***This is a working paper, last updated August 6, 2020. We are circulating this report for discussion. This is not a peer reviewed scientific paper and does not reflect the views of any organization. See www.datadividends.org for contact info and for the latest information about this initiative.***

## Executive Summary:

In his 2019 State of the State speech, California Governor Gavin Newsom proposed a "data dividend" to share the wealth generated by personal data with the users who generate it. The motivation behind the proposal is powerful and important: California's data-driven economy does not exist without California's data-generating public, and the public should receive their fair share of the benefits from this economy.

Governor Newsom's proposal raises many complex questions: such as: How should data be valued, and how should that value be distributed? These questions have led to an understandable debate and have raised doubts about what can be implemented in the near term.[1] In this report, we propose a pragmatic approach to implement a "data dividend" that is motivated by the notion that data is "our data": data is a collective good and any data dividend must account for that fact.

The governor's proposal spurred us to form the California Data Dividends Working Group - an ad-hoc team of scholars and practitioners without any political affiliation. We have drafted **a data dividends plan that can be implemented *right now and can have effects on alleviating data-driven inequality in the near future.***. Our proposal is rooted in existing practices and does not require any leaps in technological capability. It quickly creates meaningful benefits to a wide group of Californians and, at the same time, lays the foundation for longer-term reforms. The plan is also adaptable: technological advances, rapidly-evolving research into the data-driven economy, and developments in other jurisdictions demand that data dividends be implemented using a flexible approach. To achieve this adaptability, we emphasize the establishment of flexible institutions rather than specific parameters.

The critical insight motivating our plan is that the economic value of data primarily comes from the aggregation of data generated by large groups, rather than from any one individual. The rapidly emerging new generation of Artificial Intelligence (AI) based technologies do not focus on analysing one person's individual data. AI "learns" from aggregated data sets and generates value by applying these insights back to individuals. Thus, to design a data dividend, we must think in terms of **"*our* data", not "*my* data."** The same aggregation effect is true for data brokers and other benefactors of the

[1] Jazmine Uolla, "Newsom Wants Companies Collecting Personal Data to Share the Wealth with Californians," *Los Angeles Times*, May 5, 2019, sec. Politics, https://www.latimes.com/politics/la-pol-ca-gavin-newsom-california-data-dividend-20190505-story.html; "Opinion: Newsom's California Data Dividend Idea Is a Dead End," *The Mercury News* (blog), March 7, 2019, https://www.mercurynews.com/2019/03/07/opinion-newsoms-california-data-dividend-is-a-dead-end/.

data-driven economy that have already created profitable markets in personal information. Thus, the best way to mitigate the harms of this economy - such as increased inequality, lack of access to opportunities, and the rise of powerful platform monopolies - is to treat data as a collective resource that must be managed through proper institutions rather than an individual asset.

Our plan applies this insight into data's function as a "commons" to establish four principles of an effective data dividend:

1. **Institute a Data Dividend Tax:** California should create a *Data Dividend Tax* on companies based on their "data dependence" – or the extent to which their business depends on the aggregation and storage of user data. To share the burden as fairly as possible, the tax would define data dependence using regulations already in the California Consumer Privacy Act (CCPA) and be implemented similarly to existing California corporate income tax. The Data Dividend tax would also include a tax on sales of personal data to third parties.

2. **Fund public goods:** The Data Dividends Tax should fund public goods that are well-known to provide broad benefits to a wide group of Californians. Ideas in our proposal include education investment, public wifi and computing infrastructure, and universal savings instruments.

3. **Create a Data Relations Board:** A *Data Relations Board* should be instituted and funded. The board should be modeled on existing state boards and will conduct and review studies on the data-driven economy. It will be tasked with making recommendations to the California government based on this research. The DRB will be a crucial source of expertise for policymakers and the public.

4. **Prepare for the future:** The Data Relations Board should lead explorations into more transformative initiatives. In our report, we suggest that the board should explore a *data industrial policy* that promotes collaboration between government and industry towards building infrastructure from which all Californians can benefit. These initiatives also include the promotion of pro-social, data-intensive businesses through public administration of socially important data sets and the facilitation of union-like entities for data contributors.

**The Data Dividend Tax is a pragmatic step to raise revenue that is well-motivated by the notion that "our data" is a collective good**. Because data's value derives from its collective properties, forcing individuals to sell their data in a marketplace without a major change in the *a priori* bargaining power of each side of the transaction stands to exacerbate the inequalities existing in society. We do not want to see a situation where the poor and socially vulnerable are forced to sell their privacy

to make ends meet. Even if we avoid market approaches and attempt to assign dividends based on some estimate of an individual "data value", we risk creating new inequalities.[2] A progressive data policy must make sure that all citizens can equally share in the fruits of the data they generate and have a stake in the benefits of the increased productivity that stems from innovation.

The Data Relations Board is a critical complement to the Data Dividend Tax that can consider more novel, larger-scale policies that could unfold over several years. These policies can foster new relationships between firms and users and between the firms themselves. For instance, we suggest that the board explore the establishment of *public data trusts*. Companies could contribute non-personal data to these trusts that the Board deems as serving the general public good and receive a credit against their Data Dividend Tax. Such a trust would level the playing field between established platforms and other participants in the data-driven economy, enabling more Californians to start data-driven businesses and create innovative public-private ecosystems that facilitate pro-social entrepreneurship. The Data Relations Board could give similar credit to firms that work with *data cooperatives*, union-like entities that act as fiduciaries for members' data. The board would be well-positioned to help establish any regulation around these cooperatives. We believe that such initiatives are crucial next steps to take full advantage of the possibilities latent in a data dividend to restructure the institutions responsible for inequality in the data economy.

Like Governor Newsom, the California Data Dividends Working Group believes that everybody who contributes to the data-driven economy should receive substantial benefits. Data-driven technologies like artificial intelligence rely just as much on data contributors as they do on software developers. In the same way that we say *'we* went to the moon', all Californians - and all Americans - can say *we* created the amazing AI technologies that are improving our lives every day. Our proposal helps to ensure that all Californians benefit more from these technologies, not starting years in the future, but starting right now.

[2] Nicholas Vincent et al., "Mapping the Potential and Pitfalls of 'Data Dividends' as a Means of Sharing the Profits of Artificial Intelligence," *ArXiv:1912.00757 [Cs]*, November 18, 2019, http://arxiv.org/abs/1912.00757.

## Introduction: What Is a Data Dividend and How Do We Get There?

In his 2019 State of the State speech, California Governor Newsom called for California to implement a data dividend to ensure that all Californians "share in the wealth that is created from their data." Representatives from other jurisdictions such as Colorado and Canada have also expressed interest in data dividends. Governor Newsom's initiative highlighted the idea that data has an intrinsic value that is calculable in monetary terms and is in some way "extracted" from users into the political mainstream.

Returning a portion of the value embedded in data to users has had some traction in academic discussions. However, there has not yet been a reckoning with how this value can be defined and measured, and what specific steps governments should take to make sure that more of this value returns to the citizens whose inputs have helped create it. This issue is particularly perplexing since the actual economic potential that exists in the data-driven economy is still unknown. While we have seen massive valuations for firms that experiment with using big data and machine learning to optimize commercial activities, we have not yet seen these firms convert high revenues into sustainable profits. We have not seen these advances transfer into large, widespread gains in productivity or growth.

We believe that the reason for the uncertainty about the future of the data-driven economy is linked to the same dynamics that have led the data economy, and the American economy as a whole, to be extremely unequal: the establishment of a uncompetitive and extractive market by several large, early entrants. As a group of specialists comprising the California Data Dividend Working Group (CDDWG), we argue that we must conceptualize a "data dividend" as broader set of structural interventions that provide a chance to "get ahead" of an economic transformation that is in the early stage of its development: a transformation that will challenge our understanding of concepts at the heart of a capitalist economy such as property rights,

labor compensation, and public goods. A comprehensive solution to the inequalities inherent in the data economy would not only compensate users, but also restructure the innovative ecosystem in a way that encourages an open, and productive economy. *The radical potential of a data dividend is that, if properly designed, it can be a series of measures that make sure that the gains created by new potential general-purpose technologies are, at their inception, shared broadly with the general public.*

This imperative flows from our theory of how value is generated in the data economy. AI technologies do not gain utility from analysing one person's data. In most cases, AI "learns" from aggregated data sets composed of contributions from many different people and generates value by identifying relationships in these data sets. Thus, to design a data dividend, we must think in terms of **"*our* data", not "*my* data."** The same aggregation effect is true for data brokers and other benefactors of the data-driven economy. As such, the best way to mitigate the harms of this economy - such as increased inequality, lack of access to opportunities, and the rise of powerful platform monopolies - is to treat data as a collective resource that must be managed through proper institutions rather than an individual asset.

This report will proceed in the following manner:

First, we address the challenges that a "data-driven" economy presents for how we value inputs into production by individuals. Data has several properties that make it a seemingly unique asset. Data carries a vast network effect, which means that for every additional data point collected, additional value is created for the overall data set. In other words, data has a positive marginal rate of return. This means that firms that exploit user data tend to be natural monopolies whose profits come from market power rather than value added. Moreover, data is extracted from users in economic transactions in which they are often unwitting counterparties. For this reason, we believe that we should understand the generation of data as a kind of "collective labor" that we perform as clusters of individuals. The return to this labor is frequently illegible to us and that opacity may limit our personal and economic liberty. Thus, it is reasonable for the government to intercede to recoup "unpaid wages" for this labor, and to proactively structure the kind of economy we would like our data to build.

Second, we highlight actions that the State of California can take in immediately capturing revenues from data-dependent firms through new taxes on the value generated from the collection and exploitation of personal data. We examine several new taxes that resemble but do not replicate similar efforts to tax "big tech" in Europe and are consistent with the principles already outlined by the State of California in the California Consumer Data Protection Act (CCPA).

Third, we point out ways in which these revenues can be returned to the citizens of California. We argue that the most effective means to invest the revenues collected on

behalf of Californian's digital labor is to not remunerate them individually but to follow the logic of data creation as *collective* labor and create funding streams for public goods.

Fourth, we recommend that the Government of the State of California create a Data Relations Board (DRB), which serves as a mechanism to bring together representatives from government organizations, industry, and citizens groups to oversee and guide the development of California's data economy. We recommend that this organization have a research staff that can develop and propose regulations necessary to meet the challenges of this rapidly developing new sector of the economy.

Finally, we explain why part of the DRB's purview should be to help organize a structural reform of the data economy by supporting the creation of new institutions and a "data industrial policy" aimed at building an infrastructure to stimulate broad access to entrepreneurship. Two immediate actions in this space that we recommend are the establishment of "public data trusts" and the enabling of "data cooperatives." Public data trusts are state-administered legal entities that organize government data and create incentives to responsibly contribute proprietary data into the public sphere. Data cooperatives are private entities with a fiduciary duty to their members to negotiate the terms of members' data collection with large platforms. These structural changes would help shift the underlying balance of power in the data economy.

