Motivating Subjects:
Data Sharing in Cancer Research

By Jennifer Tucker

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy In
Science and Technology Studies

Barbara L. Allen
(Chair)

Janet E. Abbate
Barbara J. Reeves
Doris T. Zallen

September 9, 2009
Falls Church, VA

Keywords: data sharing, biomedical informatics, cancer research, metaphor, reversal theory

Motivating Subjects:
Data Sharing in Cancer Research

Jennifer Tucker

Abstract

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.

To explore these dynamics, this research centers on a large-scale cancer research program led by the National Institutes of Health's National Cancer Institute. The goal of the program is to encourage interoperability and data sharing between diverse and highly autonomous cancer centers across the U.S. Housed in an organization focused on biomedical informatics, the program has a technologically-focused mission; the goal is to facilitate institutional data sharing to connect the cancer research enterprise.

This focus contrasts with the more relationship-based point-to-point data sharing currently reported by researchers as the norm. 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 is both a personal decision and a social level problem. Data is both subjective and personal; it is often an extension of researcher's identity, and serves as a measure of his or her value and capability. In the search for standards and interoperable data sets, institutional and technologically-mediated forms of data sharing are perceived to ignore the subjective and local knowledge embodied in the data being shared. To explore these dimensions, this study considers the technology, economics, legal elements, and personal sides of data sharing, and applies two conceptual frameworks to evaluate alternatives for action.

Table of Contents

Table of Tables

Table of Figures

Chapter 1: Introduction

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. "Emotions...saturate human existence through the lifespan...They can be essential ingredients for, as well as overwhelming obstacles to, optimizing human potential" (Cacioppo et al 173, Ch. 11). Emotions are an omnipresent factor in all of our daily lives, including the lives and work of scientists and technologists.

This premise does not deny the presence of social factors in mediating or influencing human decision-making. On the contrary, the goal is to connect social level norms and interests to the motives and emotions expressed by scientists and technologists. A review of the literature suggests that while motivation and emotion has been richly researched in other disciplines, this area has had somewhat limited coverage by the science and technology studies (STS) community. Despite this, there is a stable conceptual STS platform from which to launch. STS literature is rich with case studies that have "unpacked the technical" to reveal deeper interests. This research aims to identify ways of unpacking these interests further to get to the more personal motivational and emotional elements driving individual decision-making related to these norms.

While questions related to motivation and emotion could be applied to a range of science and technology activities, this research focuses in the realm of a specific representative

case study: the motives and emotions associated with decisions and actions related to scientific data sharing among cancer researchers.

1.1 Defining the Problem and Research Questions

This research is designed to explore the boundaries and points of translation between *social* norms and discourse and *individual* motivation and emotion. There are two overarching problems that bound this research.

First, at the abstract level, there is the need for more actionable ways to better understand and navigate the intersection of the social understanding of how decisions are driven in science and technology (e.g., social-level norms and interests) with the more personal understanding and sense-making that leads to decision-making at an individual level (e.g., personal motivation and emotion). In the end, even if the world is indeed a social construction, it is the people in that world that make actionable decisions. The need to better understand the "handoffs" between social norms and individual decisions is an ongoing problem that deserves attention on many disciplinary fronts.

Second, at a case study level, there is the problem in the social world of cancer research related to scientific data sharing. As scientific questions and analytical tools evolve in cancer research, there is a push towards increased data sharing between scientists. At the same time, economic and legal structures raise mixed messages about the feasibility of such sharing. At an individual level, there are diverse motives and emotions that shape the researcher's decision to share or not to share, educated both by personal values systems and the surrounding

environment and reward structure. Controversies at both social and individual levels define the research problem at the case study level.

These problems lead to the driving questions explored here:

1. How do personal motives interplay with social norms and public discourse? How different are the personal motives underlying decisions and actions from the social norms related to those same decisions and activities? Where do personal motives and public arguments overlap, and where do they differ?
2. How does one constructively and practically study these questions? How does one detect connections between personal meaning and motivation with the social process of science and technology?
3. Based on findings from the first two questions, what advice might be given to those wishing to influence the motives and emotions of those engaged in science and technology?

This research begins from the standpoint that motivation is what inspires action and direction towards a certain goal, and that emotions are the subjective feelings that occur when motives are either fulfilled or not. This project does not specifically aim to explore or validate the various definitions and interpretations attached to the terms motivation or emotion, or their relationship to each other. Instead, it focuses on a more pragmatic and functional treatment of these terms, where motivation is seen as a driver for action; and emotion is the subjective feeling that that either precedes or results from that action. For the purposes of this research, a close connection is assumed between the two: positive or negative emotions may

motivate one to take an action; once that action is taken, positive or negative emotions may result.

1.2 The Case Study: Data Sharing in a caBIG World

Data sharing in cancer research has received great attention in scientific, technical and legal circles, offering a well defined baseline of social norms related to the individual and institutional decisions to share scientific data. A 1946 *Science, New Series* article entitled "Coordination of Cancer Research," sets the initial tone, arguing for a "coordinated program," rather than "lone workers" to fight the "war" on cancer:

> Faced with a problem of such magnitude and complexity, we are inclined to think that its solution must await the chance discovery of some lone worker in the field at some unknown date in the future. In the meantime thousands annually die a lingering death at the hands of this killer. Actually, this menace should be regarded in the same light as any military foe that might claim the lives of thousands of Americans before their time … What is needed here is a well planned and completely coordinated program, directed by a group of experts in the field and serving to organize the activities of all competent investigators (Pilcher 167).

Today, the National Institutes of Health's (NIH) National Cancer Institute (NCI), the primary government organization charged with "eliminating the suffering and death due to cancer," positions collaborative science as a critical element of its strategic plan:

> Never before have so many scientific tools and so much biomedical knowledge been assembled to power our ability to reach our Vision to eliminate the suffering and death due to cancer by 2015. We as a Nation will achieve this Vision by optimizing new approaches in interdisciplinary collaboration and transdisciplinary science (National Cancer Institute, "NCI Strategic Plan" 2).

Often, public statements position collaborative science and data sharing as being related, even natural, companions. Associated with scientific collaboration is scientific data

sharing. The 2006 NCI Strategic Plan notes the following activities it will undertake as examples of that linkage:

> Invest in effective infrastructures to promote a high degree of integration, coordination, and communication along the discovery-development-delivery research continuum… (40)
>
> Link preclinical research data with a comprehensive database of clinical trial results to coordinate and optimize information and data sharing…. (41)
>
> Build new partnerships and multidisciplinary collaborations to ensure the unprecedented level of integration required to realize the tremendous potential for improved cancer treatments arising from current scientific advances…. (44)
>
> Coordinate and optimize patient-information and other data sharing by creating a comprehensive database of clinical trials and results (44).

These quotes closely link the ideas of coordination, collaboration, infrastructure, and data sharing as critical to the vision of eliminating suffering and death due to cancer. Data sharing and collaborative science are, however, deceptively simple concepts. This simplicity begins with the normative nature of the term "data sharing" in the scientific context. In U.S. culture, we learn early the social norm that the action of sharing is good, and withholding is bad. Tightly coupling collaboration with data sharing thereby establishes a discourse of generosity that seems hard to argue with. The metaphor of sharing will be addressed more specifically later; however, it is interesting to highlight this choice of word at the very start. At its very beginning, we have data positioned as a public good that is to be shared; not an item of ownership that is sold or stolen.[1]

Despite this normative starting point, however, there are many factors that serve both to encourage and hinder data sharing. These factors will be considered throughout the

[1] Consider a comparison: Online "sharing" of electronic music files is usually referred to as data piracy; sharing music is stealing, sharing scientific data is a social norm.

following chapters. First, however, is the introduction of the case study that serves as a portal into these dynamics at both the social and individual level: NCI's cancer biomedical informatics grid© (caBIG®[2]) program.

The caBIG program is described by the NCI as "a 21st century information initiative that will transform the way we do cancer research." Led by the NCI's Center for Bioinformatics and Information Technology (CBIIT), the caBIG mission is "to develop and deploy a 'world wide web' of cancer research, allowing cancer researchers to share both tools and data among cancer centers to facilitate scientific discovery" (National Cancer Institute, "caBIG Community Website" par 1-2). The program's premise is that increased collaboration and data sharing through shared standards and infrastructure will lead to better science and, ultimately, faster and more targeted cures for cancer.

Understanding the significance and difficulty of the caBIG mission requires a basic consideration of two dimensions: first, the underlying structure of NCI-funded cancer research; and second, a brief introduction to the nature of cancer research today.

In its budget positioning paper, "The Nation's Investment in Cancer Research: Connecting the Cancer Community: An Annual Plan and Budget Proposal for Fiscal Year 2009," the NCI reports:

> Nearly 80 percent of NCI's budget funds extramural research activities—research taking place at institutions across the country and around the world. The Extramural Research Program supports cancer research in nearly 650 universities, hospitals, Cancer Centers, and other sites throughout the United States and in more than 20 countries. The majority of NCI's extramural funding

[2] caBIG® is a registered trademark of the National Cancer Institute. Permission to use the trademark for research purposes are granted through the following statement on the caBIG Website: "Researchers and participants in the caBIG® Initiative are encouraged to acknowledge the contributions of the caBIG® project in all abstracts, presentations and published manuscripts. Both the overall caBIG® project as well as specific tools can be acknowledged or cited."

supports investigator-initiated Research Project Grants (31).

The implication of this funding structure is that despite interest in and rewards for coordination and collaboration, much of NCI-sponsored cancer research is fundamentally decentralized, and is funded and conducted in individual laboratories across a diverse array of organizations. Each of these organizations has its own collection of biospecimen (tissue and other biological materials from humans and other animals) repositories, clinical and research tools and systems, policies, hierarchies, reward structures, and cultures.

Within this distributed landscape is the evolving nature of cancer research itself. Describing the evolution of this discipline is a study unto itself; as an alternative, this section outlines the foundational ideas that have led to today's interest and focus on data sharing.

In summary, today's cancer research enterprise is increasingly supported by the discipline of *biomedical informatics* (more broadly called bioinformatics), which integrates principles, approaches, and tools based on systems biology, clinical research, and the computational (information) sciences. Tools from bioinformatics help integrate and analyze clinical information, data from biospecimens, molecular annotations, and other information in increasingly sophisticated ways and at higher levels of detail than previously possible. The evolution of such systems-based approaches and tools has facilitated the co-evolution of cancer research into discovery focused at the genetic and molecular level. 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 the highly specialized understanding of cancer at the genetic level. NCI explains this scientific

integration in their materials explaining the benefits to donating biospecimens for cancer research (National Cancer Institute OBBR, “Patient Corner”):

> Biological research has moved into what is called the "genomic age". This designation refers to the ability of scientists to study disease at the most basic "molecular" level, by identifying genes and their function, and understanding the role genetics plays in the origin and progression of disease. Other emerging fields of study include proteomics - the study of the full set of proteins encoded by the genome - and pharmacogenomics, which seeks to link the human genome to variation in patient response to pharmaceuticals. In addition to molecular information, scientists are also analyzing a vast amount of clinical information from patient records and clinical trials. From this data, it is possible to identify patterns that provide a pathway to understanding disease sub-types, and potential strategies for diagnosing and treating disease in new and more effective ways (Par. 4).

Ironically, “personalized medicine” requires depersonalized data analysis, which requires that large data sets and/or the biospecimens from which they are often derived be available and shared across researchers, to aid in detecting patterns in disease and treatment outcomes over a large population (Dawyndt et al 249-58; Ginsburg et al 1359-61; Park, Screen 8). To achieve this personalized medicine, large scale data sharing and data management become critical needs. An introductory bioinformatics text explains (italics added):

> Cancer systems biology seeks to elucidate complex cell and tumor behavior through the integration of many different types of information…. The classical techniques of statistics and bioinformatics for analysis of the genome, biological sequences, large-scale ‘omic’ data sets and protein three-dimensional structure will continue to form an indispensable backbone for computational cancer research, whereas new systems-based approaches will extend our knowledge of the organization and dynamic functioning of the implicated biological systems…. *Complementing the methods of systems biology, new data management technologies to enable the integration and sharing of data and models are also a prerequisite for advancement* (Nagl 24).

Storing scientific data and conducting these analyses requires a robust set of software and infrastructure tools; as well as policies and standards that allow data to be shared between

and across those tools and infrastructure. These science and technology research advances, however, are emerging within the decentralized and distributed network described above, where many individually-funded Principal Investigators and departments have created their own tools of varying levels of technological sophistication for the data storage and analysis required for their specific research. These tools range in technological complexity, ranging from paper-based notebooks that catalog local inventories of biospecimens to large-scale electronic systems that store extensive data sets with multiple stores of inter-dependent variables.

The caBIG program was launched in 2004 to help connect the cancer research network across these diverse tools and infrastructure. The overall goal is to "connect the cancer research community through compatible tools and a shareable, interoperable electronic infrastructure; and deploy and extend standard rules and a common language to more easily share information" (National Cancer Institute, "Special Report: caBIG" 5). Ultimately, caBIG was designed to demonstrate that the sharing of cancer research tools (software) and data among cancer research institutions is preferable to each institution and even each laboratory continuing to develop its own tools in the traditional "Principal Investigator" model of scientific research.

A closer look at the management of biospecimens and their associated data in this evolving research landscape is an effective case study to illustrate the goals and ideals of caBIG. A tool commonly developed by individual cancer centers, and even individual departments or researchers, is a tissue banking system that catalogues and tracks biospecimens collected from cancer patients. Tissue banking systems that are *interoperable* between departments and cancer centers (locally owned systems that can connect with and exchange meaningful data

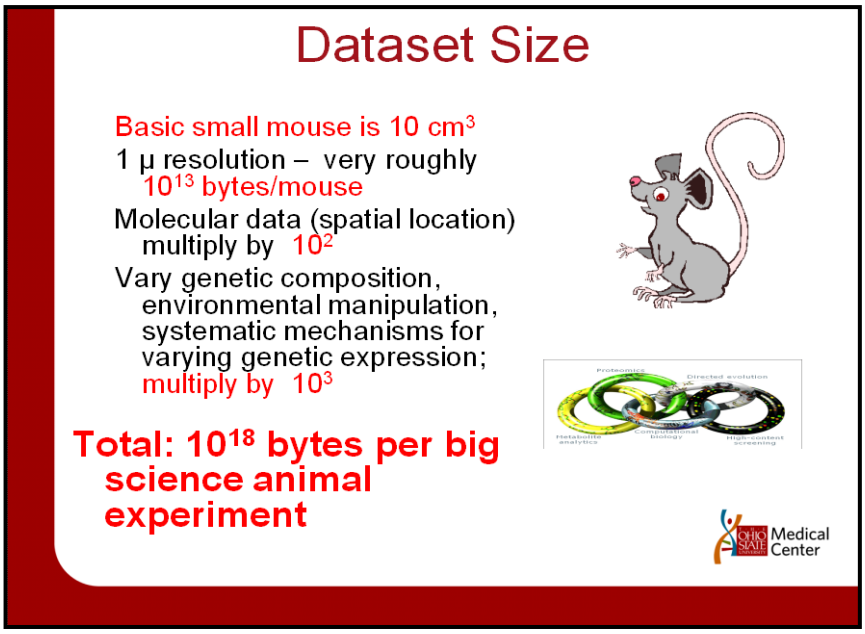
with other local systems) allow cancer researchers to identify biospecimens and associated data held by other researchers, and share information about the specimens and data that they themselves hold. This sharing increases the pool of target biospecimens and data available for specific research projects – a larger population, it is argued, will lead to better science, because more factors across diverse populations can be studied with statistically significant results. The caBIG program supports the achievement of this interoperability.

A 1999 RAND study estimated that there were more than 307 million tissue specimens from more than 178 million cases stored in the United States, accumulating at a rate of more than 20 million specimens per year (Eiseman et al 1). Tissue specimens are stored at a range of facilities, including military facilities, NIH and its sponsored facilities, other federal agencies, state collection agencies, diagnostic pathology and cytology laboratories, university- and hospital-based research laboratories, commercial enterprises, and non-profit organizations (Eiseman et al 1). It is not just the biospecimen that holds value for cancer research; it is also the associated data that comes with that biospecimen, including such information as the demographics of the donor, the treatment regimen that was used; and the outcomes from that treatment where available.

The slide below, presented at the 2007 caBIG Annual Meeting conference, provides a sense of the scale of data involved in this kind of scientific and technological activity, even when focused only on the tissue samples relevant to cancer research. This researcher estimated that one laboratory mouse could generate 10^{18} bytes of information (enough to fill many thousands of today's standard desktop computers); when extrapolated to cancerous tumors from humans

and other species, the enormous potential volume of data available to support cancer research, and the challenges involved in managing and sharing it, become even clearer and more urgent.

Figure 1: Illustration of the Data Generated from a Single Mouse[3]



From the outside, the caBIG mission of interoperability and 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. Biomedical informatics is a complex merger of biological research, medicine, and information technology. The technical difficulties of aligning data vocabularies, 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

[3] This figure is from a government conference presentation that is also posted on the Internet, Reference: Saltz, Joel. "caBIG: Envisioning the Future." *2007 caBIG Annual Meeting*. Washington DC, February 6, 2007. https://cabig.nci.nih.gov/2007caBIGconference/presentations/tuesday-february-06-2007/session-block-3/breakout-sessions/cabigTM-envisioning-the-future> It is used here under a Fair Use Determination.

top priority, and within the traditional individual contributor rewards structure of U.S. academic research science, is even more of a challenge. One researcher noted in a published article:

> "We're asking researchers at many competitive institutions to tear down barriers to sharing vast amounts of data," says Howard Bilofsky, senior fellow at the Center for Bioinformatics at the University of Pennsylvania, which participates in NCI's project. "Being able to share information in grids across the world in the arena of life science research is not something easily done" (Mashberg 2).

The caBIG program acknowledges the challenge. In an article in *The Scientist*, Kenneth Buetow, the caBIG Director for the NCI positions the enormities of the project:

> We intend for researchers to share not only their software, but also their data, where possible. There are obvious challenges associated with data sharing between industry and academia, and between academic researchers vying for the same funding pools. Also, appropriate protections need to be provided for research participants who have generously donated information and material. One caBIG workspace, the data sharing and intellectual capital workspace, focuses on these issues. It integrates expertise from technology transfer specialists, legal counsel, ethicists, security experts, institutional review boards, privacy authorities, and the advocacy community, among others, to create frameworks that guide data sharing and address security and protection of human subjects (Buetow, "Heading for the Big Time" 4-5).

Given the diverse communities and content involved in both data sharing and in cancer research, *data* and the act of *data sharing* are in fact, very diverse and differently understood terms; and “it depends on what you mean by data,” becomes a very common statement when speaking with people about data sharing. Despite this caveat, however, there is enough common ground in the definition of data and data sharing between cancer research areas for “data” to be considered a “boundary object” (Bowker and Star, *Sorting Things Out*), an object or concept that serves as an interface point between different communities or social worlds. The social worlds of cancer research, for example, include such distinct communities as clinical researchers, molecular science researchers, informatics professionals, and a range of other

specialties. There are also the different social worlds of institutional interests such as pharmaceuticals, government agencies, and academic cancer research centers. In fact, it is the positioning and flow of data as a critical potential interface activity between researchers that makes data sharing so controversial, and such fertile ground for exploration, in the multi-dimensional social worlds of cancer research.[4]

In this research, "social level" factors refers to public discourse and institutional-level arguments (which may include individuals speaking on behalf of institutions). "Individual" discourse relates to the arguments made by individual people in talking about their own data sharing experiences, or the experiences that they have seen or heard about on a personal level. Understanding social-level factors points to potentially corresponding motives and emotions at an individual level. What motives and emotions are linked to the scientist's personal decision to share cancer research data, and how do these motives and emotions align – or not align – with social norms and discourse related to the benefits or risks of data sharing? These are questions that have not been adequately addressed in the literature on data sharing, and it is the project undertaken here.

1.3 Grounding in STS Literature

Many STS case studies have focused on unpacking the norms and interests that underlie scientific decision-making and action. Feminist studies have extended this with the tools of standpoint epistemology and partial perspectives: vital forces in shaping different forms of

[4] It is the use of "data" as a boundary object in this research that leads to the default use of the term "data" in the singular, despite the common grammatical practice of the term being considered plural. This research is less concerned about the data themselves than the boundary object that data represents. This research is concerned with the personal and social dynamics of sharing of a unit of value, called data. As such, it is appropriate to keep the term singular for this purpose.

knowledge. A current research gap, however, is the next step deeper into the more personal and subjective motives and emotion that drive scientific and technology work. How does personal motivation and emotion contribute to, and how are they influenced by, social norms and interests? By learning more about how a range of motives and emotions impact the development of science and technology, we continue to broaden our understanding of the social dimensions of these activities. To this end, this section grounds the proposed research and case study in existing STS resources, integrated with a consideration of the elements related to motivation and emotion that may be useful in connecting social norms with personal experience.

The Norms and Counter-Norms

The conceptual heart of this research work lies in the foundational work of Merton and Mitroff related to the norms and counter-norms of science, as these two works reflect the tension between the impersonal and the personal character of science (Merton 270-78, Ch. 13; Mitroff, 579-95). The following table summarizes their comparative views.

Table 1: Merton's Norms and Mitroff's Counter-Norms

Merton's norms of science: *Theorized to be the key driving active factors in encouraging social and institutional controls and stability in science.*	***Mitroff's counter-norms of science:*** *Based on interviews with Apollo scientists, proposed as an alternative theory about how science operates.*
• **Universalism** – Scientific work is completed through the use of impersonal evaluation and criteria, focused on the scientific or technical problem at hand. • **Communism** – The output of scientific work is knowledge that is given to the community through publication in journals and presentations at conferences. Once	• **Emotional Commitment –** Scientists choose research questions and directions that they have passion and an emotional commitment to. • **Particularism** – Scientists do not weigh scientific research and reports equally; their assessment of the research is driven in part by their subjective knowledge of the

Merton's norms of science: *Theorized to be the key driving active factors in encouraging social and institutional controls and stability in science.*	***Mitroff's counter-norms of science:*** *Based on interviews with Apollo scientists, proposed as an alternative theory about how science operates.*
released, science is a public good. • **Disinterestedness** – Scientists must detach themselves sufficiently from the problem at hand to become separate from it. It is only through this distance that the moral integrity of scientific work is maintained. • **Organized Skepticism** – As a group, scientists must be committed to continuous questioning and debate about their work. A key need in science is to reveal new facts, which can only be done by raising new questions, and entering with the willingness to be proven incorrect.	author's personal characteristics. • **Solitariness –** Scientists often hold their research and data as private and proprietary until they are able to publish the results. • **Interestedness –** Scientists have a stake and interest in the outcome of their research questions. • **Organized Dogmatism.** There is a tension between accepting and acknowledging previous work as a foundation for one's own, and rigorously questioning its truth. One believes one's own work is true and is willing to advance it, while questioning everyone else's.

As Merton and Mitroff both acknowledged, and Mulkay ("Norms and Ideology of Science") further noted, the question is not which of the norms or counter norms are true or not true. Rather, they together provide a framework for understanding how norms and counter norms interplay to reflect the true nature as well as the ideal of science; and how they are mutually positioned in communicating about science to the public. Connecting to the research question here, how do these norms and counter norms relate to the motivational lives of scientists? In her 2007 research article, Nancy Jones (32) notes that, "little work has been done to test the adoption or practice of norms for scientists." Jones points to two studies that found that while more faculty and students report practicing the norms *themselves*, they see *others* around them acting more often according to the counter norms than the norms. While the norms may be the ideal for *self*, the impression is that counter norms are those most experienced *from others*. This invites the question, do people perceive themselves as adhering

to the norms, while others perceive them as adhering to their opposites; or are people simply more likely report a personal affinity to the idealized norms in a survey because they sound "better" than the counter norms do? How do we detect norms and counter norms in the first place? What motivates a scientist to embrace one or the other?

This is a question addressed in the research ahead, but it is worth theorizing about what motives and emotions one might hear that would align with the norms and counter norms. For example, while the norms might at first appear to signal a more non-emotive and dispassionate side of science, there are motives and emotions that might result from the motive of objectivity and contributing to the communal greater good. These might include:

- Being motivated by the pursuit of truth, outside what any one scientist is working towards. At its best, this could be experienced as transcendence, or the intellectual independence that comes from delving deeply into an objective problem.
- Being motivated by the greater good of "giving" to science and other scientists; of doing the "right thing" and belonging to a larger community. At its best, this could contribute to a sense of generosity, belonging and selflessness.

On the side of the counter-norms, there may be a slightly different set of motives and emotions:

- Being motivated by personal devotion to the research problem at hand. At its best, this could contribute to feelings of both pride and passion, and protectiveness and advocacy of the intellectual contribution of oneself and the team.
- Being motivated by the connection to the particular people involved with the research, and even a healthy sense of competition with other research teams. At its best, this

could contribute feelings of connection with collaborators, and the satisfaction of achieving personal and team goals.

Motives and emotions that could easily cross between the norms and counter-norms might include:

- Confidence – Motivated by one's own ideas, and the faith in ones' ability to evaluate their own work and others
- Fear – Motivated by concern about being outdone by others
- Curiosity – Motivated by wanting to know the "answer"
- Skepticism and Distrust – Motivated by finding flaws in existing research (one's own and others)

This is a broad list of possible motives and emotions. How have others seen such an attempt? Throughout *Fear: A Cultural History*, Joanna Bourke argues that *fear* is the most pervasive emotion of modern society, and that this emotion shapes much of how we construct our lives. Along these lines, in *Master Passions: Emotion, Narrative and the Development of Culture*, Moldoveanu and Nohria present the argument that *anxiety, envy, greed, and jealousy* are key cultural drivers in human activity. There are, of course, a host of other motives and emotions that researchers may experience, reflecting a natural motivational continuum extending from impartial rationality to a more passionate subjectivity. How are these norms and counter-norms reflected in the arguments used to discuss data sharing, both at an individual level and at a social one? One of the hypotheses for this research is that the norms and counter-norms reflect a range of motives that are *all* experienced by scientists and technologists over the course of their work. As originally envisioned by Merton and Mitroff, the

norms and counter-norms do not reflect an either-or choice; they are both present. The question then becomes: how are they expressed and exchanged, and how do they impact individual decision-making when it comes to sharing data?

Institutional and Economic Perspectives

Work related to interests and reward structures also provides a launching point for research on data sharing. While vital in revealing socio-cultural dynamics, interests and rewards generally focus on the *positive and internal* social-level motivators shaping the work of scientists, such as the community recognition associated with publication in a prestigious journal, unspoken differences due to class distinctions and accessibility; or the continued funding that may come when questions are framed in ways of interest to particular industrial or political groups (Barnes, *Interests and the Growth of Knowledge*; Merton, Ch 13). STS work in interests and rewards is generally framed at a social and institutional level and generally over a long-term perspective, rather than at the level of specific actors and the diversity of motives and emotions that may be felt within a single research project, or by someone in specific role.

One could argue that social level interests are not only linked to the positive emotion of pride associated with rewards, but also to the negative emotion of embarrassment or fear, which could come when from *not* getting rewards, if indeed one is within the social world competing for them in the first place. At an individual level, *incentives* motivate and precede action; *rewards* follow and hold the praise for that action. What are these incentives? For scientists in cancer research, it may include grants and other forms of funding, written in a way to encourage certain behaviors; or requirements from publishers to adhere to specific guidelines when submitting papers for publication. This study will consider the interplay

between individual motives and decisions, and the overarching context of institutional interests and rewards within which those decisions are made.

Other institutional perspectives are provided by the constructs of modernity and post-modernity, and by the work of Michel Foucault. Modernism and postmodernism are often proposed as two opposing constructs for conceptualizing science and technology (Latour, *We Have Never Been Modern*; Harvey, *The Condition of Postmodernity*). Modernism reflects a quest for institutional consistency and control, reflecting a drive towards the efficient management and leveraging of capital and the economies possible with large scale approaches. Postmodernism, conversely, reflects a quest to "return to the local," to recapture the diversity of approaches and knowledge residing in local systems. In the past, these constructs have been used to understand the evolution of scientific and technological knowledge systems and means of production; here, they are used to better understand the goals and practices proposed by caBIG.

The work of Michel Foucault has also been used to understand the institutional structures of science, focusing specifically on the issues of power and institutional control (*Power/Knowledge*). Foucault focused particularly on the link between knowledge and power, and the role of institutions in exerting control over individuals through the creation and definition of the institutional structures within which they operate. The importance of these dynamics lies not in overt expressions of power and hierarchy, but rather, by more subtle processes of surveillance and discipline communicated through specific choices in discourse and institutional structures. Foucault argued that those with certain types of knowledge are able to construct categories that serve to either elevate or subjugate different members of

society, in subtle but real ways. Detecting the process and outcomes of knowledge-power at work is somewhat illusive, as different sources of power are held by different parts of the network. In this research, the work of Foucault is used to better understand the labor and power dynamics that go to the very heart of where an individual researcher's control and autonomy begins and ends.

Theories related to material and moral economies also provide useful bridges for understanding the connections between the social and the individual, and the relationships between objects, subjects, and emotion. Arguments for the pivotal role that material concerns and economic structures play in the evolution of science and technology are grounded in Marxian thought, but have been demonstrated in a range of settings, generally in macro-structural terms. It can be argued that many science and technology decisions in the United States are generally driven by materialist interests, and that financial considerations are ubiquitously used to evaluate and prioritize science and technology needs. This research probes the role of economic and material interests in influencing data sharing at a personal level, including an assessment of how the value of cancer research data is perceived, and how it differs depending on the role and perspective of the researcher deciding to share or not share.

In this light, the work of Hans Joas in the *Creativity of Action* is also an effective theory construct on which to highlight both the economic and creative value embedded in the data sets generated and shared by cancer researchers. Blending Marxian theory and economic pragmatism, Joas is interested in how subjective expression is translated into production and innovation. This is a valuable shift that can also be applied in reframing how we think about emotion. Following Joas' model, emotion can be reframed from an adjective (I feel trust and

caring; or I feel fear) to an action that it inspires (I share data; or I withhold it). This work also highlights the question of what determines the existence of creative expression in the first place. Is a data set merely a collection of objective facts, or does it reflect a subjective creative expression that needs to remain linked to the data creator in order to retain its value?

It is not just material economics that influence data sharing; moral economies are at play as well. As such, another work that helps set the stage for understanding the labor issues associated with data sharing is Robert Kohler's *Lords of the Fly*. Kohler articulated the concept of "moral economy" as a driver in the distribution and sharing of scientific materials within the original *Drosophila* fly group, and as a decision criterion for how materials would be distributed beyond the group. A moral code emerged in the original fly group helped shape boundaries of collaboration: including who was in, and who was out in terms of both relationship and content.

Kohler's work, however, also points to a possible danger to be controlled for in any work on emotion when describing relationships and personalities among individuals. It is easy to over-interpret the emotions and motives of others, extending into projective storytelling rather than analysis. Extrapolating the meaning of decisions from personal motives and feelings is risky ground to be avoided.

The story of caBIG is a technological story, a 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 and streamlining of data sharing across technological systems, so that resources can be more broadly accessed and used. In *The Social Life of Information*, Brown and Duguid point to the potential deceptiveness of the "streamlining" desire and

expectations. Through case studies, these authors point to instances where labor has been made invisible, but no less demanding, by information technology. The ability to access electronic research journals, for example, hides the labor that is involved in making those available to users. The other "invisible" impact to labor is the additional time spent on transactions – the time and cost spent to manage the technology as it is introduced, and to process the information now available to us. These observations have important implications for considering how, and how much, data actually is *desirable* to be shared and what labor costs are involved, both visible and not.

Brown and Duguid also highlight the social nature of technological problem solving; social structures naturally allow for people to share specialized knowledge, leveraging information more effectively across a group. This undermines the framing of people as "information handlers;" rather, there is context underlying the sharing of information that should not be lost (Ch. 4). This dimension, we will come to see, is vital in understanding both the dynamics of and objections to data sharing.

Finally, specifically grounded in the domain of interest in this project, Devra Davis' *The Secret History of the War on Cancer* provides a compelling argument of how the driving factors of institutional politics and economics have shaped cancer research. With a few notable exceptions, often linked with a less privileged class, this history portrays scientific researchers as somewhat powerless agents reacting to events around them as a result of primarily fear and greed, and – again with exceptions - more shaped by institutional interests than shapers of them. How do scientists in the field today describe their own choices, fears and ambitions?

How do their personal stories of choice either validate or differ from social-level analyses such as Davis'?

Situated Knowledge and the Power "To": Feminist Perspectives

Feminist resources are also critical to the construction of this research. Motives and emotions are ultimately individual and subjective, naturally drawing in Harding and Haraway's ideas related to strong objectivity, standpoint epistemology, partial perspectives, and local knowledge. These authors call for an approach that blends institutional (e.g., the best objective) knowledge with the more contextual knowledge provided by the people closer to the situation at hand (Harding 145-168, Ch. 14; Haraway 169-88, Ch. 15). Knowledge generated from a particular stance or view (one interpretation of standpoint epistemology) is, by definition, personal subjective knowledge, with all the motives and emotional "baggage" put hand-in-hand with "objective experience and facts." By calling for the integration of personal knowledge with institutional knowledge, we call for the blending of traditionally distinct worlds: the public sphere of science described by objectivity and non-emotive displays of knowledge; and a more private sphere of feelings and more subjective knowledge. This research explores how these worlds interplay: how do we detect and define a partial perspective, and how do personal perspectives and social knowledge intersect with and inform one another?

Supporting these ideas, Stephanie Shields addresses the normative nature of emotion through the viewpoint of psychology. In "Politics of Emotion in Everyday Life: 'Appropriate' Emotion and Claims on Identity," Shields describes the different rules that govern the degree to which it is acceptable to display different types of emotions in different situations (amount of emotion, and fit with situation). This work echoes Sandra Harding's thesis that who owns the

knowledge often drives its acceptance (Harding 145-68, Ch. 14). In this case, Shields moves beyond "whose knowledge" to "whose emotion." Are emotional arguments, either for or against data sharing, more or less likely to impact decision making? Multiple perspectives may lead to conflict, anger, and fear. The potential emotional implications – and the potential psychological messiness - of feminist "theory in practice" is a vital element to explore in research involving motivation and emotion.

Once questions of "whose knowledge and emotion" are introduced, questions of leadership and its impact on shaping action emerge as well. On the positive side of affective leadership, Shamir et al and Conger et al propose that leaders are most effectively charismatic when they build positive self-concepts in their followers. This is consistent with general models of motivation, which presuppose that we are motivated to do what we do because we want to feel good about ourselves. Other outcomes attributed to successful charismatic leadership in these studies included organizational identification (wanting to do something for the larger organization) and values identification (the leader confirms the follower's values). Are arguments for or cautions against data sharing more likely to be accepted when they are related to a researcher's goals, when they are framed within the larger goals of cancer research or an organization's mission, or when they engage at a values level?

Actor Network Theory and Object-Oriented Analysis

Research about data sharing requires a consideration of the diverse forms that data can take, ranging from physical objects, to analytical methods and materials, to data elements in electronic databases. For example, the data associated with and derived from biospecimens

are often of vital interest to researchers, for they represent the raw material from which research can be conducted. In fact, the broadest consideration of data in the field of cancer research can include the materials used to prepare samples, the methods used to analyze different forms of data, the contextual information (metadata) that annotate the core data elements, and so on.

Given the diverse range of both people and objects involved in the cancer research network, actor network theory is therefore another useful framework for this research. In the tradition of Latour (*Science in Action*), keeping objects, from biospecimens to their associated data to cancer therapies themselves, present as actors in the research; and noticing how the presentation of motives and emotions change based on what objects are being considered, is an active goal. Are discussions concerning human tissue more emotive than discussions about data derived from those tissues? How do discussions about sharing human biospecimens differ from the discussions about sharing the large data sets that are derived from those tissues? What data sets are people more or less willing to share?

Latour's work on "quasi-objects" in *Science in Action,* beautifully informs this work. First, as a concept, it helps reflect the continuum between "subject" (human donating tissue) and "object" (a data set derived using this tissue). In fact, this research also explores whether motivation and emotion itself could be observed and exchanged as a quasi object (Latour 51-55). If motivation and emotion is conceptualized as something that can be exchanged, it may be possible to "trace" their pathways and possible impacts between individuals they traverse. Again, feminist conceptions of power and perspective support this analysis. Partial perspectives and situated knowledge are useful tools not only for cultural perspectives and

understanding local knowledge systems, they also support the understanding of how people perceive data. Data sets are personal, as are tissues from which data is derived. The idea that one data set may hold very different knowledge for one scientist than another seems useful in understanding the motives related to sharing it.

The location of the power to share also seems a particularly important question: where does the power to share "live" in today's cancer research enterprise, and how does that shift with the increasing introduction of technology tools that support data sharing? This is a question that might be usefully informed by feminist constructs as well as actor network theory, and requires the consideration of power as a dynamic element on a network that extends from patients and tissues (from which data is derived), to scientists and their materials and data, to lawyers and their intellectual property and regulations, to the technologists establishing the technical infrastructure across which data is shared.

Another focus area in actor network theory is an emphasis on practice: the "doing" of science in laboratory environments. This research offers the opportunity to integrate motivation and emotion into this landscape, by focusing on the idea of emotion and "sensemaking" in organizational settings. Weick et al, Berscheid et al, and Peters et al discuss the role of emotion with respect to task accomplishment and "making sense" of the activities and emotions that one experiences each day.

> Sensemaking in organizations will often occur amidst intense emotional experience. As interdependent partners (e.g., partners exchanging data) learn more about each other and move toward closeness by becoming increasingly dependent on each other's activities for the performance of their daily behavioral routines and the fulfillment of their plans and goals, the number and strength of their expectancies about each other increase. As a result, their opportunities for emotional experience also increase (Weick et al 419).

Peters et al note, “Seemingly trivial social talk provides fertile ground for emotion sharing (a narrator and audience’s realization that they experience the same emotional response toward a target), which in turn creates a coalition” (780). Weick at al may ask it best: Are “institutions better portrayed as cold cognitive scripts built around rules or as hot emotional attitudes built around values?” (419) This research assumes that these are variables that impact the activities of human actors within a social network, and aims to detect both that diversity, and its impact.

Metaphor

Discourse and personal stories about data sharing are full of both linguistic and graphical imagery; these metaphors help both frame problems and shape positions about data sharing. In light of this, metaphor serves as an important conceptual tool in understanding data sharing dynamics. To establish this use, this section provides a brief review of the role of metaphor as a conceptual tool in science and technology studies, focusing on three specific uses of metaphor that apply to this study. The first is the use of metaphor as a tool for cognitive understanding: a tool for explaining a new idea in terms that are already understood. The second is the use of metaphor for political advocacy: a tool for shaping opinions in a way that the metaphor presenter wishes them to be shaped. The third is the use of metaphor as a tool of physical alignment: a tool for orienting a line of thought to connect closely with the physical embodiment of an activity.

Donald Schön presents metaphor as the projection of an image or “frame” from one domain to another in order to help establish the understanding of a new idea in terms of one

that is already understood: “Metaphor refers both to a certain kind of product – a perspective or frame, a way of looking at things – and to a certain kind of process – a process by which new perspectives on the world come into existence” (137). The power of this method is that it allows for conceptual shortcuts in understanding; there is a baseline of knowledge that is carried with a term from one domain to another. This “transport of the familiar” facilitates both discovery and communication, as new ideas, practices and tools are discovered and positioned. In his writing on *Models and Archetypes*, Black writes, “A metaphor operates largely with commonplace implications. You need only proverbial knowledge, as it were, to have your metaphor understood” (239). Using commonplace explanations to describe difficult concepts helps them become both established and able to be transmitted to others.

There is benefit in this activity when it helps quickly make the “unexplained explainable;” when a new idea can be filtered through the baseline of another. Unfortunately, as George Lakoff illustrates in his work related to the Gulf War communication (“Metaphor and War”), and Emily Martin illustrates in her work related to the communication of human reproductive processes (“The Egg and the Sperm”), this is not a neutral activity. Metaphors carry values and meaning that can influence how the recipient reacts to and processes the relationship being presented. Metaphors can be used to pre-position political agendas under the cover of the seemingly objective process of establishing understanding.

This connects directly with the use of metaphor to evoke emotion. In *Metaphor and Emotion*, Kovecses describes how a variety of metaphors can cause emotional reactions as far-ranging as fear, confidence, and love. Marketing specialists, of course, discovered this long ago in their quest to increase sales through powerful imagery that move target buyers to action

(Zaltman and Zaltman). The leap from "selling an idea" to "selling a product" is not a large one to make, and is a particularly useful one in understanding how the act of data sharing is being positioned by its advocates.

Metaphor can also be closely considered with actor network theory, already introduced above as a valuable conceptual tool for focusing on the interaction between scientists and technologists and their physical environment, tools, and material objects. In *Metaphors We Live By,* Lakoff and Johnson note the close connection between the language we use and our physical being/form, linking the metaphors we use to our physical experience as human beings. This type of analysis asks us to take a closer look at the language used by researchers to reveal how our physical interactions with objects and other people shape how we think about and use them in our work. This role of metaphor has specific application to the problem and positioning of data sharing.

Different metaphors are used by different people in different contexts to describe the data being shared and those agents sharing it. This research will demonstrate that new technologies and new subspecialties in science are currently evolving that change the way that data is perceived and treated within and across the scientific life cycle of discovery and reporting. Metaphor helps construct understanding; how metaphors are constructed and exchanged helps us see both the process and product of that evolution in understanding. In this context, while metaphors are generative about technical understanding and knowledge; they are also serving to generate motivation and emotions about data, its ownership, and its exchange.

The analytical tool of metaphor is such an important one in understanding the evolving activity of data sharing in cancer research that it is dedicated its own chapter in this work.

1.4 Research Methods and Population

Research about data sharing is broad in its potential applicability; however, the research here focuses tightly on the National Cancer Institute (NCI) and the caBIG program case study described above for a number of reasons. First, data sharing across academic research institutions has been acknowledged as a vital element of the success of the program, as the benefits of shared tools and extendable infrastructure are even further maximized when people share data over them. Second, an established workgroup is chartered specifically to address the regulatory and proprietary issues associated with data sharing. This provides both a "laboratory" environment for conducting research, and access to individual researchers at cancer centers who both contribute to the public discussion and have motives and emotions of their own related to data sharing. Third, this program is an open access government program, where most meetings where data sharing issues are discussed are open to the public, and therefore are accessible for research. Finally, the program provides access to a variety of researchers from different kinds of organizations with potentially different perspectives related to data sharing. This diversity, still rooted within the domain of cancer research, facilitates both research control and flexibility.

Research conducted for this work lies at two levels: the social level of information, including public presentations, documents, articles, and open meetings; and the individual reports of personal belief and decision-making, explored during private and anonymous

interviews. Social level research included literature reviews, and attendance at public meetings and teleconferences related to data sharing. This research activity included observing group discussions related to data sharing, within the organizational construct of caBIG. These included face-to-face meetings where data sharing was discussed, and monthly teleconferences addressing the topic, attended over a time of nine months.

Individual research centered on hour-long interviews, conducted with professionals associated with the caBIG program and the broader field of cancer research. The list of interview candidates was originally developed from public sources, including caBIG program membership, meeting minutes, rosters for public meetings, and personal knowledge of people in the community.

Participants 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. From a demographic perspective, 42 individuals participated in the study. In terms of gender, 43% of interviewees for this research were men and 57% were women; ages ranged from approximately the late 20's through the early 60's, with an estimated even distribution across this spectrum.

The following table illustrates the distribution of interviewees in terms of both professional emphasis and organizational affiliation. As the table shows, many interviewees were affiliated with an NCI-designated Cancer Center; these Centers are housed at academic institutions across the country and receive direct funding from NCI to support cancer research

and clinical trials. Other interviewees represented the NCI, the caBIG program team, commercial organizations, and non-profit/other governmental organizations.

Table 2: Category Descriptions of Interviewees

Type of Organization and Profession	Total Count	Men	Women
Interviewees at NCI Cancer Centers (Total: 24)			
Principal Investigators/Scientific Researchers	9	3	6
Bioinformaticists	8	5	3
Technologists	3	1	2
Legal-Regulatory Specialists	4	1	3
Interviewees with NCI/caBIG Team (Total: 12)			
Leaders/Managers/NCI PM's	5	2	3
Researchers/Legal Specialists	7	2	5
Interviewees at Commercial Organizations (Total: 4)			
Researchers	1	1	0
Leaders/Entrepreneurs	3	3	0
Other (Non-Profit Consortium, Other Government) (**Total: 2)**	2	0	2
TOTAL	42	18	24

Structured interviews lasted between 30 and 60 minutes, and followed a three part interview protocol as follows:

- A demographic and historical component to gather organizational position/role, domain of interest (type of cancer research), reported personal experience with the

act of data sharing, and proximity with others that have had data sharing experience.

- A qualitative segment to explore the participant's views on data sharing. This was a structured interview that built upon the demographic and historical information, and moved to either perceptions about the social arguments related to data sharing, or personal emotions on the topic, depending on the path taken by the participant.
- A more structured assessment exercise to ask the participant to identify which of several factors he or she found more compelling from a structured list of arguments supporting data sharing, and arguments that cite factors that hinder data sharing.

Most of the researchers, scientists, and bioinformaticists have personally faced the decision to share data; others are in roles that impact the practice or policies of data sharing, but have not faced that decision themselves. From a disclosure perspective, it is important to note that this research was conducted from the perspective of a participant observer; I have been a member of the caBIG program team as a part-time consultant since 2006. This allowed for unique access to the research population and to open access (public) materials that support the caBIG initiative.

