BEA Working Paper Series, WP2024-1b

Private Funding of "Free" Data: A Theoretical Framework

Author Rachel Soloveichik, U.S. Bureau of Economic Analysis

Contact Rachel.Soloveichik@bea.gov

Date April 2024

Abstract One might think that the private value of data equals the net present value of future data sales revenue. In fact, this paper demonstrates that even data with zero sales revenue can still have substantial private value. This paper develops a theoretical framework in which private data can either be funded through data sales or funded by subsidies from the owner of a complementary asset. For example, maps of tropical beaches can either be funded by selling them to individuals who have already booked a local hotel room or funded by travel agencies who use "free" maps in their marketing campaigns. This paper then solves that theoretical framework, calculates the value of data by funding mechanism, and identifies parameter regions for which free data are more valuable than sold data. Free data are particularly likely to be more valuable when data are complementary to other data (Coyle 2022), when data are specific to one capital asset, or when piracy reduces data sales revenue.

This paper illustrates the importance of free data with a back-of-the-envelope calculation. To start out, the paper reviews four previous case studies which together studied $1.8 trillion of free data creation in 2017 (Soloveichik 2023) (Soloveichik 2024) (Sveikauskas et al. 2023). This paper then uses those case studies, existing input-output tables and occupational employment to extrapolate that total private creation of free data in the United States was $6.6 trillion in 2017. In that same year, including free data in the economic statistics raises measured gross domestic product by more than 20 percent and raises measured household production by more than 100 percent.

Keywords Data, capital, free, intangibles, investment

JEL codes D12, E01, and G14

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Introduction

The United States economy depends on data. Banks check credit scores before approving loans, insurers check risk factors and claims history before setting premium rates, doctors check lab results before making diagnoses or prescribing treatments, and schools check standardized test scores and grades before admitting students or recommending classes. In addition, employers check performance reviews before promoting workers and governments check tax forms before approving benefit requests. Finally, households check social media reviews before buying from a business or socializing with an individual.

Data are rarely sold in an arms-length transaction. On the one hand, some business data are kept in-house and are used by their owner to gain a competitive advantage. Previous national accounting papers studying data focused on these in-house data (Coyle and Li 2021) (Mitchell et al. 2022). On the other hand, some business data and most consumer data are shared with all authorized users for "free." For example, hotels might post maps of local beaches and hiking trails on a website for the world to see or consumer might report their credit score on a loan application website for all banks in the area to see. This paper focuses on those free data.

This paper first develops theoretical frameworks in which private data can either be funded through data sales or funded by subsidies from the owner of a complementary asset. This paper then solves those theoretical frameworks and identifies plausible parameter regions for which free data are more valuable than sold data. This paper then illustrates the importance of the free data studied in these theoretical frameworks by presenting a back-of-the-envelope calculation that estimates total private free data creation was $6.6 trillion in 2017. National accountants have started a discussion on the value of data, and the next edition of the official guidelines for national accounting may recommend tracking data as intangible capital assets (Rassier et al. 2019) (Eurostat 2020). When free data are tracked according to the recommendations discussed, the U.S. Bureau of Economic Analysis' (BEA's) 2017 economic statistics show a more than 20 percent increase to measured gross domestic product (GDP) and a more than 100 percent increase to measured household production.

At first glance, the $6.6 trillion value for data creation estimated using the back-of-the-envelope calculation seems inconsistent with previous national accounting papers that found smaller values for data investment (Statistics Canada 2019) (Coyle 2022) (Calderon and Rassier 2022) (Mitchell et al. 2022). In fact, the different results can be explained by the different study focuses. Previous research only

studied complex files that are managed by skilled data experts. This paper studies both those complex files and simple files that can be managed by ordinary workers. For example, previous research might study complex online search histories which are used to automatically target digital ads for tropical vacations. This paper studies both those complex online search histories and simple mailing lists of people who requested more information on tropical vacations. As a result, it is not surprising that this paper's estimates of data creation are substantially larger than other papers' estimates.

To be clear, $6.6 trillion value for free data calculated in this paper is only a back-of-the-envelope estimate that is based on four case studies and an extrapolation. Many more case studies need to be done before free data could be measured precisely enough to be included in BEA's published economic statistics. Nevertheless, the back-of-the-envelope estimate demonstrates that free data are large enough to change measured GDP and measured household production noticeably. These results suggest that free data deserve much more research going forward.

This paper is divided into five sections. The first section discusses the unique features of data. The second section develops a simple theoretical framework with fixed output prices, only one capital asset, and only one data type. The second section solves that simple theoretical framework for selected parameter regions. The third section extends that simple theoretical framework to allow for multiple capital assets and multiple data types. The third section then solves that extended theoretical framework for selected parameter regions. The fourth section relaxes the assumption of fixed prices to allow for output prices that depend on the quantity of output. The fourth section then solves that relaxed theoretical framework for selected parameter regions. Taken together, the theoretical frameworks in sections two to four demonstrate that every sector of the economy, under a wide range of parameter regions, is likely to produce free data. The fifth section illustrates the importance of free data by calculating that the total value of all private free data creation was $6.6 trillion in 2017. That section then describes how BEA's economic statistics might change if all types of private free data were tracked as intangible capital assets.

1. Data Definitions and Features

Defining Data

This paper defines data as information that can be copied easily and used by multiple entities simultaneously.[1] Previous national accounting papers have studied complex digital data (Statistics Canada 2019) (Coyle 2022) (Calderon and Rassier 2022) (Mitchell et al. 2022); this paper broadens the focus to include all types of data. Some types of data are simple enough to be stored in a small file, and other types of data are so complex that they can only be stored on supercomputers. Data are sometimes stored on a physical object like a CD or a piece of paper and are sometimes stored entirely electronically. Modern data are typically formatted digitally, but data can also be formatted with alphabetic characters, with DNA codons, or even with analog sound. This paper's discussion focuses on data which are used repeatedly for more than one year because those data are long-lived enough to be tracked as produced capital assets (Rassier et al. 2019) (Eurostat 2020). However, the theoretical frameworks developed in this paper also apply to data which are used repeatedly for less than one year. Those short-lived data might be tracked as intermediate inputs.

Data describing an asset are separate from the asset itself. For example, maps of the wildlife on tropical beaches are both conceptually and legally separate from both the wildlife and the tropical beaches. In national accounting terms, maps are produced intangible assets while both wildlife and tropical beaches are non-produced natural resources. Maps of a hotel are produced intangible assets while the hotel itself is a produced structure. Maps showing restaurants with talented cooks are produced intangible assets while cooking talent is a type of human capital owned by the chefs themselves. Finally, an index of all the maps mentioned above would be considered metadata (data about data). This paper studies the value of data and metadata, but not the value of other assets.

Free data are distributed at a price close to zero. To be clear, the free data are not always distributed at a precisely zero price. For example, warehouse clubs might restrict their tropical island maps to individuals who paid a membership fee to enter the store. Conversely, a timeshare association might encourage people to download tropical island maps by providing a desirable coupon to anyone who does so. Both of those prices are close enough to zero to be treated as zero in a theoretical framework.

1. Cryptocurrency is an example of information which can be copied easily but not used by multiple entities simultaneously.

This paper values private data based on the revenue associated with the data. Sold data are valued based on their sales revenue. Free data are valued based on their value to the owner of a complementary capital asset. Free data which are complementary to a non-produced beach can provide value through additional tourist visits, higher beach entry fees, and so on. Free data which are complementary to a produced hotel structure can provide value through more hotel guests, higher room rental rates, hotel guests who require less employee help to find their room, and so on. Free data which are complementary to a chef's cooking talent can provide value through more job offers, higher wages, better kitchen conditions, and so on. To be clear, these private data values do not include any positive externalities—and therefore are often lower than the social value of data studied in other research (Coyle 2022).

Unique Features of Data

The minimal cost of copying is the single most important data feature for this paper. In the extreme, digital data stored on servers can be duplicated almost instantly at virtually zero cost. However, other data formats also have a copying cost far below their original creation cost. Files stored on CDs can be burned onto another CD, text stored on paper can be photocopied, DNA strands stored on seeds can be replicated with planting, and oral traditions can be retold. Because copying costs are already minimal, technologies which make copying even easier have little impact on overall data costs. For example, the hotel production function does not change much when its marketing department switches from paper maps that can be photocopied to electronic maps that can be instantly duplicated. Accordingly, this paper does not split data by either the format they are recorded in or the medium they are stored on. Instead, it assumes that the data copying costs are always minimal regardless of format or medium.

The difficulty of excluding potential users is another important feature of data for this paper. As the old saying goes, "three can keep a secret if two of them are dead" (Franklin 1735). Very sensitive data can sometimes be protected with sophisticated cryptographic techniques (de Groot 2022), but those sophisticated cryptographic techniques are generally not feasible for ordinary data users or ordinary data sellers. Hence, all firms which buy data can provide data as secondary output. Furthermore, employees who access data on-the-job can also become data providers on their own account. The remainder of the discussion calls either firms or individuals who provide copied data "pirates." Regardless of whether the pirates sell data or give it for free, they compete with the original data seller and drive down the price of data. Accordingly, original data sellers face a choice of either restricting access (Drolet 2016), setting prices low enough to compete with pirates, or accepting the fact that many potential data buyers will choose pirated data rather than the original seller's data.

To be clear, acceptance of piracy does not mean zero privacy. Data providers are often required by contracts or regulations to only distribute data to authorized users. These contracts or regulations are relatively easy to enforce if authorized users are charged a very low fee, the set of authorized users includes every entity with a genuine need to know, and entities who access the data without a genuine need to know are punished. For example, a doctor might record medical data in a patient's chart. Those medical data are vulnerable to piracy in the sense that the original doctor is not allowed to charge a large fee for data access. Instead, a patient is only required to pay copying costs to access their own chart or send the chart to a new doctor (Baker et al. 2015). Nevertheless, medical data are protected by the law and doctors cannot give patient data to curious journalists or nosey family members without authorization from the patient.[2] For simplicity, this paper assumes that data security is always sufficient to protect genuine privacy concerns (Acquisti et al. 2016). Because the marginal cost of copying data is low and data security is always sufficient to protect privacy, social welfare is maximized when data are widely shared between authorized users (Coyle 2022) (Jones and Tonetti 2020).

2. Simple Theoretical Framework with Fixed Output Prices, One Capital Asset, and One Data Type

This paper explores general ideas about the value of data under a variety of possible funding methods. In order to illustrate those general ideas, this paper develops several related theoretical frameworks in which data are sold and used. This section starts out with a very simple economy and the later sections move on to slightly more complicated economies.