## I. Data as An Asset in the Tech Economy

### A. The Data Dependent Business Model and Its Harms

The growth of Silicon Valley's "tech" ecosystem has clearly generated a vast amount of wealth. To understand how "data" might contribute to this stock of wealth, we must clarify the particular activities that "data-dependent" firms engage in. An important caveat is that, naturally, lines of business grow out of one another and are highly complementary. Silicon Valley began manufacturing microchips, first primarily for the military and then later for consumers and businesses. This line of business quickly expanded to the production of "software" programs that could run on that hardware. These business models evolved with the rapid commercialization of the internet and adaptation of home computers toward supporting and facilitating commerce. In these iterations, the actual "data" that firms collected was not considered valuable in and of itself.[3]

This began to change in the late-1990s with the advent of what has become called "big data:" or the creation of datasets that are so large that the traditional tools of analysis no longer work to gain insights from it. The assembly of "big data" sets became possible due

[3] William H. Janeway, *Doing Capitalism in the Innovation Economy: Markets, Speculation and the State* (Cambridge: Cambridge University Press, 2012).

to the rapid rate of digitization of information and the falling cost of data storage.[4] The primary uses of "big data" in commercial applications are to cut costs for firms by finding new patterns, identifying new markets, and targeting services toward specific niches of customers. In effect, the big data revolution has made it possible for firms to observe activity that was once too minute or too private to be exploited by the commercial sphere and to use it to target products and services at potential customers. The primary source of revenue for data-intensive firms is targeted advertising. For example, in Q2 2019, eighty-five percent of Alphabet's (the parent company of Google) revenues still come from advertising.[5]

However, advances in so-called, "artificial intelligence" (AI), especially the breakthroughs in neural networks in 2013, offered the potential for other forms of business. The potential of AI has created a flood of new firms and ventures attempting to apply these technologies to areas as disparate as self-driving cars and check processing. What is significant about this potential boom in AI is that it relies on algorithms "learning" from ever-expanding data sets to achieve better results. This means that the collection and storage of data, in and of itself, will have increasing intrinsic value. Firms have the incentive to capture the value created by social interactions by creating "platforms," which establish closed spaces to record networked interactions. The records of these interactions can then be provided, on demand, to operators of machine learning technologies.[6]

The potential power of AI incentivizes even greater concentration of data gathering and storage by "platform" companies. Platform companies build interlocking ecosystems of services that create connections between users and vendors but also collect vast amounts of data on user behavior. This presents a challenge because the extraction of user data has vast *network effects*; as platform firms collect more data, they can better tailor their services to more customers, thereby aggregating more data, and capturing more market share. Machine learning tools mean that the more data is captured, the more precise services can become. Platforms that store user data *en masse* can thus develop network effects as an initial accumulation of user data allows businesses to offer better services, thus accelerating the accumulation of users and data. The large-scale collection of data by firms are often not directed by a specific analytical task or, if collected for a specific purpose, can be reused for new processes. This generates a return

---

[4] Martin Hilbert and Priscila López, "The World's Technological Capacity to Store, Communicate, and Compute Information," *Science* 332, no. 6025 (April 1, 2011): 60–65.

[5] Mike Murphy, "Google's Future beyond Advertising Starts to Become Clearer," *Quartz*, https://qz.com/1675133/alphabet-q2-2019-earnings-show-non-google-revenue-lags/.

[6] Philipp Gerbert and Michael Spira, "Learning to Love the AI Bubble," *MIT Sloan Management Review* 60, no. 4 (2019): 1–3; Daniela Hernandez, "Meet the Man Google Hired to Make AI a Reality," *Wired*, January 16, 2014, https://www.wired.com/2014/01/geoffrey-hinton-deep-learning/; Dave Gershgorn, "The inside Story of How AI Got Good Enough to Dominate Silicon Valley," *Quartz*, July 18, 2018, https://qz.com/1307091/the-inside-story-of-how-ai-got-good-enough-to-dominate-silicon-valley/.

to scale for the collection of data even when individual analytical processes might have diminishing returns to the analysis of new data points. These return to scale of data collection means that early entrants into the field have outsized returns and reduced competition. The size of these data pools make them uniquely vulnerable to security risks and create risks to personal privacy as well.[7]

Anti-competitive practices by large platform companies are common enough that developers have an industry term for them: "the kill zone." A new product is often dependent on a platform for collecting data and for distribution. Once this product begins to show potential, it enters a kill zone in which the parent platform threatens to replicate this product's services or to buy it outright. Many entrepreneurs chose the latter option instead of taking the risk of being put out of business. Thus, Silicon Valley, like the rest of the American economy, is becoming ever more concentrated around a few "superstar firms" that dominate their respective markets.[8] The stark realities of these practices were highlighted in a recent Congressional hearing which acquired internal emails from Facebook's Mark Zuckaberg related to the company's purchase of rivals, Instagram and Whatsapp. "The businesses are nascent but…if they grow to a large scale they could be very disruptive to us" read the communication several days before the acquisition of Whatsapp. The same day Mr. Zuckaberg wrote that "Instagram was our threat," but that, "one thing about startups though is you can often acquire them."[9]

The effects of this economic transformation are evident in the economic data. A variety of studies have found that, since the turn of the century, concentration in American industry has increased. Meanwhile, the rate of new business creation has fallen off a cliff since the "Great Recession."[10] Concentration has had negative effects on the American economy. One of the mysteries economists have been tackling is a slowing of productivity since 2004. Productivity – or the unit of GDP produced per worker – is the key component of economic growth. New technologies should increase the productivity of workers. Yet, despite rapid advances in technology, for some reason, this metric has

---

[7] Charles I Jones and Christopher Tonetti, "Nonrivalry and the Economics of Data," Working Paper (National Bureau of Economic Research, September 2019); Maryam Farboodi et al., "Big Data and Firm Dynamics," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, January 1, 2019), https://papers.ssrn.com/abstract=3334064; Matthew Hindman, *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy* (Princeton University Press, 2018).

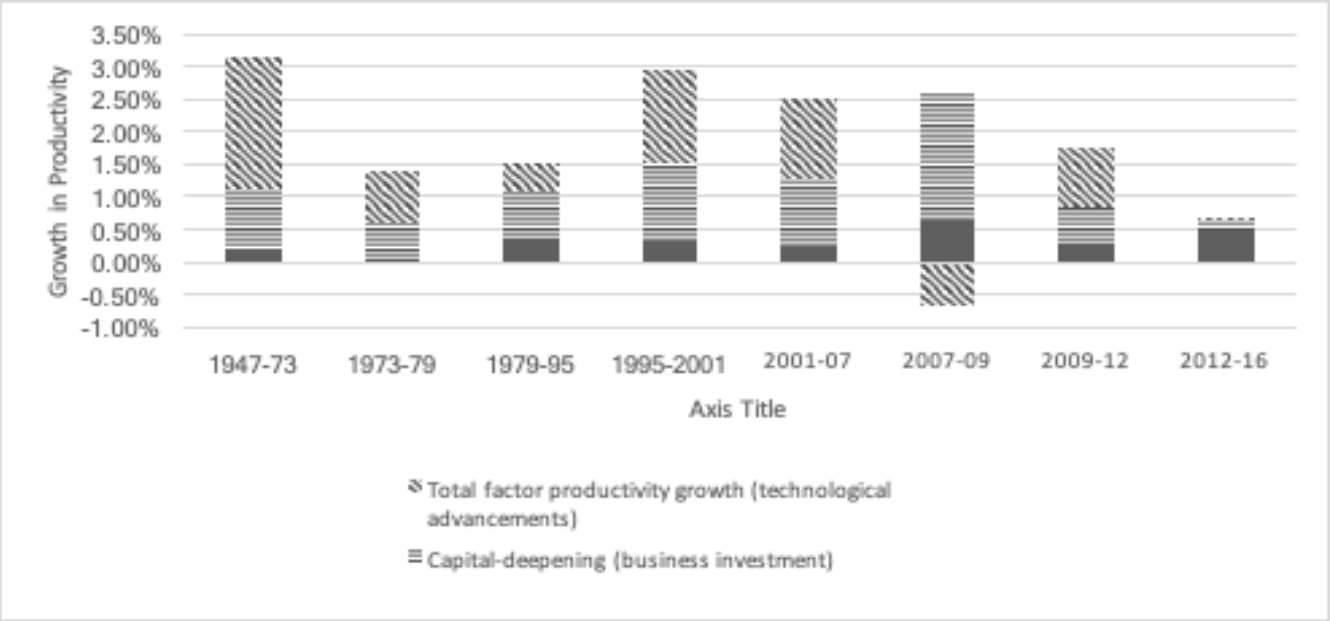
[8] Hal Singer, "Inside Tech's 'Kill Zone': How to Deal With the Threat to Edge Innovation Posed by Multi-Sided Platforms -," *Pro-Market*, November 21, 2018, https://promarket.org/inside-tech-kill-zone/.

[9] "Facebook's Zuckerberg Skewered with Internal Emails during Antitrust Hearing," *Reuters*, July 30, 2020, https://www.reuters.com/article/us-usa-tech-congress-facebook-idUSKCN24U3DG.

[10] Matias Covarrubias, Germán Gutiérrez, and Thomas Philippon, "From Good to Bad Concentration? U.S. Industries over the Past 30 Years," Working Paper (National Bureau of Economic Research, June 2019).

not only slowed to a crawl but has increasingly come from workers' investments in their skills rather than technological upgrading.[11]

*Figure 1: Sources of Productivity Growth in the United States Economy*




Source: John G. Fernald, "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity," Working Paper Series (Federal Reserve Bank of San Francisco, 2012), https://ideas.repec.org/p/fip/fedfwp/2012-19.html.

Initially, some analysts were sanguine. One popular theory was that the rise of "intangibles," such as data, simply meant that old metrics were mismeasuring productivity and investments by firms into new technology. The implication was that the challenge of the digital economy was not distribution or organization but rather a lack of skills.[12] **Despite the public prominence of this thesis, it has largely been debunked.** One of the advocates of this thesis has now quietly walked back his claims and now argues that productivity growth now *might* happen with the introduction of AI but is no guarantee.[13]

---

[11] John G. Fernald, "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity," Working Paper Series (Federal Reserve Bank of San Francisco, 2012), https://ideas.repec.org/p/fip/fedfwp/2012-19.html.

[12] Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, 1 edition (New York London: W. W. Norton & Company, 2016).

[13] Kemal Derviş and Zia Qureshi, "The Productivity Slump—Fact or Fiction: The Measurement Debate," *Brookings* (blog), August 26, 2016, https://www.brookings.edu/research/the-productivity-slump-fact-or-fiction-the-measurement-debate/; Chad Syverson, "Challenges to Mismeasurement Explanations for the U.S. Productivity Slowdown," Working Paper (National Bureau of Economic Research, February 2016); Erik Brynjolfsson, Daniel Rock, and Chad Syverson, "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," Working Paper (National Bureau of Economic Research, November 2017).

A more plausible explanation for this slowdown is the concentration of new technologies in "superstar firms" who use their advantages in gathering information to prevent technology from defusing to competitors. Concentration is responsible for the falling "wage share" – or share of profits going to workers. The slowdown in productivity has been a significant contributor to the expanding rich-poor gap because more profit is going to a narrower segment of the population.[14]

B. The Collective Value in the Data Dependent Economy

In evaluating the economic value of data, we must also face an existential issue: what exactly is data? Is it a commodity that is fungible like a natural resource? Can it be freely exchanged? And if it does have some of these qualities of a natural resource – the "new oil" – where is it "extracted" from? These questions can be answered by understanding how the data economy is constituted through the monetization of social relations.

