SHASHI SHEKHAR
Department of Computer Science and Engineering
University of Minnesota
Email: shekhar@umn.edu

Shashi Shekhar, a McKnight Distinguished University Professor at the University of Minnesota and an U.C. Berkeley alumnus, is a leading scholar of spatial computing and Geographic Information Systems (GIS). He is serving on the Computing Research Association (CRA) board, and as a co-Editor-in-Chief of *Geo-Informatica Journal* (Springer). Earlier, Shekhar served as the President of the University Consortium for GIS (UCGIS), and on many National Academies' committees. Recognitions include IEEE-CS Technical Achievement Award, UCGIS Education Award, IEEE Fellow and AAAS Fellow. Contributions include algorithms for evacuation route planning and spatial pattern (e.g., colocation, linear hotspots) mining, an Encyclopedia of GIS and a Spatial Databases textbook.

Spatial Data Science Agenda: A Perspective

Definition: Spatial Data Science[1] is a transdisciplinary field that uses scientific methods, processes, and algorithms to extract novel, useful, and non-trivial patterns and insights from spatial or spatio-temporal data.

A Historic Example: Dr. John Snow plotted 1854 Cholera locations on a street map of London (see Figure 1) and noted that the incidents were concentrated in an area around the Broad Street water pump. This led to a novel insight quite different from the prevalent Miasama theory positing that bad air caused Cholera. Formally, it provided a scientific hypothesis, linking cholera to drinking water contamination after the disease spread was checked after removal of the pump-handle. Soon, controlled laboratory experiments by Louis Pasteur's in 1860s and Robert Koch in 1880s provided strong evidence scientifically linking germs and diseases. Useful applications ranged from the Snow's 1849 recommendation that water be "filtered and boiled before it is used" to Pasteur's method to screen silk-worm eggs for infections to Joseph Lister's procedures in 1870s for medical and surgical sanitation. It also prompted cities to the create sewer systems to protect drinking water.

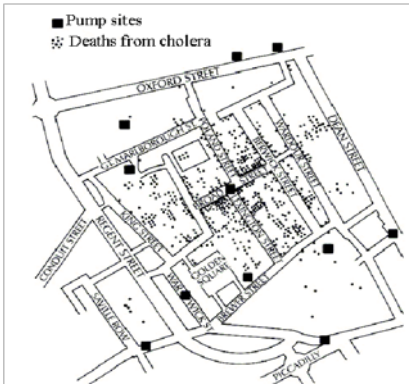

Fig. 1. Snow's 1854 London Cholera Map

Figure 2 shows the science aspects of spatial data science using four steps. John Snow collected and curated the locations of Cholera in London and discovered the hotspot pattern, i.e., high concentration of disease, in an area around Broad Street water pump yielding a

hypothesis. Subsequent work by Louis Pasteur tested the hypothesis using controlled experiments. Robert Koch developed criteria for scientifically demonstrating that a disease is caused by a particular organism bolstering the Germ Theory.

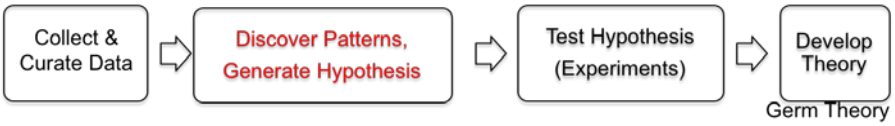

Fig. 2. Scientific Methods

What has changed? Many things have changed since the historic example described in Figure 1. First, the volume of data has increased by orders of magnitude. John Snow collected only a few hundred cases of a single disease for a small geographic area in London. Today, the Center for Disease Control collects reports of many infectious diseases from large geographic areas spanning the United States. Many other countries collect similar data and pool it together for studying diseases which may spread rapidly across country boundaries. In addition, geo-spatial data has grown tremendously with a large number of layers providing opportunities for understanding environmental impacts of a large number of factors. Beyond public health, spatial data has transformed our lives by improving monitoring of global weather and agriculture for early warning of hurricanes and inclement weather as well as food shortage risks due to crop stresses or failures. Further, with 2 billion receivers in use for location and time services, the GPS has become a critical infrastructure for the world economy for use cases ranging from precision agriculture to navigation to ride sharing to smart cities.

