Social media for large studies of behavior

Derek Ruths & Jürgen Pfeffer

Summary

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadvertently undersampled Truman supporters (*1*). Rather than permanently discrediting the practice of polling, this event led to the development of more sophisticated techniques and higher standards that produce the more accurate and statistically rigorous polls conducted today (*3*). Now we are poised at a similar technological inflection point. Recent years have seen the rise of online personal and social data for the study of human behavior. Powerful computational resources combined with the availability of massive social media datasets have given rise to a growing body of work that uses a combination of machine learning, natural language processing, network analysis, and statistics for the measurement of population structure and human behavior at unprecedented scale. ... Here we highlight issues that are endemic to the study of human behavior through large-scale social media datasets and discuss strategies that can be used to address them. While some of the issues raised are very basic (and long-studied) in the social sciences, the new kinds of data and the entry of a variety of communities of researchers into the field make these issues worth revisiting and updating.

Big data meets public health

Muin J. Khoury & John P. A. Ioannidis

Summary

In 1854, as cholera swept through London, John Snow, the father of modern epidemiology, painstakingly recorded the locations of affected homes. After long, laborious work, he implicated the Broad Street water pump as the source of the outbreak, even without knowing that a *Vibrio* organism caused cholera. "Today, Snow might have crunched Global Positioning System [GPS] information and disease prevalence data, solving the problem within hours" (*1*). That is the potential impact of "Big Data" on the public's health. ... Separating the true signal from the gigantic amount of noise is neither easy nor straightforward, but it is a challenge that must be tackled if information is ever to be translated into societal well-being.