# Labor Out of Place: On the Varieties and Valences of (In)visible Labor in Data-Intensive Science

Michael Scroggins and Irene Pasquetto

For Vannevar Bush, 1945 was an annus mirabilis. His wartime efforts at the Office of Scientific Research and Development (OSRD) to reorganize American science along lines established by industrial engineering had borne fruit in the form of scientific advancements and overflowing archives. In chemistry and physics, Bush found a model of scientific practice radically changed from the prewar era; research conducted by teams of cooperating scientists, intensive publishing schedules, and wide dissemination of research findings became the new normal. This new model of science, Bush argued, could and should be expanded to the whole of science. Cementing these changes would be a wholesale change in the organization and funding of American science (see Mirowski 2011), and much of the OSRD model would find its way into the postwar Nation Science Foundation (NSF) along lines laid out by Bush in *Science: The Endless Frontier* (Bush 1945a). One direction of Bush's vision called for inculcating a new set of sentiments within scientific apprentices. In contrast to their prewar counterparts, scientists trained in the postwar era needed to be comfortable with cooperatively communicating and disseminating research, working at the faster wartime pace normalized in physic and chemistry, and bridging the gaps occasioned by increasing disciplinary specialization. Another direction called for the flood of information produced by this new model of science to be tamed through technology and automation, such Bush's own Memex (Bush 1945b). Yet, Bush's vision was lacking in one critical respect; the army of technicians, specialists, and scientific adjuncts co-extensive with the increased pace of wartime science were, in Bush's view, passive functionaries soon to be replaced by technological advances.

Two decades after Bush's annus mirabilis, his speculation that the onerous work of science could be automated through technological advance ran headlong into an inconvenient fact. In a speech to the 1968 American Psychological Association titled "As We May Think, Information Systems Do Not," Paisley observed that despite the computational power, technology, and data being available in 1968, the archives had failed to automate:

> The missing element in information service is people. Mediators. Middlemen. We cannot abolish the archives -- if anything, we need them more as the doubling cycle of scientific knowledge moves from decades down to a handful of years. However, I think that our dreams for mechanizing the archives and making them truly responsive to researches and other users are dreams for the next century (Paisley 1968, 13)

Science, as Paisley gently reminds us, is a conversation between and betwixt collaborators and interlocutors (be they human, animal, mineral, mechanical, or digital), not a monologue given by a solitary scientist to an audience of functionaries. This is no less true today than in 1968. Our parenthetical expression (in)visible labor draws attention to the stakes of our inquiry. Behind data-intensive science's technological facade lies a bewildering array of human labor, some performed in the spotlight by star scientists, but most performed by the precariously employed in service to digital machines.

In what follows, we illustrate the valences of (in)visible labor in data-intensive science. First, we draw on an archive comprising fifteen years of continuous and comparative research across multiple domains of data-intensive science and classic works in science studies to delineate a stock of activities necessary and common to data-intensive science: authoring, administering, maintaining, archiving, and collaborating. Second, by applying concepts developed in recent labor scholarship to our data corpus, we demonstrate the valences of (in)visible labor within data-intensive science. We intend valence in its chemical and grammatical formulation, showing how labor combines with the activities of science in multivalent and often surprising ways – to transform noisy instruments into clean data sets or to magically meet diversity metrics. We animate these combinations with ethnographic vignettes drawn from our data corpus. Finally, we conclude with a summary of our argument and highlight changes to scientific labor and practice on the near horizon.

## What makes a science data-intensive?

By data-intensive science we mean scientific fields in which the quantity and velocity of data generation have led to (a) a reliance on computational power and techniques to analyze, curate, and archive data (Kitchin 2014a; Burns, Vogelstein, and Szalay 2014; Critchlow and Dam 2013; Ekbia et al. 2015; Kitchin 2014b, 5–7) and (b) a need for transdisciplinary collaboration (Bowker 2008; Bell, Hey, and Szalay 2009). Data-intensive science is heterogeneous in composition and complex in operation. A complication in studying data-intensive science is that, in practice, it combines elements of older scientific practice (observation, experimentation, laboratory work) with elements of computer-supported work as practiced in industry. Adding to this complicated environment is the capital-intensive nature of data-intensive science, typically administered through large grant-funded projects with budgets in the tens or hundreds of millions and requiring an extra layer of administration, public outreach, and data-management. Because of the large budgets and long life of the resultant infrastructure, data-intensive science requires intensive interdisciplinary collaborations that must be maintained through time and across space, often, as we demonstrate, requiring formidable amounts of social work.

As Bush played a key role in the transformation of science in the postwar period from his perch at the OSRD, another engineer, Jim Gray, played a key role in ushering in the era of data-intensive science from his position at Microsoft. In a lecture summarizing his long-term collaboration with astronomers, Gray (2009) articulated his vision for a 4th scientific paradigm based on the intensive generation, analysis, and archiving of scientific information – from lab books to grey literature to data sets - stored in databases and connected through the internet.

In this sense, the labor of data-intensive science marks both a departure and a continuation of Bush's vision of automating away the onerous parts of science. But automation has an odd trajectory. For example, the labor of typists and calculation, outsourced to specialized technicians in Bush's era, is now part of every scientist's daily work. Computers have automated much, but in terms of actual time spent, they have not saved labor hours, but only distributed it. On the other hand, equipment demands maintenance and instruments demand calibration – whether or not the final output is digital data. What is new in data-intensive science, and reliant on ubiquitous connectivity and reliable digital storage, is the work of exploring, sifting, and reanalyzing large data sets through computation and statistical means. Once enough data is accumulated, per Gray, the main analytic focus will naturally shift from the analysis of bespoke data sets to the reanalysis of extant data sets generated automatically and made publicly available. While Gray's vision has not been universally adopted (the humanities and field sciences are notable holdouts while large parts of astronomy have adopted Gray's vision), in disciplines where data-intensive science has become the normative mode, new social forms and new job descriptions have emerged in its wake.

