Two Socio-Technical Gaps of Cyberinfrastructure Development and Implementation for Data-Intensive Collaboration and Computational Simulation in Early e-Science Projects in the U.S.

Kerk F. Kee
Chapman University
One University Drive, Orange, CA 92866
kerk.kee@gmail.com

Larry D. Browning
University of Texas, Austin
2504A Whitis Ave., Austin, TX 78712
lbrowning@mail.utexas.edu

ABSTRACT

This position paper explores two socio-technical gaps that hamper effective synergy between human and machine intelligence as well as data-intensive and cognitively complex sense making and decision making in early U.S. e-science projects. Based on qualitative methods guided by grounded theory and 70 interviews, we identify limiting constrains at the levels of scientists, technologists, science funders, and technology quality, under the participatory and bespoke natures of cyberinfrastructure. More specifically, scientists cannot fully envision the new computational tools they need for their science (i.e., human intelligence in domain science), and technologists have limited knowledge of the domain science they are building cyberinfrastructure/computational tools for (i.e., representing the machine intelligence built into computational tools). Furthermore, science funders offer limited direct funding for cyberinfrastructure development, and this development is time consuming and it produces unstable software (i.e., thus affecting the quality and development of machine intelligence). A closer examination of these key limitations led to the identification of two gaps: the *specialization-synergy gap* between scientists and technologists, as well as the *science investment-technology quality gap* between funders and development. We argue that these two gaps hamper the effective cognitively-complex sense making and decision making in data-intensive e-science projects.

Author Keywords

Cyberinfrastructure; e-science; participatory and bespoke natures of cyberinfrastructure; socio-technical gaps.

ACM Classification Keywords

K.4.3 [Organizational Impacts]—Computer-supported collaborative work, H.5.3 [Group and Organization Interfaces]—Organizational design, Computer-supported cooperative work.

INTRODUCTION

The grand vision of cyberinfrastructure (CI) is to enable large-scale research using aggregated computational resources and combined datasets through the Internet, high-performance networks, and local machines and to be able to mine publicly-funded datasets accumulated over time. Data intensiveness if a key defining characteristic of CI and e-science projects [1]. When the grand vision described is achieved, there will be increased productivity and breakthrough discoveries in research. However, like most innovations, an ambitious endeavor such as CI implementation for data-intensive collaboration and computational simulation often encounters challenging conditions in the general environment. In this paper, we explore two socio-technical gaps that present challenging conditions to the synergy between domain scientists (i.e., the key human actants in e-science projects) as well as computational technologists (i.e., whose logic, algorithm, and design become of the machine intelligence for e-science) and science funders (i.e., whose funding determines the quality of CI development). As a preview, the two gaps are the *specialization-synergy gap* and the *science investment-technology quality gap*. These gaps stemmed from the participatory and bespoke natures of CI.

PARTICIPATORY AND BESPOKE NATURES OF CYBERINFRASTRUCTURE

This paper examines CI development and implementation for data-intensive collaboration and computational simulation among pioneering scientists, particularly those who work with computational technologists to develop new CI tools. This group of pioneering scientists adopts CI at the conceptual level and gives it meaning when they submit a grant proposal (often jointly with computational technologists). Once funded, they work with technologists to co-produce CI tools that do not yet exist. In other words, pioneering scientists adopt CI as a possibility, not as a fully developed tool in the physical sense. This participatory

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.
Copyright 2012 ACM 978-1-4503-1086-4/12/02...$10.00.

characteristic of CI makes it a unique case for studying innovation implementation. [Note: a separate paper focusing on scientists who develop their own CI tools and scientists who adopt existing tools has been submitted to *W2: Data-Intensive Collaboration in Science and Engineering*].