The simple economy starts out with a physical capital owner who rents their capital asset at rental rate of r. In order to facilitate mathematical solutions, this theoretical framework assumes that r is a fixed constant. In addition, there is a data owner who grants access to their intangible data asset for a price of a per data unit. The data owner is assumed to set their access rate, a, at a level which maximizes their profits. Finally, there are n separate firms which buy capital services and data access. These n firms all have constant elasticity of substitution (CES) production functions which combine capital services and data access to produce revenue. The n firms differ in their skill at using data, with firm 1's skill designated as s_1, firm 2's skilled designated as s_2, etc. For discussion purposes, the firms are ordered so that $s_1 \geq s_2 \geq \ldots \geq s_n$. Both the total supply of capital, K, and the total supply of data, D, may be low enough that firms are constrained to rent less capital or access less data than they would otherwise. Capital is

2. Doctors providing emergency medical treatment can access data without prior authorization (Page et al. 2020).

rival, so the total quantity of capital rented by all n firms must be less than or equal to K. In contrast, data are nonrival, so up to n firms can use the same amount of data simultaneously. This model assumes that the capital is split proportionally to firm demand in case of a shortage[3] and each firm is given the maximum quantity of data, D, in case of a shortage. The equations designate the quantity of capital rented by firm i as k_i and the quantity of data accessed by firm i as d_i. Finally, the price of firm output is designated as p. For any i between 1 and n, the revenue and profits of firm i can be written as:

(1) Output of firm i = $Y_i = (\alpha k_i^{\rho} + (1-\alpha) * (d_i s_i)^{\rho})^{(1/\rho)}$

(2) Profit of firm i = $\Pi_i = p * (\alpha k_i^{\rho} + (1-\alpha) * (d_i s_i)^{\rho})^{(1/\rho)} - r * k_i - a * d_i$

This simple economy can be illustrated with the example of a tropical island. The island has a hotel that rents out rooms to travel agencies at a fixed price, r, per room. In addition, there is a map owner who shares their maps for an access rate, a, per map. For example, a walking map might show recommended clothing for hikes in each season or an ecology map might show good birdwatching areas. Finally, there are n separate travel agencies which rent blocks of hotel rooms and then rent individual hotel rooms to tourists. These n travel agencies differ in their skill renting hotel rooms to tourists, but they all rent hotel rooms and use maps to help tourists plan vacations.

Capital Usage and Data Creation when an Economic Planner Controls All Outcomes

Previous pioneering research studied both the social value of data and the private value of data (Coyle 2022). In order to compare this paper's results with that research, this paper calculates the social value of data along with the private value of data under a variety of funding mechanisms. In particular, this subsection considers a world in which data are distributed by an economic planner whose goal is to maximize total value-added in the economy.[4] Given the distribution of data, each firm then decides how much capital to rent.[5] In the tropical island example, the economic planner might be a local government which wants to maximize the region's GDP or a trade association of hotels and travel agencies.

The economic planner's problem is trivial. Data sharing is assumed to have zero marginal cost. Therefore, the economic planner distributes the maximum amount of data, D, to every firm. The

3. If capital prices were flexible, then the capital owner would respond to a shortage by raising rents.

4. Business value-added is not necessarily identical to welfare, and so welfare is not necessarily maximized.

5. The economic planner does not decide capital allocations, and total value-added may be lower than the maximum possible if the capital owner sets a rental rate, r, that is too high.

economic planner then finances data creation by charging each firm a fixed data access fee that does not vary with the quantity of data used. Each firm's profit function can be expressed as:

(3) $\Pi_i = p*(\alpha k_i^{\rho}+(1-\alpha)*(D*s_i)^{\rho})^{(1/\rho)} - rk_i -$ fixed access fee for $data_i$

The only decision facing firm i is how much capital to rent. If capital is not in short supply, then that decision can be solved simply by taking the derivative and finding out when the marginal profit contribution of capital is zero:

(4) $d\Pi_i/dk_i = p*(\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)-1} *\alpha k_i^{\rho-1} - r = 0 \rightarrow (\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1-\rho)/\rho} *\alpha k_i^{\rho-1} = r \rightarrow$

$(\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1-\rho)/\rho} = (r/\alpha)*k_i^{1-\rho} \rightarrow (\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho}) = (r/\alpha)^{\rho/(1-\rho)} * k_i^{\rho} \rightarrow$

$(1-\alpha)*(Ds_i)^{\rho} = [(r/\alpha)^{\rho/(1-\rho)} - \alpha]* k_i^{\rho} \rightarrow$ **$k_i = p*(1-\alpha)^{(1/\rho)}(Ds_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha]^{(-1/\rho)}$**

(5) Total value added for firms $=\sum p*(\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)} - r*k_i$ - fixed fee for $data_i$

Value added for capital owner $=\sum r*k_i$

Value added for data owner $=\sum$ fixed fee for $data_i$

Total value added for all businesses $=\sum p*(\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)} =$

$=\sum p*(\alpha\{(1-\alpha)^{(1/\rho)}(Ds_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}\}^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)}$

$= p*D(s_1+ s_1+..+ s_n)*(1-\alpha)^{(1/\rho)}(\alpha*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1}+1)^{(1/\rho)}$

If capital is in short supply, it is rationed to firms in proportion to their demand:

(6) Demand for firm 1 $=p*(1-\alpha)^{(1/\rho)}(Ds_1)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$;

Demand for firm 2 $=p*(1-\alpha)^{(1/\rho)}(Ds_2)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$;

Demand for firms 1 to i $= p*\sum(1-\alpha)^{(1/\rho)}(Ds_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)} \rightarrow$

$k_i=p*(1-\alpha)^{(1/\rho)}(Ds_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}/\{(1-\alpha)^{(1/\rho)}[D(s_1+ s_1+..+ s_n)]*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)} \rightarrow$

$k_i = s_iK/(s_1+ s_2+ ...+ s_n)$

(7) Total value added for firms $=p*\sum(\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)} - r*k_i$ - fixed access fee for $data_i$

Value added for capital owner $=\sum r*k_i$

Value added for data owner $=\sum$ fixed fee for $data_i$

Total value added for all businesses $=p*\sum(\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)} =$

$$=p^*\sum(\alpha[s_iK/(s_1+ s_2+ \ldots+ s_n)]^{\rho} +(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)} =$$

$$p^*(s_1+ s_1+..+ s_n)^* (\alpha[K/(s_1+ s_2+ \ldots+ s_n)]^{\rho} +(1-\alpha)*D^{\rho})^{(1/\rho)}$$

Equations (5) and (7) are difficult to solve by hand, but they are trivial to solve with a computer given a full set of parameters: p, r, K, s_1 to s_n, α, ρ, D. Figure 1 in a later subsection shows the economic benefits of data for one selected parameter region. For now, the theoretical framework discusses some general patterns to the solution of equations (5) and (7). Most obviously, the economic planner's value of data is higher when the n firms have more total skill, $(s_1+ s_2+ \ldots+ s_n)$, using data. In addition, the economic planner's value of data is higher when the stock of complementary capital, K, is large enough that capital is not in short supply.

Capital Usage and Data Sales Without Piracy

This subsection considers a world in which decision-makers maximize individual profits without considering either positive or negative externalities. The rental rate for capital, r, is a fixed constant that does not depend on economic conditions. Decisions are made in three sequential steps. First, the data owner sets the data access rate, a. Second, the firms decide how much data to access. Finally, the firms decide how much capital to rent. This subsection solves by induction. First, it calculates how much capital each firm rents given their data access and the capital rental rate, r. Next, it calculates how much data each of the n firms access given the amount of capital each expects to rent and the data access price of a. Finally, it calculates the revenue-maximizing data access rate, a.

In the tropical island example, the map owner first picks an access rate for their maps that maximizes their revenue. Second, travel agencies decide how many maps to rent. Finally, each of the n travel agencies decides how many rooms to rent. The n travel agencies combine maps with rooms to sell hotel rooms to individual tourists.

Capital allocation when capital is not in short supply to get the same solution is determined by the same profit function as equation (3) and can be solved to get a similar solution as equation (4):

(8) $\Pi_i = p^*(\alpha k_i^{\rho}+(1-\alpha)*(d_i*s_i)^{\rho})^{(1/\rho)} - rk_i - ad_i$

(9) $d\Pi_i/dk_i = p^*(\alpha k_i^{\rho}+(1-\alpha)*(d_i*s_i)^{\rho})^{(1/\rho)-1}*\alpha k_i^{\rho-1} - r = 0 \rightarrow k_i = p^*(1-\alpha)^{(1/\rho)}(d_is_i)*[(r/\alpha)^{\rho/(1-\rho)} -\alpha)]^{(-1/\rho)}$

(10) $\Pi_i = p^*(\alpha\{(1-\alpha)^{(1/\rho)}(d_is_i)*[(r/\alpha)^{\rho/(1-\rho)} -\alpha)]^{(-1/\rho)}\}^{\rho}+(1-\alpha)*(d_i*s_i)^{\rho})^{(1/\rho)}$

$$- \; r\{p^*(1-\alpha)^{(1/\rho)}(d_i s_i)^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(1/\rho)}\} - ad_i =$$

$$p^*(\alpha(1-\alpha)(d_i s_i)^\rho{}^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+(1-\alpha)^*(d_i{}^*s_i)^\rho)^{(1/\rho)} - \; r^*p^*(1-\alpha)^{(1/\rho)}(d_i s_i)^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)} - a_I =$$

$$d_i \; ^*(p^*\alpha(1-\alpha)s_i{}^\rho{}^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+(1-\alpha)^*s_i{}^\rho)^{(1/\rho)} - p^*r(1-\alpha)^{(1/\rho)}{}^*s_i{}^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)} - a\}$$

Firm i's profits are linear with d_i and therefore the marginal profit from data is constant throughout the entire choice set. Accordingly, firms pick the corner solution of d=0 when the marginal profit is negative and firms pick the corner solution of d=D if the marginal profit is positive. For mathematical simplicity, this paper assumes that the data seller sets the access rate high enough that the marginal firm is precisely indifferent between buying data or not:

(11) Revenue of the data seller = $\sum ad_i$; d_i =D if

$$s_i[p^*\alpha(1-\alpha)^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+(1-\alpha)]^{(1/\rho)} - p^*r(1-\alpha)^{(1/\rho)}{}^*s_i{}^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)} \geq a$$

0 otherwise.

Revenue of the data seller if they sell to i firms = i*aD =

$$i^*\; p^*D^*s_i\{(\; \alpha(1-\alpha)^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+(1-\alpha))^{(1/\rho)} - r(1-\alpha)^{(1/\rho)}{}^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}\}$$

(12) Revenue of the capital owner = $\sum rk_i$ $= r^*i^*D^*p^*(1-\alpha)^{(1/\rho)}(s_1+ s_2 +...+s_i)^*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}$

Equation (11) is long and would be difficult to solve by hand, but all the terms in the equation are constants and so it is straightforward to solve with a computer. In many parameter regions, the data seller faces the standard monopoly trade-off between high prices and high volume. In those parameter regions, some firms will be priced out of buying data and therefore total data usage and total output are less than they would be otherwise. The deadweight loss associated with monopoly is a well-known result that will not be discussed further. Instead, this paper will focus on the fact that the revenue received by the data seller in equation (11) is less than the economic planner's value shown in equation (5). In other words, the private value of data is lower than the economic planner's value.

Equation (12) shows an externality from the data seller to the capital owner. When the data access rate is high, fewer firms choose to buy data access and therefore fewer firms rent capital. Due to this externality, the capital owner prefers that the data access rate be lower.[6] Similarly, the capital owner

6. There is also a converse externality where the data owner prefers that capital rental rates be lower. This paper assumes that capital rental rates are fixed, and therefore does not explore that externality.

prefers that the quantity of data available, D, is larger. These two externalities are not important if data and capital are weak complements, ρ is close to 1, but very important if data and capital are strong complements, ρ is close to $-\infty$.