The extraction of data from individuals about their social interaction is a transformation in how individuals engage in economic transactions with private business. This is not a new phenomenon: new productive techniques often involve a renegotiation over control of what once considered "the commons" – an area of life outside of the economy. Naturally, capitalist firms have tried to maximize their ability to collect revenues from the lopsided exploitation of the commons and the wage-labor relationships that the destruction of the commons creates. Folk songs about the company store illustrate just one of the ways that this process has played itself out.[15]

Data-dependent economic activities are such a transformation. Today, we talk about data collection as tapping into an individual's social "exhaust". This exhaust captures everything from creative content generated from social media interactions to unintended information such as location. In both examples, these once purely social interactions were at some point not thought of as an activity that produces a profit for a corporate entity. A technological change has allowed these once non-commercial individual interactions to be aggregated at scale and exploited for profit. In other words, big data is a commodification of the networks that constitute "society." This means that data is a socially produced good with natural returns to scale. **Practically, this means that your data is not as valuable as data collected from multiple individuals.** This means that there is likely no market where you could sell 'your data' in isolation -- the value of your data is only unlocked when it is combined with data from others, much in the same way as if you were working on an assembly line producing clothes at

[14] David Autor et al., "The Fall of the Labor Share and the Rise of Superstar Firms," Working Paper (National Bureau of Economic Research, May 2017); Jason Furman and Peter R. Orszag, "Slower Productivity and Higher Inequality: Are They Related?," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, June 6, 2018), https://papers.ssrn.com/abstract=3191984..

[15] Yochai Benkler, "The Political Economy of the Commons," *Upgrade* 4, no. 3 (July 2003): 6–9; Yochai Benkler, "Commons and Growth: The Essential Role of Open Commons in Market Economies," ed. Brett M. Frischmann, *The University of Chicago Law Review* 80, no. 3 (2013): 1499–1555.

industrial scale, the particular shirt seam you are sewing on each shirt is not valuable / sellable on its own -- the market exists only for the final product, which aggregates the output of all the workers on the assembly line. We can think of data aggregation as the assembly of a new final product -- an advertising base, a predictive model -- where only industrial scale operations have the ability to produce a product for which a market exists.

Thus, our interactions with data platforms should be thought of as a type of "collective labor" that has, for better or worse, displaced many aspects of the social commons. For example, purchasing a book from a bookstore, in principle, generated a similar stream of information as an individual transaction with Amazon . However, since the storage of data has become cheap and its processing has become efficient, this once ethereal information can be combined with other information streams to create a monetizable product from relationships once outside of the capitalist economy. Intimate connections like conversations between families are now mediated by social networking platforms that gain their revenues from the buying, selling, and interpretation of large sets of data.

It is the obligation of a democratic government, as the collective agent of society, to intervene to manage this new sphere of capitalist life to both maintain many useful features of the social commons and to shape markets in a manner that makes sure that their exploitation is productive and beneficial to all parties, not just the capitalist firms that hitherto accumulated these resources for their own profit at practically zero cost.[16]

C. From My Data to Our Data

The functioning of a data dividend must reflect the unique value of data as a resource in the modern economy. To build the framework for thinking about an effective policy for capturing the value of data for the public good we must shift our paradigm from "my data" to "our data." Thus, we define the goal of a "data dividend" as:

1. Creating a source for public revenue that can directly or indirectly contribute to personal wealth building.
2. Integrating with measures that will incentivize private firms to change their behavior toward the larger commons and their exploitation of social data.
3. Creating institutions that restructure the markets for data toward the public good by fostering competitiveness and an economy in which the benefits of productivity flow to as many groups as possible.

From these goals we have developed a series of "principles" which we have used to evaluate options for introducing a data dividend:

---

[16] This analysis is derived from Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action*, (Cambridge: Cambridge University Press, 1990).

**Table 1: Principles for Evaluating the Efficacy of "Data Dividend" Policies**

| **Principle (shorthand)** | **Short Description** |
|---|---|
| Inequality Reduction (*inequality)* | Data dividends should reduce economic inequality by sharing the financial winnings of data-driven technologies with all CA residents. |
| Address Structural Causes of Inequality (*structural causes)* | Data dividends should affect structural causes of inequality, in particular the monopolistic nature of data- and network-based business models. |
| Fair Burden (*fair burden*) | Data dividends should impose a lower tax burden on companies that do not rely on private data. |
| Avoid Complexity (*complexity)* | Data dividends should avoid imposing complex processes on the state or businesses that create an administrative burden, excessive business overhead, or gameable "loopholes". |
| Avoid Privacy Harms (*privacy*) | Data dividends should not create new major privacy risks or impede advancements in privacy. |
| Equitable access to AI Capabilities (*equitable AI access*) | Data dividends should equitably share and maintain the capabilities of AI for all CA residents. |
| Maintain California's AI Leadership (*AI leadership*) | Data dividends should maintain CA's position as a global leader in the AI industry. |
| Legality (*legality)* | Data dividends should not violate existing laws. |

*Appendix 1* discusses our reasoning in defining these principles in more detail.

## II. Raising Revenue from a Data Dividend

The first, and perhaps most obvious, component of a "data dividend" is a revenue-raising mechanism that returns proceeds to users for their digital labor. In our deliberations, this committee has evaluated various models of raising revenue. Some alternatives we have considered include:

- "Data dependence" tax based on the calculation of firm data exploitation and sales apportioned revenue (our recommendation)
- A tax on direct sales of consumer data to third parties (our recommendation)

- A "pay per unit" tax based on firm self-assessment and state technical guidance.
- A "pay per use" tax based on the implementation of a future data usage counting technology.

Based on the criteria presented in section I, we believe that a sale apportioned "data dependence" tax would be the most effective revenue-raising mechanism for a data dividend. We base many of the features of such a tax on the so-called "GAFA taxes" (referring to Google, Amazon, Facebook, and Apple) that are being introduced in a variety of European Union jurisdictions. However, we go beyond these measures by recommending a specifically data-dependent, rather than company revenue-driven classification of taxable income.[17]

The tax schemes we suggest return revenues from data dependence while not unduly belaboring firms that do not use massive amounts of data or smaller "startups" that have only very small data sets. A data dependence tax incentivizes firms to upgrade privacy protections and make crucial decisions about the efficacy of exploiting personal data rather than simply "vacuuming it up" as it is generated. In accordance with these goals, we also recommend that California introduce a tax on the sale of data to third parties such as advertisers and brokers.

The following discussion will paint the broad outlines of what we believe are viable taxation measures. It will not attempt to make recommendations regarding efficient rates or specific measures of dependence measures. Such concerns will require further research that we recommend the government commission if it accepts our recommendations. Appendix 4 of this report presents further clarification and suggestions on the parameters of these future research efforts.

### A. Data Dependence Taxes

While it is difficult to value individual data, we believe that the taxation of "data dependence" is possible. **We recommend California explore the design of what we term "a data dependence tax." This measure would tax the sales apportioned revenue of firms that use large amounts of personal data.** Below we outline two ways in which these taxes could be assessed: through a marginal user tax, or a flat tax based on a modified version of the CCPA's requirement.

#### 1. Marginal User Tax

California should consider implementing a marginal tax on user-base sizes. This tax would create several tax brackets that depend on the number of California users whose data is stored by a firm. The tax brackets are then assessed against the firm's revenue. This can be done through a variety of methods, including a per user value formula

[17] The European Commission Taxation Department, "Data on Taxation" (The European Commission, September 13, 2016), https://ec.europa.eu/taxation_customs/business/economic-analysis-taxation/data-taxation_en.

described in Appendix 4, or by assessing the total tax value against the firm's sales apportioned revenue.

We believe that using a marginal system would produce the a *fair burden* for firms engaged in the data space by taxing user-heavy firms at a higher rate than firms that are less dependent on accumulations of personal data. Because they serve fewer customers, most service providers--such as a car dealership using data to better schedule its customers--would be excluded from paying a tax, while large social networks that monetize their data would carry a heavy tax burden. A marginal structure could be optimized to maintain incentives for firms to grow while also assessing different taxes at each level of growth. A drawback of a marginal model is that drawing tax brackets is more an art than a science. Thus, extensive work must be conducted to understand what the optimal tax structure is. During legislation, powerful interest groups will have more time to adjust the brackets in their favor.

2. A Flat Tax based on Data Intensity

An alternative and perhaps more immediate approach to taxing data intensity can be to use a simple, "yes or no" criteria for assessing data intensity and attaching a flat tax on all firms that meet a "yes" threshold.

For example, the CCPA has a set of such criteria to determine if a firm falls under its requirements:

- Has annual gross revenues exceeding twenty-five million dollars
- Alone or in combination, annually buys, receives for the business's commercial purposes, sells, or shares for commercial purposes, alone or in combination, the personal information of 50,000 or more consumers, households, or devices.
- Derives fifty percent or more of its annual revenues from selling consumers' personal information.

The CCPA was designed as a consumer protection legislation rather than a "data dependence" measurement. *Thus, if legislators wish to pursue this action, we recommend that the CCPA be a starting point for these criteria but not an endpoint.* The legislature must consider what portion of the data economy they wish to tax and adjust the gross revenue threshold accordingly. Second, there must be at least one other criterion to compliment a revenue component. The CCPA is a strong starting point for such criteria.

While this method of measuring data intensity avoids many of the complexities of a detailed, bracket based system, it may reduce the *fair burden* of tax incidence and thus create a situation where smaller and younger firms are placed at a disadvantage compared to more established entrants. *For this reason we believe that the marginal*

*user tax is a superior policy instrument for fully addressing the evaluation criteria of a data intensity tax.*

B. Sales Apportionment

We recommend that a "data dependence tax" be applied through a "sales-apportioned" model of revenues. Sales apportionment can be achieved through a combination of two assessments:

- Percentage of revenues generated in the state of California
- Percentage of users domiciled in California as defined by the CCPA

In a globalized economy where registration is gamed by firms to reduce tax burden, apportionment, already part of California's corporate tax scheme, is the most effective tool to counter avoidance.[18]

For data-dependent firms, revenue is more appropriate to tax than profits – in the current economy, data-dependent firms often have low profitability but high revenues. Such a revenue collection schema is necessary because the strategy of many firms in the data-dependent sphere is not to create profits but rather to continually expand market share to capture the maximum volume of a market and its attendant data. Thus, a revenue-based tax has a secondary benefit of focusing firms on building profitable high margin businesses from the beginning, rather than aiming to establish a monopoly.[19]

These features give a sales-apportioned revenue tax advantages for avoiding complexity and addressing the structural causes of inequality.

C. Data Sales Tax

One reason to adopt a data dependence tax is that we believe that this measure will head off exploitation in the emerging AI ecosystem. **However, for many firms operating with user data, most profits do not come from AI but from advertisement and sales of user data. Because of this, we recommend that the state of California examine options for imposing a tax on the sale of data to third parties known as "data brokers."** These measures are consistent with the principles of the CCPA and are implicit in the above revenue schemes. However, we believe that a data sales tax should apply *across the board at a flat rate* rather than be tiered into brackets or exempted for companies whose revenue is too low to fall into a single "data dependent" category.

## III. Revenue Disbursement

---

[18] "Taxing Multinational Corporations in the 21st Century," *Economics for Inclusive Prosperity* (blog), accessed February 3, 2020, https://econfip.org/policy-brief/taxing-multinational-corporations-in-the-21st-century/.

