Scientific Research Publishing

Creative Education, 2019, 10, 524-534
http://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Big Data—Building Software: Some Thoughts on the Future of Building Sciences

Winifred Elysse Newman

Department of Architecture, Clemson University, Clemson, SC, USA
Email: elyssen@clemson.edu

How to cite this paper: Newman, W. E. (2019). Big Data—Building Software: Some Thoughts on the Future of Building Sciences. *Creative Education, 10,* 524-534.
https://doi.org/10.4236/ce.2019.103038

Received: February 11, 2019
Accepted: March 3, 2019
Published: March 6, 2019

Abstract

Building and sustaining our communities is the most significant challenge of our day. This means seeing and understanding the interrelatedness of economies, technologies, physical and natural systems. One way of achieving this objective is visualizing our options using big-data decision spaces where unique systems are represented by dynamic flows of data create patterns, relationships and context. As seminal computer scientist Jim Gray articulated in 2003, "big data describes on the one hand extremely large data sets that through analysis generate patterns and associations; on the other hand, it can be understood to encompass all potential data generated by any dynamic process, in short life on Earth as we know it". This kind of ecological "data literacy" let us see beyond a mechanistic view of the phenomenal world in which the imagined universe is a machine composed of elementary building blocks, to a system view where the dominant model of the material world is a network of systems with interrelated and interdependent patterns of behavior. Adopting the systems' approach the challenge in designing resilient buildings is: 1) educating architects and engineers capable of using multiple and varied data to create multipart decision spaces, 2) assessing the impact visualizing data will have on all aspects of the building design, construction and fabrication process and 3) making increasingly large data-sets available to the AEC professionals and researchers to asses building performance and its impact on our design, delivery and maintenance of the built environment available.

Keywords

Big-Data, Systems, Architecture, Engineering, Building Design

1. Introduction

In January of 2007 pioneering computer scientist Jim Gray gave one of his last talks to the Computer Science and Telecommunications Board where he outlined a vision of a "fourth paradigm" in scientific research. The fourth paradigm according to Gray is a methodological approach to discovery based on data-intensive science beyond experimental and theoretical research and computer simulations of natural phenomena. This isn't new per se as increasingly complex and larger amounts of data can arguably be considered part of the drive to empiricism in the 19th and the early 20th century science. What is new and as yet unclear—how the new paradigm will change fundamental modes of knowing and acting in the sciences and technology in the 21st century? The transition from material or phenomena-based research to systems research is relatively new in relation to the amount of data, so called big-data, available through the Internet-of-Things, supercomputing, and the like. This paper looks at the fourth paradigm and its relationship to *e*-science, potential modes of knowing and their bearing on architecture, engineering and construction, and the AEC of building sciences. We discuss the nature of transition from mechanical to systems science, the definitions of *e*-science and computational thinking and the potential impact this will have on the way to manage big data in the building sciences.

2. System Science

One of the fundamental conditions of the present is a growing interconnectedness and interdependence of data. Big-data describes on the one hand extremely large data sets that through analysis generate patterns and associations; on the other it can be understood to encompass all potential data generated by any dynamic process, in short life on Earth, as we know it (Hey, 2009). This means a shift in scientific methodology where the dynamic flow of data creating patterns, relationships and context is the lens through which we study a given phenomenon, not an initial hypothesis based on inductive or deductive reason. The data is the starting point, not the substantiation of a theory. This kind of ecological literacy let us see beyond the mechanistic view in science where the universe was a machine composed of elementary building blocks. Let's take a moment to compare the *systems* approach in the sciences to the *mechanistic* view.