Capital allocation when capital is in short supply is determined by the same profit function as equation (3) and can be solved to get a similar solution as equation (6):

(13) $d_i s_i * K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n) = k_i$

(14) $\Pi_i = p*(\alpha\{(d_i s_i)*K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n)\}^{\rho} + (1-\alpha)*(d_i * s_i)^{\rho})^{(1/\rho)}$

$- \; r(d_i s_i)*K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n) \; - a d_i =$

$[p*\alpha(d_i s_i)^{\rho} * K^{\rho}/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n)^{\rho} + (1-\alpha)*(d_i * s_i)^{\rho}]^{(1/\rho)} - \; r(d_i s_i)*K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n) - a d_i =$

$d_i * \{p*[\alpha s_i^{\rho} * K^{\rho}/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n)^{\rho} + (1-\alpha)*s_i^{\rho}]^{(1/\rho)} - \; r s_i * K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n) - a\}$

Equation (14) is not quite linear with d_i because the term $K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n)$ depends on d_i. When there are only a few similar firms buying data there can be multiple Nash equilibria. But when there are many firms, $K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n)$ is almost constant and firms are very likely to be at a corner solution. This paper solves for equation (14) as if it were linear with d_i and assumes that the data seller sets their price high enough that the marginal firm is precisely indifferent between buying data or not:

(15) Revenue of the data seller = $\sum a d_i$; $d_i = D$ if

$s_i\{(\alpha K^{\rho}/(D s_1 + D s_2 + \ldots + D s_i)^{\rho} + (1-\alpha))^{(1/\rho)} - \; rK/(D s_1 + D s_2 + \ldots + D s_i)\} \geq a$

0 otherwise.

Revenue of the data seller if they sell to i firms = $i*aD =$

$i* \; p*D*s_i\{[\alpha(1-\alpha)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)]^{(1/\rho)} - \; r(1-\alpha)^{(1/\rho)} * [(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}\}$

(16) Revenue of the capital owner = $\sum r k_i = rK$

Just like in a world where capital is not in short supply, the data seller earns less revenue than the economic planner's value of data shown in equation (7).

When capital is in short supply, the full supply of capital is purchased regardless of what price the data seller sets or what quantity of data, D, is available.[7] Accordingly, there are no externalities from the data seller to the capital owner. But there are now externalities between the n firms which use capital and data. By assumption, capital is rationed in proportion to each firm's demand. Each firm's demand for capital increases when they use more data or are more skilled at using the data they have. Accordingly, each firm would prefer that its rivals not use data or be less skilled at using data.

One might think that flexible capital prices could solve the externalities shown in equations (12) and (16). In fact, flexible prices simply transform quantity shifts into price shifts. In a model where capital prices are set such that supply equals demand,[8] then there are j firms which choose to buy the full quantity of data, D, and n-j firms which choose not to buy any data. The rental rate of capital is determined by the equation:

$$(17)\ r = \alpha * \{K^{-\rho}/[(s_1 + s_2 + \ldots s_j) * (1-\alpha)^{(1/\rho)} * D]^{-\rho} + \alpha\}^{(1-\rho)/\rho}$$

The equation above implies two separate pecuniary externalities. First, the capital owner receives a higher rental rate for capital when the supply of data, D, is larger. Second, each of the j capital using firms pay higher rental rates for capital when more firms buy data. Hence, the result that data create externalities is robust to relaxing the assumption of fixed capital prices.

Capital Usage and Data Sales With Piracy

This subsection once again considers a world in which each firm makes decisions to maximize their individual profits with one difference: a new decision about piracy. First, the data owner decides what price to charge for their data. Second, the firms decide how much data to buy. Third, all firms which buy data decide whether to resell pirated copies of the data. Finally, firms decide how much capital to rent.

In the tropical island example, the mapmaker sets their access price as before. But now every travel agency has a photocopier and can resell copies of whatever maps they accessed whenever they want. Once maps have been obtained, either from the original data seller or the other travel agencies, each of the n travel agencies decides how many rooms to rent. The n travel agencies then combine maps with purchased blocks of rooms to sell individual hotel rooms to tourists.

---

7. The invariance only holds locally. Large changes can reduce the demand for capital below the supply.

8. Monopolist capital owners can sometimes increase profit by withholding capital and thereby raising rental rates.

Capital allocation when capital is not in short supply can be solved using a similar profit function as equation (3) and a similar marginal profit function as equation (4):

(18) $k_i = p*(1-\alpha)^{(1/\rho)}(d_i s_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$

The third step, in which firms decide whether to pirate data, can be solved with simple logic. Data can be copied at zero marginal cost, and so every data user has the potential to sell unlimited copies of their data for any price they choose. This paper assumes that multiple firms which hold data compete to resell pirated data in a Bertrand model and therefore drive the price of pirated data down to zero. Given this piracy behavior, at most two firms pay a significant price for their data and the remainder of firms get it for nearly nothing. Each firm knows that it is better to buy pirated data last, so the n firms play a complex game to see which firms will buy data at a positive price. Such a complex game has many Nash equilibria. For now, this paper analyzes the equilibrium which yields the most revenue to the data seller. In that equilibrium, firm 1 accesses the maximum amount of data, D, first at a price a_1, then resells the maximum amount of data access, D, to firm 2 at a price a_2, and then firms 3 to n access D at a zero price. Similar to equation (11) we can solve to get the maximum possible p_2:

(19) $p*[\alpha(1-\alpha)s_2^{\rho}*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)*s_2^{\rho}]^{(1/\rho)} - r(1-\alpha)^{(-1/\rho)}*s_2*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)} = a_2$

Firm 1 knows that it will earn both its standard revenue from production and also a_2*D from reselling pirated data to firm 2. Accordingly, we can solve for a_1 similarly to (11):

(20) $a_1 = a_2*s_2[p*\alpha(1-\alpha)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)]^{(1/\rho)} - s_2*p*r(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)} + s_1[\alpha(1-\alpha)[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)]^{(1/\rho)} - s_1*r(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$

(21) Data seller revenue = $(a_1 + a_2)*D =$

$$(s_1 + s_2)*D*[p*\alpha(1-\alpha)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)]^{(1/\rho)} - p*r(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$$

(22) Capital owner revenue $\sum rk_i = r*D*p*(1-\alpha)^{(1/\rho)}(s_1 + s_2 + \ldots + s_n)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$

The critical difference between equation (21) and equation (11) is the number of firms. When piracy is a possibility, at most two firms pay a positive price for data. In contrast, a potentially large number of firms pay a positive price for data in a world without piracy. The result is much lower revenue for the data creator in most circumstances. However, there are a few parameter regions where piracy acts as an

indirect form of price discrimination and thereby raises data seller revenue. Interestingly, equation (22) shows that there are now no externalities between data access rates and capital owner revenue. Intuitively, piracy means that all n firms access data regardless of the data access rate charged to either firm 1 or firm 2. Accordingly, demand for capital and capital owner revenue are both maximized.

Capital allocation when capital is in short supply can be solved similarly to equation (12):

(23) $k_i = d_i s_i K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n)$

Just like the earlier piracy scenario, at most two firms pay a significant price for their data and there are many Nash equilibria. This paper once again analyzes the equilibrium which yields the most revenue to the data seller. In that equilibrium, firm 1 buys the maximum amount of data access, D, first at a price a_1, then resells the $d_2<D$, to firm 2 at a price a_2, and then firms 3 to n access d_2 at a zero price. Similar to equation (11) we can solve to get the maximum possible a_2:

(24) $[p^*\alpha(d_2 s_2 K/(Ds_1 + d_2 s_2 + \ldots + d_n s_n))^{\rho} + (d_2 s_2)^{\rho}]^{(1/\rho)} - r(d_2 s_2 K/(Ds_1 + d_2 s_2 + \ldots + d_n s_n) \leq a_2$

We can then compare equations (19) and (24) to see that firm 2 derives much less value from buying data when capital is in short supply. Intuitively, firm 2 pays the direct cost of data, a_2, in both equations. But in equation (24), firm 2 also pays a shadow cost due to the fact that their data purchase leads to a supply shortage for capital and therefore lowers the quantity of capital firm 2 is able to buy. For most plausible parameters, the shadow cost is sufficiently high that firm 2 maximizes its profits by taking the corner solution of accessing exactly enough data so that capital demand equals capital supply.[9] We can solve to get that quantity of capital and data:

(25) $k_1 = (1-\alpha)^{(1/\rho)}(Ds_1)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$ & $d_1 = D$;

$k_i = \{K-(1-\alpha)^{(1/\rho)}(Ds_1)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}\}*s_i/(s_2 + \ldots + s_n)$ for i>1→

$\{K-(1-\alpha)^{(1/\rho)}(Ds_1)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}\}*s_i/(s_2 + \ldots + s_n) = (1-\alpha)^{(1/\rho)}(d_i s_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$→

$\{K*(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)} - Ds_1\}/(s_2 + \ldots + s_n) = d_i$ for i>1

$a_2 = (p^*\alpha[s_2 K/(D + d_2 s_2 + \ldots + d_n s_n)]^{\rho} + (1-\alpha)*s_2^{\rho})^{(1/\rho)} - r[s_2 K/(D + d_2 s_2 + \ldots + d_n s_n)]$

9. Firm 2 buys more data than the corner solution in two rare circumstances. The first rare circumstance involves a duopsony in which firms 1 and 2 are the only major data users. The second rare circumstance involves capital and data that are very weak complements. In those two cases, capital rationing only imposes a very small cost on firm 2 and so they buy the quantity of data that they would without capital rationing.

(26) Revenue of data seller =

$$s_1 D\,[\alpha(1-\alpha)*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+(1-\alpha)]^{(1/\rho)}- r(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}+ s_2 d_2\{K*(1-\alpha)^{(-1/\rho)}(Ds_1)^{-1}*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(1/\rho)}\text{-}1\}*(1-\alpha)^{(-1/\rho)}/(s_2+ \ldots+s_n)$$

(27) Capital owner revenue $\sum rk_i = rK$

The data seller in equation (26) earns strictly lower profits than the data seller in equation (21) because firm 2 only buys part of the data rather than all of the data. In other words, even a data seller who is resigned to the existence of piracy would still prefer that the supply of capital be higher so that they can sell slightly more data. In practice, the data seller's preference for a high supply of capital is relatively weak because data sales revenue is low even when there is no shortage of capital. Just like equation (22), equation (27) shows no externalities between data access rates and capital owner revenue.

Capital Usage and Rental Rate with Free Data Funded by the Capital Owner

This subsection considers a world in which the capital owner buys complete rights to the maximum amount of data, D, and then distributes that data to maximize capital revenue. Given a distribution of data, each firm then decides how much capital to rent. For example, the tropical island hotel might hire a mapmaker to make maps and then email those maps without cost to all n travel agencies. Given that those maps are distributed, the travel agencies decide how much capital to rent.
The capital owner's distribution problem is trivial if higher data usage raises the demand for capital. Data sharing has zero marginal cost. Therefore, the capital owner distributes the maximum amount of data, D, to every firm. Given that distribution of data, we use equations (3) and (4) to calculate capital usage by each firm:

(28) $k_i = p*(1-\alpha)^{(1/\rho)}(Ds_i)*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}$ if capital is not in short supply

$k_i = s_iK/(s_1+s_2+ \ldots+s_n)$ if capital is in short supply.

(29) Capital owner revenue = $\sum rk_i$ =

$p*r(s_1+ s_1+..+ s_n)*(1-\alpha)^{(1/\rho)}(D)*[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}$ if capital is not in short supply

rK if capital is in short supply =

$$\min\{p*rD(s_1+ s_1+..+ s_n)*(1-\alpha)^{(1/\rho)}[(r/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}, rK\}$$

Equation (29) can be solved numerically. Figure 1 shows the capital revenue associated with free data for one selected parameter region. For now, the theoretical framework discusses some general patterns to the solution. Just like an economic planner, the capital owner derives more value from data when the n firms have more total skill using data, when the supply of capital is higher, and when the demand for capital is much lower than the supply of capital. For example, a hotel is more likely to distribute free maps if the n travel agencies know how to use maps, when the hotel is large, and when the hotel has many empty rooms. Conversely, a hotel which is fully booked for the next year is unlikely to bother distributing free maps because the extra demand does not help them.

