Strategic Plan (DRAFT) for a Geospatial Software Institute

This report has been produced by the Geospatial Software Institute conceptualization project (https://gsi.cigi.illinois.edu) and supported by the National Science Foundation under grant number: OAC-1743184. Any opinions, findings, conclusions or recommendations expressed in the report are those of the project participants and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

Summary

Many scientific and societal grand challenges (e.g., emergency management, environmental sustainability, population growth, and rapid urbanization) are inherently geospatial as articulated in a number of visionary reports such as the NSF's Ten Big Ideas and the United Nations Sustainable Development Goals. A variety of fields (e.g., environmental engineering and sciences, geosciences, and social sciences) are increasingly dependent on geospatial software to tackle these challenges. Critical and urgent efforts are also needed to prepare the next-generation workforce for computation- and/or data-intensive geospatial-related research and education, technological innovation, and real-world problem solving and decision making. In response, we have engaged diverse communities that develop and use geospatial concepts and software for conceptualizing a national Geospatial Software Institute (GSI). The mission of the GSI should be to transform geospatial software, cyberinfrastructure (CI), and data science across many fields to revolutionize diverse discovery and innovation by enhancing computational transparency and reproducibility. Its vision is a sustainable social and technical ecosystem to enable geospatial-inspired innovation and discovery. Overall, GSI is well-positioned to revolutionize many science domains while nurturing a high-performance, open, and sustainable geospatial software ecosystem across academia, government, and industry.

GSI should position itself to resolve extensive social and technical barriers to diverse communities by providing access to cutting-edge geospatial software and data science capabilities to enable significant research and education advances across many geospatial-inspired communities while making significant contributions to the national CI ecosystem. Most challenging scientific problems today require expertise from multiple domains and geospatial software plays a key role in successfully spanning domains, which is a crucial strength the institute needs to leverage. A key contribution of the GSI would be to act as a bridge to span domains where the solutions are dependent on and useful across multiple domains, not just one. The formation of a GSI should usher in a new era by unleashing the power of geospatial software to serve diverse science communities and numerous users by empowering them to make important contributions to (1) their own respective scientific disciplines, (2) cross-disciplinary challenges, and (3) innovation and sustainability of advanced CI, including future geospatial software. To achieve such significant and broad impacts, GSI should be organized across five focus areas: (a) Science Drivers; (b) Core Capabilities and Services; (c) Education and Workforce Development; (d) Evaluation and Sustainability; and (e) Outreach and Partnerships. These areas form a pipeline in which GSI can dynamically gather and articulate community needs; advance geospatial software and mobilize the communities to establish best practices and common standards for developing and using geospatial software and services; implement programs to develop and disseminate structural guidance of computational reproducibility; establish and evolve GSI's success metrics and pursue sustainability of its efforts; and develop connections and partnerships to achieve broader impacts.

GSI promises to become a transformative geospatial fabric to enable major scientific breakthroughs across a number of disciplines. GSI needs to represent diverse academic, governmental, and industrial institutions as well as international partners. The institute should identify new capabilities and knowledge that will have far-reaching impacts on solving real-world problems and making geospatial decisions in numerous contexts, including those with significant impacts on the nation’s economic development and security. GSI needs to strategically build upon external relationships in the museum and informal science community education to raise public awareness about geospatial software and related research and discoveries.

Contributors

Shaowen Wang, Donna Cox, Coline Dony, Ned English, Michael F. Goodchild, Daniel S. Katz, Praveen Kumar, Paul Morin, Anand Padmanabhan, Margaret Palmer, George Percivall, Mohan Ramamurthy, Eric Shook, Mary Shelley, Diana Sinton, Carol Song, David Tarboton, Victoria Stodden, E. Lynn Usery, Nancy Wilkins-Diehr, et al.

1. Overview

Geospatially heterogeneous and interdependent changes across the globe, such as emergency management, environmental sustainability, food security, population growth, and rapid urbanization pose many grand scientific and societal challenges [1-9]. Tackling these challenges, as a geospatial data deluge permeates broad scientific and societal realms, requires critical thinking about the complex interactions between their driving processes and related geospatial patterns across a number of spatial and temporal scales [10-14]. Geospatial software plays a critical role in examining and understanding such interactions and has been widely developed and used by numerous communities to transform data with geo and spatial references into valuable insights and significant scientific knowledge. The growing benefits and importance of geospatial software to science and engineering are driven by tremendous needs in numerous fields such as agriculture, ecology, environmental engineering and sciences, geography and spatial sciences, geosciences, national security, public health, and social sciences, to name just a few, and these are reflected by a massive digital geospatial industry [15-25].

In this context, the U.S. National Science Foundation (NSF) has funded a project (https://gsi.cigi.illinois.edu) to conceptualize a Geospatial Software Institute (GSI) that aims to create bridges across many geospatial-inspired domains by establishing a long-term hub of excellence in geospatial software to serve diverse research and education communities. GSI is also well-positioned to help drastically improve the landscape of scientific software by assuring quality, accessibility, sustainability, re-usability, transparency, and reproducibility, as well as reducing duplication of development efforts. Input from communities of geospatial software users and developers is critical to conceptualizing such an institute. This conceptualization project has engaged such communities, primarily through three workshops and a community survey, which has led to the following mission, vision, and goals.

Mission

- Transform geospatial software, cyberinfrastructure (CI), and data science across many fields to revolutionize diverse discovery and innovation by enhancing computational transparency and reproducibility.

Vision

- A sustainable social and technical ecosystem to enable geospatial-inspired innovation and discovery

Goals

1. Reproducible, transparent, and scalable geospatial software: Enable researchers to harness the geospatial data revolution for discovery and innovation by combining geospatial software and data at scale, in reproducible and transparent ways.
2. Geospatial digital workforce: Increase the nation's workforce capability and capacity to utilize geospatial big data and software for knowledge discovery supported by critical spatial thinking, and to further innovate geospatial software and advance related sciences.

3. Ethical and open geospatial software: Promote a culture of ethical and open geospatial software driven by diverse communities.
4. Structured guidance for computational reproducibility: Establish structured guidance for computational reproducibility in scientific research and education that are dependent on geospatial software.
5. High-performance and data-intensive geospatial software: Further the convergence of high-performance geospatial software with advancements in data-intensive and high-performance computing.

The mission, vision and goals of GSI emphasize the critical roles of geospatial software for tackling numerous computation- and data-intensive scientific and societal challenges by integrating domain scientific knowledge, cyberGIS and geospatial data science, and advanced computing and CI through critical spatial thinking and extreme digital transformation (Figure 1).

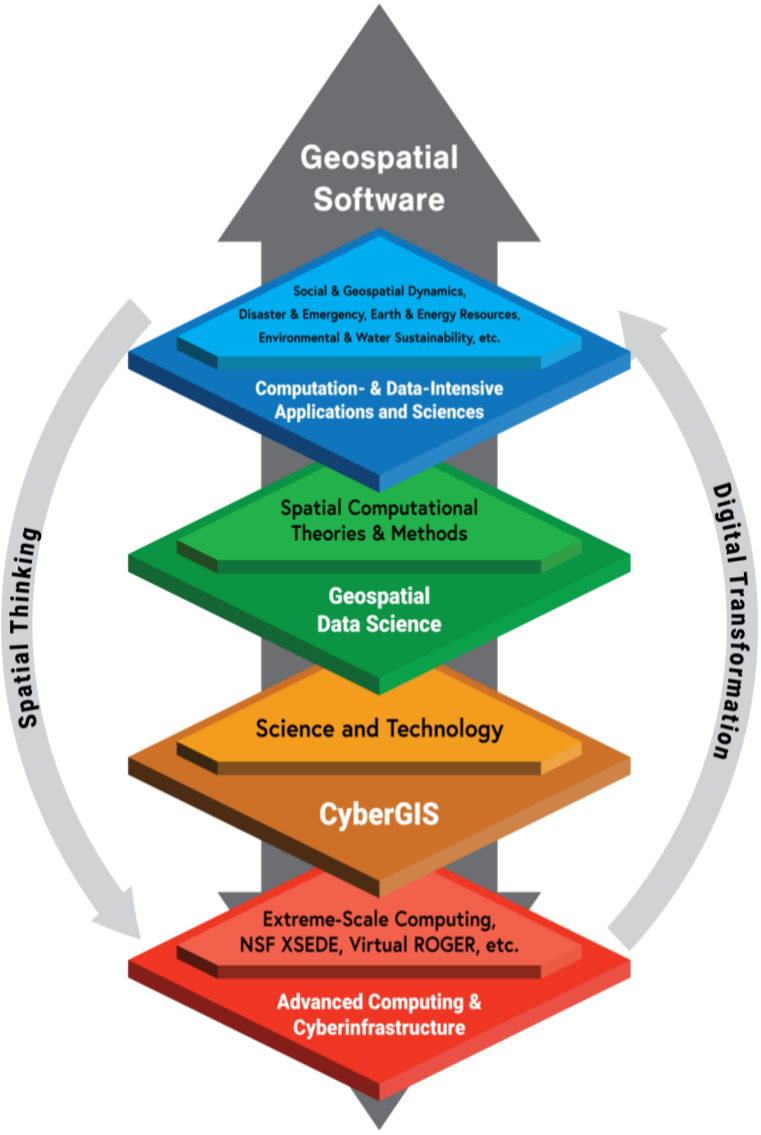

Figure 1: The context of GSI's mission, vision, and goals

2. Science Drivers

The science drivers and related communities that GSI is well-positioned to serve can be classified into two complementary types – geospatial-centric and domain-centric. Geospatial-centric research is driven by geospatial problems or questions while domain-centric research is motivated by problems and questions in specific domains and enabled by geospatial approaches. Both types of research are dependent on geospatial software, and often interdisciplinary and transdisciplinary in nature, requiring collaborative work among researchers from multiple fields. Geospatial software and associated infrastructure provide common ground and support for facilitating and enabling such research. In this context, what's common between geospatial-centric and domain-centric research is the increasing convergence among analysis, modeling, and observation approaches that can be facilitated and enabled by geospatial software to advance a variety of scientific frontiers and solutions to geospatial-related grand challenges (Figure 2). GSI promises to become a transformative geospatial fabric to synergize between the two types of research for major scientific breakthroughs across a number of disciplines.