In CI co-production, a scientist presents a scientific problem and a technologist explores ways to create a tool to help investigate it. A CI project manager in Indiana explains, "We [technologists] have to understand enough of their [scientists'] problem to be able to understand ourselves where computers can help. And then do the best we can to explain that to them [scientists] and give them options". This is a critical process that this project manager repeated during the same interview, "I can't do my work unless the domain scientist is willing to take the time to explain necessary things to me. The domain scientists can't do their work unless I can build them the right tools". Therefore, CI co-production is often driven by a problem within a particular scientific context and domain. Furthermore, because the research is also primarily defined by the scientist, the tool produced is based on the approach (i.e., theory, measurements, methodology, etc.) employed by him or her. Hence, Kee and colleagues describe CI as a bespoke innovation [4].

DATA COLLECTION AND METHOD

We conducted interviews over a period of 32 months, from November 2007 to June 2010. The data set includes 70 interviews with 66 participants from across 17 U.S. states and three other countries. The interviews were spread across four years with 10 participants in 2007, 42 in 2008, 16 in 2009, and two in 2010. Because most of the interviews were conducted in 2008, the analysis primarily reflects CI development and implementation for data-intensive collaboration and computational simulation during this period. The shortest interview was 15 minutes and the longest was 2 hours and 16 minutes. The interviews averaged approximately one hour each and were conducted in person with 19 of the participants and over the phone with the remaining 51 participants. All the interviews were audio recorded except for two, due to technical difficulty and following one participant's request. However, notes were taken immediately after these two interviews.

The 66 interview participants came from Texas (12), Illinois (11), California (10), Michigan (5), Indiana (4), Massachusetts (3), Arizona (2), Colorado (2), Louisiana (2), Washington (2), DC (1), Maryland (1), New York (1), Virginia (1), Ohio (1), Pennsylvania (1), Delaware (1), as well as Australia (2), Germany (1), and the UK (1). The geographic affiliations refer to the primary locations of the participants at the time of the interviews. Participants include 52 males and 14 females. Participants' primary professional roles were diverse, including domain scientists who used CI to conduct science (15), computational technologists who built CI (12), a range of administrative directors and program managers at supercomputer centers and national research laboratories across the country (21), US NSF program officers who helped allocate funding to CI projects (4), social scientists and policy analysts who studied and participated in CI projects (12), and experts from commercial industry (2).

This paper employs grounded theory in qualitative data collection and analysis. According to Corbin and Strauss, one of the main goals of grounded theory is to seek to uncover relevant conditions under which social phenomena manifest [2]. We argue in this paper that the relevant conditions for the two socio-technical gaps are the participatory and bespoke natures of cyberinfrastructure and e-science projects discussed earlier.

THE TWO SOCIO-TECHNICAL GAPS

The two gaps identified in this paper are socio-technical in nature because they manifest during the interactions between the main social actants (i.e., the domain scientists as primary users) and the technical tools (i.e., the logic, algorithm, design, quality, etc., resulted from computational technologists and science funders). These two gaps hamper effective synergy between human and machine intelligence, in which we also argue that machine intelligence is the result of other key social actants, such as computational technologists and science funders. The next paragraph begins with describing the first gap.

Specialization and Synergy Gap

There is a critical gap between participating in the co-production process as specialists and the need to achieve synergy under the time constraint of a funded e-science project. In this paper, this socio-technical gap is referred to as the *specialization-synergy gap*. As CI is a participatory innovation, the first part of this gap involves specialists building CI/computational tools. CI tool development is driven by a scientific problem pursued by domain scientists (as users of the tool). Therefore, the development process requires the participation of specialized scientific experts in the domain. On the other hand, a CI tool is a piece of technology that cannot be bought off-the-shelf commercially. Its production requires specialized technologists who are skilled at high-performance, distributed, and parallel computing techniques and knowledge. Their techniques and knowledge are then built into the fundamental logic and design of the machine intelligence, which interacts with domain scientists during CI development and implementation for data-intensive collaboration and computational simulation. One often has to focus on a very narrowly defined area of knowledge in order to become a specialist in domain science, computing techniques, etc.

