When Big Data Meets Life Sciences: Data Reporting Standards and Innovation [*]

Min Ren[†]
Kellogg School of Management, Northwestern University

January 27, 2014

Abstract

Cumulative innovation is a driving force of economic growth. Access costs, the time and effort scientists need to devote to understand existing knowledge, may potentially hinder new innovation. I examine the effect of a decrease in access costs resulting from the adoption of a data reporting standard—Minimum Information About a Microarray Experiment (MIAME)—on subsequent life sciences research. I take advantage of a natural experiment, in which different academic journals adopted the MIAME standard at different times, to implement a difference-in-differences estimate of MIAME on subsequent use of data in journal publications. The results show that microarray data submitted after a journal adopts MIAME is at least 50 percent more likely to be reused. Overall, the evidence suggests that the decline in access costs due to data reporting standards is important for the accumulation of knowledge in the life sciences.

[*] I am very grateful to Meghan Busse, Shane Greenstein, and Benjamin Jones for their advice. I also thank David Dranove, Alberto Galasso, Danielle Li, Mike Powell, Scott Stern, and numerous seminar participants for helpful comments. I thank Heather Piwowar for sharing the reuse data. Several individuals from the Functional Genomics Data Society and related institutions, including Warren Kibbe, Roger Kroes, Neil McKenna, John Quackenbush, and staff at the Gene Expression Omnibus, offered valuable guidance. Financial support from the Sloan Research Fund on the Economics of Knowledge Contribution and Distribution is gratefully acknowledged. All errors are my own.

[†] m-ren@kellogg.northwestern.edu

1 Introduction

Cumulative innovation is often thought as a driving force of economic growth (Romer 1990; Jones 1995). In order to build upon previous knowledge, however, scientists need to devote time and effort to understand the details and contributions. If a scientist performs a novel experiment but does not document his protocol, successive cohorts would have a difficult time replicating the experiment and innovating based on the idea. I define the costs involved in acquiring such prior knowledge as access costs.[1] If access costs present such a burden to innovation, it is important to understand how to reduce their negative effects. The economics literature has focused on the incentive to create new knowledge, but understanding how to make the knowledge stock easily accessible may also be critical to innovation.

A particular form of access costs comes along with the challenge of big data. As we have entered the digital age, the quantity of data has grown dramatically. The increasing use of digitized information has also affected scientific research. This is especially true for the life sciences. Since the Human Genome Project, which constructed a catalog of all three billion chemical letters in the human genome, the process of decoding genetic data has become a crucial part of life sciences research. Complex genetic data, however, are difficult to analyze when they are poorly documented. Therefore understanding the data produced by other researchers has created access costs. Access costs associated with big data add to the "burden of knowledge" as discussed in Jones (2009), potentially hindering long-term economic growth. From an innovation policy standpoint, understanding how to reduce the costs of access to data is important. In this paper, I study how data reporting standards serve as a way to reduce access costs and facilitate subsequent innovation.

In this paper, I analyze the standardization of microarray data reports. Microarrays were introduced in the 1990s as a key technology for the study of gene expression, i.e., the information translation from gene to protein. Understanding gene expression helps decode the genetic foundation of many diseases including cancer. Yet microarray data is meaningful only with clear data reports. For instance, when the list of genes or the physical trait of the sample is missing, the data are of little value.[2] Therefore, standardization of data reports may promise to reduce access costs, enabling researchers to discover new knowledge from the data.

So how do data reporting standards influence subsequent innovation? Does the effect of

[1] Mokyr (2005) also discusses access costs of this type: ". . . what counted for useful knowledge to play a role in generating economic growth was therefore access costs, the marginal costs involved in acquiring knowledge possessed by someone else in society."

[2] For example, Culhane et al. (2010) mention the problem of gene loss when they construct a gene expression signature database due to the nonstandard annotation of gene lists.

data reporting standards vary across different data sets? To investigate these questions, I use a natural experiment—the introduction of the Minimum Information About a Microarray Experiment (MIAME) standard—and exploit a new measure of data-based innovation, data reuse.

MIAME is a data reporting standard developed by the Functional Genomics Data (FGED) Society in 2001. It specifies the minimum information required to describe a microarray experiment. The end goal of MIAME is to ensure that every researcher can interpret the experimental results in an unambiguous way. Starting at the end of 2002, some journals such as *Nucleic Acids Research* endorsed MIAME and required authors publishing with them to comply with the MIAME protocol. Over time, more journals adopted MIAME and the variation in the timing of adoption is mostly due to journal editors' preferences. This is the source of variation that I exploit for identification. Specifically, I compare the difference in data reuse before and after journals' adoption of MIAME with comparable journals that do not adopt MIAME, to estimate the impact of data reporting standards on subsequent scientific research.

To measure the impact of data reporting standards on subsequent research, I use a novel measure, data reuse, which documents whether a data set has been reused in new research articles. Data reuse is measured by searching the full text of articles for mention of data identifiers in a public repository, Gene Expression Omnibus (GEO).[3] Scientists can reuse existing data for multiple purposes: to create new algorithms and statistical tools, conduct meta-analysis, and verify new findings. For example, Professor Atul Butte and his colleagues at Stanford University scrutinized the data on GEO with a computational method, and they discovered unexpected new uses for existing drugs. One of their findings is that a widely used, cheap, over-the-counter anti-ulcer drug may treat a form of lung cancer (Sirota et al. 2011). With the digitization of medical research since the Human Genome Project, data reuse can be very valuable, and data reporting standards can assist scientists' reuse of data.

My paper has two major findings. First, I show that data reporting standards increase data reuse. Data submitted after a journal's adoption of MIAME is at least 50% more likely to be reused. This result is not driven by the selection on article quality, as article quality, measured by the count of citations, does not change significantly. I also conduct a placebo test and find that MIAME has no impact on self-reuse of data, confirming that the result is not caused by a change in data value. Second, the impact of data reporting standards is stronger for prestigious journals and authors from top medical schools. This suggests that reputation and data reporting standards are complementary.

The paper builds on the strand of recent papers that documents the relationship between

[3] This is following the method in Piwowar et al. (2013).

access costs and cumulative knowledge production. First, Furman and Stern (2011) emphasize the impact of institutions on knowledge accumulation. They examine a biological resource center that reduces access costs by certifying, preserving, and providing access to standardized biological materials. They find an amplification in cumulative knowledge production. However, they do not separate the effect of standardization from the multiple functions of the institution. Second, Murray et al. (2009), Galasso and Schankerman (2013) and Williams (2013) discuss the effect of intellectual property rights on subsequent innovation, and they stress the crucial role of openness in knowledge accumulation. Data reporting standards could play a similar role in making data more open to third-party investigators. Finally, Agrawal and Goldfarb (2008) study the effect of a decrease in collaboration costs resulting from the adoption of Bitnet on university research collaboration. Similar to Bitnet, data reporting standards may be another way to reduce communication costs between researchers.

Understanding how data access affects innovation is crucial to government funding of scientific research. Government funding is often suggested as central to medical innovation, and the National Institutes of Health (NIH) plays a primary role in this endeavor.[4] One key question for funding agencies is to choose which projects to fund, and my empirical work offers a new perspective on this, by suggesting that there could be great returns from funding projects to standardize data reports and annotations, not just from funding new research projects. Such standardization can make data more legible, so that a large community of scientists can discover new knowledge from the data that have already been generated. When many scientists have easy access to the pool of existing experimental data, they may use it with fresh approaches which may yield answers to questions that the original data contributors have never thought of.

Moreover, data access is crucial in broad settings. As technology advances, the volume of data keeps increasing in various fields. Besides biomedical studies, other sources of health-related information, including electronic health records and hospital admissions, have also been digitized. These are useful data sources for further exploration: the McKinsey Global Institute (2011) reports that the large data sets produced by the US health care system could potentially create substantial value. Therefore, with the opportunity from big data, if society can make sure the data from various sources are reported in a standard and accessible way, scientific progress may become faster and cheaper.

The paper proceeds as follows. Section 2 provides a brief scientific background. Section 3 describes the identification and data. I discuss the empirical results and robustness tests in Section 4 and Section 5, respectively. Section 6 presents concluding remarks.

[4] For example, see Azoulay, Graff-Zivin, Li, and Sampat (2013) for a discussion on the impact of NIH funding.

2 Microarray: Background

2.1 What Is a Microarray?

A microarray has mostly been used to study gene expression since the late 1990s. Gene expression describes the transcription of the information contained within the DNA into messenger RNA (mRNA) molecules that are then translated into protein: DNA is the repository of genetic information, mRNA is the messenger between DNA and protein, and protein performs most of the critical cell functions. Every cell contains almost identical genes with a few exceptions, but only a fraction of them are turned on. In other words, only the subset of genes that are "expressed" confers the genetic information to cells, and "gene expression" is the term to describe such information transmission process. Maintaining proper gene expression is important to human health, and disruptions or changes in gene expression can lead to many diseases including cancer. Therefore scientists study the kinds and amounts of mRNA produced by DNA to learn which genes are expressed, so as to understand the function of cells.

Traditionally, scientists survey a relatively small number of genes at a time to study gene expression. However, with microarray technology, scientists can quickly analyze expressions of many genes, and even the whole human genome that includes the complete genetic information, in a single experiment. With such improvement in efficiency, microarray technology has become widely used to explore the underlying genetic causes of many human diseases.

Figure 1 shows what a microarray looks like. A microarray consists of a small membrane or glass slide containing sample of many genes arranged in a regular pattern. It works by exploiting the ability of a given mRNA molecule to bind specifically to the DNA template from which it originated. Specifically, scientists can simultaneously determine the expression levels of thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array. With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

So how does a scientist extract information about a disease condition from a dime-sized chip containing thousands of individual gene sequences? Figure 2 illustrates the key steps of a microarray experiment. The key of the process is nucleic acid hybridization, which matches the complementary molecules of mRNA, so that the signal of mRNA can be identified.

For example, suppose there are two cells: cell type 1, a healthy cell, and cell type 2, a diseased cell.[5] Both contain an identical set of four genes, A, B, C, ad D, and scientists are interested in determining the expression profiles of these four genes in the two cell types. With microarray, they can identify the signal level of each gene and store the digital information

[5] This example is from NCBI A Science Primer (2007).

on signal strength in the computer. As a result, the computer may conclude the expression levels of the four genes: both cells express gene A at the same degree, cell 1 expresses more of gene B, cell 2 expresses more of gene C, and neither cell expresses gene D. But this is a simple example; in fact, a microarray experiment can investigate the whole human genome with more than 20,000 genes. So, the volume of data generated from a single experiment can mount quickly.

2.2 Microarray Data and MIAME

The data generated from microarray experiments can be huge. Consider data generated using a single microarray chip that contains expression information of up to 20,000 genes in the human genome. Besides expression information for the 20,000 genes, annotations such as which gene corresponds to which spot on the microarray chip and the physical traits of each target are necessary in order to interpret the results. Microarray experiments are potentially helpful to improve our understanding of gene expression, but scientists must tackle the new challenges to analyze the big data that has been generated.