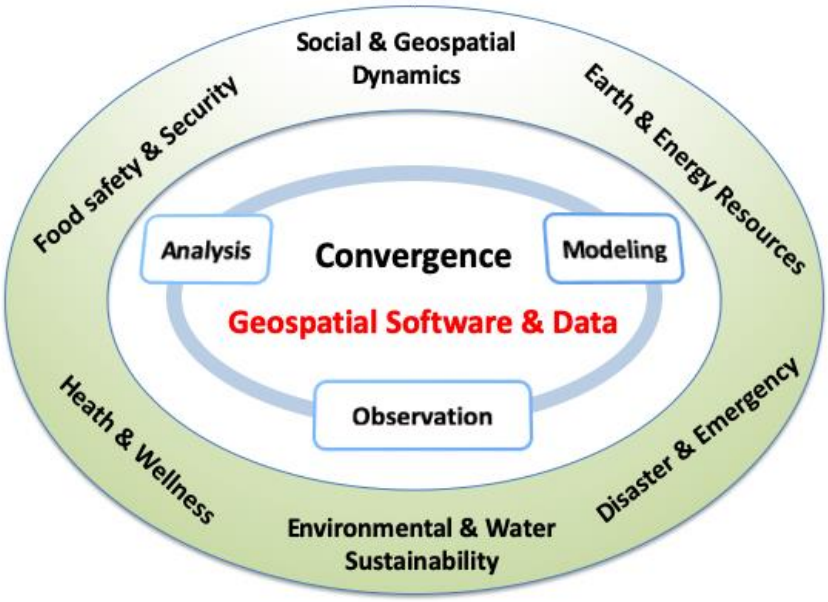

Figure 2: Science drivers

The GSI conceptualization process (detailed in Section 3) has identified a suite of scientific drivers and related communities that span hundreds of institutions of diverse types (e.g., academic, industrial, and governmental) and are expected to both contribute to the development of and greatly benefit from a GSI. Over the past decade, NSF and other federal agencies have made investments in geospatial software and data infrastructure, some of which have been identified below. A GSI should leverage such investments and strive to achieve convergence to enable significant scientific advances. While these drivers need to be the initial targets for a GSI to serve, a mechanism and structure (described in Section 7) for identifying and selecting other drivers have been formulated to dynamically engage and serve diverse geospatial-related research and education communities.

2.1 Geospatial-Centric

CyberGIS. CyberGIS represents an interdisciplinary field combining advanced computing and CI, geographic information science and systems (GIS), spatial analysis and modeling, and many geospatial domains such as emergency management, smart cities, and the food, energy and water nexus, to enable broad scientific and technological advances [1, 10, 26, 27]. CyberGIS has emerged as new-generation GIS based on holistic integration of high-performance and distributed computing, data-driven knowledge discovery, visualization, and visual analytics, and collaborative problem-solving and decision-making capabilities [13]. CyberGIS software established by the NSF software infrastructure program includes three interrelated pillars: 1) the CyberGIS Gateway for a large number of users to access online cyberGIS analytics and services, 2) the CyberGIS Toolkit for distributing open-source and scalable modules, and 3) GISolve middleware for integrating CI and cyberGIS capabilities [26, 28, 29]. The CyberGIS Gateway provides transparent access to advanced CI resources including, for example, the NSF XSEDE and various clouds. Through friendly user interfaces, the Gateway makes cyberGIS capabilities accessible to a large number of users for various research and education purposes. The CyberGIS Toolkit, on the other hand, is targeted at advanced users and provides access to scalable geospatial software capabilities within advanced CI environments. It is composed of a set of loosely coupled components to focus on exploiting high-end computing resources. GISolve is the leading spatial middleware that integrates advanced computing and information infrastructure with GIS capabilities for computationally intensive and collaborative geospatial problem solving.

While cyberGIS fulfills an essential role in enabling computation- and data-intensive research and education across a broad swath of disciplines leading to widespread scientific advances and broad societal impacts [26], GSI is urgently needed to tackle the following three challenges that remain to be daunting: 1) integration of geospatial big data, 2) computational reproducibility of geospatial workflows, and 3) adaptability to evolving CI. The integration of geospatial big data is a common need for data-intensive knowledge discovery; support for computational reproducibility is critical to the credibility of geospatial-inspired scientific discovery and innovation that often enable real-world problem solving and decision making; and adaptability to evolving CI is essential to taking advantage of and contributing to the frontiers of advanced CI. One of such frontiers is interoperability between geospatial software platforms such as GABBs based on HUBzero.

GABBs/HUBzero - A broad range of research domains is represented by the HUBzero science gateway framework [30], including 60 known gateway resources (hubs) to date. The NSF funded DIBBs project – GABBs added general-purpose, geospatial data building blocks to the widely adopted HUBzero, making geospatial software capabilities natively available on an end-to-end platform for broader communities [31]. GABBs has simplified the management, processing, and visualization of some of the most common types of geospatial data and led to rapid development of containerized web-based scientific tools. The recently funded NSF CSSI project, GABBs 2.0, is developing a plug-n-play extensible geospatial data framework (GeoEDF) to help make the existing, large, valuable geospatial datasets (such as those provided

by NASA, USGS, IPUMS) usable directly in scientific applications and tools. These efforts all work to provide seamless connections among platforms, data and tools, leading to FAIR (Findable, Accessible, Interoperable, Reusable) compliance in research practices. The open source GABBs software is by design intended to lower the barrier for a wide range of user needs. The GABBs-enabled MyGeoHub [32] geospatial science gateway has seen a rapid growth of its user base, currently serving more than 11,000 users annually. It has transformed research practices in domains that traditionally have not used geospatial software or followed the FAIR principles, e.g., agricultural economics, digital agriculture to name a few. In the digital agriculture domain, MyGeoHub is supporting cyber-physical systems of field devices and advanced CI by seamlessly connecting real-time data collected from low-cost handheld devices with analytical models and visualization tools in advanced CI, thus reducing time to results for digital and precision agriculture research. In agricultural economics and related global food-water-energy sustainability research, MyGeoHub has helped transform the traditional black-box economic modeling approach into a more transparent resource for international trade analysis and formulating sustainable development strategies. Decision tools that utilize reproducible workflows transforming climate information and data into actionable outputs are helping crop growers to deal with climate variability, and engaging policymakers and organizations involved in meeting sustainable development goals. These cross-domain, multi-scale data-driven interactive modeling workflows are demonstrative of the "long tail" needs for geospatial software capabilities and in particular seamless interoperability among software, tools and data, where GSI is well positioned to enable.

The National Map. USGS advances science about the water, energy, minerals, natural resources, and natural hazards that threaten lives and livelihoods on which we rely; the health of our ecosystems and environment; and the impacts of environmental changes. The USGS's National Map provides geospatial data to support USGS science and the 3D Elevation Program (3DEP, https://www.usgs.gov/core-science-systems/ngp/3dep), initiated in 2015, acquiring, processing, distributing, and archiving high-resolution LiDAR data that support elevation, hydrographic feature extraction, and other mapping and science modeling activities. Computation- and data-intensive geospatial analytics, such as map projection and transformation equations, require computational methods that can handle large data volumes. One of the scientific challenges of USGS includes the computational tractability of 3DEP data transformation. For example, hydrologic modeling of regional or national extent, which uses an elevation surface generated from 3DEP, can only be implemented in realistic timeframes on high-performance computing systems. To address the computational requirements needed for USGS science, researchers are using open-source geospatial solutions, such as CyberGIS-TauDEM, in parallel computing environments ([26]). USGS has internally developed a parallel computing system specifically to enable geospatial research and 3DEP LiDAR data processing and converted LiDAR processing software to the high-performance computing environment. USGS is also investigating scalable geospatial libraries and distributed computation to support various science activities, such as hydrographic feature extraction and national flood modeling. GSI offers an ideal pathway to enable these research and development efforts for the USGS-related science communities.

The Open Geospatial Consortium (OGC). OGC is an international consortium of more than 530 businesses, government agencies, research organizations, and universities driven to make geospatial (location) information and services FAIR - Findable, Accessible, Interoperable, and Reusable. Currently 99 universities are members of OGC. OGC's consensus process: creates geospatial standards; anticipates emerging tech trends; and conducts an agile research and development lab. Geospatial-inspired sciences are engaged across many of the OGC Standards Program Working Groups (e.g., Earth System Science, Hydrology, Geosciences, etc.). The University Working Group is a forum led by universities and includes discussion of curricula informed by open standards. The OGC Innovation Program initiatives advance the state of geospatial software based on interoperability testing, standards development, applications and demonstrations. Collaboration of GSI and OGC centered on geospatial standards and innovation is well positioned to advance a sustainable social and technical ecosystem to enable geospatial-inspired innovation and discovery.

2.2 Domain-Centric

Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI, https://www.cuahsi.org) and HydroShare Data and Model Sharing System. The CUAHSI was established to develop shared infrastructure, including CI, which enables the hydrologic science community to investigate the behavior and effects of water in large and complex environmental systems. HydroShare [33-36] was created as CUAHSI CI to advance hydrologic science through collaborative data and model sharing, and publication. With HydroShare, users can share, manage, access, visualize, analyze, manipulate, and publish hydrologic data and models with citable digital object identifiers (DOIs) as well as use web services Application Program Interfaces (APIs) to program automated and client access. The grand challenges that hydrology faces are by nature integrative, and require the integration of large datasets and models, which could pose serious challenges for computational reproducibility. Much of this data is geospatial. HydroShare uses the architecture of components that are loosely coupled, interact through APIs, and support extensibility in that developers can create geospatial applications through interactions with resources stored in HydroShare. A particularly powerful use of this architecture paradigm is, for example, to compute the height above nearest drainage (HAND) for large-scale flood inundation mapping [37] enabled by cutting-edge cyberGIS software and infrastructure and the USGS National Map. The hydrology community is thus poised to take advantage of GSI and contribute to achieving desirable computational reproducibility.

Polar Geospatial Center (PGC). The PGC at the University of Minnesota has pioneered providing access to multi-petabyte, high-resolution imagery from the National Geospatial-Intelligence Agency to the NSF polar science community. In addition, PGC has worked with collaborators to produce a White House directed publicly available time-dependent 2-meter posting digital surface model of the Arctic and 8m posting terrain model of the Antarctic with a spherical accuracy of ~1m using almost 1 billion core hours on the Blue Waters supercomputer at the National Center for Supercomputing Applications (NCSA). This dataset and associated

computational challenges serve as a strong motivation for GSI. For example, Terrain data represents a fundamental source of Earth knowledge, and when complete, there will be nearly 200,000 -- 17x17km to 17x120km digital elevation models (DEMs) describing the Earth's poles over an 8-year period. Such data poses significant high-performance computing and geospatial software challenges particularly related to integration with related data and models.