1.5 Research Boundaries

This research project centers on a specific segment of a broad network involved in the data sharing debate: scientific, technology, and legal professionals involved in the practice of cancer research. Furthermore, given the centrality of data sharing to its mission, the caBIG program was used a central starting point for identifying study participants. While diverse in

membership, caBIG as a large program represents a distinct social world that people select into; this work was centered in this social world (Strauss 119-28). While this removes the idea of a more "random" sample across the cancer research enterprise, with over 800 active participants, the caBIG program is considered broad enough to be considered sufficiently representative, as the social world of caBIG includes a broad range of people from a variety of other communities and social worlds, each engaged in research activities that support the mega-community of cancer research.

This research design choice, however, forced the bracketing off of an important group of stakeholders in the cancer research data sharing landscape: cancer patients who choose or not choose to share their tissues for research use, and the patient advocates that act as their voices in the policy process. As this research will demonstrate, biospecimens are considered one of the most valuable types of data in cancer research; patient consent is therefore one of the first decisions that make other data sharing decisions possible. Despite this importance, the decision was made early to focus this research on an internalist view of science: decisions related to data sharing made by the scientists and technologists engaged in cancer research. This focus does not deny the importance of the initial patient decision; it simply reflects a research emphasis that is oriented toward a different element of the cancer research life cycle.

Another research area outside the boundaries of this inquiry is the intellectual problem of defining emotions, and differentiating them from feelings, intuition, motivation and cognitive processes. The delineation between emotion, motivation and cognition is an area of great debate, and was outside the scope of this project. As noted above, this is a pragmatic and functional project, not a philosophical one; it is focused on decision and action (to share or not

to share), and how that decision or action is shaped and justified by both individual and social level arguments.

Finally, an area not addressed in this dissertation is the gendered aspects of data sharing and cancer research at a social level. Considerations of gender dynamics are social level and class level issues that are present in the social world of caBIG, but are not addressed in this research due to its scope and the composition of the participant group. This was an unfortunately, but necessary choice. Ultimately, this is a piece about understanding the intersection between personal and social action. While this intersection is influenced by class-level gender differences, as well as other classes of difference, there is as much variability in motivation within a gender category as between them. Understanding the gendered dynamics of data sharing is an area quite attractive for future research.

1.6 Summary of Chapter Contents

The promise and problems of data sharing are captured in articles and discourse that cross multiple disciplines. “Sharing” may be an ideal learned at a young age, but despite the simplicity of the term, the difficulty of the activity in practice in a range of areas is acknowledged in the cancer research community facing the challenge.

The caBIG program has been clear that data sharing is an overarching goal since the start of the program, establishing early a workgroup called the Data Sharing and Intellectual Capital (DSIC) Strategic Level Workgroup to explore the issues involved. Workgroup leaders often describe these issues in terms of the layers of an onion; each layer must be sequentially

peeled off to remove the barriers to data sharing, and achieve the goals of collaborative research. The four chapters following this introduction consider each of these layers.

Chapter 2 considers the **Technology** aspects of data sharing, reflecting that the social understanding is that data sharing is currently hindered by a deficit in shared tools and standards for managing the data generated by caBIG. This makes data sharing difficult, as even if one wants to share data; incompatible systems complicate both understanding and exchange. Chapter 3 turns to the **Economic** aspects of data sharing, considering issues such as the current institutional reward systems that impact data sharing and intellectual property rights that play a role in either limiting or encouraging data sharing. Next are the **Legal and Regulatory** aspects of data sharing, considered in Chapter 4. In cancer research, like any biomedical field involving humans, patient privacy laws and Institutional Review Board (IRB) guidelines related to human subject research can hinder the sharing of patient derived data. This chapter will consider how legal factors impact data sharing, and where the decision lies to do so.

In essence, Chapters 2-4 focus on social-level factors and provide a broad picture of the network within which individuals navigate and make individual decisions. Chapter 5 then turns to the **Personal** stories of data sharing – considering the personal side of what are named socio-cultural issues on the caBIG program. While these stories are referenced throughout the previous chapters, this section focuses on considering the deepest reasons why people report wanting to share or not share data – what are the most compelling reasons at an individual level?

Chapters 6 and 7 change the focus of the analytical framework, turning to an application of two conceptual tools that assist in understanding and influencing data sharing motives.

Chapter 6 focuses on the role of **metaphor** in the shaping of perceptions related to data sharing. What images are used to frame data sharing decisions, and how do these differ between those with different interests? Chapter 7 specifically focuses on **reversal theory**, a structural tool for understanding motivation and emotion at both the individual and social level. These two chapters act both a synthesis of the analyses from the previous chapters, and as a proposal for how this motivational model could be used to help shape future discourse related to data sharing with different audiences. Chapter 8, the final chapter, summarizes recommendations resulting from the research; and suggests future paths for both research and advocacy.

The overarching goal of this research is to detect systematic patterns of motivation and emotion in activities and language that influence the personal decisions and social norms of scientific research. By "detecting" motivational states in justification statements, objections, and metaphor at both personal and social levels, we may be able to suggest new arguments that trigger alternative motivational states and different actions. This research was conducted by a participant observer in the caBIG program, and the work is overtly values-driven. It is intended to have an activist bent, not towards sharing or not sharing, but towards the increased ability to describe a versatile "emotional toolbox" – reframing emotional resources as valuable sources of influence and power, rather than as distasteful "baggage" to be left at the door.

Chapter 2: The Technology of Data Sharing

The caBIG program is managed and administered by the NCI's Center for Bioinformatics (NCICB), which reports its mission as follows:

> NCICB plays a lead role in bioinformatics and information technology within the National Cancer Institute and serves as a focal point for cancer research informatics planning worldwide. NCICB's distinctive open access, standards-based technical approach is coupled with a firm commitment to collaboration across disciplines, institutions, and sectors. The Center spearheads critical public-private partnerships to develop and disseminate informatics for managing, analyzing, and sharing the wealth of information generated in the fight against cancer (National Cancer Institute Center for Bioinformatics, "About NCICB," Par. 1).

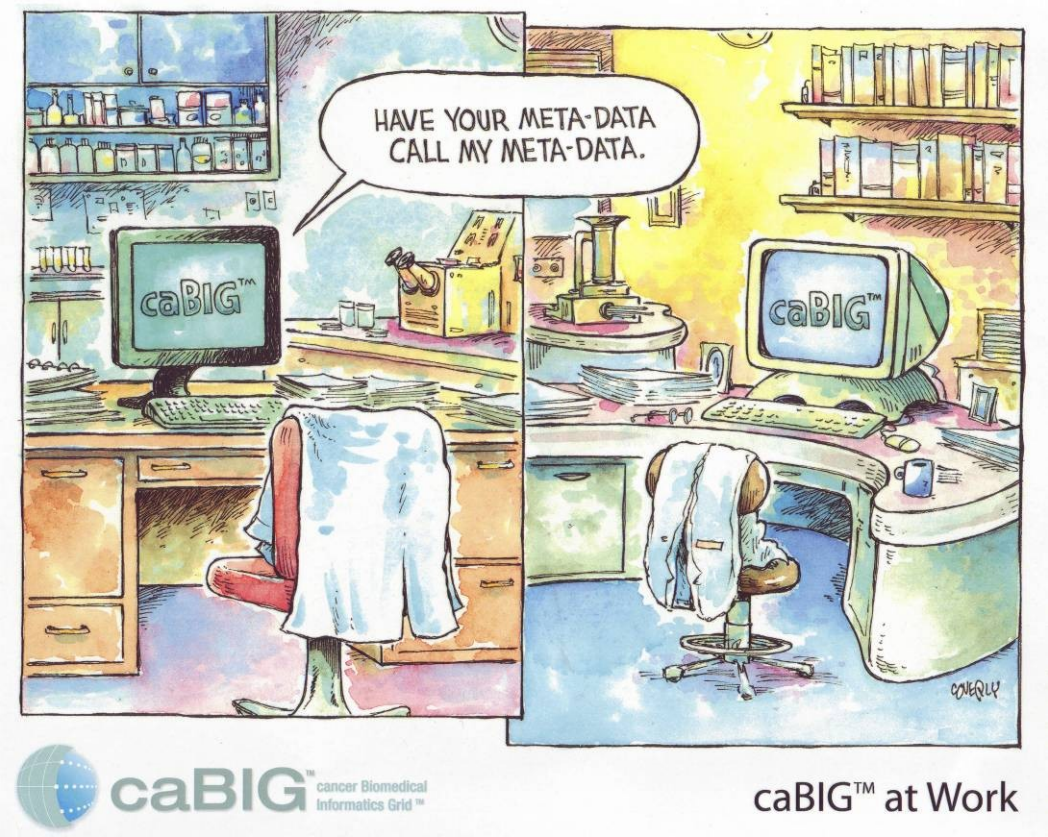
This is a technological mission, coupled with a "commitment to collaboration" across multiple entities. There are both interests and assumptions that underlie this. The implication is that data sharing and collaboration is possible because of the underlying caBIG-led technology: that *informatics* is the platform for managing, analyzing, and sharing information. Informatics, framed this way, precedes sharing; one must have the technology solutions in place in order to share data.

A cartoon commissioned by the caBIG program in 2006 below captures this core tenet. People are absent in this graphic; this is not a vision about connecting people, or about individuals sharing with other individuals. Rather, in the caBIG vision, data is shared across information technology networks or a "grid", where "meta-data" (defined as "data about data; or, definitional data that provides information about other data") facilitate the discovery and exchange of data across computer networks. Once data is formatted to meet shared standards, others can access them. This is explained in an article about caBIG technology:

> At the heart of the caBIG approach … is a Grid middleware infrastructure, called caGrid… caGrid is a model-driven and service-oriented architecture that synthesizes and extends a number of technologies to provide a standardized framework for the advertising, discovery, and invocation of data and analytical resources (Saltz et al 1910).

This technology is represented more simply and humorously in the cartoon; yet, the cartoon is also clear in its implications for a broader audience than might read about Grid technologies: technology is driving the "work of caBIG."

Figure 2: caBIG at Work[5]



This is a clear example of a project being defined by the interests behind it. The caBIG program is housed within an organization with an informatics mission; the problem of data

[5] This graphic was created by the caBIG program in 2006 and has not been formally published. It is used here as a U.S government work.

sharing is generally defined and expressed as first (though certainly not only) a technological and institutional infrastructure need and problem. This chapter unpacks this messaging, and considers how this interplays with the personal motives involved in data sharing.

2.1 Establishing the Technical Imperative of Data Sharing

Amidst all of the controversies about data sharing, the social discourse shaped by caBIG and the NCI rests on a fundamental assumption: caBIG-driven technology and infrastructure is a key enabler of data sharing; it is upon this technological imperative that other controversies play out. In the social world of informatics, grid technology is the stage upon which a revolution in cancer research is occurring, not a tool itself to be questioned. Published articles about caBIG in media that cross the boundaries of biology and technology establish the setting (Italics added).

> From *Bio-IT World*, "While the software has been crafted in response to the needs of clinical researchers in oncology, it is by and large generic…The idea is to "surround" electronic data capture (EDC) systems with management tools specific to sites and patients… It is also to demonstrate that disparate groups can *'share data in a reliable, trustworthy way'* across trials so as to spot important disease patterns" (Borfitz, Par. 2-3) .
>
> From a conference abstract at the IEEE 27th Annual International Conference of the IEEE-Engineering in Medicine and Biology Society, "caBIG is developing new *software and modifying existing software* within Clinical Trials Management Systems, Tissue Banks and Pathology Tools and Integrated Cancer Research *tools to manage the huge volume of data being generated and to facilitate collaboration across the broad spectrum of cancer research*" (Fenstermacher 743).
>
> An article in the *Scientist* quotes the NCI leader who conceptualized caBIG: "One goal we have here at NCI is to connect, by the year 2010, all NCI comprehensive and community cancer centers in the United States. *The data collected at each center will be shared (as appropriate), and all multicenter clinical cancer trials will connect to each other electronically* and to the Food and Drug

Administration" (Buetow, Par. 23).

These articles work to forge a link between software and the trust needed to share data. This position reflects a social process of framing "trust" as a technological issue; in fact, one of the central tools developed by the caBIG Data Sharing workgroup is called a "trust fabric," and contains a broad array of decision-support tools that are technologically-based – with security considerations, and plans for "click-through" agreements for sharing data in online environments (caBIG, "Data Sharing and Decision Framework").

The assumption that grid technologies and institutional-level interoperable infrastructure would serve as an underlying enabler of data sharing is not simply an interest-based positioning ploy. There is logic to the idea that the technological tools and infrastructure that are available today can be leveraged for cancer research. The key, however, is to consider how this positioning shapes the data sharing debate at the social level, and how it impacts which questions are asked. The assumption that caBIG technology forms the foundation for cancer research data sharing, in fact, removes two vital questions from the table.

The first question removed relates to whether people want or need to share data in the first place. By establishing caBIG technology as the *how*, the question of *whether* is made invisible. The second question relates to whether the technology tools and grid solutions provided by caBIG actually build on the existing ways in which people *in practice* actually share data in today's environment. With caBIG presented as the solution that establishes a basis for trust and collaboration, the question of whether that solution fits the actual need becomes veiled. In a world of diverse technological possibilities, caBIG frames the very question.

Another way in which caBIG tightly couples the ideas of grid technology and sharing is through the software that the program advocates for its participants (those who help develop the tools, infrastructure and policies), and those who consume caBIG products (generally researchers and clinicians). The caBIG project is an "open source, open access community," which means that software can be downloaded and installed, generally for free.[6] Open source software is generally developed with the mindset of transparency and open distribution:

> Open source is a development method for software that harnesses the power of distributed peer review and transparency of process. The promise of open source is better quality, higher reliability, more flexibility, lower cost, and an end to predatory vendor lock-in…
>
> The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed (Coar, Par. 2).

Open source development is generally positioned as an alternative to proprietary software, thereby helping to shape the caBIG social message. In addition to the practical aspect that it is generally technologically easier to share data if people are using common software tools available for free download through which that data can flow, there is also a role model element: caBIG is acting as an example for the sharing that it wishes to shape.

[6] Those close to caBIG will note that while most of the software available through caBIG is open source, some of it depends on non-open source dependant software to operate, and it requires a financial investment in personnel to deploy the software and maintain it. This means that while the software itself can be downloaded for free, it does involve a financial outlay both as an initial investment and for ongoing support. While this research does not intend to ignore this problem, given that these investments are already being made in other areas, it is not a central factor impacting social messaging related to data sharing, which is the focus here.

How do those in the cancer research community see this? Are institutional informatics solutions accepted as the inevitable foundation that will lead to data sharing? This, too, seems shaped by interests. The caBIG program funds Cancer Center representatives from across the United States to participate in the development and deployment of caBIG. Many of these professionals come from bioinformatics departments, and have a professional interest in deploying tools across their organizations that support the new paradigm of technology-driven personalized medicine. We would expect, perhaps, that people in this techno-scientific role would accept the baseline assumption that informatics is the platform and condition upon which data sharing will occur. Others are more skeptical. Those with an interest in bioinformatics tools tend to agree that this is the direction and need of the future; those from historically less-technologically driven research areas are more likely to question whether these tools and sharing are the only possible path forward. One neuroscientist working in the field for more than 30 years noted (interview question in italics):

> There was a day when there was much more a sense of complete ownership of the data, by the person who generated the data, in terms of lab notebooks; it was part of the culture of the time. It was really before the technology existed in the kinds of ways to share data that exist now… *What caused it to be different? How did you see it play out?* Well, I think the biggest thing was the development of the Internet. I think that's #1. We didn't have very efficient ways to connect individuals to resources any way before, you know, the mid-1980's. And that shifted everything and made it possible for us to know what was happening in other institutions, without having to call them up and talk to them about it.

This comment was not made with any sense of regret about the "good days" of ownership gone; it was a description rather of how she sees the "strands of science" changing over time, and the potential that change provides. From another experienced researcher working in clinical studies (interview question in italics):

> The idea that everything just flows and changes, it's much more prevalent than it was in the 70's. *When did the shift happen?* I think the biggest single factor was the internet; way back when Mosaic came out, and you had the ability to communicate on the fly with another investigator and send them information; it suddenly opened the door to a more loose and quick collaboration. Suddenly, it didn't need to be done by snail mail, where I had to print it out, compile it, put it in the mail.

From a younger scientist: "Before, there was no Google. We weren't really in the digital era. Right now, there is more need to know what other people are working on." In this final case, the language is particularly striking: it is the availability of technology that is driving the need: there is Google, therefore, there is need.

There is a nuance here, however. These researchers, and others like them, are talking about the Internet as a mechanism for sharing data with others, which may or may not include the more specialized mechanism of informatics tools and grid infrastructure like those provided by caBIG. While the Internet changed the dynamic, many interviewed working outside the job role of bioinformatics do not see this change as a matter of inevitable technological momentum, with greater informatics capabilities and shared institutional infrastructure leading to greater sharing.

Another scientist with a strong background in software development also agrees that while the Internet makes data transmission easier, that doesn't inevitably lead to more sharing. In fact, the social messages associated with a for-profit technology organization have, from his perspective, discouraged sharing (interview question in italics):

> *Question: "Wait, I hear a conflict – you said that old-schoolers were less likely to share, but now you are talking about resistance among younger scientists too. Where's that coming from?"* Response: "Microsoft. Everyone looked at Bill Gates becoming the richest man on Earth by restricting access to his technology,

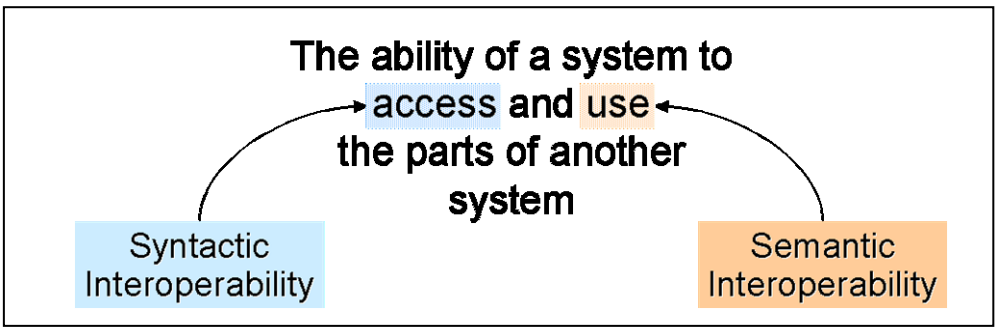
> and said, well, this is the way to do it… Before that, it wasn't…. Now, there's an even younger generation that says, you know, you get more from open development, from open source, Linux. That's a matter of open warfare – Linux, Microsoft - right now. Mainframe guys shared their stuff pretty openly – the guys on Unix originally.
>
> Look, there's always been people who are like, mine mine, mine, and there have always been people who are like, hey, let's share it. The business argument for sharing has become stronger in recent years. You look around at younger software guys and gals – all this Web 2.0 stuff – it's not just software, it's just their personal life…I see the ebb and flow on both sides. It would be absolutely incorrect to say there is any one paradigm that's completely dominant at any given time."

Consistent with these ideas, interviews suggest that bioinformatics has aided in the current interest in data sharing, but that grid-based data sharing is by no means a matter of technological momentum. This is important when comparing personal perceptions to social discourse, because of the interests that drive institutional level investments related to data sharing. A great deal of interest in grid-mediated data sharing comes from departments that have interest in developing technology solutions, e.g., biomedical informatics, and driving them across the organization. Those that have technological interests are creating demand for data sharing, by building the tools that both generate high volumes of data, and allow for the sharing of it. Interests frame action – and in the world of "Heath IT," it is not a coincidence that Health is the modifier and IT the noun. While caBIG leaders caution that there is misguided faith in the belief that "if you build it, they will come," the social discourse continues to message that institutionally-based technological investments are what will ultimately drive the scientific movement towards increased data sharing.

2.2 A Post-Modern Modernity: Federated, Standards-Based Data Sharing

A key element of the caBIG technology-based solution that helps establish an environment for data sharing is an emphasis on the shared standards that lead to *interoperability*. The following graphic is frequently used in caBIG presentations to define and state the importance of interoperability in facilitating data sharing.

Figure 3: Defining Interoperability[7]



The context behind the definition is that, in order for data sharing to occur in a truly technologically mediated way, one technology system must be able to "understand" the data in another system; both must be using a common vocabulary and similar data elements (e.g., both have the same kinds of data, and they must be able to be exchanged and interpreted correctly by each other). Without this ability, there are data sets that are exchanged, but which cannot be used together without human intervention to align the *meaning* of the data from the two systems with each other. The social message: it is not enough just to share data; they must be formatted in a way that they "understand" each other, which requires some tools and upfront labor to ensure that the data sets are aligned to the same standards. A bioinformatics journal

[7] This graphic is used in a variety of caBIG presentations. One of these is "caBIG Essentials Training Program, Lesson 1" accessible at <https://cabig.nci.nih.gov/concepts/essentials> (November 2008). It is used here as a U.S government work.

editorial quotes the caBIG Director on why this is so critical to data sharing from NCI's perspective:

> 'The lack of interoperability is now a real barrier,' Buetow says. 'The lack of this interoperability is now slowing them (cancer researchers) down.' He speaks of 'cumbersome, disconnected processes, not to mention the inevitable shift from paper to electronic data, and it's clear that he believes that technology can make a difference in fighting cancer.
>
> 'For this virtual community to succeed it was important to embrace the individual diversity of members and to connect them, as opposed to creating one big central resource where everyone needed to place their information. As such, caBIG focused on providing tools and infrastructure that could be run by individual laboratories, organizations, or institutions and connect electronically through the Internet. This strategy is called standards-based interoperability, and caBIG has realized it through a services-oriented architecture called caGrid. It is worth noting that caBIG adopted international standards where they existed and extended them as needed to address new problems.' (Editorial Staff, "Forward-Looking Systems at NCI," Par. 6)

This positioning recognizes the decentralization and independence of cancer researchers; the goal is *not* to build a giant centralized database into which data is deposited. Rather, the goal is what is called a "federated" model where individual investigators hold their own data, but format them according to common standards, and then make them available to others so that they can be shared. Agreeing to those standards is a method for building the community consensus that is needed for true interoperability and collaboration. This approach has precedence in the technological community, as standards are widely understood as the foundation for internet protocols. Given that caBIG is both conceptualized and positioned as a world wide web of cancer research, the extension of the perceived importance of standards is understandable. The same editorial as quoted above (2008) notes:

> "A geneticist by training, Buetow and the NCI probably face a governance challenge that is as gnarly as any in the land. Somehow he corralled the

> innumerable egos and agendas of federal and university-based scientists. "People can put down their proprietary and competitive forces," Buetow says. *"The one place we can get universal agreement in our community is around the structure and definition about how we collect our information and our applications can talk to each other"* (Editorial Staff, "Forward-Looking Systems at NCI," Par. 7).

It is not just caBIG that sees standards as vital to encouraging data sharing. A 2005 article in the *Economist Technology Quarterly* about biobanking (referred to also above as a biospecimen repository) expresses as a key conclusion:

> If the full potential of biobanks is to be met, there will have to be a standard way for researchers to order and access samples from databanks – just as the Internet's common standards facilitated the free flow of information across digital networks, and made available previously untapped data. Only then will be it be possible to fully exploit the mountains of samples, and reams of data, that are currently locked up in the world's hospitals, clinics and laboratories (Editorial Staff, "Report: Medicine's New Central Bankers" 19).

This idea is also captured in a 2004 *Science* article proposing an international framework to promote access to data. "Technical and semantic interoperability of databases" is listed as one of eight "operating principles for data access regimes," noting specifically that ensuring this interoperability takes both time and resources:

> Establishing and maintaining this infrastructure requires continued and dedicated budgetary planning, with appropriate financial support. The use of research data cannot be maximized if access, management, and preservation costs (including cost of documentation and metadata creation) are an afterthought or are insufficiently or inconsistently funded in research projects (Arzberger et al 1777).

How do researchers themselves see this? It is a complex answer. Many agree that interoperability is absolutely vital to support large-scale data sharing; but also do not see reaching agreement on standards truly a baseline dependency for the kinds of data sharing

they are currently doing. For many professionals, person-to-person sharing, which involves simply transferring a file in its existing format, is the norm. Some noted that the added labor required to format data to match standards for interoperability is often used as an excuse to "opt out" of sharing data.

Much of this seems to be a matter of the scale of sharing being considered. As noted above, caBIG tends to position the goal and benefits of large-scale, institutional data sharing using interoperable data and tools. Data sharing is, in reality, far more often conducted as a point-to-point or specialized community-based activity, which isn't perceived to require the same interoperable rigor. For many, the "caBIG way" (standardized and interoperable data sets and tools exchanging information across grids) is not the way that people in practice are either ready to share, or are actually sharing, data. The exception is people working on specific consortium-based or cross-institution projects, where there is an internal community or network of data sharing, supported by a technical infrastructure that supports the very specific sharing needed by the group. Even here, the sharing is specific, not seen as large scale and anonymous, but rather quite targeted in nature.

Responding to the list of research hypothesis that "technical difficulties and needs for standards is a hindrance to data sharing," interviewees respond:

> There is some technical difficulty, but I don't think that's the main issue. If people really want to share data, they will figure out a way to share information with each other.
>
> It's complex, but it's not too complex. I think that's a scapegoat answer. [Quoting other scientists] 'I don't really want to share my data, so I am going to say I can't do it because it's too hard.'
>
> Technically, we [referring to caBIG] aren't there yet. We are still in a situation where it would be easier to call someone up and say, I want your data set; what

> form is it in? And just have them stick it on a server and FTP [file transfer protocol] it down, rather than ship it around using caBIG.
>
> You know, people are busy. If you don't give them an easy way to do things, they just don't think they can. I don't want to bother; the technology is complicated, I'm busy. Unless they have a really strong motivation, like they are partnering with someone on a project, on a grant, on a drug trial - then you work through the issues, because you are focused on that outcome.

This emphasis on the outcome, or reward, being a critical motivator to data sharing will be addressed in Chapter 3; those managing large data sets are quick to note the pragmatic variable of cost in meeting standards for interoperability encouraged by caBIG and similar programs. It may not be as technically difficult as people believe, but it does require labor, resources, and specific skill sets to prepare data in the format needed to share on a large scale; and right now, many don't yet see the incentive to do so, and/or do not have the skills themselves, or access to the people with the bioinformatics tools and skills to do so.

This discussion with interviewees also raised the point that the standardization of data can 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. In addition, each data element may have its own history and context that can alter its meaning and applicability to different settings. Forcing a data set to conform to a set of standards causes the loss of this unique context, and ultimately removes the subjective knowledge of the researcher in the search for the objectivity of a standardized data set. The implications of this are discussed further in Chapter 5.

Many interviewees did concede that that is an evolving process; the Internet established a setting in which data could be more easily shared; and interoperability is the ultimate goal for allowing that data to be more easily understood. This evolution, however, takes time and

resources, and a focus on the making sense of the practicalities of implementation. This includes making sure that there are professionals available with the skills needed to introduce these new capabilities, the incentives that will encourage people to use them, and a clear understanding of the scientific problems that will benefit from the use of this technology.

A conclusion here is that while interoperability is positioned by the caBIG program at the social level as a key tool to facilitate data sharing, there is a lack of matching positive emotional energy at the individual level. What is seen as a path to community building on one level is seen generally as either irrelevant to the point-to-point sharing that they actually do; or worse, as a perceived or real deterrent and nuisance. The social messaging of "standards development as path to community building" is generally not seen as a motivator at a personal level. As a researcher, I may be motivated to share, but the way in which I do may be sending files to someone directly in the formats they are already in, rather than standardizing and then sending them through the grid network of caBIG. Scientists generally see the technical problems differently than how it is positioned caBIG; the personal stories are different than the social ones.

From an STS perspective, there is a natural tension embedded within the caBIG quest for interoperability in a decentralized environment. This is a balance between the modernist drive towards standards and uniformity; and the more post-modern drive towards diversity, individualism, and local knowledge systems. This is a balance and tension that deserves further discussion. Modernism and postmodernism are often proposed as two somewhat opposing constructs with which to conceptualize science and technology. Modernism reflects a quest for institutional consistency and control, reflecting a drive towards the efficient management and

leveraging of capital and the economies possible with large scale approaches. There is a certainty and predictability that comes with modernism structure; at its best, it provides a common platform across diverse interests, and the comfort of a "universality" that can be translated into a broad application of institutional practice. Postmodernism, conversely, reflects a quest to "return to the local," to recapture the diversity of approaches and knowledge residing in local systems. Postmodernism questions the certainty of centralized institutional approaches, calling instead for the embracing of complexity rather than the ordering of it.

In *The Condition of Postmodernity,* David Harvey proposes that the difference between these constructs, however, is far more fluid, with the compression of space and time leading to what appears to be post-modernism diversity against a modernist backdrop. This is a useful contrast for considering the path advocated by caBIG. The current decentralized cancer research landscape can be conceived as primarily a post-modern one, with researchers constructing personalized tools and research approaches in the proprietary comfort of their own laboratories. This reflects an emphasis on the local knowledge and expertise of the researcher and the team; in many cases, interviewees reported that the diversity of practice even in the same institutions results in local variability that makes data sharing between departments complex, even if one wanted to do so.

At the same time, the changes in scientific and technological tooling that permits, and in fact demands, the large scale aggregation of data in order to detect genetic variations is calling for a broader approach to data use. There is a strong argument for the efficiencies provided by the technological ability to perform large-scale data abstraction; this demands, however, the standardization of local data into forms that can be exchanged readily for others, and which

have a longer life than one researcher's experiment. Accepting David Harvey's conceptualization of the space-time compression as the motivator for post-modernism, we can in turn conceptualize caBIG as trying to take the post-modern cancer research landscape and "stretch it" over both time and space, *without losing* the speed and personality creativity offered by local control. The caBIG program wants the best of all worlds: local control of data to allow local knowledge the freedom to thrive; *and* the institutional availability of data across these local systems to allow for larger scale data use and analysis across both time and space.

The caBIG program works to balance these factors by encouraging standards within a federated model; meaning that while data need to adhere to centralized standards, the data sets remain locally held and managed. This balance appears to have led to a motivational disconnect at an individual level. From the scientist's perspective, if one is able to hold data locally (which is a clear caBIG benefit to scientists), what is the motivation to invest in formatting it (or finding someone else who can) according to centralized standards? The incentives for integrating a modernism system within current post-modern practice are unclear.

This concern is foreshadowed in Latour's critique of modernism. In *We Have Never Been Modern*, Latour argues that modernist approaches work to separate science and nature from society and self, breaking object-subject connections in the service of standardization. This process results in institutional constructs that discourage individual interpretation and hide controversies, shielding them behind the presentation of objective truth. The researcher skeptical of the caBIG emphasis on standards would appear to agree. Standardization of cancer research data loses the how and the why: how a data set was generated and for what purpose. Data is both subjective and personal; data sharing on grids removes this personal element.

The caBIG advocacy for a federated model, where individuals and institutions retain the freedom to control data locally, leads to a different set of motives than in centrally-controlled rewards-based systems, where all data is to be deposited in a central database according to a set of standards in order for a reward to be received. In the "deposit-reward" model, used by several scientific journals as a precursor for publication, the incentives and rewards of modernist interoperability are more direct and straightforward. In the caBIG case, a more diverse set of individually-mediated motivational factors appear to be at play. Ultimately, there is a motivational clash between the Mertonian and modernist-oriented drive to share data through anonymous, public biomedical grids; and the individualized and subjective post-modern orientation that appears to more typically drive the process, and which caBIG allows for in its federated model. In counter norm terms, people are particular about who they share their data with, and the communalism of an impersonal grid does not align with their personal experience.

This analysis in no way suggests or proposes that the caBIG "federated, standards-based" model be dropped or altered. Just as STS writers can fall into the trap of evaluating opposing theories as "ors" (norms or counter-norms; social or individual; post-modern or modern), it is easy to seek the clarity of a "centralized (also called aggregated) *or* federated" direction for caBIG. There are benefits of the "both and" approach, and it is one of the complexities that keeps interest in caBIG high. The quest here is to consider the motivational and emotional impacts of this complexity at the individual level, and to propose strategies for engaging with these impacts at the social level.

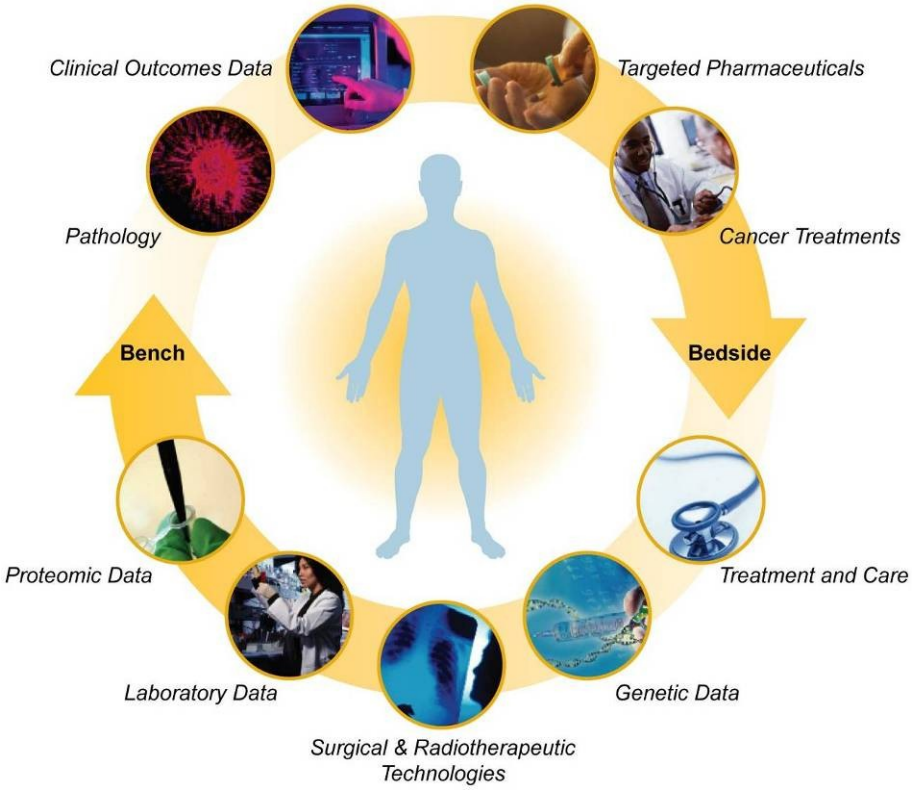
2.3 Technology: Driver for a New Model for Science?

Federated interoperability is not the only way in which caBIG is seen as working on the leading edge of technology and data sharing. There have been fundamental changes in the way biomedical research is practiced, and many see it a direct result of the introduction of today's biomedical informatics tools. First, the message from interviewees that are conceptually positively motivated to share data is that bioinformatics technology and caBIG have the potential to drive not only scientific progress in cancer research, but the very foundational ideas of the traditional retrospective "hypothesis-test" scientific method. Second, these technologies are seen as increasing the speed of scientific progress, through increased visibility into the actual work, both successful and not, of individual scientists and organizations. These changes influence how people think about data sharing, and their motivation to do so, with emotional outcomes that range from trepidation to excitement.

Describing the Science and Data: A Range of Actors and Objects

To set the context for these discussions, it is useful to introduce the kinds of science being discussed here. The caBIG program often describes itself as being a resource across the full "bench to bedside" scientific continuum. As illustrated in the following figures used frequently in caBIG outreach materials, this means that caBIG tools and infrastructure can be used to support everything from clinical research involving patients and treatment outcomes, to laboratory (or "bench") research that focuses on genetic (genomics) or proteomics, to pharmaceutical research focused on drug discovery.

Figure 4: The Bench to Bedside Cycle of Cancer Research[8]

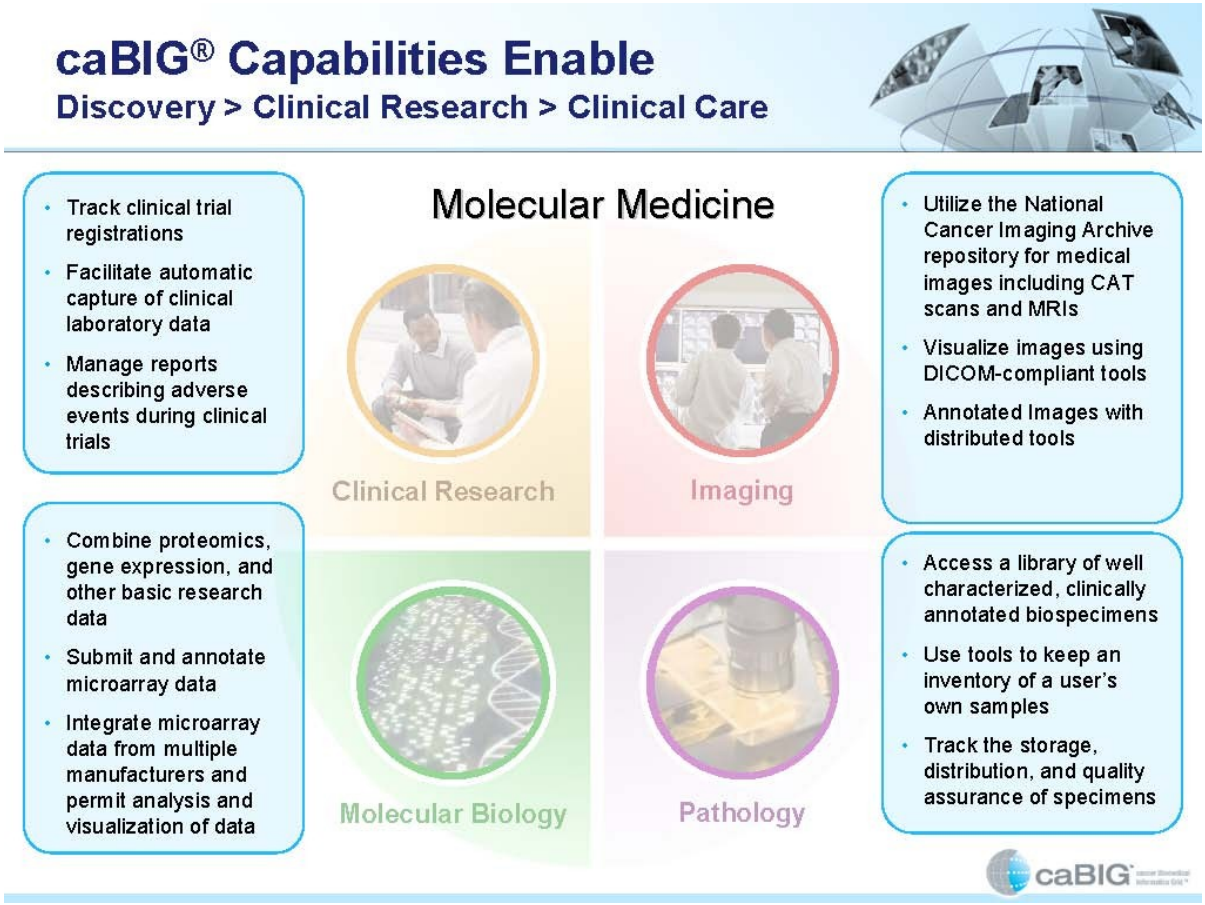


What does this mean in terms of data sharing potential? As many interviewees cautioned, that depends on what you mean by "data." The activities represented by the figure above generate a variety of kinds of activities, which caBIG generally categorizes into four capability areas: clinical research, imaging, pathology, and molecular biology. As the figure below subsequently suggests, data generated from these processes vary greatly and can include: patient records (including demographics, enrollment in clinical trials, adverse events), study protocols and methods, biospecimens and their associated descriptive information,

[8] This graphic is used in a variety of caBIG presentations. One of these was in the "caBIG 2009 Annual Meeting Newcomer's Session" accessible at <https://cabig.nci.nih.gov/2009AnnualMeeting/presentations/monday-july-20-2009/welcome-and-newcomers-session/newcomers/file_download/presentation> (July 20, 2009), available for free download. It is used here as a U.S government work.

reagents used to preserve and process biospecimens, images of tumors, cell lines and clones, genetic information, microarray data and the specialized methods used to analyze them, database structures and management techniques, trial results, and treatment outcomes.

Figure 5: From Capability to Artifacts - Data Generated from Bench to Bedside[9]



[9] This slide is from a presentation by a U.S. Government Official. Reference: Buetow, Kenneth H. "Building a 21st Century Biomedical System: The Cancer Biomedical Informatics Grid (caBIG®)." Supercomputing 2008: November 19, 2008, Austin, TX, 2008. < https://cabig.nci.nih.gov/overview/Buetow_SC08-120408.pdf> . It available for free download. The figure is used here as a U.S government work.

New Scientific Questions and Visibility

How does this all contribute to the revolution in the scientific method highlighted by interviewees? Three particularly compelling elements of this scientific change, and its impact on data sharing, are considered here:

- How research questions and analyses are created and conducted
- How the visibility of the scientific process is changing
- How negative results can be reported

A 2005 scientific article exploring new genetic research at the molecular level describes the new kinds of science made possible through biomedical informatics tools and methods:

> [Referring to breast cancer research] Until recently, evaluations of prognostic and predictive factors have considered one factor at a time or have used small panels of markers. However, with the advent of new genomic technologies such as microarrays capable of simultaneously measuring thousands of genes or gene products, we are beginning to construct molecular fingerprints of individual tumors so that accurate prognostic and predictive assessments of each cancer can be made. Clinicians might one day base clinical management on each woman's personal prognosis and predict the best individual therapies from the genetic fingerprint of each individual cancer…. The goal of comprehensive, genome-wide approaches is to identify clinically useful genetic profiles that will accurately predict the outcome of therapy and the prognosis of patients with breast cancer (Chang et al 100).

The NIH describes this evolution in scientific method at a broader level:

> With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. The tools include computerized databases that contain the reference human genome sequence, a map of human genetic variation and a set of new technologies that can quickly and accurately analyze whole-genome samples for genetic variations that contribute to the onset of a disease (National Institutes of Health, "Genome-Wide Association Studies," Par. 2).

The Human Genome Project was referenced many times in the course of interviews as a significant milestone demonstrating the success of bioinformatics as an established field. One interviewee described the impact of this project and the resources it both utilized and now provides on how scientists approach cancer research and the current need for data sharing (interview questions in italics):

> …the development of high throughput technologies made it possible to generate data which had value beyond individual experiments. Much more descriptive data, you can describe an entire human genome, or whatever, and that had immediate global interest…. It used to really be that investigator-initiated hypothesis-driven research was the be all and end all. But, I think there's increasingly an interest in the mining of data, which is a very different approach. It is much more discovery oriented, and that requires much larger amounts of data. It is more retrospective, it is about connecting individual data sets into much larger aggregates that can provide and support for that kind of data mining approaches, which require a lot of data. Those are the strands that have led us to where we are….
>
> I do feel that we, in some ways, we came up against a wall with reductionist hypothesis driven research, where we are looking at individual factors, and we sorted out a lot of, you know, human disease issues that could be sorted out in that way. Now, we are in the age where we're needing technologies and resources and information that will let us address and attack much much harder problems in biomedicine, and there are going to problems where there are a lot of interdependence between factors, and a lot more factors to sort out. That, I think, requires a very different approach, and it will be difficult to do that without sharing data, I suspect.
>
> *How is this change in science communicated?* I don't think it is explicitly communicated. It's a theme that runs through everything I see. All the way from the kinds of grant requests and announcements that go out, to individual people's conversations, in the sense that there is a kind of a snowball effect, if you will. Once you begin to get larger aggregates of data available through different resources, then things become possible that were not possible before. People can talk about things being possible, that they couldn't really articulate or really dream of before. And that shift then propagates. It's not like someone says, "Well, we are not going to focus exclusively on hypothesis research, we're going to focus on discovery processes," but the conversations have changed a lot. It's a chicken-egg problem, I'm not sure which came first, but it's feeding each other. It happens everywhere; at the level of individual discussions, the

> kinds of approaches that people write into grants, I think you see a lot of people making claims that they used discovery approaches because the data is there, that just wouldn't have been possible before.
>
> *So what happens that prevents that sharing from happening?* There is still a sense among some investigators that giving away your data is going to deprive you of your livelihood in some way, that you are depriving yourself of the ability to control the path of your research, and that is difficult to address, because it's very personal. There are tremendous benefits that aren't clear right away, that become clear only later. People have to have a little experience with data sharing to get over the hump of that personal sense of turf. That's personal, which is in some ways the hardest thing to deal with.

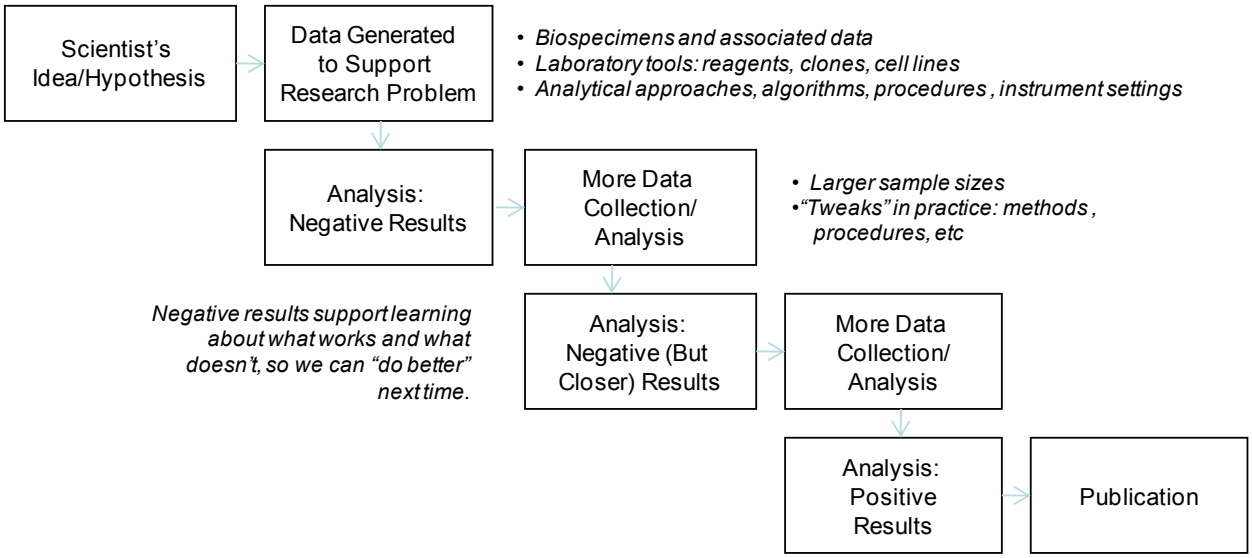
One of the areas in which this "turf concern" becomes the most evident is in the sharing of biospecimens, the original tissues from which many scientific experiments are derived. Human tissues, particularly those with certain forms of tumors, are often considered a highly valuable resource, because they contain that genetic material with which, and based on which, much analyses can occur. Discovery-based investigation requires a lot of data, which in turn, may require a lot of samples. Much of the controversy about data sharing, therefore, can be seen at the level of this artifact. Because of the value question raised here, this aspect is considered more deeply in Chapter 3, the Economics of Data Sharing.