[19] Shaoul Sussman, "Prime Predator: Amazon and the Rationale of Below Average Variable Cost Pricing Strategies Among Negative-Cash Flow Firms," *Journal of Antitrust Enforcement* 7, no. 2 (July 1, 2019): 203–19.

We believe that the state should act as an intermediary between data-dependent firms and the citizenry for its social labor. This means that a central question for the implementation of a data dividend is not just a matter of collecting revenues but deciding what means are best to disburse them. As a committee, we have deliberated on the following options examined in Appendix 3:

- State Spending on public services (recommended)
- Baby bonds or universal saving accounts (recommended)
- Per capita payments to all California residents, in other words, universal cash payments
- Meritocratic payments for data contributions

In our opinion the first two options would be most effective in reducing inequality and its structural causes and reducing complexity is through the funding of public services such as education and programs to accommodate technological unemployment *and/or a system of pre-seeded wealth building accounts that some authors refer to as "baby bonds."*

**These measures accomplish similar goals of creating mechanisms by which Californians can *build wealth* and improve their bargaining position relative to large corporate entities.** Creating such public, universal goods best reflects the fact that the underlying value of data comes from aggregation rather than individual inputs.

A. Improved Public Services

There is a strong correlation between the provision of universal services, such as healthcare and education, and the ability of households to build sustainable savings that create household resiliency. Strong universalistic public services allow households at the middle and bottom of the income distribution to use their savings to increase discretionary spending and inter-generational wealth transfer, while also being less dependent on specific employers. An increase in the quality and extent of state services serves as a "wealth building" measure.[20]

Education is an area in which access can be easily improved. The primary source of productivity growth has not been investment in installing new technology by private companies. Rather, the source of American growth has been a more educated workforce that can take advantage of investments into technologies made in the previous decades[21]

---

[20] Monica Parsad, "The Trade-Off between Social Insurance and Financialization: Is There a Better Way?," Niskanen Center Policy Essay (Washington, D.C, August 20, 2019), https://www.niskanencenter.org/the-trade-off-between-social-insurance-and-financialization-is-there-a-better-way/.

[21] Josh Bivens, "A 'High-Pressure Economy' Can Help Boost Productivity and Provide Even More 'Room to Run' for the Recovery," *Economic Policy Institute* (blog), accessed August 5, 2020,

Most of this "labor upgrading" has come at the expense of American households. The student debt crisis is the flip side of economic growth mainly being driven by improvements in education. Universities can increase prices given the higher demand for education. The importance of prestigious degrees as a signaling device for inclusion into the upper-middle class allows elite universities to increase the cost of education independent of the actual content of courses or quality of instruction. On the lower end of spectrum, for-profit colleges used the under funding of public education to charge excessive fees on poor students desperate to be included in the economy. Meanwhile, the increased economic concentration in the American economy has allowed employers to demand higher skills, and thus personal educational outlays for lower pay. Thus, the returns on education are falling just as it becomes the most important criteria for the most basic form of economic inclusion. Support for student debtors, and/or the lowering of the cost of secondary education would be an effective intervention in increasing the net wealth of households and offset harms caused by the data economy.[22]

B. "Baby Bonds"

Another option for California to pursue to reduce wealth inequality would be to use the revenues raised from a data tax to help seed "baby bonds" – or guaranteed savings accounts at birth – for all residents. Baby bonds respond to research that shows that parental net worth is the most significant predictor for an individual's wealth. Darrick Hamilton and Sandy Darity have estimated that a baby bond that would close the racial wealth gap would entail *$50,000 to $60,000* issued at birth *with an appreciation of 1.5 to 2.0%* p.a for the lowest wealth quartile. Under the Hamilton-Darity plan, higher quartiles are issued bonds between $20 and $25,000. These bonds would become available for housing, business, and educational expenses once the child turns eighteen.[23]

## IV. Institutional Measures to Democratize the Internet

The evolution of the data economy since the 1990s has been shaped by largely uncoordinated changes in technology. However, the commercialization of technologies is not neutral. The lack of regulation and guidance has created path dependencies that amplify the market power of entrants and exacerbate pre-existing inequalities. Four decades of spontaneous development have set the stage for a digital economy that is

---

https://www.epi.org/publication/a-high-pressure-economy-can-help-boost-productivity-and-provide-even-more-room-to-run-for-the-recovery/.

[22] Julie Margetta Morgan and Marshall Steinbaum, "The Student Debt Crisis, Labor Market Credentialization, and Racial Inequality: How the Current Student Debt Debate Gets the Economics Wrong," Roosevelt Institute Policy Papers (New York, October 16, 2018), https://rooseveltinstitute.org/student-debt-crisis-labor-market-credentialization-racial-inequality/.

[23] Darrick Hamilton and William Darity, "Can 'Baby Bonds' Eliminate the Racial Wealth Gap in Putative Post-Racial America?," *The Review of Black Political Economy* 37, no. 3 (September 1, 2010): 207–16.

structurally bound towards monopoly, low levels of technical diffusion, and exploitative relationships with users.[24]

The introduction of AI presents a juncture. **We can either let the system develop with the inertia it has inherited from the past or treat this crossroads as an opportunity to head off a more exploitative economy by building the institutions that will restructure the internet into a more democratically governed space.** Given California's large market and central position in this economy, state-level regulation can begin a larger process of transforming our collective future.

This section outlines our recommendation for measures that can be executed in the near and medium term to begin tackling the underlying issues that motivate this report and the "data dividend."

### A. Establishing the Data Relations Board (DRB)

The challenge of designing a policy to improve the data-driven economy is that we are still in its very early days. It is likely that within a decade, the above measures could be completely ineffective in regulating and governing the new processes of value creation that emerge from these new technologies. This uncertainty provides an opportunity: as new technologies reshape productive relations, governments can concurrently steer the new economy towards serving the common good.

For this reason, we recommend that California establish a government entity we call a "Data Relations Board" (DRB). The function of the DRB would encompass:

- Examining the effectiveness of data tax revenue usage for offsetting specific economic and privacy harms created by a data economy.
- Funding and administering (a) research studies that address critical questions about the data-driven economy and (b) incubation projects that leverage the results of these and other studies to explore improvements to the data dividend program.
- Addressing and adjudicating new data dividend-related challenges that arise, e.g. classifying which companies are data-dependent as the economy evolves.

An analog to the DRB is the Environmental Protection Agency (EPA). The Clean Air Act mandated the EPA regulate air quality and fund ongoing research into the topic. The Act was amended multiple times to update the scope of regulations. Critically, the Clean Air Act provides a model for adaptable policymaking, that implements immediate change while supporting research for understanding the implications of future regulations and technical change. Given the low information nature of the data governance, we believe

[24] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: Public Affairs, 2019).

this model carries the greatest potential for any data dividend policy and further regulation of the data economy.[25]

### B. Public Data Trusts

The DRB should be tasked with serving as the administering body for "public data trusts." These trusts are legal entities that hold rights to data with the DRB serving as the owner, on behalf of the public. The concept of data trusts has emerged in the context of the creation of "smart cities" and the problems of civic data ownership that accompany them. This report argues that the mission of these entities should be expanded and integrated into the program of data dividends by creating incentives for private firms to integrate their proprietary data sets with publicly accessible data.[26]

The DRB should make the public data trust an active instrument of governance by soliciting new, or existing data sets that are necessary for solving public interest, including establishing common pools of data that startup firms would be able to access in areas of social or economic interest. Establishing such a pool would alleviate the problem of firms gaining a first-mover advantage rather than through the most innovative and efficient systems of data analysis. An active and powerful set of data trusts creates the opportunity for the state government to encourage private industry to work toward goals that complement the public's interest and to engage in less extractive business models.

A simple first step toward establishing such entities would be to compile, large dynamic state data sets that can be supplemented by proprietary data. The DRB should create a set of warrants that specify the scope, quality, and formatting of private data to be integrated into this trust. Once the DRB approves a firm's transfer, it can offer tax breaks from the data dependence tax and create specific bounties and contracts to incentivize private firms to complement these public data sets with proprietary data for public use.

### C. A Data Industrial Policy

The state should consider a "data industrial policy" to incentivize data-dependent firms to work with the state rather than against it. Public data trusts are one component of a

---

[25] Janet Currie and Reed Walker, "What Do Economists Have to Say about the Clean Air Act 50 Years after the Establishment of the Environmental Protection Agency?," *Journal of Economic Perspectives* 33, no. 4 (November 2019): 3–26.

[26] Bianca Wylie and Sean McDonald, "What is a Data Trust"*Center for International Governance Innovation* October 18, 2018. https://www.cigionline.org/articles/what-data-trust

public-private ecosystem that is more accountable to the public. Other projects the state might consider to help advance these goals might include:

- Public internet and mobile connection
- Public technological platforms for use by individuals and firms
- An industrial plan for A.I. which seeds new firms and gives the state a stake in them.
- Dissemination and commercialization of research or development stage technologies, management techniques, or new platforms.

- Establishing ownership rights in patents that were funded by state institutions and grants.

This type of government-led innovation is at the root of Silicon Valley's success. A similar government program, the Defense Advanced Projects Agency (DARPA), provided the initial seed capital for most of the general-purpose technologies upon which the data-driven economy has been built. DARPA pursued what is now termed a "mission-driven" development strategy whereby it not only set goals for development but actively forged the networks of developers needed to pursue this mission. The critical elements of this were active coordination through DARPA and the mandatory sharing of intellectual property between competitors working under the DARPA umbrella.[27] The DRB's active participation in the data ecosystem through the creation of publicly-owned data and infrastructure perform this function in a new century and make sure that productivity moves quickly through all levels of the economy rather than just the top.

## D. Data Cooperatives

To overcome the disparity of power between platforms and users, we recommend empowering and incentivizing businesses to work with data cooperatives that represent consumer interests on an aggregated basis. Consumer data cooperatives are brokers for user data representing groups of users. These brokers will have fiduciary responsibilities over user data, defined by a contract between user and cooperative. A data cooperative solves the principal-agent problem of the platform economy by turning individual users into a collective bargaining organization that can then represent them before large data firms. Additional legal and regulatory interventions (for example by a Data Relations Board) are necessary for data cooperatives to gain a meaningful foothold in most parts of the data economy.[28]

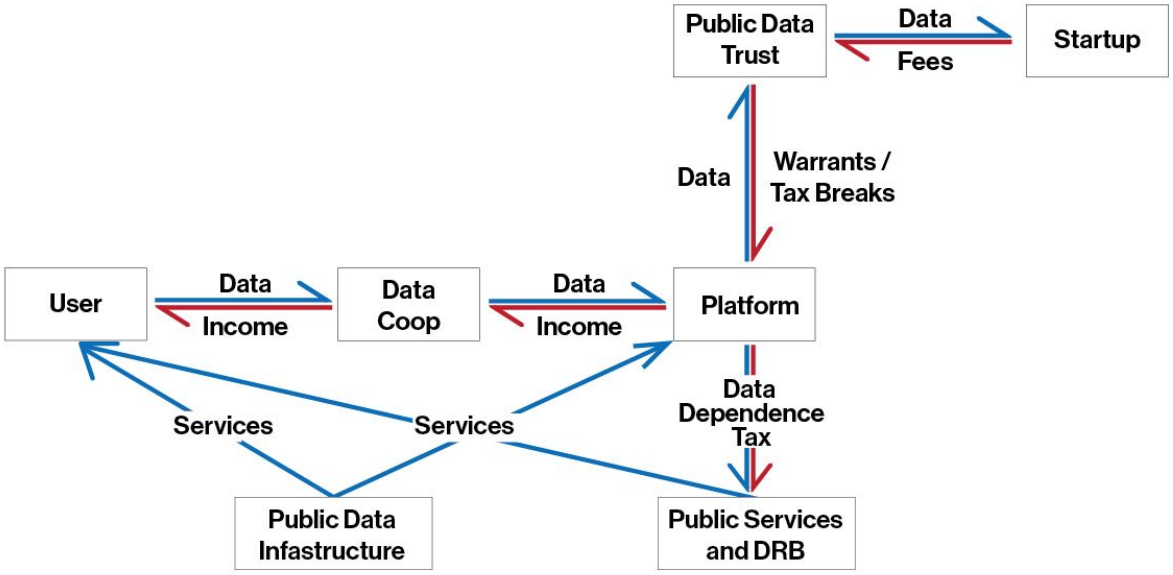
[27] Mariana Mazzucato, *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*, Revised edition (New York: Public Affairs, 2015).
[28] See Appendix VI.

