American Economic Association
Committee on Economic Statistics
https://www.aeaweb.org/about-aea/committees/economic-statistics

Successful Private Firm-Academic Researcher Arrangements for Access to and Replicable Use of Private Data

March 2020

Context

The capacity for data collection through government surveys is finite and, with decreasing response rates, declining. Data collected by private companies in the course of carrying out their business comprise an alternative source of data for research. Private data reveal economic behavior, record transactions, and generate knowledge. Use of private data for economic research within, between, and outside the companies that collect or create them can further the public good. There are a number of models for private data collaboration[i], each with different opportunities and challenges.

This report focuses on research and analysis partnerships primarily with technology and information firms, in which tech companies engage directly with academic or public-sector partners, sharing proprietary data to generate new understanding of economic phenomena. It is informed by a January 3, 2020 "Working Lunch" sponsored by the American Economic Association's Committee on Economic Statistics, in which representatives of tech and information companies, academic and federal government researchers, data intermediaries and their supporters, and economic journal editors sought "Successful Private Firm-Academic Researcher Agreements for Access to and Replicable Use of Private Data." Data intermediaries included representatives from those providing access to health expenditures data and private capital transactions data. Participants (listed in appendix) identified common challenges faced in attempts to create private-public research and analysis partnerships, and potential solutions to some identified problems.

Challenges to Private Data Sharing Arrangements for Research

Several years ago, as part of a Future of Privacy Forum project, research was conducted and interviews held with experts in the academic and industry communities to determine: The extent to which leading companies make data available to support published research that contributes to public knowledge; Why and how companies share data for academic research; and The risks companies perceive to be associated with such sharing, as well as their strategies for mitigating those risks[ii]. Participants in the AEA Working Lunch built and elaborated on challenges identified by that project.

1. **What's in it for the company?** It is not always clear to companies what the value proposition of sharing data with outside researchers is to the company. Both partners must want to share and see benefits from sharing. Some companies may perceive risks that research based on their data could put them in a bad light. Then again, some companies may value the public relations benefit when, for example, research papers get on the front page of the New York Times. But PR can be a double-edged sword; may bring unwanted attention; risk of negative reaction may outweigh benefit. On the other hand, the public benefit arising from collaborations can be a selling point. Also, some companies inherently value scholarship.

   Potential actions toward resolution:
   - Conduct research to determine the strength of alternative companies' motivations for/benefits from data sharing
   - Appeal to companies' public service ethic
   - Recognize that top management (CEO) support will be necessary. Target education to CEO's.
   - Offer to enhance the value of companies' data by cleaning, editing, organizing data
   - Develop social norms around sharing
   - Collaborate with companies' researchers to assure that the research addresses company concerns as well as academic excellence. Co-author published research with academic and company researchers.
   - Assure that benefits to companies can compensate for the risks they take to provide access
   - Pool resources in industry-wide data repositories (Private Capital Research Institute or Health Care Cost Institute, for example) where research is not identifiably associated with any one company.

2. **Misperceptions by researchers about appropriateness of private data sets for research.** Private data are not always a good fit for the research question.

   Potential actions toward resolution:
   - Better integration with researchers inside the company who understand both the data and research question and can head off inappropriate uses before researcher investment of time/money.
   - Collaboration of researchers inside and outside the company is important.
   - Imbed academic students/researchers in the company

3. **Companies' legal or administrative restrictions have associated high transactions costs.** Internal resources are needed to get access through legal departments. Standardization is difficult because each researcher wants different data with different access criteria. Standardization with respect to nondisclosure is especially difficult. Questions about which party is responsible for disclosure control. Protecting firms' data from unsanctioned groups is a paramount reason for legal protection. Risks to companies increase the more granular the data. Companies also have to consider whether they might get sued for collaboration. In the meantime, the time required to negotiate private data access can significantly delay students' completion of theses, dissertations.

   This is not universally true. There are companies that see benefits in data sharing and research publication and lower legal barriers. For example, in at least one instance of collaborative work with academic researchers, EBay and its collaborators created a data set of bargaining transactions which, after scrubbing, was made available on the NBER website. Alternatively, some companies (e.g., Google) have created public use data files using sophisticated synthetic data methods.

   Potential action toward resolution: Standardization of agreements would remove incentives for legal invention of bespoke arrangements and legal reinvention of common arrangements. But heterogeneity among private data companies makes standardization a tall order.

   Another approach is the employment of trusted intermediaries (such as the Kilts Center for Marketing, the Health Care Cost Institute, and Private Capital Research Institute) that, after conducting bargaining with firms, then provide access under specific terms to data users.

4. **Universities' Contracts and Grants and Intellectual Property offices are also often substantial barriers to companies willing to share data; substantial bottlenecks in sharing agreement.** For some, negotiating IP rights is the most difficult challenge to successfully reaching agreement with universities on data sharing. Bespoke agreements increase the cost of data provision. And in standard data agreements legal requirements also get "fine tuned" each time a new lawyer examines risks and returns. Further, in many universities, only lawyers can sign data sharing agreements; not faculty.

Potential action toward resolution: Standardization of agreements would remove incentives for legal invention of bespoke arrangements and legal reinvention of common arrangements.

5. **There are enormously high transactions costs in terms of dollars and cents as well as time and effort.** Establishing a secure data enclave, for example, is very costly. Moving large, confidential private data, if even physically possible, is costly and puts companies (and sometimes researchers) at risk of disclosure.