This change in science, from hypothesis to discovery, leads to the second reality in the shift in the scientific process that is offered by caBIG and programs like it. The technologies and infrastructure provided by caBIG allow for a level of visibility into the scientific process that was not possible before; and with it, potential visibility into the objects that scientists own and generate during the scientific process, but may not wish to share with others. Science, across many specialized fields, is an iterative process. In the case of cancer research, interviewees described a process of discovery that follows the generalized conceptual path shown in the

following figure. As the flow notes, there are a number of intellectual and physical objects, artifacts, or data that could be shared anywhere along this path.

- The idea or hypothesis itself (the research problem being worked on)
- The data generated to support the study, which could include human tissue and associated data, reagents and laboratory materials
- The analytical method being used, which could include analytical procedures and methods, or even instrument settings and practices
- Outputs from instruments that come in the form of spreadsheets or graphical outputs
- Results that fail to support the original idea, but produce learning through the attainment of negative results
- Results that support the original idea; and results that point to a new idea

Figure 6: Choosing to Share Data – What and at What Point of Discovery?



The potential to share a variety of forms of data through the tools of caBIG ultimately sets up an environment where others could potentially peer into this messiness of laboratory

life, putting new visibility into what was previously invisible labor; and potentially creating new demand for materials that were once not open for view by others. This is new potential that was not present in the traditional less technology-driven publication model of science, where knowledge was shared by sharing final results through written publications, reviewed by a set of peers. Rather, up to the final box of activity above (publication), it was generally assumed that the *practice* of science would largely stay behind the curtains, where sharing was limited perhaps to close colleagues working on the same project. Even when considering the Mertonian norm of communalism, "gift giving" has generally been interpreted as sharing scientific outcomes through published papers.

Along these lines, in conference presentations using the slide below, the caBIG Director notes that one of the challenges of the current biomedical landscape is that science is still following a somewhat archaic "workflow" established more than 300 years ago, with publications and meetings still serving as the primary forums of scientific exchange.

Figure 7: Challenging the 17th Century Paradigm[10]

In a new technologically-mediated science offered by programs like caBIG, data sharing becomes more possible anywhere along the scientific process path. With appropriate protections and security, clinical trials can be advertised, increasing patient accruals; the availability of tissues can be made visible through biospecimen databases connected through grids; analytical services can be made available through those same grids; and microarray data can be posted in databases at any point in the analytical process.

At what point along the continuum should data be shared then, to help facilitate this discovery-driven science? This was a significant point of controversy that emerged during the interviews. While many support the *ideal* that data sharing helps the advancement of science on a conceptual level, when and whether to do so it is another question. It is at this point where some of the personal resistance and reluctance to sharing data becomes evident. It is

[10] This slide is from a presentation by a U.S. Government Official. Reference: Buetow, Kenneth H. "Building a 21st Century Biomedical System: The Cancer Biomedical Informatics Grid (caBIG®)." Supercomputing 2008: November 19, 2008, Austin, TX, 2008. < https://cabig.nci.nih.gov/overview/Buetow_SC08-120408.pdf> Note the metaphor of "information islands" embedded in both the text and graphic, assumed to frame the tradition as a physically isolated, even lonely, process. The role of metaphor in communicating about data sharing is the focus of Chapter 6. It available for free download. The figure is used here as a U.S government work.

also the point at which the questions about data sharing become more refined, into not whether data should be shared, but *when.* Even in genomics, which was described by many as one of the more open fields in cancer research, the timing of data sharing is important. Says one researcher (interview question in italics):

> I don't think you are going to see a lot of people putting their cutting edge data right off their microarray array analysis right out on the web to share. *Why not?* Mostly because they are first and foremost wanting to see what they can find in the data – is there something good in there? Something important? Sometimes too, it is a question of, is the data accurate? You don't want to put stuff out there that will come back and bite you, maybe you weren't doing this quite right, maybe your procedures weren't the best, maybe your method of gathering or analyzing the data wasn't good. You don't want to put out bad data.

Another scientist notes the disincentives to sharing data pre-publication:

> With some exceptions, in the academic world, your primary business is grinding out papers. Anything that detracts from that, you are not going to pursue… In certain areas, mainly bioinformatics areas like gene chip arrays and that kind of thing, there's a requirement from the journals at the far end, if you publish in these areas you have to make your data sets available, your gene expression information available, so that somebody else could analyze it and either confirm your result or challenge your result…. So, we'll do that to get published because it's required. It's one thing to share data after you have published on it; it's another thing to share data before you publish on it. The notion that people would be willing to assemble data and then hand it off to someone, in an academic environment, without having first published, that would be craziness for them. They're not going to do it…. Incentives in the academic community run so counter to that; it would be suicide.

Another genetics researcher agrees:

> I think these guys have to get over the idea that raw data are going to get published. Because people won't do it, and I don't see a paradigm shift coming easily or soon. On the other hand, if people are putting out data once its published, you can still get a lot of value from it, and quite frankly, raw data is quite suspect until it is validated and analyzed, until you do that, all you are doing is increasing the background value of crap. You do an experiment, and later, you go back, and say, that wasn't quite the right control, or the temperature in the room was off, or stellar neutrinos impacted my lab at that exact time of the experiment (*laughs)*, and something had gone quite askew. So,

> you say later, this data set is outside the norm of my observations, I can figure out and explain why, and so it's outside what I'm going to use.

In general, it was clear that most scientists close to the data being proposed for sharing are not ready to share it pre-publication unless it is with a collaborator with whom there is a clear goal, or unless sharing directly benefits their project. Some of this has to do with the current reward structure of science, discussed in the next chapter; some of it has to do with the perceived messiness of the raw data before it is cleaned up in final analysis; and some of it is an assessment of the additional, and currently not planned for, labor and skills required to prepare data for sharing pre-publication.

> You don't usually clean data until you are ready to publish. It is the act of the final analysis that cleans up the data for publication, because you only have to upload the data that you are using for your publication, not the whole set.

A university bioinformatics analyst, who often supports scientists in their data management needs, reports that her clinical research department has chosen not to expose data through caBIG infrastructure.

> Another concern about sharing too early is the fear if data is shared, the floodgates will open and a lot of people will ask for the data sets. In addition to the work required to set the data up in the first place, there may be other questions: "Looks good! What else you got? What about *this?*" All of this sets up an expectation that we are just not ready to provide. You need to consider the labor requirements involved in all this.

All of these reactions further highlight the lack of technical inevitability that in some ways threatens the caBIG vision of an integrated world wide web for cancer research. While the ideals may be inspiring, the pragmatic realities appear to make data sharing more of a goal than an act of generosity.

Despite these perceived barriers, other interviewees – mostly those not as closely associated with the actual decision to share - cited other ways in which the ideals and tools of caBIG can change the process of science. For example, the increased visibility into resources, and ability to post data whether it has been published in a paper or not, allows for an increased opportunity to share negative results, captured in this context by one interviewee as, "things we tried that didn't work, so we'll not likely ever publish it." This comment captures again the reality that the current paradigm of information dissemination lies in published papers and results, not through the sharing of the data that led to those results, or through results that were considered to be "negative" because they did not support the hypothesis posed. In general, publications favor positive results; it is rare to have negative results published. Several interviewees noted that given the volume of results that are often generated on the path to a positive result, there are *lots* of data out there that no one is likely to ever see: on laboratory hard drives, DVDs, on graduate student laptops, even in the memory of instruments.

In the grid-enabled technology of caBIG, where with the right investment, interoperable data sets can be more easily physically shared with others, there is an emerging infrastructure for posting data, and negative results, should the researcher choose to do so. This means that negative results – what people have tried and failed - can potentially be more easily shared in the new environment, without the gatekeeping of the publication process. One interviewee states, "The practical benefit of data sharing is the furthering of science. You aren't treading ground that others are treading before, because people have published results, even negative results. Even negative results matter. If you aren't going anywhere, other people need to know that."

The conflict between the ideals and the pragmatic realities of data sharing show as people talk about these ideas. Whereas the norms of science would point to a dispassionate and objective analysis of one's own work, the idea of sharing what "didn't work" is seen as both needed, and somewhat threatening. On one hand, people want to know what others are working on, what others have learned and found out. On the other hand, sharing their own raw data may allow others to determine the research question they are working on: the ideas behind their own data. One researcher distinguishes the difference between sharing results (even unpublished) and sharing the data those results are based on.

> You want to let other people know what you have figured out, but not the raw data. If you present your raw data, people may be able to figure out something out of that that you didn't see. You then are the data provider for someone else, to advance on top of your results.

Another scientist told of a time when she wanted to share results (not data) data as a PhD student, but was prevented from doing so by her advisor:

> When I was in graduate school, I had come up with results that contradicted a previous publication from our lab…or, it didn't contradict the results, it contradicted the interpretation, so my PI would not let me publish. And, so I spent another three or four years in grad school, and ended up publishing it anyway. He - the PI - left the university, and ended up going to run a [government office], and I went to another person's lab, and she didn't agree with how the experiments had been done. So, I had to reconstruct a lot of materials with new factors in mind. My strains before were not all the same background, so I was trying to get them all aligned. He had left, he had moved on, he wasn't working with me anymore, though he had never released me, but we sent it in to be published. But a lot of time had passed, so the work wasn't as interesting as it had been before.

Again, this example points to not only scientific discovery that was only captured in a publication late in the process, but also the complexity of the data involved in that process – the strain materials, the experiment protocols, the various results, and the various

interpretations of those results. The counter-norms of holding data proprietary seem clearly invoked in the withholding decision; and the power dynamics of a PI telling a student not to publish because of questions of interpretation suggest motives that are not well aligned with the idealized norm of "open science, sharing knowledge for the greater good."

The objections to data sharing detected in this research mirror those captured in other studies related to data sharing: fear that data sharing will reveal research errors, poor quality, or even researcher fraud; fear that sharing data will allow another researcher to preempt the original researcher on future questions, or that the context of the data will be misunderstood, causing data to be used inappropriately in secondary research projects (Sterling; Ceci; Sterling; Bishop; Dawyndt et al).

In advocating the vision of caBIG, one cancer center director trained in systems biology notes, "We are testing the hypothesis that knowledge equals power." The premise is that advanced technologies coupled with data sharing yields collective knowledge that furthers science, and grants the power to find new cures for cancer through personalized medicine. For researchers, the more pragmatic questions are values-based and ask: whose knowledge, shared when, and with whom?

This chapter has pointed to the importance of institutional interests in defining the desired norms and expectations at the social level. Despite this interest-based positioning, however, individual motives are quite different. Interests are external and socially shaped; motivation is more internal and personally shaped, based on individual personality, experience within a system, and other factors. Interests may play a factor in influencing motivation, but interests and motivation are fundamentally different constructs.

One clear message from this chapter's exploration into the technological elements of data sharing was that there is nothing inevitable about it happening; the technological momentum encouraged at a social level becomes more weighted when it's time for an individual to put his or her own hard-earned data, literally, on the line. Individual sense-making of the decision to share is based on an assessment of very personally-driven values, and social messages have failed to completely make the case. Regardless of the technology and techniques available, the broader practices of data sharing advocated and made possible by the caBIG program and others like it will require incentives that do not yet seem fully in place. This is the topic of the next chapter.

Chapter 3: The Economics of Data Sharing

Interviews about the technological variables impacting data sharing often lead to a discussion about the current lack of incentives and rewards for using grid technology and infrastructure in the new way offered by caBIG and programs like it. The leads for the caBIG Data Sharing working group informally list economic factors as the second layer of the data sharing "onion of complexity;" for this research, economic variables include topics such as the value of data and the current rewards system of science.

3.1 The Value of Data

Before discussing the incentives and rewards associated with sharing data, it is important to set the stage by considering the different value attributed to different kinds of data. The previous chapter introduced the various forms that "data" can actually take, from biospecimens, to data about those biospecimens, to laboratory materials, to methods and procedures, to testing outcomes. An article about data sharing from *Law and Human Behavior* touches upon the range of data that may be involved in data sharing considerations, in this context, for replication of experiments to validate findings:

> Scientists may need access not only to the raw data on which that work is based, but also to actual samples, and to the original research techniques and procedures including "know-how," software, and other materials and devices (Sieber 200).

Some of these objects are perceived to have more value, both scientifically and economically, than others; and this differs between scientists based on their research. An

article arguing the importance of centralized biobanks sets the tone for discussing the economic elements of data sharing, focusing on the value proposition of the biospecimen itself.

> Nobody knows for sure, but there is a growing consensus that there are economically valuable and scientifically revealing deposits of biological samples and clinical data around the world, just waiting to be tapped. …. A growing interest in the notion of "personalized" medicine has spurred a growing realization, in both the health-care and information technology sectors, that biobanking could be a very good business indeed. The result has been a growing level of activity in the field (Editorial Staff, "Report: Medicine's New Central Bankers" 18).

While monetary value is not always perceived as the primary measure of exchange as directly as it is captured in this quote, discussions for this research about value and rewards related to data sharing were far more often materialist than idealist in nature. Marxian thought would argue that social conditions and relationships are embedded in products; this was certainly demonstrated in this research, where the "product" is framed as the full range of data types. Data, in all forms, carry materialist potential: the potential to answer a research question, which leads to better publications, which leads to grants, which leads to more valuable collaborations that lead to more papers and grants. Scientific discovery may be important, but someone has to pay for the equipment, and career competition is a real value that drives perceptions and decisions. The sharing of data is the exchange of potential value.

A bioinformatics scientist with experience in both software development and start-up biotech company leadership explains how different objects might have different value, and notes that sharing often differs depending on the work being done and the data being generated.

> People will say, "I have only so many slices of this specific tumor." So, you have a board, and they meet and decide about biospecimen allocations. Dr. XYZ would like to run her particular methodology on this set of tumors. Do we think

> that's worth sending those slices to Dr. XYZ? Those are intensely political often… You would hope these were conducted, you know, with scientific dispassion, or whatever [laughs], but that's just not usually the case… You can at least understand when these things are physical specimens and they may be rare, some degree of caution may be necessary. Every kid in Finland can't write to you for your genetic sequence here.
>
> Now, we are looking at broader and less concrete electronic data sharing. My gene expression data doesn't go away when I send you a copy. There's almost really no cost in that, especially when I use something like caBIG tooling….there's no cost…you just put the data out there. Now, you're saying, is there a justification for not sharing this data broadly? And the justification that people make is about intellectual property, the same old demon. It says, "there's valuable information in here, and I'm not finished mining it yet. I went to the effort of collecting this data, and I'm not sharing it until I am done with it, which may be when I'm dead. Or my university is saying it's IP that belongs to the university and it may have value, even if that value has yet to be established…."

This pattern is also demonstrated in an interview with the director of the informatics core at an NCI Cancer Center (interview questions in italics):

> Yes, we do get asked for data quite often. We have various stockpiles of data. Generally, I am very open to sharing. Of course, we do have constraints on what we are allowed to do…. If the data is in our databases, if it's another researcher's data that we have done analysis for them, we contact them and get their permission to share, and many times, yeah, they allow the data to go, especially if it's post-publication. We do have IP issues and contractual issues with various organizations that do not allow me to share data when we get requests. On the whole, I am very open. I do think that data should be shared openly. *What kinds of data are we talking about?* Generally, things like micro-array data, SNP data, sometimes analysis data we are doing, like if we created a signature for various types of cancers or outcomes, but it's mainly the genomics and proteomics type of data. *How do they know you have it in the first place?* Well, a lot is from publications. Once we publish, researchers here often publish, we say we have these data sets, so people will contact us. People will also see press releases that we are doing work with [commercial organizations].
>
> When we are talking about translational research, we are talking about Big Science, the only way we are going to make huge inroads into personalized medicine is by sharing data. No one center is going to be able to generate all the data we need to meet these lofty goals we all have for personalized medicine. We have to come together and being willing to share data. It's much like the

> Human Genome Project….. No way one center could have done that work, we would still be waiting.

These general expressions of openness shift quickly only a few minutes later when the same discussion turns to biospecimens (interview questions in italics):

> Currently, we do not make available the tissue samples we have in our bank. *Do you get asked for them?* We do all the time. We get requests all the time. *Why don't you share those?* The main reason is that we don't have the capability to make that readily available. When we do get requests, it's around the (proprietary project) and we contractually cannot provide information on those samples…. Any publically-funded collection of tissue, we do plan to make that available through caTissue on the Grid - that's caBIG's biospecimen tool.
>
> *When you say you are sharing data through caTissue for publically-funded data, are you sharing the data about the tissues, the tissues, or both?* Just the data about the tissue, but then of course, we expect to get requests. We do get outside requests periodically for tissue. We actually have a committee that all requests go through. We sit down and make hard decisions about whether the science is valid, what the value of the samples is for the center. We will share periodically the tissue, but once again, there has to be a good reason for it. Whoever we are sending it to has to have at least some affiliation or collaboration with a researcher here.
>
> *That's such a difference in the way you were talking about genomics and proteomics data, "I am open, I want to do it." When you are talking about tissue, it's much more restrictive. Summarize the difference between those scenarios, because those are very different responses.* Why is there a big difference? (*Laughs*) This is just my opinion; we tend to think of the tissue as having high intrinsic value for the research we are trying to perform, so we have a lot of plans for that tissue…. We are looking at making sure that we have the resources for the research that we want to do, that we are not just giving away samples that could be high value for us, even a year or two down the road. Not to say that we don't, we do release some tissues, but we are more careful about it.

These quotes and others reveal a pattern of value perception when it comes to data. 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.

In these contexts, biospecimens (particularly human tissues) are perceived as more valuable because they are a limited resource; whereas derivative data generated from these tissues (e.g., gene expression data) are perceived, by these particular scientists, as longer lasting, and therefore of lower value. On the other hand, that derivative data may be perceived by *other* scientists as *more* valuable than these scientists see it, because of the knowledge that may yet still be revealed from it and published upon. One scientist explains that the very format of the derivative data itself, regardless of the tissues it came from, may reveal the research question that a scientist is working on; knowledge and intellectual property that could be very valuable by another researcher competing for the same grants.

The opposite justification applies as well. Whereas the scarcity of data and materials may discourage data sharing, it can also motivate data sharing in different circumstances. A clinical research scientist, who shared data with another researcher to assess clinical outcomes with a specific colon cancer treatment, reported:

> Ultimately, the strength of my paper was stronger, because we were able to combine our data… It takes so long to do some studies, especially those involving human subjects, to get enough sample size to make profound discoveries. Science moves too quick these days to spend 8-10 years working on a study. You can't wait that long. You are going to get funding for three or four years, you need to have results…

Another person working closely with clinical scientists noted this same decision, but in the context of studying rare disease:

> There aren't as many subjects with orphan diseases. Nobody gets a lot of samples or patients, it needs some collaboration between researchers… For rare diseases, it's more needed for people to collaborate, there's a dependency on each other. You can find 10 people, I can find 20, there may be enough sample size. On my own, I may never reach significant sample size. People are more co-

dependent in this case.

These examples show that the potential value of knowledge embedded in different types of data plays an important role in discouraging data sharing, while having *enough* property or seeing gain from an exchange in terms of professional payoff can encourage data sharing on the other.

There is also value attributed to data based on the time and labor invested to collect it. One researcher, also a director of a bioinformatics core at an NCI Cancer Center, focuses on the value of time invested in generating different forms of data. She counters the idea that the longevity of gene expression data makes it an attractive object for sharing as suggested by the interviews above:

> [Talking about genomic high throughput data] There is some feeling about the data being proprietary, not like it is a trade secret, but more like that there was a lot of investment, effort in assembling it, in cleaning it, and so on… It is a huge part of the task. The analysis and writing the papers is well under half the time that people spend…. A huge amount of time is spent assembling data, so it is very valuable, and you don't want to give it away if no one is making you. That data is a resource for your shop that you have developed yourself.

This researcher is also direct about the cost implication of generating and sharing biospecimens:

> With tissue, it is very expensive to get this stuff, you've got to have a surgeon, someone to ferry the stuff away as soon as it is taken from the body, you have to handle it in a certain way. There are a lot of quality issues you have to manage…. There is definitely expense around getting the tissue and storing the tissue. You've got freezers, and they have to have power, if the power goes out, you can't lose power, and so on and so on. All tissue is expensive to get and maintain. Sometimes, tissue is rare – if it is a rare thing, if you give it to someone else, you will not have it for something that you or another researcher in your institution will want it for.

Different scientists see different forms of data as being valuable in different ways. In general, biospecimens rank the highest in value because of their nature as limited physical resources that require clear and quantifiable investments to acquire and store. The closer a scientist is to the generation and use of data for their research question; the more valuable they are likely to perceive the data to be. In STS terms, quasi-objects, combined with the practice of actors, yields items of perceived material and political value; that value determination is a subjective process shaped by a variety of institutional and personal factors.

Now, the discussion turns to the incentives and rewards that accompany these objects in order to realize value in the eyes of the researcher who holds them.

3.2 The Rewards for Sharing

The different value assigned to different objects depending on the investment made to get them, their relative scarcity against the needs, and future potential that may lie within them correlate with the rewards that are perceived from having and using them. Interviewees referenced a range of rewards in cancer research, which vary depending on the field and the role of the researcher involved:

- Research grants (generally from the government) or contracts (generally from commercial organizations)
- Publications (referencing the common “publish or perish” success criterion in academics)
- Professional recognition, such as invitations to be on peer review committees, editorial boards, or conference panels

- Solving interesting problems first in order to advance reputation and accumulate more investigative opportunities (competitive advantage among peers)
- Patents on inventions (products) or approaches
- Tenure (university systems)

One senior research professor and department head who also holds a medical degree notes the relationship between these various rewards.

> What the reward system is depends on the area of academics or industry that you are working in. If your assignment is to generate patents; if that's what your job description is, then you are going to be interested in patents. What I need to do is generate grants and external funding. Some people may say that it is publications, but publications are necessary but not sufficient. If you are working for a university, the university needs to be able to pay for the labs, the lights, the secretarial help, and all of those things that go into a working environment. The way they pay for that is overhead money on grants. I can publish all I want, I can publish hundreds of papers, but if I am not bringing in money, it doesn't matter. Like a lot of things, it boils down to money. If you can support yourself with your mind, get people to pay you for your mind, then you can stay. Once you get tenure, they can't fire you, but they can reduce your pay, take away your office space. That is what happens on the biological side; it may be different in other disciplines. Grants are really the bottom line.
>
> Now, if you get grants, but don't really publish, then ultimately, you won't get more grants. The publication is the outcome of the grant. If review groups see that you are getting paid to do research, but you are not yielding results, then that's a bad track. You won't ultimately get funded. For faculty advancement, they look at your publication record. So publications are important, but not enough by themselves for advancement.
>
> For medical professionals, they support themselves by doing medical work. They have a certain number of hours dedicated to medical work per week, and then they do academics in the rest of their time. Publications are important to them, but for those guys, if they are not the ones who are getting grants, then they are typically co-investigators or co-authors on someone else's grants, so they are continuing to contribute. They have to generate money from clinical work, and if they don't do that, they will be gone the same way someone who doesn't get grants will be.

This researcher, and others in academic research centers, mentioned that publications are a 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. People win grants in order to produce results (positive ones) captured in publications; they then find each other through the literature; once they find each other, other collaborative opportunities become available in a continuing cycle that ultimately accumulate and shape the direction of a scientific career.

What does this have to do with data sharing? One clinical researcher, who rarely shares her data with others due to her very narrow subspecialty of interest ("I don't imagine anyone would want it," she said), described a time when sharing her data furthered her professional standing in her field (interview question in italics):

> One of our professional societies was in ongoing discussions with the FDA about an approach to a form of testing. So we did actually provide a de-identified data set to our professional society to submit to the FDA as it supported the argument for a particular way to test for something. That's probably the only form of data sharing I have participated in. *What was the benefit of doing that sharing for you?* One benefit was to argue the method to the FDA. Second, making that contribution heightened my standing in the eyes of other people in the profession, that I actually had this data, that I could make a valuable contribution. When I think about my visibility within that professional society, since that time, I have been asked to hold positions and serve on committees within that organization, and I think it is reasonable to assume that my sharing the data contributed to that rise in presence.

This same scientist has chosen not to share in other cases, a decision made based on the perceived value of the materials (interview question in italics).

> There was a request from another institution, looking for tumor samples to support a GWAS [genome wide association study]. I judged not to send my numbers of tumors over there, mostly because we have a relatively limited number of tumors, and I need to take care of my own study. That's not really data sharing; it's specimen sharing.
>
> If it had been more of a collaboration, I would have been more willing to share them. *What would the minimum conditions of collaboration be for her to engage in this process?* Having been approached about participating in the design, the analysis, and the publication. The absolute minimum would have been some publication credit.

Interestingly, this researcher (as a few others did) even bracketed off biospecimens from the category of data considered to be part of data sharing. Data about physical objects "in;" the physical objects themselves "out" – despite the fact that biobankers conceded that it would be rare to want data about tissues without eventually wanting some of the tissue itself depending on the research question. Researchers investigate tissue sample data in order to find the tissues needed; the idea that data sharing would include the data but not the physical sample it represents is an interesting conceptual bracketing exercise that reaffirms how the different perception of value is handled, and how the withholding of them is justified, when talking about sharing.

One author has suggested that "data sharing, if responsibly handled, implies an arrangement that is equitable, non-harmful, and scientifically legitimate" (Sieber 202). Several scientists agreed with this sentiment, aligning "non-harmful" as "not detracting from my own research." The same researcher quoted immediately above continued with this comment:

> If sharing the resource didn't detract from my own research program, then the incentive could be less. If it was a data sharing project, and the focus of the data sharing was not directly in line with what my main research interest was, such as sharing a data set with another investigator who wasn't going to publish the same analysis that I wanted to do, then I think I wouldn't necessarily require

collaboration to the degree of being considered a coauthor as a threshold for sharing.

All of these evaluations point to a rather complex personal values-based calculus of investment-reward taken on a case by case basis, when researchers are asked to consider a data or resource sharing opportunity. If sharing helps someone more than it hurts me, then it will be done for less of a reward than if it is directly related either to the work the researcher is doing, or contemplates someday doing.

There is a time element embedded in this value consideration as well, particularly for bioinformaticists engaged in more technologically-driven research areas. For these scientists, the data to be shared may not be tissue samples or gene expression data, but rather, the software tools, methods, and algorithms that enable the analyses to occur in the first place. A bioinformatics director explains the cycle:

> This is no different from any other technological progression over time. We started with he who has fire is king. Then we went to the stone age, and then the copper age, and then the iron age. With each advancement of my tools are better than your tools, suddenly, all the preceding tools are worthless. Who cares about fire if I can walk up to you with a gun and take your fire? It's really no different. As we develop more powerful technology and analytical methods, and ways to look at science and biospecimens, suddenly, the preceding stuff isn't as sexy anymore from a research perspective. It's still necessary, but it loses some of its value.
>
> If you think of it from a monetary standpoint, with the rapidity of development, you have to use your resources well and quickly to get as much value out of them as you can. Two years from now, they could be worth - and worth is in quotes there - half or a quarter what they are today. The pace of scientific development is pushing data sharing, because suddenly your data may not be worth anything, so you might as well look like a good guy and share it. But during the time when your data has value, you want to hold on to it very closely, and do as much as you can with it during that very short time.

Again, the theme is consistent, even when time and progress are the variables rather

than the inherent value of an object in a specific time and place. That which is considered most valuable is held close, unless it is the action of sharing itself that gets the reward. That which is not likely to earn rewards, is more open to be shared because the risk of sharing is lower.

Grants and Contracts: NIH Data Sharing Policy

Given that grants are such a key source of reward for many cancer researchers, NIH has worked to influence the incentive to share through their grant evaluation processes. From the NIH's perspective, data sharing is positioned as important and tied directly to grant awards:

> NIH reaffirms its support for the concept of data sharing. We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. The NIH endorses the sharing of final research data to serve these and other important scientific goals… Starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible (National Institutes of Health, "Data Sharing Policy," Par. 2).

While many interviewees noted that this requirement has been useful in signaling NIH's belief that data sharing is important, they referenced three factors that detract from the true "across the life cycle" data sharing that is possible through the new bioinformatics tools and approaches such as from caBIG.

First, the policy specifically references "final research data," and specifically acknowledges that release may coincide with publication. The posted final policy announcement, referring to the feedback received to the draft policy, notes:

> Several groups and individuals objected to sharing of research data prior to publication. As noted earlier, NIH recognizes that the investigators who collect the data have a legitimate interest in benefiting from their investment of time and effort. We have therefore revised our definition of "the timely release and sharing" to be no later than the acceptance for publication of the main findings from the final data set. NIH continues to expect that the initial investigators may

> benefit from first and continuing use but not from prolonged exclusive use (National Institutes of Health, "Data Sharing Policy," Par. 8).

This response aligned with the belief from many interviewees, already discussed above, that it was unrealistic to expect data to be shared pre-publication. At the same time, they do concur that this does not quite get the community further towards the ideal of active and open sharing earlier in scientific processes. While many interviewees support a "data embargo" (a time during which data is protected from others' use), the NIH policy does not articulate clear guidelines for more diverse options.

Second, the policy "supports" and "endorses" data sharing and "expects" a *plan* for sharing or explanation as to why it's not being done. This is not directive language that communicates a requirement of the dreaded regulatory "you shall" statement. One research group has maximized this flexibility by identifying their "plan for sharing" as depositing the data from their study in a database mandated by the journal being published in, "which, by the way," the lead researcher reports, "we would have to do anyway based on the journal submission requirements." Since they would have to do that submission anyway, the NIH Policy hasn't really, in that researcher's mind, required them to do anything differently at all.

As another "out" highlighted by an interviewee involved in clinical studies, the researcher could simply write his study protocol in a way that would cause the Institutional Review Board (IRB) to determine that sharing isn't possible due to privacy restrictions on the data, and "there, you met the policy without sharing anything."

Third, it was pointed out that grants are not legally binding in the same way that contracts are; hence, said one interviewee, the "squishy language." The government simply can't place mandates in grants. Along with this "lack of teeth," at a practical level, many

interviewees also noted that there are few mechanisms in place to assess whether data sharing has actually occurred and what value that sharing provided. One researcher summarized the sentiment when asked about how he integrates NIH Data Sharing policy requirements into his research plan:

> That doesn't weigh much into it for me…You can find some way to fulfill the NIH requirements without actually sharing all your data. I wouldn't want to be the one to want to shine a light up NIH's skirt, but [laughs], I'm not supposed to go above 30 on the street out here either, and I and everyone else does it.

In general, academic researchers agreed that the NIH Data Sharing Policy, while useful in pointing to NIH's interest in the practice, is not strong enough to qualify as a true incentive, and there is no real reward attached for actually developing a data sharing plan where data is actually, laughs one researcher, "well, you know, *shared*." Another senior researcher notes:

> In terms of funding, you need to give incentives for sharing data, rather than just sharing results. 'I can tell you my results and conclusions, but I won't let you see my data' – that's not data sharing. One problem is the fundamental question of how you incentivize people to have data sharing behavior. Right now, according to NIH funding mechanism, there's really no – it's very competitive – everyone competing with each other; there's really no award that says that we are rewarding for data sharing… even though there is a requirement, I don't see people putting all their data out – they are just meeting the minimum threshold.

Interviewees agreed that the current NIH Data Sharing Policy and grant evaluation process in general fails to adequately incentivize the data sharing that is likely to be needed in upcoming years as biomedical research continues to expand into new technological realms. They also agreed that it would be grant-giving organizations like NIH that could make the most difference in this area if they strengthened the way grants are evaluated, both at the time of application, and after the grant is complete to determine that data were shared and that this sharing added value to the scientific process.

An NCI representative acknowledged the problems with the NIH Data Sharing Policy, noting that it is ultimately up to NIH Program Directors to ensure that data sharing actually happens on a specific research project. “Some of them advocate for sharing; others, though, they let it slide, because they don’t want others to beat them to results either; their careers are vested in the work the same as the Center PI’s [Principal Investigators] are.”

The caBIG program has recognized some of the disadvantages of grant mechanisms in incenting specific behaviors. Most of the funding awarded to caBIG participants, primarily NCI Cancer Centers, is awarded in the form of contracts, which are different than grants in the degree of specific obligations that can be imposed for the reward provided. This difference is generally acknowledged to have caused some culture clashes in the program between the cancer centers and the NCI; being accountable for specific products and deliverables against specific timeframes is a practice that grant-friendly academics admit to not being fully used to.

One caBIG community member experienced in molecular science described this tension when it came time to meet a contractual requirement for data sharing agreed to by a Cancer Center, citing the lack of alignment in expectations about when data would be shared with respect to publication status.

> The adopter team [Cancer Center contractor] ended up putting out onto the Grid not the data that they had proposed, because it was not yet been published. They ended up pulling data from somebody else’s paper, so it became this exercise in, “Well, I signed the contract, so I have to put something up there.” It ended up being busy work …. It was very frustrating, *very* frustrating. The answer was, “People have to think differently about it.” And I was thinking, “No you don’t, you just have to wait for it to get published.” It is everyone’s interest to publish your data; it’s publish or perish. There’s *no* interest in sitting on your data; it’s all a matter of timing.”

Upon hearing this story, one caBIG program member countered this perception, "We are not asking people to share pre-published data. Yes, we hope someday that will be the norm, but we know we aren't there yet." This appears to be an area of both misunderstanding and conflict; clarity in expectations, and the difference between those expectations and future hopes related to pre-published data, is an important element in shaping how caBIG is perceived.

Other Incentives: Publications, Collaborations, and Promotions

Given the lack of strength of the NIH grant policies as an incentive for sharing (or at the least given the perception that the policy creates no punishment for not sharing), and the relatively rare use of contract mechanism, interviewees reported there are three other avenues that do or could serve as alternative mechanisms to incentivize data sharing, each in a different way. These are:

- Journal publications
- Formal collaborations
- Career evaluation

The first path that already works well to incentivize data sharing comes from scientific journals in genetics and informatics. As described previously, publications are a key piece of "currency" in research academics, and as such, the signals that they send requiring data sharing are considered important to researchers who wish to publish in those journals (McCain; Marshall). In fact, researchers in the area of genomics and proteomics noted that some of the leading journals in that field set that tone early for that sharing. One researcher notes:

> The prevailing view in genomics is open sharing. To reproduce analysis, you need access to the underlying sequence…. Journals insisted that if I am going to publish the sequence of a new gene, I had to publish in GenBank and submit the ascension code. These journals insist that you deposit data that goes with your work, both so the work can be validated, and so that it can be extended. Trust me is not part of science. In order for science to properly operate, you need some kind of concrete and demonstrable evidence that what happened is what you said happened.

Published results require validation; sharing the underlying data allows for that validation. The point here, though, is that these incentives are tied to the publication of a result; credit is awarded for the paper (reward), the value is fulfilled in a very direct way. People share data when it comes to publication time, because first, I need to in order to get published; and second, once I publish, there is less need to hold on to that data, because with some exceptions, most researchers agree the potential value has been realized. Journals are *not* asking, "What data did you share along the way to the result here, and how did it further your (or someone else's) science?" The deposit made is the data that relates to the paper published; there may be extensive data left out of that process – the data generated earlier in the process, and the remaining elements of the data set not included in the published set are not engaged with by the journal-based incentive structure. As noted in the previous chapter, this incentive structure also does not pick up data and analyses that did not lead to results deemed publishable.

Often, reflecting the importance of this incentive, researchers will negotiate some kind of authorship in exchange for the value, including data, they have added to the effort. "Being an author is huge." Unfortunately, this has led to the following dilemma, described in a May 2009 article:

> **Proper recognition for authors and contributors.** The traditional form of acknowledgement is through a publication, which is also a key way of ensuring career advancement. Many journals require that data production should be acknowledged, but how this is done is largely left up to individuals, who follow the norms that exist in their particular discipline. One solution has been to publish articles with large numbers of authors, as this recognizes the involvement of many researchers and data producers in large collaborations. A difficulty arises, however, when the number of authors becomes excessive, as authorship is more a reflection of contribution to a project rather than to a publication (Kaye et al 332).

Furthermore, many interviewees noted that in a paper with a long string of authors, the first and last authors are really the only ones that are able to truly get "credit" on their career scorecard for the publications.

The second path to incentivizing data sharing is more personal in nature, and consists of the formal collaborations that are developed in order to pursue research and then publish results. Previously, it was noted that researchers are more likely to engage in point to point sharing, than the institutional network sharing propelled by caBIG. Agreements to collaborate emerged from interviews as the most positive and frequent incentive to the actual data sharing that currently occurs. Many reported that there is a ritual involved in this process: researchers with complementary interests meet each other at conferences, through reading publications, or through advisors and networks; they then agree to pursue a collaborative study; Principal Investigator status and authorship is agreed upon upfront; research is conducted; and data is shared. There is a code of honor embedded in this process, one researcher notes, saying:

> Quid pro quo – If I share with you, when your project comes along, and I need some help, I absolutely expect that you will share back. And once the word got out that you didn't share back with someone, you would be dead. That is absolutely part of the culture. It's a professional credo. You can't go out and ask people to share with you, and not turn around and reciprocate back. If that happened, no one would share with you again, and you would have to go sit at

> the back of the room. Once you get into the culture, each time it gets a little easier. As you build the network, the amount of data you are able to accumulate and give out, it gets bigger each time.

The personal side of data sharing, which includes the role of relationship-building in encouraging data sharing, is addressed further in Chapter 5; here, the point is that collaborations lead to the potential for more publications, which lead to more grants, and so on. As such, the incentive of sharing data to support a specific collaboration becomes very materialistically valuable. In this way, the potential for a different kind of accumulative advantage occurs, coming not just through the accumulation of publication credits, but through the sharing and accumulation of the underlying data as well.

The irony here is that a researcher must share data with others in order to receive it; the act of giving leads to later accumulation if you trust the "quid pro quo" process. It is no longer only more publications in more prestigious journals that come to the stars of the field; if researchers choose to share their data with others, there is the potential for the accumulation of "data wealth" – another indirect pathway to accelerate a career, if of course it leads to more papers and grants and so on. At this point in time, that is a possibility that is far easier for researchers to believe when it refers to personalized point-to-point sharing, rather than the broader grid-based sharing vision.

Given the "publish or perish" mantra of academic life, it is impossible to speak of the role of data sharing in supporting publications without touching upon the tenure process as a third variable impacting data sharing incentives. Again, this discussion applies narrowly to academic research, however, given the amount of public research conducted by these institutions, this is an important community. Unfortunately, the existing rewards system

related to tenure evaluations does not appear to particularly incentivize data sharing. Many interviewees noted that tenure is still about grants and publications; and that there is currently no standard way to capture the contribution of data sets, and what value those data sets provided to others. "Until that happens," says one researcher, "until it's down on paper as important to getting tenure, we aren't sending the message that sharing is important to your career in a meaningful way. There needs to be a scorecard, and there isn't one for that yet."

Many interviewees cited the current reality that institutional rewards such as tenure and paper citations do not directly recognize data sharing; some note that it will take the embedding of recognition for data sharing within these institutional structures before true motivational shifts will happen. A Department Chair stresses:

> We have to take our institution from the current state that rewards individuals that get grants, and that get publications where they are first author or last author, to places where those are some of the metrics of success, but where there are others. Where we reward collaborations across multiple institutions. If I could show you, look at the tenure and promotion processes – they make you calculate all the grants you got as a PI, the number of papers where you are first or last author; there's nothing in there for data sharing, for publishing papers where you are one of many collaborators and you are all sharing data. There's no metric for it – it's not included in the evaluation process.

Until these are incentives are in place, what does this imply for the kinds of data that are likely to be shared across the grids of caBIG? Many believe that, for now, it is the data that are perceived to have lower value that will be placed online for the impersonal exchanges made possible through the caBIG infrastructure.

> [Reacting to the caBIG "Have my meta-data talk to your meta-data graphic referenced above] I think that what ends up happening in this scenario is that the data that gets out there is not always the cutting edge data; you know, it's data I have had for a while, that I have gotten out of it everything I want to get out of it, then yeah, I'm willing to just put that out there for the world, I've

> beaten this as much as I can possibly do – I can't squeeze another ounce of this myself, it's freeware now. Take it, enjoy it, hopefully, it will contribute to the common good.

This attitude, "I'll put the stuff out there I'm done with," is not likely the ideal that caBIG founders envisioned when starting their journey towards a world wide web of cancer research. Unfortunately, many believe that until the incentives are better aligned to draw high quality pre- and post-published data out more systemically, this may be a significant barrier to success. A researcher engaged with caBIG concludes:

> There's been a bit of push back. There's been a little disinterest. We – caBIG - are putting all this stuff out there, and saying, we build it, they will come. It's like, hmm, no. We need a reason to come.

3.3 Specific Incentives, Specific Sharing: Drug Discovery

So far, this chapter has focused primarily on the academic cancer research community, because this is the social world most closely associated with the caBIG program. Now, the discussion expands to consider the incentives that drive or discourage data sharing by commercial organizations. Many interviewees at cancer centers, in fact, pointed to the importance of these entities in setting future directions in data sharing. As NIH grants become more and more scarce with constricting budgets, commercial entities are now a significant source of funding for academic cancer centers. As such, examining private sector positioning about data sharing is important to understanding the network overall.

While many organizations may have interest in the cancer research work being done by the caBIG community, one group that might specifically benefit should the motive exist consists of pharmaceutical companies interested in developing cancer treatments. Several cancer

centers reported holding contracts with these companies to conduct research that ultimately feeds the cancer treatment drug discovery process. Of these, many reported that most formal intellectual property restrictions on data sharing come as a result of relationships with these firms; materials that would be shared (perhaps) from publically funded data generally cannot be shared under commercial contracts. The legal aspects of intellectually property will be discussed in the following chapter; the goal here is illustrate the role of incentives and awards in driving data sharing in cancer research by commercial organizations.

Commercial company interviewees noted that even though they are not driven by the same reward systems as academics, and that researchers within a company or division are working toward common financial goals, pharmaceutical researchers generally don't share data among each other either. Why? The risks that information could be disclosed, taken to a competitor, or stolen are driving concerns in the ultra-competitive pharmaceutical space. "Corporate espionage - there are efforts at a national level in ways that would surprise you," says one commercial representative.

What are the incentives to data sharing outside the protectiveness of the walls of commercial organizations? The motive to share in this community lies in the area of pre-competitive research and data; intellectual property that could be developed and shared in collaboration to create a baseline of knowledge upon which the competitive drug development process unfolds. This sharing early in the research process distributes both cost and risk among companies, making it an incentive worth considering. One paper notes that biomarkers are a potential target for this sharing and collaboration, because biomarkers help indicate whether a cancer treatment will be effective, before the cost of advanced phase drug trials is incurred:

> Drug development in cardiovascular disease is relatively straightforward: blood pressure and cholesterol profiles are accepted surrogate endpoints. Similarly, drug development in HIV/AIDS has been simplified by the acceptance of viral burden as a surrogate endpoint for drug efficacy. Biomarkers offer this same potential for cancer drug development, but to date they do not exist. Instead, progression-free survival (sometimes difficult to assess) or overall survival (sometimes requiring significant time) are the crude 20th century endpoints for drug approval in an area of medicine in which the science has clearly entered the 21st century. There is no doubt that the development of biomarkers as an endpoint for "go-no-go" decisions within companies is reduplicative and reminiscent of the lack of standards that plagued the semiconductor manufacturers of the 1980s (Curt 3).

Sharing certain pre-competitive data in the commercial sector to save time and money downstream is seen as a powerful reason to collaborate. The conclusion to the article above is:

> The cancer drug development community should consider a new precompetitive environment in which major companies would present their biomarker programs for cancer drug development, under confidentiality, to the NCI. The NCI would gain a unique perspective unobservable to its individual industry partners, as a precompetitive safe harbor, where there is overlap and redundancy, where there are gaps, and where there is promise to validate cancer-related biomarkers that might enable cancer drug development across companies and academia. The NCI would select the most promising partners for codevelopment of biomarkers, and when these are ready as predictive or surrogate endpoints, share them with the academic and industry communities at large. The research would be precompetitive, the risks would be shared, and collaboration would replace competition: important step change in a new safe harbor (Curt 3).

This recommended approach rests on an important element: the sharing of data with a neutral party such as the NCI, called the "safe harbor." Says one commercial research director: "No company would ever share this information with each other, but they might share it with NCI. NCI has an important role here in facilitating data sharing." The conditions under which this sharing would occur in this interviewee's mind include: a high-impact problem that everyone shares that would be significantly eased through data sharing (a clear payoff), a third

party non-commercial organization to act as a neutral broker to receive the data, and sharing of data that is agreed to be pre-competitive by the Justice system.