An example of an already existing data cooperative is the MIDATA cooperative for medical data organized by ETH Zurich and the Bern University of Applied Sciences. MIDATA members voluntarily hand over their medical data to the cooperative and subscribe to levels of privacy. MIDATA then works as a broker between members and commercial end-users. MIDATA is a strictly non-profit cooperative and thus does not provide a cash payment to its members.[29]

Cooperatives can choose to share their revenues with their members by making them equity holders. Ideally, users might be able to choose which cooperative to join based on considerations of privacy versus profit, or other preferences.

In addition to enabling legislation, the state of California could, if it so desired, use the above-described revenue mechanisms to incentivize the use of consumer data cooperatives. For example, firms that agree to work with data cooperatives, or even mandate their users to be members, might be eligible for a tax break from data taxes. For example, every user that accesses a platform through a data cooperative as a broker could reduce the firm's user count.

*Figure 2: Components of our proposed Data Dividend*




## Conclusion: The Full Data Dividend Is a More Democratic Economy

The digital economy has been steadily transforming itself from one that sells products to customers to a "data-driven" economy in which the customers' data is now as much of the company's worth as are its sales. This presents a challenge for privacy, equality, and

[29] Ilse van Roessel, Matthias Reumann, and Angela Brand, "Potentials and Challenges of the Health Data Cooperative Model," *Public Health Genomics* 20, no. 6 (2017): 321–31.

economic justice. However, it is also an opportunity to restructure this economy at a critical juncture. We can make sure that wealth and productivity do not only concentrate at the top but also spread to all levels of society.

**Therefore, we believe that the data dividend should not be reduced to a revenue-raising measure. Rather, it should be a series of actions that restructure the incentives and institutions of the new economy to make it complementary instead of extractive.** We recommend that in evaluating the options that this report has presented, the government understand them all as a component of a system that will remunerate Californians for their contemporary social labor while also creating the long-term incentives to make sure that the fruits of this labor are not just shared as "wages" but as a new structure of mass participation in the new economy. This way we can transition ourselves from doing digital labor to participating in a democratic digital economy.

## Appendix 1: Principles

In this Appendix, we detail eight principles that we used to evaluate the revenue collection options (Appendix 2) and dividend disbursement options (Appendix 3) we identified.

### A1.1 Wealth Inequality Reduction

The core motivation of data dividends, as stated by Gavin Newsom, is that "consumers should also be able to share in the wealth that is created from their data."[30] In other words, the economic value created by data is being concentrated rather than broadly shared. By broadly sharing this value, data dividends can reduce economic inequality, which is itself associated with a wide variety of societally dangerous outcomes.[31] If a given data dividend implementation does not reduce wealth inequality, it is unlikely that it is achieving the goal of helping consumers "share in the wealth that is created from their data." To reduce wealth inequality a policy must not only work to capture *flows* of income but also create a more equitable distribution of *stocks* of asset ownerships. These mechanisms can be of individual savings mechanisms, publicly owned assets, or reductions in spending burdens on individuals and households.

### A1.2 Address Structural Causes of Inequality

Data dividends should consider the structural relationship between data-driven economies and wealth inequality. Autor et al. argue that the inability of labor to benefit from rising productivity is directly related to the monopolization of productivity gains by superstar firms in various sectors.[32] There is evidence that the features of the data-dependent economy exacerbate this situation.

There is strong evidence that data has a network effect, that rewards returns to scale with little to no decreasing marginal returns. This means that companies have the incentive to completely dominate a market and use first-mover advantage to lock out competition. This gives these large firms a distinct negotiating advantage vis-a-vis their users and employees.[33]

---

[30] Gavin Newsom, "State of the State Address" (2019), https://www.gov.ca.gov/2019/02/12/state-of-the-state-address/.
[31] Juliet B. Schor, "Does the Sharing Economy Increase Inequality within the Eighty Percent?: Findings from a Qualitative Study of Platform Providers," *Cambridge Journal of Regions, Economy and Society* 10, no. 2 (July 1, 2017): 263–79.
[32] David Autor et al., "The Fall of the Labor Share and the Rise of Superstar Firms," Working Paper (National Bureau of Economic Research, May 2017).
[33] Carl Shaprio and Hal R. Varian. *Information rules: a strategic guide to the network economy*. Harvard Business Press, 1998; Newman 2014, "Search, antitrust, and the economics of the control of user data" and Grunes and Stuckes 2015, Daniel L. Rubinfeld

Anecdotal evidence supports the existence of a network effect and its economic consequences. For example, Facebook has repeatedly used its user data to push competitors out of new lines of business.[34] In fact, despite the press attention to "startups" the majority of new companies established in Silicon Valley are sold to one of the major technology firms.[35] The increasing concentration of productivity into fewer firms has significant effects on wages and employment.[36]

More systematic empirical research is still evolving, with studies providing mixed results regarding the strength of big data's network effects. One reason for this degree of uncertainty is that machine learning applications are still very young and the radical change they introduce into computing will likely change the ways firms interact with previously collected data. In fact, while there is some evidence that data does not have a return to scale in *individual* machine learning processes, it does have network effects in aggregate. While a single process might require a specific data set with diminishing returns, collecting big data indiscriminately allows established firms to be the providers of any data set before the task's specific definition. In other words, while you might have less need for more data at some specific point, you still must find your data somewhere and the large platforms own it. Thus, returns to scale still hold in general even if they don't in specific processes.[37]

Understanding data as a set of assets on the balance sheet of a platform that can be monetized lets us get to the heart of how unequal access to this new resource incentivized inequality -- the legal claims to cash flow and profit generated from an uncompetitive market that is ultimately, mediated by the state. Like any asset, data carries no economic value without ownership being legally defined, i.e. by platform-level user agreements (e.g. Facebook's terms of service) or the state itself (e.g. through

---

and Michal Gal, "Access Barriers to Big Data," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, August 26, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2600051; Allen P. Grunes and Maurice E. Stucke, "No Mistake About It: The Important Role of Antitrust in the Era of Big Data," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, April 28, 2015), https://papers.ssrn.com/abstract=2600051

[34] Olivia Solon and Cyrus Favear, "Leaked Documents Show Facebook Leveraged User Data to Fight Rivals and Help Friends," *NBC News*, November 6, 2019, https://www.nbcnews.com/news/all/leaked-documents-show-facebook-leveraged-user-data-fight-rivals-help-n1076986.

[35] "American Tech Giants Are Making Life Tough for Startups," *The Economist*, accessed April 28, 2020, https://www.economist.com/business/2018/06/02/american-tech-giants-are-making-life-tough-for-startups.

[36] Mike Konczal and Marshall Steinbaum, "Declining Entrepreneurship, Labor Mobility, and Business Dynamism: A Demand-Side Approach" *Roosevelt Institute Working Paper,* July 21, 2016.

[37] Charles I Jones and Christopher Tonetti, "Nonrivalry and the Economics of Data," Working Paper, Working Paper Series (National Bureau of Economic Research, September 2019); Maurice E. Stucke and Allen P. Grunes, "Data-Opolies," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, March 3, 2017); Maryam Farboodi et al., "Big Data and Firm Dynamics," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, January 1, 2019), https://papers.ssrn.com/abstract=3334064; Di He et al., "Scale Effects in Web Search," in *Web and Internet Economics*, ed. Nikhil R. Devanur and Pinyan Lu (New York: Springer International Publishing, 2017), 294–310,

data-related legislation).[38] Other authors similarly highlight the structural basis of economic outcomes,[39] and particularly changes in wealth inequality and distribution.[40] Thus,any state intervention to build a more equitable data-driven economy must work to reduce the structural causes of inequality by actively reshaping the process of value creation in the data economy. Transformative policy is not just a simple matter of raising revenue and redistributing it. Rather, any revenue-raising and distributive mechanism must be tailored to creating a data-driven economy in which increased productivity materializes across a wide cross-section of firms and benefits working people.

The data-driven economy would not exist without the legal and physical infrastructure that taxpayers have invested in over many decades. The state, as a representative of its citizens, has the responsibility that all citizens benefit from the public's investment.[41] We argue that one means to do this is for the government to enforce competition and an equitable economy through a variety of public options that would set a floor for the economy and shape the private sector's growth strategy to benefit the greatest number of stakeholders.[42]

### A1.3 Fair Burden

Some businesses benefit much less from consumer data and AI technologies than others, and thus these firms should carry less of the costs of data dividends. For instance, if the profits of small local barbershops are used to fund a data dividend, this would not represent the winnings of AI technologies being shared broadly. More generally, data dividends should seek to impose burden relative to how dependent a company is on the collection and exploitation data for the generation of revenues. It is this classification that makes our policies a *data dividend*, rather than a tax on business in general.

Conversely, because practically all companies use a variety of data about consumers for analytics, it does not mean that self-described "technology companies" should front the costs of data dividends. Many firms with primary lines of business outside of data dependent processes also store and sell customer data.

---

[38] Pistor, Katharina. *The Code of Capital: How the Law Creates Wealth and Inequality.* Princeton University Press (2019).
[39] Lothian, Tamara. "Economic Progress and Structural Vision." In *Law and the Wealth of Nations: Finance, Prosperity, and Democracy*, 197-336. New York: Columbia University Press, 2017.
[40] James K. Galbraith, *Inequality and Instability: A Study of the World Economy Just Before the Great Crisis* (Oxford University Press, 2012).
[41] Mariana Mazzucato, *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*, 1 edition (London ; New York: Anthem Press, 2013).
[42] Ganesh Sitaraman and Anne Alstott, *The Public Option: How to Expand Freedom, Increase Opportunity, and Promote Equality* (Cambridge: Cambridge University Press, 2019).

### A1.4 Avoid Complexity

A data dividend should be designed to achieve a reduction of the structural causes of inequality via the least complicated means possible. By avoiding complexity, we hope to eliminate the possibility for large firms to exploit loopholes in tax implemenation to gain a competitive advantage over their smaller rivals. One of the ways that large firms have become more dominant in the United States economy is by investing resources into legal departments that allow them to reduce tax and regulatory burdens in ways not available to smaller firms.

