Data Sharing for Research Case Study: IBM

Executive Summary

The Future of Privacy Forum (FPF) analyzed a diverse sample of data-sharing partnerships between companies and academic researchers and produced a series of case studies distilling our findings. We learned that there is broad consensus regarding the potential benefits of industry/academic data-sharing partnerships, including the acceleration of socially beneficial research, enhanced reproducibility of research breakthroughs, and broader access to valuable data sets. At the same time, companies and academic researchers understand and take steps to mitigate risks - particularly ethical and data protection risks. Increasingly, stakeholders are identifying risks arising from re-identification threats or data breaches while acting to mitigate those risks through the use of Data Sharing Agreements (DSAs) and Privacy Enhancing Technologies (PETs).

FPF's analysis of corporate-academic data-sharing partnerships provides practical, evidence-based recommendations for companies and researchers who want to share data in an ethical, privacy-protective way. These case studies demonstrate that corporate-academic data-sharing partnerships offer compelling benefits to companies, research, and society. Risks exist, but effective mitigation strategies can reduce the likelihood of harm to individuals, communities, and society. For many organizations, data-sharing partnerships are transitioning from being considered an experimental business activity to an expected business competency. This trend is most pronounced among established firms; it is an opportunity for researchers to access new data for scientific discovery.

Data Sharing Type

Internal, Closed Trusted Partnerships, Open Data

Organization and Partners

Company

With over 288,000 employees, International Business Machines Corporation (IBM) is a multinational technology company that provides a variety of computing and communications technologies and services for businesses. IBM is the largest industrial research organization in the world, with 19 research facilities globally. In 2022, IBM reported an annual revenue of $60.5 billion.[1]

Researchers

IBM hosts an in-house research organization composed of data scientists and researchers who process data to train models and improve services, among other projects. Externally, and in accordance with applicable data and privacy laws and IBM policies and practices, IBM researchers may share data for the same or similar purposes with universities, non-profits, and research labs around the world. IBM researchers generally limit data sharing with third parties to non-sensitive data and metadata for research purposes and as part of data-sharing partnerships. For example, working with the Australian utility Melbourne Water, IBM collected and processed data to develop insights that will help cut energy emissions. In limited instances when IBM shares data it has collected that includes personal information, IBM uses Privacy Enhancing Technologies (PETs). During the beginning of the global COVID-19 crisis, IBM collaborated with researchers and scientists to process SARS-CoV-2 genomic sequences, resulting in more than three million sequences, which were made available in a repository for researchers working to identify molecular targets for drug design, test

[1] International Business Machines. '2022 IBM Annual Report.' 2022. https://www.ibm.com/annualreport/

development, and treatment. IBM is also working on big data machine learning projects using de-identified medical data (i.e., with personal identifiers removed) to advance scientific discoveries on disease progression, including diabetic kidney disease.

Partnership Considerations

Data

Representatives stated that, in limited scenarios, IBM may share a variety of non-sensitive data and metadata externally, depending on the purpose and nature of the research request. Data is shared only for the original purpose for which it was acquired, which can be found in each data acquisition's procurement statement. If IBM seeks to share data that includes personal information, before sharing, they use PETs to remove personal identifiers or render the datasets into a form that no longer constitutes personal data. IBM prioritizes PETs such as federated learning and differential privacy and has made libraries and toolkits publicly available.

Data Sharing

IBM faces ongoing demand for data sharing from inside the company, notably people in IBM Research, which supports a network of international research facilities and about 3,000 researchers. According to IBM representatives, when a researcher seeks access to third-party data, the researcher develops a proposal that is analyzed by procurement or contract professionals and counsel, who might request modifications to the governing terms and would be involved in any negotiations. Counsel, with support from IBM Privacy Office professionals and automated internal processes, analyzes the proposed uses to determine that they are consistent with the purposes for which the data was acquired, ensures compliance with IBM requirements and privacy implications, and addresses other sensitivities associated with requested data. That might include a check on the appropriateness of or permissions associated with data collected by IBM. These processes are designed to address issues of data privacy, security, and quality.

Company representatives reported that IBM researchers sometimes provide data externally through contributions to data-sharing communities and sometimes directly to third-party partners in connection with an initiative, with a preference for open terms, in situations where non-sensitive and quality data and metadata is being shared that do not favor particular users or uses. For example, IBM worked with UK Research and Innovation - the UK government agency that directs research and innovation funding - to make available under open terms certain wave-elevation data. IBM favors the Community Data License Agreement (CDLA) permissive licenses for sharing open data. Unlike other open-sharing mechanisms, the CDLA permissive license is adapted to data sharing. For open data, IBM employs a review process to ensure that no sensitive data, such as personally identifiable or health-related data, is shared and that none of the data would be subject to privacy regulations. All IBM data sharing is based on a jurisdictional approach that accounts for differences in location and legal regimes – as a multinational company, IBM is accustomed to tuning its compliance based on jurisdictional requirements.

Privacy and Ethics

Representatives stated that IBM prioritizes transparency, data stewardship, privacy, security, and ethics, implemented through a technology ethics and privacy-by-design review process. Its AI Ethics Board enables IBM to take a centralized and multi-disciplinary approach to technology ethics. Their review considers the entirety of the use case, including the uses of data, such as training AI models where the risk of bias is a known concern, and it identifies methods for mitigating harm. In the limited instances where IBM researchers share personal data with third parties, the quantities are small, and agreements specify purposes, required consents, protections for privacy and security, and other terms. Throughout the formal review process, data-sharing arrangements are subject to data constraints. A primary data constraint is the use of PETs (such as masking, encryption, or anonymization tools), which depend on the data's type, size, intended use, and source. Additionally, all arrangements are subject to IBM's security provisions.

Costs

IBM has both ongoing and fixed costs in privacy and security related to data sharing. These include staff time and internal tool development.

Next Steps

Representatives noted that because IBM is dedicated to continuing its data-sharing arrangements, it is investing in related data protection, privacy, and security. IBM has developed a tool to systematize the review of third-party data to identify, among other attributes, personal information, and is specifically dedicated to increasing data sharing in the environmental, sustainability, and artificial intelligence contexts. IBM representatives stated they want to see more data sharing, particularly in the open data space. Former IBM executives were also involved in establishing an industry group, the Data and Trust Alliance, dedicated to improving data stewardship and potentially creating a vehicle for fostering practices supporting data sharing for research.

Risks and Benefits

According to IBM representatives, data misuse can cause significant harm. IBM continuously monitors ongoing lawsuits related to data and data sharing and reassesses risks as the landscape changes. Despite those risks, the benefits of data sharing to the company and the general public justify the practice, which IBM is proud of. IBM credits data sharing as having improved many company products and services. Executives believe corporations, governments, and citizens will profit from data sharing and open data. In the AI sector, for example, they contend that increased data sharing could help make AI models more equitable.

Partnership Information

IBM: https://www.ibm.com/
IBM Research: https://research.ibm.com/
IBM Developer Datasets: https://developer.ibm.com/exchanges/data/all/

Data and Trust Alliance: https://dataandtrustalliance.org/

To learn more about data-sharing partnerships, read The Playbook: Data Sharing for Research or join the Ethics and Data in Research Working Group for updates on legislative developments and monthly calls with experts. This project is supported by the Alfred P. Sloan Foundation, a not-for-profit grantmaking institution whose mission is to enhance the welfare of all through the advancement of scientific knowledge.