Numerical Example of Data Values Under Multiple Possible Funding Mechanisms

This section shows the value of data by funding method for one selected parameter region. All the graphs in this section are calculated using the parameters p = 1, α = 0.5, r = 0.5, K=50, and firms with moderately heterogenous skills, $s_1 = 1$, $s_2 = 2^{-0.4}$, $s_3 = 3^{-0.4}$, …, $s_{100} = 100^{-0.4}$. These parameters were picked for graphical clarity and are not calibrated to any real world example.

Figure 1. Value of Data, by Quantity and Funding Method

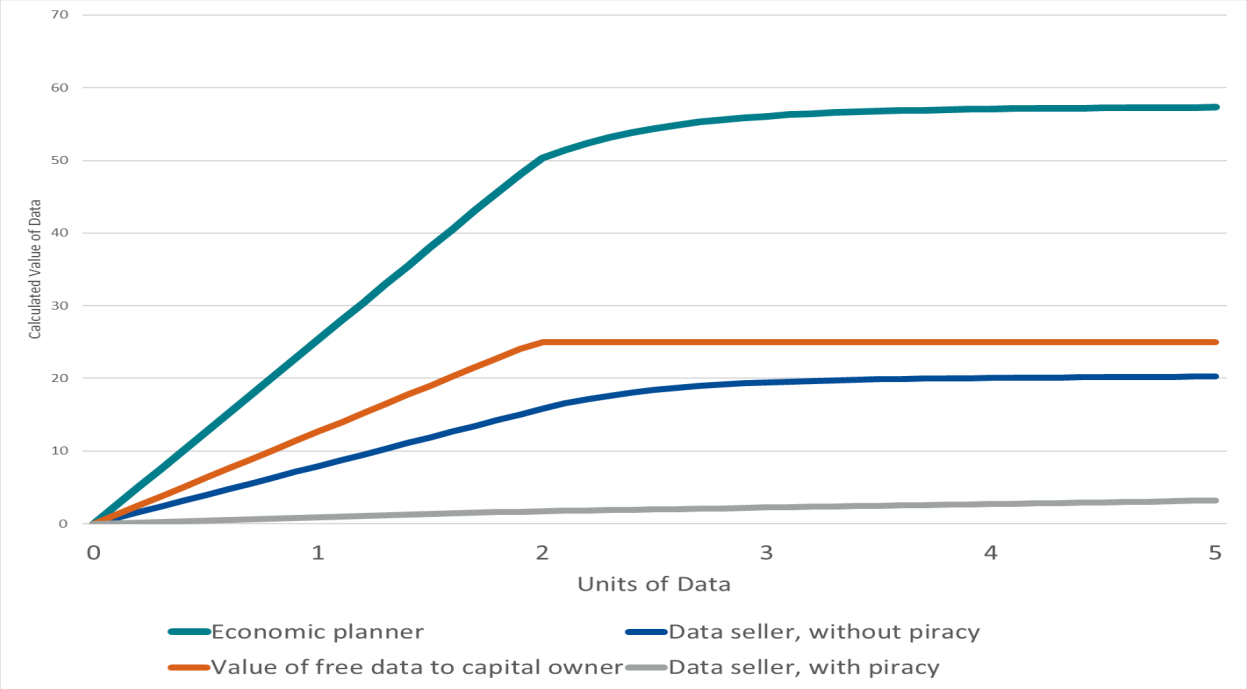


Figure 1 shows that the value of free data to the owners of a complementary asset is slightly higher than data sales revenue without piracy. This result is sensitive to the exact parameter region studied. Other plausible parameter regions show that the value of free data to the owner of a complementary asset is

slightly lower than data sales revenue without piracy. Accordingly, data which are protected from piracy can either be funded by the owner of a complementary asset or funded by data sales revenue. In practice, most of the data types which are protected from piracy by intellectual property laws are already included in GDP as intangible capital assets (BEA 2022). Hence, they are not covered by either the proposed national accounting guidelines (Rassier et al. 2019) (Eurostat 2020) or this paper's back-of-the-envelope calculations.

Figure 1 also shows that the value of free data is much higher than the value of data sales revenue with piracy. This result holds for almost any parameter region where many firms use data simultaneously. Accordingly, shared data which are vulnerable to piracy are very likely to be funded by the owner of a complementary asset. But unshared data can sometimes be funded by data sales revenue.[10]

One might think that free data yields more value to the capital owner when data and capital are strong complements. That is certainly true for some functional forms. However, the particular functional form used in this paper's graphs shows that free data yields similar value to the capital owner when data and capital are strong complements, $\rho = -5$, and when data and capital are weak complements, $\rho = 0.5$. Intuitively, the higher relative demand increase associated with strong complementarity is balanced by a lower absolute demand when capital and data are strong complements. In other words, even a small degree of complementarity is sufficient for data to have a large absolute impact on capital demand. This paper saves space by only showing graphs where data and capital are strong complements, $\rho = -5$.

3. Extended Theoretical Framework with Fixed Output Prices, Multiple Capital Assets, and Multiple Data Types

The previous section studied data values in a theoretical framework with only one capital asset and therefore only one capital owner. If that theoretical framework is slightly extended to allow for multiple capital owners, then the data values calculated in the social planner subsection and the data seller subsections still apply. However, the data value calculated in the free data subsection no longer apply because each capital owner only receives a portion of the aggregate revenue increase associated with free data. Hence, one might think that free data are only more valuable than sold data in situations where all capital is owned by a single monopolist. This section extends the previous section's simple economy to study data funding mechanisms in a world with multiple capital owners.

10. Own-account data are treated as if they were "sold" by the data production division to the data using division.

The extended economy starts out with v separate capital assets that are owned by v separate capital owners. The capital owners set their capital at rental prices, r^1 to r^v. In addition, there are w separate data owners that sell data access at fixed prices per unit of data, a^1 to a^w. Finally, there are n separate firms that use capital and data. These n firms all have v separate production functions. Each production function is a modified constant elasticity of substitution (CES) function that first combines the w data types into a single data index and then separately combines one capital asset with that data index to produce output which is sold at a price, p. The n firms differ in their skill at using data, with firm 1's skill designated as s_1, firm 2's skilled designated as s_2, and so on. For discussion purposes, the firms are ordered so that $s_1 \geq s_2 \geq \ldots \geq s_n$. This extended model assumes that each firm decides how much of each capital stock to rent, k^1 to k^v, and how much of each data type, d^1 to d^w, to buy. Just like in the simple model, capital is rival. So, the total capital supply of K^1 will be split proportionally in case of a shortage of capital asset 1, the total capital supply of K^2 will be split proportionally in case of a shortage of capital asset 2, and so on. Data are nonrival, so each firm is given the maximum quantities of data, D^1 to D^w, in case of a shortage.

(30) Output of firm i = $Y_i = \{\alpha(k^1_i)^\rho + (1-\alpha)*[(\beta^{1,1}d^1_i{}^\sigma + \ldots + \beta^{w,1}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} +$

$\{\alpha(k^2_i)^\rho + (1-\alpha)*[(\beta^{1,2}d^1_i{}^\sigma + \ldots + \beta^{w,2}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} + \ldots + \{\alpha(k^v_i)^\rho + (1-\alpha)*[(\beta^{1,v}d^1_i{}^\sigma + \ldots + \beta^{w,v}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)}$

(31) Profit of firm i = $\Pi_i = p*\{\alpha(k^1_i)^\rho + (1-\alpha)*[(\beta^{1,1}d^1_i{}^\sigma + \ldots + \beta^{w,1}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} +$

$p*\{\alpha(k^2_i)^\rho + (1-\alpha)*[(\beta^{1,2}d^1_i{}^\sigma + \ldots + \beta^{w,2}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} + \ldots +$

$p*\{\alpha(k^v_i)^\rho + (1-\alpha)*[(\beta^{1,v}d^1_i{}^\sigma + \ldots + \beta^{w,v}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} - (r^1k^1_i + r^2k^2_i + \ldots + r^vk^v_i + a^1d^1_i + a^2d^2_i + \ldots + a^wd^w_i)$

This extended economy can also be illustrated with the example of a tropical island that has many capital-intensive tourist activities in addition to the hotel. There might be a boat owner who offers fishing tours, an airplane owner who offers skydiving, or a forest owner who offers nature walks. And now there are multiple mapmakers who each sell one type of map. Some maps are nonspecific and therefore have similar impacts on all capital assets. For example, a climate map showing normal seasonal temperatures is useful for almost any outdoor activity planning. Other maps are specific to one capital asset. For example, maps showing fishing spots are complementary to the boat asset while maps showing good hiking trails are specific to the forest asset. Finally, there are n separate travel agencies which rent the v capital assets and use those capital services to create vacation packages for individual tourists. These n travel agencies differ in their skill selling vacation packages, but they all sell more

vacation packages when they can use maps to either match customers with the vacation package that's best for them or advertise those vacation packages.

Capital Usage and Data Creation when an Economic Planner Controls All Outcomes

This subsection considers a world in which data allocations are made by an economic planner whose goal is to maximize total value-added. Given the distribution of data, each firm then decides how much of each capital asset to rent. Just like in the simple economy, the economic planner's data allocation decision is trivial. Data sharing is assumed to have zero marginal cost. Therefore, the economic planner distributes the maximum amount of data (D^1, D^2,.., D^w) to every firm. Given that universal data usage, each firm's profit function can be expressed as:

(32) $\Pi_i = p*\{\alpha(k^1_i)^\rho+(1-\alpha)*[(\beta^{1,1}D^{1\sigma}+..+\beta^{1,w}D^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)}$

$+ p*\{\alpha(k^2_i)^\rho+(1-\alpha)*[(\beta^{1,2}D^{1\sigma}+..+\beta^{w,2}D^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)} + ...+$

$p*\{\alpha(k^v_i)^\rho+(1-\alpha)*[(\beta^{1,v}D^{1\sigma}+..+\beta^{w,v}D^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)} -\{r^1*k^1_i + r^2*k^2_i +...+ r^v*k^v_i\} -$ fixed data access fee$_i$

The only decision facing firm i is how much of each capital asset to rent. If capital is not in short supply, firms can solve that decision by taking the derivative and finding out when the marginal profit contribution of capital is zero:

(33) $d\Pi_i /dk^1= p*\{\alpha(k^1_i)^\rho+(1-\alpha)*[(\beta^1_1D^{1\sigma}+..+\beta^1_wD^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)-1}*\alpha(k^1_i)^{\rho-1}- r^1= 0 \rightarrow$

$p*(1-\alpha)^{(1/\rho)}[(\beta^{1,1}D^{1\sigma}..+\beta^{w,1}D^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}= k^1_i$

(34) Total value added for all businesses =

$\sum p*(\alpha\{[(1-\alpha)^{(1/\rho)}[(\beta^{1,1}D^{1\sigma}..+\beta^{w,1}D^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}\}^\rho+(1-\alpha)*(\beta^{1,1}D^{1\sigma}..+\beta^{w,1}D^{w\sigma})^{1/\sigma}s_i)^\rho\}^{(1/\rho)} +$