DNA microarray analysis has already become one of the most widely used sources of large-scale genome data in the life sciences. In the last decade, an increasing number of scientific articles that use microarray technology have been published. For example, there were fewer than 2,000 articles based on this technology published in 1999. The number has grown rapidly to over 6,000 in 2005. Cumulatively, approximately 30,000 articles were published in the 6 years between 1999 and 2005 (Mogoutov et al. 2008). These microarray expression studies are producing massive quantities of gene expression and other functional genomics data, which promise to provide key insights into gene function and interaction within and across metabolic pathways (Brazma et al. 2001).[6]

Data generated by microarray experiments are complex and they are meaningful only in the context of a detailed description of the condition under which they are generated. The descriptions of the experimental and biological conditions, in addition to the data-processing protocols, are essential to understanding the data fully. So scientists need a data reporting standard to make microarray data useful. Moreover, since it is a new technology, there is not a universal platform to perform a microarray experiment, and different platforms are used to generate data. This makes the interpretation of data from different platforms more complicated. To sum up, microarray is a useful tool for genomics studies, but it has been a challenge for scientists to understand the data it generates.[7]

[6] Moreover, these microarray studies are not confined to basic research: an increasing number of publications describe clinical applications of microarrays (Keating and Cambrosio 2004).

[7] For example, a 2008 survey of international microarray facilities and individual microarray users, conducted

As a consequence, in 2001, the Functional Genomics Data (FGED) Society,[8] a group aiming to provide a standard for the representation of microarray expression data that would facilitate the exchange of microarray information between different data systems, published the Minimum Information About a Microarray Experiment (MIAME) standard, a guideline for the minimum information required to describe a DNA microarray-based experiment (Brazma et al. 2001). MIAME is independent of the particular experimental platform and provides a framework for describing experiments performed on all types of DNA arrays. Its guidelines specify the information required to describe such an experiment so that another investigator in the same discipline could either reproduce the experiment or analyze the data *de novo*. MIAME consists of six sections: experiment design, array design, samples, hybridization, measurements, and normalization controls. These contain the main elements to describe the experiment so that the data can be understood clearly and universally.

As a result, MIAME guidelines specify the information required to describe a microarray experiment and make some implicit information codified. The codification of this information required by MIAME reduces data access costs to any third-party researcher and eases the communication between scientists.

2.3 Scientific Journals' Adoption of MIAME

After members of the FGED recommended a formal standard for the publication of microarray data (Brazma et al. 2001) in 2001, they focused on implementing the standard. So in 2002, a working group of FGED contacted editors of several genomics journals to encourage them to adopt FGED's recommended guidelines in their journals' instructions to authors.

The result is that different scientific journals responded quite differently to the MIAME standard. The adoption decisions, made by journals, were exogenous to the research community. For example, *Nucleic Acids Research* endorsed MIAME very soon. Besides, *Nature* also adopted MIAME fairly soon: in September 2002, *Nature* announced that, effective December 1, 2002, authors of manuscripts containing new microarray data must submit complete supplemental information to the editor and at an online data repository using the MIAME standard. "Harried editors can rejoice that, at last, the community is taming the unruly beast that is microarray information" (*Nature* Editorial 2002). Similarly, *Physiological Genomics* adopted the MIAME standard in 2003 “to ensure that what is cutting-edge today is not out-of-date 5 years hence” (Glueck and Dzau 2003).

by Microarray Research Group (MARG) at the Association of Biomolecular Resource Facilities (ABRF), finds that 63% of the respondents indicated that bioinformatics (data management and analysis) is the major challenge for a MA (Microarray) facility."

[8] The Functional Genomics Data (FGED) Society is formerly known as the Microarray Gene Expression Data (MGED) Society. It changed its name in 2010.

But other journals did not immediately respond positively. For example, *Science* did not adopt the standard explicitly, and a senior staff writer at *Science* said that "the journal plans to publish letters from some of the groups working with microarrays, creating a forum for feedback and discussion among a broader group of researchers—something that she said hasn't happened yet" (DeFranceso 2002). Also *Proceedings of the National Academy of Sciences (PNAS)* did not adopt the MIAME standard until October 2005.

A journal's decision on MIAME standard adoption—whether and when to implement the data standard—mostly depends on the organization of the journal and the editor's opinion about this data standard. These adoption decisions at the journal level were made outside the research community. This natural experiment offers a nice variation in journals' timing in implementing the MIAME data standard, which I can use to identify the impact of the data standard.

Table 1 lists the journals that have adopted MIAME by year 2005, and it shows that there is variation in adoption time. I will call this group of journals "treated journals" hereafter in this paper. The control group of journals includes those that have not adopted MIAME by 2005, and I will call this group of journals "control journals". I select the two sets of journals, the "treated journals" and the "control journals", because they include comparatively more articles with microarray data deposited at the public repository Gene Expression Omnibus (GEO).[9]

3 Identification and Data

3.1 Model Specification

Suppose that the journal J implements the MIAME standard and asks all its authors to comply with this microarray data reporting standard.[10] After J's adoption of MIAME, an article published in journal J should on average have better data reporting qualities than before. MIAME requires the key elements of the experiment to be clearly documented in the data, while without this requirement, some key pieces of information may not be provided. As a result, the data from articles published in journal J after MIAME should become more legible and have a relatively higher value to any third-party investigator. Third-party investigators

[9] Two journals are excluded from the empirical analysis, as it is not clear about their MIAME adoptions at the sample period. See section 3.2 for details.

[10] Ochsner et al. (2008) discuss that the implementation of MIAME is not rigorously enforced in reality. Therefore the impact of MIAME may be underestimated. Yet as long as the implementation of MIAME increases the average annotation quality of linked data, it is reasonable to expect that the data should be more legible and accessible in the treated journals than the control journals on average.

may find it easier to understand the data because of the improvement in annotation, which makes them more likely to reuse it. They can reuse the data to verify new findings or to develop new statistical models. Therefore reuse is a direct measure of the effect of MIAME.

Additionally, if the increase in data reporting quality signals the scientific credibility of finding of the article, the article may also receive more citations. However, citation is a very noisy measure of the effect of MIAME, as people cite for many different reasons. Compared with citation, data reuse involves a large investment to understanding the details of the experimental data. Therefore the effect of MIAME on citation may be too small to be significant.[11]

I test whether articles published in journals have a better outcome on average post MIAME implementation with the following specification.

$$(1) \qquad (outcome)_{d,j,t} = f(\alpha_t + \beta(MIAME)_{j,t} + \lambda_j + \gamma'(covariates)_{d,j,t} + \varepsilon_{d,j,t})$$

A unit of observation is a paper or a data set: $(outcome)_{d,j,t}$ indicates the outcome of a paper or a data set that is published in year t, journal j. α_t controls for the publication-year fixed effects (for the data, it is the data-release year fixed effects; the data are published on the public repository GEO either before or after the publication of the article). λ_j controls for the journal fixed effects. The coefficient on the $(MIAME)_{j,t}$ variable is the main estimate of interest, and it is an indicator for whether the article (or the data linked to the article), published in year t, is submitted after journal j's MIAME implementation: a transition from 0 to 1 represents the implementation of MIAME by the journal. This varies across journals over time as shown in Table 1. The variation in timing of journals' MIAME adoption provides the identification. The identification strategy is that once I have controlled for the common shocks to the journal and publication year, no other shocks will affect my estimation on MIAME. I also add $(covariates)_{d,j,t}$ —characteristics of the article (or data)—as extra controls.

One possible problem with the identification is journal selection by the authors: authors can choose to which journal to submit their paper. Having spoken to some scientists working in this field, I do not think selection is an issue, but I will conduct some placebo tests in the robustness-test part of the paper. In one placebo test, I use self-reuse to as a way to examine whether there is a selection behavior. If scientists choose to submit more–valuable data to MIAME-compliant journals, they will also be more likely to reuse the data themselves, i.e., the self-reuse of the data should increase. By looking at the self-reuse patterns, I may rule out the selection story.

I use reuse between 2002 and 2010 to measure the outcome of a data set. I apply the Logit

[11] Piwowar and Vision (2013) discuss several possible channel on how data openness may affect citation.

model to test whether MIAME-compliant data may be more likely to be reused. I use the total number of forward citations a paper has received in 2002–2008 to measure the outcome for a paper with the Poisson Model.

The specification in equation (1) is succinct, and does not allow for time dynamics in citation or reuse. Many papers have shown that the number of citations varies over the life of an article. So in the next specification, I will let data set-year or paper-year be the unit of observation:

$$(2) \qquad (outcome)_{d,j,t,\tau} = f(\alpha_{\tau-t} + \delta_\tau + \beta(MIAME)_{j,t} + \lambda_j + \gamma'(covariates)_{d,j,t} + \varepsilon_{d,j,t,\tau})$$

In this specifications, $(outcome)_{d,j,t,\tau}$ indicates the outcome in calendar year τ of a paper published in year t, journal j (or a data set released in year t, journal j).[12] δ_τ controls for the calendar–year fixed effects, and $\alpha_{\tau-t}$ controls for the age fixed effects. I include the year fixed effects and age fixed effects to control for time–specific shocks that are common across data sets, such as the popularity of microarray technology. λ_j controls for the journal fixed effects. The coefficient on the $(MIAME)_{j,t}$ variable is the main estimate of interest, and it is an indicator for whether the article (or the data linked to the article), published in year t, is submitted after journal j's MIAME implementation. Similarly, $(covariates)_{d,j,t}$ are characteristics of the data or paper and they are used as extra controls. I use the annual reuse to measure a data set's outcome in year τ (between 2002 and 2010) with a Logit model, and I use the annual forward citation to measure a paper's outcome in year τ (between 2002 and 2008) with a Poisson model.

3.2 Data Construction

As discussed earlier, the development of microarray technology made it a useful tool for life sciences research, but as more laboratories acquired this technology, it became crucial to have systematic management of the data. Not only are the data sets large, but also the content is complicated which requires good annotation. In order to support the public use and dissemination of gene expression data, the National Center for Biotechnology Information (NCBI) launched the Gene Expression Omnibus (GEO) in 2001, in an effort to build an expression data repository and online resource for the storage and retrieval of gene expression data from any organism or artificial source.

[12] The publication year of an article and the releasing year of its linked data sets sometimes are different. Thus, I use data–release year in the data reuse regression, and I use article–publication year in the article–citation regression.