The second part of the gap is the need for synergy because CI is a bespoke innovation custom-made for specific scientific problems. Synergy can be defined as the process or mechanism that enables collaborative advantages [6] among diverse specialists and participants. In order to achieve synergy among specialists, there needs to be a common language, shared understanding of basic concepts, motivation to learn, and the ability to see the outcome, all of which is impossible without synergy. However, it takes time for these four elements to fully develop before true synergy can positively impact CI co-production. As these elements cannot fully develop when projects are funded for on a limited time basis, such as a three-year or five-year term, the development of CI tools is compromised. The analysis continues with examples offered by scientists and technologists who have experienced the specialization-synergy gap.

Scientists. A scientist's ability to envision what is possible for a CI tool is limited by his/her ability to see the outcome, and this is impossible without synergy. More specifically, it is difficult for them to recognize and articulate their needs. These are new needs yet to be clearly defined for new CI tools. Because pioneering scientists are in the stage of prototyping tools, co-production becomes an experimental process in which scientists try to determine what is even possible. In other words, co-production can be interpreted as setting an agenda for cyberinfrastructure and science, involving the work to "define what cyberinfrastructure should be" (see interview quote below). A pioneering cyberinfrastructure adopter and a water resources engineering professor at Illinois shares her co-production experience working with a leading supercomputer center in the country, "I've been… trying to figure out what they are,… working with [the center] to try to understand the cyberinfrastructure needs … We're… prototyping and developing what might be possible,…to define what cyberinfrastructure should be and to prototype early cyberinfrastructure".

Pioneering co-production is time-consuming because the exploratory process requires working closely together often in face-to-face interactions. Lee and Bietz argue, "[m]ost scientists [are] reluctant to invest more than a very small amount of time to learn to use new technologies unless the benefits [are] substantial and related directly to their research" [5, p. 3]. A big part of this limitation experienced by scientists could be the result of not knowing enough about computational science to envision the potential. Therefore, the co-production is an exploratory process "trying to sort of make a match" between the needs to be discovered and the tools to be developed. A social scientist in Michigan observes, "It is time-consuming… They [technologists and scientists] really work closely… face-to-face to try to… make a match in a way between their [scientists'] needs, needs that they don't really know they have yet… so that they can do different things with their science. This quote implies that scientists often do not enter into co-production with a clear design, product, and outcome in mind; they explore and figure out the tool with technologists in a time-consuming process.

These excerpts reveal that scientists often experience difficulties in envisioning new CI tools for their science. This difficulty comes from not fully knowing what is possible for their science, and not fully knowing their needs yet. Therefore, CI tools remain in the prototype stage, and co-production is very time intensive. Technologists also experience challenges in specialization-synergy gap.

Technologists. While scientists are constrained by their ability to articulate clear needs and envision what is possible, technologists appear to be limited by their knowledge of domain science and their lack of motivation to acquire it quickly. It is understandable that most technologists (as specialists in computer science and computational techniques) do not often speak the language and understand the basic concepts of science. However, not having enough scientific knowledge can lead to developing tools that do not match the scientific problems to be investigated. Furthermore, acquiring the necessary language and concepts for synergistic communication and participating in ongoing communication both take a lot of time. A technologist in Indiana explains, "We speak different languages. It's very easy for the [technologists] to do something that turns out to be nonsensical because we don't understand the science… The domain scientists get frustrated… because [they] don't necessarily understand the complexity involved in it".

The challenge of having technologists who are unfamiliar with the science is that the co-production process often requires repetition of cycles and revisiting the design. As technologists acquire scientific knowledge in the process, they often need to re-work some early recommendations and/or parts previously built. The result is a delay and slow progress in CI projects. Moreover, the learning process can be impaired or hampered if technologists are not personally interested in or motivated by the field of science because their participation in the co-production process is short-term and they do not actually work under their scientist counterparts. A geochemist in New York reflects on her experience, "You had to go back many times… they said – Oh, if this is that way, then we cannot do it here…. They were not really that motivated and that enthusiastic about learning it. It was just part of the work".

These quotes about the technologists explain that they often do not possess either the language or the concepts of the science they are building tools to address. It becomes more challenging if technologists are not personally motivated to learn the science. Because their participation is often on a short-term basis and many of them do not work long-term and full-time for the scientists and/or in the field of science they temporarily serve, the *specialization-synergy gap* is likely to persist.

