any point reveals a label and date. It's an example of the distance between print and Web-based graphics.

The book tells the story of collecting the billions of words in all the world's books.

But let's be positive. The authors bravely do not dodge copyright questions raised when books are scanned or seemingly unethical "shadow" ways to get around those questions. There is lovely detail about a 2002 experiment by Larry Page and Marissa Mayer, who worked out how long it might take to scan all the world's books. Apparently it would take "millennia, even eons." So how did the authors get around that problem? Read the book; you'll have fun!

Nigel Holmes
Principal, Explanation Graphics, author, most recently of Wordless Diagrams *and* The Book of Everything