$\sum p*(\alpha\{[(1-\alpha)^{(1/\rho)}[(\beta^{1,2}D^{1\sigma}..+\beta^{w,2}D^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}\}^\rho+(1-\alpha)*(\beta^{1,1}D^{1\sigma}..+\beta^{w,1}D^{w\sigma})^{1/\sigma}s_i)^\rho\}^{(1/\rho)}$

$+...+$

$\sum p*(\alpha\{[(1-\alpha)^{(1/\rho)}[(\beta^{1,v}D^{1\sigma}..+\beta^{w,v}D^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}\}^\rho+(1-\alpha)*(\beta^{1,v}D^{1\sigma}..+\beta^{w,v}D^{w\sigma})^{1/\sigma}s_i)^\rho\}^{(1/\rho)} =$

$p*(s_1+ s_2+...s_n)*(1-\alpha)^{(1/\rho)}*(\beta^{1,1}D^{1\sigma}..+\beta^{w,1}D^{w\sigma})^{1/\sigma}*(\alpha[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+1\}^{(1/\rho)} +$

$p* (s_1+ s_2+...s_n)* (1-\alpha)^{(1/\rho)}*(\beta^{1,1}D^{1\sigma}..+\beta^{w,1}D^{w\sigma})^{1/\sigma}*\{(\alpha[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+1\}^{(1/\rho)} +...+$

$p* (s_1+ s_2+...s_n)* (1-\alpha)^{(1/\rho)}*(\beta^{1,1}D^{1\sigma}..+\beta^{w,1}D^{w\sigma})^{1/\sigma}*\{(\alpha[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{-1}+1\}^{(1/\rho)}$

If capital is in short supply, then it is distributed in proportion to demand:

(35) $k^1_i = s_i * K^1/(s_1+s_2 +\dots +s_n)$

(36) Total value added for all businesses =

$$\sum\{p*(\alpha[K^1*s_i/(s_1+ s_2+\dots s_n)]^{\rho}+(1-\alpha)*(\beta^1{}_1 D^{1\sigma} ..+\beta^w{}_1 D^{w\sigma})^{1/\sigma} s_i)^{\rho}\}^{(1/\rho)} +$$

$$\sum\{ p*(\alpha[K^2*s_i/(s_1+ s_2+\dots s_n)]^{\rho}+(1-\alpha)*(\beta^1{}_2 D^{1\sigma} ..+\beta^w{}_2 D^{w\sigma})^{1/\sigma} s_i)^{\rho}\}^{(1/\rho)} +\dots+$$

$$\sum\{ p*(\alpha[K^v*s_i/(s_1+ s_2+\dots s_n)]^{\rho}+(1-\alpha)*(\beta^1{}_v D^{1\sigma} ..+\beta^w{}_v D^{w\sigma})^{1/\sigma} s_i)^{\rho}\}^{(1/\rho)} =$$

$$p* (s_1+ s_2+\dots s_n)*\{(\alpha[K^1/(s_1+ s_2+\dots s_n)]^{\rho}+(1-\alpha)*[(\beta^1{}_1 D^{1\sigma} ..+\beta^w{}_1 D^{w\sigma})^{1/\sigma}]^{\rho}\}^{(1/\rho)} +$$

$$p* (s_1+ s_2+\dots s_n)*\{(\alpha[K^2/(s_1+ s_2+\dots s_n)]^{\rho}+(1-\alpha)*[(\beta^1{}_2 D^{1\sigma} ..+\beta^w{}_2 D^{w\sigma})^{1/\sigma}]^{\rho}\}^{(1/\rho)} +..+$$

$$p* (s_1+ s_2+\dots s_n)*\{(\alpha[K^v/(s_1+ s_2+\dots s_n)]^{\rho}+(1-\alpha)*[(\beta^1{}_v D^{1\sigma} ..+\beta^w{}_v D^{w\sigma})^{1/\sigma}]^{\rho}\}^{(1/\rho)}$$

Equations (33) and (36) look long and complicated, but all the parameters are constants and the equations can be readily solved on a computer. Figures 2 to 4 show those total profits for selected parameters. For now, this paper simply notes that the same general factors influence the planner's value of data in this extended theoretical framework as influenced the planner's value of data in the simple theoretical framework shown in the previous section.

Capital Usage and Data Sales Without Piracy

This subsection considers a world in which each firm makes decisions to maximize their individual profits without considering either positive or negative externalities. First, the data owners decide the prices, a^1 to a^w, to charge for their data type. Second, the firms decide how much of each data type to buy. Finally, the firms decide how much capital to rent. Just like in the simple framework, the precise solutions shown in this subsection depend on whether capital is in short supply.

In the tropical island example, the hotel owner sets a rental rate for hotel rooms of r^{hotel}, the boat owner sets a rental rate for boats of r^{boat}, the airplane owner sets a rental rate for planes of r^{plane}, and so on. Meanwhile, the map owners decide $a^{hiking\ map}$, $a^{fishing\ map}$, $a^{climate\ map}$, and so on. Third, each travel agency decides how many maps of each type to access. Finally, each of the n travel agencies decides how much

of each asset of capital to rent. The n travel agencies then combine bought maps with rented capital to sell vacation packages to individual tourists.

This subsection solves part of this model by induction. First, it calculates how much capital each firm rents given their already decided data purchase and the capital rental rates, r^1 to r^v. Next, it calculates how much of data type 1 the firms buy given the amount of capital each expects to rent and their already decided data purchases for types 2 to w. It does the same for each of the other data types. Finally, it calculates data access rates a^1 to a^w which maximize each data seller's individual profit.

When capital is not in short supply, then capital allocations can be solved by simply taking the first order conditions of the profit function for firm i. To save space, only the first capital asset, k^1_i, is shown:

(37) $d\Pi_i / dk^1_i = p^*\{\alpha(k^1_i)^\rho+(1-\alpha)*[(\beta^{1,1}d^1_i{}^\sigma+..+\beta^{w,1}d^1_i{}^\sigma)^{1/\sigma}*s_i]^\rho\}^{(1/\rho)-1}*\alpha(k^1_i)^{\rho-1} - r^1 = 0 \rightarrow$

$k^1_i = p^*(1-\alpha)^{(1/\rho)}(\beta^{1,1}d^1_1{}^\sigma+..+\beta^{w,1}d^w_1{}^\sigma)^{1/\sigma}*s_i*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha]^{(-1/\rho)}$

(38) $\Pi_i = p\{\alpha(k^1_i)^\rho+(1-\alpha)*[(\beta^{1,1}d^1_i{}^\sigma+..+\beta^{w,1}d^w_i{}^\sigma)^{1/\sigma}*s_i]^\rho\}^{(1/\rho)} +$

$p\{\alpha(k^2_i)^\rho+(1-\alpha)*[(\beta^{1,2}d^1_i{}^\sigma+..+\beta^{w,2}d^w_i{}^\sigma)^{1/\sigma}*s_i]^\rho\}^{(1/\rho)}+...+p\{\alpha(k^v_i)^\rho+(1-\alpha)*[(\beta^{1,v}d^1_i{}^\sigma+..+\beta^{w,v}d^2_i{}^\sigma)^{1/\sigma}*s_i]^\rho\}^{(1/\rho)} -$

$\{r^1*k^1_i + r^2*k^2_i +...+ r^v*k^v_i + a^1*d^1_i + a^2*d^2_i +...+ a^w*d^w_i\} =$

$p\{\alpha(1-\alpha)(\beta^{1,1}d^1_i{}^\sigma+..+\beta^{w,1}d^w_i{}^\sigma)^{\rho/\sigma}*s_i{}^\rho*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha]^{-1}+(1-\alpha)*(\beta^{1,1}d^1_i{}^\sigma+..+\beta^{w,1}d^w_i{}^\sigma)^{\rho/\sigma}*s_i{}^\rho\}^{1/\rho} +$

$p\{\alpha(1-\alpha)(\beta^{1,2}d^1_i{}^\sigma+..+\beta^{w,2}d^w_i{}^\sigma)^{\rho/\sigma}*s_i{}^\rho*[(r^2/\alpha)^{\rho/(1-\rho)}-\alpha]^{-1}+(1-\alpha)*(\beta^{1,2}d^1_i{}^\sigma+..+\beta^{w,2}d^w_i{}^\sigma)^{\rho/\sigma}*s_i{}^\rho\}^{1/\rho} + ..+$

$p\{\alpha(1-\alpha)(\beta^{1,v}d^1_i{}^\sigma+..+\beta^{w,v}d^w_i{}^\sigma)^{\rho/\sigma}*s_i{}^\rho*[(r^v/\alpha)^{\rho/(1-\rho)}-\alpha]^{-1}+(1-\alpha)*(\beta^{1,v}d^1_i{}^\sigma+..+\beta^{w,v}d^w_i{}^\sigma)^{\rho/\sigma}*s_i{}^\rho\}^{1/\rho} +$

$-r^1(1-\alpha)^{(1/\rho)}(\beta^{1,1}d^1_i{}^\sigma+..+\beta^{w,1}d^w_i{}^\sigma)^{1/\sigma}*s_i\}*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)} -$

$-r^2(1-\alpha)^{(1/\rho)}(\beta^{1,2}d^1_i{}^\sigma+..+\beta^{w,2}d^w_i{}^\sigma)^{1/\sigma}*s_i\}*[(r^2/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)} -$

$-r^v(1-\alpha)^{(1/\rho)}(\beta^{1,v}d^1_i{}^\sigma+..+\beta^{w,v}d^w_i{}^\sigma)^{1/\sigma}*s_i\}*[(r^v/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)} - .. - (a^1d^1_i + a^2d^2_i +...+ a^wd^w_i) =$

$(\beta^{1,1}d^1_i{}^\sigma+..+\beta^{w,1}d^w_i{}^\sigma)^{1/\sigma}*s_i*(1-\alpha)^{(1/\rho)}(\alpha[\ p^*\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)} - r^1[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha]^{(-1/\rho)})$

$(\beta^{1,2}d^1_i{}^\sigma+..+\beta^{w,2}d^w_i{}^\sigma)^{1/\sigma}*s_i*(1-\alpha)^{(1/\rho)}(\alpha[\ p^*\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)} - r^2[(r^2/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)})...+$

$(\beta^{1,v}d^1_i{}^\sigma+..+\beta^{w,v}d^w_i{}^\sigma)^{1/\sigma}*s_i*(1-\alpha)^{(1/\rho)}(\alpha[\ p^*\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)} - r^v[(r^v/\alpha)^{\rho/(1-\rho)}-\alpha]^{(-1/\rho)}) -$

$(a^1d^1_i + a^2d^2_i +...+ a^wd^w_i)$

Equation (38) is can be solved using first order conditions to get an interior solution:

