Classification
Major category: Computer Sciences
Minor category: Environmental Sciences

The Emergence of Spatial Cyberinfrastructure

Dawn J. Wright[*†] and Shaowen Wang[‡]

[*]Department of Geosciences, Oregon State University, Corvallis, OR 97331-5506;
[‡]Department of Geography and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801

[†]To whom correspondence should be addressed. E-mail: dawn@dusk.geo.orst.edu

An Introductory **Perspectives** piece for submission to
Proceedings of the National Academy of Science
Special Feature on Spatial Cyberinfrastructure

Manuscript information: 9 text pages (equivalent to ~4 pdf pages as is customary for an introductory perspectives paper), no figures for the manuscript but a figure for the COVER of the issue is being submitted

Author contributions: D.J.W. and S.W. wrote the paper.

The authors declare no conflict of interest.

Key words: distributed computing, geographic information science, spatial computational domain, spatial analysis

Abstract
Cyberinfrastructure integrates advanced computer, information, and communication technologies to empower computation-based and data-driven scientific practice, and improve the synthesis and analysis of scientific data in a collaborative and shared fashion. As such, it now represents a paradigm shift in scientific research that has facilitated easy access to computational utilities and streamlined collaboration across distance and disciplines, thereby enabling scientific breakthroughs to be reached more quickly and efficiently. Spatial cyberinfrastructure seeks to resolve longstanding complex problems of handling and analyzing massive and heterogeneous spatial datasets, as well as the necessity and benefits of sharing spatial data flexibly and securely. This article provides an overview and potential future directions of spatial cyberinfrastructure. The remaining four articles of the special feature are then introduced and situated in the context of providing empirical examples of how spatial cyberinfrastructure is extending and enhancing scientific practice for improved synthesis and analysis of both physical and social science data. The primary focus of the articles is on spatial analyses employing distributed and high-performance computing, sensor networks, and other advanced information technology capabilities to transform massive spatial data sets into insights and knowledge.

\body

The term "cyberinfrastructure" (CI) was first coined by a National Science Foundation Blue-Ribbon Committee (1) to reflect how the traditional modes of scientific research (e.g., experimentation in the lab, observation in the field, processing/analyzing on a single calculator or computer, even calculations on the back of an envelope) are being enhanced and even revolutionized by the integrative capabilities of high-performance computers, storage and visualization tools for very large data sets, digitally-enabled sensors and instruments in the environment, virtual organizations for collaborative problem-solving, and interoperable suites of software services and tools (2). The world of scientific publishing is being transformed as part of CI evolution (3). CI therefore represents a paradigm shift in scientific research that has facilitated collaboration across distance and disciplines, thus enabling quick and efficient scientific breakthroughs that might not be possible otherwise.

Examples include the discovery of abrupt transitions in Earth's climate and ecosystem dynamics, previously unknown properties of minerals at extreme temperatures and pressures deep within the Earth, new simulations of the development of early Universe, new discoveries and insights through improved ocean models, new understandings of individual and group behavior and its relationship to social, economic and political structures, and the creation of a comprehensive human linkage genetic map (2, 4, 5). As Benioff *et al.* (6) note, computation, along with theory and experiment, has become the "third pillar" of science and engineering. And making new scientific discoveries requires the computational ability to synthesize and analyze very large data sets, integrated across biological, physical, and social sciences and engineering, and across the science-technology interface, where Hey *et al*. (5) name "data-intensive science" as the "fourth paradigm." Indeed, CI has become more than just hardware and software, but its own evolving area of research in the realm of data-intensive science and digital libraries (5—9), with many countries investing hundreds of millions of dollars in CI research and

development (10, 11), and calls from diverse scientific communities arguing the urgent need for further levels of CI investment (12, 13). Hey *et al.* (5) point out that while we have attained high-performance computing at affordable cost and have made good progress on simulation tools, many challenges remain in effectively integrating multiple field observatories containing thousands of instruments, involving millions of users, and petabytes of data, built on a true data grid with the ability to analyze data on that grid with sophisticated data analysis.