Due to the early stage of CI prototyping, co-production is very time-consuming because of the specialization-synergy gap demonstrated. Because CI implementation implies co-producing and prototyping CI tools with technologists, implementation can negatively impact a scientist's research productivity, which requires cognitively-complex sense making and decision making.

The specialization-synergy gap is a critical challenge for CI implementation because it compromises the CI tools developed to explore important scientific problems and grand challenges. It also has a detrimental impact on the research productivity of the pioneering adopters.

Science Investment-Technology Quality Gap

The second gap, the *science investment-technology quality gap*, involves funding patterns and software developments in co-production. Kee and Browning argue that as CI is a participatory and bespoke innovation, successful implementation is often impeded by a gap between science funding and software development [3]. This often leads to negative impacts for this generation of scientists who are ushering in the cyberinfrastructure vision.

Science Funders. The first part of the gap, science investment, represents a funding condition under which CI has emerged. However, the investment pattern to fund science also represents two important challenging conditions for CI development and implementation for data-intensive collaboration and computational simulation. The first challenge is that there is not a dedicated public funding source for long-term and sustainable software development to support CI development. Funders, such as NSF, are set up to fund science, and limited on flexibility to fund technology. Because there is not a 'National Technology Foundation', software development for CI is often compromised.

The second challenge is that there is limited long-term outlook for most science (and e-science) projects funded by agencies such as NSF. A technologist from Indiana states, "Governments tend not to pour money into areas more than five or at most 10 years… Often NSF will fund the initial steps, but not the long-term sustainability". Due to these two challenging conditions, the lack of direct and stable funding for software development and the lack of long-term funding for projects, CI development, adoption, implementation for data-intensive collaboration and computational simulation based on short-term funding for science creates a challenge for quality tools.

In the short-term and science-oriented funding environment, software development for CI is often assumed to be free and/or a part of a science project that does not require direct financial support. A senior administrator who retired from a major university in California shares, "The instability of the software is due to… [people who] make funding decisions just don't think of software as something requiring a big long-term ongoing investment. It's nice to think of it as somehow being free… [it] gets created and maintains itself". This is an inherent challenge when CI is a bespoke innovation based on specific scientific problems to be funded by federal agencies.

NSF funds projects that rely on open source platforms and CI users often turn to open source software as the primary preference for software development under these funding conditions. The rationale is that the software will be created as part of the process of science and that the scientific community will maintain it as a public good. As a project manager in California explains, "Most of the organizations the NSF is involved with are using open source. [It] is more a part of cyberinfrastructure than Windows technology… They appear to be less expensive because you don't have to pay for the software licensing". However, relying on open source software development also encounters challenges that will be discussed next.

Technological Developments. The first challenge in the area of CI software development is the dependence on open source platforms. While these are free and the open source philosophy also sits well with academics, it does not have many pre-determined and standard solutions to known problems (i.e., turnkey solutions) to speed development. This disadvantage poses a challenge to timely technological development. The project manager in California quoted earlier continues, "Our project didn't come in on time because instead of using…open source software, we should have migrated to Windows technology. We would have had more turnkey solutions that… the existing team could have produced software more quickly". In the same interview he continues to explain that using open source software also requires more programmers due to the same limitation of a lack of turnkey solutions. While the open source approach has its inherent benefits, including pooling a wide range of ideas, knowledge, and expertise [8] and organizing beyond traditional boundaries [9], the organic and emergent nature of an open source approach cannot always guarantee effective outcomes under critical time pressure.

The second challenge under the current funding condition is that software development is both rushed and unstable. As previously discussed, there is no long-term funding for software development. Funded scientists and technologists have to reapply for funding in order to sustain a CI and e-science project. To secure the next round of funding, sometimes there is a need to rush through the development process. A technologist and a professor of computer science in Louisiana states, "There's so much to do and there's a tendency to rush the job, try and get systems into place, try to get scientists using them… before they're really ready to be used because of the pressure to have continued funding". This pattern affects technology quality.