Since the launch of GEO, many journals have required articles with microarray data to deposit data at GEO, regardless of whether the journal has implemented MIAME. This has provided me with a chance to identify a paper-data link. To illustrate how I construct the data, I will discuss the process in an example. In 2005, a *Nature* article "Genes that Mediate Breast Cancer Metastasis to Lung" was published (pmid =16049480), and it is a study with a transcriptomic microarray analysis. Following *Nature*'s data–depository policy, the set of microarray data used in this article was submitted to GEO under accession number GSE2603. So from the GEO repository, I am able to obtain a paper-data link. Then I use a catalog of data reuse created by Piwowar and Vision (2013) to track the reuse of the data set GSE2603 from articles on PubMed Central, and in the meanwhile, I also track the number of citations to this paper from Web Of Science (WOS). Piwowar and Vision (2013) estimate data reuse by searching the full text of articles in PubMed Central for mention of a data set's GEO accession number. Going back to the example, for the 2005 *Nature* article and the data set with accession number GSE2603, I find 11 articles mentioning accession number GSE2603 (excluding the original paper). Among the reusing papers, two of them use GSE2603 to validate new findings: one for a cross-species comparison, and one for X-ray crystal structural studies. Some reuse articles combine GSE2603 with other data sets from GEO to develop new bio-statistical algorithms, and some conduct meta-analysis together with other GEO data sets. So the reusing indicator provides a good measure of what future researchers can do based on the original data sets, which may be similar to or different from the original purpose. And the reusing cases can be interpreted as the value of data. Given that microarray data sets are complicated for a third party researcher to understand, it is not surprising that the value of the data may be enhanced by standard data reports which codify implicit information about the data.

In this paper, I collect articles published between years 2002 and 2005, and the linked data that are published before 2006. GEO started to accept deposits of data in 2001, but there are very few data sets in 2001, so I start with year 2002. As of 2006, most of the journals have adopted the MIAME policy, and GEO has changed its format to be more MIAME compatible, with some database modifications aimed at better supporting MIAME elements. Additionally, GEO has increased enforcement of the provision of raw data (Barrett et al. 2010). Thus I choose the year 2005 as the last year of publication for my sample. One note is that the year that a paper is published may be different from the year that the corresponding data is released. In most cases, the data are released on GEO after the publication of the paper.

After collecting the set of papers and the linked data sets, I collect the bibliographic information from WOS and reuse information from Piwowar and Vision (2013). To identify when and which journals adopt MIAME, first I collected information from journals' instructions to authors and checked whether MIAME is required. Second, for the group of journals that require MIAME, I wrote to the journal editors for when they adopted the MIAME standard. Third, if the editors were not sure about when they adopted the MIAME standard, I followed the method by Seamans and Zhu (forthcoming) and used web archive (http://archive.org/web/web.php) to identify the approximate time when the journal web site posted the requirement for MIAME in the instructions to authors. I also checked the information that I collected on journals' MIAME adoptions with two other sources of information, the list of journals that require MIAME-compliant data on FGED Society's website and the list in Rockett and Hellmann (2004).[13]

3.3 Description of the Reuse Articles

The mostly widely used measure for knowledge accumulation is citation. As pointed out in Jaffe et al. (1993), it is difficult to measure the flow of knowledge, and citation does provide us with the trail on how knowledge flows. In this paper, not only do I use citation to measure knowledge accumulation, but I also adopt a novel measure, data reuse, to directly capture another format of sequential innovation: how future researchers build new knowledge based on existing experimental data.

Previous data can be used in different ways. In the context of microarray studies, data can be reused either by re-analyzing the original data (raw data) or by reusing the summarized results. Based on the discussion in Wan and Pavidis (2007) and Rung and Brazma (2013), I decode data reuse in the following five categories.

First, at the most basic level, data can be referred to or compared with new results in an anecdotal way. For example, the *Nature* 2005 paper I described earlier has been reused by two articles to verify new results. The comparison could be done across the microarray data sets with different species, or it could be done across microarray and other types of data, such as X-ray crystal structural studies. This type of reuse is important for detecting fraudulent behavior as discussed in Jin, Jones et al. (2013).

Second, researchers can conduct comparative studies in a systematic way, either using the raw data or the summarized results. This type of comparative study or meta-analysis may increase the statistical power of the results. As explained in Rung and Brazma (2013), "such

[13] The link to FGED's website is http://www.mged.org/Workgroups/MIAME/journals.html. I exclude two journals in my empirical analysis, *Science* and *Journal of Biological Chemistry,* as it is not clear whether MIAME is explicitly implemented by the two journals according to different sources. *Science*, in its instructions to authors, said "For microarray data, *Science* supports ongoing efforts for data standardization (such as that represented by the MIAME project)" (Dec 2003), while FGED website lists it as a journal that has "no mention of MIAME". *Journal of Biological Chemistry* mentioned in its instructions to authors that "Submitted data is encouraged to follow the MIAME checklist" (Mar 2004), but in Rockett and Hellmann (2004), it is not listed as MIAME-compliant according to the response from the journal.

studies draw on the power in numbers: by combing many data sets, the power to detect weak signals is improved, and the large quantity of samples already assayed in conditions that are relevant to the biological questions would often be costly and time consuming to obtain in a single laboratory." For example, one article reuses the *Nature* 2005 data and conducts a meta-analysis that identifies genes regulated in certain types of cancer cells.

Third, data can be reused by researchers in statistics or computer science who are trying to improve algorithms for expression analysis. "In this case the data are generally used in cross-validation setting, often pitting one algorithm against another with relative objective measures of performances. While the use might seem of limited direct interest to biologists, it is worthwhile to consider if one wants a good algorithm for a particular type of data set."[14] The creation and use of algorithms have become a key character in the new data deluge, where researchers can use computer algorithms to identify biological patterns. In genetics studies, it is especially important. An article that reuses the data from 2005 *Nature* article generates a new algorithm called Rank–Rank Hypergeometric Overlap (RRHO) for identification of statistically significant overlap between gene-expression signatures.

Fourth, new methods and tools could be generated besides algorithms with experimental data, like statistical models and software. For instance, some of the articles that reuse the 2005 *Nature* article data develop new web tools. And finally, data could be reused by other researches for new biological findings, either to address the same questions or new ones with similar or different methods.

Based on the data set I collected, I find 366 articles reusing 205 data sets. I use the Medical Subject Headings (MeSH) terms of each reuse article as well as its title and abstract to divide these articles into the five categories shown in Table 2.[15]

Great value can be generated by reusing data. The research conducted by Atul Butte and his colleagues is a good example as described in the introduction. Generally speaking, not only can data reuse facilitate follow-on studies of previous research (i.e., study the same data for the same question with a different method), but it can also create new knowledge with new frameworks in an integrative way, e.g., to perform meta-analysis or develop new algorithms. To identify the value of the reuse articles, I look at the journals in which they are published. I find that the majority of the reuse articles are published in specialized journals,

[14] From Wan and Pavidis (2007).

[15] For example, I put a reuse article in the category "Algorithms" if one of its MeSH describor is "algorithms" or "computational biology". I put a reuse article in the category "Tools and methods" if one of its MeSH describor is "software" or one of its MeSH qualifier is "statistics & numerical data". I put a reuse article in the category "Validating findings" if it submits its own data set(s) to GEO while simultaneously reuses other data set(s). I put a reuse article in the category "Meta-analysis" if its title or abstract includes "meta-analysis". If it is difficult to put an article in the previous four categories from the text information from its MeSH terms, abstract and title, I put it in the category "New biological studies".

such as *BMC Bioinformatics,* but many of them are also published in prestigious journals, such as *Proceedings of the National Academy of Sciences (PNAS).*

Jaffe et al. (1993) also documents that knowledge spillovers, measured with patent citation data, are geographically localized. I compare the geographic location of reuse articles to those of the articles whose data were reused, and I find that data reuse is not strongly constrained by state borders. For example, of all the "data-reuse article" pairs where the reuse articles are written by third-party investigators affiliated with institutions in the United States, I find 72 percent are from different states. Therefore it seems that when data are reused by third-party investigators, geographic proximity is not a determinant factor. When the necessary information that is required to interpret the data is documented in a clear and accessible way, subsequent innovative activities that utilize the data could occur over long distances. Thus standardized data reporting may improve the flow of knowledge.

4 Descriptive Statistics

I will discuss the summary statistics in this section. Since every paper is linked to one or several data sets, and the publication year of the paper and the data can be different, I will show the article and the linked data characteristics separately.

4.1 Descriptive Statistics for Articles

Table 3 provides the descriptive statistics for the articles that generated orginal microarray data. The data set consists of 797 GEO-linked articles. I refer to these articles as "root articles" to distinguish them from "citing articles," articles that reference the root articles. I track citations to each root article from the year of its publication (median publication year = 2004), which yields 3,896 article-year observations. Of the root articles, 64.5 percent are from treated journals—journals that adopted MIAME by 2005 as shown in Table 1. Meanwhile, 53.5 percent of the root articles are submitted after these journals' adoption of the MIAME standard. An article includes 1.2 data sets on average, and Table 4 shows the number of data sets linked to each of the articles.[16] The majority of the articles, 87%, are linked to only one data set. On average, a root article receives 57.5 citations by 2008. My sample includes citations received by root articles between 2002 (the earliest publication year) and 2008 (the latest year I can track), and the average age of an article is 2.0 years. The average number of forward citations per year is 11.7.

[16] I only include the data sets released by 2006 and papers that include at least one data set released by 2006. There are two articles linked to more than 20 data sets released by 2006; I exclude these two articles as outliers.

4.2 Descriptive Statistics for Data Sets

The 797 papers are linked to 994 gene expression data sets in the public repository GEO. I refer to these data sets as "root data" to distinguish them from the "reuse articles," which reuse them. Table 5 reports the summary statistics of the data sets. I track reuse to each root data set from the year of its release (median data-release year = 2004), which yields 6,528 data-year observations. Of the data sets, 68.1 percent are linked to articles in the treated journals. Also 54.4 percent of the data sets are linked to articles submitted after a journal's adoption of the MIAME standard. Among the data sets, 22.2 percent include supplementary raw data sets, and the average number of samples in a data set is 31.7. In terms of sample species, all data sets can be categorized as follows: 33.9 percent involve homo sapiens (i.e., the data object is a human), 24.9 percent involve mus musculus (i.e., the data object is a mouse), and the remaining 41 percent involve other species, such as saccharomyces cerevisiae, arabidopsis thaliana, and rattus norvegicus. Furthermore, 90 percent involve expression profiling by array, while the rest involve other series types such as profiling by serial analysis of gene expression (SAGE) or genome variation profiling by array. Of the data, 21.6 percent have been reused at least once.[17] I divide reuse into two types—reuse by third-party investigators or reuse by some of the data contributors. I call the first type of reuse "third-party reuse", and the second type "self-reuse". Nineteen percent of the data sets have been reused by a third party, and 5 percent by a data contributor.[18]

The sample includes reuses received by root data sets between 2002 (the earliest data release year) and 2010 (the latest year I can track with PubMed central). For each calendar year 5 percent of the data have been reused, and the frequency of annual and total reuse is shown in Table 6. The distribution is very skewed. As I discussed early, I consider a reuse to be a third-party reuse if there is no overlap between the authors of a reuse paper and the data contributors. The percentage of data that has been reused by a third party is 4.4%.[19]

In the results section, I compare the two measures used for the outcome: the number of citations received by a paper, and number of reuses of the data linked to the paper. Piwowar and Vision (2013) document that there is a positive correlation between citation and data

[17] This is similar to Piwowar and Vision (2013). Also note that reuse number is much smaller than the citation, partly because the citation information is from the Web of Science (WOS), while reuse number is from PubMed Central. Piwowar and Vision (2013) document that PubMed Central only includes around one-third of all publications by PubMed, which is a smaller portion compared with WOS.

