DAY ONE PROJECT

Promoting Entrepreneurship and Innovation Through Business-to-Business (B2B) Data Sharing

Alex Leader

December 2020

The Day One Project offers a platform for ideas that represent a broad range of perspectives across S&T disciplines. The views and opinions expressed in this proposal are those of the author(s) and do not reflect the views and opinions of the Day One Project or its S&T Leadership Council.

Summary

To bolster competition, entrepreneurship, and innovation, the next administration should facilitate business-to-business (B2B) data sharing between startups and data-rich, established companies. Asymmetry in the digital economy is an existing market failure that, if left unchecked will continue to intensify to the detriment of consumer choice and our collective security.

Leveling the playing field requires policy to remove barriers to entry created by data advantages and to promote market competition through increased access to big data. Specifically, we propose that the Small Business Administration's Office of Investment and Innovation establish a data-sharing program that gives entrepreneurs access to the data they need to improve algorithms underpinning their products and services. This would support a thriving and diverse ecosystem of startups that could in time yield valuable new markets and products.

Challenge and Opportunity

The rate of new-business formation in the United States has declined sharply since the early 2000s,[1] as have the total number and amounts of seed-funding deals for technology startups.[2] The frequency of young firms entering the digital economy and their respective market shares have also fallen over the same time frame.[3] Many early investors now avoid backing startups in data-driven enterprises like e-commerce, mobile, online search, and social media. In 2017, there were nearly 23% fewer first financing rounds for startups in data-driven areas than there were in 2012.[4] These trends carry far-reaching economic consequences. A 2014 study published by the Kauffman Foundation found that job destruction in the high-tech sector has been outpacing job creation since 2005, with no sign of imminent reversal.[5]

Declining entrepreneurship in the tech sector is due in part to an anticompetitive environment created by the domination of a handful of major online platforms. The market power of a small number of players has staved off would-be competitors by creating insurmountable barriers to entry. Access to data lies at the root of this power imbalance.

[1] Akcigit, U.; Ates, S.T. (2019). Knowledge in the hands of the best, not the rest: The decline of US business dynamism. VoxEU, July 4. https://voxeu.org/article/decline-us-business-dynamism.

[2] Teare, G. (2019). Decade in review: Trends in seed- and early-stage funding. TechCrunch, March 16. https://techcrunch.com/2019/03/16/decade-in-review-trends-in-seed-and-early-stage-funding/.

[3] Hathaway, I. (2013). Tech Starts: High-Technology Business Formation and Job Creation in the United States. Ewing Marion Kauffman Foundation. https://www.kauffman.org/wp-content/uploads/2019/12/bdstechstartsreport.pdf

[4] The Economist (2018). American tech giants are making life tough for startups. June 2. https://www.economist.com/business/2018/06/02/american-tech-giants-are-making-life-tough-for-startups.

[5] Haltiwanger, J.; Hathaway, I.; Miranda, J. (2014). Declining Business Dynamism in the U.S. High-Technology Sector. Ewing Marion Kauffman Foundation. https://www.kauffman.org/wp-content/uploads/2019/12/declining_business_dynamism_in_us_high_tech_sector.pdf.

As data-driven machine learning increasingly drives innovation, companies with the largest data pipelines emerge as the most cutting-edge and secure market dominance. A company with a large network of users has access to enormous volumes of behavioral data ("big data") that the company can use to train and refine algorithms. Algorithmic improvement yields better products and services, attracting new users to expand the company's network even further. Existing, data-rich internet and technology platforms are hence rewarded with continued growth while smaller competitors are prevented from gaining a market foothold. This in turn undermines competition, narrows choices available to consumers, and disincentives innovation. These feedback loops are currently primed to self-perpetuate, exacerbating and entrenching the disparity separating a few industry leaders from aspiring companies.