Perhaps the most distinctive, and consequential, social form invented in the interval separating Bush and the 21st century, is the Primary Investigator (henceforth PI). Since the Second World War, the scientist has morphed from a singular figure responsible for the entirety of a scientific project into a principal investigator leading a multifunctional team of apprentices and technicians, coordinating their team's activity with university administrators, IRB boards, and donors. Today's PIs are responsible for a mushrooming range of management and financial duties: overseeing the PhD student whose dissertation project must harmonize with the PI's larger project, integrating the postdoctoral researcher or research scientist who brings specialized knowledge and fresh outlooks to the project, managing research scientists working on ad-hoc contracts, and attracting and retaining technicians who play an increasingly important, but largely uncredited, role in the research process. Today, the work of the scientist, as Bush romantically envisioned it, is distributed across a farrago of instruments, sensors, databases, assistants, apprentices, administrators, and technicians with everyday work of data generation and analysis largely in the hands of PhD students, postdocs, and precariously employed research scientists.

In addition to new social forms, data-intensive science has engendered new scientific workplaces. The labor of data-intensive science is performed often in front of a screen using common digital tools such as spreadsheets, databases, and code editors rather than at a field site or lab bench. Even where older norms of scientific work persist, such as in ecology, data generated in the field is converted (often laboriously) to digital form for manipulation, circulation, and storage. As well, data-intensive science is often spread across multiple locations and organizations that combine to create data-intensive projects behind shell corporations in order to shield universities and industry partners from potential liability and give fiduciary control to advisory boards and managers who often come from industry backgrounds. A common stock of tools, organizational techniques, and administrative acumen have developed between the two fields, and scientific labor in fields touched by data-intensive science has come to resemble labor in industry, and vice versa. Above all, the distinguishing characteristic of a data-intensive science is not reliance on data, but rather the intensive production, consumption, and circulation of data and data products. In a data-science specific form of the drunkard's search, Darch and Borgman (2016) have observed that the pull of data is so strong that it can be difficult for emerging fields, data scarce by definition, to secure funding and access research infrastructure.

## Research design

The analysis presented here is based on interviews and ethnographic observations conducted by members of the UCLA Center for Knowledge Infrastructures (CKI), including the authors of this paper. We draw upon data from several waves of studies run by the CKI's members over the last decade, selecting responses to questions about workforce, collaboration, automated processes for data management, data sharing, and data reuse (Wallis, Rolando, and Borgman 2013; Sands et al. 2014; Darch et al. 2015; Pasquetto, Randles, and Borgman 2017; M. Scroggins 2017; Souleles and Scroggins 2017; Pasquetto 2018). Studies were conducted in a diverse array of scientific fields, specializations, and sectors across the physical, life, and social sciences and professional fields like medicine, business, and engineering, and computer and information science. Participants span PIs, postdoctoral researchers, doctoral students, graduate students, technicians, librarians, and staff. Ethnographic research incorporated in this article includes field observations of participants performing research, laboratory and community meetings, and other events. Members of the CKI interacted with researchers during formal gatherings such as research reviews and retreats and weekly research seminars and informal gatherings such as discussions within labs and offices. Memos provided context for interpreting interview transcripts. All interviews were audio recorded, transcribed, and complemented by the interviewers' memos on noteworthy topics and themes. In sum, the corpus consists of 483 coded interview transcripts, an equal number of ethnographic observations, and thousands of pages of documents, listserv archives, and other grey literature.

Using Atlas.ti, we converted our handwritten list of activities into codes that allowed us to retrieve relevant parts of our data corpus. From the codes, we drew a selection of vignettes to highlight the contours and boundaries of these ongoing activities and to illuminate their scope, depth, limits, varieties and valences. By selecting via activity rather than job classification, social position, or place in the academic hierarchy, we are choosing to foreground the everyday complexities of scientific work and discuss how the everyday work of science cuts across job classifications, social positions, and academic hierarchies.

Differences in tools and work practices, disciplinary concerns, epistemological and ontological assumptions, the scale and centrality of data generations, and the state of standards and measurements within a given discipline all play their part in the kinds and amounts of labor required. For example, astronomy, with its standardized file formats and use of common instruments (telescopes) with consistent metrology, requires different forms and amounts of labor than a field like ecology, which lacks standardized file formats, common instruments, and consistent metrology. Though both astronomy and ecology use data intensively, they produce, consume, and circulate data in differing manners. Of course, despite differences in work practices, certain tasks are essential to scientific practice and remain stable across disciplines. These tasks form the bones of our analysis, with ethnographic vignettes providing the flesh. Confidentiality allowing, where possible we have allowed our research participants to speak in their own voices.

## Opening the black box of (in)visible labor in data-intensive science

The canonical understanding of invisible labor was stated by Daniels' (1987)  as work "that disappears from our observations and reckonings." That is, labor symbiotic to the labor channeled through classificatory schemes and metrics such as job descriptions (Star and Strauss 1999; Bowker and Star 1996, 2000). The visibility/invisibility dichotomy has acquired more nuances as scholars have (i) stressed that visibility is a multivalent and multifaceted spectrum, one with fractal folds and complexities and (ii) sought increasingly fine-grained and nuanced concepts to account for the complexities of labor in the 21st century.

In this sense, the difference between labor made visible and public through metrics and classificatory schemes and labor that "disappears from our observations and reckonings" mirrors the classic formulation of the Janus face of science (Latour 1987, 4). In all its permutations, the form invisible labor takes cannot be separated from the wider political economy in which it is embedded, as Cowan (1976) argued in the context of housework. While we do not directly address the work of funding data-intensive projects, the necessity of seeking funding within a marketplace of competing projects forms the political economy of 21st century science (Lave, Mirowski, and Randalls 2010; Tyfield 2013; Edgerton 2017) and creates the context within which labor acquires (in)visibility. On one side of the ledger is labor that contributes to winning grants, on the other labor that "disappears from our observations and reckonings" when grant applications, tenure cases, publications, and public relations are given pride of place.

