Motivating Subjects: Data Sharing in Cancer Research

By Jennifer Tucker

This dissertation explores motivation in decision-making and action in science and technology, through the lens of a case study: scientific data sharing in cancer research. The research begins with the premise that motivation and emotion are key elements of what it means to be human, and consequently, are important variables in how individuals make decisions and take action. At the same time, institutional controls and social messaging send a variety of signals intended to motivate specific actions and behaviors. Understanding the interplay between personal motives and social influences may point to strategies that better align individual and social perceptions and discourse.

Case Study Description

Cancer research is increasingly supported by the discipline of *biomedical informatics*, which integrates principles, approaches, and tools from systems biology, clinical research, and information sciences. Informatics tools help integrate and analyze clinical information, data from biospecimens (such as human tissue), molecular data, and other information in increasingly sophisticated ways and at higher levels of detail than previously possible. The result is envisioned to be "personalized medicine," where clinical interventions for individuals become highly customized (e.g., the right drug for the right person at the right time), made possible because of a highly specialized understanding of cancer at the genetic level. Ironically, "personalized medicine" requires depersonalized data analysis, which requires that large data sets and biospecimens are shared across researchers to aid in detecting patterns in disease and treatment outcomes over a large population.

Data sharing in this complex setting has received great attention in scientific, technical and legal circles, offering a well defined baseline of social norms related to data sharing. How closely do these social norms align with individual motives? To explore these dynamics, this research centered on a large-scale cancer research program led by the National Institutes of Health's National Cancer Institute (NCI). A primary goal of the program, called the cancer Biomedical Informatics Grid® (caBIG®), is to facilitate data sharing between diverse and currently highly autonomous cancer centers across the United States. The program's premise is that increased collaboration and data sharing through shared standards, technology, and infrastructure will lead to better science and faster and more targeted cures for cancer. In the caBIG vision, data vital to cancer research can be shared across researchers on information technology networks or "grids." Once data are formatted to meet shared standards and made available on the grid, other researchers at a range of different institutions can access them.

From the outside, this approach to data sharing appears to carry obvious benefits. Evolving from an individual investigator model of cancer research to a collaborative model across institutions, aided by the advancement of information technology tools and the availability of the Internet to share large volumes of data, seems natural. This shift, however, is by no means inevitable. The technical difficulties of aligning data, practices, tools, and standards across a distributed community are challenging on their own; doing this within a legal environment that sets patient privacy as a top priority, and within the traditional individual contributor rewards structure of U.S. academic research science, is even more of a challenge.

Research Approach

The goal of this research was to compare the social messaging about data sharing with the more personal motives driving this activity. As such, the research was conducted at two levels. The first was at the social level of information, and included reviews of public presentations, documents, articles, and open meetings related to data sharing. The second explored individual beliefs and decision-making about data sharing, explored during private and anonymous interviews conducted with 42 professionals associated with the caBIG program and the broader field of cancer research. Interviewees included clinical and bench scientists and research Principal Investigators (PI); bioinformaticists and information technology professionals; legal and regulatory professionals; and project and data managers involved in data sharing initiatives. The overarching goal of this research was to detect systematic patterns of motivation and emotion in activities and language that influence the personal decisions and social norms of scientific research.

Key Findings

The caBIG program has been clear that data sharing is an overarching goal since the start of the program. Many program representatives use the metaphor of an onion to describe the complexity of data sharing issues; each layer of the onion must be sequentially peeled off to remove the barriers to data sharing. The four layers, which were also used as a structural tool for this research, include: technology, economics, legal and regulatory considerations; and socio-cultural (personal) issues.

<u>The Technology of Data Sharing.</u> The story of caBIG is a technological story, where the evolution of technology drives changes in how labor is constructed, and the kinds of tasks to be performed. At the core, caBIG argues for the automation of data sharing across technological systems, so that resources can be more broadly accessed and used. This is a project that is defined by its goal, but also by the interests behind it. The caBIG program is housed within an organization with an informatics mission; the problem of data sharing is generally defined and expressed as first (though certainly not only) a technological infrastructure need and problem.

The caBIG program generally focuses on the benefits of large-scale, institutional data sharing. Conversely, interviewees report that data sharing is, in reality, far more often conducted as a point-to-point or specialized community-based activity. Researchers are motivated to share data with others under specific conditions: when there is a foundation of trust with the person or community being shared with, when the perceived reward of sharing is well-defined and of value to the person sharing and when there is perceived to be a lower risk or cost than the benefit received. Without these conditions, there are often determined to be insufficient incentives and rewards for sharing.