Enlightenment to the late 19th century science was empirical. The primary drive was to describe natural phenomena through observation and study. Science was generally performed under the so-called mechanical or mechanistic view. These assumed phenomena are isolated, able to be categorized independently with clear part to whole relationships between entities that are measurable. As Vannevar Bush outlined in his groundbreaking essay, *As We May Think* (1945), there is an underlying assumption of the self-organizing determinism of phenomena with a purposive behavior and teleology. Atoms and molecules in classical physics and chemistry relied on the notion anonymous particles moving at random give rise by their multiplicity to a statistical order and regularity. This compelled a focus on the cause/effect binary and an almost unavoidable tendency to classify everything into discrete parts analyzable only by separating them from context so to identify means and ends. To grow beyond the methodological linearity inherent in this analysis, science had to change and in Kuhnian terms, a paradigm shift is visible by the mid-20th century (Bush, 1945). To fill the lacunae of theories capable of addressing some of the problems posed by modern technology and instrumentation, science moved away from the teleology of the mechanistic model to an analytical science of systems where generating simulations and models to observe interactions between complex behaviors, predict outcomes, and generate alternatives is the norm.

Systems theory formally emerged with the publication of Ludwig Bertalanffy's *General System Theory* (1928) but arguably noted earlier in Leibniz and the monadology, Nicholas of Cusa's coincidence of opposites, and Hegel's dialectical view of history. The history isn't as important here as the terms of the new paradigm. Mechanistic analytical procedure means entities under examination can be assembled and re-assembled from their parts and the part to whole relation will bear the condition of summativity where equations describing the behavior of the total are of the same form. As Bertalanffy notes, these conditions cannot be met in systems consisting of parts in interaction where the prototype of their description is a set of simultaneous nonlinear differential equations in the general case (EIA). Systems, in contrast, are "organized complexity" (Von Bertalanffy, 1968). The problem for systems theory is to address relations between parts and wholes "in interaction"; meaning they are nonlinear or inherently dynamic, occur in time (metabolic processes, growth a decay), are open or closed (developmental processes), and rely on the calculus. Graph-, compartment-, set- and net-theory are subsets of systems theory requiring specialized mathematics to describe structural, topological or quantitative relations. Cybernetics is technically a subset of systems theory dealing with the control of mechanisms in technology and nature and founded on the concepts of information and feedback (Von Bertalanffy, 1968).

3. *e*-Science and the Fourth Paradigm

In this next section we turn to the relation between *e*-science and big data generated when computing meets systems theory. Jim Grey argues the fourth paradigm follows on the experimental, theoretical and more recent computational science. It is the overwhelming amount of observational and computationally generated data requiring new ways to manage how we analyze, visualize and store data, including data generated in the process of analyzing other data. This is relatively new and emerges along with ubiquitous computing available from the 1970s onward making models of increasing robustness and complexity possible in systems theory. Once established there is intensifying reliance on processed data: data captured by instruments and processed by software, stored in computers and managed by statistics, or metadata or both. The new paradigm shift parallels the development of *e*-science where "IT meets scientists" (Hey, 2009). The exemplar case is cybernetics. Three fundamental contributions appear at about the same time: Wiener's *Cybernetics* (1948), Shannon and Weaver's *Information theory* (1949) and von Neumann and Morgenstern's *Game*

Theory (1947). Wiener's *Cybernetics* and the others lead to developments of computer technology, information theory, and self-regulating machines. Wiener in particular carried the cybernetic, feedback and information concepts far beyond the fields of technology and generalized it in the biological (EIA). Concurrent to the emergence of systems theory and cybernetics engineer and inventor Vannevar Bush remarked,

"There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—*conclusions which he cannot find time to grasp*, *much less to remember*, *as they appear*. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial" (Italics by Author) (Bush, 1945).

Bush tried to imagine the next computation machine to manage the data needed to feed the systems analysis—reminding his reader that "such machines will have enormous appetites. One of them will take instructions and data from a whole roomful of girls armed with simple keyboard punches and will deliver sheets of computed results every few minutes. There will always be plenty of things to compute in the detailed affairs of millions of people doing complicated things". References to the human and very female "computers" of his time aside, Bush could only imagine the later generations of machines doing computations *faster*—meaning improvements were improvements of degree, not type. The truth is more startling: in the research world today *data itself* is the phenomena, not a phenomenal subject: the constant stream of dynamic information captured by instruments or generated by simulations, processed and stored in the computer's memory bank is what we study.