In the 1970s, feminist scholars began to look at household labor in relation to economic indicators. The work of maintaining a household, despite its symbiotic relationship to the general economy, was rendered invisible through non-inclusion in official metrics and statistics (Fee 1976; Himmelweit and Mohun 1977; Coulson, Maga, and Wainwright 1975). Within history of science, invisible labor has often been conceptualized as labor symbiotic to that of the scientist. Shapin's (1989) description of the invisible labor in Boyle's laboratory is the classic account, with recent accounts of the gendered and intersectional complexity of scientific work and theory (Roberts 2018; Suchman 2011; Harding 2016) carrying this work forward into contemporary science.

By the 1980s, as the economy shifted from its industrial base towards the service sector, scholars such as Hochschild (1983) conceptualized emotional labor as a new type of invisible work that paralleled the division of labor in the domestic household as a transformation of domestic work into a commodity necessary to the operation of the service economy. Following on Hochschild's pioneering work, the concept of care has emerged as a critical standpoint from which to examine labor in all its manifested forms (Star 1990; Mol 2008; Tronto 1993), but particularly the labor of maintaining human relationships. Tronto defined care as "all we do to continue, maintain, and repair 'our world' so that we can live in it as well as possible (1993)." While care has been widely taken up as an ethical stance in terms of human relationships, Jackson (2017) emphasized that care also has material implications. In data-intensive science, care is important as both an ethical stance towards both mending scientific machinery and furthering the relationships that animate scientific inquiry.

In the digital domain, the ethereal nature of digital work creates distinctive difficulties determining whose labor should be measured in official metrics, be they the mythical "man hours" in software development or scientific papers written and grants won, and whose labor is symbiotic to the metric. For example, Nardi and Engeström (1999) conceptualized digital work as a "web on the wind," structured in practice yet lacking institutional recognition. Building on the earlier CSCW and HCI scholarship is a thread of research into the difficulties of collaboration in digital environments. Though technological tools can sometimes be used as the main medium of collaboration, more commonly, personal relationships and an ethic of care must carry the collaboration forward, particularly in the case of transdisciplinary collaborations (Ribes and Bowker 2008). Similarly, Leonelli (2010) and Plantin (2018) argued that in data-intensive science, the work of cleaning data sets, such as assigning metadata, developing ontological schemes, and checking instrument readings, is typically devalued as "non scientific work," yet both Leonelli's (2010) research on biocurators' work and Plantin's (2018) on "data cleaners" in a social science archive demonstrate that the creation of ontologies and metadata and the cleaning and documenting of data sets, respectively, is essential for data-intensive science.

While the canonical definition and foundational works on invisible labor emphasize the negative ramification of invisibility, two works (Orr 1996; Allen 2014) focusing on professional labor implicitly argue that not all invisible work is negative. Work performed by nurses within a hospital is invisible in the canonical sense, but Allen (2014) suggested that the demands of professional practice in environments where sensitive information is archived, reused, and deliberated over can make a virtue of invisibility. Allen found that nurses, as part of their invisible yet essential professional practice, work across the boundary separating the formal elements of hospital care, such as scheduling medications, assigning beds, consulting with doctors, and filling out paperwork, from the informal aspects of hospital care, such as soothing patients and family and discussing cases and colleagues with fellow nurses. This work is essential to care but also to patient privacy. Likewise, in a study of Xerox copier repair technicians, Orr (1996) found that the repair technicians' talk about machines, customers, and salespeople, more so than technical ability, to be their primary form of labor. Talk was used by the technicians to train apprentices, understand how and why copiers break down, and manage customer's and sales manager's expectations. In both instances, the professional talk of nurses and repair technicians police and repair the boundary between customer and company and carve out a class of professionals whose job becomes translating across that boundary. In data-intensive science, the professional work of science qua science occurs in negotiations over authorship credit, discussing the future of controversial lines of research, and mentoring apprentices.

Visible labor is the unmarked category of labor made visible through classificatory schemes (Bowker and Star 1996, 2000; Foucault 2002). Writing for publication is the canonical visible work of the scientist. The work that technicians, administrators, and staff perform according to the standards negotiated in job classifications, descriptions, and annual reviews can be considered visible. New standards bring new forms of visible work. For example, publishing data sets and research software has quietly become part of the visible work of science in some fields, as granting agencies have begun requiring data used in writing papers to be made publicly available (National Institutes of Health 2017).

Within the labor literature, hypervisible (Crain, et al. 2016) work has emerged as a category denoting labor in which the aesthetic enjoyment of observing work, as a skillful achievement or spectacular failure, is the main attraction. The work of celebrity chefs in open kitchens, actors on a stage or screen, and servers at restaurants are all examples of hypervisible work. Like all work, hypervisible work carries both positive and negative ramifications. University donors attending academic conferences or spending time “on the mountain” with astronomers or in the lab with scientists are paying to experience the hypervisible work qua work of science. Work featured in the popular press, the work of public intellectuals, work with strong and direct ties to areas of policy, and work of interest to major foundations and donors we also consider hypervisible. (Crain et al. 2016). On the other hand, scandals and controversial research may become hypervisible in a negative manner. The other side of credit for success is blame for failures. An example of negative hypervisible scientific work is the recent spate of “outings” over the difficulty of reproducing studies in social psychology and other fields (Dominus 2017; Marcus and Oransky 2018). A classic example is Diane Vaughan’s narration of the Challenger accident (Vaughan 1996).

## The varieties and valences of (in)visible labor in data-intensive science

In the following section we employ vignettes to explicate the valences of labor in data-intensive science in service to understanding the terrain of this quickly evolving scientific paradigm. We highlight the invisible, visible, and professional labor of data-intensive science in its complex permutations. Though our corpus contains several instances of hypervisible labor, in the interest of confidentiality we have excluded them.

