Fundamentals of Big Data for CyberGIS
Michael F. Goodchild
Emeritus Professor of Geography, University of California, Santa Barbara

Keywords: Big Data, new media, data quality, conflation, volunteered geographic information

Massive data volume is one of the three frequently mentioned Vs, the defining characteristics of Big Data, and thus Big Data is often cited as a driver for CyberGIS. It is important, however, that students recognize the other characteristics, notably velocity and variety, and understand their implications for computation. Moreover Big Data is also often associated with highly variable and sometimes questionable quality, and thus with a fourth V, either validity or veracity. Big Data is also raising numerous issues of a fundamental nature, and stimulating much discussion about changes in the nature of science. If we believe that Big Data is intimately associated with CyberGIS, then I believe it follows that a CyberGIS curriculum must include the topics discussed in detail in what follows.

The defining characteristics of Big Data

Volume: What techniques have been used in the past to work around the problems caused by massive data volume, for example in remote sensing?

Velocity: What new opportunities are afforded by data that are available in close-to-real time, and what does this imply about the changing nature of science? What new methods of analysis are needed?

Variety: In the world of Big Data there are almost always many sources of the same information, with different properties including quality. How can traditional methods of analysis be generalized to handle multiple sources?

Data quality

How can geospatial data quality be characterized? What metrics are available, and how can they be used to address the propagation of uncertainty through analysis? How can uncertainty be simulated?

Conflation

How can multiple sources of geospatial data be combined to produce a synthesis that is more reliable than the individual sources? Topics include fusion of raster data, conflation of vector data, and conflation of both vector and raster data. What new methods of conflation and synthesis are needed?

Volunteered geographic information

VGI, and crowdsourcing more generally, is an important source of Big Data. How can its quality be assessed and modeled?

New media

The topics discussed above are all important in applying CyberGIS to Big Data. But they do not address the long-term implications of doing so, and rather focus on the ways in which Big Data can be brought into traditional approaches. In my view the more fundamental impacts of Big Data and CyberGIS are if anything more important in a curriculum than the technical ones, especially if what is learned through the curriculum is to be useful to students in the longer term, and is to prepare them for a lifelong career.

First, Big Data is challenging the traditional role of theory, and writers about data-driven science have even suggested that theory is dead, that we should “let the data speak for themselves.” Second, the traditional limits that have been placed on publication in order that discoveries fit within the norms of journal articles or books no longer apply, and Web-based dissemination provides a very different model
in which it is possible to publish everything relevant to a topic, not just carefully distilled and synthesized results. Similarly, a project no longer ends with publication, since it is easy to edit and modify publications as new research emerges.

I think it is very important that students of CyberGIS and Big Data understand the distinct differences between geospatial data and other types of data that may well have figured in the classic articles about Big Data and data-driven science. We know that no geospatial data are perfect, that all are to some degree subject to uncertainty. This means that if we “let the data speak for themselves” we make discoveries about the data, but not necessarily about the real geographic world that the data are trying to represent. We need to be aware that our discoveries may be due to artifacts of the data, such as sampling interval or raster orientation, rather than of the real world.

Much of the literature on Big Data is concerned with successful prediction, but prediction has always occupied a somewhat contested place in science. A traditionalist would argue that successful prediction does not necessarily imply the acquisition of new, useful knowledge, and that one prediction may well not generalize to other predictions. But to take a positive view, we need to decide what is meant by prediction in a geospatial context. I would argue that geospatial prediction is concerned with predicting *where* something will happen, and possibly when. Thus far the literature in GIScience is very thin on prediction.