Moreover, a complex tax scheme involves establishing administrative overhead for the state. In an area as cutting edge as data, this overhead may not only be expensive but also underdeveloped and technically unreliable.

### A1.5 Avoid Privacy Harms

The extraction and analysis of personal data have been directly tied to a variety of privacy harms. A comprehensive data dividend should balance raising revenues with steps to minimize threats to consumer privacy. Fortunately, we believe that these goals may not only be compatible but mutually-enforceable.

In evaluating policy options, this committee examined how they might influence company behavior to increase harms to privacy. If state policies incentivize outsourcing data collection and processing or making datasets more widely available to the public, they also raise the risk of privacy harms due to added steps in the chain of custody. Two areas of special concern are: (1) storing sensitive data insecurely, leading to hacks and leaks; (2) publicizing datasets that appear to be private, but can be de-anonymized.[43] Thus in assessing policy options, we pay attention to ways that the state can regulate published data such as requiring the use of privacy-preserving statistical techniques such as differential privacy.[44]

By treating data as a collective, social asset we can re-imagine the way governments address privacy. Most popular proposals to monetize data treat it as individual labor and thus center it around the creation of market places. These proposals thus assume that "privacy is a luxury good" meaning that it will be the poor and disadvantaged that remain most exploited. In contrast, our framing and policy recommendations establish public representation and distribution that ensures that Californians will have more

[43] Arvind Narayanan and Vitaly Shmatikov, "How To Break Anonymity of the Netflix Prize Dataset," *ArXiv*, 2006.
[44] Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy* (Hanover, MA: Now Publishers, 2014).

options when they interact with the data market and do not have to sell their individual privacy to fully enjoy its benefits.

### A1.6 Equitable Access to AI Capabilities

For all Californians to fully share in the winnings of their data, they should also be able to share in the non-monetary benefits that advances in data science might bring. Unfortunately, without a reform of the institutions of the internet, we fear that any policy adopted by the state might reinforce the "algorithmic bias" that amplifies inequalities. As well, we believe that our policy recommendations should work to make sure that potential AI applications do not widen the "digital divide" across race, economic status, gender, and geography.[45]

Thus, we propose that lawmakers consider the "data dividend" not only as a mechanism to raise and distribute revenue but to fundamentally restructure how the tangible and intangible winnings from new economic activity are dispersed. For example, our discussion will outline a "digital industrial policy" that can be funded via data dividends that would close the gaps between access to broadband, data sets, and education.

### A1.7 Maintain California's AI Leadership

Driven heavily by the economic success of "Silicon Valley" since the 1970s, California has since been a global leader in the tech industry. Many of the "tech giants" are headquartered in Silicon Valley (e.g. Google, Facebook, and Apple). Through the passage of the CCPA, California is leading the U.S. in developing policies to regulate the data-driven economy.

In evaluating policy options, we have paid special attention to introducing taxes and models that will not encourage the relocation of economic activity out of California. For example, a "headquarters tax" would incentivize firms to move their headquarters elsewhere.

For this reason, our recommendations are geared toward incentivizing innovation and reducing the potential of entrenched actors to collect rents. Thus, we encourage lawmakers to view their policy options not as passive tools to piggyback on private innovation but to create incentives that might create clusters of innovation and guide actors to serve the public good.

---

[45], A. J., van Deursen and J. A.van Dijk, The first-level digital divide shifts from inequalities in physical access to inequalities in material access. *New Media & Society,* 21(2) 2009: 354–375 Sen, Shilad; Ford, Heather; Musicant, Dave; Graham, Mark; Keyes, Oliver; Hecht, Brent "Barriers to the Localness of Volunteered Geographic Information", *CHI* 2015.

Finally, we believe that leadership in the new economy is not just a matter of technical development but social innovation. California's existing advantages in technological development and giant market give it the potential to become a global public policy leader. The "California model" will incentivize other jurisdictions to consider similar steps.

### A1.8 Legality

Data dividends must obey existing laws and survive legal challenges. Several potential challenges are important to consider.

The *Internet Tax Freedom Act*[46] prohibits state and local governments from taxing internet access or placing multiple or discriminatory taxes on Internet commerce. The language of that Act focuses on barring differential treatment between the same commercial activity occurring electronically versus non-electronically. The Data Dividend outlined here would not constitute such differential treatment: the tax rules can and should apply equally to non-electronic businesses that collect customer data.

*Non-Discrimination Against Out-of-State or International Commerce:* Data dividends should avoid discriminating against out-of-state commerce to avoid conflicting with exclusive federal jurisdiction: the burden of data dividend-related laws must not be "clearly excessive in relation to the putative local benefits"[47]. While a California-specific data dividend would likely have significant impacts on interstate commerce, it should take care to minimize them, and must not discriminate against out-of-state actors. Similarly, to comport with the USMCA[48] the rules must not disfavor international actors.

*Constitutionality of Regulating Commercial Speech:* In the event that taxes on data transfers are found to be restrictions on commercial speech, data dividends may need to pass the constitutional "intermediate scrutiny"[49] test. This analysis for commercial speech restriction is outlined in the *Central Hudson Gas & Elec. v. Public Service Commission* case[50]. If applied here, a data dividend would likely be found to advance a "substantial government interest", but should take care not be "more extensive than necessary to serve the interest asserted" in order to pass constitutional muster.

---

[46] 47 U.S.C. 151 *et seq.*
[47] Pike v. Bruce Church - 397 U.S. 137, 90 S. Ct. 844 (1970)
[48] https://ustr.gov/sites/default/files/files/agreements/FTA/USMCA/Text/19-Digital-Trade.pdf
[49] https://www.law.cornell.edu/wex/intermediate_scrutiny
[50] Central Hudson Gas & Elec. v. Public Svc. Comm'n, 447 U.S. 557 (1980)

## Appendix 2: Collecting Revenues for a Data Dividend

The most obvious plank of a *data dividend* is that it raises funds from data related commercial activities. In practice this would be some form of tax. In our deliberations, we evaluated a variety of tax mechanisms that could be implemented by California to collect revenues and evaluated them against the criteria described in Appendix 1.

***Our recommendation is to introduce a sales apportioned revenue tax on firms based on a marginal user count as a proxy for "data dependency".*** In our opinion, this type of tax has balanced our concerns better than alternatives. In this appendix, we will examine various options legislators have and test them against our criteria to help put our recommendations in context.

***In addition to this tax, we recommend levying a tax on the sale of data by "data brokers" to third parties.***

We have divided this section into two groups: immediately feasible taxes and solutions that might require advances in technology that do not currently exist but can be reasonably imagined in the near future.

### A2.1 - *Immediately Feasible Taxes*

#### A2.1.1 - *Sales-apportioned Data Dependency Tax*

Our recommendation is for California to introduce a sales apportioned data Dependency tax based on the sales-apportioned corporate income tax.[51] A sales apportioned tax assesses a rate through the share of global revenue generated by sales in a jurisdiction. An example of such a tax is California's current corporate tax, which assesses a rate based on the revenue generated in the state and calculates it as a portion of the overall global revenue of the firm.

The data dependency tax substitutes users for sales. Tax rates should be based on a marginal user so as not to penalize companies for expanding but still compensating for externalities that we have discussed in this report. If properly defined, the user count can act as a rough proxy for a firm's data intensity. Parameters for this definition and other conceptual details of a marginal user tax will be discussed further in appendix IV.

By applying the evaluation criteria described in Appendix I, we believe this tax carries many advantages . First, it raises revenue from activity that contributes to structural inequality by compensating for the network effects associated with large data sets

[51] Zucman, 2018. "Taxing multinational corporations in the 21st century". https://econfip.org/wp-content/uploads/2019/02/10.Taxing-Multinational-Corporations-in-the-21st-Century.pdf

generated from user data. Second, the tax is progressive and its *burden* falls more on firms with larger data sets. As such, it is the closest to taxing the actual value of data.

### A2.1.2 - *Sales Taxes on Data*

Revenues from data-driven commerce and other economic activities can be collected at various points in the value chain with taxes that have well known precedence and predictable impact.

***We recommend that the state tax the sale of data to third parties***. This type of tax would complement the data intensity tax by closing off the possibility of companies selling their data to lower their user counts. Moreover, a data intensity tax is biased to data as it is used in the training AI technologies. However, a significant, if not the major, source of large technology company revenues comes from the selling of data for targeted advertising. As well, it adds an additional layer of taxes on so-called *data brokers* – or company's whose sole business model is the buying and selling of personal data. These firms are infamous for loose security and the dangers they present to privacy.

## A2.2 - Taxation Requiring Technical Advances

Below, we review taxes we believe are currently not feasible without major technical advances, but offer promising strengths that make them worthy of future consideration.

### A2.2.1 - *Pay Per Data Unit (per unit tax)*

A *Pay per Data Unit Tax* taxes firms based on a unit of data that they *collect* and/or *hold* in a manner like an excise tax on tobacco products.[52] One advantage of a pay per use tax is that it is extremely precise and thus satisfies the *fair burden* requirements, as well as treating the *structural causes of data driven inequality*. Furthermore, by specifying different tax rates for the collection of data and the holding of data, this tax can be designed with very precise incentives in mind.

However, a considerable barrier to a *Pay per Data Unit Tax* is that we have no reliable means to make data *fungible*. In other words, it is very difficult to compare the monetary values of individual units of data, or even to choose a neutral standard for a unit's worth to a machine learning algorithm. Data is stored in a variety of formats, new uses of data are frequently developed, and data management practices are not

[52] See discussion here: https://tobaccocontrol.bmj.com/content/25/4/377

standardized across firms. We believe that there is a likelihood that these issues might be overcome with future research and/or regulation. For instance, if additional data standards are instituted more broadly (analogous to standards around medical data)[53] and data valuation techniques become standard[54], this technique may be feasible. However, we believe that this tax cannot be implemented based on current technical standards. One thread of research that the government and its agents may wish to pursue is research into the feasibility of developing such technologies and standards.

### A2.4 - *Pay Per Data Use* (*Use tax*)

As AI becomes more common in commercial applications, it will boost the value of large sets that these processes can "train" on. Under future technological infrastructures, it may be possible to directly track the exact amount each unit of data is used by a firm for a discrete analytical purpose. This would allow for a direct "sales tax" on business to business sales or exchanges of data.

Unlike tax-based approaches or data-counting approaches, which can lean on precedents of tax law and privacy policies, this schema is unsupported by existing infrastructure and is unlikely to be viable in the near term. Such infrastructure could also introduce major *privacy harms* to companies or individuals, e.g. if an attacker can learn about the fraud detection algorithm a bank uses or identifies an individual's location based on which models her data appears in. These privacy harms may be avoidable, but because the idea itself is still mostly speculative.

[53] Institute of Medicine (US) Committee on Data Standards for Patient Safety; Aspden P, Corrigan JM, Wolcott J, et al., editors. *Patient Safety: Achieving a New Standard for Care*. Washington (DC): National Academies Press, 2004) https://www.ncbi.nlm.nih.gov/books/NBK216088/

[54] Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1885-1894. JMLR. org, 2017.

## Appendix 3: Distribution Options for a Data Dividend

This section discusses the committee's deliberations on options for disbursing revenue collected from taxes on data usage. We believe that the most efficient and equitable way to distribute revenues from a data dividend would be to use it to invest in improved social services, education, and in elements of a "data industrial policy," which we will describe in detail in appendix V. We believe that collective spending on public goods is superior to individual remuneration.