### Authoring

By authoring we intend the work Foucault ([1969] 2012) glossed as “enunciating the truths of science.” Today, those truths are spread among instruments, papers, data sets, and software rather than concentrated in a singular location. In data-intensive science, authoring is both writing papers and grants and authoring data sets and research software (Mayernik et al. 2015; Green 2009; Hills et al. 2015). Authorship is also a dividing line separating those whose name appears on the byline, and hence registered as a citation, from those in the acknowledgements. Adding to the complexity, in disciplines such as astronomy, it is customary to place the instrument that took the observations on the byline. A contested form of authoring is creating datasets and research software. Data sets are often authored by graduate students, postdoctoral researchers, and technicians. Despite being central to data-intensive science and playing a key role in ensuring data are able to circulate and remain interpretable, data and software authorships goes uncredited in publications and uncounted in tenure cases (Howison and Bullard 2016; Velden et al. 2014). Below we present three vignettes drawn from our data corpus that illustrate the valences of authorship in data-intensive science.

*When postdoctoral researchers from different disciplines are joint authors, professional labor is required to assign credit:* Authorship norms vary between disciplines and negotiating the authorship order between authors hailing from different disciplines, in this case physics and astronomy, requires balancing the sometimes competing claims of the work on the paper against disciplinary norms for distributing credit. This is professional labor proceeding, as Allen (2014) and Orr (1996) observed, through talk and negotiation. One participant described the delicate dance of arranging an article’s byline according to the field a postdoctoral researcher is applying in: “It was a negotiation primarily between the particle physicist and the astronomers…we can't just ignore the fact that astronomy departments have different criteria than physics departments about when people are applying for faculty jobs. We have to accommodate that and since the particle physicists don't care who's first author, we say, ‘All right, let the astronomers be first author.’ ... The way it works in particle physics you might say, ‘How do particle physicists ever get faculty jobs?’ The answer is you rely on letters from the people in the collaboration who know what they did. So the procedures are just significantly different. It requires, it's a delicate thing, how do you balance the needs of the two different groups.”

*Technicians’ labor is not always invisible, but at the PI’s discretion technicians can be authors:* In data-intensive science, inclusion on the byline is often determined by the judgment of the PI, as it is a project’s PI, not individual researchers working on the project, who own the data. The PI’s judgment is especially important when a potential co-author has moved on from the research team, lacks the academic credentials to justify authorship, or their contribution is unclear. Here a postdoctoral researcher explains the decision to include a laboratory technician on the byline of a recent paper: “He was actually their lab technician, so he did a majority of that particular work. And, of course, he’ll be included in on the paper because he did... [the PI] is very good about, at least worst case scenario, making sure you get credit for what you did because they're not here to defend themselves."

*Formerly invisible authors of datasets can be made visible, with caveats:* Cleaned and archived data sets are the lifeblood of data-intensive science. Yet, cleaning and archiving data sets is often uncredited and unrewarded work within scientific disciplines. One optimistic possibility is an emerging technique of data publishing and authorship used in some astronomical projects that grants authorship credit to all those, no matter their role or status, who had a hand in producing a dataset. As one of our interviewees explains: “And then you'll find at least one and sometimes two alphabetical tiers of authors which are indicating people whose work was not devoted to that particular science effort, but they've earned their authorship rights by virtue of helping enable the survey by doing kind of broader infrastructure work and made that work possible. And so they get what's called ‘architect status.’"

## Administering

Administering, from the Latin, means to “to help, assist, manage, control, guide.” Administering is not ordinarily visible by those outside the project, and many aspects of administration are invisible to non-administrators inside a given project. Often this means taking responsibility for softer, nonscientific parts of a large grant – education and diversity requirements - and satisfying the demands of IRB boards, data privacy and security policies, and HR mandates. For staff, administering often requires developing interactional expertise, being able to talk authoritatively about a scientific domain  (Collins and Evans 2007), while for faculty, PIs, and research assistants administering often means developing contributory expertise (Collins and Evans 2007) in administration, learning to contribute to the state of the art in project management. Below we present three vignettes illustrating the complexities of administering in data-intensive science.

*Administrative assistants labor invisibly, doing the housekeeping of data-intensive science:* The labor of administering is often spread across multiple positions. Official organizational charts that simply divide academic from non-academic staff are of little help in discerning who administers what and how. One administrative assistant described the experience: “I am an administrative assistant. I work on, mainly, event coordination and event planning. I’m also responsible for development of the center’s website, and doing the cyberinfrastructure research to advise the Executive Committee, and then, sort of, everything else under the sun from administrative paperwork to event setup. We’re sort of a catch-all kind of job… when I applied to this job, there was nothing in the job description about what sort of organizations this was or what they did… [I am] figuring out how to combine the administrative and the scientific databases with the public website, so establishing a website with data portals of various kinds to sort and to filter different entities based on their metadata.”

*Administrators do the visible labor of meeting metrics required by funders:* The broader impacts of scientific work, important for securing funding but difficult to institute in practice, are often the responsibility of administrative staff on large projects. Particularly in projects with educational and diversity goals, administrators are often at the leading edge of creating new pathways into scientific careers for underserved communities. One project coordinator explained how she organized a gender equity program: “One of our big things that we're also looking at was gender equity, especially in computer sciences... Sort of trying to demystify some of the stereotypes that revolve around women and the sciences.... Our hope was that we could educate people and provide enough positive experience for the outcomes to be that people feel as if there are pathways for making that more equal. But at the same token also, we were also gauging perception of what people felt. What is the capabilities, I suppose, or stereotypes of women in the sciences.”

*Administrators must be autodidacts, educating themselves in the professional labor of science:* Also common is the autodidact administrator who, once in the position, must teach themselves enough domain science to be help, manage, assist, or guide a project. Another administrative assistant describes the struggle of coming up to speed with a fast moving scientific field: “[The first day] was like, ‘Okay, you’re the Education and Diversity Director. Here’s what the grant said. Here’s what this document says. Go for it.’ …[the PI] told me once that she considered professional development to be the ultimate sign of an employee being able to figure out what it was that they needed to do and go and do it. And when I came in, I had no background in science. So, I recognized that as the point of my greatest learning curve. And I began to attend the weekly lab meetings. Nobody made me go. Nobody suggested that I go. I just said, ‘I’ve got to go hear these students talk.’ So, they would share and I would ask questions. And frequently, I would talk to the grad students or the postdocs and afterwards, I’d make lists. I come back to my office. I Google the words that I had down on paper.”