   Potential actions toward resolution:
   - Create a data sharing environment in the Cloud, and give researchers notebooks.
   - Populate Administrative Data Research Facilities with private data so that they can act as secure intermediaries
   - Utilize existing data enclaves like NORC's.

6. **Commercial data inconsistencies and lack of archiving hamper possibilities for research replication.** One participant relates experiences studying venture capital and private equity, which are almost impossible to study through public statistics alone as they are not tracked in any systematic way. As a result, commercial databases exist but are not consistent across providers. Access is a deal-by-deal thing, and replicability is an issue. These databases are dynamic and updated frequently, and thus the results look different when you come back.

   Efforts toward resolution include:

   - Partnerships with private information compilation firms, such as Burgiss, can make such data available in a "cleaned up" manner.
   - Harvard's Private Capital Research Institute seeks to obtain licenses to a number of commercial sources and link to other sources, including Census data. Researchers there are also trying to link to certificates of incorporation, but this is difficult to access despite their being public documents.

7. **Many agreements stipulate that private data cannot be made public, which conflicts with increasingly transparent requirements of journals for reproducibility.** This stipulation can prevent publication of research findings in journals that strictly require that data and code be made openly available or available upon challenge so that published research can be replicated or validated. Or, when research

is published in journals without such stipulations, the reproducibility or replicability of the research may be questioned.

A partial resolution: The AEA Data Policy encourages, but does not absolutely require open access to private or proprietary data. However, it stipulates that to the extent possible, the *characteristics* of the data - its metadata, its appropriateness for the research undertaken, the conditions for using the data and the process leading to its access should be transparent to other researchers.

Other actions toward resolution:

- Synthetic data is a possible solution. Conduct the research with actual and synthetic data, but make only the synthetic data available.

A General Conclusion and Specific Follow Up Possibilities

No One Size Fits All: An undeniable conclusion from discussions at the AEA Committee on Economic Statistics Working Session is that there is great heterogeneity among technology and information companies with respect to if, how, and under what conditions they can collaborate with public sector or academic researchers. There is variation among companies' willingness to share, and the rules by which they share. For example, while Google gives access to specific data to all, or none, other firms have tiered access according to researchers' experience and relationships with that company or companies like them.

Some Promising Approaches:

a. **Typologize and create associated, standard templates by typology:** Identify important characteristics of firms that make them more or less willing and able to share data with researchers. Such characteristics might include: Higher vs lower transaction costs; Those with and without researchers in house; Degree of experience with a pattern of external engagement; Whether the company already sells data (suggesting that its data are likelier to be documented and consistent); Whether or not they already have public use data sets, etc.

   With the active participation of both tech/information firms and researchers, a matrix of important characteristics might be developed and used to identify 4-6 relatively distinct typologies. This, in turn, facilitates the development of 4-6 associated standardized collaboration agreement templates. Such standardization, if possible, would substantially reduce transaction costs.

b. **Create a Community of Practice and accelerate information transfer among members of the community:** The idea of a Wiki, in which successes and failures of various agreements are documented and discussed in a monitored environment, would be a first

step in creating a Community of Practice (COP). Beyond that, we can imagine the development and distribution of blogs on relevant news, and new models of collaboration, API's, or technologies, creating, in essence, an online community of interested parties. Sessions at professional meetings could be proposed to accelerate adhesion among community members in the private and academic settings. The more knowledge and experience shared, the higher the probability that norms begin to shift toward a model of mutual respect and understanding for productive sharing relationships.

c. **Celebrate successes by giving an esteemed prize to encourage sharing.** The newly announced Future of Privacy Forum Award for Research Data Stewardship recognizes a research partnership between a company that has shared data with an academic institution in a privacy protective manner, thereby driving the use of privately held data for academic research (https://fpf.org/award-for-research-data-stewardship/). A tangible and well-advertised prize from the AEA and/or another professional association could bolster this somewhat more obscure prize, and clearly convey to economic researchers professional recognition of the value of successful collaboration between private companies and academic researchers.

d. **Conduct a visioning exercise to stipulate the infrastructural characteristics and types and roles of people within that infrastructure that allow successes to "scale up**." In the words of one public sector participant, the success stories we heard at the AEA Working Session are successes that don't necessarily scale, and many have high transaction costs. At some point someone has to say what is the end point, what do we want to achieve? How do you get access to useful data? What does that infrastructure look like? Architecturally what does that look like, what are the roles of the people that participate, and who are they? Nickel and diming in the short term is fine but we need a vision.

   Perhaps we need a "five-year plan."

---

[i] Verhulst, Stefaan G., Andrew Young, Michelle Winowatan, and Andrew J. Zahuranec. *LEVERAGING PRIVATE DATA FOR PUBLIC GOOD A Descriptive Analysis and Typology of Existing Practices.* GovLab, October 2019. https://datacollaboratives.org/static/files/existing-practices-report.pdf

[ii] *UNDERSTANDING CORPORATE DATA SHARING DECISIONS: PRACTICES, CHALLENGES, AND OPPORTUNITIES FOR SHARING CORPORATE DATA WITH RESEARCHERS.* Future of Privacy Forum, November 2017. https://fpf.org/wp-content/uploads/2017/11/FPF_Data_Sharing_Report_FINAL.pdf

Acknowledgments: Thanks to Kitty Evans, Marty Gaynor, and Josh Lerner for organizing the event, Cathy Buffington for invaluable notes, Kitty Evans for drafting this report, John Haltiwanger and Lars Vilhuber for review and comments on the draft, and to all of the Working Lunch participants for their openness in sharing observations and ideas.