[18] Note that some of the data sets have been reused by both third-party investigators and some of the original contributors.

[19] Note that the annual self-reuse rate, which equals 4.4, is smaller than the overall self-reuse rate, which equals 5. This is because overall self-reuse dummy is set to 1 if this data set has ever been reused by a data contributor, while annual self-reuse dummy is set to 1 if this data set is reused by a data contributor in the calendar year.

reuse, and I explore this relationship in Figure 3. The x-axis is the number of times a data set is reused, and the y-axis is the number of total citations received by the paper to which that data set is linked. Each dot represents a data set, and green solid line represents the fitted value with a linear regression excluding the two outliers that receive more than 20 reuses.[20] Consistent with Piwowar and Vision (2013), there is a positive relationship between citation and reuse in my data set. Table 7 shows the mean citations to articles, grouped by whether the linked data sets have been reused. For articles whose data sets have been reused, the mean of total citations is 85.483, while for articles whose data sets have never been reused, the mean of total citations is 47.865. The difference between the two groups is 37.618, which is statistically different.

5 Results

5.1 Impact on Data Reuse

In Table 8, I look at the effect of MIAME on data reuse estimating specification (1) with a Logit model.[21] The unit of observation is a data set. The regression includes 994 data sets deposited at GEO linked to the journals shown in Table 1. The dependent variable is a dummy that equals 1 if the data set has been reused at least once between its release year and 2010. The robust standard error clustered by journal is reported in parentheses.

Column (1) reports the results when I control only for release-year fixed effects and the treated journal dummy. The variable Treated Journal is a dummy that equals 1 if the journal has adopted MIAME by 2005. The coefficient on Treated Journal measures the difference between the two groups of journals before the treated journals' adoption of MIAME. It is equal to -0.32 and is not statistically significant. The coefficient on MIAME is 0.583 and statistically significant, which indicates that data linked to articles in treated journals have a higher probability of reuse after the journals' adoption of MIAME compared with the control journals. I calculate the average marginal effect of MIAME (listed in the row AME of MIAME) to be 0.096. This implies that after the treated journals adopted the MIAME standard, the probability of reuse increased by 9.6 percentage points compared with control journals. The average percent of reuse is around 20 percent, so an increase of 9.6 means a rise in the likelihood of reuse by around 50 percent.

Column (2) adds controls for whether the data include supplementary raw data, the number of samples in the data, whether the samples involve human (Homo sapiens), mouse (Mus

[20] The red dot line is the fitted value with a linear regression including the outliers.

[21] I also use a Poisson model for the specification, and it generates similar results.

musculus) or other species, and the series type. These controls are important and add considerable explanatory power to the regression, as measured by the log likelihood. I interpret the Logit result as the average marginal effect: including supplementary raw data increases the probability of reuse by 0.24; adding one sample increases the probability of reuse by 0.0008; compared with data involving non-human & non-mouse species, homo sapiens samples have a higher probability of reuse by 0.129, and data of mus musculus have a higher probability of reuse by 0.07; compared with other series type, array profiling has a higher probability of reuse by 0.08. However, adding all of the data characteristics as controls does not have much of an effect on the relationship between MIAME and reuse.

Column (3) includes journal fixed effects instead of the data-characteristics control, and journal fixed effects control for the common shock within a journal. The journal fixed effects add some explanatory power compared with column (1); however, once again, the journal fixed effects do not have much of an effect on the relationship between MIAME and reuse.

Column (4) includes both the journal fixed effects and data-characteristics controls. Similarly, the control variables matter and help explain the variation in probability of reuse. The coefficients on the data-characteristics controls are similar to those in column (2), and the log likelihood is larger than in previous columns. Nonetheless, all these controls together do not affect the estimated relationship between MIAME and reuse.

Finally, column (5) adds another control variable—the number of total forward citations received by the linked articles. It is reasonable to expect that when a data set is linked to a highly cited article, it may get more attention and more potential reuses: when people cite the article, they may also know the data. The coefficient on the citation variable is equal to 0.00394 and is statistically significant. The average marginal effect is 0.0006. So when the number of times an article is cited increases by 10, or 16 percent of the average total citation, the probability of reuse increases by only 0.006. This seems to be a small effect. So the impact of citation on data reuse is relatively small, although I show a positive correlation between reuse and citation. It is noteworthy that, even with all these control variables, the average marginal effect of MIAME on the probability of reuse is 0.099, which indicates a 50 percent increase in terms of the likelihood of reuse.

In Table 9, I show the Logit regression results under specification (2). In this specification, the unit of observation is data-year. The dependent variable is a dummy that equals 1 if the data is reused in that year. The robust standard error clustered by journal is reported in the parentheses. Compared with Table 8, I control for both the calendar-year fixed effects and data-age fixed effects, and this controls for the age dimension of reuse. Again, column (1) reports the results with year and age dummies and the Treated Journal dummy. The coefficient on MIAME is 0.73, and the average marginal effect is 0.0358. This indicates that data linked to articles in the treated journals are 60% more likely to be reused after the journals' MIAME adoption (since the mean of annual reuse is 0.051). So in this specification with age and year fixed effects, the effect of MIAME is larger than in specification (1).

Figure (4) shows the relationship between MIAME and the probability of reuse. I compare the percent of reused data by age between the treated journals and control journals. The left graph shows the reuse of the data released in 2002 and 2003 for the two groups of journals, when MIAME had not yet been adopted by most of the treated journals.[22] This left graph does not exhibit a clear distinction in terms of reuse between the treated and control journals. The right graph shows the reuse of the data released in 2005, the year that most of the treated journals adopted the MIAME standard. It is clear that the treated journals on average have a higher rate of reuse compared with control journals in 2005.

In the other columns of Table 9, I add more controls, but the effect of MIAME on reuse is not affected. For example, in Column (5), I add journal fixed effects, data characteristics and article citations as controls. The average marginal effect of MIAME is 0.31, which is very similar to column (1). The data characteristics and article citations are of similar signs as in Table 8. In summary, this is consistent with MIAME having a causal positive effect on reuse.

5.2 Impact on Third-Party Data Reuse

Next I explore the effect of MIAME on third-party reuse. I discussed in previous sections that the purpose of MIAME is to improve the quality of data annotation, which helps codify the tacit knowledge about the data and thus reduces data access costs to any third-party investigators. I define third-party reuse as a case when there is no overlap between the data contributors and the authors of the articles in which the data is being reused.[23] Third-party reuse is a more direct measure of data value for the general research community. Table 10 shows the results under specification (2) with a Logit model, when the dependent variable is the third-party reuse dummy. I include the full set of controls that were used in the column (5) of Table 9.

Similar to Table 9, column (1) shows that the impact of MIAME on third-party reuse is statistically significant: the average marginal effect of MIAME on third party reuse is 0.03.

[22] The earliest adoption of MIAME occurred at the end of 2002, and data released in 2002 are linked to articles published before MIAME adoption. Although some of the treated journals adopted MIAME by 2003, the majority of the 2002 and 2003 cohorts from the treated journals were submitted before those journals adopted.

[23] This is done based on Piwowar and Vision (2013) data. Piwowar collected the last names of the data contributors and the authors of the reuse articles. This is a bit of a stretch for "third-party authors", because some common names may indicate different researchers. But I verified that there are not many common names, which should not constitute a problem in the data.

Given that the average annual third-party reuse is 0.044, this is almost a 63 percent increase. The effect is a little stronger compared with the results in Table 9, and this indicates the major impact of MIAME is due to third-party reuse. The control variables have similar signs as before.

In column (2) of Table 10, I use multi-data third-party reuse as my dependent variable; it is equal to 1 when the reuse is from a third party, and the reuse article use multiple data sets. This is to illustrate that one particular use of a MIAME-compliant data set is that it makes investigators more likely to combine different data sets together, i.e., to allow subsequent innovators to build new knowledge based on many previous data sets. The average marginal effect of MIAME is equal to 0.0349 and statistically significant. Therefore, compared with non-MIAME compliant data sets, MIAME-compliant ones are 3.5 percentage points more likely to be used in analysis involving multiple data sets.

5.3 Impact on Article Citation

Table 11 reports results for article citations based on specification (1). I use a Poisson model, and the unit of observation is a paper. The dependent variable is the number of total citations received by a paper between its publication year and 2008, and the robust standard error clustered by journal is reported in parentheses.

Column (1) reports the results when I control only for publication-year fixed effects and the treated journal dummy, and the coefficient on MIAME equals 0.129 and is not statistically significant. The incidence-rate ratio of MIAME (IRR of MIAME) is 1.138. The coefficient on Treated Journal equals 0.09 and is not statistically significant. Column (2) adds controls for the number of data sets linked to an article, whether at least one of the linked data sets has supplementary raw data, the total number of samples, the species of the linked data, and series type. These controls add explanatory power to the regression, as measured by the log likelihood. But again, the coefficient on MIAME is not statistically significant. I find that articles with raw data, with more samples, and with human subjects are more frequently cited. Including raw data increases citation by 47%, adding an extra sample increases citation by 0.3%, and compared with non-human and non-mouse subjects, including a data set with a human subject sample increases citation by 28%. Unlike previous regressions in which the dependent variable was data reuse, the coefficient on the series type dummy is negative. This implies that compared with articles with other types of data, articles linked to array profiling data have on average 33% less citation. One possible explanation could be that the profiling-by-array data is the major target data type by the MIAME standard (89 percent of the sample), and the implementation of the MIAME reporting standard increases the reusability of these data. However, the link between data reusability and citation is weak. Moreover, this also shows the difference between data-reuse value and the value of the article. Column (3) adds journal fixed effects, and column (4) includes both characteristics control variables and journal fixed effects. The coefficients on MIAME are not statistically significant in these two columns.

Table 12 shows the regression results under specification (2), with a similar structure as Table 11 in terms of controls. The unit of observation is paper-year, and the dependent variable is the annual forward citation. The specification includes the calendar and age fixed effects. The results are very similar to those in Table 11, with no significant effect of MIAME on citation. The non-significant coefficient on MIAME also suggests that there is no selection effect after journals implement MIAME, and article quality between treated journals and control journals, measured by citation, does not change much after MIAME adoption.[24]

Figure (5) shows the relationship between MIAME and citation. In Figure (5), I compare annual forward citations by age between articles in treated and control journals. As in Figure (4), the left graph shows the citation of articles published in year 2002 and 2003, when most articles in treated journals are not MIAME-compliant. The articles in treated journals receive slightly more citations in the first three years post publication, but then the control journal articles catch up. The right graph shows citations of articles published in 2005, a year by which most treated journals had adopted MIAME. The difference between the two groups is similar to the left graph, which is consistent with the regression results in Table 11.