The Institute for Social Research and Data Innovation (ISRDI). ISRDI is an interdisciplinary research institute that provides the infrastructure and services to support the work of its four centers: the Minnesota Population Center, IPUMS, the Life Course Center, and the Minnesota Research Data Center. The Minnesota Population Center is a University-wide interdisciplinary cooperative for demographic research that serves over 100 faculty members and research scientists from ten colleges and 26 departments at the University of Minnesota and nearly 200,000 researchers worldwide. In collaboration with 105 national statistical agencies, nine national archives, and three genealogical organizations, IPUMS has created the world's largest accessible database of census microdata that describes 1.4 billion individuals drawn from over 750 censuses and surveys. Many of these microdata records include geographic identifiers that can be linked to polygons. Notably, IPUMS International includes nearly 20,000 polygons representing first- and second-administrative level units in over 90 countries. IPUMS also integrates and disseminates the nation's most comprehensive database of area-level census data and electronic boundaries describing census geography from 1790 to the present. IPUMS NHGIS (Natural Historical GIS) includes 366 billion data points and 28 million map polygons describing U.S. Census geographic units ([38]). Interactions with GSI will advance scientific discovery in social sciences in at least four ways: 1) provide open rich demographic and spatial data to motivate scalable software development and reproducible results; 2) broaden workforce training and education opportunities to social sciences; 3) gain expertise and guidance in developing open data and computation standards from a long-standing global institute; and 4) act as a bridge between over 200,000 diverse social scientists worldwide and the equally vast geospatial community to advance scholarship across many domains of science.

The National Socio-Environmental Synthesis Center (SESYNC). SESYNC serves the needs of a diverse community of scholars undertaking data-driven interdisciplinary synthesis research on a wide range of socio-environmental (S-E) problems. Research teams from all around the world are supported to work collaboratively at the center's facility where SESYNC provides a range of science facilitation and support services from project inception through results dissemination [39]. Fifteen postdoctoral fellows are in residence and the center hosts nearly 1,000 visiting participants per year, with equal representation between the natural and social sciences. The center provides infrastructure, advice, and training to enable researchers to reuse existing data and software in novel ways so they can ask and answer innovative, data-driven questions and scale up their analyses. The vast majority of projects at SESYNC require geospatial data management, analysis, and visualization to accomplish their project goals, but existing resources are not adequate to meet the demand posed by the wide variety of needs across the entire S-E research community. There is a clear need in the S-E research community for geospatial software that can facilitate streamlined processing pipelines for reconciling and integrating heterogeneous types of geospatial data. Remotely sensed data products, household

surveys, ecological plot data, weather station observations, stream measurements, hydrological models, fire models, census data, and social media data are a few examples of the types of geospatial data that many teams seek to synthesize. Interactions with GSI will accelerate discovery in socio-environmental sciences in at least three ways: 1) develop robust tools to assist in geospatial S-E synthesis; 2) extend training and support for sustainable, open-source software contributions and reuse from S-E scientists; and 3) expand awareness of and ability to seize on new opportunities presented by advanced CI.

3. Conceptualization Process

The GSI conceptualization project has sought input from the broad communities of geospatial software developers and users, primarily through three workshops, one survey, and the AAG-UCGIS Summer School on Reproducible Problem Solving with CyberGIS and Geospatial Data Science (https://www.ucgis.org/summer-school-2019). The first workshop focused on connecting big data with geospatial discovery and innovation and was held in January 2018 in Los Angeles (https://gsi.cigi.illinois.edu/workshop). This workshop brought together fifty participants representing diverse research and education communities who helped identify a set of key science domains (e.g., biosciences, environmental science and engineering, hydrology, polar sciences, social sciences, and the nexus of food, energy, and water) that the potential GSI could serve together to enable convergence research and education. The second workshop, held in July 2018 in Chicago (https://gsi.cigi.illinois.edu/workshop2), focused on gaining an in-depth understanding of which scientific use cases could benefit from advances in geospatial software and identifying a suite of core technical capabilities necessary for geospatial data transformation and associated scientific problem-solving. The second workshop brought together fifty-five representatives of both domain sciences and geospatial software communities and began the planning of how GSI could lead the communities to sustain and benefit from an open geospatial software ecosystem. The third workshop, held in July 2019 in Annapolis, Maryland (https://gsi.cigi.illinois.edu/workshop3/), focused on developing the strategic plan and governance model for GSI to serve the broad and diverse geospatial communities. The third workshop brought together forty-five representatives of the diverse communities and began developing the strategic plan of GSI.

The participants of the three workshops collectively: 1) represent key stakeholders from diverse geospatial-related communities; 2) recognize the need for developing the type of overarching infrastructure that the GSI should provide; and 3) represent both contributors to and consumers of GSI's capabilities and services. More than half of the participants were selected from an open application process to ensure broad community input and diversity with respect to ethnicity, gender, and discipline. Overall, seventy-six position papers were accepted to capture the contributions and inputs of the participants addressing the foci of the three workshops. These position papers are available on the website of the GSI conceptualization project as follows:

- The first workshop's position papers: https://gsi.cigi.illinois.edu/workshop/position-papers/
- The second workshop's position papers: https://gsi.cigi.illinois.edu/workshop2/position-papers/
- The third workshop's position papers: https://gsi.cigi.illinois.edu/workshop3/position-papers/

In addition to the three workshops, one community survey was conducted to further reach members of the geospatial software community who were not able to participate in our workshops. The primary goal of this survey was to gain an in-depth understanding of the geospatial software community in support of the GSI conceptualization process, as the success of the GSI will be dependent on both engaging and supporting a variety of geospatial software experts and users with diverse and evolving needs. By analyzing the results of the survey, we aimed to map the concerns and limitations of geospatial software to key areas of need for the GSI.

Additionally, to also reach out to graduate students and junior scholars and understand their needs from an educational perspective, in July 2019 we conducted a week-long summer school on *Reproducible Problem Solving with CyberGIS and Geospatial Data Science* that was held on the campus of the University of Illinois at Urbana-Champaign (UIUC) and co-led by the CyberGIS Center for Advanced Digital and Spatial Studies (CyberGIS Center) at UIUC, American Association of Geographers (AAG), and University Consortium for Geographic Information Science (UCGIS) (https://cybergis.illinois.edu/aag-ucgis-summer-school-2019/, https://www.ucgis.org/summer-school-2019). More than 30 graduate students and early-career scholars attended the summer school and learned to collaborate in developing novel solutions to complex problems and to take advantage of geospatial data science and cutting-edge scientific advances and geospatial software capabilities based on advanced CI (e.g., CyberGIS-Jupyter ([40-42])). Participants experienced the types of collaborative and professional interactions that are key to addressing reproducible geospatial problem-solving in the context of computation- and/or data-intensive research involving confidential geospatial data. The program had students working in teams on interdisciplinary and transdisciplinary topics. There were six teams that worked on the following areas: (a) GeoAI: Mapping Safety Features from Street View Images Using Deep Learning Approaches; (b) A GIS-Based Terrain Analysis Approach for Estimating Water Quality; (c) Stream Line Detection Using Deep Learning Techniques; (d) Digital Humanities - Corpus Linguistics; (e) Spatial Simulation Modeling of Disease Spreads; and (f) Measuring Spatial Accessibility. Even though none of the participants had sufficient expertise in all of domain science, computer programming, geospatial software knowledge, and finding the right data sources, by working as a team with a mentor, each of the teams was able to develop and present CyberGIS-Jupyter notebooks by the end of the Summer School. The participants (both mentors and students) were highly satisfied and the program was an unequivocal success. GSI should be able to learn from this program and create similar learning experiences for broader communities.

Broad Community Engagement

Although it is well understood that a variety of communities face challenges in terms of scaling, integrating, accessing, processing, analyzing and visualizing geospatial data using geospatial software, there is a lack of clear understanding of which aspects of these challenges are perceived as central to the developers and users of geospatial software. Given the sheer breadth of challenges, understanding individuals' perceptions of the most pressing challenges can help us begin to understand *which* challenges in the geospatial software world will hold the

potential to beneficially impact the largest swath of the user base (and, ideally, produce the largest return on investment in terms of improving the pace/quality of science using geospatial software).

We designed the questions of the survey to identify the axes of variance over which participants have had different experiences and interactions with geospatial software. Our aim was to distinguish among different types and intensities of interactions with geospatial software. There are two key foci: potential use cases for GSI that could benefit from organized community support, and geospatial software limitations that could be resolved to increase the quality and quantity of geospatial-related research and education. One journal manuscript has been submitted to articulate the design and findings of the survey ([43]), several key findings from which are articulated below.

Availability of and access to needed geospatial software, data, and computational resources are major limiting factors for related research and education. Respondents were hindered by overly expensive software, the need to develop custom tools, and reliance on prohibitively expensive datasets. Some, by necessity, depend on free or relatively inexpensive software solutions. Large spatial and temporal coverage gaps frustrate those doing regional or site-specific work, as well as not recent and generally unsuitable data. Data that are available and suitable are often stored in different poorly-compatible formats. Finally, many respondents primarily used desktop computers and noted difficulties gaining access to high-performance computing resources.

The geospatial software learning curve and development rate also presented a set of challenges. Many reported simply not having enough time to learn to use newer tools for their work, especially as tools rapidly change. It was also considered hard to find support or outside expertise to aid in gaining competency. For some respondents, geospatial software developed too fast, making it hard to stay current. For others the rate of development was not fast enough, where recent technical or algorithmic developments were slow to be included in major software applications.

Finally, geospatial software scope and ease of use were reported as issues. Many respondents preferred commercial solutions as they were familiar, while others used only open-source tools, as black box procedures in commercial software can hinder computational reproducibility. Some preferred more one-size-fits-all geospatial software applications to simplify analysis workflows. Respondents viewed software tools with too broad of coverage as lacking depth, with a poor fit for specific use cases. Similarly, greater general standardization could help to support greater software interoperability. Also, the current geospatial software ecosystem is perceived as buggy and fragile. Tenuous connections are easily broken by operating system, programming language, and other software updates, often leading to cascading failures.