3.1. Computational Science versus Data-Intensive Science

Computational science and data-intensive science are dissimilar enough to warrant a quick explanation. Computational science focuses on how a system works—for example, computational neuroscience simulates how the brain works where general neuroscience collects data about how the brain performs, computational ecology simulates ecological systems, while eco-informatics collects and analyzes information gathered during experimentation. One of the effects of the computational paradigm is "computational thinking" or applying computer processes to problems at hand like, "reformulating a seemingly difficult problem into one we know how to solve, perhaps by reduction, embedding, transformation, or simulation". Computational thinking involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science. Computational thinking includes a range of mental tools that reflect then breadth of the field of computer science (Wing, 2006).

Complementing computational science, data-intensive science consists of collecting, curating and analyzing data (Hey, 2009). Why these activities are critical and emerging as a second path to computational science becomes clear when we imagine how scientists share data these days, not dissimilar from the way we in the AEC share building information (Figure 1). For instance, using building information model or BIM-models work as long as everyone has the same data, meaning the components for buildings are standardized and represented in an algorithmic way. In many sciences like astronomy and ecology, the software costs needed to process data exceed capital investments in instrumentation. High-powered computing (HPC) infrastructure is an equal to or larger investment than the instrumentation (telescopes, colliders, high-resolution microscopy). By 1996 more than 70% of America's top 500 companies were using AI (Goodman, 1996). This is rapidly becoming the case in the AEC professions as software and visualization tool costs make up significant percentage of overhead costs for AE firms. Among all AE and EA firms accurate project cost forecasting (52.3%) and collaboration communication (47.4%) were the top ranked project management challenges. Better decision-spaces using data-intensive computing enabled by artificial intelligence (AI) are poised to change the way we build.

3.2. Collecting, Curating, Analyzing and Archiving Data

Data must be given definition, shape or form—in effect more than files, but a database with a schema (e.g. ductwork) making the data self-describing enough

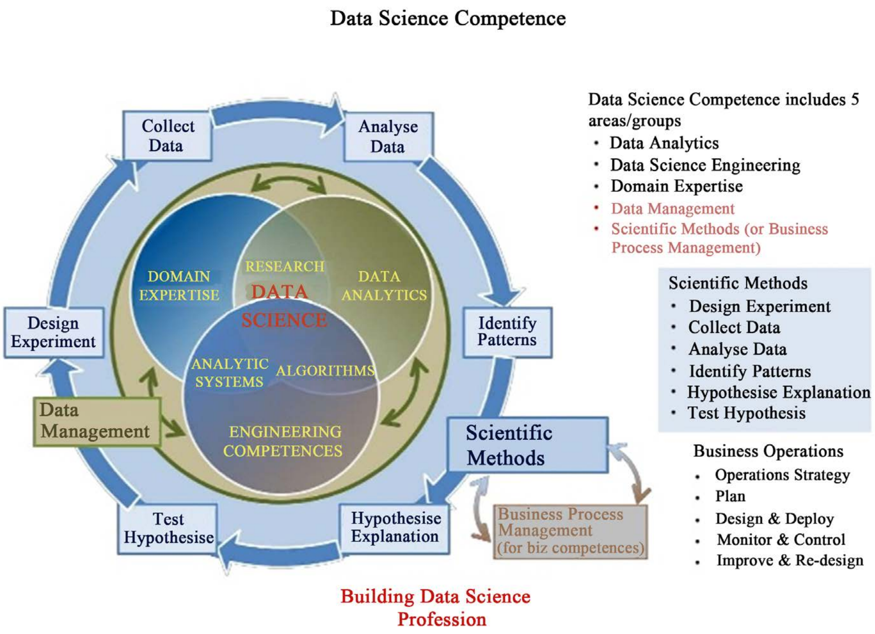

Figure 1. The diagram outlines the data science competence needed in the building data science profession between engineers, architecture and construction. The relationship between hypothesis explanation and testing is often now determined by patterns in the data or from patterns generated through data analysis. Source: Data Science and EUDAT User Forum Data Science Competence Framework (CF-DS) and Community engagement Yuri Demchenko University of Amsterdam EDISON Education for Data Intensive Science to Open New science frontiers EUDAT Workshop 5 February 2016, Rome Grant (INFRASUPP: CSA).