“Spatial CI” is an emerging term in the literature (14—16) and is defined as a specific type of CI that synergistically integrates the capabilities of CI, geographic information systems or GIS (e.g., 17, 18) and spatial analysis (19, 20) for geospatial problem solving and decision-making. By "spatial” or “space," we mean both real, physical space (i.e., on the surface of the Earth, in the atmosphere, or under the ocean), and virtual space (e.g., digital worlds, or understanding how and where computers are connected worldwide). Nearly all of our knowledge about the world can be classified according to space (location, area, distance or spatial interaction), as well as time. But while time is divided into the globally-understood units of seconds, hours, years, and so forth, spatial units and associated relationships are much more complex, multi-dimensional (e.g., x,y,z), at multiple scales and resolutions, often heterogeneous (even in the representation of a single variable), and always changing over time. Without a clear understanding of space, any associated models, structures, and hypotheses may be erroneous (especially those about relationships among variables).

In particular, the complexity of geographic space poses significant computational and intellectual challenges in distributed spatial data access, sharing, and analysis, government-sponsored spatial data information infrastructures (21), and the “geospatial semantic web” (22) (i.e., locating and integrating information without human intervention, including providing the ability to search for geographic information within web pages), all of which are part of a spatial CI. However, many of these challenges are already well known to those working on spatial data, and a variety of approaches not involving spatial CI have arisen to address these challenges. Spatial CI is going beyond these existing approaches by anchoring solutions in more sophisticated thinking about the representation and the implications of space, coupled with the latest in sophisticated mathematical and statistical models (e.g., 23—26), and forging more intimate collaborations between computer and information science and the domain disciplines of geography, geology and geophysics, oceanography, ecology, environmental engineering and sciences, and the social sciences, to name a few (e.g., 5, 8, 27—28). Such cross-disciplinary collaborations are making possible new knowledge systems that are leading to, at long last, a partial realization of a "Digital Earth,” as first envisioned by Vice President Al Gore (29), and now epitomized in products such as Google Earth, Microsoft Bing Maps, and NASA World Wind.

The deluge of spatial data collected in an accelerated pace in the foreseeable future from sensor networks, satellites, and even cell phones, continues to be driven by the tremendous needs of the aforementioned domains, and cannot be well used or understood unless they can be properly managed, analyzed, and shared through spatial CI. The dynamic nature of the Earth system (e.g., waves, tides, atmospheric turbulence, movements in the Earth’s crust) further complicates our efforts to accurately and

precisely measure the system. Massive datasets are common in the spatial analysis of human systems as well, including population and transportation systems, risk assessment, disease vectors, human mobility, and much more. Spatial analysis (broadly including spatial modeling) itself has traditionally encompassed a variety of approaches, including but not limited to spatial statistics (30, 31), heuristics and optimization (32, 33) and simulation for spatial problem solving and decision-making (34, 35). These methods, have been extensively applied in many fields (e.g., 36—39), but have been difficult to implement for large- and multi-scale problems that are computationally intensive and require collaborative input. This is a limitation that has existed despite the advances already made to deal with the challenges associated with the complexity of geographic space mentioned earlier. But spatial CI promises to remove this limitation and, thus, transform spatial analyses into powerful and accessible computational utilities for enabling widespread scientific breakthroughs. Spatial CI is also proving invaluable in the estimation of errors that propagate from measurements through to the analyses, and is facilitating the development of better models for error representation, propagation, and management throughout large, distributed computational networks (40).

The articles in this Special Feature address how the coupling of CI with spatial thinking and geographic approaches offers a promising path forward for solving scientific problems and improving decision-making practices of significant societal impact (e.g., assessing impacts of global climate change, understanding the complexity of coupled human-natural systems, sustaining ecosystem services, preserving and accessing digital resources in humanities and social sciences, and managing transportation infrastructure). They are far from inclusive of all aspects and current interests of spatial CI, as the field is growing quickly. But they are representative of current research addressing longstanding problems of the complexity of spatial datasets and spatial analysis, as well as the necessity and benefits of sharing spatial data flexibly and securely. This research highlights some of the new discoveries and insights that can be gained, results that could not have readily occurred without spatial CI.