Broader generalized data sharing has been done by a different company in the area of diabetes research (Tapscott and Williams):

> Earlier this month, Swiss drugmaker Novartis did something rather unusual—and almost unheard of in the high-stakes, highly competitive world of Big Pharma. After investing millions trying to unlock the genetic basis of type 2 diabetes, the company released all of its raw data on the Internet. This means anyone (or any company) with the inclination is free to use the data—no strings attached (Par. 1).
>
> So why the giveaway? "These discoveries are but a first step," says Mark Fishman, president of the Novartis Institute for BioMedical Research. "To translate this study's provocative identification of diabetes-related genes into the invention of new medicines will require a global effort" (Par. 3).
>
> It's worth noting that Novartis didn't reveal everything. For example, it didn't give away three years' worth of its own observations on the data, which gives it a substantial lead time on other companies attempting to exploit the research. Meanwhile, the close ties and goodwill that Novartis has fostered with the research community studying diabetes will give it an advantage over competitors as it moves to the next stage of research (Par. 5).

This is another example of the complex calculus that feeds into data sharing decisions. Whereas cancer researchers in academia will share with collaborators to increase grant opportunities, drug companies may share data to better position themselves for the payoff of drug discovery downstream. Again, in both these cases, the argument for sharing is at its heart, a materialist one. The discourse is not about the greater good of science and finding cures for cancer (though all interviewed agree that this is the ultimate goal, *of course*); rather, it is about the economic terms that make data sharing either compelling, or not. Referencing this connection between data sharing and competitive advantage, a researcher who frequently

collaborates with industry comments on the potential impact of the 2008-2009 economic down turn on data sharing:

> [Referencing data sharing] We've traversed the field far enough; the genie's out of the bottle on that one, I think. On the other hand, I am a little afraid that in a significant economic down turn, and money going away from science has the potential to make everybody entrench. At the same time, what I have seen in at a number of organizations is that people are banding in teams more than they ever were before. People are banding together more, in part to search for dollars. They have used the team aspect of this, it's more effective to do that than to do it alone. I am hopeful that things will play out on that way, but there is no guarantee.

3.4 Networks of Rewards and Exchange

Critics of actor network theory (ANT) argue that this conceptual tool removes the importance of the agency of human actors, that it misses the overarching institutional and structural power dynamics that impact technical systems, and that it removes the element of advocacy from STS analysis (Pickering; Winner). Rather than discard the tool based on these objections, the analysis here argues that blending ANT with a consideration of institutional value and rewards across a network yields insights pertinent to the problem of data sharing.

First, the analytical tool of the "quasi-object" proposed by Latour in *Science in Action* (51-55) is a useful one for considering the continuum of data involved in cancer research. A research focus on data sharing, by definition, focuses on the "object" of the data being exchanged. Identifying the diverse data types involved in cancer research, however, blurs the line between subject and object. A laboratory method employed by a specific scientist, and held only in her mind, is procedural data that might be considered of value to another scientist, and is therefore, a quasi-object that could be exchanged. A piece of human tissue retains an element of the identity of the person from whom it is taken, even after it goes through a

process of de-identification. Scientifically, one could say that the tissue changes from subject to object the moment it is removed from the body; others might argue that as long as the tissue can be linked to an individual, it remains a subject representation.

Data sets on their own are generally considered meaningless without the context of human participation; that interpretation, which can be communicated to others wishing to use the data, could also be seen as a quasi-object. This is not just a thought exercise; blurring the lines between subject and object accomplishes the seemingly contradictory act of identifying objects in subjectivity that can actually be exchanged; and demonstrates the subjectivity of data that is typically seen as objective.

How can the agency of human actors be removed in a study on motivation and emotion, as is suggested by an ANT analysis? When actors are, for the moment, reframed as "quasi-objects," it becomes easier (again, for the analytical moment) to conceptually bracket off the individual motives and emotions of individuals, and turn to more foundational questions about the flow of data and value between the quasi-objects in the network. Is data flowing, or not, and where has it flowed from or to? What causes that data to flow or not, and what happens when two quasi-objects are integrated (an actor analyzing a data set, for example)? How do both the subject and object transform, and how does that impact its movement? Does action along the network stop, does it proceed, and under what conditions does this happen?

It was this ANT conceptual frame and resulting questions that revealed the importance of the value assessment in predicting whether data generally will be shared. Once said, it sounds simple: data perceived as materialistically valuable is not shared; data that is of less value and lower risk is more likely to be. This, however, is not a simplicity that has appeared to

be applied in the social messaging about data sharing on the caBIG program; and which interviewees themselves were sometimes taken aback by. "Well, it sounds so straight forward for someone to say it like *that*," said one scientist, "but when you are living with it every day, it isn't as immediately clear."

ANT also removes the distinction between science and technology, in a way that is helpful in thinking about data sharing in cancer research. Does data that comes off a gas chromatograph qualify as a scientific output, or a technological one? Biomedical informatics has blurred the line between science and technology such that the answer is really no longer important. This was useful in removing the conceptual stovepipes between scientists and technologists involved in cancer research. These players may engage with different tools and data at different points in the cancer research life cycle, but the incentives impacting their willingness to share are often very much alike.

The next conceptual step is to layer on the institutional considerations of value and rewards over this network of quasi-objects. Conceptually, institutional incentive and reward structures, such as tenure committees, publication rules, and funding mechanisms, can be positioned as "magnets" within the network that shape the sharing of data across that network. When incentives encourage specific behaviors at certain times, data is shared. In the case study of academic cancer research, this is seen most productively in the difference between pre-published and published data.

This metaphor of the magnet unfortunately suggests a directional flow to the network (data is flowing from one side to the other), which is not the intention here. The conceptual goal is simply to consider network movement against an institutional background. This is

ultimately a figure-ground question. While ANT focuses on the figures (or quasi-objects) in the network, the institutionally focused perspective adds in the normative ground within which these figures are working. We need both for a complete view. Here, the question is, what institutional factors cause data to flow among actors (or quasi-objects), and what factors cause it not to flow? There is no judgment in this specific analysis about how the data *should* flow; for the moment, in the spirit of ANT, it is a descriptive look, not an activist one.

This analysis again highlights the usefulness of the idea of a quasi-object, because it allows for the same object to be seen fundamentally differently depending on the context: a data set may change significantly in value depending on whether funding and resources are available to support its continued use (the difference between valuable tissue and "tissue rotting in the freezer") and whether the results generated from that data are interpreted as positive or not (e.g., whether it will be shared in a publication's sanctioned database, or whether it will remain on a graduate student's hard drive). The object itself has not changed; but the institutional "magnets" of reward have changed its prominence and placement within the network. The institutional backdrop against which the quasi-object is placed changes the context of that object, and its movement.

Did the choice of the ANT analysis predispose this research to draw conclusions that focus on economic metaphor and arguments? Philip Mirowski, D. Wade Hands, and Chris McClennen's each critique actor network theory when it is taken too far, warning of the dangers of framing scientific activity in terms of economic metaphor. In short, all three highlight the risk that this form of analysis can obscure other possible interpretations of motive and action. McClennen in particular expresses concern that an ANT focus invites a particularly

strong emphasis on the economic aspect of human behavior. By focusing on artifacts and material objects, the capitalistic elements of science and technology become overemphasized, minimizing the social and idealist dimensions of scientific choice and action.

This is a valid concern; however, the outcome from this research suggested that an economic and materialist perspective was actually reported as the most prevalent one in driving researcher choice when it comes to data sharing, and that *despite* the foundational use of the idealist idea of data *sharing,* idealist and social concerns are generally secondary motivators. The nature of this qualitative research did not predispose economic and materialist concerns as the most prominent; rather, it was these dimensions that were raised most explicitly by interviewees, with even collaborative relationships framed as a form of capital. This direct feedback from interviewees preceded the analysis through the ANT frame; the theory did not drive the analysis, it flowed from it. Economics is a metaphor, but it is also a reality facing researchers on a daily basis. While STS has added great value in unpacking metaphor, in this case, the metaphor reflects reality; it is making that reality acceptable, visible, and overt that becomes the more important challenge.

Given this economic reality, norms and counter norms can also then be applied: which appear more prominent in facilitating movement across this conceptual network of quasi-objects? The evidence above suggests that the materialist emphasis of data sharing puts the counter-norms of science front and center, far more so than the norms. Movement is shaped by the different perception of the value of data, the degree to which it should be held in propriety, and the particular collaborators invited to connect together. Interviewees consistently discussed the value of data in terms of the effort and resources that went into

generating that data, and the potential value that might be generated from its analysis and applications. Data sets are seen as material goods that cost money and time to generate, and have the potential for material or professional gain. While the "gain" may differ, the message is consistent: these discussions are framed based on a subjective and pragmatic economic evaluation, not on a basis of an altruistic "gift giving" scientific ideal.

This analysis, however, also reframes the traditional Mertonian interpretation of the publication of scientific papers as motivated by gift giving to the greater good of science. On one hand, the normative perspective would argue that scientists aren't paid for publications; they are ultimately public goods, so there is no formal value attached that would signal a counter-norm motive. Interviews demonstrated, however, that publications are indeed seen as a currency to be transacted. Rather than being framed as a gift, they are seen as both the reward and the investment for a new round of funding.

This chapter has considered the economic variables involved in data sharing decision-making by scientists and technologists in cancer research. There is another layer to be added at an institutional level, however, that considerably broadens the size and scope of the network considered here. It is time to bring in the lawyers, and the influence of legal and regulatory factors on the motives and emotions involved in data sharing.

Chapter 4: The Legal Side of Data Sharing

This is a study about the motivational factors that play into a researcher's decision to share data. To this point, this has included an evaluation of technological and economic factors that inform that process. This chapter now turns to legal, regulatory, and intellectual property issues. While the 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 area considered here is the legal and regulatory area of patient privacy, and the balance between the protection of human subjects in research, and the public good that comes from the research conducted with these human subjects. The second area relates to intellectual property rights, which builds on the incentives discussed previously, but formalizes the value assessment into components of legal ownership. Together, these factors serve to continue to broaden the broad social network in which data sharing activities occur, as the focus moves from the scientist's choice to share, to the various offices and stakeholders that consider the feasibility of that sharing in light of regulatory constraints and intellectual property potential.

This broadening of the network has implications that force a reexamination of the foundational research questions asked here. Early in this research, a university technology transfer office representative cautioned that the very research topic (expressed as "exploring the motivational factors associated with researchers' decisions about data sharing") was an

irrelevant one. "You are asking the wrong question," she said, "Sharing or not sharing is not up to the researcher anymore." Whose decision is it, in her mind? She referenced a complex network of institutional interests, from review boards to privacy offices to technology transfer offices, who now distribute decision-making among them, and then advise the researcher on what to do.

The statement was provocative, and certainly drove some panic into this researcher's head. In the end though, even when asked about it directly, most other interviewees did not fully agree with the comment. One lawyer notes:

> Of course the first line of whether something is shareable is the scientist, because the regulators and technology transfer office and research administrators don't know scientists have data unless the scientist tells them that they do. Most offices aren't set up to do looking at every progress report on every project. There is no way for them to assess the data. Generally, it is the scientist that says, 'I have something of interest to someone else,' and it would be the scientists that would determine who the audience might be, and who the target of the sharing might be. 'I might have something scientifically interesting, and I'd like to share with others.'

Even while agreeing that the decision does begin with the scientist, as this lawyer did, most did not reject the sentiment associated with the original objection about the researcher's lack of ultimate agency. It is clear that the legal and regulatory variables profoundly impact decision-making around data sharing. This chapter considers these dynamics.

4.1 Protecting Human Subjects

The first legal and regulatory dimension examined here relates to the protection of human subjects in research, and the protection of patient privacy, which places specific

constraints and boundaries on data sharing. Interviewees generally cited the following critical tools as instrumental in shaping the legal-regulatory landscape that they operate within:

- Title 45, Part 46 of the **Code of Federal Regulations, Protection of Human Subjects** (45 CFR 46), also referred to as the "Common Rule," which provides for review and approval of human subject research activities, including appropriate informed consent procedures.
- At the institutional level, compliance with Protection of Human Subjects regulations are managed by **Institutional Review Boards (IRB)**, committees whose primary mandate is to protect the rights and welfare of humans who are the subjects of research.
- The **Health Insurance Portability and Accountability Act (HIPAA)**, enacted by the U.S. Congress in 1996, implements specific patient privacy regulations that carry the potential for civil or criminal penalties if violated.

These regulatory structures were implemented to ensure that people are informed of and given the opportunity to consent to research participation in the medical sciences; and that their privacy is protected if they choose to participate. These structures reflect the widely held belief that study participants need to understand the exact nature of their participation in research, understand the potential risks involved in that participation, consent to the research based on that understanding, and then have their information kept protected and private throughout the course of the research. On the positive side, it is hoped that if research subjects feel protected, they may be more likely to participate in research; on the negative side, it is noted that it can lead to such complex notifications that subjects may opt out simply because

they do not understand all the implications of what is involved in participation. Ultimately, the complexity of these structures and the means by which compliance is assured, tend to also discourage scientists from sharing data once they are able to acquire it.

These are vital problems in the area of cancer research. Involving human subjects in research, with all the steps involved to ensure protection, is a complicated matter even when no risk is anticipated and candidates are considered healthy. Consider, in a moment of research reflexivity, the steps undergone in this specific project in order to simply interview people over the phone about their viewpoints about data sharing. This research project required interviewees to read and agree, both in written form and verbally, to an informed consent form; and for the researcher to separate records that identified the interviewee's identity from the content that was in those interviews. Even with these informed consent and protection procedures, there were times in actual interviews when interviewees paused to ask, "are you sure my name can't be traced back to this?"

Now, imagine those same steps with a human subject that is about to undergo a biopsy or treatment procedure related to cancer, and who is asked to contribute tissue towards research. Immediately, the potential impacts change dramatically: Will my employer or insurance company have access to this? Who else might see my information? Will it change how I am treated, both now and later? What else might be done with my tissue beyond what I am consenting for now?

Privacy advocates and experts point to core philosophical concerns that underlie these concerns, which rest in interests related to both control and autonomy. The conceptual argument is straightforward: any data generated from the use of my body should be within my

control, and it should be my choice to share that data with others (autonomy). The logistical realities make this difficult: many scientists would conceptually prefer to have the broad ability to use human-derived tissues in a range of research studies given their value; however, unless this consent is given, this may not be possible. Additionally, there is the question of how much a tissue has to be modified before it is no longer truly a representation of the human that gave it. Tissues are modified with reagents; they go through complex transformation and analyses as the result of different tests. Are these derivative products covered by the same protections as the original tissues? What is the risk that these products could be traced back using emerging scientific methods to the original donor's identity through alternative mechanisms and analysis?

De-Identification of Data

In the incentives discussion in the previous chapter, a clear line of distinction rose between pre-published data and post-published data in terms of attitudes about data sharing. When the discussion turns to legal and regulatory issues, a different line emerges: identified versus de-identified data. Separating data about the human from whom tissue is derived, from the tissue itself, or from other data about the tissue (separating the object from the subject that generated it), is called de-identification. This means that data that can identify a subject is separated from the rest of the data set, so that others can use the de-identified data without being able to trace it back to an individual. Generally, researchers started their discussion about data sharing with the presupposition that we were talking about de-identified data; without that step, "game over" when it comes to sharing with others.

The counter-concern raised by those with an interest in patient privacy is that data can never be completely de-identified; as genetic data analysis becomes increasingly advanced, it may be possible to identify a person from data associated with them. Even though data may be "unidentifiable" today, with now unknown future analysis techniques, that traceability may be possible. This concern is particularly relevant for biospecimens, since physical tissues hold DNA that could be used to identify a unique individual using today's scientific methods. A 2005 report in the *Economists Technology Quarterly* illustrates the concern:

> Bioethicists are quick to point out that the very thing that makes biobanks enticing and powerful to health care professionals and drug companies make them equally so to law enforcement, the insurance industry, and government officials with a different agenda. This fear is not without foundation. The Swedish government, which created one of the world's first biobanks in 1975 – now has at least one blood sample from all its citizens - used a loophole to gain access to the biobank a couple of years ago, in order to track down a killer (18).

While many may not be opposed to tracking down killers, examples like this signal the potential risk of data sharing for scientists, who have a vested interest in having patients feel comfortable donating tissues. While researchers may be well intentioned, there is a perceived (and perhaps actual) risk that data that they hold and share could someday be used to "re-identify" someone's data and subsequently use it for more subversive purposes. In effect validating that potential, in 2008, the NCI restricted a previously open database from public access, because it was determined that a new genetic test could allow for the identification of personal information from de-identified data held in the data store:

> The National Institutes of Health (NIH) announced new procedures for researchers to access previously public databases from genome-wide association studies (GWAS) in light of recently published research that describes a technique that can pick out an individual's genetic fingerprint from a mixture of many

DNAs, even if the individual's DNA is only 0.1% of the total (Clayton, Par. 1).

TIME Magazine summarizes the concern in a recent article about biobanking:

> [Referencing biobanking] Sounds easy, but will it work? That all depends on how comfortable people can get with sharing their DNA. "Having all of your DNA out there where organizations or governmental institutions have access to it makes people nervous," says Dr. Randall Burt of Huntsman Cancer Institute in Utah. The medical incentives are certainly great — scientists are convinced that only by mining the riches of the human genome will we uncover the next generation of treatments for disease (Park, Screen 8).

Scientists and Lawyers: Conflicting Motives

Scientists need data in order to conduct research; in cancer research, donated tissues from consenting research participants are vital in feeding a life cycle of research activity that extends from "bench to bedside," hopefully culminating in treatments and cures. Science needs the donation of human materials in order to benefit the public good. This tension is captured in a legal paper on the topic, which strongly criticizes the implementation of HIPAA because of its negative impact on the balance between patient privacy and the public good need for research:

> As alarmed as Americans remain about medical privacy and unauthorized use of their health information, they also overwhelmingly agree that the United States should be a world leader in medical and health research….. (despite this), HHS (has made a) conscious policy decision to decisively tilt the scales in favor of individual privacy regardless of the detriment to communal progress (Nosowsky and Giordano 576).

There is significant controversy about the benefits and costs of HIPAA, and the perceived conflicts between these tools and other regulations that interplay with the same concerns, such as from the Food and Drug Administration (FDA) when it comes to clinical trials. This project is not to conduct an evaluation of these tools, but rather, to consider their impact

on data sharing decisions, and the network within which these decisions are made. To examine this element from both sides, interviews included both legal professionals engaged in interpreting and applying these legal and regulatory frameworks within their institutions, and researchers that both come to these professionals for help, and who must act on their decisions.

The first perspective comes from the legal professionals charged with interpreting the regulations on behalf of their institutions, who generally report having the motive of protecting patients, researchers, and institutions from the negative consequences of non-compliance, or even the perception of such non-compliance. Reports one lawyer:

> The problem with HIPAA was, among other things, the people writing it were not thinking about research. It imposes a lot of administrative requirements on those who are regulated by the rules. Those requirements are costly, and concerns about violations are extremely high, because for egregious violations, you can be subjected to legal penalties. All of these lead into a very conservative environment. The specter of penalty is out there; and because even if there was no government enforcement, there could be a public relations fiasco that could come if you violate the regulations. All of this feeds into a very conservative environment in most places. That environment has caused institutions to stop sharing data where they used to do so, or to place strong restrictions on their investigators, faculty, physicians' use of data for totally legitimate purposes: public health research.

Another university professional that works closely with IRB's notes:

> I am not going to say that IRB's are a barrier, but there are different interpretations. Even within and between boards, you could have different interpretations. There is a lot of subjectivity still involved. Maybe it's changed a little bit, but there can be differences in a board. Of course, across them, different institutions interpret things differently too. That's a big issue, and I think it's being addressed because people are aware of the issues, but I am not sure how quickly that is happening.

The perceived power of the IRB became very clear in research interviews, both from legal professionals and from scientists. Scientists are clear that the IRB sets the final word for their research's data sharing direction, and some suspect that other scientists even use this to their advantage to some degree in order to avoid data sharing. A question that remained open, however, relates to who the IRBs report to: Who holds these boards accountable for their decision-making; can they be held accountable for the negative impact of discouraging data sharing? One interviewee close to IRB governance issues confirmed that, in fact, there is little guidance about how IRB's are overseen. IRB's are charged with making sure regulations are followed, but there appear to be few standards related to "who watches the watchers" other than the publicity concerns than can result if something goes wrong. As one researcher notes, "the public risks of getting a patient angry because data were inappropriately shared and used is greater than the public risk associated with having *not* done a research project because you did not have data. Risks are higher from harm than from omission." Another researcher simply notes on a listserv dedicated to data sharing issues: "We REALLY need to discuss what sharing over the grid means to IRBs. Without being able to articulate this, it will be impossible to get our IRBs to approve this."[11]

From the scientist's perspective, legal concerns seem most apparent on the clinical side of research rather than the bench side, which is more where the more safely de-identified molecular work is done. A data manager that works with both clinical and molecular scientists notes:

> Clinical trials have to deal with potentially identifiable data, so there is a fear-factor, this I am sure you have heard many times, this prevents data sharing.

[11] "This" refers to connecting an institution's database holding human tissue inventories to the caBIG Grid.

> There's an over-reaction to HIPAA sometimes, they have to write into their consent forms that they won't, that they will collect only minimal amounts of data. They might then be missing some meta-data, they go overboard in getting as little data as is needed, which makes sharing hard. People are very afraid of HIPAA, and that trumps everything. This trumps the NIH Data Sharing Policy any time; you can't get sued for *that*.

From scientists, all of these concerns have led to some adversarial perceptions when it comes to lawyers. The general sentiment, shared in some interviews and even in some comments in general conference sessions, is that when it comes to data sharing, the perception is that "the bias is always towards no when the lawyers get involved." One scientist said it more strongly, with the frustration clear in her voice as she spoke of problems encountered when trying to share data with a colleague at a different institution:

> Once you get lawyers, involved, they don't care; they just don't care. All of us have found them just very frustrating. Because they weren't hearing what we wanted to do, what we were doing was so far out of their experience. I don't know, I don't deal with clinical data, so I understand that they need to be involved with that, but this was different.

One lawyer understood the frustration, saying, with what sounded like sympathy in her voice:

> The complexity of the regulation, the reactions to HIPAA, have really hindered the ability of researchers to exchange information for appropriate purposes. So when Joe Researcher at University A wanted to share with Jane Researcher at University B, they might have once been pretty lax about it, maybe query about whether there was a risk or not. Now, they go through 20 people approving it, and no good system for exchanging it, and it has become such a hassle, that where things were very collegial and informal once, it has become such a hassle that someone who was willing to do it before, may not think it's worth it anymore. It's not that I don't want to share it, it's that I don't want to go through the hassle of sharing it.

There is an emotional divide here between the social level institutional factors that complicate data sharing (the legal-regulatory frameworks) and the reactions to those

structures. Researchers become irritated, even angry, when talking about lawyers; the conflict becomes personal. The lawyers report that, they too, generally think HIPAA has crossed a line of reasonableness when it comes to research, and is being applied in ways that were never intended, but they ultimately feel little control over fixing the institutional problems. The result is both a scientific impasse and a personal perception of conflict, neither of which appear to advance the personal or public good. Even returning to the overall goal of protecting the patients at the heart of the research doesn't seem to help much, because few believe that the current requirements actually do much to help patients when it comes to the research that could save lives.

4.2 Data Ownership and Intellectual Property

One critical assumption has underpinned the discussion to this point: the data generated to support cancer research is the researcher's to share, assuming they have the approval of their IRBs. The reality, in fact, is that this is not true; if a researcher leaves an institution, he or she may be required to leave materials behind. This clarity has not always been present. A 2006-2007 legal case that was raised all the way to the Supreme Court (Washington University v. William J. Catalona) generated great interest among caBIG-connected researchers, as it judged that a scientist leaving academic institution could not take a collection of tissue samples with him, despite years of having built the repository and conducting extensive research with it. This decision determined that institutions own data, not researchers; it also affirmed that patients do not hold the power to make decisions about the

institutional transfer of their tissues post-donation. A 2007 editorial in the *Annals of Neurology* reports:

> On June 20, the Eighth Circuit Court of Appeals unanimously upheld a lower court ruling that cells, blood and DNA donated by patients belong to the University, not Catalona (who developed a valuable test using the materials) or even the donors themselves. Two months later to the day, Justice Alito denied Dr. Catalona's request to bar Washington University from using thousands of tissue and blood samples in question until he has petitioned the full Supreme Court for a review of the appeals court's ruling. Justice Alito's thumbs down was the latest development in perhaps the highest-profile case involving ownership of biomedical research specimens donated by patients to institutions.
>
> The precedent-setting Eighth Circuit ruling appears to give research institutions the right to dictate if, and under what circumstances, researchers from other institutions and former employees can use donated biomedical specimens. Perhaps of even greater importance, the ruling indicates that institutions have the authority to prevent researchers from taking samples with them when they leave for other institutions (Editorial Staff, "Researcher Access to Patient Samples Reaches Supreme Court," A12-A14).

While "science" as a field has worked hard to maintain autonomy and independence, individual researchers are ultimately employees of their institutions, and it is their institutions in addition to the researchers that reap the rewards of scientific research. This reality transitions this research to the topic of intellectual property (IP), and the impact of legal property rights on data sharing decisions. These issues were touched on previously in the discussion on rewards, as material with high value is likely to be considered to have greater IP potential. This discussion, however, dives more deeply into the legal implications involved in a quasi-object that is also categorized as legal intellectual property.

As with other topics, biospecimens are a useful place to begin in considering the legal issues involved in intellectual property, particularly given the high value attributed to these

objects, and the resistance expressed by researchers when asked about sharing them. Is tissue an item of intellectual property? Legally, as noted above in the case of Washington University versus Catalona, once a biospecimen is donated, it no longer belongs to the donor. Instead, it generally is categorized as a "gift" to the institution to which it is given, which then has rights related to that sample. This right, however, is both constrained and governed by the patient consent and privacy rules discussed above, even though the patient no longer "owns" the tissue.

> Consenting patients for future use of biospecimens is further complicated by the discrepancies between relevant federal regulations. Although 45 CFR 46 Subpart A (the Common Rule) allows consent of patients to future unspecified research, the Health Insurance Portability and Accountability Act Privacy Rule requires that each authorization by the patient for release of protected health information includes a specific research purpose. In addition, under the Health Insurance Portability and Accountability Act Privacy Rule, the creation of a biospecimen resource or database with protected health information and subsequent disclosures for research purposes are considered separate activities. Each activity requires authorization from the research participant unless a waiver or alteration of authorization is obtained from a Privacy Board or Institutional Review Board. Because support of future research is a major purpose of biospecimen resources, this lack of harmony among federal regulations has had a significant effect on, and created a great deal of confusion within, the biospecimen community (Vaught et al 2522).

The impact of this is that even if an institution maintains and invests in the retention of donated human tissue, it may not be able to be used for downstream research activities if the consent process has not specifically allowed for it. In this case, stewardship of the biospecimen may not lead to the freedom to use that resource for research.

There are a number of forms of IP generated by the cancer research process, and as such, the caBIG program remains concerned with considering these variables in its Data Sharing and Intellectual Capital workgroup. The caBIG program itself is considered open source and

open access; software produced by the program carries an NCI or caBIG license, but is provided free of charge. That said, there are other elements of IP associated with cancer research that are not expected to be free. Unfortunately, other than IP controlled by a patent or license specifically, there do not appear to be overarching rules in academic science for clearly defining what should be publically available (after being de-identified), and what should be allowed to remain proprietary for academic competitive advantage, even if funded by public money. An NCI representative shares:

> You do have funded programs like the TCGA [The Cancer Genome Atlas], which NCI said from the beginning, this is a community resource project. All of the data needs to be deposited; it is essential to the funding. There, they have gotten away with it. There is certainly nothing legally there to say that, 'this is publically funded, you have to share your data.'
>
> There are also public policy decisions here. One of the reasons that universities and companies don't want to share data prematurely, it could stop your ability to get a patent, because patents are useful things to have for commercializing findings. Where's the line? If we encourage scientists to share data widely and freely, because it is a commodity, how do we define the attributes of data that should *not* be available as freely, that should be protected, for at least some period of time? How do you tag that, so the right data gets to the right people at the right time?

Why isn't all publically funded intellectual property just open and free? The primary legal tool protecting intellectual property funded by the government is the Bayh-Dole Act, which was passed in 1980, and relates to the designation of intellectual property associated with research funded by the federal government. In summary, this rule gave research universities, including those conducting cancer research, more control of their inventions, even those created using public funding. This allows both university and commercial organizations to protect their work through patents, a primary tool of controlling intellectual property.

Scientists, as noted above, tend to be interested in grants and publications in order to advance their careers; motivation comes from being the first to share their findings in competition with their peers. On the other hand, in addition to these academic rewards, *patents* are considered valuable to *organizations* and the technology transfer offices that manage these legal tools, because they reflect proprietary value that could potentially be leveraged for money, gained through licensing fees (e.g., licensed access or subscriptions to databases) or future contracts. Just because a database is formatted to meet caBIG standards and be interoperable does not mean that the data within it must all necessarily be available for free.

This quickly becomes another example of why researchers may not have the control that they previously had when it comes to sharing. The researcher may be the first to raise the possibility of sharing something with a colleague at another institution, but the possibility of being rejected lies in other offices. Technology transfer offices, with interests in protecting the financial potential in research can deny the ability to give something away for free; and IRB's can deny sharing (or confirm a researcher's choice not to share) based on human subjects and privacy concerns.

An intellectual property expert explains also that "reach through value" needs to be considered – it is not just the value of the data itself, but the value of an invention that could lie in the data, and is just waiting to be discovered:

> There is a big difference in the calculation between looking at the proprietary nature of the data set itself, versus the proprietary value in an *invention* that could be *derived* from access to that data set. Those are very different things. There are data sets that are so inherently valuable, that the institution or the scientists are interested in mining the proprietary value of the data set, and that may be just be because they can, because there is money to be made. And

> administrators like me can be very creative in how we do that, selling subscriptions, or selling licensing rights. There's also value in reach through rights, though universities don't tend to go there; it isn't considered very collegial in academic circles, and it would be hard to stand up and claim those rights under current property law.

The lines are far clearer, it appears, in commercial environments. Here, interviewees concurred that there is a clear delineation between pre-competitive data (discussed in the previous chapter) and competitive data, and cited patents as the defining tool for protecting and licensing materials for commercial gain. A researcher experienced with biotechnology inventions in the commercial sector discusses a strategic use of patents, and how the protections can actually support data sharing. This story also illustrates the point that just because it is patented, doesn't mean it can't be openly shared – the user just might have to pay for something down the road (interview question in italics):

> We were collecting data, and as we were doing it, we made a tremendous investment in high throughput DNA sequencing apparatus, machines that cost a half a million bucks each, and they were churning through human genome sequences looking for genetic variation that could be associated with risk of cancer and efficacy of drugs. We had a whole system, I spent three years building that system and running it. Once we had the data, and the techniques that were used to collect that information, they were proprietary, I hold patents on them. When I wrote these algorithms, I also wrote patents, and I got granted those patents. We had deep intellectual property… This was like a machine for generating intellectual property.
>
> Now, here's what's interesting. I made a case very very strongly to the company, and I had a lot of influence, that we should publically share that data. Now, that sounds a little odd, doesn't it? But here was the argument I made, and I believe in it. Once we had the patent rights, it made no sense for us to sit on the data. We should put that data out there, and make it as broadly useful as possible, because if someone figures something useful out that is within the scope of the patent, we would have the commercialization rights associated with it. We wanted to get it out there under a broad license, a reminder saying, 'Hey, we hold certain patent rights. If you use this, you acknowledge that.' Let's imagine they find things out, and they go develop a drug based on it, and they don't

> license the patented intellectual property from us to do so. Well, guess what, we have a whole army of lawyers standing ready to present to them the agreement that they made, the patents we hold, you owe us! The point is, from a research standpoint, we don't care – we would rather people are doing these things, as long as we get the value from the derivative product. It's a little radical.
>
> *What are the objections to this approach? Why is it radical?* How do we know we are getting paid fairly – what if they go do something with our stuff that we don't know about or can't tie back to our stuff? What if they patent things that could piggyback on top of ours? This mirrors the "we're not done with this yet" objection in academia. And there were questions about whether these kinds of agreements will actually hold – this was all new, it wasn't clear if it was going to work or not. Is there value locked up here? Could we charge them for this? Could we force a payment, rather than doing the click-through agreements? I just thought we would get more out of it by putting it more out there. It's ultimately a cost model question. We ended up in the pharmaceuticals business – we took what we wanted, and everything else, was sort of on its own. It's more positive from a scientific standpoint, and neutral from a monetary standpoint on balance, you should aim for scientific value, because it is the right thing to do.

This discussion is a compelling argument for data sharing, but in a protected and calculated kind of way; the sharing is done under legal agreements that dictate that a monetary exchange will occur if something valuable is found on top of someone else's work. Unlike in the academic calculus of rewards and publications in exchange for data sharing, the great leveler here is the legal technology of the patent itself. The patent may not define the actual value of the data, but it ensures that the value is realized if sharing leads to something good.

This model also reflects the most interesting mix between the norms and counter-norms when considering scientific data sharing. Here, we have patented data that is distributed in a communal style, as broadly as possible, for no fee at all – an apparent "gift" to science, open to objective analysis by whoever might use it, with no pre-supposition of what that end result might be – a seeming tribute to the Mertonian norms of disinterest and communalism, as long

as you accept the click through acknowledgement of the rights held. On the back-end of that gift, however, we have commercial interests in whatever you happen to find out, and the lawyers to back that claim up. Merton picks up Marx and a legal team on the way to drug discovery.

Not every commercial organization takes this approach. One very valuable piece of cancer research IP controlled by a commercial organization is OncoMouse®[12], a genetically altered research mouse line that is particularly susceptible to cancer, making it attractive for cancer research. The DuPont website reports: "OncoMouse® technology[13] is in wide commercial use. Free academic licenses have been executed with nearly 300 nonprofit universities and research institutions worldwide." DuPont also holds a Memorandum of Agreement with the NIH for non-commercial use of the mouse. Non-academics, however, points out one commercial researcher, pay a large fee to license OncoMouse use, one that helps DuPont maintain a significant competitive advantage.

4.3 Public Goods versus Patentable Goods: Sensemaking of a Case Study

The role and use of patents in science and technology post World War II is well travelled territory in STS, and is not the point of this research. Rather, the question here is how the social level discourse about data sharing, data ownership, and patents intertwines with personal opinions and decision-making related to data sharing. Shared social stories help to shape those personal stories and sensemaking, and there is no more popular story for academic cancer

[12] OncoMouse® is a registered trademark of E. I. du Pont de Nemours and Company or its affiliates.
[13] Note the use of the word technology; a living organism is advertised clearly as a technology for scientific research, both blurring the lines between science and technology in cancer research, and clearly objectifying the licensing of a living being.

researchers on the "bench side" (genomics and proteomics, primarily) than the Human Genome Project. This "community resource project" – not just "Big Science" but "Mega Science" or "Huge Science" - was referenced often in interviews. While there were certainly clinical and privacy questions associated with the Human Genome Project, it was the messaging about data ownership that was referenced most often in this research.

The Human Genome Project has been well documented and analyzed; this research is not concerned with reanalyzing the history. Rather, this project is interested in how cancer researchers describe that project in reference to their own work today; and how that shared story impacts the way they think about their own research. One interviewee provides an overview of the project, flavored with personal commentary, to initiate this discussion (interviewer questions in italics):

> The prevailing and overall view in genomics is completely open data sharing, and there are a number of things that have promoted this. Number 1, we are talking about in order to reproduce analysis results, one needs access to the underlying gene sequences. The questions early on about the patentability of the sequence of genes I think played into that, but also, the culture of places like the NIH Genome Research Institute, and the Human Genome Project where these guys pushed openness as far as it has ever had been before. *How did they do that?* How practically? Here's an example. When they would do a DNA sequence run, the raw files, along with the base codes, along with the finished sequence, everything was made available together on the Human Genome site and in GenBank. So, number one, the capability was there to share this data broadly; all you had to do was upload the data.
>
> Number 2, the culture was pushed very strongly, to say this is an open activity, and a little bit, I think folks saw themselves as fighting the good fight, keeping science open, as opposed to companies like Celera and others that were intent on privatizing the genome, making it through patents and so on, restricted access. And I think the scientists thought it was the wrong thing to do from a scientific perspective. And the way they fought against that was to make their data open and accessible, and it in fact, it worked. There may have been doubts about that in 2000, but by now it is clear. They took the air out of private ownership of the genome, and in some sense, and again in my personal opinion,

> thank God they did. No one could have afforded to do the genome twice. And this was what Craig Venter [lead for the commercial effort competing with the NIH public project] was always saying, "We are going to do it, we are going to do it first, and we are going to have it, and no one will bother to do it again. And we can charge whatever we want." Look, that's sort of, a somewhat skewed representation, but I don't think it's actually that far from the truth. So here you have an environment of data sharing.

Many in genomics today saw the controversy, called by many in both the social and personal worlds the "Genomics War," play out in real time with emotional reactions, and it has impacted how they "make sense" of data sharing questions and their own choice to share. On one side, the NIH team that funded scientists at centers across the country to generate well-defined, publically-available, free data about the human genome. On the other side, there were the commercial interests represented by a company called Celera, taking a different and less expensive analytical approach to achieve the same end, and proposing a subscription service to the genome that could be used by researchers and commercial interests to conduct the same types of research as people would with the public data. In the end, while both sides published their success in a joint release, the government was seen to have won the war, both scientifically and philosophically; and biotechnology stocks took a considerable hit.

This controversy positioned the ideals of the norms of science advocated by the public project ahead of the more materialist approach (subscription service) advocated by commercial interests, and also led to controversy about how articles related to the project were selected for publication in the leading journals *Nature* and *Science* (Marshall). In the midst of the controversy, the Celera leader, Craig Venter testified before Congress to clarify the company's motives for the public (Venter, "Congressional Testimony"):

> I would like to address the confusion that has arisen over the accessibility of our data, in particular the accessibility of our data on the human genome. We have and will continue to react to claims that Celera intends to withhold information and delay progress, particularly when our fundamental mission is to accelerate the dissemination of high quality, accurate information. Let me emphasize--our data on the human genome is currently available to those subscribing. Our vision is that the list of subscribers will be very long. Let me draw an analogy Mr. Chairman. When you pick up a newspaper at your doorstep, you consider it quite accessible. You probably do not even remember that you are paying a subscription to have that access and you certainly don't claim that the newspaper company publishing is being secretive or restricting access to news about current events just because you pay a subscription fee.

This testimony brings clarity to the difference between being available under a patent, and being available for free. This is not a question, necessarily, of sharing data or not; it is a question of how access is paid for. Clearly, the word "sharing" complicates the debate; we learn early in US culture that sharing generally doesn't come with a price tag. Given that, it is uncomfortable in the normative ethos of science to start talking about subscription services for the human genome, as if somehow the work is science is similar to the work of your local newspaper outfit.

An editorial at the time captures the tension of intellectual property and data withholding issues "invading" the purity of science, highlighting several (at the time) recent examples of the negative impact of such materialist concerns in academic science. Rather than an institutional condemnation, however, restrictive behavior is categorized simply as "not wicked, but tacky." It is an interesting choice of words, framing the delivery not as an objective evaluation to be debated among peers, but rather as a more informal, almost degrading reproach.

> I decided to list some recent actions or events that qualify as tacky….. Refusal to share materials used in a published experiment is not a sin that journals (or

> anyone else) can punish effectively. But it is a particularly tasteless exercise of scientific competitiveness, and it is reaching epidemic proportions…. "Everybody does it" is a familiar excuse, but it's still tacky.
>
> The exchange of materials has been made even more difficult by the institutionalization of Material Transfer Agreements (MTAs): paperwork required by universities and industries to accompany the cells, reagents, etc. Not only is this helping to shrink the "knowledge commons" that was once an academic feature, the MTAs may contain provisions that act as deadfalls. Investigator at A reads B's paper, likes it, gets his cell line. A does a nice experiment and publishes the results; C asks for A's cells, but A then learns that B's MTA (which of course he had been too busy to read) prohibited redistribution to scientists in industry. C complains to the journal that published A's paper, but it can't do much about this perfectly predictable but tacky result.
>
> …Anything that has the look of a publicity stunt or of self-interest takes away from the credibility of the process and the reputation of the scientific endeavor. …what we have here is a growing list of behaviors that, taken together, exemplify the gradual retreat from generosity and straight dealing in a community that is usually known for those qualities. Perhaps the core element of "tacky" in these examples is that they all eat away at the sense of community, shared understanding, and public trust that are crucial to science (Kennedy 1237).

The phrasing here suggests that this editorial was working to reassert the Mertonian norms of science, a science that values openness and the communal sharing of ideas. It does this assertion not by debating the true issues involved, but by simply writing off the alternative views as “tacky.” This is a subtle but effective means of maintaining the public perception of the boundaries of science as the disinterested objective field the public would like to think it is; dismissiveness allows science to keep the high road without getting into the messier aspects of why the withholding being commented on (the tacky behavior) is being done in the first place. The message to other scientists is clear: here are some examples of tacky behavior; if you too are taking these actions, we think you are tacky too.

Today's academic researchers navigate somewhere along the continuum between the Human Genome Project's fully open "sharing right off the machine" philosophy, and the more pragmatic materialist withholding that was deemed to characterize the Celera effort. It is clear in interviewee comments that the researchers saw specific benefits in, and specific drivers for, the Human Genome Project. It was, however, a special case. Interviewees cited the project as a "game changer" that helped prove the success of bioinformatics and shape the next steps of genomic science. It was also considered so large in scope that it could not be done *without* the open sharing that took place. The Human Genome Project was so large and so significant; it required open sharing as a baseline of knowledge production that many other projects could build on.

This singularity of the project, however, also sets its status as an ideal case. This leads researchers to note quickly that it is not a representative model for the way things actually work. Recall, may people agree with the *ideal* of sharing data, but when it comes to the practical aspects of it, the objections about timing begin. Researchers may not mean to be tacky, as suggested by the editorial, but they do have certain practical career and funding interests that may lead them not to post pre-published data for others to benefit from before they themselves can. For the Human Genome Project, scientists were *paid* to generate human genome data and post it immediately using very well-defined and easy to accomplish deposit processes; recalling the previous discussion, they were essentially paid to participate in what was constructed and managed as a modernist endeavor. The universal processes, technology, and rewards of this "Mega Science project" were easy to understand and do; and reflected a very different kind of scientific practice than many now operate under on a day-to-day basis.

Often described as a "race to the starting line," now that the project is complete, the longer-term and more post-modern marathon appears underway.

There is a difference between a culture and an ideal. While several interviewees noted a "culture of openness" in genomics and bench science in general, the experiences shared suggest more that this openness is a goal towards which the community is striving: an ideal type, as it were, rather than a cultural description. The Mertonian ideal of openness as currently accepted was not inevitable; had the Celera commercial approach succeeded, some noted, it would likely be a radically different landscape today.

An ideal is a vision, a goal. Culture, on the other hand, is a reflection of where a social world is right now: the collection of structures, practices, beliefs, expectations, and consequences that ultimately identify who we are at this point in time. An ideal is a sense of shared future; culture is a collective identity. Legal structures are one element of a collective identity; the question remains, how can they also help shape the path to the ideal?

4.4 Copyrights and Licensing: An Absent Discussion

The research to this point suggests the need for different kinds of incentives to motivate academic researchers and universities to share data. Unfortunately, there is a significant range in the recognition continuum between authorship on paper based on data contribution, or the citation of a data set in a bibliography, and the commercial power of a patent, which is not always "the point" for researchers and universities. What is needed is a way to link a data set to an author to create an institutionally recognized process and tool whereby the author gets downstream credit for sharing data, while also allowing that sharing to happen for free (e.g.

non-commercially), as is the most frequent model in academic research. How might a middle ground support the needs for greater clarity in data ownership, while also enhancing the non-commercial sharing of these data in academic environments? What would such an intermediate tool "look like" in terms of an institutionally recognized instrument?

A topic that, with one exception did *not* emerge in research interviews is the use of copyright and licenses in providing a clearer sense of ownership for the people and institutions generating data that could be shared. The caBIG program for example, despite being a federally funded program that advocates open and free access to its software, has copyrighted both its name and logos, and attaches a license to each piece of software the program generates. Use of the caBIG copyright is given to licensed organizations that undergo a negotiation process with the NCI. With the exception of "fair use," such as media usage for reporting caBIG in news stories or research projects like this one, caBIG is a restricted term.

What does this have to do motivating data sharing in cancer research? At this point in the discussion, perhaps nothing. STS, however, encourages us to examine not only the practices and discourses that are visible in exploring a research area, but also the elements that are not. At its most fundamental level, copyrights are used to identify and protect the ownership of works that express creativity; licenses allow for the distribution of rights associated with that ownership. Since copyrights and licensing tools exist for the express purpose of *allowing* sharing in a protected way, and are in heavy use on the commercial side, its absence in the discussion about data sharing in academic cancer research is somewhat striking. Several interviewees noted the need to be able to track the usage and impact of data generated by cancer researchers; and to have that use "count" in career evaluation decisions,

so that there is a reward assigned for such sharing. Only one interviewee mentioned copyrights and licensing as a tool to support this. Why do these tools appear to be so absent from the dialogue?

One immediate answer may lie in the limitations on what can be copyrighted under U.S. law. One of the difficult aspects of scientific data ownership and protection is that facts and concepts themselves cannot be copyrighted, and furthermore, research is considered "fair use" of data. As such, in the purest sense of the word data, data can be owned, but cannot be copyrighted or licensed unless embedded in a form that includes some form of creative expression. In one sense, this would seem to eliminate this legal realm from consideration when it comes to cancer research as much of the data considered so far, biospecimens, hypotheses (concepts), and facts (such as the human genome sequence), cannot actually be copyrighted in the current system. On the other hand, it is clear that some kind of mechanism is needed to assign more "ownership-aware" rights in the practice of sharing data, and the rewards to accompany such rights and transfers.