## Maintaining

Data-intensive science requires a dazzling array of technical skills, most of which require ongoing education, both formal and informal, to master. Yet, people with technical skills, no matter the importance of their contribution, are often in precarious positions and paid with the least stable forms of funding. The technicians required by data-intensive science range in skill and experience from graduate students who know a little more Python than anyone else to electrical and software engineers with decades of specialized domain knowledge. Their jobs range from building instruments, ranging in size from telescopes to discrete sensors, to maintaining equipment and codebases and to repairing all of the above and then some. Not until key components break, fall out of calibration, or fail to be constructed on schedule, does maintenance work come to the forefront.

*The invisible work of technicians eases the PI's managerial burdens:* The need to maintain equipment is a constant companion in data-intensive science. Many scientists we have interviewed, such as the one quoted below, have drawn a surprisingly old distinction (Shapin 1989; Morus 2016) between technical staff and scientists over their relationship to equipment. In this case, a PI positively describes the role technicians played at a former institution and laments the additional labor required of him at his current institution: "[There] you have a person for everything. You have a person who does orders. You have a person that if you cannot order something via the net or whatever, that person actually drives around to buy stuff. You have computer people that help with everything computer. You have technicians like mechanical and electronic technicians that help with all kinds of equipment. Here, we don't have that."

*Students and postdoctoral researchers do the invisible work of maintaining analytic pipelines:* A technician is also a *bricoleur*, skilled at combining the odds and ends of various systems and infrastructures into a workable whole. Here a PI describes how a data stream originating from a robot is rendered useful for analysis: "So the problems are that it requires a fair amount personal intervention in the sense that I could draw a nice picture of this data flow but it's not anywhere near as automated as anyone might believe it is, if it involves typically... In this case it involved [a graduate student] doing a fair number of things manually… It's not automated, it's just enough to do experiments of this kind, but it doesn't really translate into a system or anything like that that we could just give to other people to do, right? So, and all the tools to massage the data once it comes off the robot are all sort of home brewed, right? And they tend to live because two or three students or postdocs sort of maintain them. But they aren't systematically maintained or archived or sort of curated in any way."

*The metrology of large infrastructure is maintained through the invisible labor of skilled technicians:* In astronomy and physics, the cost of physical infrastructure is measured in the hundreds of millions and is intended to serve thousands of scientists over several decades. To hold the measurement standards of such complex scientific instruments steady, specially trained technicians must assist in the operation. A technician at an observatory explains the process: "There are only two other telescopes built like this and we did this for a number of different reasons, primarily, to maintain what's called laminar flow between the optical elements primary and secondary mirror, to maintain image quality and temperature control…We have a crew that operates it, we collect the data, we give the data away, and you'll find that out that basically, we're a factory. We produce the data; we give it to people to use it… I do the mechanical, [another technician] does sort of the software side of things…then there's a series of technicians that work [in operations]. One electronic tech, one mechanical... and then a series of pluggers. So we have like three pluggers, and what that means is these are the people who actually plug the plates during the day for observation purposes at night and that's pretty much the crew."

**Archiving**

Archiving is the work of cleaning, wrangling, curating, and preserving data for future reuse (Pasquetto, Randles, and Borgman 2017). The technical and fiscal cost of generating data has fallen over the last few generations but the asymmetry Paisley identified when he observed that “information specialists” were needed to mediate between data banks and research groups has only grown wider. Further adding to the complexity is the velocity at which data is generated, making the process of fixing data in place for archiving more difficult.

*The professional labor of science is often at odds with the professional labor of archivists:* Archiving in data-intensive science means addressing the collision between file formats suited for cutting-edge research and file formats suited for archiving and preservation. Though astronomy is the rare discipline where one file format, FITS, predominates, localization of the file headers can cause problems for archivists, who need fixity to preserve data over the long term. An astronomical archivist working with a ground-based telescope explains: “We didn't necessarily want to force [astronomers] into a standardization, because it tends to quell innovation and cleverness, and things like that… But then, we get different operational software or different detectors that collect the data in different ways, and, so all these [file] headers are slightly different from instrument to instrument and from era to era. That's one of our problems.”

*Behind the dream of automated data generation is the invisible labor of cleaning and munging data sets:* Another common difficulty is overcoming the fragility of automatically generated data. Automation can save labor but automation can also be a cause of additional labor. The following vignette describes the case of a malfunctioning environmental sensor that caused malformed data to be automatically generated, necessitating cleaning by hand: “When we collected the data in Bangladesh, we had this really ad hoc way of saving the data…it’s just what some guy came up with when he wrote the software. … [When the sensor malfunctioned] I had to spend hours and hours just cleaning the data because we had duplicate packets that were shown, data was printed out of order. And because of the software it was really hard to get it back in order. So if a node rebooted, any kind of node in the network rebooted, then the time stamps were screwed up, the sequence numbers were screwed up…I had to use anecdotal information, like okay, I know that I rebooted this node at this time…I did a lot of manual, like I think this time stamp should actually be this, so a lot of writing scripts to manually set time stamps, which never feels good.”

*Authoring metadata is invisible work that renders scientific papers and datasets visible:* If authoring is the visible work of science, authoring the metadata that makes data discoverable and usable is its invisible accompaniment. Compared to excitement of a published article or winning a grant proposal, success in archiving is decidedly unglamorous and made difficult by the emphasis on publishing over preservation: “You want some kind of connection that’s permanent, so when you see this link in whatever form it takes, it’s gonna be good 20 years from now. It’s that technical challenge…But there’s also a social problem… People submit their papers and they don’t provide the links [to the underlying data]. Partly it’s culture, that’s how astronomy has always been done. But there’s another reason which is a little more self-serving, which is astronomers do not want to give other astronomers a leg up when doing research…There’s also another difference, a big difference in the nature of data now and 20 years ago. Data used to be really expensive, and therefore people protected them. Data are now cheap, people don’t need to protect them as much. This is the age of the big sky surveys, which are supported by very fast high quality computers that are capable of managing and processing huge amounts of data.”