The next administration can intervene by introducing business-to-business (B2B) data sharing, giving startups access to the informational resources they need to compete. Big data is essential for artificial intelligence and machine learning (AI/ML). A federal initiative to expand big-data access would therefore help equitably proliferate the benefits of AI/ML throughout society. What's more, evidence suggests that the introduction of high-quality training datasets (e.g., ImageNet, WordNet, and the MNIST database) have contributed to the most consequential breakthroughs in AI/ML. In the words of Alexander Wissman-Gross, a fellow at Harvard University's Institute for Applied Computational Science, "Perhaps the most important news of our day is that datasets—not algorithms—might be the key limiting factor to development of human-level artificial intelligence."[6] Government-led data sharing could do more than just kickstart entrepreneurship and competition: it could very well foster quantum leaps in state-of-the-art AI/ML.

Plan of Action

The next administration should create a public-private partnership (PPP) for big-data sharing. This PPP—which we call DataShare—should be designed to stimulate competition and innovation in the technology industry by making AI/ML, specifically deep learning, more feasible for startups. DataShare would facilitate temporary data "grants" from companies with ample user networks ("grantors") to data-driven startups ("grantees"). Grantees would be able to choose training data that are integral to their proposed AI/ML models, and then use the allotted data to train AI/ML algorithms in a precompetitive, collaborative ecosystem. These data grants would allow their recipients to realize the benefits of deep learning: e.g., unlocking new business opportunities such as personalized recommendations and predictive analytics.

In short, DataShare would give startups access to an informational resource that is currently only available to digital platforms with massive user networks, promoting competition and encouraging entrepreneurship. By relying on existing datasets, DataShare would enable grantees to avoid high-cost and labor-intensive aspects of new-data processing, such as manual

[6] Wissman-Gross, A. (2016). Datasets Over Algorithms. Edge. https://www.edge.org/response-detail/26587.

labeling. DataShare would also save grantees from the burden of developing in-house infrastructure to store and manage large volumes of data.

A selection committee could determine whether a prospective grantee should be eligible to participate in DataShare, focusing on the following criteria:

- Demonstrated need, particularly a prospective grantee's inability to compete without certain data.
- Feasibility of a prospective grantee's algorithmic products and its overarching business model.
- Availability of data that meets the prospective grantee's requested specifications, such as volume and variety.

If a prospective grantee is deemed eligible for DataShare it would then receive access to appropriate data for a specified length of time: a period long enough for the grantee to train its models until a suitable level of accuracy is achieved during testing.

DataShare would constitute a public-private partnership where grantors retain ownership over the data assets but delegate operational responsibility to the government. Precise terms of the data-sharing agreements will need to be determined by the next administration. Options include having the government pay companies directly for data access, offering fiscal or other incentives in exchange for access, issuing a legal mandate for access to certain data, or some combination thereof.

The Small Business Administration (SBA), specifically the SBA Office of Investment and Innovation (OII), is well-positioned to administer DataShare, as the agency's mission and core functions align with the partnership's proposed scope. Among the SBA's many relevant functions are (1) working with lenders to provide and set guidelines for small-business loans, and (2) assisting small businesses to find resources that suit their needs.[7] DataShare would resemble some of the SBA's existing efforts supporting entrepreneurs looking to start high-tech businesses. Through the Small Business Innovation Research Program (SBIR), for instance, nearly 5,000 small businesses already receive more than $2.5 billion total in federal grants intended to help entrepreneurs conduct research and develop high-tech products. Colloquially known as "America's Seed Fund", the SBIR offers "the largest source of non-dilutive, early-stage seed capital in the world."[8]

While the SBA's existing initiatives for high-tech small businesses provide monetary resources and guidance, more needs to be done to foster greater competition in tech. Simply injecting more capital does not give startups access to the large user networks that yield behavioral data.

[7] U.S. Small Business Administration (n.d.). About SBA. https://www.sba.gov/about-sba.

[8] Beesley, C. (2017). Starting a High-Tech Business? You May be Eligible for Government Funding. U.S. Small Business Administration, June 5. https://www.sba.gov/blog/starting-high-tech-business-you-may-be-eligible-government-funding.

And when startup competitors cannot access such data, dominant incumbents retain their quasi-monopsony[9] on user information.

