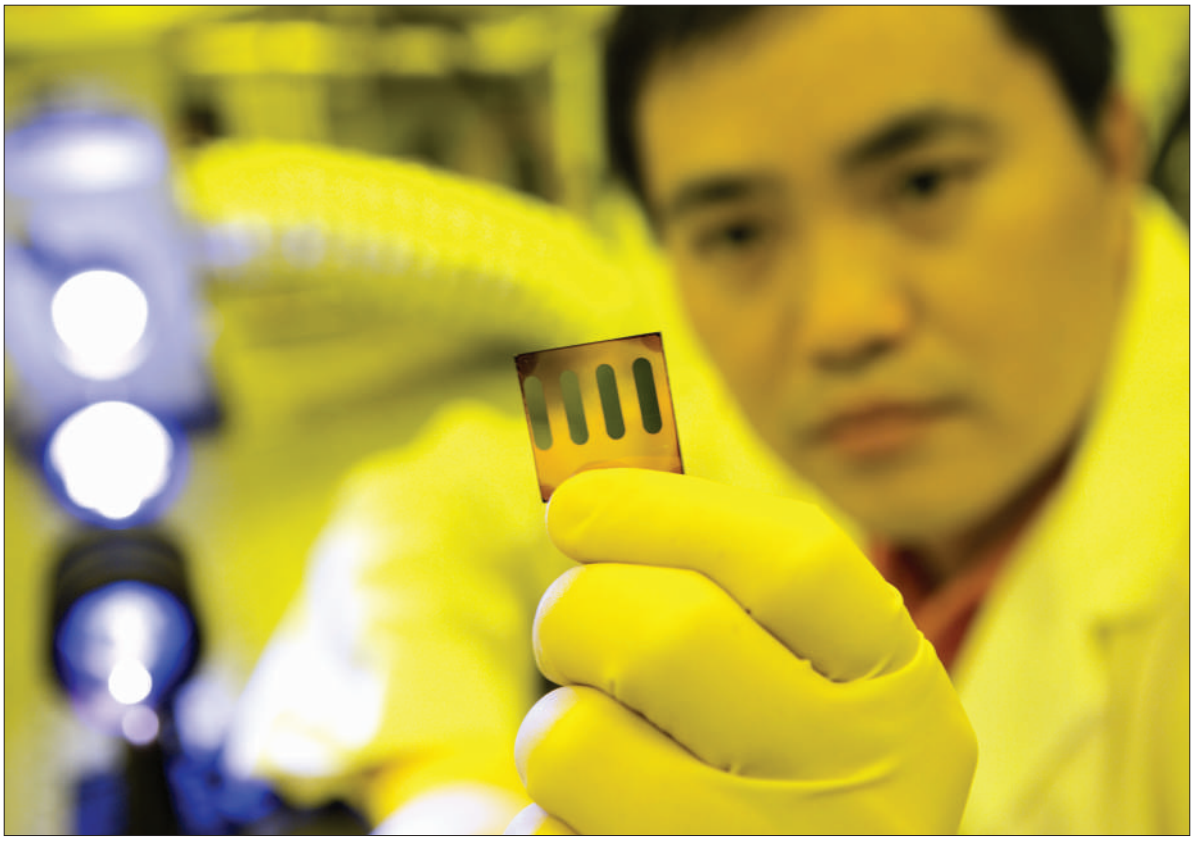

Singapore Journal of

Scientific Research

ISSN: 2010-006x

science alert

http://scialert.net/sjsr

Using Digitized Books as 'Cultural Genome,' Researchers Unveil Quantitative Approach to Humanities

Researchers have created a powerful new approach to scholarship, using approximately 4 percent of all books ever published as a digital "fossil record" of human culture. By tracking the frequency with which words appear in books over time, scholars can now precisely quantify a wide variety of cultural and historical trends.

The four-year effort, led by Harvard University's Jean-Baptiste Michel and Erez Lieberman Aiden, is described this week in the journal Science.

The team, comprising researchers from Harvard, Google, Encyclopaedia Britannica, and the American Heritage Dictionary, has already used their approach -- dubbed "culturomics," by analogy with genomics -- to gain insight into topics as diverse as humanity's collective memory, the adoption of technology, the dynamics of fame, and the effects of censorship and propaganda.