(39) $d\,\Pi_i / dd^1_i = 0$ at optimum $\rightarrow a^1 =$

$$s_i(\beta^{1,1}d^1{}_i^{\sigma}+..+\beta^{w,1}d^w{}_i^{\sigma})^{1/\sigma-1}*\beta^{1,1}*d^1{}_i^{\sigma-1}(1-\alpha)^{(1/\rho)}[p\alpha\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^1\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}+$$

$$s_i(\beta^{1,2}d^1{}_1^{\sigma}+..+\beta^{w,2}d^w{}_1^{\sigma})^{1/\sigma-1}*\beta^{1,2}*d^1{}_i^{\sigma-1}(1-\alpha)^{(1/\rho)}[p\alpha\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^2\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}$$

$$+.+$$

$$s_i(\beta^{1,v}d^1{}_1^{\sigma}+..+\beta^{w,v}d^w{}_1^{\sigma})^{1/\sigma-1}*\beta^{1,v}*d^1{}_i^{\sigma-1}(1-\alpha)^{(1/\rho)}[p\alpha\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^v\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}$$

$$\rightarrow a^1 d^1{}_i^{1-\sigma} = s_i(\beta^{1,1}d^1{}_i^{\sigma}+..+\beta^{w,1}d^w{}_i^{\sigma})^{1/\sigma-1}*\beta^{1,1}*(1-\alpha)^{(1/\rho)}p[\alpha\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^1\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}$$

$$+s_i(\beta^{1,2}d^1{}_1^{\sigma}+..+\beta^{w,2}d^w{}_1^{\sigma})^{1/\sigma-1}*\beta^{1,2}*(1-\alpha)^{(1/\rho)}p[\alpha\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^2\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}+.+$$

$$s_i(\beta^{1,v}d^1{}_1^{\sigma}+..+\beta^{w,v}d^w{}_1^{\sigma})^{1/\sigma-1}*\beta^{1,v}*(1-\alpha)^{(1/\rho)}[p*\alpha\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^v\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}\rightarrow$$

$$d^1{}_i = (a^1)^{1/(\sigma-1)}*$$

$$\{s_i(1-\alpha)^{(1/\rho)}*[(\beta^{1,1}d^1{}_i^{\sigma}+..+\beta^{w,1}d^w{}_i^{\sigma})^{1/\sigma-1}*\beta^{1,1}*[\,p*\alpha\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^1\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}$$

$$+(\beta^{1,2}d^1{}_1^{\sigma}+..+\beta^{w,2}d^w{}_1^{\sigma})^{1/\sigma-1}*\beta^{1,2}*[\,p*\alpha\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^2\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}+.+$$

$$(\beta^{1,v}d^1{}_1^{\sigma}+..+\beta^{w,v}d^w{}_1^{\sigma})^{1/\sigma-1}*\beta^{1,v}*[p*\alpha\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^v\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{(-1/\rho)}]\}^{-1/(1-\sigma)}$$

Equation (39) is very long and does not generally have a simple solution. Nevertheless, it shows a few important ideas. Most obviously, the optimal quantity of data type 1, d^1_i, increases when the access rate, a^1, is lower. This is a standard result of almost any demand function. More important for our analysis, the optimal quantity of data type 1 for firm i, d^1_i, depends on the quantities of data types 2 to w already owned by firm i. If data types are strong complements, then a higher price for data types 2 to w reduces demand for all other data types and therefore imposes a negative externality on all the other data sellers. If data types are substitutes and capital is in short supply, then a lower price for data types 2 to w creates capital shortages and therefore imposes a negative externality on all the other data sellers. In almost every circumstance, the w data owners could increase their combined revenue by bundling the w data types into a single data asset that they sell together.

Capital Usage and Data Sales With Piracy

This subsection once again considers a world in which each firm makes decisions to maximize their individual profits, with one difference—a new decision about piracy. The earlier section examined two separate scenarios: one in which capital is not in short supply and one in which capital is in short supply. This subsection considers the same two scenarios as before and finds qualitatively similar results to the earlier subsection with only one capital asset and only one data type. Just like before, at most two firms pay a positive price for each type of data and therefore the data sellers often earn less revenue than they do in a world without piracy. In order to save space, this paper does not repeat the derivation of those results.

Capital Usage and Rental Rates with Free Data Funded by the Capital Owners

This subsection considers a world in which a capital owner buys complete rights to the maximum amount of a specific type of data, and then distributes that data to maximize capital revenue. For example, a boat owner might distribute maps of fishing spots, or a forest owner might distribute maps of hiking trails. The example with multiple separate capital owners can be broken down into the same two steps as the economic planner: how much data and how much capital should each firm rent.

The second question is trivial to solve. Data sharing is assumed to have zero marginal cost. Therefore, each capital owner distributes all the data that they own to every firm. This paper designates the data owned by data owner 1 as the ($d^{1,1}$, $d^{2,1}$,…, $d^{w,1}$), the data owned by capital owner 2 as ($d^{1,2}$, $d^{2,2}$,…, $d^{w,2}$), …, and the data owned by capital owner v as ($d^{1,v}$, $d^{2,v}$,…, $d^{w,v}$). The total quantity of data owned by all capital owners is designated as (D^1, D^2,…, D^w). If capital is not in short supply, then the capital demand is easy to calculate:

(40) k^1_i with all data = $p*(1-\alpha)^{(1/\rho)}[(\beta^{1,1}D^{1\sigma} ..+ \beta^{w,1}D^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}$

(41) k^1_i without capital owner 1's data = $p*(1-\alpha)^{(1/\rho)}[(\beta^{1,1}\{D^1 - d^{11}\}^{\sigma} ..+ \beta^{w,1}\{D^w - d^{w,1}\}^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}$

(42) Additional revenue for capital owner 1 associated with their data =

$$\sum p*(1-\alpha)^{(1/\rho)}[(\beta^{1,1}D^{1\sigma} ..+ \beta^{w,1}D^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)} -$$

$$\sum p*(1-\alpha)^{(1/\rho)}[(\beta^{1,1}\{D^1 - d^{11}\}^{\sigma} ..+ \beta^{w,1}\{D^w - d^{w,1}\}^{w\sigma})^{1/\sigma}s_i]*[(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)]^{(-1/\rho)}$$

Equation (42) can be solved numerically for any selected distribution of data ownership. This paper focuses on a few selected parameter regions and distributions of data ownership across capital owners which correspond to plausible examples. If capital is in short supply, then the capital demand is even easier to calculate:

(43) $k^1_i = s_i * K^1/(s_1 + s_2 + \ldots s_n)$ with or without capital owner 1's data

(44) Additional revenue for capital owner 1 associated with their data = 0

Capital Usage and Rental Rates with Free Data Funded by Another Data Owner

This subsection is very similar to the previous subsection. The only difference is that free data are now owned and distributed by a data seller rather than a capital owner. For example, a fishing map owner might buy fish recipes and then distribute those recipes for free. Given a distribution of data, firms decide how much capital to rent using the standard first order conditions.

The data seller's optimal data distribution strategy depends on the specific parameters selected. It may be true that data sharing is assumed to have zero marginal cost. But two data types can be sufficiently substitutable that one data owner can earn higher revenue if their competitors are eliminated. In that circumstance, the profit-maximizing option might be for one data owner to buy complete rights to the other data types, hold those data types without distributing them at all, and then sell their data at a very high price. This option can result in similar outcomes as the w data sellers colluding to set prices and quantities. However, that option is difficult to implement because the original data seller may secretly retain copies of their data and sell it to firms despite selling complete rights to the data owner earlier. Furthermore, antitrust law generally discourages this type of data transaction. The remainder of this subsection focuses on data types that are sufficiently complementary that a single data buyer, data owner i, chooses to distribute all purchased data for free to every firm. Equation (39) can be solved to get the optimal quantity of the data type i.

$$(45) \qquad d^1_i=(a^1)^{1/(\sigma-1)}$$

$$\{s_i(1-\alpha)^{(1/\rho)}*[(\beta^{1,1}D^{1\sigma}+..+\beta^{w,1}D^{w\sigma})^{1/\sigma-1}*\beta^{1,1}*[p*\alpha\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^1\{(r^1/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}+(\beta^{1,2}D^{1\sigma}+..+\beta^{w,2}D^{w\sigma})^{1/\sigma-1}*\beta^{1,2}*[p*\alpha\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^2\{(r^2/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{-1/\rho}+.+(\beta^{1,v}D^{1\sigma}+..+\beta^{w,v}d^{w}{}_1{}^{\sigma})^{1/\sigma-1}*\beta^{1,v}*[p*\alpha\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha\}^{-1}+1]^{(1/\rho)}-r^v\{(r^v/\alpha)^{\rho/(1-\rho)}-\alpha)\}^{(-1/\rho)}]\}^{-1/(1-\sigma)}$$

It is difficult to solve equation (45) mathematically, but the value of free data can be solved numerically.

Numerical Examples of Data Values Under Multiple Possible Funding Mechanisms

The three graphs shown in this subsection are all special cases of figure 1. Just like that figure, p=1, data and capital have similar weights, α = 0.5, the firms are moderately heterogenous, s_1 = 1, $s_2 = 2^{-0.4}$, $s_3 = 3^{-0.4}$, …, $s_{100} = 100^{-0.4}$, and data and capital are strong complements, $\rho = -5$. All three graphs have 100 separate capital owners and 100 separate data types. All 100 capital owners charge the same rent, r =0.5, and have the same quantity of capital, K = 5,000. The difference between the graphs is the complementarity between the different data types and their specificity to particular capital assets. Figure 2 shows a parameter region in which the data types are perfect complements, $\sigma \sim -\infty$. Because these data types are all perfect complements, they are automatically specific to every single capital asset. Figure 3 shows a parameter region in which the data types are perfect substitutes, $\sigma \sim 1$, and data types are nonspecific, $\beta^{1,1} = \beta^{2,1} = .. = \beta^{100,1} = \ldots = \beta^{1,100} = \beta^{2,100} = .. = \beta^{100,100}$ = 0.01. Figures 4 shows a parameter region in which the data types are perfect substitutes, $\sigma \sim 1$, and data types are mostly specific, $\beta^{1,1} = \beta^{2,2} = .. = \beta^{100,100}$ =0.75; & …$\beta^{1,2} = \beta^{2,1} = \beta^{2,100} = .. = \beta^{100}{}_{99}$ = 0.00245.

By construction, all of the three graphs shown in this subsection are symmetrical. Hence, all of them have a Nash equilibrium in which each capital owner earns the same value from free data and a Nash equilibrium in which each data seller earns the same revenue from data sales. This paper focuses on those particular equilibria and shows total data value across all data owners. The graphs shown in figures 2 to 4 are all calculated with simplified methods that approximate the true solution. The precise lines shown may not be robust to alternative simplification methods, but the qualitative results are.