a mechanical engineer can access the files sent from the architect's office. Or for researchers the data can be indexed, aggregated, and searched (Hey, 2009). In the building sciences one advantage over the other sciences is our affinity for visualization: we already work with a set of visual conventions for abstracted representations of phenomenal conditions. Remember, the AEC professions rarely work on the physical object, but determine everything through simulacra: models and graphic representations. The three activities, collecting, curating and analyzing data are worth closer inspection to understand how this is part of a paradigm shift in the AEC industries and why we should think about the parallel shift from mechanical to systems theory as the emerging decision space in design.

3.3. Collection and Data Validation

Data collection happens at every scale in our industry, but we typically are not using it. In the sciences, people collect data either from instruments or sensors, or from running simulations (Hey, 2009). In the building industry and especially in the area of improving building practices for resiliency and disaster mitigation, we collect data for safety and security, accessibility, cost effectiveness, water use and indoor environmental air quality. In the wake of the 2011 announcement by the U.S. Energy Information Administration (EIA) they were suspending work on the Commercial Building Energy Consumption Survey, the National Institute of Building Sciences established the High-Performance Building Data Collection Initiative to collect and curate data for multi-dimension data. ASHRE established DASH, the Database for Analyzing Sustainable and High-Performance Buildings in 2004 also co-managed by the Green Building Alliance to facilitate consistent collection for measured data (**Figure 2**). Their sources include existing building information databases, organizations, companies and researchers. Stakeholders included the real property industry, researchers and analysts in the academic-military-industrial complex, and consultants, services and products – in short, anyone involved in designing, manufacturing, fabricating, installing or evaluating buildings or building products (Read, 2018).

What wasn't anticipated by these stakeholders in 2004 was the scale of databases made possible by the Internet-of-Things. The IoT is the network of physical devices, vehicles, and other items including buildings and building components embedded with electronics, software, sensors, actuators and network connectivity enabling these objects to collect and exchange data. This interconnection via the Internet of computing devices embedded in everyday objects enables them to send and receive data, in effect—they are information producers. Linking information to a real-time computational simulation allows us to refine a decision space instantaneously. The IoT is behind automation in many areas including the smart grid, city, and house. At a methodological level this necessitates developing protocols for collecting, tagging, and ordering data, but implications about how we theorize the boundary between a "thing" and information about the thing is less clear. La Diega & Walden (2016) of Queen Mary University