There are many types of data sharing that DataShare could adopt. Examples provided below are neither mutually exclusive nor collectively exhaustive:

- **Data pooling.** DataShare could create a pool of anonymized data, made up of contributions from different grantors, while maintaining a fiduciary obligation to uphold grantor trade secrets.
- **Application programming interfaces (APIs).** Grantees could receive access to an API that enables them to directly call a grantor's dataset and train their algorithms accordingly.
- **Online directory.** Grantors could publish categorized data on a password-protected site, akin to a GitHub, where grantees could search for relevant datasets and pull those datasets for local use.
- **Data accounts.** Grantees could receive curated accounts populated with designated allotments of data. The category and volume of data would be based on demonstrated need.
- **Federated data and learning.** Participating grantors could continue to store their training data locally within their own infrastructure, while a grantee trains its model across each dataset remotely.

There is already a private-sector analogue for a government-sponsored DataShare program, albeit one for a different stage of the AI/ML workflow. The research lab OpenAI operates an open-source toolkit, known as Gym, for "developing and comparing reinforcement learning algorithms."[10] Gym's series of environments enable users to teach their algorithms, or "agents", a variety of deep-learning tasks, such as playing games. Gym users can access this toolkit to actually implement their algorithm in a novel environment: just as a grantee participating in DataShare would be able to access a dataset to refine their algorithm's parameters.

DataShare could also provide additional services to reinforce its mission and the overall goals of the SBA's offerings. Examples of such services include:

- **Grantor matching.** DataShare could work with prospective grantees to determine which grantor possesses data best suited to improve the prospective grantees' services and product offerings, similar to how the SBA helps small-business applicants find suitable lenders.
- **Technical assistance.** DataShare could assist grantees in navigating the full machine-learning workflow and identify areas for support before committing to a specific data grant on a project.

[9] A monopsony refers to a market condition where a single buyer purchases goods and services offered by many sellers. In this case, a single company offers a free platform to effectively "purchase" behavioral data from billions of users.

[10] OpenAI (n.d.). Gym: Documentation. https://gym.openai.com/docs/.

- **Challenge awards.** Recognizing the monetary value of grantors' data, DataShare could facilitate merit-based challenges in which prospective grantees must demonstrate a return on investment in order to qualify for a data grant.

While there are few market incentives for data-rich firms to participate as grantors in the partnership, the next administration could pursue several policy approaches to secure grantor participation. These approaches are summarized below:

- **Grantor compensation.** The next administration could establish mutually beneficial terms for data grants that provide grantors with some level of equity or remuneration. In essence, DataShare could provide opportunities for joint ventures or corporate venture capital, particularly where grantees' product offerings complement those of grantors.
- **Fiscal incentives.** The next administration could work with Congress to develop subsidies or tax breaks for companies participating as grantors. Big data holds significant monetary value and sharing it could constitute a positive, quantifiable contribution to the economy. Conversely, the next administration and Congress could explore a Pigouvian tax on data that market leaders refuse to share.[11]
- **Data-sharing mandates.** The next administration could call on Congress to pass legislation that mandates B2B data sharing through a public-sector initiative like DataShare. Two distinct resolutions could underpin this mandate:
    - Classifying data as a public good. Congress could assert that individual companies do not retain ownership over the data they collect from their consumers. While this assertion alone will not affirm whether an individual owns their personal data, it would provide the legal basis for DataShare to require data-rich companies to participate as grantors.
    - Setting thresholds for anti-competitive market share. Congress could determine that data-rich firms with a predefined market share are benefiting from anticompetitive practices. The amount of data a company must make available would depend on the market share it has captured.
- **Antitrust law.** The next administration could unilaterally pursue compulsory data sharing by making data sharing a remedy available to the Federal Trade Commission (FTC) and the Department of Justice (DOJ) in enforcement actions. In antitrust investigations and lawsuits, a defendant could opt to settle by effectively reducing its anticompetitive power through participation in DataShare.

Considering the immense resources of dominant tech firms, tax incentives and financial compensation alone are unlikely to motivate grantor participation in DataShare. The next administration should therefore be prepared to couple these "softer" strategies with compulsory data sharing through antitrust enforcement. A data-sharing mandate from Congress would

[11] A Pigouvian tax refers to a tax levied against a private entity for activities that generate negative externalities for the public. When data is controlled by a single company and not shared with competitors, the resulting anticompetitive conditions could be considered negative externalities.

complement this approach by expanding the administration's scope to firms that are not necessarily subject to antitrust investigations.