GSI is well positioned to address many of the pain points outlined by the respondents in the survey specifically as it relates to development/cataloging of core technical capabilities that can be relied upon by the community and the development of education and learning materials that would help advance many diverse fields.

4. Major Challenges and Opportunities

The current state of geospatial software is evolving rapidly with significant fragmentation. It is often difficult to integrate various proprietary and open-source software tools for solving scientific problems. There are also related data challenges such as geospatial data sharing and reproducibility of geospatial research (including research based on confidential data and/or commercial data). As geospatial software is used in so many scientific domains, lack of education support and geospatial scientific guidance causes serious challenges to current and future generations of scholars. Oftentimes, various campus-level geospatial centers struggle to serve ever-growing needs largely through ad hoc efforts. On the other hand, these challenges highlight the tremendous potential for GSI to synergize related efforts across diverse communities and institutions while nurturing a sustainable social and technical ecosystem to enable geospatial-inspired innovation and discovery. Specifically, the community inputs from the GSI conceptualization process has identified two types of major challenges with one focused on computation & data and the other on research & education.

Computation & Data Challenges

- Uncertainty quantification, representation, and communication (goal 1: reproducible, transparent, and scalable geospatial software)
- Computational reproducibility and transparency (goal 4: structured guidance for computational reproducibility)
- Conflating and fusing multiple data sources (goal 1: reproducible, transparent, and scalable geospatial software)
- Integrating geospatial data with domain-specific models (goal 5: high-performance and data-intensive geospatial software)
- Incorporating geospatial characteristics and principles into artificial intelligence and machine learning (goal 3: ethical and open geospatial software)
- Scaling computational capabilities to solve increasingly complex geospatial problems using an increasing amount of geospatial data (goal 5: high-performance and data-intensive geospatial software)

Research & Education Challenges

- Goals 1 (reproducible, transparent, and scalable geospatial software), 2 (geospatial digital workforce), and 5 (high-performance and data-intensive geospatial software) require collaboration between disciplines that currently do not have common communication practices or research drivers.
- Goals 3 (ethical and open geospatial software) and 4 (structured guidance for computational reproducibility) require changes in the current academic incentives, which require the involvement of university administrators, publishers, and academic societies, as well as the development of technical specifications, software, and tools including platforms. This will mainly be accomplished through the collaborative efforts of the education and workforce development and the core capabilities and services teams.

To tackle these challenges while pursuing the goals, GSI can greatly benefit broad communities by:

- Developing collaborative practices between the geospatial research, and software developer communities are necessary, and relevant more broadly in the related sciences.
- Connecting research with industry and government in community discussions that require broad communities of researchers, software development and applications.
- Delivering the following leadership roles for the diverse geospatial communities.
    - Enable discovery and innovation
    - Advancing geospatial technologies enabled advanced cyberinfrastructure
    - Enable the development of structured guidance for geospatial software development, including facilitating computation reproducibility
    - Focus on fundamental scientific and societal challenges
    - Prepare the future workforce
    - Bridge the digital divide between diverse geospatial and advanced CI communities
    - Foster open collaboration

GSI is well-positioned to tackle the challenges and seize the opportunities of enabling the US research community to harness the geospatial data revolution to advance scientific discovery and innovation by employing a three-pronged strategy: **Deep, Wide and Transparent**.

- **Deep**: To help users of geospatial data and software to tackle the challenges of scale and complexity of utilizing geospatial data in modeling, analytics, visualization and decision making to solve high-impact research problems and societal challenges.
- **Wide**: To enable broader use of geospatial software and data especially by non-traditional and diverse geospatial software users, and train the next generation of researchers and workforce in creating and using geospatial software following the FAIR principles.
- **Transparent**: To promote and enable transparency and reproducibility of data-driven research and innovation by engaging both geospatial data producers and consumers in developing structured guidance, and employing an inclusive governance model.

5. Social and Technical Ecosystem

Geospatial software conducts analysis, modeling, processing, simulation, and visualization; represents a broad spectrum spanning both geospatial data and computing functionalities; and is developed by diverse communities including academia, industry, government, and open source communities. Therefore, our plans for nurturing a social and technical ecosystem synergizes four critical dimensions of advanced geospatial capabilities: data transformation, integration and interoperability, cyberGIS, and collaborative problem solving

with an overarching focus on computational reproducibility and transparency (as shown in Figure 3).

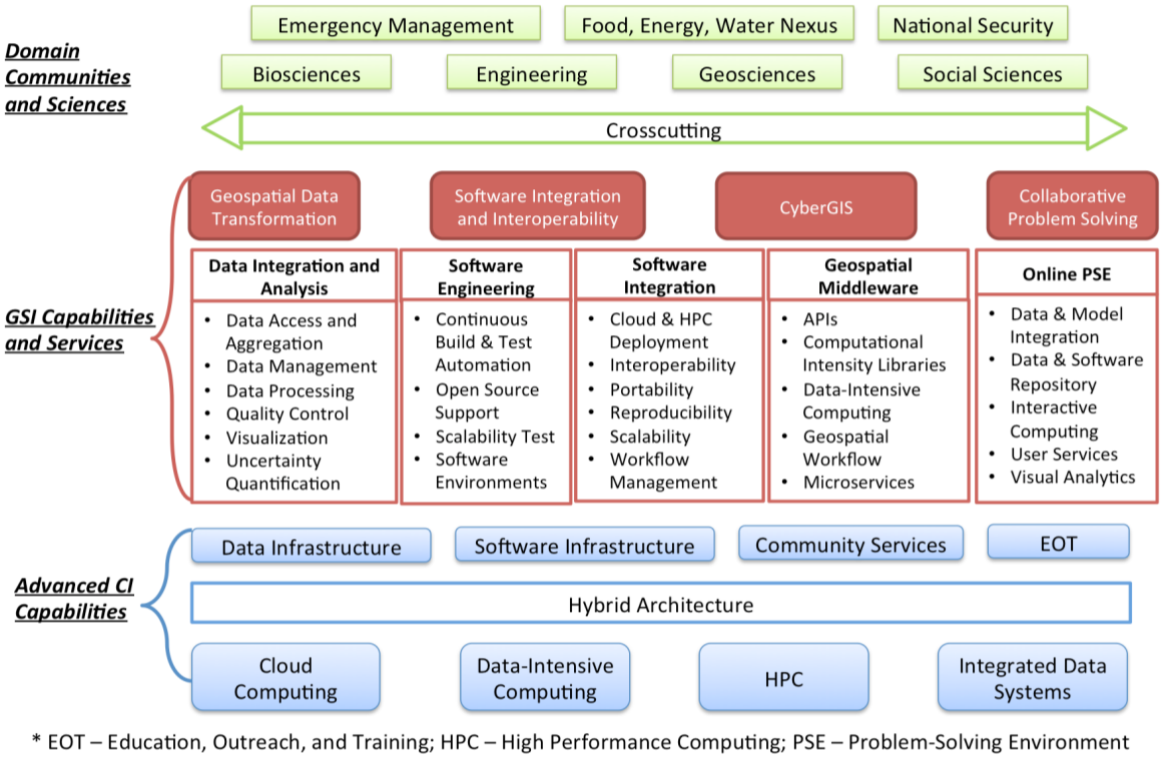

Figure 3: GSI technical architecture and building blocks ([44])

Social Ecosystem

GSI needs to include a set of domain-science researchers and software experts, mixing geospatial, reproducibility, and CI software development and maintenance expertise. These experts need to be distributed across institute partner sites as is done by the Science Gateways Community Institute (https://sciencegateways.org). In addition, the institute needs to offer fellowships and a visitor program, where software researchers and developers from projects outside the institute can work with the institute and its staff members, either remotely or at an institute site, to develop or improve particular software. This program needs to engage broad communities.

Technical Ecosystem

GSI needs to leverage existing NSF and other federal agencies' investments in geospatial software that have occurred over the past decade. This, for example, should include cyberGIS software, GABBs, National Map, EarthCube's Geosemantics, US Federal Resources for Social Science GIS Data. More details about the core capabilities GSI should focus on are discussed in Section 6. Additionally, there is a substantial open-source geospatial software community that is developing geospatial software tools, mainly driven by commercial interests. The institute needs

to ensure requirements and concerns coming from science domains are well communicated, and where appropriate, contribute software back that has been modified or optimized to work on large geospatial data and/or advanced CI. The technical ecosystem should include a focus on computational reproducibility as part of the software engineering and integration lifecycle.

6. Core Capabilities and Services

Geospatial data transformation functionalities include but are not limited to access, management, analysis, and visualization, where geospatial characteristics have long been recognized to be critical to the quality, performance, and validity of the functionalities. Given the diversity of software tools and complexity of workflows for geospatial data transformations as well as the *Major Challenges and Opportunites* identified in Section 4, it is important to ensure that the geospatial data transformation process employed is transparent, verifiable, and computationally reproducible. Typically, geospatial data compiled for a given application comes from multiple geospatial data sources with heterogeneous spatial data representation, extent, projection, resolution, scale, and accuracy. For example, satellite remote sensing data provided by NASA and USGS is central to various hydrological and environmental analysis tasks. The NASA data, in turn, is hosted at various DAACs (Distributed Active Archive Centers), and is organized based on the frequency of data collection (8-hourly, daily, etc.) and the remote sensing instrument used. For example, MODIS data is organized based on spatial coverage and uses different spatial projections for different resolutions. Various data processing functions often need to be invoked to prepare data as inputs to diverse applications. Data catalog, metadata, and visualization information are often tightly coupled with data functionalities. Correct execution of these functions is important to avoid accuracy and uncertainty issues that may significantly alter analysis results if not treated properly.

