CIF21 DIBBs: Scalable Capabilities for Spatial Data Synthesis (NSF 1443080)

PI: Shaowen Wang; Co-PIs: Kate Keahey and Anand Padmanabhan

Massive spatial data collected from numerous sources are increasingly used to understand our natural, human, and social systems at unprecedented scales while providing tremendous opportunities to gain dynamic insight into complex phenomena. Though such big data streams play crucial roles in many scientific domains and promise to enable a wide range of decision-making practices with significant societal impacts, exploiting them successfully poses significant challenges. On one hand, spatial and location attributes serve as a common key to many types of data such as census and population, land use and cover, flood plain, and vegetation distribution. Oftentimes perceived as significant benefits, spatial data synthesis can be used to link disparate pieces of data that pertain to common spatial references and units. On the other hand, however, there are diverse spatial references and units for data collection and management and they are based on different representation models and assumptions. Furthermore, the quality, validity, and applicability of spatial data synthesis are dependent on scalable and timely discovery, correlation, and transformation of various sources of data. To tackle these challenges, this project has created a suite of scalable capabilities for spatial data synthesis – resolving both data aggregation and integration problems – enabled by innovative cloud computing and cyberGIS (aka geospatial information science and systems (GIS) based on advanced computing and cyberinfrastructure) and driven by multiple scientific communities.

CyberGIS has recently emerged as a vibrant interdisciplinary field, and is advanced in this project to enable a large number of users to benefit from innovative capabilities for scalable spatial data synthesis through novel cloud computing infrastructure and strategies. Such capabilities are designed to support integration with cyberGIS analytics and workflow. The project has established a set of core capabilities through a spiral approach by initially developing the capabilities for solving specific scientific problems and later moving on to engage broader communities for validating and improving the core capabilities. Current scientific problems revolve around two interrelated cyberGIS applications: 1) TopoLens for making extraction, processing and visualization of national elevation data easily accessible to broad scientific communities; and 2) UrbanFlow, an interactive online cyberGIS environment that facilitates exploration and examination of population dynamics.

Intellectual Merit: The project has made solid progress on creating novel and scalable capabilities for spatial data synthesis enabled by cloud computing and cyberGIS. These capabilities are established on innovative data models and transformation processes for taking into account quality and uncertainty across diverse and massive data sources. Scientists spanning bio, computational, engineering, geo, and social sciences can access an open cyberGIS environment to aggregate and integrate distributed spatial data sources with rigorous scientific principles applied. Massive datasets are handled by coupling the synthesis capabilities with cutting-edge cyberGIS approaches while novel cloud computing solutions are developed to deliver desirable quality of service characteristics (e.g. stable response time). Multiple scientific collaborations solving emergency management and urban problems are working actively to exploit the capabilities.

Broader Impacts: Spatial data synthesis is a widely used process across numerous scientific domains. The spatial data synthesis capabilities have significant potential value beyond scientific problem solving to support making important decisions in many real-world applications. Opportunities exist for applying the capabilities to generate data (e.g., social media and elevation data) that are of great interest to general public. Undergraduate students, and minority and underrepresented groups are engaged in our research while learning materials derived from the research activities are openly accessible through the CyberGIS Gateway. Several universities have used the capabilities for teaching courses related to such topics as GIS and hydrology. A summer school and a series of training workshops have been offered to benefit over 300 attendees from broad scientific communities. We plan to extend our education and outreach through targeted development of a massive open online course (MOOC).