## Collaborating

In data-intensive science, data, storage, and computing power takes pride of place. But, as Paisley recognized and a generation of researchers have documented, it is human relationships, mediated through data, storage, and computing power, that produce scientific knowledge. And unlike computers, human relationships require constant care and attention to hold cross-disciplinary research together. It is the emotional labor (Hochschild 1983) and relationship building we gloss here as collaborating. Collaborating is an active verb, expressing action as well as attitude and ethical orientation. As Jackson (2017) has argued, the work of caring and cultivating an ethic of care has both symbolic and material dimensions and can be directed towards human relationships or relationships to materials and equipment that mediate between human relationships.

*Administrators often do the emotional labor of counselling and mentoring students:* One participant described an administrator who took on the emotional spadework required to make good on claims of furthering diversity and career formation: “One of the things that I think was dramatically underappreciated by the management at large is the critical role that [the administrator] has played as a big brother, as a mentor towards lots and lots of students. And every program needs someone like him, whoever the official role is. As a person, he has been critical to the success of the center…. one of the remarkable sociological things about the center was that, there was probably at one point that half of the full-time staff, whether we were administrative or technical or whatever, were gay. For this department, engineering and so on, that was, I think, a quiet watershed event…in a very behind-the-scenes way and out-of-working-hours way, was somebody that many of our male students, at least gravitated to, to just be able to deal with that side of their lives. None of this ever was above the surface. There were never any gay bashing issues.... It was all very professional sort of a thing, but I think [he] has just had this incredible role.”

*PIs often do the emotional work of holding research teams together:* A common refrain in long-term research teams is the PI who does the emotional spadework of making sure each member of the research team feels valued and appreciated. In an economic climate where many researchers and scientists work on short-term, grant-dependent contracts in precarious positions, the emotional work of making everyone feel valued is important to retaining key personnel. Here a research scientist on a short-term contract talks about of a PI who went out of her way to create the social conditions required for successful collaborations in data-intensive science: “One of things that I've appreciated about her is that more than any of the other faculty that I've worked at the School of Engineering, she has a degree of caring about people at a personal level that was very refreshing and rewarding.”

*Research scientists on temporary contracts often do the emotional work of bringing peers into conversation:* One of our interviewees worked as a mediator, building trust and friendliness while avoiding “flame wars” between colleagues in the same discipline but with divergent viewpoints on method, research, and analysis. Here the interviewee describes the difficulties of reshaping relationships built on competition into relationships built on cooperation: “There were some really harsh emails, people didn't hold back on being critical of one another, and again, it just didn't help in terms of trying to build a cohesive team that was really trying to work together. So a big part of what I did in the early years when I was on the project was, I spent a lot of time traveling between the sites, getting to know people …we'd take people from two institutions, and we'd sit them next to each other in the same room, and it's like, 'we're working together guys.' So it was really challenging…and even if at the end of the day they still didn't fully respect each other or there was still some mistrust, I think I was able to develop a sense of rapport and trust with people that they relied on me to make sure that, to bridge it."

## Conclusion

Bush's annus mirabilis of 1945 cemented a sea change in the political economy of science. Out was the prewar model of science as the contemplative activity of a solitary scientist, in was the wartime models of teams of scientists and assistants working on solutions to common problems and desires. Unseen by Bush, but noted by Paisley a generation later, was the active role in enunciating scientific truths played by the army of technicians, specialists, and scientific adjuncts engendered by the new political economy of science. We argue that Paisley's observation about active role played by "mediators" has only intensified as science has become more data-intensive. Our argument hinge on the assumption that researchers, policy makers, and scientific funders require a fuller and more nuanced accounting of scientific labor in order to understand how scientific work has changed and how it might change in the future.

Throughout, we have drawn examples from the extensive data corpus accumulated by several cohorts of researchers at the CKI over the last fifteen years. Though extensive, it can only be a start in understanding the varieties of valences of labor in data-intensive science. In particular, more work needs to done to understand the funding process and how seeking funding bends and shapes the contours of labor in data-intensive science. For example, we have not addressed other consequential forms of (in)visible labor, such as funding, building maintenance, instrument building, or IT support for the screens where data-intensive science takes place that fall outside the purview of our data corpus.

In closing, we observe that data-intensive science is a rapidly evolving field with new forms of automation on the horizon. Tools and techniques common to the CKI's early studies of ecology, for instance, are now obsolete. Nor is science immune to broader currents in the economy (see Mirowski 2018). As Bush placed American science on an industrial footing during the Second World War, new techniques and organizational ideas, such as platformization (Rahman and Thelen 2019; Gillespie 2017; Kelkar 2017) and artificial intelligence and machine learning (Irani 2015; Ekbia and Nardi 2014), are presently being adopted into some scientific disciplines, bringing with them new forms of labor. Neither tools and techniques nor organizational theories and business models cleanly replace each other; rather they accumulate and accrete. Elements of the old, prewar style of American science exist alongside industrial and data-intensive styles, as elements of platformization are beginning to take their place alongside them. The need is for more empirical studies of scientific labor that trace connections between scientific practice and wider economic trends.

## Acknowledging our (in)visible labor

Funding provided by: Alfred P. Sloan Foundation
Executive Producer: Christine Borgman
Lead Writer: Michael Scroggins (member, UAW Local 5810)
Assisting Writer: Irene V. Pasquetto
Data Curator: Milena Golshan
Copy Editing: Yookyung Bang

Developmental Editors: Bernie Boscoe, Morgan Wofford, and Milena Golshan of the Center for Knowledge Infrastructure at the University of California, Los Angeles. Peter Darch and Cheryl Thompson, University of Illinois at Urbana-Champaign. Special thanks to Tim Elfenbein and Shreeharsh Kelkar.

Written in Los Angeles, California. Set provided by the University of California, Los Angeles. With special thanks to the staff of the GSEIS department for logistical support.

## Bibliography

Allen, Davina. 2014. *The Invisible Work of Nurses: Hospitals, Organisation and Healthcare*. Routledge.

Bell, Gordon, Tony Hey, and Alex Szalay. 2009. “Beyond the Data Deluge.” *Science* 323 (5919): 1297–98. https://doi.org/10.1126/science.1170411.