The current practice of geospatial data transformation by non-geospatial communities, which is common given the broad use of these data, has two gaps. First, geospatial expertise is needed to understand how to choose and apply data transformation software, given the variety of data and software tools. Second, data integration through a stack of geospatial libraries/tools needs rigorous geospatial inspection while ensuring computational reproducibility. Geospatial data workflows thus far often involve a mix of desktop tools, remote computation, and intermediate data staging and transfers. Chief among the reasons for the persistence of such cumbersome workflow practices is a barrier to entry for developing completely code-based data transformation software that is designed for and leverages advanced CI and high-performance computing resources. **GSI can provide a unique and critical capability to fill the gaps through software innovation, computational reproducibility guidance, and related community support services by closely working with related data-driven projects and programs (e.g., EarthCube and Unidata ([45])) and CI resources (e.g., CyberGIS and HUBzero).**

Let us consider a typical application from social sciences that needs integrated data from multiple sources, a common requirement for many geospatial analysis applications. Data might

come from federal agencies such as the US Census Bureau, the Centers for Disease Control and Prevention (CDC), and others that create and disseminate social science geospatial data and related mappable output to support researchers and practitioners. For example, the US Census Bureau creates official boundary information to support standard demographic programs, as available through web resources ([38, 46]), with related demographic data available at data.census.gov. Similarly, the Centers for Disease Control (CDC) provides a number of online public health and epidemiology-related resources (https://www.cdc.gov/gis/public-health-maps.htm). All of the above resources contain multi-level data along with various spatial units of aggregation, with some being compatible across spatial unit hierarchies and others not. For example, data may be published both at census units such as tract, block group, and block, which are nested, as well as ZIP code or municipality, which are not. Consequently, typical applications embody multi-layered many-to-many relationships that need to be considered using geospatial processing ([38, 47-49]). Moreover, such data are constantly being updated with releases occurring at various times throughout the year. For example, new official boundaries are expected to be created as part of the 2020 decennial census, in addition to ZIP code boundary changes that regularly occur ([50]). The geospatial data transformation and integration challenges outlined for such an application are, to a large extent, in-line with the gaps identified in the previous paragraph. Desirable progress on bridging the identified gaps is expected to greatly benefit social science researchers and data users with **improved geospatial software to facilitate data management, analytics, and integration**.

There is a pressing need for structured guidance of geospatial data transformation software that includes commonly used utilities, documentation, best practices, and a knowledge base to aid domain application developers to ensure computationally reproducible data handling and shortening development time. GSI should help to satisfy this need by addressing common but often ignored pitfalls in geospatial data transformation by working with developers to improve software when appropriate and providing the appropriate educational and training resources as well as structured guidance for geospatial software developers. For example, the basic re-projection function in the popular PROJ.4 library is applied to a single spatial object sequentially, which means the re-projection of a million points would create a million of spatial objects and project them one by one. Structured and scalable design for exploiting geospatial parallelism is thus needed to support the fast processing of a large number of projection calls in an efficient way. Furthermore, because extensive geospatial data libraries and tools are designed based on sequential computing approaches, **scalable designs based on high-performance and scalable computing approaches, along with workflow tools to manage computational research and enable reproducibility**, are urgently needed as part of broad geospatial software practices.

Geospatial computing functionalities, on the other hand, are often based on various geospatial algorithms and models with cartographic, geometric, mechanistic, statistical, and topological foundations using a variety of computer architecture and programming models. High-performance computing has been largely under-represented in geospatial software. A variety of existing geospatial software tools are not able to take advantage of advanced CI resources because the underlying programming model can at most exploit CPU, memory, and IO

within a single computing node. **High-performance functions need computational profiling, efficient spatial data structure design, performant implementation, and thorough testing on scalability**.

Given the diversity of geospatial software, **GSI should focus on identifying cutting-edge geospatial software functionalities (e.g. geospatial data integration), architecture (e.g. hybrid), elements (e.g. agent-based modeling), frameworks (e.g. cyberGIS), services (e.g. computational scalability), and workflows (e.g., GISolve ([28]) and Swift ([51])) commonly needed by the communities and projects**. GSI needs to focus on developing a roadmap to resolve geospatial computation and data challenges, establishing structured guidance for computational reproducibility, and achieving sustainable adaptability to advanced CI. To facilitate usability and ease of development, GSI needs to work with partners like OGC to help design standardized software interfaces for geospatial data integration and transformation libraries. These interfaces will not just aid in a higher-level conceptualization of a complex geospatial workflow as a set of data building blocks, but also simplify the composition and reusability of these building blocks in diverse workflows that are completely executable in advanced CI environments. There is already a wealth of scientific workflow frameworks (e.g., Pegasus ([52]), Apache Beam ([53]), Airflow ([54]), Parsl ([55])) that can seamlessly manage execution across different execution environments. Hand-in-hand with facilitating effective use of high-performance computing, **GSI should also provide guidance for adapting these building blocks into appropriate scientific workflows (and frameworks)** for particular use cases.

GSI needs to play leadership roles in **harnessing geospatial data revolution for diverse discovery and innovation by promoting the FAIR principles** when facilitating any new geospatial software development for data-driven research and education. The role of GSI is centered on enabling the **evaluation, integration, and scaling of geospatial software by exploiting advanced CI** and holistically engaging geospatial-related communities. Through rigorous software engineering and guidance of critical geospatial knowledge, we envision focusing on integration, interoperability, and scalability while partnering with geospatial software development and standardization efforts (e.g., ASF (Apache Software Foundation), OGC, and OSGeo (Open Source Geospatial Foundation)) for individual software development and evaluation.

One primary technical deliverable of GSI is a **geospatial software community hub** that bridges researchers and software experts by provisioning an integrated set of software development, testing, and evaluation cases; community adoption practices; sample data; and quality and performance benchmarks. Three categories of software infrastructure services have been identified as critical to the communities: 1) a continuous and scalable integration sandbox; 2) deployment services based on containerization/virtualization via cloud computing approaches; and 3) an online interactive platform for testing, research & education, and community feedback built on CyberGIS-Jupyter ([42]). It should be noted that computational reproducibility is a challenge for each of these categories and will be an important consideration when core capabilities and technical services are provisioned. These services are expected to be deployed

and executed on hybrid CI architecture (e.g., Virtual ROGER: https://cybergis.illinois.edu/infrastructure/hpc-user-guide/) that integrates data-intensive computing, high-performance computing, geospatial databases, and cloud computing components (Figure 3).

7. Organizational Structure and Governance

In the workshops conducted as part of the conceptualization process, there were a number of open discussions regarding how GSI should govern itself and how it should engage with the diverse communities. The institute should be a multi-institutional entity with leaders and participants coming from various academic institutions, government agencies, and industry. Based on the discussion we have converged to the following governance model (Figure 4).

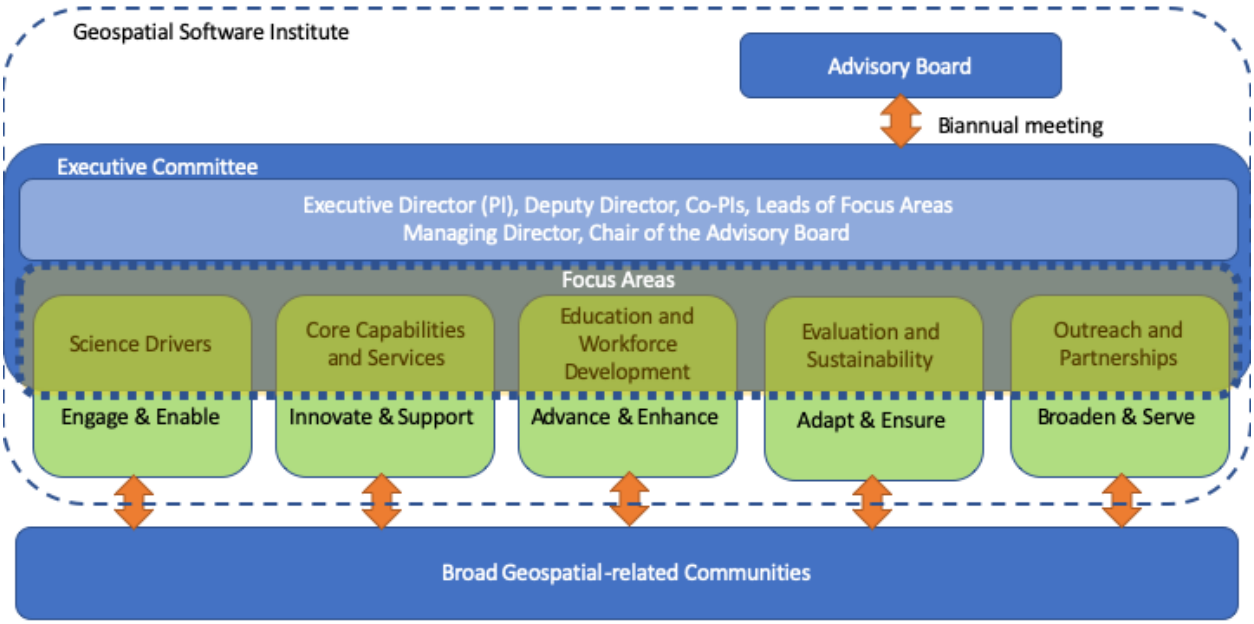

Figure 4: Organization Structure of GSI

PI/Co-PIs: The PI should serve as the Executive Director (see details below) and oversee one focus area, and each of the Co-PIs should be responsible for overseeing one of the focus areas. We expect the PI to be committed at least 50% of her/his time to the institute activities. Each of the Co-PIs is expected to commit 1 to 3 months of time per year on GSI activities.

Executive Director: The PI should serve as the Executive Director, providing vision and ensuring that the overall functioning of the institute is in line with its commitments. The Executive Director leads the Executive Committee and strives to achieve consensus within the Committee on strategy and priority decisions based on advice and recommendations by the Advisory Board. The Executive Director makes the final decision in case consensus cannot be reached.

Executive Committee: The Executive Committee should oversee the programs and activities of the institute and ensure that progress is on track. The Committee comprises the Executive Director, Co-PIs, Deputy Director, Managing Director and leads of focus areas. The Executive Committee should meet weekly to manage the regular activities of the institute and to make short-term plans. The chair of the Advisory Board needs to be an ex-officio member of the Committee and should attend at least one meeting per month to provide timely guidance. In addition, the Committee needs to meet twice during each project year to assess project progress and evolve long-term plans. The Executive Committee, with guidance from the Advisory Board, should be responsible for making changes to focus areas as needed to meet the institute's goals.

Advisory Board: The Advisory Board reviews the institute's activities and provides timely guidance to the Executive Committee. The Advisory Board represents the domain sciences, geospatial sciences and technologies, and CI communities. The Board should include 11 members and be appointed based on soliciting diverse nominations from the broad communities and consulting with NSF. The duration of the Board membership is 2 years. The Board should have a chair and a co-chair elected by the Board members with a 1-year term for each. The Board should meet biannually co-located with the face-to-face meetings of the Executive Committee for desirable coordination and collaboration.