Spatial Principles

The Special Feature begins with a technical treatment by Yang *et al.* (41) who examine the spatial principles governing the interaction of different parameters and phenomena in a variety of physical geographic studies (e.g., of the Earth's lithosphere, hydrosphere, atmosphere, pedosphere, and global flora and fauna patterns). Chief among them is the development of architecture and algorithms for distributed geographic information processing within a spatial CI (drawing in part upon spatial CI theory introduced by Wang and Armstrong (24)), to enhance the understanding of ecosystem dynamics and improve the forecasting of the onset and extent of dust storms in the U.S. southwest. As a result of the experiments, scientists were able to predict the onset of dust storms at higher resolutions (3 km x 3 km) over longer time periods (5-10 days).

Physical Science Applications

Helly *et al.* (40) describe the evolution of a set of methods and software tools to integrate multi-scale, -source, and -disciplinary oceanographic data over several recent research cruises to the Antarctic. Their initial goal was to investigate several scientific hypotheses about the movement of sea ice and meltwater plumes from icebergs, but an important parallel effort was the creation of a near real-time geospatial decision-support framework.

As they developed a spatial CI to support this framework, they were led to the innovation of a new sampling scheme, optimized to capture smaller scales of interest with respect to the broader scale of the study area. This sampling strategy overcame the limitations of the conventional sampling methods used previously (i.e., using a research ship as a static platform for sampling a single parameter on a station-by-station basis), thereby allowing for more rapid characterization of the surface of the ocean using multiple data streams at sea and in outer space, and simultaneously over multiple spatial and temporal scales. Thus, without the spatial CI, the authors would not have been able to make the first direct observation and characterization of meltwater plumes from individual icebergs and would not have been able to effectively integrate these individual results with regional- and global-scale data. The results lend new insights as to the influence of meltwater from icebergs on carbon flux from the surface of the ocean to sediments on the ocean floor, as well as the role that icebergs play in controlling biological productivity in the Weddell Sea. Their results also illustrate the importance of spatial CI in the overall scientific enterprise, and identify key architectural and design considerations in the development of current and future Earth observing systems, especially as oceanographers and other Earth scientists move into an era of petascale computing.

From Physical to Social Sciences and the Humanities

A goal of this Special Feature is to demonstrate that spatial CI is not only about hardware and software or enabling the physical sciences, but about "distributed knowledge communities" that serve the needs of the social sciences and humanities, as well as the multiple stakeholders and decision makers of citizen groups from differing social, economic, and political backgrounds. Building a CI is also very much a social endeavor, as well as scientific. As such, Sieber *et al.* (42) report on a spatial CI incorporating the China Biographical Database (the largest in the world), the China Historical Geographical Information System (part of China's original Electronic Cultural Atlas Initiative), and the McGill-Harvard-Yenching Library Ming Qing Women's Writings database. The study is one of the first to focus in general on a CI for humanities data, and specifically on a spatial CI that aids research on Chinese women writers, their kinship networks, publishing venues, and their literary and social communities. The article provides a critical examination of and recommendations on related issues of conflicting data that researchers may not necessarily want to eliminate, differing data models and geographic scales. This case study show the value of spatial CI in removing difficulties arising from spatial, but also multilingual, biographical, and temporal ambiguities in these databases, solutions that, again, would not be possible without spatial CI.