Figure 2. Value of Perfectly Complementary Data, by Quantity and Funding Method

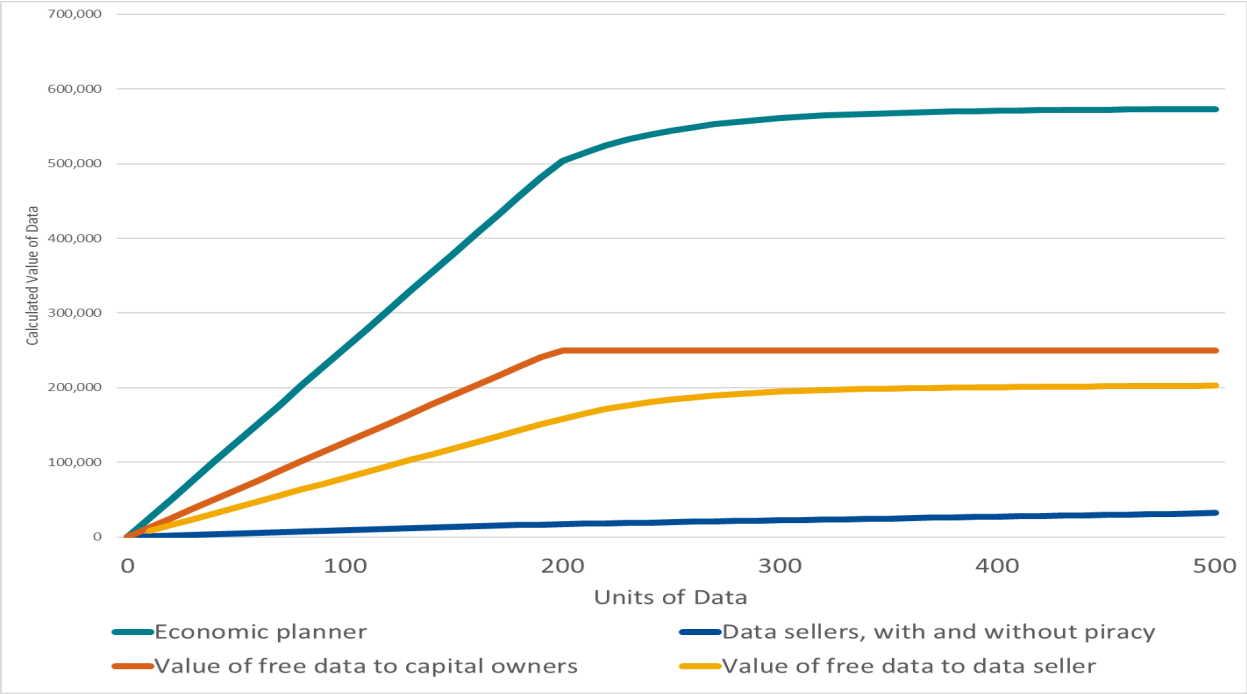


Figure 2 shows the revenue earned by data sellers when data are perfectly complementary to other data. The basic issue is simple. When the data types are complementary enough, the total number of firms buying data depends on the total cost of data. If the data sellers could collude, then they would collectively pick a moderately high price for data and sell to multiple firms. But collusion is not stable because each data seller has a strong incentive to raise their prices and capture a much larger share of a slightly smaller total data demand. In the tropical island example, the data types might be safety maps that show the activities in each zone. For example, skydivers cannot parachute into water if fishing boats are potentially below them and fishing boats cannot dock at a beach if swimmers are potentially in the water. So, a travel agency cannot put together a package without all the safety maps and each safety map seller has the power to raise prices very high. For many plausible parameter regions, the only Nash equilibrium is one where all the w data sellers charge such a high price that only the highest ability firm, firm 1, buys data. At that Nash equilibrium, no firm can raise its prices without reducing demand to zero. Because there is only one buyer, total data sales revenue can be much lower than it would be if all data types were sold by a single monopoly data seller.

In contrast, figure 2 shows that free data have a high value when data are perfectly complementary to other data. The basic issue is simple. If one free data owner withheld their data, then total data services fall to zero and demand for either capital or other data falls to the minimum possible. In other words, collusion is stable because the free data owner has a strong incentive to supply whatever data they hold

to the market. As a result, the total value of free data to the 100 capital owners is just as high as it could be of all the capital was owned by a single monopoly owner. In the tropical island example, total tourist activity completely shuts down if even one data owner did not distribute their safety map. Similarly, one data owner could be assigned to purchase all the other data types, make those data types free, and then earn as much money from the single data type that it sells as all 100 data owners could earn if they colluded to set data prices together.

Figure 3. Value of Nonspecific Substitutable Data, by Quantity and Funding Method

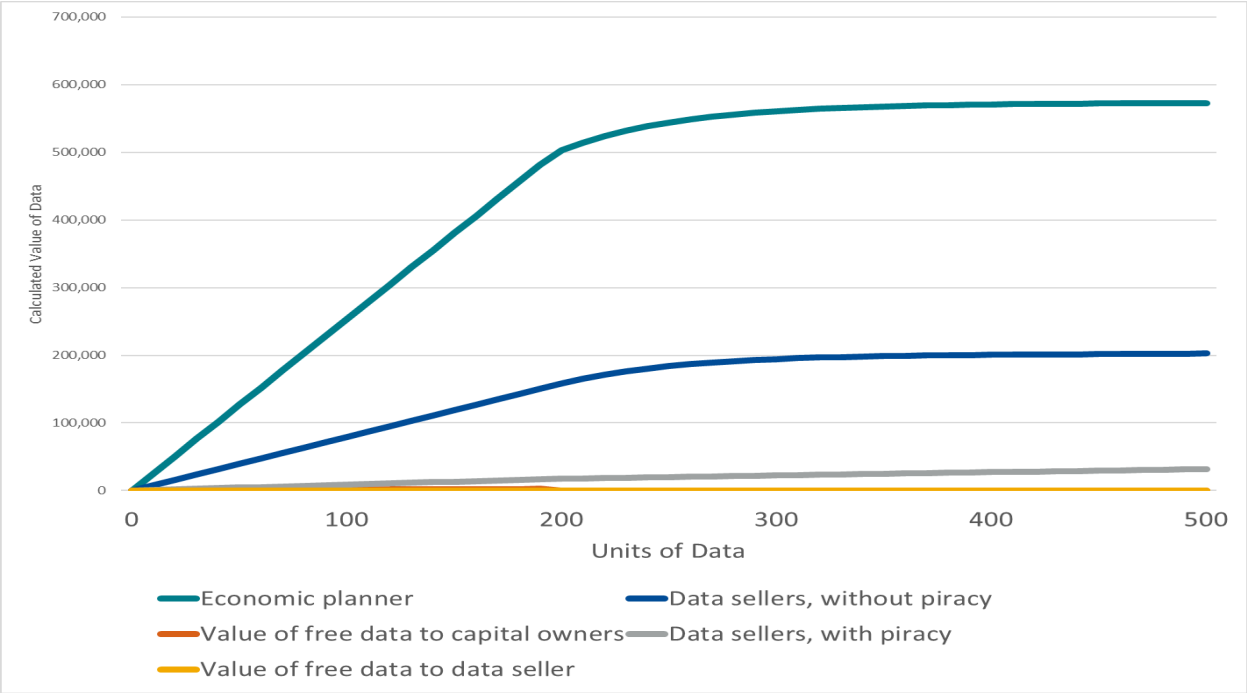


Figure 3 shows that data seller revenue is high when data are perfectly substitutable with other data. For most plausible parameters, firms pick either a corner solution of zero data purchases or a corner solution of maximum data purchases. At those two corner solutions, demand for one type of data is independent of the price of other data types. Accordingly, each data seller picks a price that maximizes their individual profits without causing any externalities for the other data sellers. Total data sales revenue is equal to the data sales revenue that would be earned if the data sellers could collude.

In contrast, figure 3 shows that the value of free data to either capital owners or other data owners is very low. The basic issue is simple. Neither total data services nor total capital demand fall much when one capital owner withholds their free data. Similarly, demand for sold data does not change at all if one data owner withholds their free data. In the tropical island example, the data types might be weather reports on different days. These weather reports are not specific to any particular capital type—and so a

capital owner who stops publishing the weather reports for their assigned day suffers only a small revenue loss. Similarly, there is little complementarity between weather reports for Tuesday and weather reports for Monday—and so a seller of Tuesday weather reports does not benefit from Monday weather reports being free. As a result, the value of individual types of free data to individual capital owners is much lower than the value of all types of free data combined to a consortium of capital owners. And the value of free data to other data owners is zero whether the data are combined or not.

The differences between the results shown in figure 2 and figure 3 are consistent with previous research showing that widely shared data are especially valuable when data are complementary to other data (Coyle 2022). However, these figures put a new twist on it. If data sellers could collude, they could earn substantial revenue from complementary data. The problem is that complementary data gives each of them an individual incentive to deviate—but substitutable data does not. Conversely, capital owners have huge penalties for deviation in figure 2 but small penalties for deviation in figure 3. Therefore, it is much easier to sustain a Nash equilibrium of free data when data are complementary and much easier to sustain a Nash equilibrium of sold data when data are substitutes. However, piracy may make it impossible for data sellers to earn much money from sales of non-specific data. In that case, the value of private data is small regardless of the funding mechanism. The combination of high social value and little private value may mean that non-specific substitutable data need government funding (Reiss 2021).

Figure 4. Value of Moderately Specific Substitutable Data, by Quantity and Funding Method

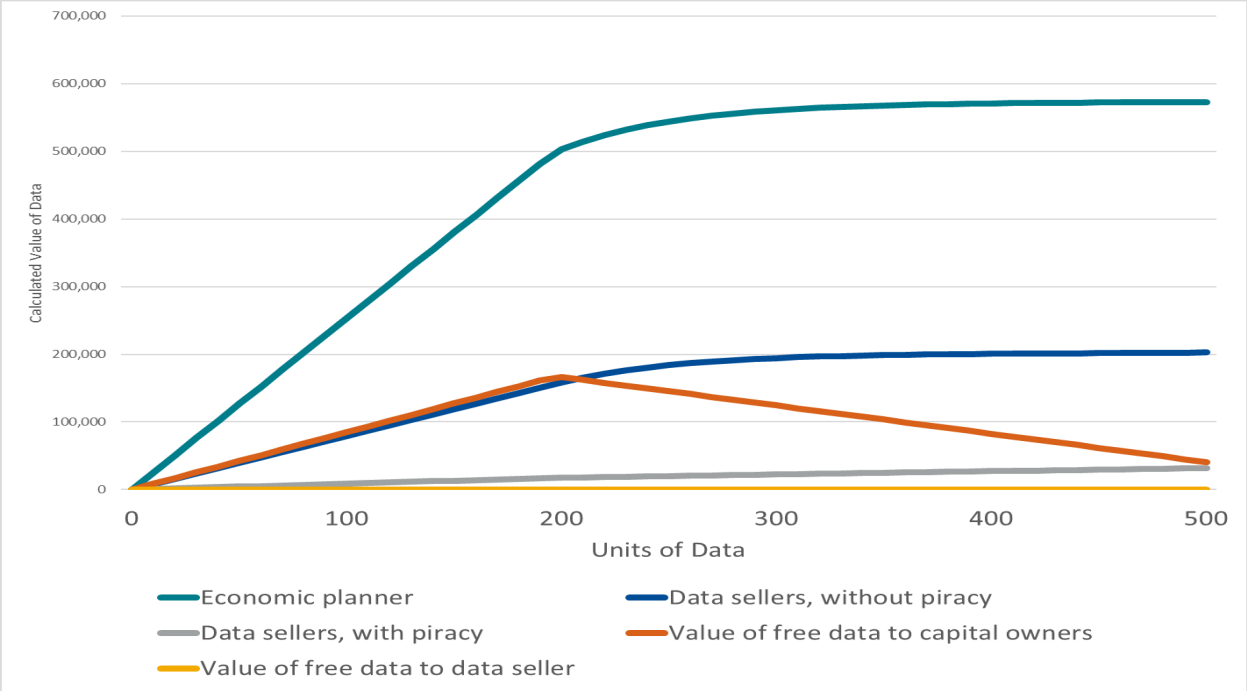


Figure 4 shows the exact same data seller revenue as figure 3. Intuitively, firms which buy data only care about the total benefits associated with data. So, their demand for data is the same whether data is completely non-specific or moderately specific. Just like before, total data sales revenue equals the data sales revenue that would be earned if all data types were sold by a single monopoly seller. And just like before, no data owner derives any benefit from other data types being available for free.