The bracketing of data sets as outside the boundary of what can be copyrighted deserves to be unpacked. One of the central proposals of this research is that data generated in cancer research is a personal and subjective output of an actor-driven research process. A dataset is ultimately a personal expression of both knowledge and creativity; Hans Joas (1996) would likely argue that generating cancer research data is an action that in itself is generative and creative. Why, then, would a unique collection of facts in a data set not be legally seen as a creative work, worthy of copyright protection? Ultimately, this question recalls the earlier discussion of the modern-postmodern divide; by positioning data as objective facts, that data

becomes separated from the researcher; it becomes a public good, rather than a personal expression deserving of protection. From activist perspective, this calls for new legal structures that allow local knowledge in the form of data to be both acknowledged and protected for what it ultimately is: the creative and original expression of the researcher(s) generating it.

Literature on the institutional rewards and interests of science focuses a great deal of attention on "gift giving" in the form of scientific publications. Ironically, until recently these "gift delivery mechanisms" (scientific publications) were almost universally controlled by commercial organizations, making a profit on the deposit of "gifts" deemed worthy of inclusion. Licensed non-commercial open access is a relatively new and still controversial concept. In fact, the emergence of open access academic publications, such as the Public Library of Science (PLoS), has only occurred in the past decade. Supported in part by an emerging legal tool called the "creative commons" licenses, published in 2002 by a non-profit organization called Creative Commons, open access publications call into question the traditional role of commercial publishing houses in coordinating peer review and publishing subscription-based journal articles that have been the lifeblood of academics (Editorial Staff, "Let Data Speak to Data" 531).

Open access journals are redefining both the peer review procedures and the economic structures associated with the sharing of scientific publications. A 2006 editorial in PLoS (Public Library of Science) declares:

> Most science is not published in Science, Nature, Cell, or even PLoS Biology. Indeed, the increasing pressure of submissions, limited page budgets, and the existing reward system by which the value of a paper is placed not on its content but on the venue in which it is published has led most journals to reject a substantial fraction of papers before peer review. The reasons given for rejection are various: the editors may claim that the paper is beyond the scope of a

> journal, too specialized, of insufficient general interest, or lacking a sufficiently novel advance—even too complicated. The basis for such decisions is inevitably subjective.
>
> But in just a few weeks, the Public Library of Science will launch a new "journal," PLoS ONE (http://www.plosone.org/), that will initiate a radical departure from the stifling constraints of this existing system. Its aims are not only to provide a more inclusive open-access platform for scientific literature—papers will not be rejected on the basis of such subjective justifications as those invoked above—but to reflect far more closely the way that scientific research is conducted by taking advantage of the increasing functionality and flexibility of internet-based communication. All papers that make a valuable contribution to the scientific literature, that are replicable, that are clearly written, and whose conclusions are supported by the data deserve publication. PLoS ONE will provide the means to do that swiftly and efficiently (MacCullum e401).

Publication services such as PLoS retain the step of peer review, but reviews are structured in a way that incentives publication over rejection. Not everyone is supportive of this new approach. A 2007 editorial in *Nature* comments:

> A radical project from the Public Library of Science (PLoS), the most prominent publisher in the open-access movement, is setting out to challenge academia's obsession with journal status and impact factors. The online-only PLoS One, which launched on 20 December, will publish any paper that is methodologically sound. Supporters say the approach will remove some of the inefficiencies associated with current peer-review systems — but critics question whether a journal that eschews impact factors will manage to attract papers. PLoS One faces some significant challenges. Many new journals struggle to attract papers until they are given an impact factor (a measure of the citations its papers receive), but a journal that accepts everything can't usefully be classified in this way. Critics also point out that referees may be reluctant to review potentially trivial papers, and that existing journals have had little luck persuading readers to comment on papers after publication (Giles 9).

The open access publication movement is consistent with caBIG's philosophy, but focuses on the narrower realm of scientific publications, rather than the broader sharing infrastructure explored by caBIG. A broader movement to encourage greater access to a

broader set of scientific materials is being facilitated by Science Commons, an extension of the Creative Commons organization initiated in 2005. Science Commons describes its mission as follows:

> Science Commons designs strategies and tools for faster, more efficient Web-enabled scientific research. Primary Focus: (1) Making scientific research "re-useful" — We develop and promote policy and tools to help people and organizations open and mark their research and data sets for reuse. As part of this work, we released an "open data" protocol to enable the global scientific community to pool and use data created under different legal regimes. (2) Enabling "one-click" access to research tools — We offer a suite of standardized contracts to bring the efficiencies and economies of scale from e-commerce to the world of scientific tools, so researchers can easily replicate, verify, and extend research. (3) Integrating fragmented information sources — We help researchers find, analyze, and use data from disparate sources by marking and integrating the information with a common, computer-readable language (Science Commons, "About Science Commons," Par. 4-8).

Although representatives from the Science Commons assist the caBIG program in developing data sharing-related white papers and articles, licensing tools being advocated by entities such as Science Commons were only once raised in interviewees on data sharing. Additionally, this discussion is not intended to suggest that introducing new licensing techniques will automatically encourage data sharing. Data sharing remains a personal activity; copyrights and licensing simply help protect the individuals that choose to share. The personal motives that inspire this activity are considered in the next chapter.

Chapter 5: Personal Sides of Data Sharing

Chapters to this point have focused on the technical, economic and legal aspects of data sharing, framed primarily through social messaging, and highlights of the gaps between social themes and personal experiences. This chapter places these social factors in the background, and focuses on more personal aspects of the question of whether to share data or not, and how this question is navigated and felt about in practice. To begin, a paper in 1990 foreshadowed the theme expressed above in the "Not Wicked, but Tacky" editorial:

> The high-sounding principle to which all scientists pay obeisance, that science and its data are public, is not always honored in practice. The ability of scientist A to link to scientist B's data is influenced by a large number of economic, social, and political factors which will likely continue to determine the willingness or unwillingness of one scientist to give another scientist access to his data (Sterling 116).

This quote, which is fairly overt in claiming that "scientists do not always honor principles," sets a tone that is critical of scientists that are not willing to share. Eighteen years later, a 2008 New York Times essay by a researcher at a prominent NCI Cancer Center echoes the judgment:

> …Most scientists doing research on how best to help those in pain, or at risk of death, want to keep their data a secret….Their reasons were entirely trivial: one cited the difficult of putting together a data set (wouldn't this have to be done anyway in order to publish a paper?); another was concerned that the data might be analyzed using invalid methods (surely a judgment for the scientific community as a whole). This is something of a clue that the real issue here has more to do with status and career than with any loftier considerations. Scientists don't want to be scooped by their own data, or have someone else challenge their conclusions with a new analysis. Yet this is exactly what cancer patients need. They want new results to be published as quickly as possible and to

encourage a robust debate on the merits of key research findings (Vickers, Par. 8-10).

The claim here is that the "trivial needs" of the career of a scientist are less important than loftier needs; that scientists are selfish when they put their career ahead of people in pain. This places scientists on the defensive, because it positions an "either-or" dichotomy: you are either interested in your career (so you don't share), or, you are interested in patient care (so you do). This kind of positioning removes the "both-and" possibility: that with the right structures and incentives, advancing careers could in fact coincide with sharing data, and through that, helping patients. Instead of targeting those institutional questions by calling on Deans, regulatory bodies and funding agencies to drive systemic change, these publications target the researchers who are acting out their roles in a larger political and social process. The social becomes personal in these public exchanges, but it unclear that these types of commentaries actually shape the change that the writers clearly wish to see. "If we berate the researchers enough, perhaps they will share" seems to be the strategy here; it is unclear that it is effective.

A more nuanced editorial admonishment was published in 2001 by a regular Cell Biology columnist named "Caveman," who frequently writes on topics highlighting the divide between the norms of science and actual scientific practice.[14] In this particular editorial, "'Send me all of your reagents and ideas. We want to work on the same experiments,'" Caveman highlights the typical objections to sharing – losing competitive advantage, taking away time from experiments to invest labor in the preparation, and risking misuse of materials – and then

[14] The "Caveman" view called out in the typical column is generally one of an "old school" selfish sensibility that does not support the evolution towards a clearly Mertonian scientific ideal.

strongly advocates, “send the reagent immediately.” Despite the rebuke of those who ask for materials inappropriately (captured in the title of the article), the writer is clear as to where responsibility should lie; it is with the researcher who is asked to share: “Do not ask questions, do not demand a collaboration or co-authorship on papers, or restrict the work that will be done by the other group. It is not easy to make this (right) choice of response.” (Caveman 1038)

Starting from the values-laden term “data sharing,” researchers are put on the defensive if they do not. This chapter dives more deeply into the personal decision-making of researchers, revealing a more complex set of variables than the social judgments of “selfish” or “selfless.”

5.1 Data Sharing as a Service: The Investment and Extension of Self

Setting aside the legal ownership of data considered in the last chapter, the first step in a more personal analysis of data sharing is to consider the relationship between a researcher and his or her data, as this positioning helps anticipate how sharing of that data will be viewed, beyond the rewards that can be accessed by sharing that data. One source of data, such as a tissue sample or image with the personal demographic and treatment information associated with it, can enter into a pipeline of research that is ultimately used by many people, from clinical researchers to bench scientists working with de-identified derivatives of that sample that lead to molecular analysis outputs. This cross-utilization potential reflects a change that has occurred over time; the implications of this are considered in the following quote from a May 2009 article about data sharing in genomics:

> **Acknowledging individual contributions**. In the past, a data set would have been used primarily by the researchers who had created it, and would provide the basis for many publications. There would have been a direct relationship between the creation of the data and control over usage and the publication of results. However, with data sharing policies, the fact that particular researchers have created a data set no longer gives them an enduring priority or control over its use and resulting publications. The challenge then is how to reward and acknowledge the production of a data set (Kaye et al 332).

There are different ways to feel in control of or connected to a data set. For clinical researchers that have been involved in enrolling study participants, there are both legal and ethical motivators for ensuring that data is kept in tight control. One interviewee noted that responsible clinicians take ownership of the protection of their participant's information. The ethical position of such a scientist may be *not* to share, because sharing would mean a loss of control, and possible subsequent use of patient data in ways that they would not support. For clinicians, there is also the question of how much metadata to request patients to provide for possible future studies. More contextual data about collected tissues may make the tissue more valuable, but it also takes more time from an often suffering patient and represents more data that must be captured and managed for no immediately clear meaningful use. For these researchers, the relationship with the data is through the patient.

For bench scientists, the relationship with the data is different. By the time materials or data come to these scientists, the tissue or other materials (if any is involved) has often already been processed and even transformed in some way, perhaps even de-identified (if it is even human at all). For these scientists, the relationship is not through the patient, but rather through the patterns and in the very structure of the data. Says one scientist, "People look for things that they recognize in the data – it's how they get to know it." Another interview notes

that the structure and type of data in a file could signal the researcher's hypothesis or question, revealing important intellectual work in the form of a dataset. In this way, data sets are personal and subjective creative expressions; different people may see different elements of interest depending on the question they are asking, their background, and training.

Personal worry that a researcher might get "scooped" by his or her own data becomes prominent in this discussion. Many researchers used this metaphor: much as a reporter gets "scooped" by another reporter that gets the story first, they fear getting "scooped" because someone will see something in their data that they did not see. This is a deep-seated fear that interviewees admitted is only partly acknowledged by the "published versus pre-published" data release debate. Certainly, the risk is higher with sharing pre-published data: if someone sees something special in the data, it could lead a second researcher to get a finding published earlier than the data's original owner.

In the competition for grants and publications, this is bad enough; however, there is a deeper potential embarrassment that lies just beneath the surface. A researcher's reputation lies in the results generated from understanding his or her own data; someone else finding a pattern in data that one was supposed to know well, whether is pre-published or after publication, is a potentially negative commentary on one's skills and abilities. A data set is an extension of scientist's intellect and identity. When there is the possibility that one could be scooped by one's own data, one's intellect could be called into question by others: "Well, gosh, I can't believe you missed *this."*

This was not a connection that interviewees shared independently, but there was agreement once the idea was said. One researcher made the point clearly through her

resistance to the "scoop" concern, "People should not be scooping you on your own data… If you know your own data, if you have done the cleaning, the analysis, no matter what you do, no one is going to know your data as well as you do. I don't think it is as big a problem as people represent it to be." The nuance is important; the deeper suggestion here is that if someone *is* able to scoop you, then you clearly have not done the work needed on your data. The stakes are high.

Data elements are not just numbers in a spreadsheet; they are expressions of personal knowledge that are controlled as an extension of the researcher him or herself. Deciding to share is an expression and measure of personal autonomy. This is not just a question of the rewards that come from sharing with others, facilitated through technology. The data defines the value of the researcher. One researcher summarized the idea at the end of our interview:

> It's a question of value. What's really the value of a scientist? How many publications they provide? What kind of data do they provide? What's the measure for evaluating their contribution?

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. This investment leads to hesitation to share, even among those that are the greatest proponents of sharing. From a researcher who has authored papers on the importance of data sharing (interview question in italics):

> I advocate data sharing, and make an attempt to always share my data, because otherwise I feel hypocritical, but I myself experience hesitations, so for completeness sake I should say that. *What are your hesitations?* It takes time, my data isn't clean and neat, and my code isn't clean and neat, so it is a little embarrassing to put it out there. It's not always organized, so it takes time to pull it together in a format to put it somewhere. And since it is esoteric data,

> there aren't well defined ways to present it, so I have to make decisions, which I put off. [Speaking more slowly and with regret] It's not easy, not even for someone who thinks it is very important and that it should be done. At some level, I would rather not, but then I do it anyway, though I am late in putting it out. But I feel guilty about that. [Sighs]

The same person reported an experience that she had with an established researcher when she informally spoke with her about her data set:

> There are prominent researchers out there that have spent years and a lot of energy compiling very large and important data sets, so large that they couldn't possibly examine everything there is to be examined in there. I asked them if they were interested in sharing that, and they said, "No. No, we are not. We are using it. It's our blood, sweat and tears, and eventually, we will be forced to, and then we will do it, but we're not going to do it before then. It's our bread and butter." They were actually insulted when people asked them for access to their data, because they had spent so much effort and energy creating the data set, and they viewed that as the "not fun" part of their job. To have people swoop in and want to do the just fun part, they found fairly insulting.

All of these examples drive home the point: in systems that reward through publications, grants, and patents, the time involved in both originally generating and sharing data is not time that is rewarded. Sharing data may earn "points" in supporting the ideal of science, but that general framework does not translate well to the day-to-day sensemaking of task and career management. This is a tension: scientists need to share data to appear to be "a good guy," but data sharing is also a service that takes time and resources to support, and may result in a real loss in investment energy and material; there are few rewards for that for those on an academic research track.

This reality was true across the research population, from clinical scientists to bench scientists to bioinformatics tools developers. As an illustration, a bioinformatics researcher has spent much of his career creating tools that support the infrastructure of today's biomedical

research: the software tools that allow other people to get more grants, do analyses faster, and so on. This researcher is pragmatic about the fact that maintaining the tool and making it available to others, a form of data sharing from an infrastructure perspective, does not help his career progression on its own, even though it is a service that supports many others' careers. He nonetheless continues to develop the tools, but compensates for this "unrewarded service" by publishing papers that document the tool and its benefits for those in the field.

There are both benefits and downsides to this. One of the downsides is the investment it takes to help others when they find out about the tool through his publications. The following comments illustrate both the cost of the investment for him, and also the effort needed at the receiving end of the sharing in some cases.

> So, I go out to XYZ Cancer Center, and I'll give a talk and a demonstration of some system we have built, and then people come up to me after the presentation, and they say, "So, is your software available?" And I say, "Absolutely! Absolutely!" And so I get back and I gather it all into a big file, all the software, the directories, the source files, tar it all up into a big file, and put it as a link on a website, and say, "Here it is!" [Laughs] "There's no documentation, there's no user manual, there's not even any installation instructions. Good luck! I hope you have some really sharp Unix people, because there is no shrink wrap, there's no one button installer." There's a lot of material in there that is specific to our use of the tool and data.

> Sometimes I feel like we are sharing things, but what we are sharing is so hard to use in the format given. Why is that? Was it because we were trying to make it hard for others to use? Not at all. It's that the cost of producing tools and data that are well documented and shareable and usable across organizations is just beyond what anybody would guess. How much do you think it would cost to take this thing and make it shareable? And usually, people are off by a factor of 10. Making it shareable is expensive. Standards for shareability are pretty high.

It takes time to help researchers install the tools that will help their work; this activity adds value to science because it supports the analyses that lead to new knowledge, but the

scientific labor itself is invisible because it does not "count" in the scientific reward system. Data sharing, in many forms, is a service; and yet, despite the fact that science is ultimately framed as a pursuit of knowledge (a service for the greater good); it is actually fundamentally artifacts and credits driven.

There is even invisible labor in the service of connecting people with each other for research projects; the catalysts who build the relationships that subsequently encourage sharing. Even this form of human networking labor is not overtly acknowledged if it does not lead to some form of publication credit. Says a population sciences researcher:

> If you are not lead author, because you are always one of the people who, you are a networking person who is able to bring people together to have a project happen, it wouldn't happen without you. But you aren't the lead author? After a while, over time, your visibility drops, because you are not really being seen as doing a lot of work, and yet that's such important work. That's one of the publication and tenure-related disincentives to sharing. There's work to be done out in the field to re-incentivize data sharing.

Becoming scientifically intimate with one's data, and investing the labor and energy to prepare it for others; these are tasks that take a great deal of time and specialized skills, and yet are part of the invisible work of science. In a social world where outcomes of paper authorship and grants are the official currency, the *service* to others and to science that is captured in the practice of sharing goes mainly unrecognized. Data itself is subjective; it takes personal energy to generate, and has specific meaning to the researcher generating it. Just "putting it out there" on institutionally-shaped technological grids leads to the perception of the loss of person-specific knowledge, and an investment made, but not rewarded.

The current incentives for data sharing come from the quest for individual achievement; yet, when probed more deeply, it becomes clear that for researchers, the actual motivation to

share comes most often based on a relationship with another person. The ideal of science may be "sharing for the greater good," but in the experience of many researchers, the reality is, "I will share for the right person."

5.2 Data Sharing as Relationship Building

Almost all the researchers interviewed for this research suggested that data sharing is currently, in their eyes, first a values-based subjective choice; there is more often than not a specific trusted target for data sharing, and a collaboration that forms the foundation for the trust required to share. Once that is established, the researcher turns to the technical and legal issues. One legal professional was clear: "They will share their data with who they want to share it with." This theme is captured by an expert that helps scientists navigate the legal side of data sharing, when asked why people generally say they want to share data:

> The most common reasons for wanting to share data? All over the map. It ranges from people wanting to be collegial; I have a data set and so-and-so needs it for their research, to people who are participating in formal multi-site networks, to people who wanted to send data to a manufacturer or to NIH or to someone else who might be willing to give them a grant.

Despite the "all over the map" characterization, there is a common thread within these answers, revealed in a follow-up question (Interview question in italics):

> *When you think about the reasons that people come to you, is it "I want to share with a specific person or multi-site network," versus, "I want to make it generally available through this system for other people to access…?"* It is much more the former than the latter – I think that caBIG and projects like it are in their infancy. I don't see a lot of "I want to make this generically available…"

This was echoed and extended in an open panel on data sharing, where one expert noted, "Data sharing happens the fastest in consortiums – people taking themselves out of university systems and engaged in an enterprise beyond the ivory tower." This is a reference to a consortium as a specific "social world" in which sharing happens; it was also a reference, however, to the perceived limits of the "ivory tower," a topic already discussed in depth. Another researcher in a conference setting shares that, "We've had good luck in data sharing in groups, because there is trust between them." Another conference speaker put it most simply, "Person-to-person bandwidth is low – but ultimately, it is trust that connects us."

Materialist or rational actor exchange theory views would argue that value is increased and transaction costs (invisible labor and risk) decreased if a researcher works with other researchers that are known and trusted. The comments from researchers suggest, however, that there is a more elusive personal element involved here; it is a calculus, but it is also a connection. Interviewees capture the theme (interview question in italics):

> *What are the conditions under which data sharing happens?* I think that in some ways, it's dependent on the nature of the data, the nature of the institutions and the nature of the investigators. I don't think interpersonal relationships can be underestimated, or that is, the importance of them cannot be overestimated. This might be my personal bias, but it's a lot easier to get along with people who are nice to you. If you have a good relationship, it is just human nature that you would want to help them out, especially if you get help down the road, like your name on a paper or invited to do an abstract. Motivations for data sharing can be reciprocal help, even if you don't get a publication.

A molecular biologist talks about a time she shared data with others, based on a personal relationship her PI had with another laboratory (interview questions in italics):

> As a graduate student, my PI collaborated with another PI at a lab in France. They were like best buddies, and so, people from our lab would go over there,

> they would come to the U.S. There would be exchanges of certain sets of information…. There's a potential to be scooped, sometimes people don't want to be explicit about their research questions….. here, there was this sense of trust, there wasn't a lot of hiding. *What led to the trust?* The personal relationship between the PI's. Yeah, that's it. When you are in a lab together, you aren't going to publish without me, because then we would have to live with each other every day.

A researcher who has significant concerns about the viability of broad-based grid technologies for cancer research talks also about the conditions under which grid data sharing might occur (interviewer questions and comments in italics):

> You need the established community to make it worthwhile. You need the guy in California to want to see what the woman in Oregon is doing. You need a problem that has a high degree of connections between the nodes…. There's a more subtle point, one of the reasons I think caBIG is hard for people is that it is so enormous. It is much more practical to think about a grid of grids that focuses on specific tasks. I think that's a real barrier to people being able to use the technology, because they don't immediately see how it can be used to support communities, as opposed to the entire world. *So, an example like the Prostate SPORE [a community of researchers working on prostate cancer].* That's much more reinforcing for people. My friends are on the prostate Grid, I should be on the prostate Grid too.

Another scientist with the same concern noted:

> *How do you find out who might be interested in your data?* That's part of networking, when I am at scientific meetings, and people are talking about their projects, and there's a logical next step – I am sitting on some data that could help them. That leads to collaborations, but it is that same process that helps you figure out what people need, what people want, what would be helpful.
>
> *What is your reaction to the caBIG model of data sharing, on the grid?* It's certainly a non-traditional model, in that once the kinks get worked out and it is socialized, it would streamline the approach to giving data to medical students, or to high ups in societies, with lower cost to me. There is a lack of incentive to do that, because there isn't that person-to-person collaboration, or earning brownie points. For investigators that have a relatively low threshold for sharing, it does make things easier. It's good for stuff you were going to share anyway. My thought process hasn't really changed enough culturally to have caBIG pop up in my mind when I think about the kind of data sharing I do.

> *What signals would make that cultural change more clear, what would help make it pop up your mind?* The incentive system would need to change, more NIH requirements, the current Data Sharing Policy is not so clear.

Another researcher goes further with the resistance:

> Why should I share my data? What does it buy me? It's such a headache! People I work with are pretty open about it. They don't understand what is in it for them, which is a very legitimate question. When you are up against the wall, people do not want to do extra work without clear benefit…. Grid data sharing is absurd, it's even insulting. I could never get into the notion that this kind of level of automation in interoperability was ever really possible. There is so much that is indefinable. People need to be able to communicate. It's all about personal interaction, where people develop a sense of trust, where it's not just someone who's going to take my data and publish it. I need something back for it. Data sharing is like dating, in the sense that it is fine for single instances, but what you are hoping for is the scientific marriage, where you have longer term collaborative goals being met. There's an exchange that goes beyond the data. There's silliness in the idea that we can take people out of this; it is the people that are the motivators for this. And they have wishes, needs and desires that go way beyond the data.

Another clinical researcher talked through the process of deciding to enter a data sharing relationship:

> There's a personal factor involved…that sorts of says, do I trust this guy? Does the work he's done seem to be upright and forthright… has he performed appropriately? You sort of get that feeling that, yes, I would trust this guy. I will let him borrow my car if he needed something, along those lines, kind of. I would be much less willing to just blindly share that kind of data at that stage with just anyone.

This subjective knowledge, or lack thereof, has also led him to not share data. He acknowledges that he has not shared data with "very young people, people who had not published, hadn't done anything is the field… I couldn't find out anything about them. It's one step away from the people calling you on the phone asking if you will give them your social

security number… No, I really don't know you, don't know who you are. I'm not comfortable at this point sharing with you." What would motivate him to share data in this case?

> They need a cosigner….somebody at their institution or someone they are associated with who has established a professional reputation, who will basically vouch for them: 'I know little Johnny here; I know him, he's working with us… we are confident in him.' Yeah, if I know I have someone who will vouch for Johnny, then, yeah, I'll go there.

Often, social messages about data sharing are focused on the incentives to the giver; the personal relationship aspect reframes data sharing more as a collaborative partnership, with both a giver and receiver. The giving aspect of data sharing by scientists is largely shaped by the person or entity with whom this data is being shared.

Data sharing does not begin with the grant and the publication; it begins with the interpersonal connection that sparks someone to collaborate and/or share the data that leads to those publications in the first place. According to the group interviewed for this research, publication decisions are actually made at the beginning of the data sharing relationship, including even agreement on who the lead researcher is (and who the lead author will be); this is part of the ritual dance that establishes the relationship. It is also clear that if a collaborative relationship has not been pre-defined, sharing is unlikely to happen before that step is taken.

This emphasis on the relationship tie seems obvious on second look given that the research is about sharing with others; yet, the social discourse about data sharing in cancer research is largely silent on the "recipient" of the data. Rather it emphasizes the benefits of "gift giving" of data to the general (and given its scope, necessarily impersonal) cause of science.

Interestingly, the only example of when "gift giving" was given as a central motivator for sharing data was when researchers talked about sharing data or materials for educational purposes, as part of the development process of the next generation of scientists. This included sharing between graduate students, or the donation of tissues that were not going to be used to an educational program.

Two researchers active in caBIG actually reported that the caBIG program, despite its focus on impersonal grid sharing, was ultimately successful long before success seemed certain *because* of its actions to bring people together to form a new community that had not previously existed. One researcher told the story of his experience with caBIG:

> Early on, in 2005, it became obvious to all of us that if we wanted to get funding from caBIG, we had to line up a bunch of people and agree that we were all going to work together. caBIG from the very beginning developed this model, that if you want to be successful, come up with a collaborative plan…. When they first came around in 2004, they said, it's really swell that you have something to offer the community. It's also swell if you are weak in an area, if you don't have a lot of tools in something, you can get it. At first, no one believed it. We were all used to saying that we had everything for everyone. caBIG created this model, and they funded it that way. It's a very big cultural change.
>
> One of the most remarkable things about caBIG is that we [cancer researchers involved in biospecimens management] all know each other now [laughs]. We talk to each other at least once a week, we get together every few months. It's been a wonderful relationship.
>
> I'll tell you about a little sharing program we had here that preceded caBIG; it was in 2002-2003 sometime. There was a state-funded program …. multiple cancer centers were told to get together to determine how to distribute…the NCI component of the funding. They wanted a state-wide biobanking kind of thing. We all got together in one room. I was there, there were people from [lists other institutions]. I knew the people, but we never worked together. The institutions themselves are very competitive, you don't let on anything, you are competing for patients, etc.
>
> [Director X] was leading up the meeting, and we all went around the room, and each institution was supposed to say what their strengths were. We went

> around the room, and all of us could do everything [laughs]. I remember [Director X], exasperated, said, "Look folks, let's all stand up and open up our kimonos. Tell us what you can do, and forget the things that you really aren't the best at." We went around again, and then divided up the work, and we were successful. It was kind of a pre-caBIG type thing. It was a nice experience. caBIG has magnified this on a national scale. What has caBIG produced? A collaborative atmosphere that is very unique.

Another researcher, who still questions the feasibility of the success of the grid system, notes that this does not detract from caBIG's success overall:

> caBIG has already been successful – the very idea that you could bring together all the NCI cancer centers, with the idea that they could share data, fundamentally changed the landscape, and they realized they had a tremendous amount in common. It was a super-saturated solution already; all they had to do was drop something in there, and all of the sudden, people were collaborating. caBIG was phenomenally successful, even a year, two years into the project. The second question is whether the technical aspects will be successful, I think the jury is still out on that.

One nuance deserves to be highlighted here. In "Data Sharing and Technology," the caBIG Director was quoted as saying that it was the goal of interoperability and setting standards that was building community. Interviewees would agree that communities are being built, but with the exception of specialists working on vocabularies and the specific technical elements of interoperability, the community focus is not on the interoperability factor, but rather, shared research problems and the needs and challenges they are facing in their own work. This is different, and more personal, than a focus on interoperability as a goal unto itself.

Leadership is also indirectly referenced here; the caBIG Director is regularly referred to as a visionary that is driving change in a way that has led to the successes caBIG has had, and

certainly, the director that insisted that everyone "open up their kimonos" was demonstrating leadership that led to change at a specific time and place. It is ironic that a program that focuses first on technology as the outer layer of the onion of complexity is celebrated so clearly for its ability to connect people, even as the program positions socio-cultural difficulties as the final and most difficult layer of that same onion.

The Patient Relationship: Motivator for Sharing?

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.

In all of these examples from a range of interviews, however, another type of relationship is conspicuously absent. Not once did any interviewee report that a central motivator for sharing data was to improve patient care and outcomes. This observation is *not* intended to suggest that researchers don't care about patients; it is just that this was classified by many as "too big" or "not related enough" to guide daily choices. References to patient care, when mentioned, were specific to two topics: (1) the need for larger sample size in clinical data (you need more patients to detect disease and treatment outcomes); and (2) the negative impacts of HIPAA's patient privacy requirements that actually discourage data sharing.

When asked specifically about the benefits of data sharing for patients, these four representative quotes from four people capture the general trend across interviewees:

> I suppose, from a high level altruistic, yeah, any scientific advance is always good for mankind, the world, and patients overall. It's just not, you know…. the first thing that comes to my mind every day.
>
> There's so much heterogeneity between data sets; you need to focus on one specific question for scientific quality. Focusing on the broad category of patient care loses the focus on the specificity in data sets that is really the benefit of data sharing.
>
> The question feels like it might come from the patient advocate side. Patient advocates want investigators to think like that. I'm not sure that they necessarily would at this point of time, maybe down the road. It's not quite so compelling. It's more, how do I get my research forward, how do I get published. You would think it would be, maybe other people are far more generous, but I don't see it.
>
> I don't think people think about it that way. If more people thought of it that way, we might be in a better position. The people who think of it that way are the patient advocates, they border on fervor in some cases, that we have got to stop isolating ourselves, that we are stopping the progress of science. Most researchers don't think of it that much.

The suspicion that the question about the connection between data sharing and patient care came from the patient advocate perspective is, in fact, true. The following slides were presented in a 2007 workshop on biospecimens management best practices, sponsored by NCI OBBR. The voice that accompanied the slides noted that the people in the photograph were cancer patients, and that some of them have died since the photograph was taken. The message was clear: biospecimen sharing is needed to prevent even more deaths, and the interests of any particular researcher or scientific question are not as important as this overarching goal.

Figure 8: Sharing to Support Patient Care[15]

As a participant observer in the research I am conducting, it seems appropriate to inject one of my own experiences into the discussion at this point. I was in the audience when this talk was given, sitting among a group of scientists and technologists. I found the slides above very moving, and the talk led me to feel very proud to be part of the caBIG project. I started tearing up in hearing about the death of these patients, and looked around to see if the reaction was similar among others.

I did not see signals of compassion that I believe the presenter was aiming for. Rather, there was some boredom as people reached for Blackberries, and even subtle signs of annoyance. I wondered why, and then suddenly realized that the advocate was proposing that the personal interests and passions of the researchers present were somehow less important than the needs of people that, while ill, none of them had ever met. I started feeling empathy

[15] These two slides were used in Patient Advocate presentations at the NCI OBBR Biospecimen Management Forums held at NIH in Bethesda, Maryland and in Boston, Massachusetts in Summer 2007. Reference: Kim, Paula, "The Importance of Best Practices to Patients, Advocates & The Public." Biospecimens Best Practice Forums. Summer 2007. It available for free download at <http://biospecimens.cancer.gov/practices/forum/boston2007/pdf/Paula_Kim-The_Importance_of_Best_Practices_to_Patients_Advocates_and_The_Public.pdf> It is used here under a Fair Use Determination.

for the researchers too, but for completely different reasons. Two years later, the emotional discord of that moment was verified through the interviews supporting this research through the quotes above.

Patient advocacy messaging at the social level doesn't appear, at least among this interviewee group, to be a compelling motivator at the personal level. At the same time, motivators that are compelling at the personal level (the relationships that build trust and encourage sharing) are not referenced in social discourse. This is intriguing from an advocacy perspective; is the social level "missing out" by not encouraging data sharing in terms used at the personal level, or should the social messaging continue as it is, in the hope that it will eventually be integrated at the personal one?

Again, the tensions between coexisting norms and counter norms provide a possible theory-based explanation for this disconnect between social and personal levels. The Mertonian norms emphasize communalism, which represents science for the larger community and greater good, but not for a specific person or for the benefit of a specific scientific relationship. Patient advocates want scientists to care about the patient, which in ways is contrary to a broader communalism. At the same time, they ignore the benefits of the relationships that the scientists have with each other as a possible and equally valid path for achieving the same goals as patients have.

The counter norms explain the particularism and commitment that is more specific to personal experience and personally-known relationships, but in their focus on commitment to theories and explanations, also fail to address the more personal commitment of the relationships between researchers. Consistent with their theoretical positioning and original

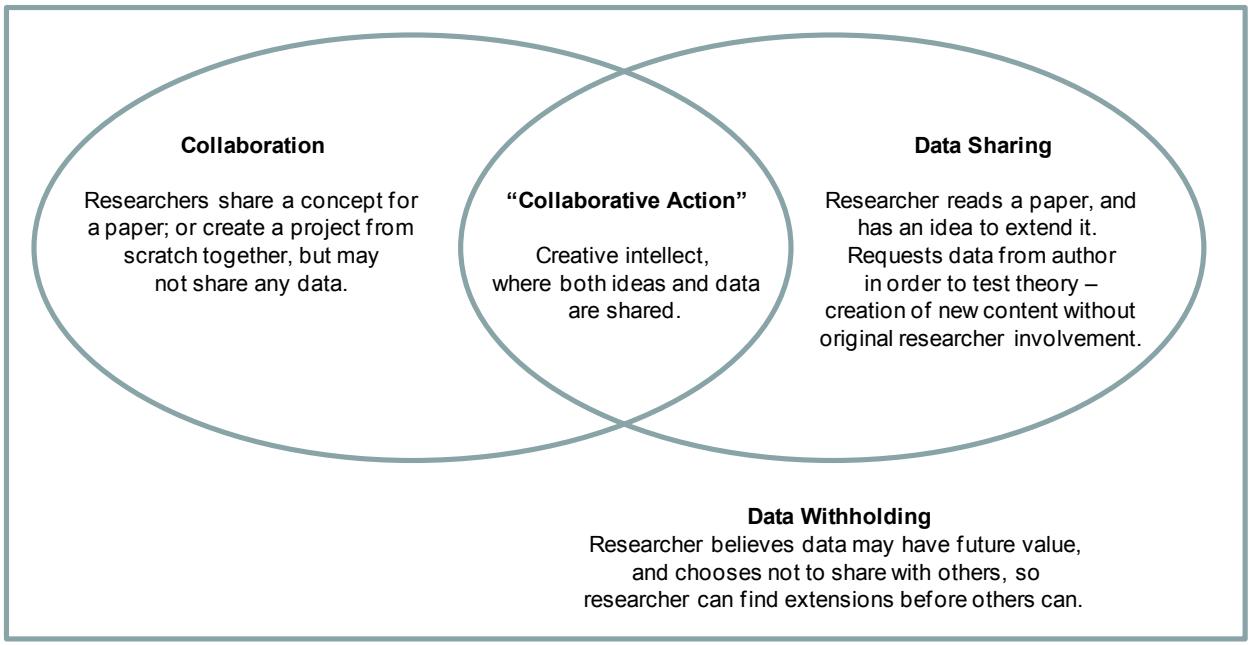
processes of definition, the norms and counter norms of data sharing remain segregated: the Mertonian ideal in the public sphere; and the counter norms in the privacy of the interview. Both, however, miss the personal relationships between researchers that seem to ultimately serve as one of the stronger motivations to share.

Collaboration + Data Sharing = Collaborative Action

The interviews about the interpersonal aspects of collaboration and data sharing also revealed a nuance about the differences between the two. While collaboration and data sharing are ideas that are often positioned as closely related in social discourse, they are seen as different elements of activity by researchers.

The figure below shows the model that emerged from interviews. The point behind the model for interviewees was that caBIG needs to identify ways to expand the area of overlap in the Venn diagram; this means both a focus on interpersonal connection of collaboration *and* the technology and security of data sharing. When social messages focus on the institutional and technological elements, they are focusing on the right hand side – not the more productive nexus referred to as "collaborative action" in the middle. Researchers acknowledged that there was an area outside the Venn diagram of data withholding, and that this is the space where the incentives would have to change to truly act to make that area smaller, and "get people to join the circle."

Figure 9: Collaboration, Data Sharing, and Creative Intellect



Most people interviewed indicated that they had indeed shared data with others, almost invariably when there was a point-to-point collaborative connection. There is no reason to believe these people were lying, since many acknowledged *not* sharing under certain circumstances, such as when there were restricted amounts of biospecimens, commercial agreements that prohibited sharing, and not knowing the requester. Despite this generally positive response, social level references to the personal side of data sharing focus more often on the *bad* motives that lead to not sharing, rather than the positive motives that support it. For example, one speaker at a caBIG open meeting on data sharing expressed,

> There are private acknowledgements that 'I don't want to share data with that person because I hate their guts.' We are never going to stop that kind of feeling. There's a certain level of irrationality that we can't address – but we can at least get it out there on the table to discuss.

At the social level, there tend to be admonishments of individual scientists that are unwilling to share for what are deemed childish, irrational reasons. The problem is that this kind of discussion ignores the positive and subjectively rational data sharing that *is* occurring, and fails to focus adequately on the incentives that would need to be in place to encourage data sharing – even among people who presumably "hate each other's guts." It's easy to say "people don't share because they don't like each other" and to discount what are considered overly emotional responses, consistent with what Stephanie Shields found in her research on social acceptability of emotion displays ("The Politics of Emotion in Everyday Life"). It's harder to ask, "How do we incentivize researchers who *do* like to share to do so more broadly, and in new ways?" From the research conducted here, this latter question would do far more to advance science towards the ideals it puts forth.

5.3 Research Grids: The New Panopticon

As with the Actor Network Theory (ANT) analysis conducted above, it may appear odd to draw in an institutional theorist such as Michel Foucault to reflect on the personal motives related to data sharing. In reality, however, questions about data sharing in a caBIG world are ultimately labor and power questions that go to the very heart of where an individual researcher's control and autonomy begins and ends.

Foucault was particularly interested in the link between knowledge and power, and the role of institutions in exerting control over individuals through the creation and definition of the institutional structures under which they operate. Discourse and classification help both build

and maintain these structures, by suggesting what the boundaries of "normalcy" are. Within the structures, holders of knowledge carry power that others do not have. The importance of these dynamics lies not, however, in overt expressions of power and hierarchy, but rather, by more subtle processes of surveillance and discipline communicated through specific choices in discourse and institutional structures.

Foucault argued that those with certain types of knowledge are able to construct categories that serve to either elevate or subjugate different members of society, in subtle but real ways. Detecting the process and outcomes of knowledge-power at work is somewhat illusive, as different sources of power are held by different parts of the network. No one person "holds the button;" instead, the specter of institutional surveillance and control, represented by the metaphor of the Panopticon (a prison structure that allows for the constant surveillance of prisoners without them being aware of being watched), governs action. If the power dynamics were more obvious, they might not be so all-encompassing.

These theories have direct implications when considered in light of the goals and discourse of caBIG and the related act of data sharing. The normative push to standardization and mass sharing of data through institutional Grid networks reflects an institutional shift that impacts the tools, practice, and professional of being a scientist. It is a new discipline, in multiple senses of the word. As the bioinformatics discipline expands, so must the disciplining of researchers so that everyone is following the same rules to allow for large scale data sharing, integration, and use. Ultimately, questions about data sharing choice and technique become fundamental labor questions about the defining value of a scientist. How does the standardization required to seamlessly exchange data on a broad scale change how any one

scientist's contribution is defined and valued? Most interviewees see the labor associated with formatting data for sharing as a nuisance that undermines their value rather than enhancing it.

The metaphor of "sharing" itself reflects a discourse shaped by power dynamics. As referenced previously, on its surface, the term "data sharing" seems somewhat straightforward. It is, however, a term that is heavy with social expectation and classification. First, it classifies data as a public good that should be shared, not sold, exchanged (in traditional quid pro quo spirit) or otherwise transacted. This social level positioning establishes a somewhat pejorative though unspoken expectation: people who share are "good," people who do not share (those that withhold) are "bad." This judgment, or course, is not explicitly stated, but rather subtly communicated through the simplicity of the term sharing.

With all the discussion about the decision to share or not share, it is also useful to consider what researchers do when they don't want to share their data. How easy is it to say no? From a power perspective, this is one of the most interesting potential impacts of the technologies offered by caBIG and similar programs. Currently, in an environment that is not "on the grid," the easiest response is simply to not respond to the e-mail request or phone message, or to defer to another group that has to decide – not difficult given all the institutional players generally involved in a project and this type of decision. A non-response is far easier than a no. One researcher noted that this is, of course, easier when you are not going to run into the requester at a conference the following month; as such, younger researchers who are not yet established in the community or known for their work, are often the ones that are more likely to go unanswered. Those that were in, or had been, at that stage when asking for data reported needing to go through their advisor to make progress.

Projects like caBIG add the possibility of surveillance to the problem of non-responsiveness, which is somewhat threatening. There is a personal consequence of a networked cancer community where intermediate research data can be more visible, as discussed above in "Technology and Data Sharing." The consequence is that, ultimately, says one researcher, "it makes visible the 'no.'" The researcher continues:

> The Grid has caused some of the divisions between the people who like to share and those who don't to become sharper because, like with the tissue folks, now, if they have this out on the Grid, and if they say what they are willing to share or not, it makes it far more visible. It makes it uncomfortable because you can't hide in obscurity anymore.

As with the introduction of the assembly line more than a century ago, technologies like the caBIG grid threaten to serve as a means of production and control. The Grid is a new delivery mechanism for the production of data across the research life cycle; and the visibility that it grants exerts a form of institutionalized control; as if it were a Panopticon, the data become suddenly more visible, and the autonomy of the researcher decreases.

What is the suggested resolution for this shift in sharing visibility? Increasing the visibility of "non-sharing" seems to invite punishment, despite data sharing incentives not being supportive of the researcher. At the same time, optimizing appropriate sharing seems difficult without making the "data market" more visible, a benefit provided by grid technologies. For now, the recommendation is to focus grid technology introduction efforts in research communities that are already sharing data, in essence, identifying areas of the Panopticon where the inhabitants are already gathered and sharing knowledge. It is in these preexisting collaborations that the burden of a new mechanism of "surveillance" may be seen not as intrusive, but as a benefit, if it will help reduce the labor burden associated with activities that

those involved are already committed to and engaged with (e.g., makes sharing easier among those already sharing). This could also be effective in future recruitment – if the Panopticon actually helps me achieve my goals, I may be more likely to invite others in as well.

Issues of autonomy and control in data sharing play out in other ways that theories of power help unpack. Two such issues include the question of "who decides" (who "holds the button" for decision making); and the question of identity and subjectivity embedded within researcher data being considered for sharing. Foucault's work again helps frame both issues. Personal knowledge of data yields personal power; putting that data out on that grid may dilute that power because of the possibility of being "scooped," it could become someone else's knowledge too. On the second item, the site of decision-making, the overtaking of scientific knowledge by expanding legal requirements (a different form of knowledge), also shifts that balance of power from the scientists to a more distributed and institutionalized system; no one holds the button on making a decision because everyone has a separate piece of knowledge.

No one theory of power adequately explains motives, incentives and rewards related to data sharing. The institutional frameworks of norms, counter norms and a Marxian focus on the control of production have already helped analyze the undercurrents of data sharing motives. Both the counter norms of particularism and interestedness are active in discussions about point-to-point data sharing; the elements of connecting with colleagues and interpersonal trust are findings that are not adequately captured by either the norms or counter-norms. And even in the discussions of the importance of personal relationships, the materialist considerations remain. Trust and connection is a prerequisite to sharing; then we enter the give and take that defines what rewards will make that sharing worth the investment.

Understanding the social-personal connections in data sharing is also greatly aided by feminist critical theory related to power and subjective knowledge. Evelyn Fox Keller's biography of Barbara McClintock's investigative style in "getting to know" the data at a highly specific and thorough level is much like today's researcher's relationship to today's high throughput genomics data. Getting to know one's data, in fact helps ensure that you won't be scooped by someone else; researchers "get to know" the data through careful analysis and cleansing; it becomes an extension of themselves and their identity. It is not the corn (in McClintock's world) or biospecimen (in today's researcher's world) that matters as much as the individualized and subjective knowledge that is gained through interacting with that material.

In this context, Harding and Haraway's work become of particular interest. Sandra Harding's conceptualization of standpoint epistemology marries well with the subjective nature of cancer research data sets: knowledge of each data set is individualized to the standpoint of a specific researcher, forming a specific local knowledge set that extends beyond the data on a page. Subjective understanding of one's data is the ultimate form of local knowledge, and sharing that data with others that one knows and trusts is the path towards creating local knowledge systems, the communities and consortia referred to across interviewees. Acknowledging the importance of local knowledge is not just an issue for multi-cultural environments or a local public's understanding of data; it can refer to the highly specific forms of knowledge and interpersonal power in the most traditional ivory towers of western science.