Deputy Director: One of the Co-PIs should be identified as the Deputy Director who in addition to her/his other roles should assume the activities of the Executive Director during periods of unavailability of the Executive Director to ensure the smooth continuity of the institute. This co-PI should commit at least 25% of her/his effort for the institute's activities and work closely with the Managing Director on day-to-day basis to ensure institute's commitments are met.

Managing Director: The Managing Director should be in charge of the day-to-day activities of the institute and responsible for translating the vision and achieving the goals with activities on the ground. S/he should be responsible for mapping and implementing the strategic decisions to tactical actions and ensure that they are implemented on schedule. The Managing Director should also work closely with the leads of the focus areas to ensure they are meeting their goals and to facilitate synergistic collaborations. The Managing Director should spend at least 70% of her/his effort on GSI activities.

Focus Areas: The following five interrelated key focus areas were identified as requiring consistent foci of the institute: (a) Science Drivers; (b) Core Capabilities and Services; (c) Education and Workforce Development; (d) Evaluation and Sustainability; and (e) Outreach and Partnerships. Each of these focus areas should have a team contributing to it and be responsible for progress on one or more of the institute goals.

The team in the *Science Drivers* focus area should gather and articulate requirements, informing what specific advances in geospatial software will be needed to advance research and enable discovery (goals 1, 4). In addition to the science communities identified in Section 2, this team should reach out and regularly conduct community events that endeavor to engage new science communities by highlighting the offerings of the institute and collecting their geospatial

software requirements. New communities and their geospatial software requirements as identified by the *Science Drivers* team need to be communicated with the Executive Committee. The Executive Committee, in consultation with the Advisory Board, will identify new science domains the project should engage with and the amount of effort the institute should put in. The Executive Committee should review and make decisions regarding continued engagement with current science communities and new potential engagements every six months.

The *Core Capabilities and Services* area should have a crosscutting team responsible for advancing geospatial software to serve the requirements of the communities and specific science drivers while mobilizing the communities to establish common standards for geospatial software interoperability, reusability, and reproducibility. This team is expected to push the state of the art of high-performance geospatial software and collaborate with the broader advanced CI communities for achieving the following institute's goals: reproducible, transparent, and scalable geospatial software, ethical and open geospatial software, structured guidance for computational reproducibility, and high-performance and data-intensive geospatial software. This team should engage closely with the *Science Drivers* area to ensure that the requirements coming from the science communities directly inform and drive the technical work in this area.

The *Education and Workforce Development* area team should be responsible for significantly increasing the community capacity to effectively use and efficiently advance geospatial software by creating curricula and developing programs such as hands-on workshops, and summer schools through collaboration with community organizations (e.g., AAG, the Carpentries, and UCGIS) (goals: geospatial digital workforce, structured guidance for computational reproducibility). They should be the main conduits of transmitting structural guidance developed by the institute (e.g., on computational reproducibility in scientific research, standard-based software practices) to broad communities.

The *Evaluation and Sustainability* team should focus on establishing and evolving the success metrics of the institute as well as pursuing its sustainability and sustained impacts. The team should assess GSI's core capabilities and services to contribute towards the goal 3: ethical and open geospatial software. This team will be cross-cutting and should engage closely with all other focus areas.

The *Outreach and Partnerships* team should develop connections and partnerships to achieve the broader impacts of the institute while helping the partners to achieve impacts that are mutually beneficial to the institute (goals: ethical and open geospatial software, and structured guidance for computational reproducibility).

While these five core areas encompass the breadth of expertise and programs necessary to achieve the institute's goals, the Executive Committee should be empowered to evolve this structure in conjunction with discussions with and recommendations from the Advisory Board.

Leads of Focus Areas: Each focus area should have a lead who is responsible for defining, evaluating, and refining the goals of the area and ensuring they are met by specific programs and activities. Based on team size and associated deliverables, co-leads may be identified to assist the leads to assure robust and timely deliverables and assume the activities of

a lead when a lead is unavailable to ensure minimal disruption in GSI operation. The areal leads should work closely with each other as part of the Executive Committee to ensure that the areal teams work in a complementary fashion to achieve the goals of the institute.

8. Education and Workforce Development

The primary goal for GSI in education and workforce development is for open, standard-based, reproducible geospatial software development to be a prominent curricular component in GIS, geospatial data science, and domain sciences that develop and use geospatial software and/or collect and use geospatial data. GSI needs to use three categories of services to achieve this goal: 1) curriculum development and instructor training; 2) graduate student and early-career scholar training; and 3) self-paced and guided virtual training for all levels of education (two-year institutions to life-long learning professionals) that incorporate principles and strategies for inclusive teaching. These three broad categories can enable GSI to transform education and workforce development in several ways. First, through curriculum development and instructor training, GSI can position geospatial software development, principles, and practices to be part of every “Introduction to GIS and/or geospatial data science” course and reach a broad and diverse audience by "teaching the teachers”. Second, students and scholars need to be trained to work in teams and become well-versed in cross-disciplinary communication and convergence research. GSI needs to develop communication practices that are currently lacking between the geospatial research and software developer communities, but which are necessary, and relevant more broadly in diverse sciences. Third, GSI needs to produce activities and curriculum materials to educate more diverse audiences with combined computational and spatial thinking skills.

The curriculum development and instructor training service aim to transform the ways in which geospatial software and sciences are taught across the country and potentially the world. Traditional introductory courses focus on commercial and open source geospatial software using Graphical User Interfaces (GUIs) and “point-and-click” approaches, which are time consuming and whose results are difficult to reproduce. GSI needs to work with partners such as AAG and UCGIS to shift geospatial software development to be up-front and center in geospatial software and data science curriculum. A key challenge in this shift is that many geospatial instructors have limited programming expertise. GSI needs to facilitate the collective creation of top-tier materials for lecture, lab, and outside activities and train instructors in how to effectively use the materials in their introductory, intermediate, and advanced courses. GSI needs to enable pertinent communities to better understand student attitudes and challenges in order to develop a curriculum that will be effective and broaden participation in this context.

Curriculum materials can infuse the latest research and best practices including software specifications and standards that support reproducibility of geospatial research and that ensure the ethical use of geospatial data and software. The curricular materials should be based on Principles of Human-Computer Interaction for the development of software for use by geospatial communities and include not only software development, but also communication practices and

collaborative problem solving. Materials should include new instruments such as usability metrics relevant to meet the software needs of pertinent communities and new protocols and practices to check the reliability, and validity of geospatial software tools. The curriculum materials should leverage the latest UCGIS Body of Knowledge (https://www.ucgis.org/gis-t-body-of-knowledge, https://gistbok.ucgis.org/), the Geospatial Technology Competency Model [56], and Cyber Literacy for GIScience [57]. GSI needs to organize a series of virtual and in-person workshops that bring together a diverse audience of scholars of the geospatial research [58] and software developer communities to address overarching curricular and education needs in these communities, which include software development that withstands reproducibility challenges for ensuring ethical use and representation of geospatial data.

GSI needs to support cross-departmental collaborations and synergies among geography, GIS, computer science and engineering faculty to teach such curriculum together, and in doing so, prepare the future workforce for interdisciplinary and transdisciplinary work. An additional challenge in higher education and workforce development in terms of acquiring advanced skills and in terms of diversity is that there is no gradual learning pathway between K-12 and college. Educators in K-12 and higher education rarely collaborate or communicate, and lack the incentives to do so. Over time, this continues to widen the gap between the knowledge and skills that high school graduates acquire, and the prerequisite knowledge and skills that are expected from them in college. Underrepresented and disadvantaged groups are increasingly impacted by this widening gap. To address this challenge, GSI needs to support collaboration around education across all levels. This can come in the form of collaboration on professional development materials, or of initiation of Researcher-Practitioner Partnerships [58] that would improve learning outcomes around geospatial data and software and associated awareness of careers before college.

Face-to-face training is paramount to transforming the current practice of geospatial-related sciences to ensure current and next-generation scholars are producing the best scholarship, and that it can be computationally reproducible, and accessible and open to the broad research and stakeholder communities. Face-to-face training can also serve as a testing ground for curriculum material development to get initial feedback and test new approaches. GSI needs to host regular summer schools, following the successful model of the CyberGIS Summer Schools held in 2017 and 2019. This should be a core service of the GSI. Second, GSI needs to host face-to-face training sessions at a major conference each year. These mini-training sessions need to provide participants a solid starting point to serve as community champions. Training materials need to be available for volunteers and champions to teach their own training events at their own institutions (similar to the carpentries model), enabling GSI to scale and reach many students and early-career scholars. These education and workforce development activities should target a diverse audience in terms of gender, race/ethnicity, and background.

Self-paced and guided virtual training can reach learners at all levels, life-long learners, and students who cannot participate in face-to-face training for a variety of reasons. GSI needs to host a bi-monthly hour-long online seminar series discussing the latest research and problems being solved by GSI using real-world examples to highlight key concepts and skills. Virtual

training sessions can use GSI infrastructure and CyberGIS-Jupyter notebooks to enable self-paced learning. The use of such notebooks can also demonstrate solutions to core principles of transparency, collaboration, and reproducibility in scientific research, even with complex and large geospatial data.

NSF has funded a number of cyber training projects (e.g., Cyber Training for FAIR Science on mygeohub.org, CyberGIS Fellows, and Hour of CI) that have developed learning modules and courses on, among other topics, open source software development, advanced computation, geospatial data science competency for use in domain science classrooms, and informal training settings such as workshops, summer schools, and online self-paced tutorials. GSI should be in an excellent position to leverage and integrate these resources and expertise, and make substantial impacts to benefit broader communities beyond what any of the individual projects can achieve independently.

9. Outreach and Partnerships

GSI needs to develop connections and partnerships to achieve broader impacts and extend outreach efforts to diverse communities. Given the broad reach and impact of geospatial software, the institute should extensively engage with partners from a large array of academic, governmental, industrial, public and international organizations. These organizations include not only those involved in the representative communities and projects, but also extend to other academic organizations (e.g., EarthCube and Unidata), professional organizations (e.g. ESIP), U.S. federally funded research and development laboratories (e.g. Oak Ridge National Laboratory), international organizations (e.g. the Association of Geographic Information Laboratories in Europe (AGILE)), major industry players (e.g., Esri, Google, and Microsoft), public institutions (e.g., California Academy of Sciences, Adler Planetarium), and consortia (e.g. OGC) that bring together these communities for consensus development of standards and innovative software ecosystem. During the conceptualization process, we have already engaged with a number of partners across academia, government, and industry, many of whom actively participated in our workshops. GSI needs to invite participants from these partners to contribute to appropriate programs and activities matched to their interests and expertise, and expects they will actively contribute to the implementation of the institute.