5.4 Impact on Third-Party Data Reuse by Journal or Institution Quality

Does the effect of MIAME vary by journal and institution quality? In this section, I explore this question by running specification (2) in subsamples divided by journals' ISI Impact Factors and authors' institution rankings. In column (1) of Table 13, I only include journals with ISI Journal Impact Factor greater than or equal to 10.[25] I call this group of journals "high-ranked journals." In column (2), I only include journals with ISI Journal Impact Factor less than 10; and I call this group of journals "low-ranked journals." I find that the coefficient on MIAME is statistically significant in column (1) but not significant in column (2). Therefore the impact of MIAME is statistically stronger for prestigious journals. In columns (3) and (4), I only include the articles that are written by authors who are affiliated with institutions located in

[24] I also use a negative-binomial model and obtain similar results.

[25] I use the average ISI Impact Factor of a journal between 2002 and 2004 as that journal's impact factor. Two journals are missing impact-factor information as they were established in 2003 and the impact factor is calculated based on the citation history of the last three years. The median impact factor of the journals in the sample is around 7.

the United States. In column (3), I only include articles of which at least one of the authors is from a top-20 medical school in the US,[26] while in column (4), I only look at articles for which none of the authors are from a top 20 medical school. Similarly, the coefficient of MIAME in column (3) is statistically significant, while the coefficient in column (4) is less significant, which suggests that the impact is stronger within top medical schools. These results suggest that reputation and data reporting standards are complementary.

6 Robustness Tests

When I identify the effect of MIAME, the key argument is that journals' implementation of MIAME, although not rigorously enforced, improved the average quality of data documentation and thus made data more likely to be reused. One potential selection problem could be that researchers may submit high-reuse-value data to the journals with a MIAME requirement, and so the effect I find for MIAME may be due to selection. By talking to the scientists in this field, I learned that this is not likely.[27] Consistent with this, the results on article citation imply there is no significant change in article quality after MIAME. Nevertheless, I will tackle the selection problem directly with a placebo test on self-reuse.

Suppose there is a selection problem: namely, that researchers submit more valuable data to treated journals that adopt MIAME. If this is the case, not only should I observe an increase in third-party reuse, but I should also find a rise in reuse by the data contributors. Accordingly, I define self-reuse as the situation in which there is an overlap between the data contributors and the reuse-article authors. So if the effect of MIAME is due to selection, I should find a positive effect on self-reuse. However, if I do not find that self-reuse increases with MIAME, the effect should be from the reduction of access costs due to standard data reports.

I conduct such a placebo test in this section to see whether self-reuse is affect by MIAME. I run specification (2) with a Logit model, and the dependent variable is self-reuse. Table 14 shows the regression results. I cluster standard errors at the journal level.

In Table 14, none of the coefficients on MIAME are statistically significant, and this implies that self-reuse does not change significantly post-MIAME. So the evidence suggests that selection is the not the main issue here. In particular, some coefficients on certain

[26] I use the ranking information from *US News* "Best Medical Schools: Research" (2013) as the source of information. Note that some teaching hospitals are affiliated with medical schools, and I try to include these hospitals when I can identify the affiliation relationship (http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-medical-schools/research-rankings).

[27] Having talked to several researchers in microarray-related fields, I learned that the major factor that affects researchers' choice of journal for publication is the article content.

control variables are no longer significant. For example, the coefficient on the supplementary raw data dummy becomes statistically insignificant. This is reasonable since self-reuse does not count on the public availability of raw data. The coefficient on article citation is no longer statistically significant, which is reasonable given the self-reuse content.

Next I explore an institutional characteristic of GEO to conduct another robustness test. Previous sections investigated reuse of the original data submitted by data contributors.[28] However, GEO not only provides the platform to store the original data, its staff also manually curate part of the original data and transfer them into a more uniform format. Because of this, researchers can easily compare different data sets across species or across platforms with the curated ones.

A researcher can either reuse original data submitted by data contributors, or data curated by GEO staff. For each of the curated data sets I know which original data set it is based on. As discussed earlier, standardized data reports increases the reuse of the data. For the data sets curated by GEO staff, there should not be a difference in terms of reuse by whether the original data set is from a MIAME-compliant journal.[29]

In Table 15, I test whether reuse of curated data varies with MIAME by using specification (2) with full controls. The coefficient of MIAME is not significant, and the result confirms that the effect of data reporting standards: reuse of curated data is not significantly different between MIAME-compliant journals and non-MIAME compliant journals once they are curated into a uniform reporting format.

7 Concluding Remarks

In this paper I document how MIAME, a data reporting standard that ensures microarray data is more legible, reduces access costs and facilitates subsequent use of experimental data in scientific research. I explore the variation in journals' adoption of MIAME, and I find that microarray data submitted after a journal adopts MIAME is at least 50 percent more likely to be reused. The result is not likely to be driven by selection on the quality of the data or the article to which the data is linked.

This analysis does not evaluate the overall welfare consequences of data reporting stan-

[28] The information on data contributors is collected from the GEO website. For the majority of cases, the data contributors are the article authors. But in some cases, the data contributors are a subset of the article authors.

[29] It is not clear which data sets are curated and which are not. GEO staff mentions (in correspondence) that they only curate microarray gene expression data with raw data and enough information about the data. Therefore, we can expect that the data that has been curated must have enough descriptions. Only looking at the subgroup of data that has been curated, I do not observe that MIAME plays a significant role in data reuse.

dards. In principle, it is possible that the increased availability and accessibility of existing data may discourage scientists from performing new experiments. If this effect dominates the finding that data reporting standards facilitate subsequent innovation, there would be welfare loss. However, with the creation of new experiment data held constant, these results suggest that data reporting standards make the data that have been created more accessible to a wide community, which may be socially beneficial.[30] Additionally, the potential costs of standardizing data reports are not considered.

As technology advances, the costs to generate large-scale data have been decreasing and the quantity of experimental data has grown rapidly, particularly in the field of life sciences research. Understanding how to facilitate the reuse of experimental data is important, as data reuse provides a new, and potentially cheaper and faster, way of making scientific discoveries. For example, Professor Atul Butte and his colleagues, discussed in the introduction, were able to accelerate the clinical trial processes of drugs out of their data exploration. They were able to start recruiting patients for a phase-2 clinical trial within 15 to 20 month of a discovery because the drugs are already FDA-approved. This is a shortcut compared with the traditional method of "trial and error" drug discovery, which usually takes several years. Moreover, the costs to start such clinical trials are also much lower.

Certainty, it is too early to see the full impact of data reuse on scientific innovation yet. MIAME is an experiment for the community of gene expression studies, but many similar standards, such as Minimum Information About a Proteomics Experiment (MIAPE), have also been initiated. Though it may take years for such data reporting standards to become the norm, the prospect from subsequent use of existing data is very promising in the long run.

[30] Yet an investigation of public repository GEO shows fast growth of data that has been submitted.

References

[1] Agrawal, Ajay, and Avi Goldfarb. "Restructuring Research: Communication Costs and the Democratization of University Innovation." American Economic Review, 2008: 98(4): 1578-1590.

[2] Azoulay, Pierre, Joshua Graff-Zivin, Danielle Li, and Bhaven Sampat. "Public R&D Investments and Private Sector Patenting: Evidence from NIH Funding Rules." September 2013.

[3] Barrett, Tanya, Dennis B. Troup, Stephen E. Wilhite, et al. "NCBI GEO: archive for functional genomics data sets–10 years on." Nucleic Acids Res, 2011(39):D1005-10.

[4] Brazma, Alvis, Pascal Hingamp, John Quackenbush, et al. "Minimum Information About a Microarray Experiment (MIAME) - Toward Standards for Microarray Data." Nature, 2001: 365-371.

[5] Chervitz, Stephen A., Eric W. Deutsch, Dawn Field, et al. "Data Standards for Omics Data: The Basis of Data Sharing and Reuse." In Bioinformatics for Omics Data: Methods and Protocols, by Bernd Mayer, 31-69. Springer, 2011.

[6] Culhane, Aedin C., Thomas Schwarzl, Razvan Sultana, et al. "GeneSigDB–A Curated Database of Gene Expression Signatures." Nucleic Acids Research, 2010(38): D716-725.

[7] DeFranceso, Laura. "Journal Trio Embraces MIAME." Genome Biology, 2002.

[8] Editorial. "Microarray Standards at Last." Nature, 2002: 323.

[9] Furman, Jeffrey L., and Scott Stern. "Climbing Atop the Shoulder of Giants: The Impact of Institutions on Cumulative Research." American Economic Review, 2011: 1933-1963.

[10] Galasso Alberto. and Mark Schankerman. "Patents and Cumulative Innovation: Causal Evidence from the Courts." CEPR discussion paper 9458, 2013.

[11] Glueck, Susan B., and Victor J. Dzau. "Our New Requirement for MIAME standards." Physiological Genomics, 2003: 1-2.

[12] Harrington, Christina A., et al. "The ABRF MARG Microarray Survey 2008: Sensing the State of Microarray Technology." 2008 MARG Survey Poster available online at http://www.abrf.org/ResearchGroups/Microarray/EPosters/020808Final2008MARGposter.pdf.

[13] Jaffe, Adam B., Manuel Trajtenberg and Rebecca Henderson. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." Quarterly Journal of Economics, 1993(108): 577-598

[14] Jin, Ginger, Ben Jones, Susan Feng Lu, and Brian Uzzi, "The Reverse Matthew Effect: Catastrophe and Consequence in Scientific Teams." September 2013.

[15] Jones, Benjamin F. "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" Review of Economic Studies, 2009: 283-317.

[16] Jones, Charles I. "R&D-Based Models of Economic Growth." Journal of Political Economy, 1995: 759-784.

[17] Keating, Peter, and Alberto Cambrosio. "Clinical Hematology Meets the New Genetics (1980-2000)." New Genetics and Society, 2004: 15-45.

[18] Manyika, James, Michael Chui, Brad Brown, et al. "Big data: The Next Frontier for Innovation, Competition, and Productivity". McKinsey Global Institute Report, May 2011

[19] Mogoutov, Andrei, Alberto Cambrosio, Peter Keating, and Philippe Mustar. "Biomedical Innovation at the Laboratory, Clinical, and Commercial Interface: A New Method for Mapping Research Projects, Publications and Patents in the Field of Microarrays." Journal of Informetrics, 2008: 341-353.

[20] Mokyr, Joel. "The Intellectual Origins of Modern Economic Growth." The Journal of Economic History, 2005: 285-351.