*Contextualizing our discussions with ballpark estimates:* Throughout this section, we refer to some ballpark estimates of how much funds could be raised to provide context. To obtain a ballpark figure, we considered that California is projected to receive $13.1 billion in corporate tax in 2019[55] and estimated that a very modest data dividend tax might bring an additional 10% of $1.31 billion[56].

### A3.1 - State Spending

We believe the most effective method with which to disburse revenues generated by a tax on data dependence would be through funding of social and public goods that contribute to the expansion of wealth. The generation of value from the data is a form of exploiting the public commons and thus, the most logical way to return these benefits is through restoring this extracted wealth through the funding of public programs. Research has shown that extensive provisioning of social services is one of the most effective methods of reducing wealth inequality. Spending on social services and public infrastructure allows those at the bottom of the income distribution to contribute less of their income to daily needs (e.g. transportation), and thus, due to the larger portion of incomes spent on such items, has an outsized effect on the ability of these households to build wealth.

---

[55] Data gathered from The 2019-2020 California State Budget: http://www.ebudget.ca.gov/2019-20/pdf/BudgetSummary/RevenueEstimates.pdf

[56] This estimate is on the same order of magnitude as the estimated total amount of revenue California loses to "profit offshoring" from Philips and Proctor's work on the sales-apportioned corporate income tax. Philips and Proctor estimate this amount is $2.8 billion, i.e. if profit offshoring was completely eliminated, California would receive an additional $2.8 billion in tax revenues. Richard Phillips and Nathan Proctor, "A Simple Fix for a $17 Billion Loophole: How States Can Reclaim Revenue Lost to Tax Havens" (Washington, D.C: Institute for Taxation and Economic Policy, January 17, 2019), https://itep.org/a-simple-fix-for-a-17-billion-loophole/.

Because the data dividend is a limited source of funding we believe that there are certain priority areas into which the government should invest revenues to address the *structural causes* of wealth inequality and to promote *equitable access to AI* such as:

- Debt free higher education
- A "data industrial policy" that closes the digital divide, as described in section V
- The seeding of "baby bonds" or at birth savings instruments for all Californians to close the wealth gap.

### A3.2 - *Individual Per Capita Payments*

Another disbursement option reviewed by this commission is a universal, yearly payment to Californians such as done by the Alaska Oil Fund. This policy has some advantages. For example, it is very simple to implement and will likely be politically popular.

However, we believe that this option should not be considered by the state government. First, these payments are likely to be extremely small. Consider our ballpark lower-bound estimate: If an additional data dividend tax raised just $1.31 billion and this was distributed amongst California's roughly 40 million residents, each resident would receive a check for around $33 annually. If instead distributed to California's roughly 13 million households[57], each household could receive $100. Even if accounting for higher marginal propensities to spend -- or the larger effect that a dollar of income has for a lower income household -- this would not place any dents into the state's income inequality.

Moreover, there are conditions under which such a payment will be regressive. Revenue is highly variable and correlated with macroeconomic cycles. When the economy is in a downturn, it is likely a data tax will fall short of the levels of funding needed to finance payments expected by households. As in Alaska, this might cause legislators to cut vital state programs that help the poor, such as state funding on education, to fund the popular individual payments.

### A3.3 - "*Meritocratic" Payments*

An alternative method of distribution is what we term *meritocratic payments*. These disbursements calculate the relative contribution of each individual to the AI ecosystem,

[57] Population figures taken from the US census bureau https://www.census.gov/quickfacts/CA

i.e. the amount of value that AI models derive from the data generated by each user. The relative contribution may be calculated using model-specific calculations or just by counting a user's number of contributions.

We highly discourage the consideration of this mechanism. First, it is highly complex and depends on a set of models that we believe are not yet reliable for the valuation of individual data contributions. While some academic literature has explored specific data valuation techniques, there is no clear consensus on a method to value data across processes.[58]

Second, we fear that this type of disbursement mechanism may entrench, or even worsen wealth inequality and deepen the digital divide. If we attempt to disaggregate the value of each individual unit of data to an algorithmic process, we are highly likely to worsen *algorithmic bias* -- the fact that algorithms not only reflect but amplify the assumptions of societies underlying biases. Moreover, it may incentivize the worse off to significantly modify their behavior to maximize earnings in ways that *reduce their privacy* and make them targets for predatory business practices.

[58] See for instance Eric Posner and Glenn Wyle, *Radical Markets: Uprooting Capitalism and Democracy for a More Just Society* (Princeton: Princeton University Press, 2018)*;* Arrieta Ibarra, I. et al. *2018.* Should We Treat Data as Labor? Moving Beyond "Free," *American Economic Association Papers & Proceedings.* 1, 1 (2018), 1.

see Ruoxi Jia et al., "Towards Efficient Data Valuation Based on the Shapley Value," *ArXiv:1902.10275 [Cs, Stat]*, December 21, 2019, http://arxiv.org/abs/1902.10275 and Ruoxi Jia et al., "Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms," *Proceedings of the VLDB Endowment* 12, no. 11 (July 1, 2019): 1610–1623, https://doi.org/10.14778/3342263.3342637 for a machine-learning perspective on individual user data. Note that these discussions are not specific to a state-run data dividend, but rather refer to more general data-dividend related ideas about paying for data. Thus, given the nascent nature of the discussion around ata dividends, discussions about data dividend-related ideas have been *implicitly* meritocratic.

## Appendix 4: Mechanics of Data Dependence Taxation

In our recommendations, we have highlighted "data dependence" as the most critical variable for tax design under a "data dividend." We believe that the number of registered users is a metric that simulates a firm's dependence on user data in forming revenues. The design of a data dependence tax must balance the following factors:

- To build in incentives for firms to grow their user base to improve services and develop new technologies but avoid doing so simply to hoard data to undercut competitors (the network effect)
- To establish a fair burden of tax incidence so that larger "tech giants" have a heavier tax burden than smaller firms
- To encourage and regulate firms to make decisions that protect user privacy
- To prevent the manipulation of use counts and tax avoidance

This committee is not recommending specific tax rates. Rather, we believe the best use of our resources is to outline concerns and issues for legislators and future specialists should consider when creating legislation and setting specific rates.

### Tax Assessment

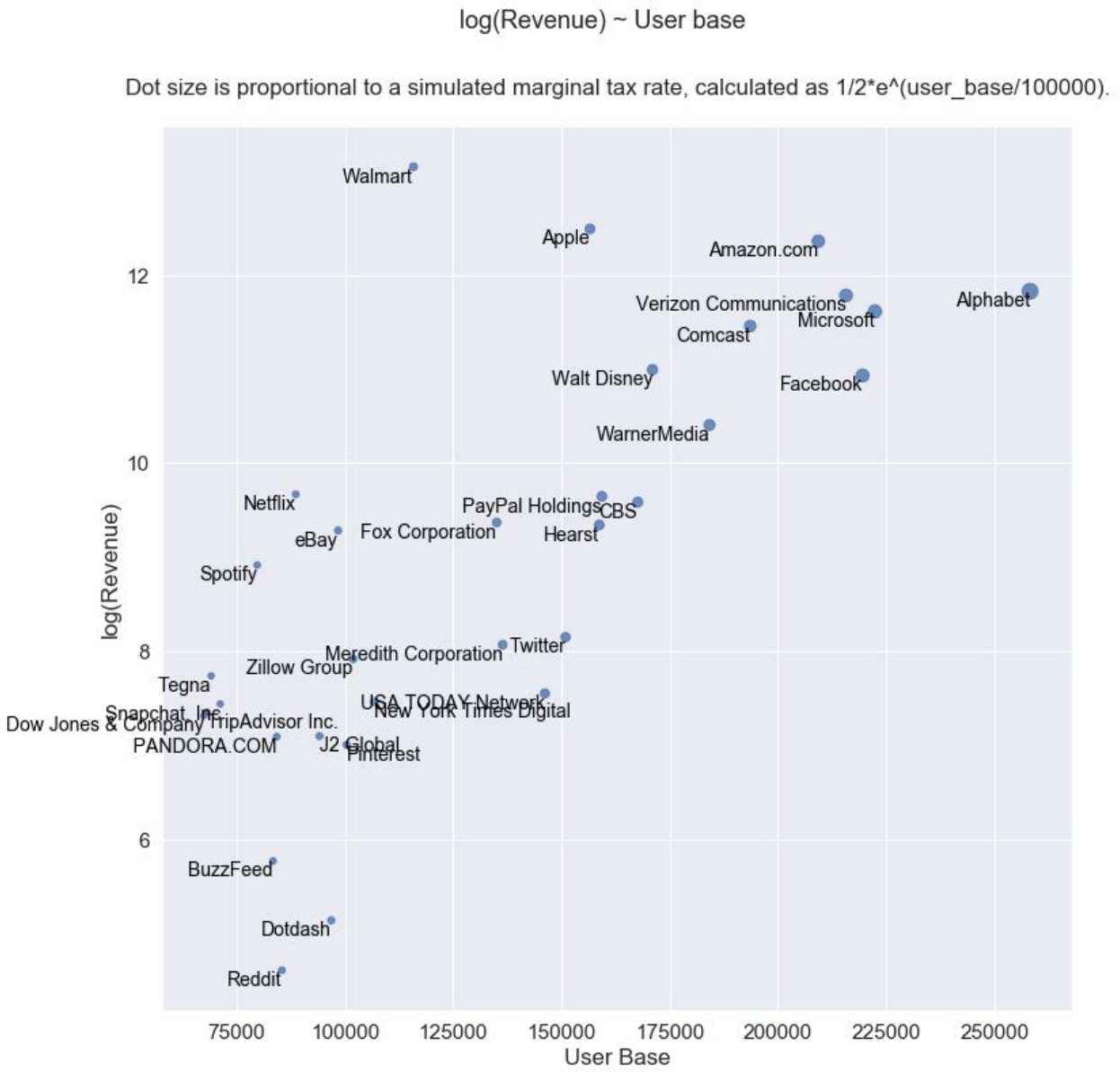
### A4.1 The Advantages of a Marginal, or Tiered Tax

We believe that a marginal tax on users that is assessed against the firm's revenue is the most effective way to raise revenue that captures a firm's exploitation of user data, while also incentivizing firms to grow. A marginal tax would establish a certain tax rate for each set of users after a certain untaxed amount, with each new tax bracket of users taxing at a diminishing rate. This would incentivize firms to continue growing, while taxing larger firms at a higher rate than a smaller firm.

A marginal tax rate on user count helps rectify two elements of data's value that have become relevant to the increasing utility of AI. In a *specific* machine learning process, data has diminishing returns. In other words, each additional unit of data has increasingly little utility for each specific task. However, the same is not true for how data is monetized and collected. Firms do not collect data for specific tasks but rather attempt to establish large sets that can be adapted to multiple tasks by virtue size and breadth. This incentivizes firms to attempt to take advantage of network effects, wherein users contribute their data due to the size of the firm's network, rather than long-term innovation. Firms that build their business model on using data to generate most of their value thus tend to try to offset profits to maximize revenue and borrowing to capture an early, dominant position in a niche and exclude competitors.