The third challenge that results from the funding conditions is that software development can hinder scientific investigation since the tools are neither robust nor fully developed when applied. In other words, the science that was originally funded can be compromised when the tools that were supposed to enable the research investigation are actually experimental tools being prototyped in the process of serious research. Therefore, both technological development and the scientific research will be compromised. A technologist at a commercial software company in Washington shares a compelling argument that illuminates this problem:

> We saw too many projects in the past that – where experimenting with technology at the expense of helping scientists do science. So cyberinfrastructure should be… services and technologies and computing-related infrastructure that just work… That, to me, is the critical factor of a success…. It shouldn't get in the way [of science]…. Cyberinfrastructure really has to not get in the way... [Cyberinfrastructure projects] have tried to come up with new ideas while being applied during [e-science]… Those new ways should not be part of a science-related project or a new scientist-created project. Those should be on their own. Only when they prove themselves, then we can apply them to e-science… The same way that the industry will not use experimental methods in order to do critical business tasks, in the same way, we shouldn't be using experimental methods to do critical to science tasks.

This excerpt reveals the unfortunate outcomes of many cyberinfrastructure co-production efforts, thus the actual implementation of CI tools and data-intensive collaboration. Since scientists adopt cyberinfrastructure at the conceptual level before the tools exist, efforts to bring forth cyberinfrastructure compromises the science the technology was supposed to serve. This is a direct outcome of cyberinfrastructure being a participatory and bespoke innovation. Scientists have to be active participants of the development of the CI tools they have adopted conceptually. The science investment-technology quality gap is the result of CI development being dependent on short-term funding for science and the CI tools being developed in the middle of a serious research project.

CONCLUSION

The co-production process of CI/computational tools for data-intensive collaboration and computational simulation often involves exploration of what is possible, an attempt to match an emerging need with technological development, learning of the science, iterative cycles, finding solutions of open source software developments, and working on a limited short-term budget for science, and producing unstable tools that face the possibility of no sustaining NSF funding or community support. The participatory and bespoke natures of cyberinfrastructure hampers effective cognitively complex sense making and decision making for important scientific endeavors. The specialization-synergy gap and the science investment-technology quality gap complicate the process.

We argue that while examine the synergy between human and machine intelligence in CI, e-science, data-intensive collaboration, and computational simulation, it is important to recognize the role of computational technologists and science funders as they impact the logic and design CI/computational tools, which represent the material actants that interact with domain scientists during data-intensive collaboration and computational simulation.

ACKNOWLEDGMENTS

We thank all the interview participants for their insights.

REFERENCES

1. Bird, I., Jones, B., & Kee, K. F. The organization and management of grid infrastructures. *Computer, 42,* 1 (2009), 36-46.
2. Corbin, J., & Strauss, A. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13,* 1 (1990), 3-21.
3. Kee, K. F., & Browning, L. D. The dialectical tensions in the funding infrastructure of cyberinfrastructure. *Computer Supported Cooperative Work, 19,* 3-4 (2010), 283-308.
4. Kee, K. F., Cradduck, L., Blodgett, B., & Olwan, R. Cyberinfrastructure inside out: Definitions and influencing forces shaping its emergence, development, and implementation. In D. Araya, Y. Breindl & T. Houghton (Eds.), *Nexus: New intersections in Internet research* (pp. 157-189). New York: Peter Lang. 2011.
5. Lee, C. P., & Bietz, M. *Barriers to the adoption of new collaboration technologies for scientists*. Paper presented at the ACM Conference on Computer-Human Interaction (CHI), 2009.
6. Roz, D. L., Elisa, S. W., & Rebecca, M. Partnership Synergy: A Practical Framework for Studying and Strengthening the Collaborative Advantage. *The Milbank Quarterly*, *79*, 2 (2001), 179-205.
7. Shirky, C. *Here comes everybody: The power of organizing without organizations*. New York: Penguim. 2009.
8. von Hippel, E. *Democratizing innovation*. Cambridge, MA: The MIT Press. 2005.