- Academic and international (e.g., AAG, AGILE, AGU, CyberGIS, ESIP, GIScience, UCGIS, and XSEDE)
- Government (e.g., CDC, DOE, EPA, NASA, NGA, NIH, and USGS)
- Industry (e.g., DigitalGlobe, Esri, Google, HDF, Kitware, Microsoft, and LimnoTech)
- Public Outreach Institutions (e.g., museums, science centers, and the California Academy of Sciences)
- Consortia (e.g. OGC)
- Others (e.g. WGIC: https://wgicouncil.org/about-us/#governance)

Critical and urgent efforts are needed: 1) to attract and prepare the next-generation workforce for computation- and/or data-intensive geospatial-related research, technological

innovation, and real-world problem solving and decision-making; and 2) to effectively communicate geospatial-inspired sciences and raise public awareness of important GSI capabilities and services. This science communication extends beyond the geospatial software community by cultivating opportunities where "science meets society" through museums, science centers, and public media.

One concrete example of a project in progress that is both useful to geospatial software community while providing opportunities for museum exhibits and informal science public outreach is a collaboration with the ArcticDEM project in which the Advanced Visualization Lab (AVL) at NCSA is developing geospatial and real-time visualization software techniques to process and visualize time-evolving satellite data over the Arctic and Antarctic. The real-time techniques will enable time-evolving geospatial animations that will enhance GSI capabilities for visualizing satellite data. This work provides informal and formal educational capabilities. Given the existing AVL partnerships with outreach programs at museums and documentary television, such GSI capabilities can be employed to expand the reach and communication to the general public.

Part of the informal science educational task needs to be accomplished through a three-pronged approach. First, GSI needs to take an important step to unleash the power and influence of innovative geospatial software to attract diverse and underrepresented future scientists through strategic partnerships with museum exhibitors and informal science specialists. Some GSI software tools are useful to provide interactive and other visualization exhibits to informally educate and provide outreach to the broad public. Second, many opportunities exist to raise public awareness of geospatial discoveries while attracting a new generation of developers/researchers and empowering the public with new geospatial knowledge that has a direct impact on societal grand challenges. GSI needs to extend, sustain and innovate advanced CI by targeting outreach opportunities to incorporate geospatial software into museum exhibits and feature GSI discoveries in documentary television programs. 3) GSI needs to target marginal communities and work with diverse groups to help broaden participation in GSI software research and development.

GSI is well positioned to impact education, outreach, and workforce development activities because geospatial software tools often provide intuitive representations and visualizations that are very attractive to students. GSI needs to represent diverse academic, governmental, and industrial institutions and to identify new capabilities and knowledge that will have far-reaching impacts on addressing real-world problems including those with significant impacts on the nation's economic development and security. The outreach activities described above are useful across diverse domains. For example, socially relevant problem-solving might be relevant to multiple fields such as archaeology, hydrology, social sciences, and atmospheric sciences.

GSI needs to partner with OGC to advance its goals in a consortium setting that brings together the research, industry, and government communities that share the objectives of FAIR geospatial software. GSI leadership in initiatives in OGC's Innovation Program is expected to result in interoperable software development, testing and applications on key science-driven

advances. GSI leadership is vital to OGC's annual testbed process where related GSI goals can be rapidly advanced through co-sponsorship of activities for pertinent communities. Alternatively, GSI leadership can be focused on initiatives tuned to meet the specific goals of GSI. The results of the Innovation Program become the basis for consensus standards development in OGC's Standards Program. The most successful standards for geospatial software have been based on fundamental geospatial science research that leads to wide-spread adoption based on the effectiveness of science-based standards.

Annual regional group meetings/workshops need to be held to facilitate face-to-face engagement of scholars and students within a particular region or institution. Regional champions should be selected to organize short workshops which could be co-organized with another regional meeting to make such events achieve the highest cost-benefit level (e.g., regional AAG meetings or statewide GIS meetings).

10. Evaluation and Sustainability

Enabling researchers and scientists in diverse domains to harness the geospatial data revolution is a key focus of GSI. Conducting and fostering research and development that lead to high-impact geospatial software should help advance both geospatial sciences and domain sciences. Focus areas that were identified in the section of *Organizational Structure and Governance* provide a mechanism to make progress on achieving the institute's goals. Within each of the focus areas, initial success metrics are outlined below and is expected to be annually updated in terms of projected impact towards the institute's goals and vision.

In the *Science Drivers* area, a suite of questions will be used to evaluate progress with the following two as initial examples:

- How many scientists/communities has GSI engaged to generate technical requirements from science drivers?
- Have the technical solutions developed by the *Core Capabilities and Services* team been successfully translated and applied to scientific problem solving?

For the *Core Capabilities and Services* area, the questions to assess impact should foucs on technical innovation and initially include the following examples:

- Has a substantially modified geospatial algorithm made execution faster?
- Have new geospatial data models been developed to effectively process and analyze geospatial big data?
- How many solutions have been developed to meet technical requirements generated by science drivers?

For the *Education and Workforce Development* area, a series of metrics need to be developed to capture the impact of education and learning materials and how they get disseminated. Some specific metrics include:

- How many curriculum development and instructor training activities have been conducted and how many courses/instructors adopt the materials being offered by GSI?

- How many graduate students and early-career scholars have been engaged by GSI's training and workshop activities (e.g. summer schools)?
- What is the levels of satisfaction of instructors and participants engaged at various training events?

The *Outreach and Partnership* area needs to be evaluated on how well this area is able to amplify the impact of the project outcome and generate new partnerships. Example metrics include:

- How many community events (e.g. webinars) have been held?
- How well are the institute's research and education activities being communicated with pertinent stakeholders and the public?
- How are the partnerships with other academic, government, industry and public institutions being developed and fostered?

These metrics will need to be reviewed and revised regularly as advances will be made by GSI. For example, the *Core Capabilities and Services* area is technically focused with activities being driven by science drivers, and hence the process must be flexible to allow incremental changes. This is a reason why we have *Evaluation* included as a key focus area. The evaluation team should work closely with the *Executive Committee* to regularly develop and update metrics so that these are aligned to ensure progress will be made towards achieving the overall goals of the institute. They need to also work with area leads to ensure that individuals working on various focus areas clearly understand evaluation metrics to help better focus efforts on achieving the institute's goals.

While we have largely discussed evaluation up to this point, the focus area covers both *Evaluation and Sustainability*. This combination is deliberate for assuring the evaluation activities to serve the purpose of achieving the institute's sustainability and sustained impacts. We recognize technological, financial, and social sustainability as the three pillars that will support GSI's sustainability plan.

- Technological sustainability needs to be achieved by utilizing open standards and technologies that can better adapt to inevitable changes in geospatial software and CI evolution. Specifically, GSI needs to 1) adopt from and contribute to, best practices and standards within OGC; 2) minimize the effort required to enable third-party (e.g. other CI communities) interactions; and 3) manage software intellectual properties using appropriate open source licenses.
- Financial sustainability needs to be realized by fostering a community ecosystem and delivering value crucial to enable scientific advances. GSI needs to employ the following concrete strategies: 1) solutions developed by GSI are focused on science domains where the challenges are clearly long term and require multidisciplinary communities to tackle them; 2) long-term partnerships with diverse geospatial-related agencies (e.g., DOE, NASA, NIH, and USGS) can assure the alignment of GSI with the long-term plans of the agencies for the development and use of GSI's capabilities and services; and 3) a transition plan, formulated for core software components to be incorporated as community-led projects.

- Social sustainability can be accomplished by employing community-driven and participatory approaches. GSI needs to: 1) leverage diverse partnerships to benefit broad communities; 2) provide participatory workshops and surveys to regularly gather community requirements and feedback, thereby developing a strong sense of community ownership; and 3) offer valuable software capabilities and access to advanced CI resources to enable high-performance and collaborative scientific problem solving.

GSI should become a center of technical expertise that can be contracted out for consulting on projects and contribution to software development that are written into other projects and activities. The institute can also provide expertise for education and training in addition to related materials developed to generate additional revenue.

References

[1] S. Wang, "A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis," *Annals of the Association of American Geographers*, vol. 100, no. 3, pp. 535-557, 2010.

[2] K. C. Seto, J. S. Golden, M. Alberti, and B. L. Turner, "Sustainability in an urbanizing planet," *Proceedings of the National Academy of Sciences*, vol. 114, no. 34, pp. 8935-8938, 2017.

[3] K. C. Seto and N. Ramankutty, "Hidden linkages between urbanization and food systems," *Science*, vol. 352, no. 6288, pp. 943-945, 2016.

[4] Y. Zhang, A. T. Murray, and B. Turner Ii, "Optimizing green space locations to reduce daytime and nighttime urban heat island effects in Phoenix, Arizona," *Landscape and Urban Planning*, vol. 165, pp. 162-171, 2017.

[5] D. Z. Sui, "GIS and urban studies: positivism, post-positivism, and beyond," *Urban Geography*, vol. 15, no. 3, pp. 258-278, 1994.

[6] J. Heissel, C. Persico, and D. Simon, "Does Pollution Drive Achievement? The Effect of Traffic Pollution on Academic Performance," National Bureau of Economic Research, 0898-2937, 2019.

[7] Y. Cai *et al.*, "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches," *Agricultural and Forest Meteorology*, vol. 274, pp. 144-159, 2019.

[8] A. Ramaswami, Bettencourt, L., Clarens, A., Das, S., Fitzgerald, G., Irwin, E., Pataki, D., Pincetl, S., Seto, K., Waddell, P., "Sustainable urban systems: articulating a long-term convergence research agenda. Available: https://www.nsf.gov/ere/ereweb/ac-ere/sustainable-urban-systems.pdf Access date: December 15, 2019.," 2018.

[9] S. L. Cutter *et al.*, "Disaster resilience: A national imperative," *Environment: Science and Policy for Sustainable Development*, vol. 55, no. 2, pp. 25-29, 2013.