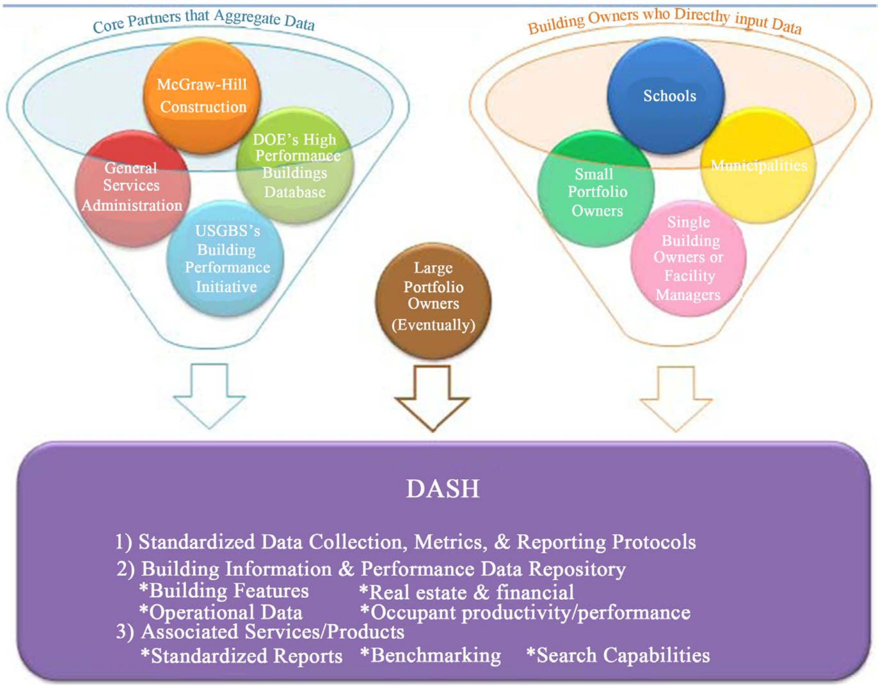
Core Partners that Aggregate Data
McGraw-Hill Construction
General Services Administration
DOE's High Performance Buildings Database
USGBS's Building Performance Initiative
Building Owners who Directly input Data
Schools
Municipalities
Small Portfolio Owners
Single Building Owners or Facility Managers
Large Portfolio Owners (Eventually)
DASH
1) Standardized Data Collection, Metrics, & Reporting Protocols
2) Building Information & Performance Data Repository
*Building Features
*Real estate & financial
*Operational Data
*Occupant productivity/performance
3) Associated Services/Products
*Standardized Reports
*Benchmarking
*Search Capabilities

Figure 2. The database for analyzing sustainable and high-performance buildings started in 2004. The diagram is from comments prepared by Doug Read, ASHRAE for the Green Building Alliance and the American Society of Heating, Refrigerating & Air-conditioning Engineers for the representative hearing of July 18, 2011. Courtesy of the national building data collection initiative. https://www.nibs.org/?page=hpbdata

in London School of Law looked at the case of the Nest^{TM} an array of IoT products monitoring cars, washing machines, lights, locks and communication devices. They focused on the contractual documents relevant to the Nest ecosystem including notices, declarations, reports and licenses. They concluded it is legally difficult to separate the "thing" from a mixture of hardware, software, data and service (La Diega & Walden, 2016). The implications of these issues are far from determined, but they suggest serious reconsideration in the building industry of how we contract for services over the course of a building's use. One potential outcome is a continued maintenance and upgrade contract opposed to a onetime fee for services.

3.4. Curation

If you have ever sorted through a set of product or machine specifications, you understand the problems entailed by curating. Organizing and cataloging information is a deceptively complex procedure. For example, when preparing a cost estimate for a project should you consider the cost of material now or for on-time delivery when the project starts? How to factor in the transportation costs or use comparative data for similar building, instrument or device types from the previous ten years? Curating the information, you will need covers a wide range of activities, starting with finding the right data structures to map into various stores (Hey, 2009). It includes the schema and necessary metadata for longevity and integration across contexts, types and professional offices. Without this schema and metadata, interpretation is implicit and depends on the programs used to analyze it. For example, in the building industry DASH identified the following focus for data intake: building characteristics, including site and operational data such as energy, water, thermal comfort, indoor air quality, lighting, and acoustics. This potentially represents petabytes (PB) of information. One petabyte is 1024 terabytes or a million gigabytes of information. The San Diego Supercomputer Center (SDSC) at the University of California, San Diego supplies computational power to the scientific community. SDSC established its Data Central site holding 27PB of data in more than 100 specific databases (e.g., for bioinformatics and water resources). In 2009, it set aside 400 terabytes (TB) of disk space for both public and private databases and data collections that serve a wide range of scientific institutions, including laboratories, libraries, and museum (Hey, 2009). Today the number is 36 petabytes or 36 thousand trillion bytes of information.