Bowker, Geoffrey C. 2008. *Memory Practices in the Sciences*. Cambridge, Mass.: The MIT Press.

Bowker, Geoffrey C., and Susan Leigh Star. 1996. “How Things (Actor-Net) Work: Classification, Magic and the Ubiquity of Standards.” *Philosophia* 25 (3–4): 195–220.

Bowker, Geoffrey C, and Susan Leigh Star. 2000. *Sorting Things out: Classification and Its Consequences*. MIT press.

Burns, Randal, Joshua T. Vogelstein, and Alexander S. Szalay. 2014. “From Cosmos to Connectomes: The Evolution of Data-Intensive Science.” *Neuron* 83 (6): 1249–52. https://doi.org/10.1016/j.neuron.2014.08.045.

Bush, Vannevar. 1945a. *Science the Endless Frontier*. U.S. Government Printing Office.

———. 1945b. “As We May Think.” *The Atlantic*, July 1945. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Collins, Harry M., and Robert Evans. 2007. *Rethinking Expertise*. Chicago: University of Chicago Press.

Coulson, Margaret, Branka Maga, and Hilary Wainwright. 1975. “‘The Housewife and Her Labour under Capitalism’ - a Critique.” *New Left Review; London* 0 (89): 59–71.

Cowan, Ruth Schwartz. 1976. “The ‘Industrial Revolution’ in the Home: Household Technology and Social Change in the 20th Century.” *Technology and Culture* 17 (1): 1–23. https://doi.org/10.2307/3103251.

Crain, Marion G, Winifred Poster, and Miriam A Cherry. 2016. *Invisible Labor: Hidden Work in the Contemporary World*.

Critchlow, Terence, and Kerstin Kleese van Dam, eds. 2013. *Data-Intensive Science*. 1 edition. Boca Raton: Chapman and Hall/CRC.

Daniels, Arlene Kaplan. 1987. “Invisible Work 1987 SSSP Presidential Address.” *Social Problems* 34: 403–15.

Darch, Peter T., and Christine L. Borgman. 2016. “Ship Space to Database: Emerging Infrastructures for Studies of the Deep Subseafloor Biosphere.” *PeerJ Computer Science* 2 (November): e97. https://doi.org/10.7717/peerj-cs.97.

Darch, Peter T., Christine L. Borgman, Sharon Traweek, Rebekah L. Cummings, Jillian C. Wallis, and Ashley E. Sands. 2015. “What Lies beneath?: Knowledge Infrastructures in the Subseafloor Biosphere and Beyond.” *International Journal on Digital Libraries* 16 (1): 61–77. https://doi.org/10.1007/s00799-015-0137-3.

Dominus, Susan. 2017. “When the Revolution Came for Amy Cuddy.” *The New York Times*, October 18, 2017, sec. Magazine. https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-for-amy-cuddy.html.

Edgerton, David. 2017. “The Politcal Economy of Science.” In *The Routledge Handbook of the Political Economy of Science*, 1st ed. Abingdon, Oxon ; New York, NY : Routledge, 2017.: Routledge. https://doi.org/10.4324/9781315685397.

Ekbia, Hamid, Michael Mattioli, Inna Kouper, G. Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R. Sugimoto. 2015. “Big Data, Bigger Dilemmas: A Critical Review.” *ArXiv:1509.00909 [Cs]*, September. http://arxiv.org/abs/1509.00909.

Ekbia, Hamid, and Bonnie Nardi. 2014. “Heteromation and Its (Dis)Contents: The Invisible Division of Labor between Humans and Machines.” *First Monday* 19 (6). http://firstmonday.org/ojs/index.php/fm/article/view/5331.

Fee, Terry. 1976. “Domestic Labor: An Analysis of Housework and Its Relation to the Production Process.” *Review of Radical Political Economics* 8 (1): i–8. https://doi.org/10.1177/048661347600800101.

Foucault, Michel. 2012. *The Archaeology of Knowledge*. Knopf Doubleday Publishing Group.

Gillespie, Tarleton. 2017. “Governance of and by Platforms.” *Sage Handbook of Social Media. Sage*.

Green, Toby. 2009. “We Need Publishing Standards for Datasets and Data Tables.” OECD Publishing White Paper. Paris: Organisation for Economic Cooperation and Development. http://dx.doi.org/10.1787/603233448430.

Harding, Sandra. 2016. *Whose Science? Whose Knowledge?: Thinking from Women’s Lives*. Cornell University Press.

Hills, Denise, Robert R. Downs, Ruth Duerr, Justin Goldstein, Mark Parsons, and Hampapuram Ramapriyan. 2015. “The Importance of Data Set Provenance for Science.” *Eos* 96 (December). https://doi.org/10.1029/2015EO040557.

Himmelweit, Susan, and Simon Mohun. 1977. “Domestic Labour and Capital.” *Cambridge Journal of Economics* 1 (1): 15–31.

Hochschild, Arlie. 1983. “The Managed Heart (Berkeley, University of California Press).” *Hochs Child The Managed Heart1983*.

Howison, James, and Julia Bullard. 2016. “Software in the Scientific Literature: Problems with Seeing, Finding, and Using Software Mentioned in the Biology Literature.” *Journal of the Association for Information Science and Technology* 67 (9): 2137–55. https://doi.org/10.1002/asi.23538.

Irani, Lilly. 2015. “The Cultural Work of Microwork.” *New Media & Society* 17 (5): 720–39. https://doi.org/10.1177/1461444813511926.

Jackson, Steven J. 2017. “Speed, Time, Infrastructure.” *The Sociology of Speed: Digital, Organizational, and Social Temporalities*, 169.

Kelkar, Shreeharsh. 2017. “Engineering a Platform: The Construction of Interfaces, Users, Organizational Roles, and the Division of Labor.” *New Media & Society*, 1461444817728682.

Kitchin, Rob. 2014a. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles, California: SAGE Publications.

———. 2014b. “Big Data, New Epistemologies and Paradigm Shifts.” *Big Data & Society* 1 (1): 205395171452848. https://doi.org/10.1177/2053951714528481.

Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Lave, Rebecca, Philip Mirowski, and Samuel Randalls. 2010. “Introduction: STS and Neoliberal Science.” *Social Studies of Science* 40 (5): 659–75. https://doi.org/10.1177/0306312710378549.

Leonelli, Sabina. 2010. “Packaging Small Facts for Re-Use: Databases in Model Organism Biology.” In *How Well Do Facts Travel?*, edited by Peter Howlett and Mary S. Morgan, 325–48. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511762154.017.

Marcus, Adam, and Ivan Oransky. 2018. “The Data Thugs.” *Science* 359 (6377): 730–32. https://doi.org/10.1126/science.359.6377.730.

Mayernik, Matthew S., Sarah Callaghan, Roland Leigh, Jonathan Tedds, and Steven Worley. 2015. “Peer Review of Datasets: When, Why, and How.” *Bulletin of the American Meteorological Society* 96 (2): 191–201. https://doi.org/10.1175/BAMS-D-13-00083.1.

Mirowski, Philip. 2011. *Science-Mart*. Harvard University Press.

———. 2018. “The Future(s) of Open Science.” *Social Studies of Science* 48 (2): 171–203. https://doi.org/10.1177/0306312718772086.

Mol, Annemarie. 2008. *The Logic of Care : Health and the Problem of Patient Choice*. Routledge. https://doi.org/10.4324/9780203927076.

Morus, Iwan Rhys. 2016. “Invisible Technicians, Instrument-Makers and Artisans.” In *A Companion to the History of Science*, edited by Bernard Lightman, 97–110. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118620762.ch7.

Nardi, Bonnie A, and Yrjö Engeström. 1999. “A Web on the Wind: The Structure of Invisible Work.” *Computer Supported Cooperative Work (CSCW)* 8 (1–2): 1–8.

National Institutes of Health. 2017. “Availability of Research Results: Publications, Intellectual Property Rights, and Sharing Research Resources.” https://grants.nih.gov/grants/policy/nihgps/HTML5/section_8/8.2_availability_of_research_results_publications__intellectual_property_rights__and_sharing_research_resources.htm.

Orr, Julian E. 1996. *Talking about Machines: An Ethnography of a Modern Job*. Cornell University Press.

Paisley, William J. 1968. “As We May Think, Information Systems Do Not.”

Pasquetto, Irene V. 2018. “From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses.” Ph.D. Dissertation, Los Angeles, CA: UCLA. https://escholarship.org/uc/item/1sx7v77r.

Pasquetto, Irene V., Bernadette M. Randles, and Christine L. Borgman. 2017. “On the Reuse of Scientific Data.” *Data Science Journal* 16 (March). https://doi.org/10.5334/dsj-2017-008.

Plantin, Jean-Christophe. 2018. “Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science.” *Science, Technology, & Human Values*, June, 0162243918781268. https://doi.org/10.1177/0162243918781268.

Rahman, K. Sabeel, and Kathleen Thelen. 2019. “The Rise of the Platform Business Model and the Transformation of Twenty-First-Century Capitalism.” *Politics & Society*, March, 003232921983893. https://doi.org/10.1177/0032329219838932.

Ribes, David, and Geoffrey C. Bowker. 2008. *Organizing for Multidisciplinary Collaboration: The Case of the Geosciences Network*. https://repository.library.georgetown.edu/handle/10822/557393.

Roberts, Celia. 2018. “Practising Ambivalence: The Feminist Politics of Engaging with Technoscience.” In *A Feminist Companion to the Posthumanities*, 199–210. Springer, Cham. https://doi.org/10.1007/978-3-319-62140-1_17.

Sands, Ashley E., Christine L. Borgman, Sharon Traweek, and Laura A. Wynholds. 2014. “We’re Working on It: Transferring the Sloan Digital Sky Survey from Laboratory to Library.” *International Journal of Digital Curation* 9 (2): 98–110. https://doi.org/10.2218/ijdc.v9i2.336.

Scroggins, Michael. 2017. “Ignoring Ignorance: Notes on Pedagogical Relationships in Citizen Science.” *Engaging Science, Technology, and Society* 3 (0): 206–23. https://doi.org/10.17351/ests2017.54.

Shapin, Steven. 1989. “The Invisible Technician.” *American Scientist* 77 (6): 554–63.

Souleles, Daniel, and Michael Scroggins. 2017. “The Meanings of Production(s): Showbiz and Deep Plays in Finance and DIYbiology.” *Economy and Society* 46 (1): 82–102. https://doi.org/10.1080/03085147.2017.1311134.

Star, Susan Leigh. 1990. “Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions.” *The Sociological Review* 38 (1_suppl): 26–56. https://doi.org/10.1111/j.1467-954X.1990.tb03347.x.

Star, Susan Leigh, and Anselm Strauss. 1999. “Layers of Silence, Arenas of Voice: The Ecology OfVisible and Invisible Work.” *Comput. Supported Coop. Work* 8 (1–2): 9–30. https://doi.org/10.1023/A:1008651105359.

Suchman, Lucy. 2011. “Subject Objects
Subject Objects.” *Feminist Theory* 12 (2): 119–45. https://doi.org/10.1177/1464700111404205.

Tronto, Joan C. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. Psychology Press.

Tyfield, David. 2013. “Transition to Science 2.0: ‘Remoralizing’ the Economy of Science.” *Spontaneous Generations: A Journal for the History and Philosophy of Science* 7 (1): 29–48. https://doi.org/10.4245/sponge.v7i1.19664.

Vaughan, Diane. 1996. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press.

Velden, Theresa, Matthew J. Bietz, E. Ilana Diamant, James D. Herbsleb, James Howison, David Ribes, and Stephanie B. Steinhardt. 2014. “Sharing, Re-Use and Circulation of Resources in Cooperative Scientific Work.” In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 347–350. CSCW Companion ’14. New York, NY, USA: ACM. https://doi.org/10.1145/2556420.2558853.

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. 2013. “If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology.” *PLOS ONE* 8 (7): e67332. https://doi.org/10.1371/journal.pone.0067332.