Data sharing requires labor, resources, and specific skill sets to prepare data in the format needed to share on a large scale. Right now, many don’t yet see the incentive to do so, do not have the skills to do so, and/or do not have access to the people with the bioinformatics tools and skills to do so. Researchers also argue that technologically-based large scale data sharing can also lead to the loss of vital contextual information and researcher knowledge. A data set is generated as an outcome from a question and a subsequent process; these drive how the data is generated and how they appear. Forcing a data set to conform to a set of standards causes the loss of this unique context. Data are both subjective and personal; technologically-mediated data sharing removes this personal element.

The Economics of Data Sharing. The sharing of data is the exchange of potential value. Data, in all forms, carry materialist potential, and economic impacts were consistently raised as leading considerations by individuals in making data sharing decisions. Key variables emerge as *scarcity (availability of data), endurance (longevity),* the perceived *past investment,* and the potential *knowledge* locked up in the data or resource being shared by the person who holds it. Different scientists see different forms of data as being valuable in different ways. The closer a scientist is to the generation and use of data for their research question; the more valuable the data set is perceived to be.

Data sets are seen as material goods that cost money and time to generate, and have the potential for material or professional gain. Economic variables lead to a rather complex personal values-based calculus of investment and reward taken on a case by case basis when researchers are asked to consider a data or resource sharing opportunity. Publications, for example, are both a reward and mechanism to support getting more grants, but they are also a mechanism for developing a professional reputation and as a path to professional collaborations, which are growing increasingly essential in a "team science" funding environment. In this way, publications are a currency for connecting with others, which may lead to interesting collaborations, more grants, more papers, and so on.

The economic and materialist perspective was the most prevalent one in driving researcher choices about data sharing. Unfortunately, the current reality is that institutional rewards such as tenure and paper citations do not directly recognize data sharing; many note that it will take the embedding of recognition for data sharing at the institutional level before true motivational shifts will happen.

The Legal-Regulatory Sides of Data Sharing. While the economic incentives and rewards of data sharing lead to a complex calculus of decision-making about data sharing, they generally appear clear to the people making the decisions. This is not the case with the legal and regulatory factors, which emerged in the research as the most complex and misunderstood area impacting data sharing.

The first legal-regulatory dynamic impacting data sharing is the protection of patient privacy, which places specific constraints and boundaries on data sharing, sometimes with legal punishments if privacy is violated. All acknowledge the balance required between the protection of human subjects in research and the public good that comes from the research conducted with these human subjects. However, the costs of these privacy measures are that they sometimes thwart data sharing where it otherwise might be happening. The second legal-regulatory issue raised in interviews relates to intellectual property rights. Because data may carry both real and potential value, they are often protected by institutions wishing to capitalize on that value. Without adequate ways to protect data ownership, the impulse is to restrict the sharing of that data with others.

Legal and regulatory factors are seen at an individual level as discouraging data sharing. In addition, it is these factors that also are perceived to most minimize the autonomy and control that an individual researcher has over his or her research. Rather than being the researcher's individual choice, the locus of control has been dispersed, as the focus moves from the scientist's choice to the various offices and stakeholders that consider the feasibility of sharing in light of regulatory constraints and intellectual property potential.

The Personal Side of Data Sharing. Social messaging consistently signals that personal concerns, such as authorship and career progression, should be less important than loftier needs such as the progression of science and patient care. At the social level, there are admonishments of individual scientists that are unwilling to share for what are deemed childish, irrational reasons. Scientists are put on the defensive if they choose not to share. Unfortunately, in the U.S. scientific systems that rewards through publications, grants, and patents, the time involved in both originally generating and sharing data is not time that is concretely valued. Sharing data may earn "points" in supporting the ideal of science, but that general framework does not translate well to the day-to-day activities of task and career management. Data sharing is ultimately a service; and yet, despite the fact that science is ultimately framed as a pursuit of knowledge, it is actually fundamentally artifacts and credits driven.

Fear is also a powerful hindrance to data sharing at a very personal level. Many researchers used the same metaphor: much as a reporter gets "scooped" by another reporter that gets the story first, researchers fear getting "scooped" because someone will see something in their shared data that they did not see. A data set is an extension of scientist's intellect and identity. Data elements are not just numbers in a spreadsheet; they are expressions of personal knowledge and an extension of the researcher him or herself. There is value and invisible labor associated with knowing one's data at the intimate level one presumably should, on top of the invisible labor involved in preparing data for others to see. The data defines the value of the researcher.