[21] Murray, Fiona, Philippe Aghion, Mathias Dewatripont, et al. "Of Mice and Academics: Examining the Effect of Openness on Innovation." 2009 March, NBER Working Paper 14819.

[22] NCBI (National Center for Biotechnology Information) A Science Primer, "Microarrays: Chipping Away at the Mysteries of Science and Medicine", available online at http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html, July 27, 2007.

[23] Ochsner, Scott A., David L Steffen, Christian J Stoeckert Jr, and Neil McKenna. "Much Room for Improvement in Deposition Rate of Expression Microarray Datasets." Nature Methods, 2008; 5(12): 991.

[24] Piwowar, Heather A., and Wendy A. Chapman. "A Review of Journal Policies for Sharing Research Data." Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing (ELPUB). Toronto, Canada, 2008.

[25] Piwowar, Heather A., Todd J. Vision, Michael C. Whitlock. "Data Archiving Is a Good Investment." Nature, 2011: 285.

[26] Piwowar, Heather A, and Todd J. Vision. "Data Reuse and the Open Citation Advantage", PeerJ, 2013, Preprint.

[27] Seamans, Robert, and Feng Zhu. "Responses to Entry in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers." Management Science (forthcoming).

[28] Rockett, John C., and Gary M. Hellmann. “Confirming microarray data–is it really necessary?” Gemonics, 2004 83(4): 541-549.

[29] Rogers, Susan, and Alberto Cambrosio. "Making a New Technology Work: The Standardization and Regulation of Microarrays." Yale Journal of Biology and Medicine, 2007: 165-178.

[30] Romer, Paul M. "Endogenous Technological Change." Journal of Political Economy, 1990: S71-102.

[31] Rung, Johan, and Alvis Brazma. "Reuse of Public Genome-wide Gene Expression Data." Nature Reviews Genetics, 2013(14): 89-99.

[32] Sirota, Marina, Joel T. Dudley, Jeewon Kim, et al. "Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data." Science Translational Medicine 2011(3): 96ra77.

[33] Trivedi, Pravin K., and A. Colin Cameron. "Microeconometrics Using Stata". Revised Edition 2010, Stata Press.

[34] Wan, Xiang, and Paul Pavlidis. "Sharing and Reusing Gene Expression Profilling Data in Neuroscience." Neuroinformatics, 2007; 5(3): 161-175.

[35] Williams, Heidi L. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." Journal of Political Economy, 2013 121(1): 1-27.

Table 1: List of MIAME-Compliant Journals and Control Journals

Treated Journals		Control Journals
MIAME-Compliant Journals (by 2005)		Non-MIAME Compliant Journals (by 2005)
Journal Title	MIAME	Journal Title
Nature	2002 Dec	Proceeding of National Academy of Sciences
Nature Genetics	2002 Dec	Journal of Bacteriology
Nucleic Acids Research	2002 Dec	Blood
Nature Immunology	2002 Dec	Development Biology
Nature Medicine	2002 Dec	Molecular Biology Cell
Nature Biotechnology	2002 Dec	Molecular Endocrinology
Nature Methods	2002 Dec	Plant Journal
Nature Cell Biology	2002 Dec	FASEB Journal
Physiol Genomics	2003 Jan	Molecular Microbiology
Cell	2003 Feb	Oncogene
Molecular Cell	2003 Feb	Invest Ophthalmology Vis Sci
Lancet	2003 Feb	Journal of Neuroscience
Journal of Immunology	2003 July	Genomics
Genome Biology	2003 Sep	Journal of Virology
American Journal of Pathology	2003 Sep	Biology Reproduction
PLOS Biology	2003 Oct	Infection Immunology
BMC Bioinformatics	2003 Oct	Bioinformatics
BMC Genomics	2003 Oct	Environment Health Perspective
Cancer Research	2004 Jun	Europe Journal of Neuroscience
Plant Cell	2004 Aug	Genes Development
Plant Physiology	2004 Aug	
Development	2005 Jan	
Human Molecular Genetics	2005 Mar	

Note: This table lists the group of journals (Treated Journals) that implemented MIAME standard before 2005 and the group of journals (Control Journals) that did not implement MIAME before 2005. Data collected from journals' instruction to authors, correspondences with journal editors, and the web archive (http://archive.org/web/web.php).

Table 2: Decoding the Reuse Articles

Purposes of reuse	Frequency	Percent
Algorithms	55	15.03
Tools and methods	92	25.14
Validating findings	50	13.66
Meta-analysis	42	11.48
New biological studies	127	34.70
Total	366	

Table 3: Summary Statistics for Articles

Variable	Mean	Median	SD	Min	Max
Article characteristics (n=797 articles)					
MIAME	0.645	1	0.479	0	1
Treated Journal	0.535	1	0.499	0	1
Publication year	2004.112	2004	0.941	2002	2005
No. of data sets	1.247	1	0.777	1	8
Supplementary	0.243	0	0.429	0	1
Total sample	39.514	18	74.974	1	1336
Taxonomy–Homo sapiens	0.349	0	0.477	0	1
Taxonomy–Mus musculus	0.253	0	0.435	0	1
Taxonomy–Other	0.398	0	0.490	0	1
Series type–Profiling by array	0.886	1	0.318	0	1
Total Forward Citations	57.541	30	83.542	0	758
Article-year characteristics (n=3896 article*year observations)					
Year	2005.965	2006	1.545	2002	2008
Age	2.035	2	1.545	0	6
Forward Citations	11.771	6	19.129	0	243

Note: MIAME is a dummy variable to indicate whether the article has been published after the journal's adoption of MIAME. Treated Journal is a dummy variable to indicate whether the article is published in a journal that adopted MIAME by 2005 shown in Table 1. Supplementary is a dummy variable to indicate whether at least one of the article-linked GEO data set includes the supplementary raw data. Total sample is the total number of samples in the article-linked data set(s). Taxonomy dummies are about the sample species in the data linked to the article. Homo sapiens is equal to 1 if any of the data sets to which the paper is linked involves a human subject, and mus musculus is equal to 1 if any of the data sets to which the paper is linked involves a mouse subject (and none involves a human subject). Series type is a dummy variable to indicate whether the linked data set involves expression profiling data by array. Total forward citation is a count variable for the number of cites that the paper has received by 2008.

Table 4: Number of Linked Data Sets per Article

No. of data sets per article	Frequency	Percent
1	695	87.20
2	49	6.15
3	27	3.39
4	18	2.26
5	4	0.50
6	1	0.13
7	2	0.25
8	1	0.13
Total	797	

Table 5: Summary Statistics for Data Sets

Variable	Mean	Median	SD	Min	Max
Data characteristics (n=994 data sets)					
MIAME	0.544	1	0.500	0	1
Treated Journal	0.681	1	0.477	0	1
Data release year	2004.315	2004	0.953	2001	2006
Total reuse	0.216	0	0.412	0	1
Total third-party reuse	0.189	0	0.392	0	1
Total self-reuse	0.051	0	0.221	0	1
Supplementary	0.222	0	0.416	0	1
Sample count	31.683	14	58.614	0	741
Taxonomy–Homo sapiens	0.339	0	0.474	0	1
Taxonomy–Mus musculus	0.249	0	0.433	0	1
Taxonomy–other	0.411	0	0.492	0	1
Series type–Profiling by array	0.891	1	0.311	0	1
Data-year characteristics (n=6528 data*year observations)					
Year	2007.147	2007	1.995	2001	2010
Age	2.853	3	1.995	0	9
Reuse	0.051	0	0.220	0	1
Third-party reuse	0.044	0	0.205	0	1
Self-reuse	0.009	0	0.095	0	1

Note: MIAME is a dummy variable to indicate whether the data set is linked to a paper that is published after the journal's adoption of MIAME. Treated Journal is a dummy variable to indicate whether the data set is linked to a paper published in a journal that adopted MIAME by 2005 shown in Table 1. Supplementary is a dummy variable to indicate whether the data include the supplementary raw data. Sample Count is the number of the sample. Taxonomy dummies are about the sample species. Homo sapiens is equal to 1 if the sample involves a human subject, and mus musculus is equal to 1 if the sample involves a mouse subject. Series type is a dummy variable to indicate whether the data set is expression profiling data by array.

Table 6: Frequency of Total and Annual Reuse

	Total reuse		Annual reuse	
	Frequency	Percent	Frequency	Percent
0	779	78.37	6194	94.88
1	125	12.58	262	4.01
2	47	4.73	36	0.55
3	18	1.81	22	0.34
4	1	0.10	6	0.09
5	6	0.60	3	0.05
6	8	0.80	0	0
7	1	0.10	0	0
8	1	0.10	1	0.02
9	5	0.50	0	0
11	1	0.10	1	0.02
13	0	0	2	0.03
28	1	0.10	0	0
29	0	0	1	0.02
59	1	0.10	0	0
Total	994		6528	

Table 7: Citation Means between Reused Data and Non-Reused Data

	(1) reused mean	(2) non reused mean	(3) difference (1)-(2)	(4) p-value of difference
total citations	85.483	47.865	37.618	[0.000]
N	205	592		

Note: This table compares the total number of citations for articles whose data has been reused at least once relative to articles whose data has never been reused. Article-level observations. Sample in Column (1) includes articles whose data have been reused at least once. Column (2) includes articles whose data sets have never been reused. The p-value reported in Column (4) is from a *t-test* for a difference in means between column (1) and column (2).

Table 8: Logit Model: Impact on Total Reuse

	(1)	(2)	(3)	(4)	(5)
[AME of MIAME]	[0.096**]	[0.096**]	[0.108**]	[0.083**]	[0.099**]
MIAME	0.583**	0.666**	0.687**	0.589*	0.715**
	(0.047)	(0.045)	(0.042)	(0.041)	(0.045)
Treated Journal	-0.321	-0.385			
	(0.382)	(0.392)			
Supplementary		1.365***		1.349***	1.339***
		(0.185)		(0.196)	(0.197)
Sample Count		0.006**		0.006*	0.005
		(0.003)		(0.003)	(0.003)
Taxonomy–Homo sapiens		0.831***		0.760***	0.642***
		(0.198)		(0.232)	(0.226)
Taxonomy–Mus musculus		0.477*		0.539*	0.487
		(0.280)		(0.304)	(0.312)
Series type–Profiling by array		0.650*		0.830**	0.927***
		(0.386)		(0.360)	(0.317)
Citation to article					0.004**
					(0.002)
Observations	994	994	936	936	936
Data Release Year FE	YES	YES	YES	YES	YES
Journal FE	NO	NO	YES	YES	YES
Log Likelihood	-509.1	-454.1	-462.8	-417.6	-411.4

Note: Data set–level observations. All estimates are from Logit models. The sample includes all microarray data sets released on the Gene Expression Omnibus (GEO) by 2006, which are linked to research articles published in the journals in Table 1. Two articles, each of which includes over 20 data sets, are dropped. Robust standard errors clustered by journal are shown in parentheses. ***: p<0.01, **: p<0.05, *: p<0.1. The dependent variable is a dummy variable to indicate whether the data set has being reused by articles on PubMed Central by 2010. Definitions of independent variables are the same as in Table 5. AMEs of MIAME show the average marginal effect of MIAME in Logit models. Note that some of the observations are dropped from column (3), (4) and (5). This is because when journal dummies are controlled, some journal dummies perfectly predict the dependent variable so that some of the observations are dropped.