In Haraway's case (2001), the concept of situated knowledge serves the same purpose; data is not objective, it is instead an extension of a researcher's mind. Haraway's thinking on "cyborg cultures" ("A Cyborg Manifesto" 149-181, Ch. 8) also adds a theoretical dimension that,

with actor network theory, helps to blur the line between researcher and the technology that defines their contributions and even identities as the techno-scientists of today's cancer research field. A researcher's tools and data are extensions of his or her identity; the very real question becomes: what criteria apply to make that data no longer part of a specific identity? When does a line get crossed that leads to data becoming an objectified commodity, easily accessible by others for whatever scientific purpose it will support? Conversely, how could we work to integrate the subjectivity of the researcher's understanding in the technological network of sharing itself? This is a two-way integration of people and technology; it is not just about humans becoming cyborg; it is about the grid becoming human: being interhuman on the internet.

Currently, the caBIG program remains largely silent on the topic of connecting people across the Grid. While the "caGrid Portal" allows a visitor to "see" which institutions have nodes on the Grid, and who the caBIG points of contact at that institution are, this is largely an exercise in labeling, not connection. There is no caBIG social network, or mechanisms by which to specifically designate the people or institutions with whom they wish to share. Data sharing is largely framed by the program as a supply-oriented issue; the nature of the data and its sensitivity drive the degree to which sharing is allowed. The demand side of the equation, who would want the data for what purposes and what would qualify them to receive it over the grid, has gone largely ignored.

What would happen if this were suggested? Program patterns suggest that privacy concerns generally trump the benefits of interpersonal connections. As an example, the caBIG team won't post its public Annual Meeting attendee list on the Internet, fearing that attendees

will get “spammed” by unsolicited e-mail. “We need to protect the people working on caBIG from getting overwhelmed with requests for help,” says one caBIG program team member. This concern is consistent with a culture of anonymous data sharing over grids: if you format your data so it can be easily exchanged, no one will bother you. Unfortunately, this means that finding you for potentially collaborative opportunities may be more difficult as well.

As one researcher directly pointed out, “data sharing is not about someone deciding to be nice.” In many ways, the sharing or withholding of data is about the understanding and expression of power by a wide range of distributed actors, each representing both their personal motives and in the interests of the social group they belong to:

- the power of institutions to drive or discourage data sharing activities through incentives and rewards;
- the power that is becoming more distributed through the scientific system through new technologies of visibility (or surveillance), where anyone can see what is being shared and potentially withheld;
- the new institutional structures of review and control of data, by a range of offices beyond the reach of scientists;
- the ultimate power to share or not share held by the patients who consent to share their bodies or not; and
- by the researchers who will ultimately know their data at the most complete and personal level, and by the technologists who have the skills required to make that data available to others.

These are power struggles that are constructed and expressed in different ways at different levels of discourse, both personal and social. Often, these expressions take the form of metaphor, a vital tool for better understanding and sorting the language and images used to both explain and position the arguments on data sharing by different actors. As such, the *metaphor* of data sharing is the focus of the next chapter.

Chapter 6: Images of Data Sharing

"Academic medical centers are a community of fiefdoms, united by a common parking problem." (Quoted in Havenstein, Screen 2, Par. 3). This organizational metaphor, delivered by the director of cancer information systems at Duke University, is a vivid but humorous snapshot of the individualist – yet also institutionalized - nature of science. Metaphor is a powerful tool for communicating about data sharing both at social and personal levels, and is the focus of this chapter.

Scientific work in cancer has been described as a war (Pilcher; Davis); cancer research as a highway (Reimann), and genes as a book of life (Kay), notes of music (Lopez), or as diverse as codes and targets of attack (Prior). Public discourse and personal descriptions of the controversies surrounding data sharing in cancer research are also filled with metaphors that both elevate and minimize the impact or agency of actors (both human and technological), the relative value of artifacts, and the meaning of activities. While previous chapters have indirectly referenced the use of metaphor in data sharing discourse, this chapter more closely examines how metaphor is used to help define the data sharing landscape, at both social and personal levels.

Chapter 1 provided an introduction to both the value and constraints offered by metaphor as a conceptual tool, highlighting its uses in (1) establishing cognitive understanding; (2) advocating certain actions or positions; and (3) capturing the embodied nature of scientific action. Different metaphors are used by different people in different contexts to describe the data being shared and those agents sharing it, which in turn helps us to learn more about the

nature of the value assigned to the data as it moves across the network through the action of various agents. To begin to see how and which metaphors are used in data sharing, the following table lists the most common metaphors detected over the course of this research. These are simply examples that "set the stage" (so to speak) for the following analysis.

Table 3: Data Sharing Metaphors

Agent, Object of Discussion	Phrase	Sample Source	Type of Metaphor
Artifact: Unmanaged Data	Only then will be it be possible to fully exploit the ***mountains of samples***, and ***reams of data***	Article	Samples and data as mountains to be navigated or reams of paper to manage.
Artifact: Unmanaged Data	We are in the midst of an ***explosion of knowledge*** about cancer as a disease process….. Within the cancer research community there exists a "***Tower of Babel***" problem. Researchers (will be able to) ***tap into an ocean*** of raw published data.	Article	Knowledge as unmanageable or potentially violent if not controlled; or as a resource that can be accessed; Languages as not understandable.
Action: Improving Research	We need – and intend – to ***move at warp speed*** to serve the patient community.	Article	Research as high speed journey
Action: Improving Research	Biomedicine has experienced ***explosive growth***… Biomedicine is at the ***precipice of unlocking*** the very essence of biologic life	Article	Biomedicine as being at the start of significant and dramatic change
Artifact: Databases/ Technology	A lung image database is ***breathing life*** into "medical grid" vision"	Article	Database as life giving
Artifact: Databases/ Technology	Think of it as an organic ***bank account.*** You put your biomaterial in and ***earn medical interes***t …that grow out of that ***deposit***	Article	Resources as collections for repeatable use

Agent, Object of Discussion	Phrase	Sample Source	Type of Metaphor
Artifact: Databases/ Technology	***Libraries of resources*** for cancer research	Article	Resources as public and reusable goods
Artifact: Databases/ Technology	Our applications (software) can ***talk to each other;*** ***Let Data Speak to Data.***	Quote in Article; Article Title	Databases as human; able to communicate.
Artifact: Used Data (Data With Little Value)	Sharing data that has been ***wrung out*** or ***beat to death.***	Interviews	Data as an exhausted resource, taken to its potential by a researcher.
Action: Act of NOT Sharing	I'm not supposed to ***go above 30 out on this street,*** but I and everyone else does it.	Interview	NIH data sharing rules as traffic laws to be broken.
Action: Act of NOT Sharing	You don't actually want all that ***raw*** data out there. You'll just end up with a bunch of ***noise***; a bunch of ***crap*** that people then have to ***wade*** through.	Interviews	Data as potential garbage or excess
Action: Act of Data Sharing	Data sharing is like ***dating***; data sharing is a ***scientific marriage***	Interviews	Data sharing as a form of intimacy
Action: Building Communities	We are ***embracing*** the individual diversity of members and connecting them	Article	Data sharing as a form of intimacy
Action: Incentivizing Data Sharing	We need both ***carrots*** and ***sticks*** to make this happen. We need a ***scorecard*** that counts data sharing.	Interviews	Incentives as needing to be both rewards and punishments.
Structural Construct	Data sharing occurs most often in ***data clubs.*** My ***friends*** are on the grid; I should be too.	Interviews	Club as a community of people who will share data

Agent, Object of Discussion	Phrase	Sample Source	Type of Metaphor
Action: Act of Data Sharing	We are all just ***waves in the kiddie pool***.	Interview	Data sharing as a chaotic game.
Action: Act of Data Sharing	This is the opposite of ***nine women making a baby in a month.***	Interviews	Data sharing as community project, science as creating life
Action: Act of Data Sharing	Scientists are afraid of getting ***scooped*** if they share their data**.**	Interviews and Article	Science as competition for publishing.
Action: Act of Data Sharing	Data sharing is a ***headache.***	Interviews	Data sharing as painful.

6.1 The Goal of Control: Explosions, Tsunamis, Sponges, and Crap

Communication about the need for caBIG, and related large-scale biomedical research efforts, is often framed using naturalistic metaphors and imagery that communicate the sense of being overwhelmed or being taken over by the data. Conceptually, this use of metaphor both helps explain the foundational problems that lead to the need for data sharing (establishing cognitive understanding) and encourages the perception of technology as the best solution to the problem (impacting interpretation). In *Metaphor and Emotion*, Kovecses describes one form of the fear metaphor as being "fear as a natural force" (23-24). 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 itself can be shared and used, under the control of scientists, through the structure and order provided by caBIG and biomedical informatics. Even the acronym of ca***BIG*** seems designed to communicate this idea; only a program as big as caBIG could help solve these large scale

problems. Consistent with this idea, several images appear to work to trigger emotions of feeling overwhelmed about the volume of data, all manageable through caBIG:

> We are in the midst of an ***explosion*** of knowledge about cancer as a disease process….. Within the cancer research community there exists a "**Tower of Babel**" problem…. caBIG is being developed specifically to enable and accelerate the "bench-to-bedside-and-back" cycle (vonEschenbach and Buetow 22-24).

The naturalistic imagery echoes the Baconian tradition of dominance over nature; however, it is also an ironic displacement of the material being shared. As already described, the data generated to support cancer research is often micro in size, not macro. More often than not, the data being considered for sharing is generated from inside the body at an exceedingly small scale, not the large scale suggested by the imagery of natural disaster.

Despite this displacement, there is a unifying idea: a large-scale entity that is made up of individual parts, termed as a "part-whole metaphor" (Kövecses, *Metaphor: A Practical Introduction* 145). This is a shared theme: an explosion is made up of many small pieces of debris; water (including tsunamis and oceans) is made up of countless water droplets – just like a biospecimen comprises cells, and data sets comprises multiple elements. The size may be different, but the "one comprising many" idea is consistent - the imagery simply objectifies the personal data into a larger scale to communicate a sense of urgency and scale that small-scale imagery can't capture at the emotional level assumed to be needed to inspire action.

In *Making Truth: Metaphor in Science,* Brown notes that "large complex problems in science inevitably involve multiple metaphors operating on different levels (13)." This is very true when it comes to data sharing metaphors; in particular, different metaphors in this case communicate perceived levels of human agency or control, at differing levels of scale.

For example, large scale metaphors like the explosions and tsunamis used in social discourse are outside the agency of any specific actor. 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. This highlights the use of metaphor to describe the embodied nature of human activity, suggested by Lakoff and Johnson in *Metaphors We Live By.* 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, leaving it as "freeware" (a term used to describe free software). Other researchers described "raw" data (a very common science metaphor), as "noise," "crap," or as additional stuff you have to "wade through" (data as excess or garbage). In this category also fell the metaphor of data sharing as a "headache." The point of these metaphors was to communicate the point that just because data is available does not mean that data is valuable, and in fact, the transaction costs (or pain) of reviewing data that has not been "cleaned" may not be worthwhile time.

In some of these cases, such as those where large scale is positioned as *outside* control and small scale is positioned as scale *within* control, many of the metaphors share the common theme of a part-whole container, or some other flexible object from which something else can be extracted. Lakoff and Johnson would likely categorize these kinds of terms as falling within the category of "orientational or spacialization metaphors" (30-36); waves and explosions flow or happen over us resulting in small elements that come from above to us, but when we beat or squeeze something, we are on top and the elements fall below. While there is a shared linkage

of violence, the site of control differs with scale – in essence, capturing the control and autonomy that each researcher has over the data held in that researcher's possession. The notable exception are metaphors that communicate the "data as excess or garbage" image; when talking about the messiness of excess and un-cleansed data, the distinctiveness of the part-whole metaphor is not needed or desired in communicating the point.

Titling and graphics also set the stage at the very beginning of articles, with pictures of natural disasters and explosion-like artwork, or simply stark large fonts setting the tone early, presumably to shape the motivational context in which something is read.

It is not just disaster metaphors that set the urgent need; while they are less common, there are also metaphors that capture the positive outcomes that come from biomedical informatics when it is leveraged well. The March/April 2006 edition of Technology Review refers to caBIG with the title, "Cancer's 'World Wide Web': A lung image database is breathing life into 'medical grid' vision." Clearly, the "breathing life" metaphor is an attempt to invoke a feeling of positive affect for caBIG as collaborative science and data sharing are enabled through technology. Another article speaks of "data dreams coming true" through the new informatics technologies now available (Kaiser, "Editorial: Making Data Dreams Come True" 239).

6.2 Mechanisms to Share: Libraries and Banks

Metaphors are often used to explain; the goal is to provide a common frame of reference so that people can understand a new idea by comparing it to one that is already understood. Early in the caBIG project, the NCI Bulletin published an article to explain the

purpose of the program: “Currently in its early stages of development, this new system will offer a ***library of tools and resources***—from clinical trial management systems to tissue bank and pathology tools—that are all built to common standards and are interoperable with other existing systems. It will also allow researchers to ***tap into an ocean o***f raw published data (NCI, “caBIG: The Launch of a Bioinformatics Community” 5) Here, the library reference gives a sense of how researchers would “tap into” the data – creating, perhaps, a container metaphor that leads to the image of retrieving and cataloguing information from a sea or wave of library catalogue cards.

The library metaphor extends to the way in which NCI writes about biospecimen management. “Biorepositories (or biobanks) are ‘libraries’ where biospecimens are stored and made available for scientists to study for clinical or research purposes” (NCI OBBR, “Patient Corner,” Par. 3). Again, the term suggests a collection of reusable resources; one borrows a resource, and then brings it back based on pre-agreed conditions.

Other articles about biospecimens use a different metaphor, that of a bank, to educate readers about these resources and their benefits. In some of these cases, the repositories are called “biobanks.” Metaphorically, this is an important distinction, because the use of the different terms can set different expectations related to the resource. The figure below, for example, is from an *Economists Technology Quarterly* 2005 editorial on the benefits of centralized biobanking. The imagery helps communicate the complexity of the data sharing problem, as well as subtly shaping the variables that are in and out of scope.

Figure 10: Images of Biobanking[16]

In the figure, a human hand holds a key, which will presumably unlock the stored potential of the biobanks building. Ironically, despite the metaphor of a "bank" inherent in "biobank," generally associated with financial or economic ideas, there are no monetary symbols in this graphic. Rather, we see only symbols of science (test tube, beaker, and presumably, a cell), medicine (a syringe), and people (a heart and human figures). These are the materials that will lead to the future payoff, both monetary and scientific; in fact, the text of the article directly references the economic *potential* that may lie in the data associated with one of these repositories; the potential is not captured in the image.

TIME Magazine is more direct in explaining the metaphor in a description of biobanks as one of "10 Ideas Changing the World Right Now." The article, accompanied by a photograph of a gloved scientist lifting sample boxes out of a freezer, describes the relationship between a

[16] This figure is from the following reference: "Report: Medicine's New Central Bankers." *Economist Technology Quarterly* December 10, 2005 (2005): 18. It is used here under a Fair Use determination.

biobank and traditional bank, and is explicit about the motivation one should have to make a "deposit:"

> Think of it as an organic bank account. You put your biomaterial in and earn medical interest in the form of knowledge and therapies that grow out of that deposit — no monetary reward, just the potential that you might benefit from the accumulated data at some later date (Sorry, no shiny new toaster to inspire you to open up such an account either — just an appeal to the greater medical good.) (Park, Screen 8, Par. 3).

One deposits something into a bank because one hopes that it will yield value over the long term. Researchers take a person's deposit of human tissue and create therapies and knowledge that benefit the greater good. This is different from the image of a library. First, it suggests an economic value proposition that the library image misses. The other difference between the metaphor of "bank" versus "library" is that libraries suggest that objects will remain somewhat in perpetuity; that is, they can be reused multiple times by different researchers. Banks, on the other hand, communicate a different proposition; once a resource is removed from a bank, it may or may not be returned. In fact, the emphasis here is actually on the change that will take place. When a person deposits tissue, science is invested in and changes.

In *Marketing Metaphoria,* marketing specialists Zaltman and Zaltman call this a "transformational metaphor;" that the investment in a product or service will lead to a deep change for the "consumer" involved (64-68). In this case, the "consumer" in investing in science; instead of paying in dollars for some good, they pay with tissue for the good of science.

Is the use of the word "bank" for purely marketing purposes, invoking a deep metaphor that will be more likely to gain buy-in than the vision of the library? It may not be this direct; in

the end, the difference between the library and bank metaphors also reflect functional differences in these types of institutions that are also aligned with different forms of data. While both of the examples above related to biospecimen management, the application of the metaphors might better be applied to more distinct data types. The image of the library works well for persistent data that can be reused limitless times, such as a data set generated from a micro array; the image of the bank seems to work better for biospecimens, where the material is a finite resource that can be exhausted if there are too many "withdrawals." Metaphors can be used to generate understanding – and the difference between libraries and banks in the imagery show a tension in how data is both viewed and positioned in explaining the tools to others.

One article on data sharing positions digital research data as a "third stream of scientific capital," extending the bank metaphor into broader images of economic tools (Beaulieu 6). This is a double metaphor, positioning data as economically valuable, and also as a flexible moving object that travels in order to manifest that value. In interviews, this was also reflected in the popular metaphors of the need for both "carrots and sticks" and "scorecards" to change the culture: tools to be wielded by organizations in order to encourage sharing and to punish those that don't share.

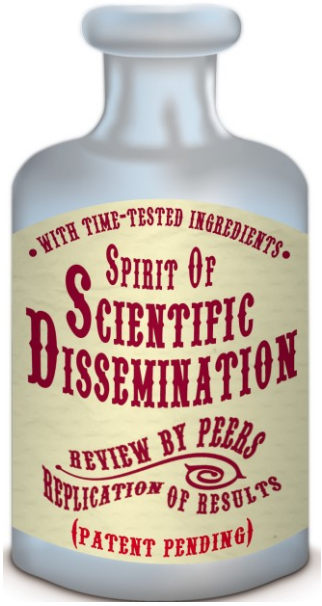
6.3 The Process of Data Sharing: Puzzles, Potions, and Traffic Signals

Other metaphors and images refer more to the process of science, rather than its artifacts (like banks and libraries). For example, one article relates data sharing to putting together pieces of a puzzle (again, repeating the part-whole metaphor) – in this case, referring

to data shared by scientists through collaborative authoring platforms called wikis (Waldrop 22).

A *Journal of Clinical Investigations* journal article addressing the role of material transfer agreements (MTA) as a mechanism for sharing scientific reagents uses the image of a medicine bottle as a visual metaphor to communicate that intellectual property rights mixed with the Mertonian norms will create data sharing that will result in medical cures (Dove 427). Again, it is hard to ignore the part-whole messaging here. In this case, it is the distinct interests that together form the container of science itself.

Figure 11: Combining Data Sharing "Ingredients" Leads to Cures[17]



Another class of metaphor encountered in both interviews and caBIG tool relates to the idea of traffic controls. For example, the caBIG Data Sharing and Security Framework, a

[17] This figure is from the following reference: Dove, Alan. "When Science Rides the MTA." *The Journal of Clinical Investigation* 110.4 (2002): 425. It is used here under a Fair Use determination.

decision-making matrix that guides different protections when it comes to data sharing, is organized around a "green, yellow, orange" color metaphor that clearly reminds the reader of stop light settings;[18] the green category includes data that can be easily shared; the yellow category contains more sensitive data; and orange is the highest level of sensitivity. Interviewees also used traffic metaphors when discussing data sharing decisions, with frequent references to the NIH Data Policy as a traffic rule that has little enforcement power, and which is frequently violated without consequence.

6.4 Building Communities: Dates and Data Clubs

The previous chapter on the personal aspects of data sharing described the importance of interpersonal connections in facilitating data sharing. When turning to these personal discussions of what motivates researchers to share data, the metaphors become dramatically different from explosions, tsunamis, and sponges to be squeezed out. When the discussion turns from the data itself to the human actors *sharing* data, the metaphors themselves also become far more personal and positive. For example, one interviewee delightfully turns the tsunami imagery on its head when talking about the relationships between researchers engaged in biomedical informatics. He says,

> caBIG is trying to capitalize on the back side of the wave – if the wave is the advance of science, and the height of the wave is some measure of the value of the data in your possession. As the wave moves forward, the data becomes less and less valuable as you ride the back side of the wave. As your data becomes worth less, and you want to be a nice guy, you can share that data pretty easily

[18] Originally, the framework incorporated the idea of an "EZ-Pass" for the green lane; this was abandoned quickly due to trademark infringement and copyright concerns. The model purposefully does not use the color red, because it would suggest an absolute stop or dead end, which was never the intent.

> through caBIG. It could be that this data may be the front end of someone else's wave, exactly what the next person needs for *their* discovery. Your back end of the wave could be someone else's front of the wave. To advance your own wave, you may need some data from someone else… you have to make the connection and the collaboration to advance your own work. Being a nice guy might help move the wave forward. We're all little waves bouncing around in the inside of a kiddie pool.

Here, the threat of the wave of uncontrolled data is instantly reframed into a chaotic but fun "kiddie pool" community, where everyone is creating his or her own waves and interacting with each other. It is not a metaphor of control, but rather of spontaneous and unpredictable give and take interaction, a post-modern reaction to the modernist need to both control and harness the giant wave as it overtakes us all. This aligns well also with the comment that caBIG will only work when people are able to see others they care about on the network, "my friends are on the grid; I should be on the grid too."

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.

Human pregnancy lasts nine months; in nine months, one woman makes a baby. This is a rather basic scientific fact that two women interviewees twisted in a creative way to illustrate the importance of data sharing. One notes, "This is the opposite of nine women making a baby in a month. You can't do it. There are some problems that can only be solved in a collaborative

setting; there is no single individual that can truly take ownership; the findings need to be communal. There are some people who understand that, they are the ones that are sharing." It's a complex metaphor, stressing through a very personal process the joining of forces that must occur for a new finding or idea to be born. In addition to the fact that pregnancy cannot be divided among nine women and completed in a shorter time, scientific work cannot be split up and done in isolation by individuals; it must be done by one communal body in order to create something new.

These are all human and relationship oriented metaphors, but the images are devoid of the power differentials that are seen in metaphors of structure and control. These are not "parent-child" metaphors or "teacher-student" metaphors; these are metaphors that signal equality and non-organizational relationships: "marriages," "dating," "friends," "clubs." Intriguingly, this is not an organizational set of metaphors described in a core work of metaphor in work settings, *Images of Organization* by Gareth Morgan. Morgan documents the phenomena of organismic and patriarchal organizations (33-71), but not of metaphors that signal a more egalitarian and social pattern of informal organization. In contrast, the metaphor of family was detected as one of nine distinct metaphor groupings in a 2003 study of metaphors used among information sciences teams to describe the systems development process (Kendall and Kendall 149-171).

While considering the gendered dimensions of data sharing is outside the scope of this work, it is interesting to note at this point that all instances of family and relationship metaphors (including the dating and pregnancy metaphors) were offered by women interviewees; none came from men. This is not to suggest that relationships are not important

to men; men often referred to the importance of relationships in data sharing. Men, however, did not use the specialized metaphor discussed here. This is also not to suggest that all women use these metaphors; as many women did *not* use these metaphors as well. Future study oriented specifically towards the gendered aspects of data sharing would be needed to explore this observation further.

While not addressing this gender question, marketing metaphor theory provides context as a possible interpretive backdrop. Zaltman and Zaltman describe inter-personal imagery as deep metaphors of "friends as a resource, as an economic exchange, and as a valuable commodity" (148-149). Close relationships allow us to complete deficiencies in ourselves, and to "extend our brain" by using others for support (150). This generalization appears to work well in terms of how researchers talk about data sharing; whereas marketing specialists often use the metaphor to highlight the *emotional* resources offered by friends and families, researchers use the term as a way to describe the ways in which they seek data and intellectual resources of others. Even with this distinction, however, the foundations of the metaphor – relationships as a resource and commodity to meet a need – are consistent across cases.

Other metaphor extensions relating to the safety of friends and families are seen in formal terminology used by data sharing specialists. "Safe Harbor" is a metaphorical term of art referring to a third-party organizational structure or protections that allow for the sharing of data without compromising intellectual capital advantage among competitors (Curt 3; caBIG, "Data Sharing and Decision Framework"). This metaphor also effectively links the "safety of a relationship" idea with the additional structure of protection from the explosions and tsunamis above. Friends, even competitors, protected by a safe harbor help us avoid the overwhelming

data exploding around us. Another product from the caBIG workgroup on Data Sharing and Security is the "trust fabric," which is a collection of security tools, protocols and governance that ensure that data shared on the Grid is secure and protected, wrapped in a soft comfortable metaphor to connote the sense of safety that one can have in using it. These are both conceptual and physical metaphors that shape both understanding and affect.

Relationships are also present in the social level discourse about bioinformatics; balancing the nature disaster images are images of connection and unity. Articles such as "Fostering Better Connections" (Borfitz); "Uniting Efforts in Molecular Medicine" (Buetow); "Agencies Join Forces to Share Data" (Butler); and "ONE for All" (McCallum) all carry positive relationship-oriented affect for the tools and structures that support open data sharing.

6.5 Being Scooped: The Risk of Sharing

Despite these positive metaphors stressing relationships, one of the most common metaphors encountered in interviews is the researcher's fear of "being scooped" if someone were to use his or her data to publish a paper before he or she were able to do so. The most remarkable element of this metaphor was its prevalence, as it was referenced in more than half of the interviews conducted. The birth of the metaphor in data sharing discourse was not evident, though it was clear that it was uniformly associated with the idea of "the scoop" in journalism, where a motivated reporter unearths a story first and has the advantage of releasing it first. "Scoop" is a physical metaphor associated with extracting something; in journalism, it is the true story uncovered. In scientific data sharing, it relates to someone

locating something in one's data that one did not find – getting the scientific story or result before you did (indirectly suggesting their intellectual and competitive superiority).

This metaphor was also the title of a 2002 *Science* editorial in which it was reported that a scientist was strongly objecting to the use of his DNA data for a publication without being consulted (Marshall 1206). The editorial (like the "getting scooped" metaphor itself) positions science as a battle rather than a club. A quote from the article layers on additional metaphor reaffirming this idea:

> The dispute is the latest in a string of clashes between those who collect and those who interpret data, and it brings into focus some questions that have been festering in the genome community. Among them: How much control should DNA sequencers wield over the data they gather? And should they be forced to share preliminary results— as many are now required to do—before they publish their own analysis (Marshall 1206-1207)?

The emotional response of the researcher is also referenced in the article, with descriptions such as, "being enraged" and "going ballistic." Being "scooped" is being beaten at a contest, which leads to anger (1206-1207).

The editorial also goes a next step, positioning the differences in data consumption between scientific researchers and bioinformaticists, and questioning whether the ideal of open sharing established by the Human Genome Project (led by the bioinformaticists) are, in fact, ideals that are appropriate:

> The researchers who benefit most from public access to genome data are the computer wizards of bioinformatics. They are sometimes seen as "parasites" because they rely on others for raw material, says Sean Eddy of Washington University in St. Louis, Missouri, a leader in this field. Eddy has championed free access to genome data collections in the past. But he says he has become more aware of the need to protect the publication rights of DNA sequencers. He even suggests that it may be time to "revisit the rules" that demand prompt public release of raw data, laid down during the heyday of the Human Genome Project

(1206-1207).

The term “parasite” is a particularly strong one in referring to a member of a specialized discipline. None of the interviews revealed language that was this emotionally charged, either directly or through metaphor. This may be because none of the interviewees reported ever having *been* scooped; again, because they reported that publication roles are generally determined right up-front, that hasn’t been an issue among those spoken with.

It is uncertain whether this article helped to bring about the universality of the “I’ll be scooped” metaphor, and related sense of fear and negative affect, to the prominence that it has among researchers. Certainly, *Science* is a leading journal in the field, and the article was published shortly after the “Genome War” (the Human Genome controversy between NIH and Celera) was at its height. It was even a popular enough article that more than one interviewee referred me to it. Regardless, the potential for “being scooped” is a deeply engrained image today, and interviews often reported the image as the first reason that people are afraid to share data. Before privacy concerns, before labor issues, the fear associated with “losing the story” is a clear motivational factor negatively impacting the decision to share.

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 competition and contest; and the previous partner is now positioned as a potential foe.

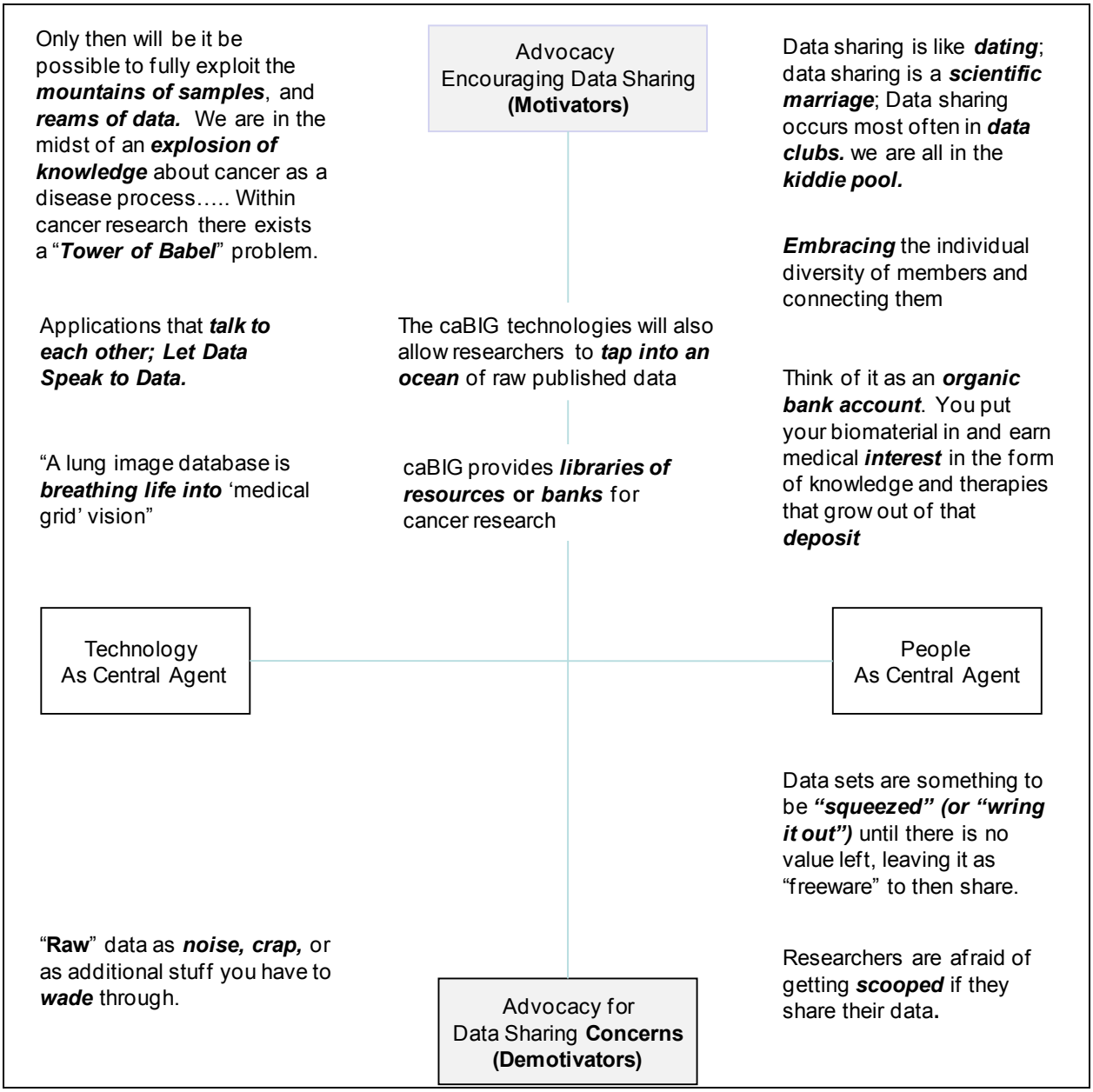
The following figure presents a metaphor map that shows the relationships between the metaphors used above. One axis captures agency: is the metaphor focused on technology as the agent, or people as the agent of data sharing? The other axis captures which arguments the

metaphors are being leveraged to support: arguments to motivate data sharing, or arguments to highlight the difficulty of sharing. For example, the metaphor statements on the upper left are metaphors being used to advocate sharing, where technology is positioned as the key agent in facilitating that sharing; statements on the bottom right capture metaphors that communicate human agency, but where the problems of data sharing are being positioned.

In general, it appears that the right side of the chart (where people are agents), the scale of metaphors are smaller than on the left; on the right side, the metaphors reflect a scale that people can control: dating, marriages, being scooped (or scooping others), depositing into banks. On the left side, where technology is being positioned more as the agent, the metaphors are larger and more amorphous and uncontrollable: tsunamis, explosions, breathing life, and unmanageable crap.

The patterns have interesting implications for how the use of metaphor might be advocated to different people working to motivate data sharing. To motivate researchers to share data, perhaps increasing the imagery that illustrates their control and connection may be more effective than the more macro-level imagery of explosions. In the end, the scale of the metaphors are also consistent with the most personal motives of interviewees for sharing data or not: nearby relationships with other researchers motivate me to share; the larger "good of science" or even patient care is not a motive that is of great presence in the daily thinking. Reframing metaphor to focus on smaller scale motives and experience may help better communicate the "what's in it for me" when it comes to data sharing.

Figure 12: Images of Data Sharing



6.6 Data Sharing, World Peace and Personalized Medicine

Language is imperfect, and it shows in the term data sharing. As discussed in the previous chapter, data "sharing" is itself a values-laden metaphor that automatically establishes

a lack of parallelism in alternative evaluation. While there is a choice involved here, 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. Several interviewees raised this problem in their discussion, but had no solution they felt was better. Alternative words were considered and discarded because they were deemed to inaccurately capture what is going on: data *exchange* insinuates reciprocal activity from both parties, which is not accurate; data *distribution* suggests information push to multiple parties, which is also often not the case. In the end, data sharing seems the closest to accurately communicating what is physically or electronically happening, but many agree that the "baggage" associated with the term "sets up" those who do not share as a "bad guy."

Two final metaphors deserve a mention to close this chapter. First, there is the "data sharing as world peace" image, raised in a few interviews and in two conference presentations. Each time, the delivery of the metaphor, so important to context setting, was not framed as an idealist goal, but rather as a pragmatic commentary on an unrealistic pageant contestant's wish. One interviewee notes:

> Sharing data is like world peace – everyone wants it, it's a good thing, but it really depends on the details and how it changes the way people do things. It is very personal, it's legal, it's cultural, it's financial. And because the benefits of sharing data do not cross the ethical boundaries of doing better research, it's a grey line. People are incentivized to not share data, on the financial and personal and professional levels. Especially the legal side, it is all about disincentives. To share data, you have to jump through hoops.

Like world peace, data sharing is something that everyone wants and thinks is a good thing, but is highly complex and difficult to bring forth in reality. Certainly a world without data

explosions and tsunamis would be a good place to start, but as most using the metaphor noted, "the devil is in the details."

Second, it seems appropriate to close this chapter with the metaphor used in part to drive today's increasing emphasis on data sharing: the metaphor of "personalized medicine." Personalized medicine is positioned by caBIG and other biomedical research efforts as a goal facilitated through large data sets and institutionalized sharing: the personal gone technical. This research has not explored how patients perceive this metaphor, and how it is contrasted with the personal care they currently receive in clinical environments from the people administering tests and treatments. What has been made clear in this particular research project to date, however, is that the quest for personalized medicine is currently being fed most by "personalized research" – research that embraces the subjective, and that is built, date by date, with trust.

Chapter 7: Motivating Data Sharing

This chapter focuses specifically on the motivations and emotions of data sharing. The intent is to provide a new analytical layer that can help organize and further analyze the diverse ideas encountered to date, and point to possible actions that might be taken to influence the data sharing landscape in different ways. The intent is not to engage in a philosophical work on the nature of motivation and emotion; this work has been done from many different perspectives reaching a range of conclusions (Lutz and White 405-36; Ambrose and Kulik 231-92; Lewis 221-34; Averill 571-80). This is a pragmatist project; the question is: how can motivation and emotion be detected in a systematic way, such that they can be strategically influenced for different outcomes?

Despite this backdrop, however, it is useful to begin with an accepted definition of both emotion and motivation to frame this discussion. In *The Lived Body,* Williams and Bendelow's conceptualization of emotion, developed to support their sociological research, is a definition that supports well both personal and social perspectives:

> Emotions are complex, multi-faceted phenomena which are irreducible to any one domain or discourse. Emotions, in other words, are thinking, moving, feeling "complexes," which sociologically speaking are relational in nature and linked to "circuits of self-hood;" comprising both corporeal, embodied aspects, as well of socio-cultural ones. While basic emotions – rooted in our biological make-up and shared among all human beings as embodied agents – are involved, they are endlessly elaborated, like colors on a painter's palate, though time and culture (137).

This definition is particularly relevant to this research in two ways. First, it acknowledges the link between self-hood and culture; emotions at a personal level are shaped

in part by the cultural landscape. Second, its reference to "embodied agents" is useful in allowing emotions themselves to be exchanged as quasi-objects, as people elaborate their understanding and decision-making processes.

Turning to motivation, motivation is generally defined as a force that that arouses someone to act towards a desired goal. Motives give purpose and direction to behavior. Like various treatments of emotion, motivation can also be conceptualized from a highly individualized perspective (personal motivation towards "fight or flight" for example) to a more social level view (where people are motivated to conform to social norms). The middle-ground approach, linking personal motives to social action, is most appropriate to this work.

When selecting a theory and framework to help analyze the motives and emotions of data sharing, the three primary criteria were: (1) the ability to be effectively applied to both personal and social dynamics; (2) the ability of the theory to address change; and (3) the presence of a structure that can be used to teach others to create that change. These criteria are important for both pragmatist and activist reasons. This research is not only concerned with understanding the subjectivity of motives and emotion; the goal is to ultimately devise possible strategies for changing them. This requires a theory that incorporates the idea of changeability in a structured and systematic way that can effectively deployed in "the real world."

7.1 An Introduction to Reversal Theory

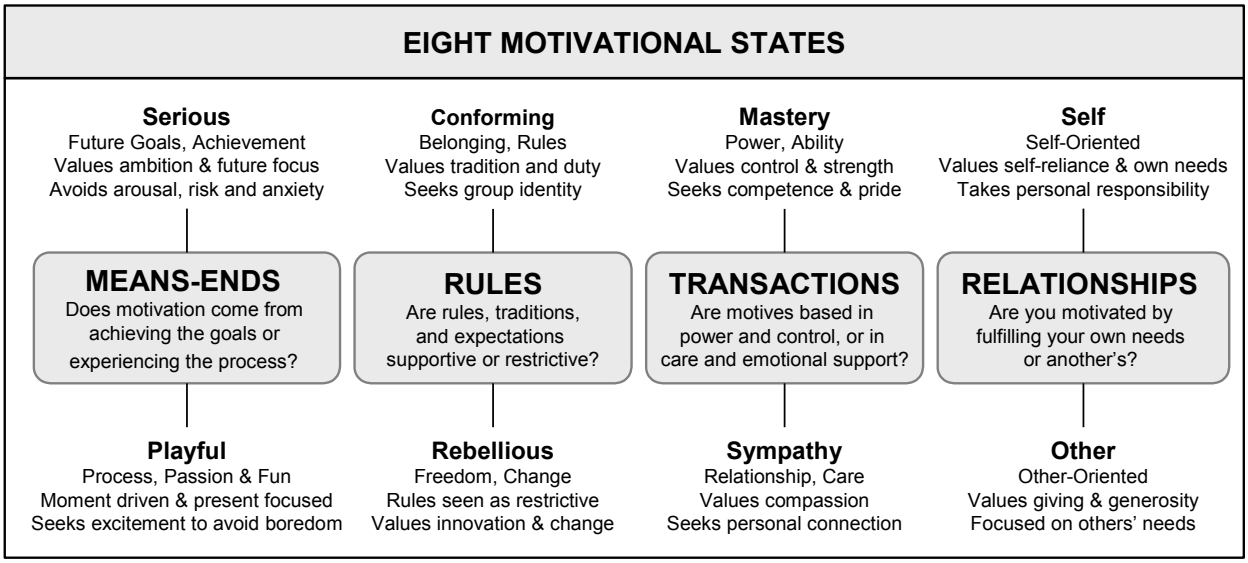
Reversal theory is unique among theories of motivation in its ability to support both personal and social explanations of motivation, and in meeting the three criteria listed above.

Based on a structural phenomenological model from psychology, reversal theory introduces a structure and taxonomy that emphasizes the complexity, changeability, and inconsistency of motivation, emotion, and behavior. Validated and applied across a range of disciplines in both psychological and social research contexts, the theory has been applied to better understand such phenomena as violence in sport, smoking cessation, addiction, weight management, organization change, engagement in high risk sports, and the human attraction to danger and war (Apter, "Motivational Styles"; Carter).

The reversal theory model proposes that humans regularly experience different emotions across four motivational domains; each domain is comprised of two opposing motivational states. There are eight total motivational states (two per domain), which yield eight core emotions: eight sets of "positive" emotions felt when things are going well (motives are being fulfilled), and eight sets of "negative" emotions felt when things are not going well (motives are not being fulfilled). The four domains are as follows; the states within the domains are shown in the figure that follows.

- **Means-End Domain** – Motives/emotions related to goals achievement versus engagement with process for its own sake.
- **Rules Orientation Domain** – Motives/emotions related to either belonging and adhering to rules/expectations, or rebelling and breaking outside of norms.
- **Transactions Domain –** Motives/emotions related to power and control (e.g, pride), or care and sympathy (e.g. compassion).
- **Relationships Domain -** Motives/emotions related to an individual's own interests, or the needs/interests of others.

Figure 13: The Domains and States of Reversal Theory



The following table lists the motives and emotions, both positive and negative, associated with each reversal theory state. Again, positive emotions occur when the motives of the state are fulfilled (called "winning"); negative emotions occur when the motives are not fulfilled ("called "losing"). One of the key benefits to reversal theory as a model is that the theory proposes that individuals "reverse" states within each domain, with a corresponding change in emotions, depending upon the specific values, meanings, and motives held by an individual in a situation. To impact one's own, or someone else's emotions, one must move oneself, or another, into a different motivational state.

Table 4: The Reversal Theory States, Motives and Emotions

State	Description/Motive	Positive Emotions (Winning)	Negative Emotions (Losing)
Serious	• Motivated by goal achievement • Engaging in activities for a future benefit or payoff.	Relaxation, Ambition	Anxiety, Fear
Playful	• Motivated by fun and the enjoyment of process and experience itself • Engaging in activities for the moment's sake only, unaware of any future impact.	Passionate, Playful	Boredom
Conforming	• Motivating by belonging and fitting in. • Experiencing rules and conventions as desirable and supportive.	Tranquility, Belonging	Embarrassment
Rebellious	• Motivated by pushing against the rules, and feeling free of norms and expectations • Experiencing rules and conventions as restrictive.	Free, Independent	Stuck, Resignation
Mastery	• Motivated by power and ability • Engaging in activities in order to gain control and mastery	Pride, Self-Confidence	Humiliation, Shame
Sympathy	• Motivated by care and relationship • Engaging in activities in order to gain personal relationships and openness – treating people as people.	Caring, Compassionate	Rejected, Guilty
Self	• Motivated by one's own needs. • Experiencing what is happening primarily in terms of its effects on oneself.	Grateful, Self-Reliant	Selfish, Resentful, Lonely

State	Description/Motive	Positive Emotions (Winning)	Negative Emotions (Losing)
Other	• Motivated by the needs of others. • Experiencing what is happening primarily in terms of its effects on others with rather than oneself	Modest, Selfless, Altruistic	Dependence, Martyrdom

Reversals (changes between states and emotions) occur when the situation changes, or when one's interpretation or reaction to that situation changes. These can happen in state combinations, as one state in each domain is active at any given time. The following examples illustrate reversals that are consistent with the research topic of data sharing:

- A scientist is working on a grant application, and is focused both on goals, time requirements, a desire to achieve, and on demonstrating the skill and ability needed to win the grant over the competition (Serious and Mastery states). In the midst of writing, the scientist becomes excited by the research problem at hand, and in the process and passion of the work, time and the consequences of the grant melt away (Playful and Mastery state). (Note, researcher is also likely in Self state, and if aware of grant submittal rules and feeling supported by them, Conforming state.)
- A researcher is generally content to follow university protocols when it comes to data sharing requirements, and in fact, feels supported by the safety net it provides (Conforming state). In this situation, however, the researcher wants to quickly get a data set to another institution, and becomes frustrated by the rules (Rebellious state).

- A legal professional is researching the guidelines for a material transfer agreement, making sure that all the variables have been captured and accounted for (Conforming and Mastery state). Later, the professional is talking to the researcher waiting for the materials on the other end, and feels empathy for the delay that has been caused by the requirements of the agreement process (Sympathy state).
- A researcher is at a conference to learn about the latest work in her field, to help generate new ideas for her next project (Self state). In the midst of a talk, she realizes that she has some data that may help advance the speaker's work, and decides to approach her during a break to offer her help (Other state).

7.2 The Motives and Emotions of Data Sharing

A range of emotions were self-reported during interviews, each of which can generally be traced back to motives related to data sharing based on what was being discussed at the time. Examples, with linkages back to the reversal theory state(s) that they suggest include:

- **Anxiety and fear** about getting grants and publications in today's competitive funding environment - captured in the phrase that "sharing pre-published data is committing academic suicide;" or being scooped by others ("Serious losing" – fearing the negative consequence of losing a grant or competitive advantage)
- **Guilt** over delaying sharing data and **embarrassment** about the messiness of data ("Conforming losing" – feeling like one "should" share data and that it should be neater)

- **Frustration** (extending to **resigned despair**) at the legal barriers that stand in the way of productive sharing ("Rebellious and Mastery losing" – feeling stuck by the requirements and not being able to push back)
- Feeling **insulted, irritation**, and **resentment** toward people who underestimate the time required to prepare data for sharing, or who ask for data so that they can just do the "fun" part ("Rebellious and Self losing" – wanting to break free of the expectation of sharing but not being able to do so; feeling like one's self interests are threatened).
- **Regret** and **sympathy** about the negative impact of legal requirements on a well-meaning researcher's ability to share ("Sympathy Other" losing – wanting to help and support someone but not being able to do so)
- Potential **humiliation** of getting "scooped" by one's own data ("Mastery losing" – feeling like one's competence could be questioned, and someone else would get ahead on one's own data if it were to be shared)
- Concern about sounding **selfish** because data weren't shared ("Self losing" – wanting to be autonomous, but aware of how that could be interpreted in a negative way).