Buetow (4) notes that while team or "big science" will continue to be necessary to achieve research goals, the small, independent investigator is still "the engine of innovative research," and that the widespread adoption of CI will allow the two approaches to blend harmoniously. Poore (43) expands upon this theme in a final perspectives article on the needs and contributions of individual users within a spatial CI. The author notes that in particular, that as human geographers and other social scientists, as well as geographic information scientists, actively participate in spatial CIs as users, there is a great opportunity to make spatial CI a truly user-centered enterprise. Spatial CI should make room for not only the scientists who will use cybertools to collaborate at a distance, but also the educators who will teach with CIs. This also applies to "citizen

scientist" users who will contribute data and insights to CI projects on some of the most important scientific questions of the day, such as global climate change.

Concluding Perspective

Citizen scientists may, along with professional scientists, increasingly participate in the now ubiquitous "cloud computing," which uses service-oriented architecture to control the life cycle of virtual machines and data archives for everything from one's personal address book to the largest of multi-dimensional, multi-disciplinary scientific modeling systems. However, rather than federating autonomous entities (computing centers) into virtual organizations as computational grids do, clouds (Microsoft, Amazon, Google) instead focus on delivering infrastructure as a service, software as a service, and so on. Huge commercial investments in clouds make it likely that these systems will dominate large-scale computing hardware and software in the next decade (44, 45). Spatial CI is an important subset of the more general CI, spanning both the computationally intense and interdisciplinary usage requirements such as service hosting, virtual computing environments, and virtual data sets. The special requirements of spatial CI are a good match for the many common capabilities of clouds, thus warranting further fundamental and empirical research.

Indeed, the notion that "spatial is special" within CI introduces several interesting research challenges for physical and social scientists alike. Many geographic applications are interdisciplinary, and involve multiple stakeholders and decision-makers who have diverse social, economic, and political backgrounds, thereby making collaboration critical yet challenging. For example, how do we effectively and securely share and integrate spatial data, information, and analytical methods to develop and sustain evolving geographic knowledge? How do we facilitate collaborative spatial problem solving and decision-making through virtual organizations?

Given the promise of spatial CI, for some the effort in mastering it may still not be balanced by the apparent benefits, suggesting that the technology will always be the reserve of a highly technical group of experts. What will it take to popularize spatial CI beyond these experts, especially if it is to benefit the social sciences and humanities? Perhaps spatial CI will follow the path of GIS and eventually become as transparent as GIS is becoming in the world of Google Maps and Google Earth. Studies such as Yang et al. (40) and Poore (43) seek to distill the principles of spatial CI into simpler concepts that lend more obvious value-added to a broader range of users. Another approach may be to deal with conceptually- and computationally-unmanageable problems by dividing them spatially, understanding the resulting pieces, and then stitching the results back together. This divide-and-conquer approach, as initially popularized in the literature of computational geometry (e.g., 46), mirrors the way that society often solves its spatial problems. In the context of spatial CI, this implies spatially-heterogeneous data, and spatially-explicit consideration for parallel and distributed processing within individual high-performance computers and/or across the grid, and, again, clouds.

Although this Special Feature provides a small sampling of a much broader scientific and engineering enterprise, we hope it will help to elucidate some important issues and research questions, thereby accelerating scientific progress in this emerging area. As the size of spatial datasets and complexity of spatial analysis and modeling continues to

increase, and the need for virtual collaboration in scientific research becomes compelling, the transformative research to establish user-centric, efficient, and extensible spatial CI becomes ever more important and timely. The intellectual merits of spatial CI stem from the complexity of the challenges, the dangers inherent in not fixing the errors that may propagate, the profound need to develop solutions that will benefit many fields of societal relevance, the continuing vision of achieving access to a complete Digital Earth, and the next generation of GIS – CyberGIS – with integrative high-performance, distributed, and collaborative capabilities (25). We have sought to make the case that spatial CI leads to new discoveries in science, discoveries that would not have been made as readily without spatial CI. It is our hope that articles in this Special Feature have shown that spatial CI has facilitated such advances, and made them more replicable, more readily distributed, and certainly better visualized. It is only by advocating spatial CI that we will see the cyber-enabled approaches emerge that can make further scientific advances possible. We urge the scientific community to wait and see.

