Data Wealth, Data Poverty, Science and Cyberinfrastructure

Steve Sawyer

This essay argues that the limits of science are increasingly set by data wealth. This is diverting attention away from the many valuable contributions of relatively data-poor sciences. This heightened attention comes at the risk of mistaking volumes of data with insight. The rapidly increasing availability of data, due primarily to new forms of digital sensing, collection and representation, is magnifying ongoing changes to the conduct and expectations of science. Data wealth and data poverty are terms which mean the volume of data available to the scholarly community. Recently, the pursuit of data wealth is paired with the development of cyberinfrastructure.

Two sets of forces are helping us to rethink the goals, if not the nature of science: globalization and computerization. In regards to globalization, science is seen increasingly as both a source of innovation and as demanding ever higher levels of intellectual and capital resources. Globalization is in part the result of a market-driven, economic model of life that permeates much of the developed west and frames much of the rhetoric for the developing east and global south. The digitization of data involves the uses of computer-based sensors and digital representations of phenomena in a form consumable by computers.

There are three examples of cyberinfrastructure being developed to handle rapid data-wealth: (1) the Physicist's Open Grid, (2) the National Center for Microscopy and Imaging Research (NCMIR), (3) Long Term Ecological Research (LTER) networks. While these examples all draw from natural science, cyberinfrastructure for various social sciences and the humanities are also emerging.

Data rich fields can be identified by three characteristics: (1) access to data sets is shared, (2) method choices are limited, (3) and only a small number of theoretical camps have been validated. Data poor fields are defined by three characteristics as well: (1) data are a prized possession, (2) the types of data available often dictate the methods taken, (3) there are many theoretical camps. Data poor fields have always existed, but the small amount of data matters more now because of the interest in many scientific fields and those who are primary sources of funding towards doing science through cyberinfrastructure.

There are four proposed actions that data-poor scholars can take. First, better connect micro-studies to macro-data. Second, focus more on theory elaboration and less on theory testing or theory borrowing. Third, draw on expertise in analyzing mixed data sets to do more multi-method research. Fourth, focus on data pooling. These are not the only four solutions, however they reflect the issues with data poverty and the opportunities to leverage skills honed by many data-poor scholars to date.