By taxing users on the margin, we attempt to bring into line the diminishing returns to individual machine learning applications with firm incentives. Thus, a marginal user tax will help restructure the incentives firms have for growing from monopoly control of markets to specific innovations in machine learning and AI.

Using comScore data, we found a strong, exponential relationship between user counts and revenue. This relatively smooth, curvilinear relationship replicates a distribution that matches the shape of a marginal tax's decreasing returns.




As the figure shows, the firms with the highest tax rate under this schema are large platform enterprises whose primary business is the capturing of user data. We believe

that our exercise may have underestimated the relative tax burden on the tech giants versus other firms because comScore data does not consider firm subdivisions.

### B.1 The Definition of a "User"

What qualifies as a user of a platform is not a straightforward question. For "two-sided" platforms such as companies in the "gig economy" the user can theoretically both be the customer and the worker. For our definition, we believe that the "user" is indeed the customer, and the worker should be understood as a platform's employee.

Moreover, a formally registered user does not cover the extent of a firm's information. In fact, many companies will attempt to register "power users" who might have information on multiple individual's data in her network. Thus, if we only define a user by a metric like a completed "end-user agreement" or registration, we risk undercounting a firm's market power. The CCPA offers a guideline to defining a user:

> a persistent identifier that can be used to recognize a consumer, a family, or a device that is linked to a consumer or family, over time and across different services, including, but not limited to, a device identifier; an Internet Protocol address; cookies, beacons, pixel tags, mobile ad identifiers, or similar technology … or other forms of persistent or probabilistic identifiers that can be used to identify a particular consumer or device.

This is a starting point for an expanded definition of user information as any identifiable member of a dataset, even once that individual identifier was anonymized.

### B.2 Reporting

Data dependence taxes must be assessed as any other tax: reporting via tax documents. Upon formation, the DRB should investigate a procedure by which a firm can easily submit its user counts to tax authorities. The DRB must also establish a method for random auditing of firms that does not compromise user privacy. An added benefit of this procedure is that it will complement the DRB's other efforts to make data more transferable between formats and thus ameliorate network effects implicit in the collection and storage of "big data." This effort is already embedded into the CCPA which states that, upon request, a firm must provide personal information to a user "in a portable and, to the extent technically feasible, in a readily usable format that allows the consumer to transmit this information to another entity without hindrance." Further, the legislature should take special care to make sure firms' subsidiaries are taxed as part of one large organization rather than separately.

# Appendix 5: Toward a Data Industrial Policy through the Data Relations Board

We believe that the fiscal mechanisms of the data dividend should be part of a comprehensive strategy for reforming the institutions that undergird the functioning of the modern digital economy. Revenue raising measures can help offset some of the harms associated with the underlying systems of data exploitation. However, to truly build a democratic digital infrastructure, we will need to proactively create new forms of public-private cooperation. We see the mission of the Data Relations Board (DRB) to create this new environment through what we term a Digital Industrial Policy (DIP). A Californian DIP will work to incentivize firms to act in a prosocial manner and open the benefits of a data economy to a wider range of people and companies. Our suggestions for a DIP mirror the lessons taught by government policies that made California a leader in computing and software in the first place.

## A5.1 Industrial Policy

This appendix details a Digital Industrial Plan (DIP) whose goal is the creation of vital digital public infrastructure in California. Vital public infrastructure includes the following:

- Public internet and mobile connection
- Vital public datasets and public database management
- Public technological platforms for use by individuals and firms
- Merging digital infrastructure of state agencies with local governments
- An industrial plan for A.I. that seeds new firms and returns equity to the state via ownership of shares or other instruments
- Dissemination and commercialization of research or development stage technologies, management techniques, or new platforms

A crucial role of the DRB should be to begin exploring options for the creation of these state investments that can integrate public and private systems to achieve prosocial goals. Our aim here is not to stymie entrepreneurship. Rather, we want to change the environment entrepreneurs operate in and the incentives that they respond to. A DIP's goal is to direct the private sector away from models that rely on the exploitation of network effects to mutually reinforcing actions that keep the ownership of shared infrastructure, including data itself, and let firms derive value from innovation rather than from collecting rents.

This type of policy not only has strong historical precedence but replicates the kinds of measures that were responsible for making Silicon Valley the center of innovation in computing technology. One important reason for the centrality of the Bay Area in software development is that the companies established there were actively cultivated by Federal and State agencies such as the Defense Advanced Research Projects Agency. One of the critical lessons that these programs offer us is that they put the sharing of know-how between competitors at the center of their operations. For example, electronics with DARPA contracts were obligated to share new

technical advances with the larger community of specialists leading to the rapid dissemination of new technology and ideas. By taking the emphasis off patenting new technology and gating it off from competitors, DARPA incentivized its firms to make profit from *value added* innovation rather than the collection of economic rents.[59]

Silicon Valley also benefited from outsized infrastructure investment. Between the 1940s and early 1960s, both strategic and political considerations, California, and Northern California in particular, received a vast amount of funding to build physical infrastructure such as roads, and scientific infrastructure, such as university laboratories. These "greenfield" investments helped make the initial entrepreneurial culture in Silicon Valley far more open and flexible than competing research clusters in established centers of science and engineering such as Boston' Route 128 corridors. This set of networked infrastructure investment and close cooperation with local government made the Valley's firms uniquely open and knowledge was widely shared, again incentivizing new innovation rather than the hoarding of valuable information.

New research has shown that since the mid-1990s, the openness that dominated Silicon Valley has actually decreased. As we have already documented in this report, the headlines hide a stunning decline not only in entrepreneurship but the value added of tech firms. One reason for this problem is that neither the state nor the federal government has maintained a systemic policy to build entrepreneurial infrastructure. While there has been continued, yet still insufficient, investment into human capital, we have underinvested into public capital.[60]

This is why we recommend that the DRB begin to formulate a plan that will allow for the construction of new data ecosystems across California. We are especially interested in using state policy to close the gaps between access to information technology via increased training in schools, community centers, and libraries and actual construction of new infrastructure such as public wifi and rural broadband as well as state-operated platforms that can allow new firms to find and share important technical data. The next section of this appendix will outline some ways that the state can make this a reality through incentives built into our proposed tax mechanism.

### 5.2 Creating a Public Data Trust

The most vexing economic problem offered by the data economy is that its value grows as the size of the data set increases. This incentivizes firms to become platforms, or monopolies, that dominate a particular space by using their first-mover advantage to continuously increase the size of their data pool compared to new entrants. In other cases of such "natural monopolies" governments have transformed firms into utilities.

---

[59] Mariana Mazzucato and Gregor Semieniuk, "Public Financing of Innovation: New Questions," *Oxford Review of Economic Policy* 33, no. 1 (January 1, 2017): 24–48.

[60] Peter Cook, World Turned Upside Down: Entrepreneurial Decline, Its Reluctant Myths and Troubling Realities. *J. Open Innov. Technol. Mark. Complex*. 2019, 5(2), 22;.https://www.mdpi.com/2199-8531/5/2/22

We believe that the best way to replicate this arrangement in the context of the commercialization of data is for the state to establish Public Data Trusts (PDTs). Public Data Trusts are state administered data banks that collect public data and make it commercially available, either for free or for a fee, to any qualified firm. Though a new idea, PDTs are beginning to find a place in the governance of data generated by public utilities and "smart cities." PDTs not only help create responsible and equitable commercialization of data but also act as fiduciaries that protect the privacy of the public. In fact, we believe that if leveraged, this last feature of a data trust can be levered to create standards for privately gathered data by creating incentives to integrate procedures for these trusts to be the nodes of public-private networks.

### B.3 Incorporating Public Data Trusts into A Data Industrial Policy

We recommend that one of the first tasks of the DRB should be to begin establishing California-wide PDTs. To do this, it would conduct a comprehensive survey of all data collected by the state government and municipalities, including dynamic data that is currently not catalogued by the California Open Data Portal. It will also examine the latest research in authentication technology to establish a protocol to transmit and store this data. Once this research has been conducted, the DRB would build the PDTs and require all firms interested in working with its data to adopt a common network standard for transmission. In doing so, the PDT will incentivize industry to begin addressing problems of portability that are at the heart of initiatives such as the CCPA and may produce greater results for de-monopolizing data in the future.

A strong PDT or collection of trusts will open a vista onto more aggressive steps to creating an open and democratic data economy and to increase the productivity of California's entrepreneurs by acting as an open depository for publicly important data. PTDs should become the centers of a larger innovative ecology that restores the spirit of cooperation that made Silicon Valley an economic engine that grew the overall economy.

To do this, the next phase of PDT development will be to begin to create incentives for private firms to share their data with other entrepreneurs. To begin this process, we argue that firms that agree to work with the PDT to publicize their data should be given a tax break on their data dependence tax. To prevent the problem of firms contributing unusable or not valuable data, the DRB must review each request for sharing against a set of guidelines that also emphasize safety and privacy. As well, the DRB should be given the power to issue contracts to private firms for data that its specialists and industry advisors deem necessary for the public good or the development of critical new technologies.

Such actions would help firms to focus on maximizing value-added innovation over establishing monopolies and circumventing labor protections. Moreover, it will condition firms to view data as a common pool resource and encourage businesses to

return to cooperative practices that once made Silicon Valley an agent of highly productive, shared growth.

## Appendix 6: Data Dividends and Data Cooperatives

Data cooperatives (alternately known as data coalitions, data trusts, or data unions) are at the heart of a related, complementary proposal on data regulation, available here. In summary, that proposal argues that a new class of regulated business entity should be established at law in order to facilitate fair and efficient bargaining over data. These new entities would have special obligations to maintain independence from data-using businesses, to refrain from permanent data alienation, and to uphold strict fiduciary obligations to their members, who would assign to them the exclusive right to bargain over certain categories of their data. Thus, Data Cooperatives would serve as collective bargaining entities and necessary counterparties for data-using businesses who wish to use data generated by Data Cooperatives' members.

The importance of data cooperatives flows from the fact that in the absence of collective bargaining, markets for data fail. The data of individuals implicates and includes information pertaining to other individuals so that conventional data transactions between individuals and platform businesses always undermine the negotiating interests of third parties. This results in platform businesses capturing the increasing returns to the data of numerous individuals, an asymmetry which worsens with the scale of the platform. In such conditions, individuals have inefficiently low bargaining power with which to vindicate not only their economic interests but also their interests in keeping their data private or influencing its downstream uses.

Permitting individuals to form medium-to-large collective bargaining entities would dramatically increase their exposure to the upside of the data economy as well as their ability to control the fate of their information. However, legal and regulatory change is necessary to facilitate this collective bargaining ecosystem. In addition to the features sketched above, the regulation would need to establish a dispute resolution forum enabling Data Cooperatives to make injunctive or damages claims against other Data Cooperatives whose transactions wrongfully undermine the interests of non-members.

Once established, Data Cooperatives would address many of the same problems that a Data Dividend seeks to ameliorate. For example, users represented as part of a Data Cooperative would not be entirely unable to negotiate a fair share in the network value that their data helps to generate. Therefore, we envision a scenario in which members of data cooperatives count less, or not at all, for the purposes of a data dependence tax. By combining a data dependence tax with a data cooperative ecosystem, businesses would be able to grow to indefinitely network size when efficient--but would not be able to do so without either paying a substantial tax, or bargaining (on comparatively even footing) with users represented by Data Cooperatives.