Acknowledgments

We thank all of the contributors to this special feature for their enthusiasm and skill in authoring these articles. We thank also the many reviewers for their thoughtful insights, which improved the manuscripts. We are grateful to our many colleagues in the Association of American Geographers (AAG) Cyberinfrastructure Specialty Group and in the University Consortium for Geographic Information Science (UCGIS) for valuable discussions and inspiration, as well as PNAS editorial board member Susan Hanson. And finally, we thank editors Michael Goodchild and David Stopak for their encouragement, assistance and helpful reviews. This material is based in part upon work supported by the National Science Foundation under Grant Number BCS-0846655. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Atkins DE, et al. (2003) *Revolutionizing Science and Engineering through Cyber-infrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, National Science Foundation Publication NSF0728 (National Science Foundation, Washington, DC), 84 pp.

2. Crawford D, et al. (2007) *Cyberinfrastructure Vision for 21st Century Discovery*, National Science Foundation Publication CISE051203 (National Science Foundation, Washington, DC), 57 pp.

3. Renear A, Palmer CL (2009) Strategic reading, ontologies, and the future of scientific publishing. *Science* 325:828-832.

4. Buetow KH (2005) Cyberinfrastructure: Empowering a "third way" in biomedical research. *Science* 308:821-824.

5. Hey T, Tansley S, Tolle K, eds (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, Redmond, WA), 286 pp.

6. Benioff MR, et al. (2005) *Computational Science: Ensuring America's*

Competitiveness, Report to the President by the President's Information Technology Advisory Committee (National Coordination Office for Information Technology Research and Development, Arlington, VA), 116 pp.

7. Wang S, Zhu X-G (2008) Coupling cyberinfrastructure and geographic information systems to empower ecological and environmental research. *BioScience* 58: 94-95.

8. Hey T, Trefethen AE (2005) Cyberinfrastructure for e-science. *Science* 308:817-821.

9. Foster I (2005) Service-oriented science. *Science* 308:814-817.

10. Bohannon J (2005) Grassroots supercomputing. *Science* 308:810-813.

11. Clery D, Voss D (2005) All for one and one for all. *Science* 308:809.

12. Lazowska ED, Patterson DA (2005) An endless frontier postponed. *Science* 308: 757

13. Craglia M, et al. (2008) Next-generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. *Int J Spatial Data Infrastructures Res* 3:146-167.

14. Blais JAR, Esche H (2008) Geomatics and the new cyberinfrastructure. *Geomatica* 62:11-22.

15. Wang S, Liu, Y (2009) TeraGrid GIScience gateway: Bridging cyberinfrastructure and GIScience. *Int J Geog Inf Sci* 23:631-656.

16. Zhang T, Tsou M-H (2009) Developing a grid-enabled spatial Web portal for Internet GIServices and geospatial cyberinfrastructure. *Int J Geog Inf Sci* 23:605-630.

17. Maguire DJ, Goodchild MF, Rhind DW, eds (1991) *Geographical Information Systems: Principles and Applications* (Wiley, New York), 1056 pp.

18. Gewin V (2004) Mapping opportunities. *Nature* 427: 376-377.

19. Taaffe EJ (1974) The spatial view in context. *Annals Assoc Am Geogr* 64:1-16.

20. Miller HJ (2004) Tobler's First Law and spatial analysis. *Annals Assoc Am Geogr* 94:284-289.

21. Onsrud H, ed (2007) *Research and Theory in Advanced Spatial Data Infrastructure Concepts* (ESRI Press, Redlands, CA), 306 pp.

22. Egenhofer M (2002) Toward the geospatial semantic web, in *Advances in Geographic Information Systems, International Symposium*, eds Makki Y, Pissinou N (Association for Computing Machinery, McLean, VA) pp 1-4.

23. Anselin L, Florax R, Rey S, eds (2004) *Advances in Spatial Econometrics: Methodology, Tools and Applications* (Springer-Verlag, Berlin), 513 pp.