Table 9: Logit Model: Impact of MIAME on Annual Reuse

	(1)	(2)	(3)	(4)	(5)
[AME of MIAME]	[0.036**]	[0.032**]	[0.044***]	[0.027**]	[0.031**]
MIAME	0.730***	0.716**	0.890***	0.598**	0.693**
	(0.283)	(0.281)	(0.298)	(0.273)	(0.290)
Treated Journal	-0.636	-0.670*			
	(0.416)	(0.372)			
Supplementary		1.387***		1.245***	1.227***
		(0.138)		(0.155)	(0.146)
Sample Count		0.005***		0.005***	0.004**
		(0.002)		(0.002)	(0.002)
Taxonomy–Homo sapiens		0.913***		0.788***	0.720***
		(0.177)		(0.256)	(0.234)
Taxonomy–Mus musculus		0.471		0.568*	0.525
		(0.299)		(0.324)	(0.329)
Series type–Profiling by array		0.605		0.833**	0.909***
		(0.398)		(0.348)	(0.291)
Citation to article					0.003**
					(0.001)
Observations	6329	6329	5964	5964	5964
Year FE	YES	YES	YES	YES	YES
Age FE	YES	YES	YES	YES	YES
Journal FE	NO	NO	YES	YES	YES
Log Likelihood	-1222	-1094	-1127	-1028	-1019

Note: Data set-year-level observations. All estimates are from Logit models. Same sample construction as in Table 8. Robust standard errors clustered by journal are shown in parentheses. ***: p<0.01, **: p<0.05, *: p<0.1. The dependent variable is a dummy variable to indicate whether the data set is reused by articles on PubMed Central in the calendar year after its release. Definitions of independent variables are the same as in Table 5. AMEs of MIAME show the average marginal effect of MIAME in Logit models. Also note that the some of the observations are dropped in column (3), (4), and (5). This is because when journal dummies are controlled, some journal dummies perfectly predict the dependent variable so that some of the observations are dropped.

Table 10: Logit Model: Impact of MIAME on Third-Party Reuse

	(1)	(2)
[AME of MIAME]	[0.030**]	[0.035*]
MIAME	0.775**	0.884*
	(0.383)	(0.472)
Supplementary	1.389***	1.432***
	(0.150)	(0.192)
Sample Count	0.004**	0.003**
	(0.002)	(0.001)
Taxonomy–Homo sapiens	0.862***	0.985***
	(0.246)	(0.279)
Taxonomy–Mus musculus	0.407	0.683*
	(0.387)	(0.380)
Series type–Profiling by array	0.749***	0.622**
	(0.287)	(0.308)
Citations to the article	0.003***	0.003***
	(0.001)	(0.001)
Observations	5775	4783
Age FE	YES	YES
Year FE	YES	YES
Journal FE	YES	YES
Log Likelihood	-868.9	-729.5

Note: Data set-year-level observations. All estimates are from logit models. Same sample construction as in Table 8. Robust standard errors clustered by journal are shown in parentheses. ***: p<0.01, **: p<0.05, *: p<0.1. In column (1), the dependent variable is a dummy variable to indicate whether the data set has been reused by articles written by third-party investigators in the calendar year after its release. Third-party article is defined as the situation in which there is no overlap between the data contributors and authors of the reuse article. In column (2), the dependent variable is a dummy variable to indicate whether the data set has been reused by articles written by third-party investigators in the calendar year after its release, and one of the reuse articles reuses more than one data set. Definitions of independent variables are the same as in Table 5.

Table 11: Poisson Model: Impact of MIAME on Total Citations

	(1)	(2)	(3)	(4)
[IRR of MIAME]	[1.138]	[1.135]	[0.947]	[0.860]
MIAME	0.129	0.127	-0.055	-0.151
	(0.225)	(0.206)	(0.127)	(0.126)
Treated Journal	0.092	0.112		
	(0.330)	(0.333)		
No. of data sets		0.025		0.010
		(0.057)		(0.028)
Supplementary		0.386***		0.223*
		(0.106)		(0.117)
Total sample		0.002***		0.002**
		(0.001)		(0.001)
Taxonomy–Homo sapiens		0.248***		0.389***
		(0.096)		(0.087)
Taxonomy–Mus musculus		-0.021		0.160
		(0.109)		(0.108)
Series type–Profiling by array		-0.388		-0.241**
		(0.243)		(0.118)
Observations	797	797	797	797
Publication Year FE	YES	YES	YES	YES
Journal FE	NO	NO	YES	YES
Log Likelihood	-27582	-24860	-14669	-12879

Note: Paper-level observations. All estimates are from Poisson models. The sample includes all articles published in journals in Table 1 between 2002 and 2005, which have linked microarray data sets released on GEO by 2006. Two articles, each of which includes over 20 data sets, are dropped. Robust standard errors clustered by journal are shown in parentheses. ***: p<0.01, **: p<0.05, *: p<0.1. The dependent variable is a count variable to indicate the number of total forward citations that the article has received by 2008. IRRs of MIAME show the incidence-rate ratios of MIAME in Poisson models. Definitions of independent variables are the same as in Table 3.

Table 12: Poisson Model: Impact of MIAME on Annual Citations

	(1)	(2)	(3)	(4)
[IRR of MIAME]	[1.082]	[1.077]	[0.925]	[0.836]
MIAME	0.079	0.074	-0.078	-0.180
	(0.240)	(0.219)	(0.126)	(0.126)
Treated Journal	0.123	0.144		
	(0.353)	(0.355)		
No. of data sets		0.023		0.010
		(0.058)		(0.026)
Supplementary		0.385***		0.215*
		(0.105)		(0.123)
Total sample		0.002***		0.0012**
		(0.001)		(0.001)
Taxonomy–Homo sapiens		0.248**		0.398***
		(0.098)		(0.087)
Taxonomy–Mus musculus		-0.029		0.164
		(0.111)		(0.108)
Series type–Profiling by array		-0.399		-0.244**
		(0.247)		(0.120)
Observations	3896	3896	3896	3896
Age FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Journal FE	NO	NO	YES	YES
Log Likelihood	-35303	-32640	-22229	-20443

Note: Paper-year-level observations. All estimates are from Poisson models. Same sample construction as in Table 11. Robust standard errors clustered by journal are shown in parentheses. ***: p<0.01, **: p<0.05, *: p<0.1. The dependent variable is a count variable to indicate the number of forward citations that the article has received in a calendar year. Definitions of independent variables are the same as in Table 3.

Table 13: Logit Model: Impact of MIAME on Annual Third-party Reuse by Journals/Institutions

	(1)	(2)	(3)	(4)
[AME of MIAME]	[0.053**]	[0.020]	[0.080***]	[0.058*]
MIAME	1.194**	0.492	1.616***	0.952*
	(0.506)	(0.412)	(0.553)	(0.535)
Supplementary	1.356***	1.485***	1.262**	0.893***
	(0.324)	(0.141)	(0.513)	(0.170)
Sample Count	0.005***	0.010***	0.006*	0.003
	(0.002)	(0.001)	(0.003)	(0.003)
Taxonomy–Homo sapiens	1.114***	1.080***	1.327***	0.941**
	(0.352)	(0.263)	(0.307)	(0.434)
Taxonomy–Mus musculus	0.367	0.842*	0.804**	0.218
	(0.578)	(0.448)	(0.403)	(0.678)
Series type–Profiling by array	1.046*	0.643	0.707	0.415
	(0.575)	(0.501)	(0.471)	(0.555)
Citation to article	0.003***	0.009**	0.001	0.003**
	(0.001)	(0.004)	(0.002)	(0.001)
Observations	2166	2917	1311	1167
Year FE	YES	YES	YES	YES
Age FE	YES	YES	YES	YES
Journal FE	YES	YES	YES	YES
Log Likelihood	-352.8	-444.8	-243.7	-252.8

Note: Data set-year-level observations. All estimates are from Logit models. Sample analysis similar to column (1) of Table 10, except the sample construction. Column (1) only includes data sets linked to articles published in high-ranked journals, i.e., the average ISI Journal Impact Factor between 2002 and 2004 is higher than or equal to 10, and column (2) includes low-ranked journals. Columns (3) and (4) only include data sets linked to articles written by authors in the United States, and column (3) is only for the articles where at least one of the authors is from a top-20 medical school (ranking information from *US News* 2013), while column (4) is for the remaining articles.

Table 14: Logit Model: Impact of MIAME on Annual Self-Reuse

	(1)	(2)	(3)	(4)	(5)
[AME of MIAME]	[0.005]	[0.004]	[0.008]	[0.003]	[0.003]
MIAME	0.507	0.339	0.524	0.181	0.215
	(0.440)	(0.464)	(0.434)	(0.439)	(0.424)
AME of MIAME	0.005	0.004	0.008	0.003	0.003
	(0.005)	(0.005)	(0.007)	(0.006)	(0.006)
Treated Journal	-0.626	-0.494			
	(0.533)	(0.536)			
Supplementary		0.722**		0.473	0.468
		(0.303)		(0.351)	(0.347)
Sample Count		0.006***		0.007***	0.007***
		(0.001)		(0.002)	(0.002)
Taxonomy–Homo sapiens		0.029		-0.194	-0.195
		(0.396)		(0.461)	(0.449)
Taxonomy–Mus musculus		0.466		0.780	0.790
		(0.399)		(0.480)	(0.483)
Series type–Profiling by array		1.866**		2.114**	2.162***
		(0.928)		(0.883)	(0.811)
Citation to article					0.001
					(0.001)
Observations	5662	5662	4026	4026	4026
Year FE	YES	YES	YES	YES	YES
Age FE	YES	YES	YES	YES	YES
Journal FE	NO	NO	YES	YES	YES
Log Likelihood	-326.7	-306.4	-298.6	-276.4	-276.1

Note: Data set-year-level observations. All estimates are from Logit models. Same sample construction as in Table 8. Robust standard errors clustered by journal are shown in parentheses. ***: p<0.01, **: p<0.05, *: p<0.1. The dependent variable is a dummy variable to indicate whether the data set has been reused by articles written by some of the data contributors in the calendar year after its release. Definitions of independent variables are the same as in Table 5.