Based on the interviews conducted for this research, it appears that relationship-related motivators for data sharing all share one common element: they are highly specific and closely related to the researcher in question. People are motivated by sharing with other known people with whom there is a foundation of trust and criteria for credit assignment established.

The Metaphor of Data Sharing. Metaphor is an important conceptual tool in this research to understand data sharing dynamics. Data "sharing" is itself a values-laden metaphor; the terminology pre-disposes the desired answer: share, or withhold. Because of the wording, the question of data sharing is established within the backdrop of a right-wrong, good-bad judgment.

Communication about the need for large-scale data sharing is often framed using naturalistic metaphors and imagery that communicate the sense of being overwhelmed or being taken over by the data. Images such as tsunamis, explosions, standing on a cliff, information islands, or facing an ocean of data send signals that the natural world is a dangerous place that must be mastered so that the data can be shared and used through the structure and order provided by the tools of biomedical informatics. Once researchers get control of the data, however, their metaphors are different. When individuals talk about data, images return to a scale that can be controlled or at least handled by the researcher. In these contexts, researchers describe data sets as something to be "wrung out" as if they were a wet sponge, or as objects to be beaten to death. Data sets are something to be "squeezed" until there is no value left – until all the knowledge has been extracted through personal use by the scientist.

When the discussion turns from the data itself to the people *sharing* data, the metaphors themselves also become far more personal and positive. For example, one interviewee delightfully transformed the tsunami imagery, by reframing the wave of scientific progress into the metaphor of mini-waves bouncing around a kiddie swimming pool. Other more personal metaphors are common when individuals talk about their data sharing experiences. Data sharing is described as "a dating process" as people get to know each other and establish roles and rules of authorship; and even as a "scientific marriage" when talking about two collaborators that come together for creative action over time. Another, less intimate, metaphor was the image of a "data club," small communities of researchers interested in similar research problems and data sets, with their own norms and practices for sharing both data and credit.

There is a tension in these metaphors at a personal level. On one hand, positive relationships terms are used to describe the relationship in which data sharing occurs, when trust is present. On the other hand, when that trust is absent, the metaphor changes to images of "being scooped" evoking the feelings of competition and contest where the previous partner is now positioned as a potential foe.

Conclusions

This work was intended to be a work of both analysis and activism; the goal was to both understand the dynamics of data sharing, and to make recommendations that could better align personal and social perceptions and actions related to data sharing with one another. It is *not* a conclusion of this research that data sharing by definition is a good thing, and that the goal is to shape messaging in a way that will encourage this sharing.

Rather, the primary conclusion is that there are a number of motives impacting a researcher's willingness to share data that are not currently "discussable" at the social level of discourse. The most pressing need, therefore, is to both acknowledge and openly facilitate discussion about the more personal and relationship-based aspects of the data sharing decision: the need for the data sharing that supports relationships and people-oriented trust, the need for career incentives that reward sharing in a meaningful way when it is done, and the need for expanded legal tools that allow the subjective elements of data to be traced and acknowledged as a personal expression of creativity rather than simply as an objective collection of facts. These recommendations share a common focus: better integrating what is happening at an individual level into the institutional systems in which these individuals work; and refocusing communication at the institutional level to better address the concerns and decision-making at a personal one.

Data sharing is both a personal decision and a social level problem. Trusted relationships ultimately form the building blocks of public trust. Data are both subjective and personal; a data set is often an extension of a researcher's identity, and serves as a measure of his or her value and capability. Instead of seeing socio-cultural factors as the final and innermost core of an onion of data sharing complexity, the caBIG program and other bioinformatics initiatives need to reframe the "people issues" as the catalyst that peels the onion in the first place. How can caBIG help get "friends on the grid," where the personal and subjective essence of data is not lost to the technology that will move it to the next place? It is vital to refocus communication to acknowledge the local and personal nature of knowledge, so that grid technologies become not only vehicles for moving data, but also tools facilitating the human connections that ultimately form the heart of cancer research.

Link to full dissertation: http://scholar.lib.vt.edu/theses/available/etd-09182009-161937/

Contact:

Jennifer Tucker
jtucker@tuckertalk.net