24. Wang S, Armstrong M (2009) A theoretical approach to the use of cyberinfrastructure in geographical analysis. *Int J Geog Inf Sci* 23:169-193.

25. Wang S (2010) A cyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals Assoc Am Geogr* 100(3): 535-557.

26. Penninga F, Van Oosterom PJM (2008) A simplicial complex-based DBMS approach to 3D topographic data modelling. *Int J Geog Inf Sci* 22:751-779.

27. Baker KS, Chandler CL (2008) Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. *Deep-Sea Res II* 55(18-19): 2132-2142.

28. Yang C, Raskin R, Goodchild M, Gahegan M (2010) Geospatial cyberinfrastructure: Past, present and future. *Comput Environ Urban Sys* 34:264-277.

29. Gore A (1999) The Digital Earth: Understanding our planet in the 21st Century. *Photogr Engr Remote Sens* 65:528.

30. Anselin L (1989) *What Is Special about Spatial Data? Alternative Perspectives on Spatial Data Analysis*, Technical Report 89-4 (National Center for Geographic Information and Analysis, Santa Barbara, CA), 10 pp.

31. Anselin L (1995) Local indicators of spatial association-LISA. *Geogr Anal* 27: 93-115.

32. Turton I, Openshaw S (1998) High-performance computing and geography: Developments, issues, and case studies. *Env Planning A* 30:1839-1856.

33. Krzanowski R, Raper J, eds (2001) *Spatial Evolutionary Modelling* (Oxford University Press, Oxford), 304 pp.

34. Batty M (2005) *Cities and Complexity: Understanding Cities Through Cellular Automata, Agent-Based Models, and Fractals* (MIT Press, Cambridge, MA), 565 pp.

35. Jankowski P, Nyerges T (2001) GIS-supported collaborative decision making: Results of an experiment. *Annals Assoc Am Geogr* 91:48-70.

36. Goodchild MF, Janelle DG, eds (2004) *Spatially Integrated Social Science* (Oxford University Press, Oxford), 480 pp.

37. Miller HJ, Wentz, EA (2003) Representation and spatial analysis in geographic information systems. *Annals Assoc Am Geogr* 93:574-594.

38. Wright DJ (2009) Spatial data infrastructures for coastal environments, in *Remote Sensing and Geospatial Technologies for Coastal Ecosystem Assessment and Management*, Lecture Notes in Geoinformation and Cartography, ed Yang X (Springer-Verlag, Berlin), pp 91-112.

39. Tesfatsion L (2002) Agent-based computational economics: Growing economies from the bottom up. *Artificial Life* 8:55-82.

40. Helly J, Kaufman RS, Vernet M, Stephenson GR (2011) Cyberinfrastructure in oceanography: Multi-source characterization of the melt-water field from icebergs in the Weddell Sea. *Proc Natl Acad Sci USA* 108:XXX-XXX.

41. Yang C, Wu H, Huang Q, Li Z, Li J (2011) Spatial computing: Using spatial principles to optimize distributed computing for enabling physical science discoveries. *Proc Natl Acad Sci USA* 108:XXX-XXX.

42. Sieber R, Wellen C, Jin Y (2011) Spatial cyberinfrastructures, ontologies, and the humanities. *Proc Natl Acad Sci USA* 108:XXX-XXX.

43. Poore B (2011) Going cyber: Users as essential contributors to spatial cyberinfrastructures. *Proc Natl Acad Sci USA* 108:XXX-XXX.

44. Fox A (2011) Cloud computing—What's in it for me as a scientist? *Science* 331:406-407.

45. Gao X, Ma Y, Pierce M, Lowe M, Fox G (2010) Building a distributed block storage system for cloud infrastructure. *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science* (CloudCom 2010), 1-9.

46. Preparata FP, Shamos MI (1990) *Computational Geometry: An Introduction* (Springer-Verlag, Berlin and Heidelberg GmbH & Co. K), 412 pp.