Table 15: Logit Model: Impact of MIAME on Annual Third-Party Reuse of Curated Data

	(1)
[AME of MIAME]	[-0.006]
MIAME	-0.178
	(0.403)
Supplementary	1.077**
	(0.544)
Sample Count	0.014***
	(0.002)
Taxonomy–Homo sapiens	1.675**
	(0.724)
Taxonomy–Mus musculus	0.439
	(0.743)
Citation to article	0.002
	(0.002)
Observations	1285
Year FE	YES
Age FE	YES
Journal FE	YES
Log Likelihood	-160.2

Note: Curated data set-year-level observations. All estimates are from Logit models. The sample includes the curated data sets by GEO staff from the sample described in Table 8. Robust standard errors clustered by journal are shown in parentheses. ***: p<0.01, **: p<0.05, *: p<0.1. The dependent variable is a dummy variable to indicate whether the curated data set is reused in the calendar year. Definitions of independent variables are the same as in Table 5. Note that GEO staff only curate expression profiling data by array, so the dependent variable Series type cannot be estimated.

Figure 1: Microarray Examples

Note: Figure from Atul Butte's presentation slides at the Big Data in BioMedicine Conference.

Accessed at http://www.slideshare.net/atulbutte/2013-05-atul-butte-big-data.

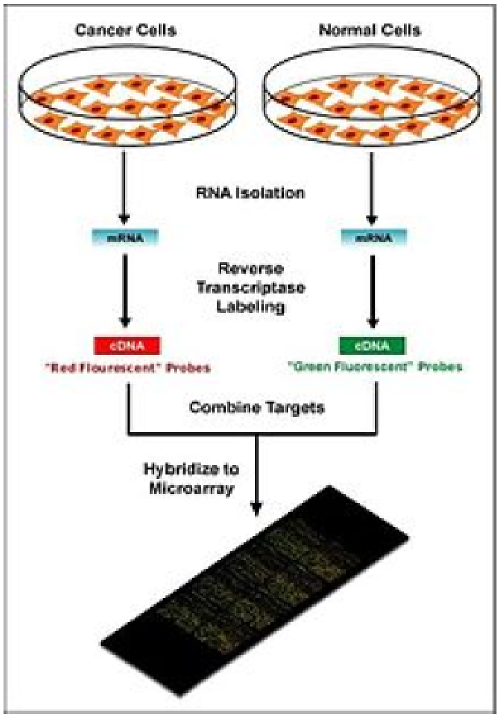

Figure 2: Illustration of a Microarray Experiment
Note: Figure from Wikipedia: DNA Microarray.
Accessed at http://en.wikipedia.org/wiki/DNA_microarray.

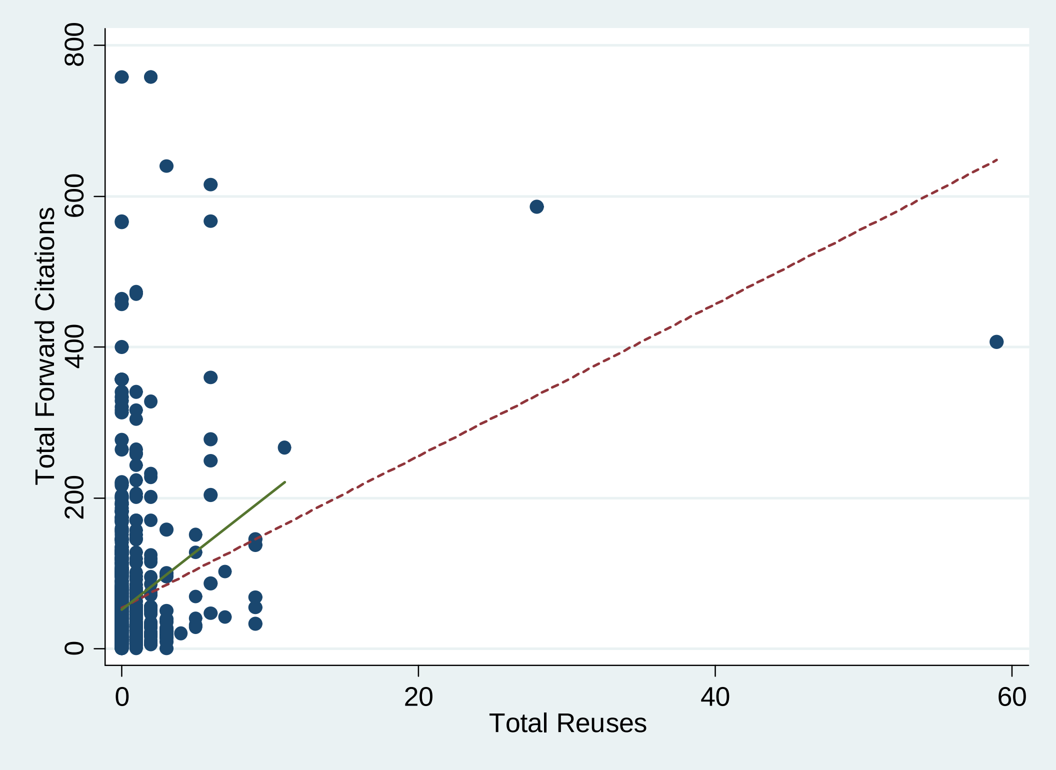

Figure 3: Documenting the Link between Article Citation and Data Reuse

Note: This figure provides the descriptive statistics that document the link between article citations and data reuse in the sample. The x-axis is the number of times a data set is reused, and the y-axis is the number of total citations received by the paper to which that data is linked. Each dot represents a data set. The dashed line represents the linear fitted value for the full sample, and the solid line represents the linear fitted value excluding the two data sets that have been reused more than 20 times.

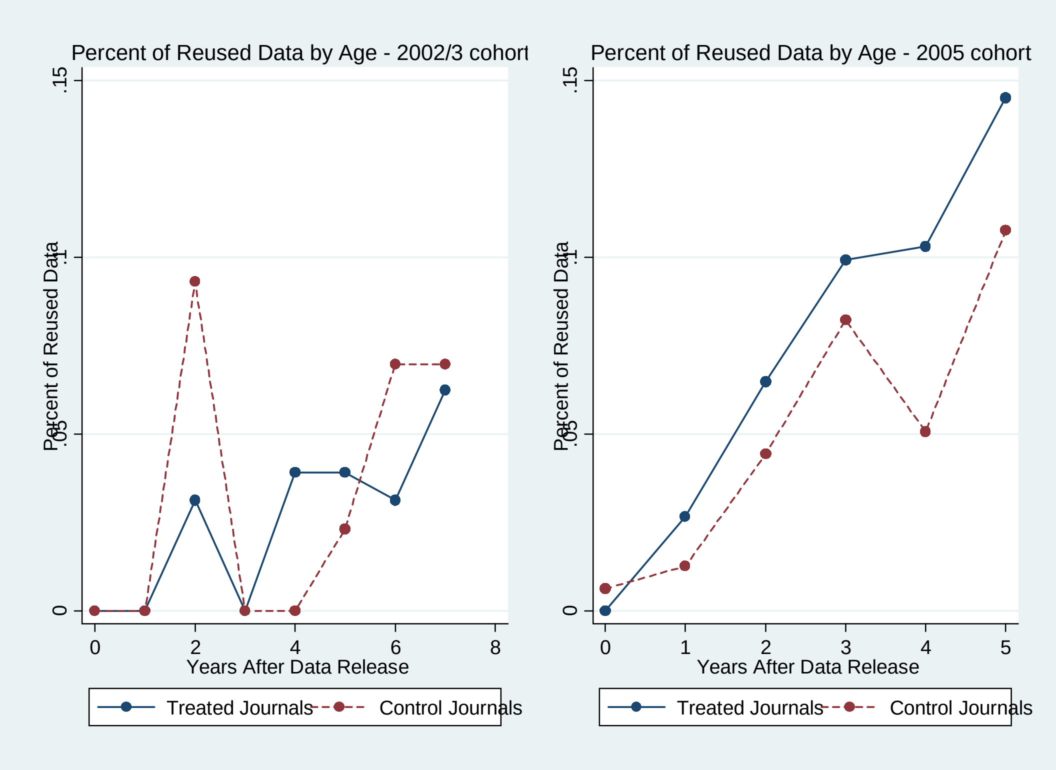

Figure 4: Impact of MIAME on Data Reuse

Note: This figure plots the descriptive statistics described in Section 5.1.

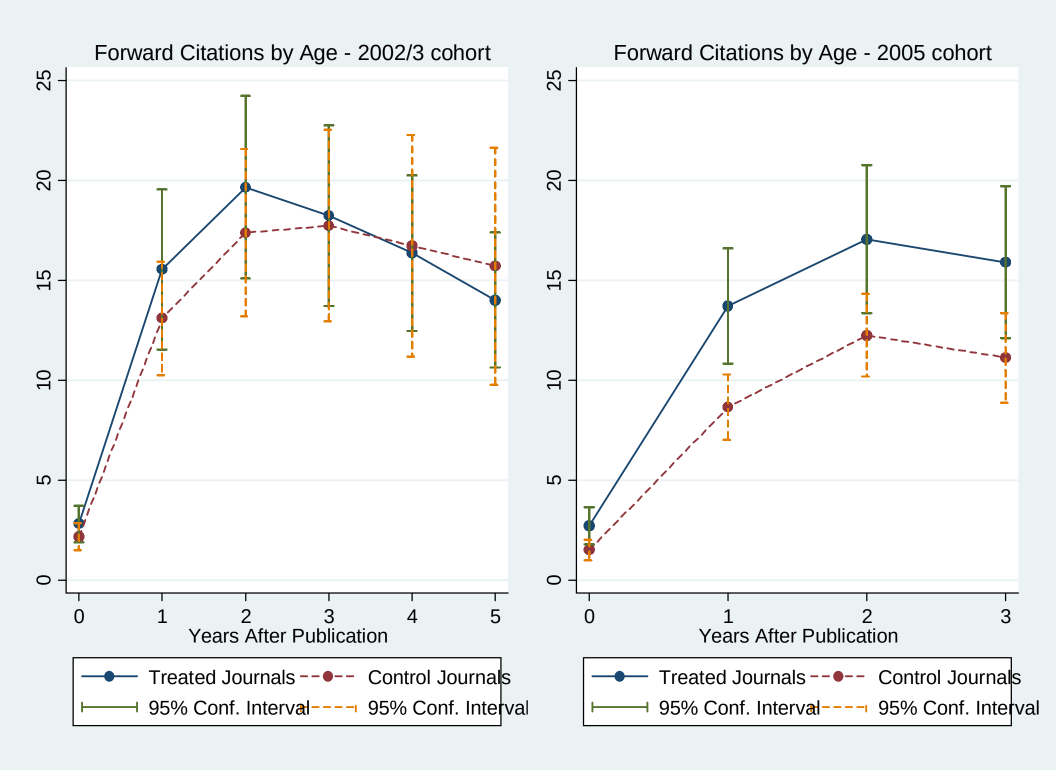

Figure 5: Impact of MIAME on Annual Citation (with 95% confidence intervals)
Note: This figure plots the descriptive statistics described in Section 5.3.