3.5. Analysis

In the building sciences data analysis, as in many disciplines, covers a range of activities including analysis, modeling, visual simulations and data visualization. The plethora of software, hardware and sub-routines possible in analyzing data are immense. Part of the difficulty is in knowing what to use, when, and how to apply the results. Analysis using modeling and data visualization is recent but fast changing the landscape of data possible to include in the design decision space. In the sciences databases may only to hold various aspects of the data rather than represent the location of the data itself. This is because the time needed to scan all the data makes analysis infeasible. As Jim Gray noted, a decade ago rereading the data was just barely feasible. By 2010 disks were 1000 times larger, yet disk record access time improved by only a factor of two (Hey, 2009). The EIA main data sets available through an Application Programming Interface (API) have 30,000 State Energy Data System series organized into 600 categories, the hourly electricity operating data, and 11,790 natural gas series to name only a few. Their stated goal is to encourage the public sector to harness and find new ways to innovate can create value-added services through public data (EIA, 2018).

One additional area worth noting is data archiving. The EIA is a good example of a big data library, but there is need for more and larger cloud storage. One of the key issues is the fair and accurate attribution of data creation. In the report, “Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century” (2005), the National Science Foundation highlighted not only the importance of data preservation but introduced the issue of the care and feeding of an emerging group they identified as “data scientists”. The report noted the interests of data scientists, the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection, lie in having their creativity and intellectual contributions fully recognized. Today in the AEC disciplines, we see increasing numbers of students interested in data—how to visualize, catalog, create and maintain datasets related to building performance at all levels. Educating and crediting this generation for their contributions to the data will be critical for fostering a skilled and knowledgeable workforce in the AEC disciplines.

4. What Does This Mean for the Building Sciences?

We looked earlier at the impact of systems theory on the methodological approach to science and the shift to a focus on dynamic processes and morphological change. The data intensive decision spaces made possible through computation validate and corroborate a systemic approach across science and engineering, including architecture, engineering and their allied disciplines. Buildings were isolated objects in the mechanistic view—in the new paradigm, buildings are data. The IoT makes this not only inevitable but also necessary. Data generated through sensors will increasingly control and determine building design, delivery and performance. The Smart City Platform requirements listed in Figure 3 outline a systemic approach to big building. The need for data infrastructure, sources, and analytics will coincide with legal, economic, and regulatory applications affecting every aspect of the city. Buildings linked to the IoT will provide real-time data about energy use, air quality, maintenance and so forth—in effect, cities will be agglomerations of built areas interconnected, monitored and managed in a dynamic parts-to-whole relationship where it will be difficult to determine the boundaries between infrastructure, bricks and mortar, and data.

Returning to the issues raised by a fourth paradigm in science, Jim Gray outlined a specific set of needs critical in the future to the computation sciences for governments, industry and research:

- Fund both development and support of software tools;
- Invest at all levels of the finding "pyramid";

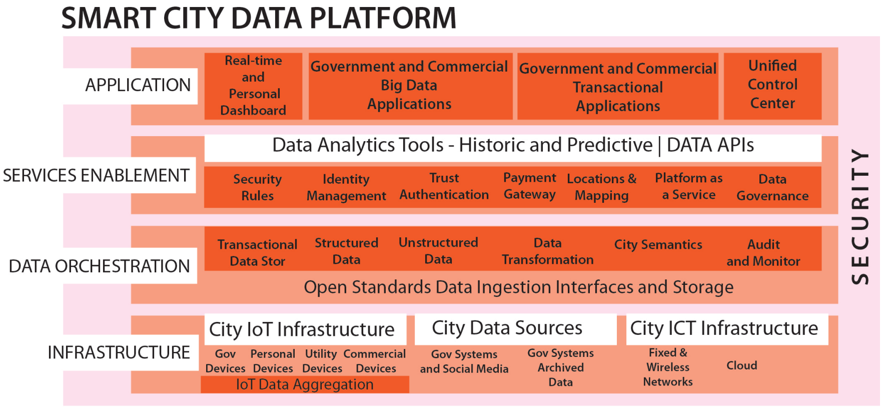

Figure 3. City platform requirements divided by layers to illustrate the complexity of the relationship between physical objects, data, environment, government regulation, economics and social spaces in the city. Source: Author, based on EMC diagram.