Negative emotions were more prevalent in discussing data sharing, though positive emotions were also displayed during the interviews:

- **Pride** about research projects that had been completed successfully as a result of data sharing ("Mastery winning" – motives of skill and competence fulfilled)
- **Pleasure** and **caring** about the interpersonal connections that had been developed as a result of data sharing and through association with the caBIG community ("Sympathy winning" – motives of connection and compassion fulfilled through relationships)

- **Amazement** and **wonder** at the new bioinformatics tools and the change they are allowing for in science ("Playful winning" – openness to the process and possibilities of science, without specific concern for the implications or outcomes)
- Many interviewees reported the interview itself as being **fun** ("Playful winning" – enjoying the engagement of the moment, laughing at "the absurdity of truth" as interviewees reported and analyzed their own and others' behaviors)

There are patterns revealed by these emotions and the discussion to date, which can be used to point to recommendations about future motivational positioning about data sharing. The following matrix works to illustrate these patterns in one table. There are two categories on each axis; the columns separate the arguments that support and that illustrate challenges with data sharing. The rows subdivide these categories into arguments that are seen as compelling by researchers, versus those that are not seen as compelling.

Table 5: Revealing Patterns in Data Sharing Discourse

	Factors/Motives that Support Data Sharing	**Factors/Motives that Illustrate Challenges with Data Sharing**
Arguments Seen as Most Compelling by Researchers	• The ability to build mutually beneficial relationships based on trust with other researchers that lead to both collaboration and sharing on the path to grants and publications (emotionally positive). • Many agree that data sharing, *in theory,* supports greater scientific discovery and progress (emotionally neutral).	• Incentive and rewards systems (grants, publications) are generally (with exceptions like caBIG) seen as rewarding individual work, not collaborative (emotionally negative when it comes to sharing that does not support reward; emotionally positive when it does) • There are no direct incentives for sharing data with others – there is no way to capture the labor

	Factors/Motives that Support Data Sharing	Factors/Motives that Illustrate Challenges with Data Sharing
	The first item points to both Sympathy and Mastery, and Self and Other states, as data sharing supports both performance and relationship building, and requires both a focus on one's own interests and on another's interest.	involved in preparing data, and no formal career reward for doing it. Data sharing may threaten future ability to publish or otherwise gain professional benefits (emotionally negative). • Privacy regulations, institutional policies, and/or other legal barriers make data sharing too difficult (emotionally negative). • Too much pre-published data pushed out on Grids may increase the level of noise and data to sort through, leading to more data available but less efficiency because of the investment to sort through it (emotionally negative). • If I allow others to see my pre-published data, I may get scooped if they publish findings before I can (emotionally negative). *These items generally call upon the Serious, Mastery and Self states – usually pointing to the possible implications of lost rewards and capability at the individual level.*
Arguments Seen as Least Compelling by Researchers	• Data sharing is needed to manage and maximize the potential of the large volumes of data generated by today's technology-based scientific tooling (emotionally neutral, except when talking about the new technology itself, which is seen as neutral to positive	• The lack of common technical standards makes data sharing too technically complex (Emotionally neutral to negative). (Note: While the added labor involved in preparing data is a disincentive, it is not the *standards* themselves that are the issue, as generally

	Factors/Motives that Support Data Sharing	Factors/Motives that Illustrate Challenges with Data Sharing
	depending on person). • Data sharing is needed to improve patient care and outcomes. Data sharing is an ethical obligation (Emotionally neutral to negative). (Note: Because of the current structural disincentives, and the lack of clear evidence that data sharing truly does advance science, only ethical dimension of data sharing is seen as ensuring sharing as part of educational efforts for the next generation.) *These items generally point to the Other state- bringing benefits in areas that are not directly in the control of the researcher.*	those are not perceived as critical for the type of sharing being done.) *This is an argument that positions the need for caBIG through the Conforming and Other states – the argument is that we need caBIG to establish the standards to overcome this barrier. This is generally not seen as compelling for researchers.*

7.3 Domain Contrasts: Opposing States at Social and Individual Levels

The mapping above reveals certain gaps in motivational potential – where negative or neutral emotions might be converted to positive motivation and action if the appropriate incentives and potentially alternative messaging were put in place. In terms of communication, there are areas where it appears that social messaging is attempting to create positive motivational states and emotions related to data sharing, but where those states and emotions are *not* actually experienced at an individual level. On the other hand, there are also states and emotions felt at the personal level that are not expressed at the social level. This motivational misalignment appears to occur across the scales of each domain, creating an intriguing

interplay between opposing states within a domain between the social and personal levels. This pattern is considered across the four domains here.

Means-Ends Domain (Serious-Playful States): As the metaphor analysis above suggests, social messaging about data sharing works to instill fear about the waves and explosions of data that will overtake science if not managed and shared effectively through the tools of bioinformatics. Fear is an emotion felt only in the Serious state. At the personal level, this is viewed with some skepticism, as many researchers focus on more personal point-to-point data sharing; and counter the idea that because data is available means it must be shared. The anxiety of the public image that tries to motivate sharing is countered by skepticism at the local level. From a state perspective, this skepticism is expressed often through the opposite state of Playful, expressed through the many jokes about the NIH Data Sharing Policy. Humor occurs in the Playful state, where there are no consequences – the exact opposite of the social state (Serious) being advocated. This is not to say that the Serious state is not felt by researchers; it is, but it is felt at the more personal scale of anxiety about competing for grants and publications, which the NIH Data Sharing Policy is not perceived to directly impact, or the even greater fear of the humiliation of being "scooped."

Rules Domain (Conforming-Rebellious States): The second example where social messaging is not matched with a positive motivational response at the personal level relates to interoperability itself, and introduces a social-personal interplay between the opposing Conforming and Rebellious states. As discussed previously, the social level messaging about the benefits of interoperability are trying to establish a new mode of conformity, where data is made to fit a set of standards that allow for the efficient interchange of data across impersonal

networks. At the individual level, however, the norm is currently *not* to engage in this kind of sharing, but rather to share with a select group of collaborators, selected by the researcher. At the individual level, discussing the grid-based data sharing of this new approach sometimes triggers the Rebellious state among researchers, with objections about the effort involved and feeling insulted by the idea that the personal dimension can be removed from the process. Once the focus switches to this personal dimension, the Conforming state reappears, in the form of "quid pro quo" comments; if I share, it is expected that you will too, and that we will agree on authorship up front. These are "rules" that govern data sharing at the personal level, and invokes the Conforming state associated with meeting the expectation of peers. Researchers rebel against the new model of Grid based data sharing, but conform to each others' expectations.

This dynamic has interesting ramifications for change management strategies as grid technologies become more ubiquitous (which many feel they will inevitably become). At what point will Grid technologies become the new norm, with enough researchers "jumping on board" to make non-participation an act of rebelliousness that is no longer accepted by the peer group? Right now, it appears socially acceptable for researchers as a group to push against the grid model; at what point will that become the opposite, where enough researchers will be "on the grid" that *not* being on it becomes the unacceptable rebellious behavior?

Transaction Domain (Mastery-Sympathy States): This social-personal state contrast also exists between the opposing Mastery and Sympathy states. At the social level, impersonal grid sharing to enhance power and ability to cure cancer across the network is a Mastery-oriented goal; whereas at the personal level, building the collegial relationships that form the

basis for data sharing is fundamentally a Sympathy-based approach. Positive emotions from data sharing lie in the Sympathy state; proponents of grid technologies have not yet tapped into this dynamic in advocating the new approach.

The opposite dynamic occurs when the social discourse invokes the Sympathy state by pointing to patient care as a motive for sharing. When researchers were asked about the goal of patient care as a motive for data sharing, the broad response was generally one of non-emotive ambivalence. While no one indicated that patient care was unimportant, it is clearly not a motivator that inspires action with respect to sharing, as it would appear that patient advocates would wish. In reversal theory terms, patient advocates appear to work to invoke the states Sympathy and Other (compassion for others) to motivate people to share. In reality, at the personal level, the Sympathy state was seen far more often in discussions about building relationships with other researchers; relationships with patients were seen as too remote to be motivating in an actionable way. Rather, the emphasis at the personal level turns to the opposing Mastery state – a focus on how data sharing does or does not benefit a researcher's ability to get grants.

Relationships Domain (Self-Other States): Closely linked with the Transaction domain, the contrasts between social and personal state expressions extend to the Self-Other states. At the social level, the arguments are often framed to invoke the Other state: sharing should be done to benefit the greater good of science and patient care. At the individual level, sharing is more often motivated by the promise of personal gain (Self state); the Other state is invoked when talking about helping other researchers achieve their goals – a vital element of establishing personal trust. Both Self and Other states are clear at the personal level;

interestingly, there are very few Self-supportive statements at the social level. This is a clear motivational gap in motivational discourse, likely driven by the emphasis on the Mertonian ideals, which are clearly a tribute to the Other state.

Taking this analysis the next step, this discussion also helps demonstrate how the norms and counter norms play out in motivational practice, and highlights a potential overall gap in these two complementary theories. In general, the Mertonian ideals of science correlate with the Conforming-Mastery-Other states: conducting science using a rational objective approach that stresses the greater good and carrying out science in an orderly and repeatable way. The counter-norms correlate more closely with the Rebellious-Mastery-Self states: stressing the particularism of evaluating work not only on the idea, but on who had that idea, a solitary proprietary approach to science, and the interestedness of keeping one's findings to oneself. Both the norms and counter-norms appear capable of calling on either the Serious or the Playful state, as both goal achievement and "science for the sake of science" (a process focus) are suggested by each.

Considering the reversal theory states alongside the norms and counter-norms does reveal a gap in these core theories. This gap is revealed in the presence of the Sympathy state. While the counter norms do address the personal nature of science, the description is a very Mastery-Self focused state combination, focused on the competitive relationships between scientists rather than trusting and collaborative ones. Given that the original Apollo interviewees were positioned in the paper's setting as primarily individual contributors working towards a common cause (Mitroff 579-95), the specific dimension of inter-personal trust and

connection may have been perceived as less of requirement or area of interest in that investigation.

While the other counter-norms detected in the Apollo study were also detected in this research, there was also the additional detection of this inter-personal counter-norm, perhaps standing as an opposite (or complement) to communalism (which can be conceptualized as supporting a broad community of scientists rather than any one particular scientist). This research into data sharing validates not only the more personal counter norms, but also introduces *inter*-personal ones.

Inter-personal counter norms relate to the personal connections that form between specific people that encourage the products of science through team science collaborations that achieve more than a single scientist can produce alone. This emerging counter norm reflects a regularly reported subjective evaluation that leads to the development of trust between researchers; this is not just about researchers supporting or refuting each other's ideas and worthiness, but also their joining together in partnerships in order to produce results. This is a dimension of science not captured by either the norms or the counter norms, but is ultimately a primary motivator for data sharing in the science and technology of cancer research.

7.4 Reversal Theory and Metaphor

Metaphor and reversal theory are two conceptual tools that have been introduced to better understand the motivational factors associated with data sharing. This section integrates the two together, with two benefits. A first benefit is that the combination of tools

provides possible strategies for altering social messaging to achieve different motivational outcomes at the individual one. Second, it raises a dimension in the application of reversal theory that has not yet been explored by researchers working with the framework: that of scale.

Examples of metaphor as a type of symbol connecting individual understanding and social communication in data sharing were provided in the previous chapter. Jacque Carpenter provides a summary of how structural theories, metaphor, and emotion each can support each other in a research program that may ultimately lead to advocacy: “Researchers can employ metaphors as a mechanism to structure data or to help the researcher understand a familiar process in a new light. Their implications may suggest appropriate interventions, and they can be used as a rhetorical tool to evoke emotion (275).” In "Reversal Theory, Victor Turner and the Experience of Ritual,” Michael Apter suggests how metaphor and the structural model of reversal theory together can help connect social and individual worlds:

> When we come to examine social processes, the interest in (reversal theory) is in the way the motivational states play out in society – how they are induced, harnessed, expressed, encouraged, satisfied and symbolized… Social processes impose themselves on the individual…. At the same time, the structures of experience, as identified through reversal theory, are essential for a full understanding of social institutions…. We therefore see a two-way feedback process between the individual and the social (188).

Symbols – metaphor - have both social meaning and personal impact. Metaphors invoke motivational states; changing metaphors can also change states, and therefore emotion. To make this connection for data sharing, the following table matches groups of metaphors

against different states of reversal theory, and notes the source (social or personal) and where social, the perceived effectiveness by interviewees.

Table 6: Reversal Theory States and Metaphor

Metaphor	States Invoked and Effectiveness
Large scale metaphors indicating data overload, and signaling the need for informatics tools: Explosions, tsunamis, mountains of samples, oceans of data, moving at warp speed.	**Serious, Conforming, Mastery** – Researchers must use bioinformatics tools in order to manage the influx of data and prevent being overwhelmed. Social message not seen as compelling.
Small scale metaphors about incentives and disincentives: using "carrots and sticks" and "scorecards" to incentivize data sharing or punish a lack of sharing	**Serious, Conforming** – Meaningful incentives for sharing and penalties for not sharing (unlike those perceived now) would create Serious consequences for actions that don't support a norm of sharing. This is a metaphor used at a personal level to describe the actions that would need to be taken for data sharing to be compelling.
Small scale metaphors about exerting control over data: wringing or squeezing the value out of data, beating data set to death.	**Mastery, Self** – Researchers need to be sure to extract value from data before making it available. Metaphor used at individual level, but not social.
Small scale metaphor about getting "scooped:" someone publishing first on someone else's data set.	**Serious, Mastery, Self** – Researchers must "clean" their data and extract value out of it before making it available to others to avoid having someone find something they did not. This is a powerful metaphor at the individual level, and was seen also at the social one.
Metaphor about relationships in data sharing: data clubs, dating, scientific marriages, having friends on the grid, making a baby.	**Conforming, Sympathy, Self and Other –** Relationship building and trust are essential to allowing data sharing – and are the "antidote" to being "scooped." These are strong metaphors at the individual level. Seen less often and more abstractly at the social level: "embracing diversity" and "data speaking to data."

Metaphor	States Invoked and Effectiveness
General traffic metaphors referring to ineffectiveness of data sharing policy: going above 30 when one shouldn't on the street; policy isn't enforced.	**Playful, Rebellious –** Seeing little consequence in Data Sharing Policy, acknowledging breaking the rules. Metaphor used at individual level.

Motivational diversity and balance is considered an important dynamic of reversal theory: it is of benefit psychologically to experience all the different states over time; and to be able to exert the self management to match states to situations appropriately so that positive emotions are felt rather than negative ones. This is the purpose of reversals; if a person is in a specific state that is not supporting positive emotions, he or she can self-create a reversal to move to a more positive one. This can be done by changing the situation or changing the personal meaning of that situation, which can be done through the reframing often created by a new metaphor. This was done most clearly by the interviewee who reframed the "wave of science" to be "bouncing around in the kiddie pool," a clear reversal from Serious to the opposing Playful state, and a likely reversal from Mastery (invoking power and ability) to Sympathy (invoking connection and care as we play in the pool).

This idea can be used as a foundation for advocating a range of different communication strategies – and institutional changes in the incentive structure - related to data sharing at the social level. As noted previously, many of the metaphors related to data sharing fall into the Serious and Mastery categories, manifested as a concern for maintaining control and competence so as to not be scooped by others. At the same time, however, the Conforming and Sympathy states can also be powerful forces if used to invoke feelings of both connection

and a bit of peer pressure between trusted colleagues. On the other hand, more generally broad metaphors and imagery – the overtaking of science by a large data set, the importance of patient care in general – do not appear to have the intended impact.

7.5 Altering States: Recommendations

The patterns revealed through reversal theory point to possibilities for altering and broadening social messaging about data sharing in order to invoke a broader range of positive states, and reframing perceived risks in a more positive light. Many of these social messaging changes would require institutional changes as a foundation, in order to ensure that the messaging matches the reality. Examples could include:

- Converting concerns about being scooped to the confidence and pride that comes from others wanting to use the data that a researcher holds, therefore signifying its importance (converting Mastery losing to Mastery winning). This could be illustrated by conveying case studies of former – or potential - competitors that have come together to specifically share data for specific projects, demonstrating the trust that has developed between them, which ultimately discourages scooping between mutual colleagues. The purpose of this project would be to make the grid mechanism secondary to the actual sharing between colleagues. In this reframe, grid sharing is not the point; it is simply a mechanism for allowing sharing between collaborators that were previously competitors. While routed in the Mastery state, it would also invoke the Sympathy-oriented elements of collaboration. Regardless,

the point is that the discussion would become about the people sharing data, rather than vehicle by which the sharing is done.

- Communicating the importance of data sharing and *consequences* of not sharing at the research program level; this would require that program sponsors at NIH both practice and invoke the Serious and Conforming states. This would require ensuring that meaningful data sharing actually occurs as a part of research projects, and that this sharing is made visible as a positive example to other researchers.
- Reframing and extending the Sympathy Other oriented concern for patient care to a more personal level, highlighting and facilitating the collaborative side of science by ensuring that research oriented gatherings include adequate time and structures for networking and informal information sharing, so that people with common interests have the time to meet and identify potential point of interest for future efforts (leaving time for Sympathy Other at a proximity that is close to researchers than patients).
- Increasing knowledge and training related to the labor and skills needed to accomplish the technical dimensions of data sharing in the new technology-driven environment, so that researchers understand who will need to be brought on board for the project, and what needs to written into grants to cover this labor involved in preparing data to be shared. The goal of this would be to make the labor of data sharing more visible and rewarded through the grant itself, and would also help in educating the field as a whole about the new techniques and expertise needed. This would invoke Conforming and Mastery Self states in a different way than is currently

invoked – increasing the institutionalization of data sharing activities and practices and ensuring that the skills are visibly present and the labor rewarded for doing so. If combined with learning through sharing data sets that are perceived as unimportant to the researcher (e.g., already "squeezed out" but of potential use to someone else, and therefore low risk), it could invoke the Playful state: there's no consequence of sharing, I'm just gaining some experience.

- Hosting "sandbox" workshops that allow researchers to experiment and "play" with the new technology tools and low impact (to them) data that can maximize the benefits that come with data sharing. These demonstrations would show in a hands-on way the possibilities associated with data sharing on a larger scale than point-to-point sharing. This would invoke the Playful state – exploring sharing without consequence, to generate future possibilities for data sharing in the mind of the researcher. Again, if targeted towards data perceived as low risk, it could also start to help lower the barriers to entry for making it available.
- Creating "safe havens" for data sharing among communities where these is already trust; laying grid sharing within established communities where there is already the motive to share. Establishing a protected place where people can share data for specific goals and even gain specific benefits from doing so that others do not enjoy may help set both precedence and provide examples. This action is in fact already underway in the early adopters of caBIG, both at organizational and consortium level. Grid sharing works when I am in a special sandbox with my friends.

- Formalizing the recognition of data sets as a source of scientific value at a comparable level to publications, and altering incentive structures within the academic system to ensure that the generation and sharing of meaningful data sets are rewarded. This would convert being scooped into being recognized, would make invisible labor more visible, and would help advocate the sharing that is currently happening on a far more personally-driven basis. What is currently Rebellious Mastery losing would be converted to Conforming Mastery winning, because the effort that is wanted would actually be rewarded, rather than advocated in what is perceived to be an unsupported way.

Of all the recommendations here, the final one is the one most cited by interviewees as having the highest potential impact, where change is most badly needed.

It could be easily argued that these recommendations were suggested already by the previous analysis, and that nothing has been added necessarily by this dimension. This is likely true in this case; however, the larger goal here is to test the use of reversal theory as a practical tool in the research community for analysis and action planning. In very real ways, reversal theory provides a somewhat modernist (universally abstract and efficient) structure that can systematically be applied to a range of problems to better describe what is going on motivationally, to assess blind spots in current thinking, and to subsequently brainstorm possible courses of action that have not yet been identified. In this case, reversal theory was helpful in uncovering the overemphasis in the social messaging on ends (patient welfare, science in general) rather than means (the pleasure of doing research, a major motivator for scientists); in uncovering the importance of personal relationships in the data sharing choice;

and in uncovering the apparent omission of the importance of collaborative relationships in both the norms and counter norms.

7.6 The Scale of the State

As noted previously, this analysis reveals a dynamic of reversal theory that has not previously been explored: the role of scale and proximity in assessing the effectiveness of state imagery in invoking a desired emotional response. Some metaphors are large in scale and not in close proximity to the personal researcher: "explosions, tsunamis;" others are smaller, and more directly related to a person: "data clubs," and "dates." Scale and proximity appear to influence how effective the metaphor or justification is perceived to be.

For example, the Other state in reversal theory represents the feeling of being motivated by the needs of someone other than oneself. What is both consistent and compelling in the discussion of data sharing is that closer forms of "other" seem more consistently important than those that are more conceptual or distant. For example, data sharing to benefit another researcher or consortium seems to be a compelling reason to share data; data sharing to benefit patients in general or humanity at large is less so. As a second example, goal achievement (an indicator of the Serious state) is compelling as a motivational argument when discussing addressing specific research questions and one's ability to win grants and have papers published; it seems less motivating on the macro level of overcoming massive volumes of data or more broadly "winning the war on cancer." This pattern can be demonstrated across all scales as demonstrated in the following table. (Note: Those marked

with a star * are not perceived to be true in the current environment, but interviewees agreed that they *would* be compelling if they were true.)

Table 7: The Effectiveness of Scale Focus

State	Small Scale, Near Proximity – Effective	Large Scale, More Distant – Less Effective
Serious	Sharing data is important for career advancement *	Sharing data is important to achieve scientific advances
Playful	Emerging bioinformatics tools are interesting and will help you explore new areas	The bionformatics field allows for the extension of science into endless areas of exploration
Conforming	Creating interoperable data sets will save time over the course of your long-term project	Interoperable data sets facilitate large-scale data sharing
Rebellious	Data sharing will separate you from the rest of the pack *	Data sharing is needed to revolutionize science
Mastery	Data sharing will help you compete for grants and get published *	Data sharing makes science better
Sympathy	Data sharing develops trust with other researchers, leading to long-term collaborative relationships.	Data sharing is needed to support the development of cures for patients
Self	If you share data now, you'll probably get something back later	Data sharing increases your visibility and success
Other	Data sharing will help another researcher who needs you	Data sharing supports the greater good

This distinction of scale is not one that has been explored by reversal theory researchers, and deserves to be considered more closely in future work in using this model in the area of both personal coaching and the broader application of strategic communications. This element also has implications for translating the theoretical tool of actor network theory

into a tactical tool for those communicating with scientists. If the activist's goal is to encourage the movement of quasi-objects across a network; in this case, the movement of data between actor-objects, relocating discourse and metaphor to a level that is seen as actionable and of direct benefit to a researcher across the different states might help translate theory into a model for action.

It will likely be noticed that the smaller scale statements presented here far more often call on the Self state combined with states from other domains rather than the Other state. While it would be nice to theorize that there is more than self-interest at play here, that does seem to be one of the implications of this scale comparison exercise: self-interest is in fact a significant driver in data sharing. There is nothing in the previous analysis that necessarily contradicts this. It does however, suggest future areas for reversal theory research: how do the Self and Other states play out across scales of action?

Where individuals are considered and focused on as the central element of communication, emotions themselves become a quasi-object that can either help or hinder data movement. Good feelings such as trust and pride may come from joining with another researcher to successfully pursue a grant; that good feeling may cause one of the researchers to enter into another collaborative relationship based on this success, this time involving even more data. The positive motives and emotions increase data sharing across this emerging "club" in a series of "dates that lead to a marriage." At a micro-level, the opposite can occur as well. If a researcher exposes data and then is "scooped," it can lead to negative motives and emotion, which stops the flow of quasi-objects (connections and data) across the network.

Personal feelings can help drive social change, but only if the social structures help encourage the flow.

Chapter 8: Conclusions and Directions

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. I am advocating, therefore, that the caBIG program and others like it 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 (rather than technology-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.

Socio-cultural elements are currently positioned by caBIG as the innermost layer of a complex onion, to be encountered after other factors such as technology, legal-regulatory concerns, and economic concerns have been at least partially peeled away. Instead, the importance of people and relationships should be reframed as a primary factor, so that the discourse becomes as much about the people with valuable data to share, as the technology that will be used to share it.

This research has purposefully not engaged in statistics related to data sharing, because of how misleading these variables can be. It is impossible to know the true scientific benefits that have come from past and current data sharing; and the scientific cost that has come from data being withheld. It is also impossible to know the transaction costs that would have been incurred if more data were made generally available for sharing, both in terms of the labor that would have been required to do so and the costs involved in managing and mining that data for potential useful information (matching data to the need). One 2002 survey reported one perspective of the current state of data sharing in genetics:

> Our survey found that data withholding is fairly common in academic genetics. In the last 3 years almost half of academic geneticists (47%) had been denied access to published information, data, and materials by other academic scientists. However, only 12% of scientists reported engaging in such behavior.
>
> Biomaterials were the resource most commonly withheld. Of people seeking access to biomaterials referred to in published work (such as the cell lines, tissues, antibodies, and reagents), 35% were denied access to these materials at least once in the last 3 years. Other forms of data that were less frequently denied were unpublished phenotypic information, information about lab procedures, and pertinent findings not included in a paper.
>
> The motivations most frequently cited by investigators who withheld data were that sharing required too much effort (80%) and that scientists needed to protect the ability of a graduate student, postdoctoral fellow, or junior faculty member to publish (64%). About half (53%) denied requests for data in order to protect their own ability to publish in the future. Nearly half (45%) withheld data because of the financial costs of providing the requested information or materials (Campbell and Blumenthal, Par. 6-10).

Another study concluded that, “Researchers who were most likely to be victims of data withholding were those who have withheld research results from others, published more than

20 articles in the last 3 years, to have applied for a patent, or spent more than 40 hours per week in research activities" (Campbell et al 303).

While the reasons given for not sharing were all also seen in this study's interviews, the interviews also revealed a more personal and emotional face: the desired pride of achievement, the protectiveness associated with holding a precious resource, and the enjoyment of shared relationships with other researchers. The interviews also reveal why the statistics above may be misleading. Interviewees in this study who are in a position to share data report doing so under well-defined circumstances. They also report *not* doing so under other circumstances. Just because someone withheld data in one circumstance does not mean that they withheld in another, leading perhaps to the apparent statistical disconnect between those who have been not shared with, and those who report having not shared in the statistics above. In the end, based on the research here, any one researcher could easily fall both into the group that shared, and the group that did not.

The term "victim" is also a pejorative one in the second quote above. What if a "victim" of data withholding was in fact a researcher who was known to "scoop" others, or withhold data him- or herself? Would that reverse the perception from sympathy (suggested by the term victimhood) to more of a sense of mastery-based fairness and even vindication? Sharing is personal, and context is vitally important to understanding the pragmatic reality behind the statistics and judging metaphors.

There are well understood and justified reasons for not sharing data, and while many of these may be "written off" at the social level as selfish concerns about career success, if one accepts the premise that the success of individual researchers also contributes to the success of

science overall, it is clear that data sharing is a shared personal and social problem that needs attention at institutional levels.

There are a myriad of actions that could be taken to help better align social expectations for data sharing with the personal decision to do so; some of these have already been referenced in previous chapters. Here, three different recommendations are presented as entry points to starting the changes that must occur for data sharing to occur on the broader scale hoped for by programs such as caBIG:

- Altering incentive and rewards structures related to data sharing
- Making more visible the labor and specialized skills required to institutionalize data sharing
- Refocusing the emphasis on data sharing to focus on specific collaborations and communities rather than general communalism

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. The goal is ultimately to better integrate the personal with the social. Siemsen et al suggest that data sharing happens best when there is an intersection of motivation, opportunity and ability: I want to, there is a reason to, and I am able to do it. All three emerged in interview comments in the course of this research, and are embedded within the recommendations that follow.

8.1 Altering Incentive and Rewards Structures

Much of the literature about data sharing (National Research Council; Gardner; Arzberger et al; Wasko et al; Piwowar, "Towards a Data Sharing Culture"; Kaye et al) and many of the interviews conducted for this research pointed to the need for revised incentives and rewards for data sharing in the academic research community. There are two elements embedded in this recommendation. First, the contribution of a data set must be traceable to the researcher(s) who generated it, and to support traceability, the citation of that data set must become as much a part of practice as citing published papers. Second, the meaningful contribution of datasets must be first incentivized, and then rewarded across the institutions of science. Implementing both parts of this recommendation would require and introduce fundamental changes in how scientific work is identified, tracked, and recognized.

In terms of ownership and citations, while data is generally legally owned by organizations that fund researchers, rather than specific researchers as individuals, the attachment of a researcher's name to a data set is a reasonable equivalent of an academic form of ownership. Raban and Rafaeli conducted a laboratory experiment about data sharing, using a series of controlled variables and then observing the sharing or withholding of data against these variables. The study found that, "sharing was higher for privately owned expertise than it was for organizationally owned content. Ownership makes a difference. It serves to increase sharing of information. Ownership can and should be framed by system design (2380)."

An institutionally-based "system design" for data sharing might include the addition of specific structures for referencing, citing, and then subsequently tracking the use of data sets that are formally associated with the researcher(s) generating them, even if those data sets are

combined from multiple sources. The end outcome of this would be that publications would not only include a section of cited papers, but also a section for the data sets used to generate the findings. Instead of embedding data references in methods sections and in acknowledgements, data sets would stand on their own in a structured way, allowing for the creation of the citation indices that are today so popular for determining the "value" of scientific publications. "Classification does indeed have its consequences," write Bowker and Star (319), and the consequence of making visible data sets in a structured, systematic and acknowledged (e.g., rewarded) way would demonstrate their importance in the changing landscape of biomedical sciences. This type of system would also help eventually distinguish the data sets of greatest use and interest, through citation index methods, hopefully leading to academic forms of credit that are currently limited to the realm of published papers.

The benefit of this recommendation is that it would place data sets on a similar level as papers themselves, elevating both their visibility and perceived importance. It also might support a selection and sorting process allowing for the ranking of the sources and types of data sets perceived as most useful or valuable by those using them. This might also help with some of the "data management overload" problem that some researchers cited as a concern if large amounts of raw data suddenly start being posted that others must then search through. This approach, of course, is still limited in its scope because it would only report the data sets used in the research activities that lead to published papers. Expanded, however, it could ultimately be used to elevate certain data repositories outside the publication process.

There are, of course, enormous logistical challenges in this type of institutionalized system and practice. It would require standard ways of identifying, storing, and accessing data

sets, such as through central repositories as mandated by existing publications or through federated systems like caBIG; and identification techniques, such as Creative Commons licensing discussed previously, to communicate how the data can be used and the protections that are associated with it. Journals and scientific societies and communities would need to agree to add a data-focused section to published papers, and researchers would need to implement the practice in their writing. Despite these difficulties, for researchers who spend a significant amount of time producing and/or integrating data sets, this is a step that would begin the process of recognizing effort that now goes unseen.

Once these data sets are more visible in terms of their contribution to the scientific process, they would need to be rewarded on the other side of that use. In addition to being "housed" in a section in published papers, data sets would also need to be made visible, and be counted as valuable, in the curriculum vitae (CV) of researchers for whom data sets are an important element of academic production. One researcher notes in a follow up e-mail after our interview,

> A thought on the whole "incentives" issue…I was wondering what mechanism various entities like Promotion and Tenure Committees could use to "assess" the value of datasets shared/contributed and had one thought: If the investigator included a section on the CV with the datasets shared, and in some way could easily assemble publications of OTHER investigators that resulted from use of that dataset, that would really help. I don't know if caBIG has built in any kind of mechanism of "registering" who uses datasets or requirement for reporting to NCI publications resulting from use of datasets to facilitate tracking what "value" datasets have. That would be a good project.

The generation and sharing of data sets must become a part of the academic evaluation process, including the tenure evaluation process, before these activities are seen as incentivized enough to warrant extra investment at the individual level. Who would need to act to

introduce this change, to make data sharing a visible element in the "scorecard" of academic career management? There was agreement among interviewees that, in academic research, this responsibility lies directly with University Deans and Presidents. One NCI representative notes, "We've moved from reductionist approaches to team science, and that needs to be rewarded. It lies ultimately with the Deans and University Presidents." Who could influence these leaders? Most agreed that the leaders of funding agencies, such as the NIH, the National Science Foundation, and other large funding organizations and foundations could do it, but no one at the research level appeared to know whether these conversations are happening or not.

Extending beyond and above funding agencies and universities, a university Principal Investigator believes (Interview question in italics):

> It is going to take messages from the very top to incentivize sharing, to drive the culture change. *Who can take the right first step? Is there somebody that holds the key?* Well, Barack Obama, I think? [laughs] He holds a lot of keys right now for a lot of things. Probably somebody like Tom Daschle.[19] It would have to come from that level, and then you would need to have somebody that could look at all these issues. You need to bring together pharmaceutical companies that have their own issues with intellectual capital, with all the other institutions that are involved in these kinds of things. It has to come from the top. In Europe, governments set the rules for how things are going to be done. Here, it's not as dictated like that. Here, we need to pull together lots of people to set the game to move forward. It's going to take 5, 10 years to make a difference.

One manager who coordinates a biobank notes that data sharing will become far more common when the government itself shares among itself, and mandates that others do so too. This interviewee cited a number of government-held biobanks that hold highly valuable

[19] Tom Daschle is a former US Senator who in early 2009 was briefly under consideration as nominee for Secretary of Health and Human Services (HHS) under President Barack Obama. This interview was conducted after his nomination, but before Daschle withdrew his name from consideration.

specimens, but which are not shared even when directly asked. "If some of these really high profile resources were opened up," it was stated, "that would change the game. Someone in charge has to just make it happen."

Incentivizing and rewarding data sharing was one of the most called for actions that emerged during this research project. This will require a top-down effort formalizing and institutionalizing data sharing by making the activity more visible, and by then subsequently rewarding this visibility. Without these steps, researchers may continue to be chided in editorials for not sharing, but will likely continue their current practices of point to point sharing, where the reward is clear but where the exchange is contained within a trusted relationship, rather than a public trust.

8.2 Increasing the Visibility and Professionalization of Data Sharing Labor

Another common theme in this research was the perceived burden of labor in preparing data for sharing, both in terms of time and skills. Many classified the labor involved in data sharing as an "unfunded mandate" – it is invisible labor that often goes unacknowledged and uncompensated, if it happens at all. More often, given that point to point data sharing is the more frequent mode of data sharing, the labor involved to fully prepare data for potentially broader standardized sharing (e.g., over interoperable tools such as those provided by caBIG) is never performed at all, if it is not framed as a part of the project to begin with.

To broaden the scope and range of voluntary data sharing, two actions need to occur, in close connection with the first recommendation above. First, the scope and type of labor associated with data sharing would need to be clearly defined and understood, so that the cost

of that labor could be both articulated and reflected in grant applications and other economic measures of scientific labor. It is important to realize that although researchers refer to *their* sharing of data, it is not typically the researcher who actually prepares the data for this activity; there likely are others that would be assigned this task. While caBIG tools are designed to remove some of the labor by introducing interoperable tools through which data can be shared with minimal additional investment, there are always likely to be some transaction costs associated with preparing data to be shared. These are costs that must be taken into account in order to be rewarded, so that the labor does not continue to be as invisible as it currently is.

Some interviewees noted that this is already happening in the institutions that have adopted caBIG internally. Often, these are larger organizations that have matrixed resources available to support scientific departments in data sharing. One researcher at a smaller Cancer Center pointed to some of the most successful caBIG deployments, "They are happening at the large matrixed organizations, where there are resources to do this stuff, and there are large grants that encourage scientists to do it. Sure, I'd love to make all this happen, but look at them, they have 10-12 people working on the team. I don't have those kinds of resources. You can't just fund one person and expect this stuff to happen."

The recommendation above would measure outcomes; this recommendation would measure process. "We tend to reward what we can count," says one government official. The labor associated with data sharing must be quantified and made both visible and public before the true cost of this activity can be planned and acknowledged in compensation and rewards.

Second, once the labor is made visible and rewarded as a part of the scientific process, it is vital that the labor and associated skills are *available* to perform the tasks required.

Unfortunately, these are also the professionals deemed as “parasites” in one of editorials cited previously; bioinformaticians that take the data provided by others to make them shareable and more broadly used. For those that find grid-based approaches to data sharing “insulting,” it will take time and new relationships to overcome the motivational barriers associated with data sharing. Professionals with the skills to help demonstrate the benefits of grid-based data sharing for higher-payoff kinds of scientific work, and more importantly, who could take on the technological burden of this effort will grow in demand as the incentives to share are put in place and made more clear.

Despite the overarching focus on cancer research explored through this study, there are multiple colliding and overlapping social worlds here that do not see fully eye-to-eye. When done well and visibly, as demonstrated by programs like caBIG, the marriage of science with bioinformatics begins a brand new inter-world “dating” process, with new research relationships spanning a broader range of the bench to bedside continuum, and connecting previously separate social worlds. It will take both education and outreach to convert the metaphor from “parasite” to “symbiont,” overcoming the negative perceptions about data sharing in interoperable grid environments.

Those on the technology side of the cancer research spectrum are building careers by creating the vocabularies and tools that facilitate interoperable data sharing in a meaningful and streamlined way. STS has identified many case studies where the integration of labor with new technology has led to the emergence of new professions and professional standards. Biomedical informatics and data management subspecialties are areas that continue to define themselves in this way, and the professionals in these specialties need to continue to work to

establish the value of biomedical informatics technologies in the more established research communities. As demonstrated above, a "if we build it, they will come" positioning will not necessarily make interoperable grid-based technologies of caBIG as inevitable as hoped.

There are, of course, challenges here: new professions can threaten to make obsolete existing ones, and new skills need to be developed and distributed as demand for them grows. Technological momentum, if and when it exists, cannot proceed without the people with the skills to make that technology work. The caBIG vision is only possible through the community of people that turn it into reality; this labor must be acknowledged and rewarded.

Clearer incentives and rewards (with appropriate carrots, sticks, and scorecards) for data sharing, as discussed above, will likely continue to expand the demand for technologies such as those offered by programs like caBIG, which will in turn continue to support the demand for and emergence of new specialties, which are made up people with specific skills and motives. Participants in caBIG note that there is a great need for people who can provide the scientific consulting services that match available bioinformatics tools to the scientific questions that need answering, but that a market for these "social world bridge services" has not yet evolved – there must be both more demand for the capability and more supply of the labor and skills, before these new inter-disciplinary clubs begin to evolve and mature.

The caBIG program actively recognizes the need for training and education in its tools and infrastructure, regularly providing "boot camps" and online training for its resources. Continued training, education, and professional development activities to both position the need for data sharing and to support the development of the skills that will make it possible, are necessary elements to create the cultural change that was so broadly called for in this

research. An additional dimension, however, lies in helping to establish the "blind dates" that can help these diverse professionals meet, in order to craft new kinds of collaborative relationships that both maximize the infrastructure provided by caBIG, and the data that could flow across it.

8.3 Moving from Communalism to Collaborations and Communities

Data sharing, while becoming more visible as a need in cancer research, is not yet seen by interviewees as an ethical obligation. This is an apparent indicator of the counter norms of science winning over the norms of science, as the norms, specifically the idea of communalism, would suggest that data sharing is an ethical act to support the public good, with sharing as a part of the scientist's obligation to the field and peers. With the absence of the ethical dimension of large scale data sharing, what remains is the inter-personal obligation and accountability that leads to and sustains trust between individuals, the inter-personal counter norm established above. It is one thing to refuse to share data on an anonymous grid; it is another to refuse to share data with a colleague who helped out on a recent project. Ethics are exercised between people, not on grids.

The Mertonian ideal stresses *communalism*; cancer research occurs in *collaborations* and *communities*. Overcoming the fear of being scooped, being willing to take the risk of sharing to begin a relationship, investing the time to build scientific trust: all of these are motivational barriers that must be overcome at a personal level in order to maximize the potential of the institutional recommendations above. The caBIG program started this process by bringing together researchers from academic cancer centers that had not yet interacted

before; this initiated collaboration and point-to-point sharing that had not previously been done. Through this, and almost *despite* its interest-based focus on grid technologies, caBIG demonstrated the value of building communities as a precursor for building networks.

As institutional grid technologies and capabilities mature, however, it is important maintain this community focus. Interviewees noted that the caBIG program needs to emphasize its benefits for specific tasks in terms recognizable to specific targeted communities, to ensure that the program does not get "so big" that people can no longer see the potential positive impact at a local and individual level. How can caBIG help demonstrate the ability to "squeeze sponges" of data, as well as manage tsunamis? Right now, people are implementing caBIG tools and connecting to the grid, but data is not flowing across that grid in the broad-based way that was envisioned. With a focus on technological grids and security trust fabrics, interviewees report that the tactical benefits of sharing the data for specific research communities have been somewhat lost in the technological and legal layers of the onion.

The caBIG program has focused on technology development as a path to data sharing, a natural consequence of the program's organizational positioning at NCI. It is, as a contrast, useful to consider an alternative example where specific and broad data sharing in cancer research has been successful *without* the benefit of any specific tools to support it. This example was shared by a representative of a commercial pharmaceutical company, who is close to a group called the "CEO Roundtable on Cancer." The Roundtable focuses on resolving issues common to a group of normally highly competitive pharmaceutical companies. These companies came together with each other and with representatives from academic cancer centers and NIH, with the blessing of the Justice Department, to share data in order to develop

a set of common contract clauses that could be standardized across contracts between these commercial organizations and academic centers and other organizations. The agreed-upon benefit of this effort's outcome was to streamline and dramatically decrease the amount of time required for drug trial contracting processes. "This was not high science," said the interviewee, "It was actually quite mundane in the end, but the idea of these companies all coming together and sharing data to do this was tremendous, and the benefits will be seen in huge savings in time and money."

This was broad-based data sharing across typically competitive organizations that has achieved great benefits, and was relatively low risk; it just took the step of agreeing to do it, and the willingness to provide the labor to then sift through the data to find the common elements that could be shared.

A previous project that this consortium took on is even more compelling in revealing specific criteria that could be used to identify other large-scale demonstration data sharing efforts. The consortium's first project was to share information about the participating companies' employee benefits programs in the area of cancer prevention, and developing a shared "gold standard" model benefits program. As with the "common contract clause" project, the target of sharing (in this case, data related to benefits programs) carried two key characteristics: (1) there was a perceived high payoff from sharing that could not be achieved without that sharing, and (2) there was a very low risk of sharing that data to start the process. "Getting scooped" on this data was not a significant concern; this made the motivational "barrier to entry" for the data sharing project quite low.

Two guidelines for starting a data sharing project flow from these examples. First, locate a research community that has a shared need that, if filled, would demonstrate a clear benefit across the community, even if that community is an internally competitive one. Second, find a type of data associated with that need that could be shared with very little investment and risk, thereby lowering the pain of making that data available for sharing. Finding a low risk data sharing opportunity that can benefit an intact, even if competitive, community is therefore a proven approach that could be used to demonstrate the power of sharing. In STS terms, find a way to make the Panopticon itself a place of safety and payoff, and make the perceived price of surveillance low.

The caBIG program successfully locked into the first criteria identified above – identifying shared needs – when the program started. It did this by prioritizing tools development efforts to meet the greatest needs across the academic cancer center data community. This is part of what led to the initial success of the community building that interviewees spoke so positively about. A next step, to lock in the second criteria (low risk of sharing), is to locate a specific kind of data set that would be perceived as easy or low risk to share, where multiple centers could contribute data to lead to a clear shared payoff for the community as a whole.

In some ways, it is unfortunate that one of the most popular early tools from caBIG was its biospecimens management tool, caTissue. This software tool filled a very specific site-specific need across the cancer center community, but sharing data *associated* with the tool across centers involves data that is perceived as highly valuable and legally sensitive. This discourages the sharing that might have otherwise occurred with a data type perceived to be

less risky. The distinction is clear: tool sharing has been successful because it has filled a shared need; data sharing has been less successful because of the higher risks and barriers involved. The same principle holds for caBIG's clinical trials software suite: the tools may be popular for internal use due to site and trial specific needs, but due to HIPAA and IRB restrictions and fears, allowing clinical data to flow through them with others is more problematic.

Would a case study in genomics be more successful, given the lower potential risk associated with the data? To meet the shared criterion above, there would need to be a shared problem that would have such benefits if solved that people would accept the risks of sharing in order to achieve it. This was, in fact, the case of the Human Genome Project; not only were people incentivized to share, but it was perceived as such a high impact project, the risks were lower. This was a unifying project with clear rewards and well-structured processes. What may be needed is another unifying problem. For the commercial side, this has played out in common contract clauses, and now, in the quest for well-defined biomarkers during the pre-competitive phase of drug development. What is an equivalent high impact need among academics in the various "omics"?

In a final note related to community development, it is important to acknowledge that the downside of "my metadata talk to your metadata." Metadata alone leave the people behind. Data is both subjective and personal; it is often an extension of 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 the 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 completely lost to the technology that will move it to the next place? "A cyborg is a cybernetic organism, a hybrid of machine and organism" (Haraway 149). If a cyborg integrates technology within a human, then how might we integrate the human onto the grid - so that we are interhuman, as well as interoperable?

Refocusing communication to acknowledge the local and personal nature of knowledge, and ensuring that grid technologies do not devalue personal communities, are next steps forward. Overtly supporting the development of inter-personal connection and trust will allow programs like caBIG to maximize both the benefits of collaboration, and the benefits of data sharing, leading to the "collaborative action" that lies at the intersection of both.