Now is an opportune moment for the next administration to leverage DOJ and FTC antitrust actions to secure participation in DataShare. After a 16-month investigation into the practices of the country's largest technology companies, the House Judiciary Committee published a report in October 2020 asserting that Amazon, Apple, Facebook, and Google had abused their monopolistic positions.[12] The report calls for restoring competition in the market, in part by bestowing greater power onto the government entities responsible for enforcing antitrust policy. Judiciary Committee Chairman Rep. Jerrold Nadler (D-NY) affirmed that the report establishes a "clear and compelling need for Congress and the antitrust enforcement agencies to take action that restores competition, improves innovation, and safeguards our democracy."[13] While the House report offers transformative policy recommendations for restoring competition in the digital economy (e.g., prohibiting future mergers and requiring interoperability), it misses an opportunity to propose government-mandated data sharing as a remedy.

Other recent events make the case for the next administration to pursue data sharing as a component of antitrust enforcement. Publication of the aforementioned report from the Judiciary Committee coincided with news that DOJ was expected to file an antitrust complaint against Google for its internet search dominance.[14] This development follows an ongoing FTC antitrust investigation into Amazon, Apple, and Facebook.[15] Impending legal action creates a pressing need for dominant tech platforms to take measurable actions that reduce their monopolistic market share, such as relinquishing exclusive privileges over their user data. The Judiciary Committee's report validates this approach by explicitly tying monopoly power to the "data advantages that dominant online platform companies have over smaller competitors and startups, and how those data advantages…reinforce dominance and serve as a barrier to entry."[16]

The origin of one dominant platform—Google—illustrates why investing in data accessibility, rather than merely suing for antitrust violations, could more effectively promote competition. In the 1990s, the FTC and DOJ repeatedly investigated and litigated Microsoft's business practices, which the Federal Government deemed monopolistic. When the U.S. Court of Appeals ruled that Microsoft was an unlawful monopoly, the company managed to settle, avoid a breakup, and maintain its standing in the tech industry.

[12] U.S. House of Representatives (2020). Investigation of Competition in Digital Markets. Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary. 116th Congress, 2nd Session. https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf.

[13] Kang, C.; McCabe, D. (2020). House Lawmakers Condemn Big Tech's 'Monopoly Power' and Urge Their Breakups. The New York Times, October 6. https://www.nytimes.com/2020/10/06/technology/congress-big-tech-monopoly-power.html.

[14] Kang, C.; Benner, K.; Lohr, S.; Wakabayashi, D. (2020). Justice Dept. Case Against Google Is Said to Focus on Search Dominance. The New York Times, September 22. https://www.nytimes.com/2020/09/22/technology/justice-dept-case-google-search-dominance.html.

[15] Kang, C.; McCabe, D. (2020). House Lawmakers Condemn Big Tech's 'Monopoly Power'.

[16] U.S. House of Representatives (2020). Investigation of Competition in Digital Markets. Page 32.

The lasting check on Microsoft's market power ultimately came from business competition, through the emergence of Google. The Federal Government played a key role in Google's founding. In 1994, the National Science Foundation (NSF) launched the Digital Library Initiative, a project to create interfaces for collecting data. The initiative funded two Stanford graduate students, Larry Page and Sergey Brin, to develop an algorithm for ranking web pages. The PageRank algorithm formed the bedrock of Google, one of the world's most successful tech firms and ultimately one of Microsoft's biggest competitors.[17]

We expect that DataShare would similarly facilitate essential market competition in the tech sector by providing the tools, infrastructures, architectures, and governance mechanisms for a thriving data-sharing ecosystem among companies with disparate market shares. This public-private partnership will support a thriving ecosystem of data-intensive companies in the United States and will accelerate digital transformation in the broader American economy. In the longer term, DataShare may also catalyze widespread deployment of data-sharing tools and platforms; establishment of data governance frameworks; and improvement in data quality, availability, and interoperability.

[17] Hart, D. (2004). On the Origins of Google. National Science Foundation, August 17. https://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=100660.