John Snow made the map by hand in 1854. However, today map makers increasing use computer hardware and software to make maps. Popular software such as ESRI ArcGIS provide tools to quickly create many types of maps after querying and summarizing relevant subsets of spatial data from spatial database management systems such as Oracle Spatial which provide common spatial data types (e.g., points, line-strings, polygons) as well as spatial indexing methods (e.g., R-tree). In addition, discipline of spatial statistics is maturing providing new methods to deal with unique geospatial challenges such as spatial auto-correlation, spatial variability, edge effects, etc. There are insights about modifiable areal unit problem suggesting that results of *A Position Paper for the 2019 Spatial Data Science Symposium (Setting the Spatial Data Science Agenda)* many statistical methods may be altered by choice of spatial partition boundary. The computing eco-system (e.g., Hadoop, SpatialHadoop, GPGPUs) has evolved to provide substantial compute power to support increased computational needs of spatial statistical methods such as algorithms to estimate parameters of spatial auto-regression models or to reduce spurious chance patterns leveraging Monte Carlo simulations.

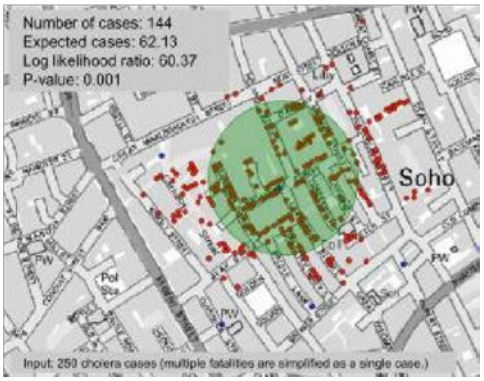

Fig. 3. SatScan output for London Cholera Data

Spatial statistical software, e.g., SatScan[2], is available to quickly identify (circular) disease hotspots from very large data sets while filtering out chance patterns. Further, such spatial data science is increasing being used in newer societal applications such as crime hotspot detection. As a result, a large number of disciplines ranging from archaeology to environmental sciences to public health to social sciences are encouraging their scholars and practitioners to spatial data science methods.

Gap Analysis: The teaching and research in spatial data science is scattered across numerous disciplines, departments and colleges, which use different names such as Remote Sensing, Spatial Statistics, Spatial Databases, Spatio-temporal Informatics, Spatial Data Mining, Geo-Statistics, Geomatics, GIS, etc. While such slicing and dicing works for insiders, it makes it very difficult for broader audience from outside our field to access spatial data science knowledge. This is unfortunate given the recent explosion in data science jobs as well as degrees on university campuses leading to hundreds of graduate degrees and dozens of undergraduate degrees attracting a very large number of students and employers. Lack of a cohesive view of spatial data science for the broad audience of data science practitioners, students and scholars means lost opportunities for spatial data science to reach out to new audience and attract new resources.

What is needed? The spatial data science community needs to come together to address the current opportunity in context of the booming area of data science. One may consider the recent statement[3] from the University Consortium for Geographic Information Science as a first step in this direction. This statement has made data science leaders aware of the need to include spatial data science content in their degrees. However, many say that they do not have resources (e.g., spatial data science faculty members, courses, curricula) towards this. Thus, it will be timely for the spatial data science community to explore ways to address this urgent need. It may start with design of spatial data science courses and curricula at multiple levels. Perhaps, it may include a spatial data science 101 course for all data science students to appreciate the special nature of spatial data and be aware of commonly available methods and tools. At the next, level, one may design a small number of courses for more interested data science students to either take a minor or specialize in spatial data science area or sub-areas. It will also help to create pedagogical material to support such courses at increasing number of spatial data science programs. Such material may include lecture outlines, slides, lecture notes, and textbooks. It may also include educator training workshops and activities. We also need to identify ways to help educational institutions without spatial data science expertise in the short run by making spatial data science courses available and in the long run by helping them identify and hire suitable faculty candidates.

This educational effort should be complemented by engagement in early stages of research to bring resources to advance spatial data science research. For example, many countries are currently starting research initiatives around Artificial Intelligence (AI) as illustrated by recent roadmap for AI research in US[4]. It is important to articulate the spatial data science research topics which are critical for success of AI.

References

[1] Y. Xie, E. Eftelioglu, R. Ali, X. Tang, Y. Li, R. Doshi, and S. Shekhar, Transdisciplinary Foundations of Geospatial Data Science. *ISPRS Int. J. Geo-Inf*. 2017, *6*: 395.

[2] SatScan: Software for the spatial, temporal and space-time scan statistics, https://www.satscan.org/.

[3] University Consortium for Geographic Information Science, *A UCGIS Call to Action: Bringing the Geospatial Perspective to Data Science Degrees and Curricula*. Summer 2018.

[4] *A 20-Year Community Roadmap for AI Research in the US,* Association for the Advancement of Artificial Intelligence and Computing Community Consortium, Aug. 7th, 2019.