8.4 Opportunities for Future Research

The fact that there are only three core recommendations in this final chapter is not meant to underemphasize the enormous personal and social changes that must occur for open data sharing to become a norm rather than an ideal in cancer research, or in any other field. Continued policy and legal work to raise researcher concerns related to HIPAA; a better understanding of the differences in perceptions between patients and researchers related to data sharing; and understanding the personal, social and technical implications of the Health IT policy agendas ahead are all areas in which lessons and tools from STS can, and should, have a voice.

In particular, a necessary but unfortunate limitation of this research was its tight focus on bioinformatics and cancer research. This omitted any number of other social worlds struggling with data sharing issues and decisions, both inside and outside science. Today's

technologies introduce a variety of questions related to data ownership, privacy, and the motivation to share. Are the trends discovered here consistent with other fields? If so, what are the common factors that contribute to either sharing or withholding? Is the category of "data" a useful boundary object in other fields that are also becoming as inter- and trans-disciplinary as cancer research? Could an analysis of what is counted "in" and "out" of the category of "data" help define the boundaries of conceptual social worlds in a practical and pragmatic way? How does a study of data within a field help STS scholars expand its understanding of labor issues in an information economy?

Another natural extension of this research would be into the cancer patient and patient advocate perspective. Given that patients are ultimately the ones who consent to have data enter into the research cycle - in the form of clinical information, personal histories, treatment courses and outcomes, and tissues themselves – their motives for sharing or not sharing certain types of data is an area of investigation worth exploring. Informal interactions with a range of patient advocates indicates that some of them are strong advocates of personal data sharing for the betterment of research; others are far more wary of widespread sharing that could compromise their control and autonomy. As in the researcher community, both views have sound foundations; what drives motivation one way or another is a research question worth exploring. Just as there is diversity among researchers, there is diversity among patients – understanding the diversity within groups as well as between them might help shape advocacy approaches in the future from researchers to patients, and from patients to researchers.

This research has attempted to stay at the boundary between the individual and social to better understand the intersection between individual motivation and social messaging and

norms. Most interviewed agreed that attitudes about data sharing come from a mix of personality, past experiences and influences, and the interests and reward structure of the environment one is in. It was impossible to separate these variables to determine which factors are greater influences; a conclusion is that there is ultimately no definable intersection between the individual and the social to locate.

The lesson here is that influencing social worlds includes influencing individual hearts and minds. The messaging that "data sharing is good for science" must be accompanied by the reasons that "data sharing is good for you." The same integration between individual and social dimensions should apply to the way in which data is treated: data is personal; data and the richness of its context and meaning cannot be fully separated from the subject that generated it. Grid technologies will be most successful if they can integrate both subjects and objects.

There are also continued opportunities to explore the role of gender and other sources of difference in attitudes towards data sharing. In this research, neither men nor women were more or less likely to have shared data, and demographically, men and women were generally equally distributed across different jobs and roles. As discussed previously, the only subtle difference between men and women in interviews occurred in the use of metaphor: women used the metaphors dating, marriage, and pregnancy; men did not. This does not suggest that men do not find the relationship element of data sharing of importance; several men referred to the trust and connections between researchers that they appreciate. The only distinction was in the metaphor use itself. Other sources of difference were not explored during this research; assessing differences in data sharing attitudes in different cultural settings and socio-economic climates would be productive avenues of research.

As a final observation related to the field of STS, this was a research project conducted through participant observation. As noted at the beginning of Chapter 1, I have been a consultant to caBIG since 2006. This was of great benefit in accessing materials and research subjects for this work, but was also at times difficult to navigate. Even in this writing, I struggled with drafting statements that may appear critical of the program I work hard daily to support, with setting previous assumptions and judgments aside to stay open to each interviewee's perspective, and finally, to establish a comfort level with interviewees such that they would disclose personal feelings despite my affiliation with the caBIG program team and NCI. While background reading on ethnographic methods was useful in anticipating these challenges, a practical, more personal "do's and don'ts" field guide from those who had tread this ground before might be of value for future researchers in this position.

8.5 Conclusion: Cancer Research as a Creative Commons

When conceived as a win-lose-scoop proposition, data sharing reflects the possible transfer of value from one person to another. When conceived as an act between collaborators, data sharing reflects a creative act that builds and creates relationships and interpersonal connection. Programs like caBIG want to transform data sharing into action on a larger scale, across relationships, where value is magnified because of the possibility of multiple forms of data use to answer a growing range of scientific questions. Unfortunately, the metaphor of sharing that positions data as a public good, also hides the economic realities of the activity, the labor associated with generating the data and then preparing it for sharing, the potential loss of revealing one's proprietary secrets, and the emotional cost of being scooped

on another's journey to publication. There are many conceptual and actionable steps that must be taken to shift this paradigm. One possible step is to accompany the act of data sharing with a more value sensitive conception of "shares" as elements of scientific knowledge, shares that can be both assigned ownership and exchanged for value. If one takes an abundance position, investing shares leads in time to more accumulation, both of personal achievement and of scientific knowledge. This idea reflects the reality that it may be time for a new "generation of metaphor" that begins to "turn the tide" of the data sharing debate.

Data sharing takes trust and faith: faith in oneself, trust in others, and faith in the future. One interviewee captured best both the fear and the potential:

> No one wants to be the first, or, the cost of being the first is high, but the cost of whatever the action is drops as more and more people do it. It's a network effect... It doesn't do you any good; it does everyone else good once you have done it, because then, people know how to do it. Ultimately, the more people that become involved, the more power you get. The basic way to get around that is agreeing all at once to take it on.

In reversal theory terms, the rebellious state (the state where it feels good to explore something that no one else has) - joined with a sense of ability, a concern for others, and the personal confidence to take the first step - can start a process that leads a community to a whole new level, and consequently a whole new set of norms. A caBIG conference speaker described the early stages of the project before caBIG had completed its pilot phase, "Pioneers get the arrows," he said, "settlers get the land." While this may not have been the most politically correct statement, it reflects how the founders of the project and the vision saw themselves: breaking ground for a new community where data is shared and cures for cancer are found.

The Creative Commons was referenced earlier as an organization "that works to increase the amount of content (cultural, educational, and scientific) in 'the commons': the body of work that is available to the public for free and legal sharing, use, re-purposing, and remixing" (Wilbanks and Boyle, "Introduction to Creative Commons"). In this sense, the "commons" is an abstract impersonal collection of data, publications, and other sources of intellectual expression, distributed across the Internet, filed away in cabinets, and stored in libraries for others to retrieve. The *creative commons* is also, however, a compelling metaphor for the present and future community of cancer research. Instead of the abstracted commons of data sources and publications, picture the more traditional physical town commons found in the center of many cities, where people bring both their goods and their thoughts in order to engage with other people. This conceptually replaces a content-driven yet abstracted "body *of* work," with people-driven "bodies *at* work," coming together as a shared community for both personal gain and common purpose. In this creative commons, both the subject and the object are included in the sharing that is done, allowing the curiosity and courage of individuals to combine to create new knowledge - one person and one data point at a time.

Appendices

Appendix A: Acronyms

ANT	Actor Network Theory
caBIG	Cancer Biomedical Informatics Grid
CBIIT	Center for Bioinformatics and Information Technology
DSIC	Data Sharing and Intellectual Capital
DSSF	Data Sharing and Security Framework
FDA	Food and Drug Administration
FTP	File Transfer Protocol
GWAS	Genome Wide Association Studies
HIPAA	Health Insurance Portability and Accountability Act
IRB	Institutional Review Board
IT	Information Technology
NIH	National Institutes of Health
NCI	National Cancer Institute
NCICB	NCI Center for Bioinformatics
OBBR	NCI Office of Biorepositories and Biospecimen Research
PLoS	Public Library of Science
RPG	Research Project Grants
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas

Appendix B: Bibliography

Ambrose, M. L., and C. T. Kulik. "Old Friends, New Faces: Motivation Research in the 1990s." *Journal of Management* 25.3 (1999): 231-92. Print.

Apter, Michael J. *Danger: Our Love of Living on the Edge*. Oxford, England: Oneworld Publications, 2007. Print.

---. *Motivational Styles in Everyday Life: A Guide to Reversal Theory*. 1st ed. Washington, DC: American Psychological Association, 2001. Print.

---. "Reversal Theory, Victor Turner and the Experience of Ritual." *Journal of Consciousness Studies* 15.10-11 (2008): 184-203. Print.

Arzberger, Peter, Peter Schroeder, Anne Beaulieu, Geof Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir, and Paul Wouters. "An International Framework to Promote Access to Data." *Science* 303.5665 (2004): 1777-78. Print.

Averill, J.R. "Everyday Emotions: Let Me Count the Ways." *Social Science Information* 43(4) (2004): 571-80. Print.

Barbalet, J.M. *Emotion, Social Theory & Social Structure*. Cambridge, UK: Cambridge University Press, 2001. Print.

Barnes, B. *Interests and the Growth of Knowledge*. London, UK: Routledge and Kegan Paul, 1977. Print.

Barnes, B., and D. MacKenzie. "On the Role of Interests in Scientific Change." *On the Margins of Science: The Social Construction of Rejected Knowledge.* R. Wallis, ed. Staffordshire: University of Keele, 1979: 49-67. Print.

Baron, James N. "Data Sharing as a Public Good." *American Sociological Review* 53.1 (1988): vi-

viii. Print.

Beauchamp, Tom L. "Does Ethical Theory Have a Future in Bioethics?" *The Journal of Law, Medicine & Ethics* 32.2 (2002): 209-17. Print.

Beaulieu, A. "Research Woes and New Data Flows." *Promise and Practice in Data Sharing*. Paul Wouters and Peter Schröder, eds. Amsterdam: The Public Domain of Digital Research Data (NIWI-KNAW), 2003. Print.

Berscheid, E., H. Ammazzalorso. 2003. “Emotional experience in close relationships.” *Blackwell Handbook of Social Psychology: Interpersonal Process*. G. J. Fletcher, M. S. Clark, eds. Malden, MA: Blackwell, 2003: 308-330. Print.

Black, Max. *Models and Metaphors; Studies in Language and Philosophy*. Ithaca, N.Y.: Cornell University Press, 1962: 219-243. Print.

Borfitz, Deborah. "caBIG: Fostering Better Connections with Open Source Software". 2009. *Bio-IT World*. <http://www.bio-itworld.com/news/2009/01/05/caBIG-open-source-software.html>. Accessed: February 2009. Web.

Bourke, Joanna. *Fear: A Cultural History*. Emeryville, CA: Shoemaker Hoard: Distributed by Publishers Group West, 2006. Print.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Inside Technology Series. Cambridge, MA: MIT Press, 1999. Print.

Brown, John Seely and Paul Deguid. *The Social Life of Information.* Boston, MA: Harvard University Press, 2000. Print.

Brown, Theodore L. *Making Truth: Metaphor in Science*. Urbana, IL: University of Illinois Press, 2003. Print.

Buetow, Kenneth H. “Building a 21st Century Biomedical System: The Cancer Biomedical Informatics Grid (caBIG®).” Supercomputing 2008: November 19, 2008, Austin, TX, 2008. < https://cabig.nci.nih.gov/overview/Buetow_SC08-120408.pdf> Conference Presentation/Web.

---. "Cyberinfrastructure: Empowering a ‘Third Way’ in Biomedical Research." *Science* 308 (2005): 821-24. Print.

---. "Heading for the Big Time." *The Scientist* 22.4 (2008): 60-67. Print.

---. "The NCI Center for Bioinformatics (NCICB): Building a Foundation for in Silico Biomedical Research." *Cancer Investigation* 22.1 (2004): 117-22. Print.

---. "Uniting Efforts in Molecular Medicine." *Genetic Engineering News* (2007): January 15, 2007: 44-45. Print.

Butler, Declan. "Agencies Join Forces to Share Data." *Nature* 446.22 (2007): 354. Print.

---. "Data Sharing Threatens Privacy." *Nature* 449.11 (2007): 644-45. Print.

Cacioppo, John T, and Gary G. Berntson, Jeff T. Larson, Kirsten M. Poehlmann, and Tiffany A. Ito “Chapter 11: The Psychophysiology of Emotion.” *Handbook of Emotions*, *2nd ed.* Lewis, Michael and Jeannette M. Haviland-Jones, eds. New York: Guilford Press, 2000: 173-191. Print.

Campbell, Eric, and David Blumenthal. "The Selfish Gene: Data Sharing and Withholding in Academic Genetics". *Science Careers*. (May 31, 2002). <http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2002_05_31/noDOI.5822398718525511595> Accessed: February 2009. Web.

Campbell, Eric G., et al. "Data Withholding in Academic Medicine: Characteristics of Faculty

Denied Access to Research Results and Biomaterials." *Research Policy* 29.2 (2000): 303-12. Print.

Cancer Biomedical Informatics Grid (caBIG®). "Data Sharing and Decision Framework". 2009. Ed. DSIC Knowledge Center. <https://cabig-kc.nci.nih.gov/DSIC/KC/index.php/Data_Sharing_and_Security_Framework>. Accessed: February-March 2009. Web.

---. "Welcome to the caBIG® Community Website". Rockville, MD, 2009. Ed. MD National Cancer Institute; Rockville. <https://cabig.nci.nih.gov/>. Accessed Multiple Dates between February 2007 - May 2009. Web.

Carpenter, Jacque. "Metaphors in Qualitative Research: Shedding Light or Casting Shadows?" *Research in Nursing and Health* 31.3 (2008): 274-82. Print.

Carter, Stephen, and Jeremy Kourdi. *The Road to Audacity: Being Adventurous in Life and Work.* New York: Palgrave Macmillan, 2003. Print.

Caveman. "'Send Me All of Your Reagents and Ideas. We Want to Work on the Same Experiments.'" *Journal of Cell Science* 114 (2001): 1037-38. Print.

Ceci, Stephen J. "Scientists' Attitudes toward Data Sharing." *Science, Technology, & Human Values* 13.1/2 (1988): 45-52. Print.

Chang, Jenny C, Susan G. Hilsenbeck, and Suzanne Fuqua. "The Promise of Microarrays in the Management and Treatment of Breast Cancer." *Breast Cancer Research* 7.3 (2005): 100-104. Print.

Chin, George, and S. Lansing Carina. "Capturing and Supporting Contexts for Scientific Data Sharing Via the Biological Sciences Collaboratory." *Proceedings of the 2004 ACM*

conference on Computer supported cooperative work. Chicago, Illinois, USA: ACM, 2004. Print.

Chiu, C. M., M. H. Hsu, and E. T. G. Wang. "Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capital and Social Cognitive Theories." *Decision Support Systems* 42.3 (2006): 1872-88. Print.

Clarke, Adele E., et al. "Biomedicalization: Technoscientific Transformations of Health, Illness, and U.S. Biomedicine." *American Sociological Review* 68.2 (2003): 161-94. Print.

Clayton, Charles P. "NIH Limits Access to GWAS Databases Due to Privacy Concerns (September 5, 2008)". *Alliance for Academic Internal Medicine*. <http://www.im.org/PolicyAndAdvocacy/PolicyIssues/Research/NIH/Pages/NIHLimitsAccesstoGWASDatabasesDuetoPrivacyConcerns.aspx>. Accessed: January 3, 2009. Web.

Coar, Ken. "The Open Source Definition". 2007 Open Source Initiative. <http://www.opensource.org/docs/osd>. Accessed: March 2009. Web.

Conger, Jay A., and Rabindra N. Kanungo. "Toward a Behavioral Theory of Charismatic Leadership in Organizational Settings." *The Academy of Management Review* 12.4 (1987): 637-47. Print.

Covitz, Peter A., et al. "caCORE: A Common Infrastructure for Cancer Informatics." *Bioinformatics*. 19 (2003): 2404-12. Print.

Curt, Gregory. "Step Change in Safe Harbors: Public–Private Partnerships." *The Oncologist* 14 (2009): 1-3

Davis, Devra. *The Secret History of the War on Cancer*. New York, NY: Basic Books, 2007. Print.

Dawyndt, Peter, Tom Dedeurwaerdere, and Jean Swings. "Contributions of Bioinformatics and

Intellectual Property Rights in Sharing Biological Information." *International Social Science Journal (ISSJ)* 188 (2006): 249-58. Print.

Dove, Alan. "When Science Rides the MTA." *The Journal of Clinical Investigation* 110.4 (2002): 425-427. Print.

DuPont. "Testing Methods Using Oncomouse® Transgenic Models of Cancer". 2009. Ed. DuPont Technology Bank. <http://dupont.t2h.yet2.com/t2h/page/techpak?id=26128&sid=0&abc=0&page=development>. Accessed: March 2009. Web.

Editorial Staff. "caBIG Delivers Federated Access to Imaging Informatics Toolkit ." December 1, 2008. *Health Imaging*. <http://www.healthimaging.com/index.php?option=com_articles&view=article&id=15316>. Accessed: February 2009. Web.

---. "Compete, Collaborate, Compel." *Nature Genetics* 39 (2007): 931. Print.

---. "Editorial: 'Good Citizenship' or Good Business?" *Nature Genetics* 36.10 (2004): 1025. Print.

---. "Forward-Looking Systems at NCI". July 29, 2008. *ClinPage*. <http://www.clinpage.com/article/forward_looking_systems_at_nci/C10>. Accessed: January 3, 2009. Web.

---. "Let Data Speak to Data." *Nature* 438.7068 (2005): 531. Print.

---. "Report: Medicine's New Central Bankers." *Economist Technology Quarterly.* December 10, 2005 (2005): 18-19. Print.

---. "Researcher Access to Patient Samples Reaches Supreme Court." *Annals of Neurology* 62.3 (2007): A12-A14. Print.

Eiseman, Elisa, Gabrielle Bloom, Jennifer Brower, Noreen Clancy and Stuart S. Olmsted. *Case Studies of Existing Human Tissue Repositories: "Best Practices" for a Biospecimen Resource for the Genomic and Proteomic Era.* Santa Monica, CA: RAND, 2003. Print.

Fenstermacher, D., et al. "The Cancer Biomedical Informatics Grid (caBIG™)." Trans. IEEE Engn Med and Chinese Acad Engn Sci Biol Soc. *27th Annual International Conference of the IEEE-Engineering-in-Medicine-and-Biology-Society*. Shanghai, PEOPLES R CHINA: IEEE, 2005. 743-46. Conference Abstract.

Foucault, Michel. *Power/Knowledge: Selected Interviews and Other Writings, 1972-1977*. Colin Gordon, ed. New York, NY: Pantheon, 1980. Print.

Fujimura, Joan H. "The Molecular Biological Bandwagon in Cancer Research: Where Social Worlds Meet." *Social Problems* 35.3 (1988): 261-83. Print.

Gardner, Daniel, Arthur W.Toga, Giorgio A. Ascoli, Jackson T. Beatty, James F. Brinkley, Anders M. Dale, Peter T. Fox, Esther P. Gardner, John S. George, Nigel Goddard, Kristen M. Harris, Edward H. Herskovits, Michael L. Hines, Gwen A. Jacobs, Russell E. Jacobs, Edward G. Jones, David N. Kennedy, Daniel Y. Kimberg, John C. Mazziotta, Perry L. Miller, Susumu Mori, David C. Mountain, Allan L. Reiss, Glenn D. Rosen, David A. Rottenberg, Gordon M. Shepherd, Neil R. Smalheiser, Kenneth P. Smith,Tom Strachan, David C.Van Essen, Robert W.Williams, and Stephen T. C.Wong. "Towards Effective and Rewarding Data Sharing." *Neuroinformatics* 1.3 (2003): 289-95. Print.

Gardner, William. "Compelled Disclosure of Scientific Research Data." *Information Society* 20.2 (2004): 141-46. Print.

Giles, Jim. "Open-Access Journal Will Publish First, Judge Later." *Nature.* 445. 4 January 2007: 9.

Print.

Ginsburg, Geoffrey S., Thomas W. Burke, and Phillip Febbo. "Centralized Biorepositories for Genetic and Genomic Research." *Journal of the American Medical Association (JAMA)* 299.11 (2008): 1359-61. Print.

Goodwin, Jeff, James M. Jasper, and Francesca Polletta. *Passionate Politics: Emotions and Social Movements*. Chicago, IL: University of Chicago Press, 2001. Print.

Hands, D. Wade. "Caveat emptor: Economics and contemporary philosophy of science" *Philosophy of Science.* 64.4 (1997): S107-S116. Print.

Haraway, Donna. "Chapter 15: Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspectives (1988)." *The Gender and Science Reader*. Eds. Muriel Lederman and Ingrid Bartsch. London and New York: Routledge, 2001. 169-88. Print.

---. "Chapter 8: A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century," *Simians, Cyborgs and Women: The Reinvention of Nature.* New York; Routledge, 1991: 149-181. Print.

Harding, Sandra. "Chapter 14: Feminist Standpoint Epistemology." *The Gender and Science Reader*. Muriel Lederman and Ingrid Bartsch, eds. London and New York: Routledge, 2001. 145-68. Print.

Harvey, David. *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change.* Malden, MA: Blackwell Publishers, 1990. Print.

Havenstein, Heather. "'World Wide Web of Cancer Research' Exploits Human Genome Map". March 26, 2008. *Computerworld*. <http://www.computerworld.com/action/article.do?command=viewArticleBasic&taxon

omyName=Grid+and+Utility+Computing&articleId=9072101&taxonomyId=65&pageNumber=1>. Accessed: January 3, 2009. Web.

Hayden, Erika Check. "Personalized Cancer Therapy Gets Closer." Nature 458 (2009): 131-32. Print.

Health and Human Services. "The Biomedical Informatics Grid: A Platform for 21st Century Biomedicine". November 2008. *Personalized Health Care Community Case Studies*. <http://www.hhs.gov/myhealthcare/news/community.html#cmuny2>. Accessed: February 2009. Web.

Hess, David J. *Science Studies: An Advanced Introduction.* New York: New York University Press, 1997. Print.

Hilgartner, Stephen. "Biomolecular Databases: New Communication Regimes for Biology?" *Science Communication* 17.2 (1995): 240-63. Print.

Hine, Christine. "Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work." *Social Studies Of Science* 36.2 (2006): 269-98. Print.

Holland, Janet. "Emotions and Research." *International Journal of Social Research Methodology* 10.3 (2006): 195-209. Print.

Hughes, J. "Bringing Emotion to Work: Emotional Intelligence, Employee Resistance and the Reinvention of Character " *Work, Employment & Society* 19(3) (2005): 603-25. Print.

Ihde, Don. *Bodies in Technology*. Minneapolis: University of Minnesota Press, 2002. Print.

Jewels, Tony, and Marilyn Ford. "Factors Influencing Knowledge Sharing in Information Technology Projects." *e-Service Journal* (2006): 99-117. Print.

Joas, Hans (translated by Jeremy Gaines and Paul Keast). *The Creativity of Action*. Chicago:

University of Chicago Press, 1996. Print.

Jones, Nancy L. "A Code of Ethics for the Life Sciences." *Science and Engineering Ethics* 13.1 (2007): 25-43. Print.

Kaiser, Jocelyn. "Editorial: Making Data Dreams Come True." *Nature* 428.18 (2004): 239. Print.

---. "Von Eschenbach Revises the NCI Agenda." *Science* 303 (2004): 1952. Print.

Kay, Lily E. *Who Wrote the Book of Life? : A History of the Genetic Code*. Stanford, Calif.: Stanford University Press, 2000. Print.

Kaye, Jane, et al. "Data Sharing in Genomics - Re-Shaping Scientific Practice." *Genetics* Volume 10 (May 2009): 331-35. Print.

Kendall, Julie E., and Kenneth E. Kendall. "Metaphors and Methodologies: Living Beyond the Systems Machine." *MIS Quarterly* 17.2 (1993): 149-71.

Kennedy, Donald. "Not Wicked, Perhaps, but Tacky." *Science* 297.5585 (2002): 1237. Print.

Kohler, Robert E. *Lords of the fly: Drosophila genetics and the experimental life.* Chicago: University of Chicago Press, 1994. Print.

Kligyte, Vykinta, et al. "Application of a Sensemaking Approach to Ethics Training in the Physical Sciences and Engineering." *Science and Engineering Ethics* 14.2 (2008): 251-78. Print.

Kövecses, Zoltán. *Metaphor and Emotion: Language, Culture, and Body in Human Feeling.* Studies in Emotion and Social Interaction. Second Series. London: Cambridge, 2003. Print.

Kövecses, Zoltán. *Metaphor: A Practical Introduction.* New York, NY: Oxford University Press, 2002. Print.

Lakoff, George, and Mark Johnson. *Metaphors We Live By*. Chicago: University of Chicago Press,

2003. Print.

Lakoff, George. “Metaphor and War: The Metaphor System Used to Justify War in the Gulf” Unpublished Paper, 1991. Print.

Langella, Stephen, et al. "Sharing Data and Analytical Resources Securely in a Biomedical Research Grid Environment." *Journal of the American Medical Informatics Association* 15.3 (2008): 363-73. Print.

Latham, G. P., and C. C. Pinder. "Work Motivation Theory and Research at the Dawn of the Twenty-First Century." *Annual Review of Psychology* 56 (2005): 485-516. Print.

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

---. *We Have Never Been Modern*. Cambridge, MA.: Harvard University Press, 1993. Print.

Lawler, Edward J., and Shane R. Thye. "Bringing Emotions into Social Exchange Theory." *Annual Review of Sociology* 25 (1999): 217-44. Print.

Lewis, Kristi M. "When Leaders Display Emotion: How Followers Respond to Negative Emotional Expression of Male and Female Leaders." *Journal of Organizational Behavior* 21.2 (2000): 221-34. Print.

Lopez, José Julian. "Notes on Metaphors, Notes as Metaphors: The Genome as Musical Spectacle." *Science Communication* 29.1 (2007): 7-34. Print.

Lutz, C. and G.M. White. "The Anthropology of Emotions." *Annual Review of Anthropology* Vol. 15 (1986): 405-36. Print.

Lyon, Margot L. "Missing Emotion: The Limitations of Cultural Constructionism in the Study of Emotion." *Cultural Anthropology* 10.2 (1995): 244-63. Print.

MacCallum, Catriona J. "One for All: The Next Step for Plos." *PLoS Biology* 4.11 (2006): e401. Print.

Maitlis, S , and H Ozcelik. "Toxic Decision Processes: A Study of Emotion and Organizational Decision Making." *Organization Science* 15(4) (2004): 375-93. Print.

Marshall, Eliot. "DNA Sequencer Protests Being Scooped with His Own Data." *Science* 295.5558 (2002): 1206-07. Print.

---. "Sharing the Glory, Not the Credit." *Science* 291.5507 (2001): 1189-93. Print.

---. "The Upside of Good Behavior: Make Your Data Freely Available." *Science* 299.5609 (2003): 990. Print.

Martin, Emily, "The egg and the sperm: How science has constructed a romance based on stereotypical male-female roles," *Signs* 16:3 (1991): 485-501. Print.

Mashberg, T. "Cancer's "World Wide Web:" A Lung Image Database Is Breathing Life Into "Medical Grid" Vision." *Technology Review* March-April 2006: 1-3. Print.

Mathew, Jomol P., et al. "From Bytes to Bedside: Data Integration and Computational Biology for Translational Cancer Research." *PLoS Comput Biol* 3.2 (2007): e12. Print.

McCain, Katherine W. "Mandating Sharing: Journal Policies in the Natural Sciences." *Science Communication* 16 (1995): 403-31. Print.

McClellan, Chris. “The Economic Consequences of Bruno Latour.” *Social Epistemology*. 10:2 (1996): 193-208. Print.

Melton, Gary B. "Must Researchers Share Their Data?" *Law and Human Behavior* 12.2 (1988): 159-62. Print.

Merton, Robert King. "Chapter 13: The Normative Structure of Science" (1942) In: Robert King

Merton, *The Sociology of Science : Theoretical and Empirical Investigations*. Chicago, IL: University of Chicago Press, 1973; 267-278. Print.

Meyer, Eric T. "Moving from Small Science to Big Science: Social and Organizational Impediments to Large Scale Data Sharing." Unpublished paper. 2007. <http://ess.si.umich.edu/papers/paper218.pdf> Web.

Mirowski, Philip. "On playing the economics trump card in the philosophy of science" *Philosophy of Science*. 64:4 (1997): S127-S138. Print.

Mitroff, Ian. "Norms and Counter-Norms in a Select Group of Apollo Moon Scientists: A Case Study in the Ambivalence of Scientists." *American Sociological Review* 39 (1974): 579-95. Print.

Moldoveanu, Mihnea C., and Nitin Nohria. *Master Passions: Emotion, Narrative, and the Development of Culture*. Cambridge, MA: MIT Press, 2002. Print.

Morgan, Gareth. *Images of Organization*. Executive ed. San Francisco, CA; Thousand Oaks, CA: Berrett-Koehler Publishers; Sage Publications, 1998. Print.

Mulkay, Michael J. "Norms and Ideology in Science." *Social Science Information* 15 (1976): 637-56. Print.

Nagl, Sylvia. *Cancer Bioinformatics: From Therapy Design to Treatment*. Wiley Life Sciences, 2006. Print.

National Cancer Institute. "The Nation's Investment in Cancer Research: Connecting the Cancer Community: An Annual Plan and Budget Proposal for Fiscal Year 2009." January 2008. <http://plan2009.cancer.gov/pdf/nci_2009_plan.pdf>. Accessed: January 2009. Web.

---. "The NCI Strategic Plan". January 2006. <http://strategicplan.nci.nih.gov/>. Accessed:

January 2009. Web.

---. "Special Report: caBIG: The Launch of a Bioinformatics Community." *NCI Cancer Bulletin* 1.9 (2004): 5-7.

National Cancer Institute Center for Bioinformatics (NCICB). "About NCICB" <http://ncicb.nci.nih.gov/about> Accessed: December 2008. Web.

National Cancer Institute Office of Biospecimens and Biorepositories Research (OBBR). "Patient Corner: What Are Biospecimens and Biorepositories?" March 18, 2009. <http://biospecimens.cancer.gov/patientcorner/>. Accessed: April 2009. Web.

National Institutes of Health. "Fact Sheet: Genome-Wide Association Studies". January 27, 2009. Ed. National Human Genome Research Institute. <http://www.genome.gov/20019523 >. Accessed: February 8, 2009. Web.

---. "Final NIH Statement on Sharing Research Data." February 26, 2003. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>. Accessed: November 2009. Web.

---. "Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS) - Notice Number: Not-Od-07-088". August 28, 2007. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html > Accessed: November 2008. Web.

National Research Council. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council, 2003. Print.

Niu, Jinfang. "Incentive Study for Research Data Sharing: A Case Study on NIJ Grantees."

Undated unpublished paper. 1-40. Print.

Nosowsky, Rachel, and Thomas J. Giordano. "The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule: Implications for Clinical Research." *Annual Review of Medicine* 57 (2006): 575–90. Print.

Ochsner, Scott A., et al. "Much Room for Improvement in Deposition Rates of Expression Microarray Datasets." *Nature Methods* 5.12 (2008): 991. Print.

Park, Alice. "10 Ideas Changing the World Right Now - #8: Biobanks" 2009. *TIME Magazine*. (6 April 2009). <http://www.time.com/time/specials/packages/0,28757,1884779,00.html>. Web.

Parker, B. "The Advocate Role in Clinical Study Development and Partnering with Patient Advocates in Your Local Institution." *Cancer Treat Res*. 132 (2007): 131-41. Print.

Peters, Kim, and Yoshihisa Kashima. "From Social Talk to Social Action: Shaping the Social Triad with Emotion Sharing." *Journal of Personality and Social Psychology* 93.5 (2007): 780–97. Print.

Pickering, Andrew. "Chapter 25: The Mangle of Practice: Agency and Emergence in the Sociology of Science (1993)." *The Science Studies Reader*. Mario Biagioli, ed. London and New York: Routledge, 1999: 372-93. Print.

Pilcher, K. S. "Coordination of Cancer Research." *Science, New Series* 104.2694 (1946): 167-168. Print.

Piwowar, Heather A., et al. "Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers " *PLoS Med* 55.9 (2008): e183. Print.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. "Sharing Detailed Research Data Is

Associated with Increased Citation Rate." *PLoS ONE* 2.3 (2007): e308. Print.

Prior, Lindsay. "Talking About the Gene for Cancer: A Study of Lay and Professional Knowledge of Cancer Genetics." *Sociology* 41.6 (2007): 985-1001. Print.

Raban, Daphne R., and Sheizaf Rafaeli. "Investigating Ownership and the Willingness to Share Information Online." *Computers in Human Behavior* 23 (2007): 2367–82. Print.

Rafaeli, A, and I Vilnai-Yavetz. "Emotion as a Connection of Physical Artifacts and Organizations." *Organization Science* 15(6).Nov-Dec. 2004: 671-86. Print.

Rai, Arti K. "Open and Collaborative Research: A New Model for Biomedicine." *Intellectual Property Rights in Frontier Industries: Biotech and Software.* AEI-Brookings Press (2005). Print.

Reimann, Stanley P. "Highways and Byways of Cancer Research." *Cancer Research* 13 (1953): 493-98. Print.

Ross, D, and P Dumouchel. "Emotions as Strategic Signals." *Rationality and Society* 16.3 (2004): 251-86. Print.

Rubin, Daniel L., Nigam H. Shah, and Natalya F. Noy. "Biomedical Ontologies: A Functional Perspective." *Briefings in Bioinformatics* 9.1 (2008): 75-90. Print.

Saltz, Joel. "caBIG: Envisioning the Future." *2007 caBIG Annual Meeting*. Washington DC, February 6, 2007. <https://cabig.nci.nih.gov/2007caBIGconference/presentations/tuesday-february-06-2007/session-block-3/breakout-sessions/cabigTM-envisioning-the-future> Presentation/Web.

Saltz, J., et al. "caGrid: Design and Implementation of the Core Architecture of the Cancer

Biomedical Informatics Grid." *Bioinformatics* 22.15 (2006): 1910-16. Print.

Schön, Donald. "Generative metaphor: A perspective on problem-setting in social policy," In Andrew Ortony, ed., *Metaphor and thought*, 2d ed. Cambridge and New York: Cambridge University Press (1993): 137-163. Print.

Science Commons. "About Science Commons." <http://sciencecommons.org/about/> Accessed: April 2009. Web.

Shamir, Boas, Robert J. House, and Michael B. Arthur. "The Motivational Effects of Charismatic Leadership: A Self-Concept Based Theory." *Organization Science* 4.4 (1993): 577-94. Print.

Shields, Stephanie A. "The Politics of Emotion in Everyday Life: 'Appropriate' Emotion and Claims on Identity." *Review of General Psychology* 9.1 (2005): 3-15. Print.

Shrum, Wesley, Ivan Chompalov, and Joel Genuth. "Trust, Conflict and Performance in Scientific Collaborations." *Social Studies of Science* 31.5 (2001): 681-730. Print.

Sieber, Joan E. "Data Sharing: Defining Problems and Seeking Solutions." *Law and Human Behavior* 12.2 (1988): 199-206. Print.

Siemsen, E., A. V. Roth, and S. Balasubramanian. "How Motivation, Opportunity, and Ability Drive Knowledge Sharing: The Constraining-Factor Model." *Journal of Operations Management* 26.3 (2008): 426-45. Print.

Sokol, R. I., and S. L. Strout. "Metaphor and Emotion: Language, Culture and Body in Human Feeling." *Culture & Psychology* 12.1 (2006): 115-23. Print.

Srinivas, Krishna Ravi. "Intellectual Property Rights and Bio Commons: Open Source and Beyond." *International Social Science Journal (ISSJ)* 188 (2006): 319-34. Print.

Stanley, Barbara, and Michael Stanley. "Data Sharing: The Primary Researcher's Perspective." *Law and Human Behavior* 12.2 (1988): 173-80. Print.

Sterling, Theodor D. "Analysis and Reanalysis of Shared Scientific Data." *Annals of the AAPPSS* 495 (1988): 49-60. Print.

Sterling, Theodor D., and James J. Weinkam. "Sharing Scientific Data." *Communications of the ACM* 33.8 (1990): 112-19. Print.

Strauss, Anselm. "A Social World Perspective." *Studies in Symbolic Interaction* 1 (1978): 119-28. Print.

Tapscott, Don, and Anthony D. Williams. "The New Science of Sharing." *Business Week*. March 2, 2007 <http://www.businessweek.com/innovate/content/mar2007/id20070302_219704.htm> Web.

---. *Wikinomics : How Mass Collaboration Changes Everything*. Expanded ed. New York, NY: Portfolio, 2008. Print.

Thacker, Eugene. *The Global Genome: Biotechnology, Politics, and Culture*. Cambridge, MA: MIT Press, 2005. Print.

Theologis, Athanasios, and Ronald W. Davis. "To Give or Not to Give? That Is the Question." *Plant Physiol* 135 (2004): 4-9. Print.

Tucker, Jennifer and David Hile Rutledge. *"Motivation and Emotion in Technology Teams."* *CrossTalk – Department of Defense Journal of Software Engineering*. November 2007: 10-13. Print.

Turner, Jonathan H., and Jan E. Stets. *The Sociology of Emotions*. Cambridge [UK]; New York:

Cambridge University Press, 2005. Print.

Vaught, Jimmie B., et al. "Ethical, Legal, and Policy Issues: Dominating the Biospecimen Discussion." *Cancer Epidemiol Biomarkers*. 16 (2007): 2521-23. Print.

Venter, J Craig. "Prepared Statement of J. Craig Venter, PhD - President and Chief Scientific Officer, Celera Genomics - April 6, 2000." *Testimony to Subcommittee on Energy and Environment*. Washington DC, 2000 of *US House of Representatives Committee on Science.* Testimony Statement.

Ventura, Beverly. "Mandatory Submission of Microarray Data to Public Repositories: How Is It Working?" *Physiological Genomics* 20.2 (2005): 153-56. Print.

Vickers, Andrew. "Essay: Cancer Data? Sorry, Can't Have It." *New York Times.* January 22, 2008. Print.

vonEschenbach, Andrew C., and Kenneth Buetow. "Cancer Informatics Vision: caBIG." *Cancer Informatics* 2 (2006): 22-24. Print.

Waldrop, Mitch. "Wikiomics." *Nature* 455.4 (2008): 22-25. Print.

Wasko, Molly McLure, and Samer Faraj. "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice." *MIS Quarterly* 29.1 (2005): 35-57. Print.

Weick, Karl E., Kathleen M. Sutcliffe, and David Obstfeld. "Organizing and the Process of Sensemaking." *Organization Science* 16.4 (2005): 409-21. Print.

Wilbanks, John, and James Boyle. "Introduction to Science Commons." August 3, 2006. <http://sciencecommons.org/wp-content/uploads/ScienceCommons_Concept_Paper.pdf> Accessed: March 2009. Web.

Williams, Simon and Gillian Bendelow. *The Lived Body: Sociological Themes, Embodied Issues*. London: Routledge, 1998. Print.

Winner, Langdon. "Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology." *Science, Technology, & Human Values* 18.3 (1993): 362-78. Print.

Zaltman, Gerald, and Lindsay H. Zaltman. *Marketing Metaphoria: What Deep Metaphors Reveal About the Minds of Consumers*. Boston, MA: Harvard Business School Press, 2008. Print.

Ziman, John. "New Modes of Knowledge Production." *Real Science: What It Is and What It Means*. Port Chester, NY: Cambridge University Press, 2000. 56-82. Print.

Appendix C: Annotated List of Figures with Copyright Use Determinations

Figure 1: Illustration of the Data Generated from a Single Mouse [Fair Use Determination]

This figure is from a government conference presentation that is also posted on the Internet, Reference: Saltz, Joel. "caBIG: Envisioning the Future." *2007 caBIG Annual Meeting*. Washington DC, February 6, 2007. <https://cabig.nci.nih.gov/2007caBIGconference/presentations/tuesday-february-06-2007/session-block-3/breakout-sessions/cabigTM-envisioning-the-future>

Fair Use Factors:

- Purpose: Research, scholarship
- Nature of Copyright: Fact-Based, Important to educational objectives
- Amount of Material: Only one slide of material used from a full conference presentation; amount is appropriate to the research purpose
- Effect: Slide is already posted on government website for open and free access; no known market impact

Figure 2: caBIG at Work [Public Domain – Government Work]

This graphic was created by the caBIG government program in 2006 and has not been formally published. It is used here as a U.S government work.

Figure 3: Defining Interoperability [Public Domain – Government Work]

This graphic is used in a variety of caBIG government presentations. One of these is "caBIG Essentials Training Program, Lesson 1" accessible at <https://cabig.nci.nih.gov/concepts/essentials> (November 2008). It is used here as a U.S government work.

Figure 4: The Bench to Bedside Cycle of Cancer Research [Public Domain – Government Work]

This graphic is used in a variety of caBIG presentations. One of these was in the "caBIG 2009 Annual Meeting Newcomer's Session" accessible at <https://cabig.nci.nih.gov/2009AnnualMeeting/presentations/monday-july-20-2009/welcome-and-newcomers-session/newcomers/file_download/presentation> (July 20, 2009). It is used here as a U.S government work.

Figure 5: From Capability to Artifacts - Data Generated from Bench to Bedside [Public Domain – Government Work]

This graphic is used in a variety of caBIG presentations. A recent reference: Buetow, Kenneth H. "Building a 21st Century Biomedical System: The Cancer Biomedical Informatics Grid (caBIG®)."

Supercomputing 2008: November 19, 2008, Austin, TX, 2008.
< https://cabig.nci.nih.gov/overview/Buetow_SC08-120408.pdf>The figure is used here as a U.S government work.

Figure 6: Choosing to Share Data – What and at What Point of Discovery? [Original Creation]

Original graphic created by the dissertation author.

Figure 7: Challenging the 17th Century Paradigm [Public Domain – Government Work]

This graphic is used in a variety of caBIG presentations. Reference: Buetow, Kenneth H. "Building a 21st Century Biomedical System: The Cancer Biomedical Informatics Grid (caBIG®)." Supercomputing 2008: November 19, 2008, Austin, TX, 2008. < https://cabig.nci.nih.gov/overview/Buetow_SC08-120408.pdf>The role of metaphor in communicating about data sharing is the focus of Chapter 6. The figure is used here as a U.S government work.

Figure 8: Sharing to Support Patient Care [Fair Use Determination]

These two slides were used in Patient Advocate presentations at the NCI OBBR Biospecimen Management Forums held at NIH in Bethesda, Maryland and in Boston, Massachusetts in Summer 2007. Reference: Kim, Paula, "The Importance of Best Practices to Patients, Advocates & The Public" Biospecimens Best Practice Forums. Summer 2007. It available for free download at <http://biospecimens.cancer.gov/practices/forum/boston2007/pdf/Paula_Kim-The_Importance_of_Best_Practices_to_Patients_Advocates_and_The_Public.pdf> It is used here under a Fair Use Determination.

Fair Use Factors:

- Purpose: Research, scholarship, criticism
- Nature of Copyright: Fact-Based, Important to educational objectives
- Amount of Material: Only two slides of material used from a full conference presentation; amount is appropriate to the research purpose
- Effect: Slide and presentation was created for an open public government forum and were made freely available to participants; no known market impact. Presentation available online for free download on a government website.

Figure 9: Collaboration, Data Sharing, and Creative Intellect [Original Creation]

Original graphic created by the dissertation author.

Figure 10: Images of Biobanking [Fair Use Determination]

This figure is from the following reference: "Report: Medicine's New Central Bankers." *Economists Technology Quarterly* December 10, 2005 (2005): 18. It is used here under a Fair Use determination.

Fair Use Factors:

- Purpose: Research, scholarship, criticism
- Nature of Copyright: Fact-Based, Important to educational objectives
- Amount of Material: Only one graphic from a journal article; amount is appropriate to the research purpose
- Effect: No significant impact to the market or potential market for this research journal article is anticipated

Figure 11: Combining Data Sharing "Ingredients" Leads to Cures [Fair Use Determination]

This figure is from the following reference: Dove, Alan. "When Science Rides the MTA." *The Journal of Clinical Investigation* 110.4 (2002): 425. It is used here under a Fair Use determination.

Fair Use Factors:

- Purpose: Research, scholarship, criticism
- Nature of Copyright: Fact-Based, Important to educational objectives
- Amount of Material: Only one graphic from a journal article; amount is appropriate to the research purpose
- Effect: No significant impact to the market or potential market for this research journal article is anticipated

Figure 12: Images of Data Sharing [Original Creation]

Original graphic created by the dissertation author.

Figure 13: The Domains and States of Reversal Theory [Original Creation]

Original graphic created by the dissertation author.

Appendix D: Institution Review Board Research Approval Documentation

VirginiaTech

Office of Research Compliance
Institutional Review Board
2000 Kraft Drive, Suite 2000 (0497)
Blacksburg, Virginia 24061
540/231-4991 Fax 540/231-0959
e-mail moored@vt.edu
www.irb.vt.edu

FWA00000572(expires 1/20/2010)
IRB # is IRB00000667

DATE: September 24, 2008

MEMORANDUM

TO: Barbara L. Allen
Jennifer Tucker

Approval date: 9/24/2008
Continuing Review Due Date:9/9/2009
Expiration Date: 9/23/2009

FROM: David M. Moore

SUBJECT: **IRB Expedited Approval:** "Motivation and Emotion in Collaborative Science" , IRB # 08-554

This memo is regarding the above-mentioned protocol. The proposed research is eligible for expedited review according to the specifications authorized by 45 CFR 46.110 and 21 CFR 56.110. As Chair of the Virginia Tech Institutional Review Board, I have granted approval to the study for a period of 12 months, effective September 24, 2008.

As an investigator of human subjects, your responsibilities include the following:

1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects.
2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.
3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date.
4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects.

Important:
If you are conducting **federally funded non-exempt research**, please send the applicable OSP/grant proposal to the IRB office, once available. OSP funds may not be released until the IRB has compared and found consistent the proposal and related IRB applicaton.

cc: File