Frequently Asked Questions

What is the market failure in data-driven enterprise?

By amassing dedicated user networks for a growing number of services, the dominant internet platforms have effectively created an oligopsony for behavioral data.[18] Control over this key informational resource enables dominant platforms to lock in their market share and exclude prospective market entrants. Without access to large-scale behavioral data, would-be entrants lack a fundamental component of AI/ML needed to create competitive products. This is the market failure that DataShare seeks to address.

Is there a private-sector solution capable of addressing this market failure?

Some companies offering machine learning as a service (MLaaS) have already begun making deep learning more widely accessible. Customers upload data and a goal, then receive a trained algorithm from the MLaaS company. But MLaaS is only as good as the data provided. Many startups currently resort to synthetic data and/or data from foreign suppliers in order to get off the ground. These data are sub-par substitutes for genuine behavioral data from target users, which only large user networks can generate. Government support is needed to make this informational resource available to young companies.

Are public data resources currently available to startups?

Yes. There are many existing datasets that could be used to train and refine algorithms, a large number of which are freely available. The Federal Government also stores publicly available datasets on Data.gov. However, these data are best suited for research and analysis, such as providing an objective benchmark for state-of-the-art advances in deep learning. Public data resources provide limited opportunities for profitable AI/ML.

Are there public-private partnerships that have leveraged data sharing?

Yes. The Accelerating Medicines Partnership (AMP) presents an illustrative case study on how public-private partnerships like DataShare engender mutually beneficial innovation. The AMP seeks to kickstart the current model for developing therapeutics through a partnership that includes the National Institutes of Health (NIH), the Food and Drug Administration (FDA), multiple biopharmaceutical companies, and nonprofits. By cooperatively setting goals, the AMP aims to increase the number of new therapies and streamline their development. AMP participants have agreed to make the resulting data publicly accessible.[19]

[18] Similar to a monopsony, an oligopsony refers to a market condition where a small number of buyers purchase goods and services, which are offered by a comparatively large number of sellers.

[19] National Institutes of Health (n.d.). Overview. Accelerating Medicines Partnership. https://www.nih.gov/research-training/accelerating-medicines-partnership-amp.

Is there existing legal precedent for mandating data sharing?

The Thomson Reuters merger offers an example of the Federal Government enforcing B2B data sharing. In 2008, the DOJ approved the merger of financial data providers Thomson Corporation and Reuters Group, conditional on Thomson selling proprietary datasets and licensing intellectual property to competing firms. The DOJ argued that exclusive ownership of the datasets by a single company "likely would have led to higher prices and reduced innovation."[20]

What are some obstacles the federal government may encounter in launching DataShare?

One key issue is privacy protection. California and the European Union have already outlawed sharing of personal data with third parties without user consent. However, these regulations allow sharing of anonymized data. While anonymizing data can limit the utility (and value) of datasets, anonymized data can still be used to train myriad algorithms. For example, anonymized data can train algorithms to recommend products based on certain input features. By providing anonymized datasets, DataShare could respect privacy concerns while still promoting competition. This accommodation would also preserve the marketing reach that industry leaders have cultivated (since data grantors would retain access to the original, non-anonymized data) while still affording smaller competitors a fighting chance.

[20] Biancotti, C.; Ciocca, P. (2019). Opening Internet Monopolies to Competition with Data Sharing Mandates. Peterson Institute. Peterson Institute for International Economics. https://www.piie.com/system/files/documents/pb19-3.pdf.

About the Author

Alex Leader is a Senior Associate at Luminary Labs, where he helps organizations adapt to emerging technologies and the evolving future of work. He developed technical expertise as a data scientist at Amenity Analytics, a machine learning and AI startup. At Amenity, Alex designed natural language processing solutions for a diverse array of business challenges, using both frontier technology and statistical tools to identify best practices and advance client goals. Alex holds a B.A. in Public Policy from the University of Michigan and a master's in International Affairs from the University of California, San Diego.

.

About the Day One Project

The Day One Project is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans, and readying them for Day One of a future presidential term. For more about the Day One Project, visit dayoneproject.org.