"Interest in computational approaches to the humanities and social sciences dates to the 1950s," says Michel, a postdoctoral researcher based in Harvard's Department of Psychology and Program for Evolutionary Dynamics. "But attempts to introduce quantitative methods into the study of culture have been hampered by the lack of suitable data. We now have a massive dataset, available through an interface that is user-friendly and freely available to anyone."

Google will release a new online tool to accompany the paper: a simple interface that enables users to type in a word or phrase and immediately see how its usage frequency has changed over the past few centuries.

"Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena in the social sciences and humanities," says Aiden, a junior fellow in Harvard's Society of Fellows and principal investigator of the Laboratory-at-Large, part of Harvard's School of Engineering and Applied Sciences. "While browsing this cultural record is fascinating for anyone interested in what's mattered to people over time, we hope that scholars of the humanities and social sciences will find this to be a useful and powerful tool."

This dataset, which is available for download, is thousands of times larger than any previous historical corpus. It is based on the full text of about 5.2 million books, with more than 500 billion words in total. About 72 percent of its text is in English, with smaller amounts in French, Spanish, German, Chinese, Russian, and Hebrew.

It is the largest data release in the history of the humanities, the authors note, a sequence of letters 1,000 times longer than the human genome. If written in a straight line, it would reach to the moon and back 10 times over.

"Now that a significant fraction of the world's books have been digitized, it's possible for computer-aided analysis to reveal undiscovered trends in history, culture, language, and thought," says Jon Orwant, engineering manager for Google Books.

The paper describes the development of this new approach and surveys a vast range of applications, focusing on the past two centuries. The team's findings include:

* Some 8,500 new words enter the English language annually, fueling a 70 percent growth of the lexicon between 1950 and 2000. But many of these million-plus words can't be found in dictionaries.

"We estimated that 52 percent of the English lexicon -- the majority of words used in English books -- consist of lexical 'dark matter' undocumented in standard references," the researchers write in Science.

* Humanity is forgetting its past faster with each passing year. The Harvard-Google team tracked the frequency with which each year from 1875 to 1975 appeared, finding that references to the past decrease much more rapidly now than in the 19th century. References to "1880" didn't fall by half until 1912 -- a lag of 32 years -- but references to "1973" reached half their peak just a decade later, in 1983.

* Innovations spread faster than ever. For instance, inventions from the end of the 19th century spread more than twice as fast as those from the early 1800s.

* Modern celebrities are younger and more famous than their 19th-century predecessors, but their fame is shorter-lived. Celebrities born in 1950 initially achieved fame at an average age of 29, compared to 43 for celebrities born in 1800. But their fame also disappears faster, with a "half-life" that is increasingly short.

"People are getting more famous than ever before," the researchers write, "but are being forgotten more rapidly than ever."

* The most famous actors tend to become famous earlier (around age 30) than the most famous writers (around age 40) and politicians (after age 50). But patience pays off: Top politicians end up much more famous than the best-known actors.

* Culturomics is a powerful tool for automatically identifying censorship and propaganda. For example, Jewish artist Marc Chagall was mentioned just once in the entire German corpus from 1936 to 1944, even as his prominence in English-language books grew roughly fivefold. Evidence of similar suppression is seen in Russian with regard to Leon Trotsky; in Chinese with regard to Tiananmen Square; and in the US with regard to the "Hollywood Ten," a group of entertainers blacklisted in 1947.

* "Freud" is more deeply engrained in our collective subconscious than "Galileo," "Darwin," or "Einstein."

Michel, Aiden, and Orwant's co-authors are Aviva Presser Aiden, Adrian Veres, Steven Pinker, and Martin A. Nowak at Harvard; Google's Matthew K. Gray, Dan Clancy, Peter Norvig, and the Google Books Team; Yuan Kui Shen at the Massachusetts Institute of Technology; Joseph P. Pickett, executive editor of the American Heritage Dictionary; and Dale Hoiberg, editor-in-chief of Encyclopaedia Britannica.

The work was funded by Google, a Foundational Questions in Evolutionary Biology Prize Fellowship, Harvard Medical School, the Harvard Society of Fellows, a Fannie and John Hertz Foundation Graduate Fellowship, a National Defense Science and Engineering Graduate Fellowship, a National Science Foundation Graduate Fellowship, the National Space Biomedical Research Institute, the National Human Genome Research Institute, the Templeton Foundation, the National Institutes of Health, and the Bill and Melinda Gates Foundation.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, the Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 16 December 2010 DOI: 10.1126/science.1199644