- Fund development of "generic" Laboratory Information Management Systems (LIMS);
- Fund research into scientific data management, data analysis, data visualization, new algorithms and tools. Three key areas for action relate to the future of scholarly communication and libraries;
- Establish digital libraries supporting the other sciences (and I would add engineering and architecture) like the NLM does for medicine;
- Fund development of new authoring tools and publication models;
- Explore development of digital data libraries containing scientific data (not just the metadata) and support integration with published literature.

If his laundry list sounds foreign, consider the increased scientization of our disciplines:

- Demands for research-based methodologies including benchmarking, protocols for acquiring, analyzing, and reporting data;
- Publications moving to a solely online presence;
- Developers demanding the building industry provide data-driven business outcomes for better cost-estimation and post-occupancy analysis;
- Professional firms investing time and money in research and publication;
- Standardization of Building Information Modeling and integrated building delivery;
- Increase in number of performance-based software since 2000.

Given the dominant model of the material world is increasingly a network of systems with interrelated and interdependent patterns of behavior, the challenge for creating resilient buildings, smart cities and good ecological management plans is the education of architects and engineers capable of using multiple and varied date to create complex decision spaces. This means big building won't mean a tall structure, but a dataset describing the conditions of a built environment. The use of artificial intelligence (AI) to manage and predict environmental conditions, safety and security, energy performance and use and land management will mean buildings and their context merge into a complex of interrelated systems. It is yet to be determined how this will effect change in the legal definitions of objects and environments or the limits of contracts in the service sector. Challenges for future study include addressing the legal definitions of property where objects and data coincide, determinations of how, when and where data are public or private, when data are proprietary and what maybe the consequences of a breach in cybersecurity across such large systems. Some of the legal considerations were briefly discussed, but there is much work to do.

Finally, the interplay of the real and virtual will become commonplace, ubiquitous and persistent in the developed world. Models, simulations, immersive virtual experiences and mixed reality where the virtual and real interact in real-time are at play in the construction industry now. Site excavation such as cut-and-fill are managed using earthmovers geo-located using GPS and coordinated with BIM models. Typically, a drone-flight over the site at the end of the days' work is used as double-check and to identify target points for the next day pass. The virtual site-model and the real site intersect in a real-time decision space updated daily. The sea-change suggested by these new processes are part of what this article aimed to suggest. The significance of bringing this discussion forward now while we are still somewhat indeterminate about future use of big data in the building sciences is precisely to outline what needs to be addressed.

5. Recommendations

Our next steps are to make sure we educate the present generation of designers, demand our professional organizations to address the impact of these changes on our contractual procedures, promote, share and disseminate findings, develop protocols for the "pyramid" of collecting, curating and analyzing data, and participate in the ethical, aesthetic and social impacts of these changes to the built world.

Acknowledgements

Thanks to Clemson University and the Mickel Endowment for the funding support to complete this article.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Bush, V. (1945). As We May Think. *The Atlantic Monthly,* 101-108.

EIA (2018). *Open Data*. https://www.eia.gov/opendata/

Goodman, W. C. (1996). The Software and Engineering Industries: Threatened by Technological Change? *Monthly Labor Review, 176,* 37-45.

Hey, A. J. G. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery.*

La Diega, G. N., & Walden, I. (2016). *Contracting for the "Internet of Things": Looking into the Nest.* Queen Mary School of Law Legal Studies Research, Paper No. 219/2016. https://ssrn.com/abstract=2725913

Read, D. (2018). *Testimony on Data Needs to Achieve High-performance Buildings*. https://c.ymcdn.com/sites/www.nibs.org/resource/resmgr/HPBDATA/Read_Summary.pdf

Von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Applications*. New York: G. Braziller.

Wing, J. M. (2006). Computational Thinking. *Communications of the ACM, 49,* 33-35. https://doi.org/10.1145/1118178.1118215