[10] L. Anselin and S. J. Rey, "Spatial econometrics in an age of CyberGIScience," *International Journal of Geographical Information Science*, vol. 26, no. 12, pp. 2211-2226, 2012/12/01 2012.

[11] Z. Zhang *et al.*, "A cyberGIS-enabled multi-criteria spatial decision support system: A case study on flood emergency management," *International Journal of Digital Earth*, vol. 12, no. 11, pp. 1364-1381, 2019.

[12] F. Lyu *et al.*, "Reproducible Hydrological Modeling with CyberGIS-Jupyter: A Case Study on SUMMA," presented at the Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning), Chicago, IL, USA, 2019.
[13] S. Wang *et al.*, "CyberGIS software: a synthetic review and integration roadmap," *International Journal of Geographical Information Science,* vol. 27, no. 11, pp. 2122-2145, 2013/11/01 2013.
[14] S. Wang, N. R. Wilkins-Diehr, and T. L. Nyerges, "CyberGIS-toward synergistic advancement of cyberinfrastructure and GIScience: a workshop summary," *Journal of Spatial Information Science,* vol. 2012, no. 4, pp. 125-148, 2012.
[15] D. Schimel, W. Hargrove, F. Hoffman, and J. MacMahon, "NEON: A hierarchically designed national ecological network," *Frontiers in Ecology and the Environment,* vol. 5, no. 2, pp. 59-59, 2007.
[16] W. Tu *et al.*, "Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns," *International Journal of Geographical Information Science,* vol. 31, no. 12, pp. 2331-2358, 2017.
[17] C. E. Catlett, P. H. Beckman, R. Sankaran, and K. K. Galvin, "Array of things: a scientific research instrument in the public way: platform design and early lessons learned," in *Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering*, 2017: ACM, pp. 26-33.
[18] J. R. Wolch, J. Byrne, and J. P. Newell, "Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'," *Landscape and urban planning,* vol. 125, pp. 234-244, 2014.
[19] C. Burton, S. Rufat, and E. Tate, "Social Vulnerability," *Vulnerability and resilience to natural hazards,* p. 53, 2018.
[20] G. C. S. W. M. Hwang and A. P. Z. Z. Kiumars, "A scalable framework for spatiotemporal analysis of location based social media data."
[21] Y.-d. Wang, X.-k. Fu, W. Jiang, T. Wang, M.-H. Tsou, and X.-y. Ye, "Inferring urban air quality based on social media," *Computers, Environment and Urban Systems,* vol. 66, pp. 110-116, 2017.
[22] A. Soliman, K. Soltani, J. Yin, A. Padmanabhan, and S. Wang, "Social sensing of urban land use based on analysis of Twitter users' mobility patterns," *PloS one,* vol. 12, no. 7, p. e0181657, 2017.
[23] L. Xu, M.-P. Kwan, S. McLafferty, and S. Wang, "Predicting demand for 311 non-emergency municipal services: An adaptive space-time kernel approach," *Applied geography,* vol. 89, pp. 133-141, 2017.
[24] A. J. Collins, P. Foytik, E. Frydenlund, R. M. Robinson, and C. A. Jordan, "Generic incident model for investigating traffic incident impacts on evacuation times in large-scale emergencies," *Transportation Research Record,* vol. 2459, no. 1, pp. 11-17, 2014.
[25] G. Cao, S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang, and K. Soltani, "A scalable framework for spatiotemporal analysis of location-based social media data," *Computers, Environment and Urban Systems,* vol. 51, pp. 70-82, 2015.
[26] S. Wang, Y. Liu, and A. Padmanabhan, "Open cyberGIS software for geospatial research and education in the big data era," *SoftwareX,* vol. 5, pp. 1-5, 2016/01/01/ 2016, doi: https://doi.org/10.1016/j.softx.2015.10.003.

[27] S. Wang and M. F. Goodchild, *CyberGIS for geospatial discovery and innovation*. Springer, 2019.
[28] S. Wang, M. P. Armstrong, J. Ni, and Y. Liu, "GISolve: a grid-based problem solving environment for computationally intensive geographic information analysis," in *CLADE 2005. Proceedings Challenges of Large Applications in Distributed Environments, 2005.*, 24-24 July 2005 2005, pp. 3-12.
[29] S. Wang and Y. Liu, "TeraGrid GIScience Gateway: Bridging cyberinfrastructure and GIScience," *International Journal of Geographical Information Science*, vol. 23, no. 5, pp. 631-656, 2009/05/01 2009.
[30] M. McLennan and R. Kennell, "HUBzero: a platform for dissemination and collaboration in computational science and engineering," *Computing in Science & Engineering*, vol. 12, no. 2, p. 48, 2010.
[31] L. Zhao, Song, C.X., Kalyanam, R., Biehl, L., Campbell, R., Delgass, L., Kearney, D., Wan, W., Shin, J., Kim, I.L., & Ellis, C., "GABBs - Reusable Geospatial Data Analysis Building Blocks for Science Gateways," *IWSG*, 2017.
[32] R. Kalyanam *et al.*, "MyGeoHub—A sustainable and evolving geospatial science gateway," *Future Generation Computer Systems*, vol. 94, pp. 820-832, 2019.
[33] D. G. Tarboton *et al.*, "HydroShare: advancing collaboration through hydrologic data and model sharing," 2014.
[34] J. S. Horsburgh *et al.*, "Hydroshare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain," *JAWRA Journal of the American Water Resources Association*, vol. 52, no. 4, pp. 873-889, 2016.
[35] J. Heard *et al.*, "An architectural overview of hydroshare, a next-generation hydrologic information system," 2014.
[36] D. G. Tarboton *et al.*, "HydroShare: an online, collaborative environment for the sharing of hydrologic data and models," in *AGU Fall Meeting Abstracts*, 2013.
[37] Y. Y. Liu, D. R. Maidment, D. G. Tarboton, X. Zheng, and S. Wang, "A CyberGIS integration and computation framework for high-resolution continental-scale flood inundation mapping," *JAWRA Journal of the American Water Resources Association*, vol. 54, no. 4, pp. 770-784, 2018.
[38] S. Manson, J. Schroeder, David Van Riper, and S. Ruggles. "IPUMS National Historical Geographic Information System (NHGIS)." https://www.nhgis.org/user-resources/data-availability#boundary-files (accessed December 16, 2019).
[39] M. A. Palmer, J. G. Kramer, J. Boyd, and D. Hawthorne, "Practices for facilitating interdisciplinary synthetic research: the National Socio-Environmental Synthesis Center (SESYNC)," *Current Opinion in Environmental Sustainability*, vol. 19, pp. 111-122, 2016.
[40] A. Padmanabhan, D. Yin, F. Lyu, and S. Wang, "Bridging Local Cyberinfrastructure and XSEDE with CyberGIS-Jupyter," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*, 2019: ACM, p. 95.
[41] D. Yin *et al.*, "CyberGIS-Jupyter for reproducible and scalable geospatial analytics," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 11, p. e5040, 2019.
[42] D. Yin, Y. Liu, A. Padmanabhan, J. Terstriep, J. Rush, and S. Wang, "A CyberGIS-Jupyter framework for geospatial analytics at scale," in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, 2017: ACM, p. 18.

[43] B. Vandewalle, W. C. Barley, A. Padmanabhan, D. S. Katz, and S. Wang, "Understanding the multifaceted geospatial software ecosystem: a survey approach.," *International Journal of Geographical Information Science,* vol. under review, 2019.
[44] S. Wang, "Cyberinfrastructure. ," in *The Geographic Information Science & Technology Body of Knowledge* J. P. Wilson Ed., 2nd Quarter ed., 2019.
[45] "Unidata." https://www.unidata.ucar.edu/ (accessed December 16, 2019, 2019).
[46] "2010 Census Tract Reference Maps." https://www.census.gov/geographies/reference-maps/2010/geo/2010-census-tract-maps.html (accessed December 16, 2019).
[47] E. Mulrow, Becki Curtis, Ned English, Yongheng Lin, and I. Ventura., "Challenges in Linking Demographic Data at Different Geographic Levels. ," in *Poster presented at the Joint Statistical Meetings Conference,*, Baltimore, MD., 2017.
[48] J. Schroeder, "Hybrid Areal Interpolation of Census Counts from 2000 Blocks to 2010 Geographies," *Computers, Environment and Urban Systems,* vol. 62, 03/31 2017, doi: 10.1016/j.compenvurbsys.2016.10.001.
[49] J. P. Schroeder and D. C. Van Riper, "Because Muncie's Densities Are Not Manhattan's: Using Geographical Weighting in the Expectation–Maximization Algorithm for Areal Interpolation," *Geographical Analysis,* vol. 45, no. 3, pp. 216-237, 2013, doi: 10.1111/gean.12014.
[50] S. Ruggles, C. Fitch, D. Magnuson, and J. Schroeder, "Differential Privacy and Census Data: Implications for Social and Economic Research," *AEA Papers and Proceedings,* vol. 109, pp. 403-08, 2019, doi: 10.1257/pandp.20191107.
[51] M. Wilde, M. Hategan, J. M. Wozniak, B. Clifford, D. S. Katz, and I. Foster, "Swift: A language for distributed parallel scripting," *Parallel Computing,* vol. 37, no. 9, pp. 633-652, 2011.
[52] E. Deelman *et al.*, "Pegasus: A framework for mapping complex scientific workflows onto distributed systems," *Scientific Programming,* vol. 13, no. 3, pp. 219-237, 2005.
[53] "Apache Beam." https://beam.apache.org/ (accessed December 16, 2019).
[54] "Apache Airflow." https://airflow.apache.org/ (accessed December 16, 2019).
[55] Parsl. "Productive parallel programming in Python." https://parsl-project.org/ (accessed December 16, 2019).
[56] D. DiBiase *et al.*, "The new geospatial technology competency model: Bringing workforce needs into focus," *URISA Journal,* vol. 22, no. 2, p. 55, 2010.
[57] E. Shook *et al.*, "Cyber Literacy for GIScience: Toward Formalizing Geospatial Computing Education," *The professional geographer,* vol. 71, no. 2, pp. 221-238, 2019.
[58] C. Dony, A. Magdy, S. Rey, A. Nara, T. Herman, and M. Solem, "Encoding Geography RPP: Building Capacity for Inclusive Geo-Computational Thinking with Geospatial Technologies," in *2019 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, 2019.