In contrast, figure 4 shows a much higher value of data to capital owners than figure 3. Assuming that each capital owner distributes the data type which are specific to their particular capital asset, then withholding data reduces the data services relevant to their asset by 75 percent. When capital is not in short supply, this large fall in data services noticeably lowers revenue for the capital owner. In the tropical island example, the data types might be rules for specific activities. For example, potential skydivers very much want to know if they will be allowed to skydive before they get in the air and potential fishermen very much want to know if they will be allowed to fish before they take a boat out on the water. So, the airplane owner has a large incentive to distribute skydiving rules and the boat owner has a large incentive to distribute fishing rules.

Real world examples of free data often belong in the parameter region shown in figure 4. For example, advertisers often give data like uses for a product, proper care of a product, and so on. These data types are specific enough that demand for a particular product would likely fall noticeably if advertisers did not distribute their data for free. But the data types are not perfectly specific, and therefore may raise demand for similar products that are sold by another company. Alternatively, high level job candidates often provide data like published articles, open-source software (Leppamaki and Mustonen 2009), talks to public interest groups, and so on. These types of data are specific enough to noticeably raise demand for a particular worker. But they are not perfectly specific, and therefore have spill-over benefits to firms which are considering hiring similar workers or firms which need a small amount of information but do not want to hire anyone. The international guidelines for national accounting are clear that neither positive externalities nor negative externalities are included in the national accounts (United Nations 2008 sec. 3.92). Accordingly, externalities associated with free data are excluded from the measured value of either sold data or free data.

4. Extended Theoretical Framework with Flexible Output Prices, Multiple Capital Assets, and Multiple Data Types

This section has almost the same parameters as the previous section. But now there is one key difference. Demand for output is now downward sloping so that output prices depend on the total quantity of output produced.

(46) Output of firm i = $Y_i = \{\alpha(k^1_i)^\rho + (1-\alpha)*[(\beta^{1,1}d^1_i{}^\sigma + .. + \beta^{w,1}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} +$

$\{\alpha(k^2_i)^\rho + (1-\alpha)*[(\beta^{1,2}d^1_i{}^\sigma + .. + \beta^{w,2}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} + ... + \{\alpha(k^v_i)^\rho + (1-\alpha)*[(\beta^{1,v}d^1_i{}^\sigma + .. + \beta^{w,v}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)}$

(47) Profit of firm i = Π_i = p(Total Output) $*\{\alpha(k^1_i)^\rho + (1-\alpha)*[(\beta^{1,1}d^1_i{}^\sigma + .. + \beta^{w,1}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} +$

p(Total Output)$*\{\alpha(k^2_i)^\rho + (1-\alpha)*[(\beta^{1,2}d^1_i{}^\sigma + .. + \beta^{w,2}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} + ... +$

p(Total Output) $*\{\alpha(k^v_i)^\rho + (1-\alpha)*[(\beta^{1,v}d^1_i{}^\sigma + .. + \beta^{w,v}d^w_i{}^\sigma)^{1/\sigma} * s_i]^\rho\}^{(1/\rho)} -$

$(r^1k^1_i + r^2k^2_i + ... + r^vk^v_i + a^1d^1_i + a^2d^2_i + ... + a^wd^w_i)$

These flexible prices can be illustrated with the example of a tropical island for which the supply of tourists is limited. If one firm puts together too many vacation packages, then tourists will demand lower prices for each vacation package. As a result, all the n firms have an incentive to collude and collectively produce fewer vacation packages.

Capital Usage and Data Creation when an Economic Planner Controls All Outcomes

This subsection considers a world in which data allocations are made by an economic planner whose goal is to maximize **nominal** value-added. To be clear, prices fall with increasing output—and therefore real value-added may not be maximized when nominal value-added is maximized. Given the distribution of data, each firm then decides how much of each capital asset to rent.

Unlike the world with fixed prices, the economic planner's data allocation decision is not trivial. On the one hand, distributing additional data gives the n firms more output to sell. On the other hand, distributing additional data reduces the price for each unit of output. The net impact of additional data on nominal value-added depends on the elasticity of demand and may very well be negative at larger data quantities. Conceptually, the economic planner's problem is very similar to the standard monopolist problem of how much output to sell and can be solved with the same techniques.

(48) Total value-added = p(Total output) * (Total Output) →

dTotal Profits/d(Total Output) = 0= [dp/ d(Total Output)]) * (Total Output)+ p →

-p/[dp(Total Output)/ d(Total Output)] = (Total Output)

Once the economic planner has picked its target level of output, it can then distribute the data necessary to achieve that level. It is difficult to calculate exactly how the economic planner should distribute data in order to achieve its target level of output. One major concern is that individual firms know that prices depend on total output and therefore may choose to produce less than they would in a competitive world. For simplicity, this paper assumes that the planner distributes data equally among all n firms. An equal distribution like that is easier to implement because it does not require any restrictions on data resale between the n firms. In other words, the economic planner may use data distribution choices to benefit firms at the expense of consumers.

Capital Usage and Data Sales Without Piracy

This subsection considers a world in which each firm makes decisions to maximize their individual profits without considering either positive or negative externalities. First, the data owners decide the prices, a^1 to a^w, to charge for their data type. Second, the firms decide how much of each data type to buy. Finally, the firms decide how much capital to rent. Just like in the simple framework, the precise solutions shown in this subsection depend on whether capital is in short supply.

This example is very difficult to solve mathematically and often does not have a simple solution. Nevertheless, this paper will highlight a few important points. Most obviously, output prices are lower when the total quantity of output is higher. If data types are strong complements, then a higher price for data type 1 imposes both the negative production externality examined earlier and a positive price externality. The net externality is theoretically uncertain. Similarly, the net externality is also uncertain if data types are substitutes and capital is in short supply. But if data types are substitutes and capital is not in short supply, then there is no negative production externality and only a positive price externality. Whether the net externalities are positive or negative, the w data owners could increase their combined revenue by colluding to set prices and quantities jointly.

Capital Usage and Data Sales With Piracy

This subsection once again considers a world in which each firm makes decisions to maximize their individual profits, with one difference—a new decision about piracy. The earlier section examined two separate scenarios: one in which capital is not in short supply and one in which capital is in short supply. This subsection considers two related scenarios: one in which neither capital is in short supply nor is potential output large enough to reduce prices significantly and one in which either capital is in short supply or potential output is large enough to reduce prices significantly. Just like before, at most two firms pay a positive price for each type of data and therefore the data sellers often earn less revenue than they do in a world without piracy. However, the quantity of data sold to firm 2 (and therefore pirated by firms 3 to n) is now smaller because firm 1 worries about both the extra competition creating capital shortages and the extra competition lowering output prices.

Capital Usage and Rental Rates with Free Data Funded by the Capital Owners

This subsection considers a world in which a capital owner buys complete rights to the maximum amount of a specific data type of data and then distributes that data to maximize capital revenue. For example, a boat owner might distribute maps of fishing spots, or a forest owner might distribute maps of hiking trails. The example with multiple separate capital owners can be broken down into the same two steps as the economic planner: how much data and how much capital should each firm rent.

Just like with the economic planner, the question of data distribution is no longer trivial to solve. Even though it has a zero marginal cost, a capital owner who distributes data may drive down output prices and therefore reduce demand for their capital asset. This paper assumes that capital owners are able to withhold data, and therefore additional data ownership never has a negative impact on their profits. But it will have a zero impact if they hold more data than necessary to achieve their profit-maximizing solution. The graphs use numerical techniques to calculate a value for data in select circumstances.

Capital Usage and Rental Rates with Free Data Funded by Another Data Owner

This subsection is very similar to the previous subsection. The only difference is that free data are now owned and distributed by a data owner rather than a capital owner. Just like before, this paper was not able to calculate a mathematical solution for the value of data to other data owners in all circumstances. But the graphs use numerical techniques to calculate a value for data in select circumstances.

Numerical Examples of Data Values Under Multiple Possible Funding Mechanisms

The three graphs shown in this subsection all use similar parameters to figures 2 to 4. The only difference is that p = min[1, 200,000/(Total Output)]. By construction, this price schedule results in the same value of data for small quantities of total output and a smaller value of data for large quantities of total output. Just like figures 2 to 4, figures 5 to 7 are all calculated with simplified methods that approximate the true solution. The precise lines shown may not be robust to alternative simplification methods, but the qualitative results are.

Figure 5. Value of Complementary Data, by Quantity and Funding Method

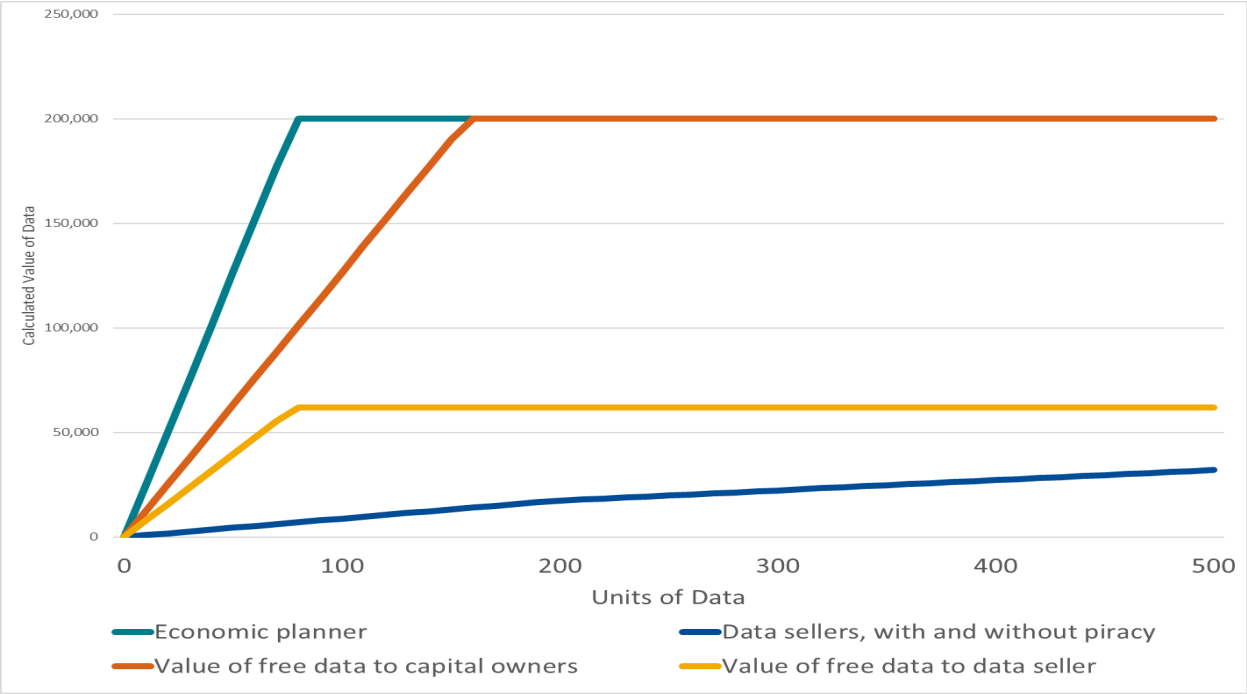


Figure 5 shows a consistently lower value of data than figure 2. This lower value reflects the fact that output prices fall when total output is greater than 200,000 and so the maximum revenue that can be earned from production is lower. Of course, the falling output prices benefit consumers even if they do not benefit businesses. This trade-off between business value-added and consumer welfare has been discussed by many previous papers studying monopolies and oligopolies.

Figure 5 also shows that capital owners receive a larger share of the economic value from free data than they did in figure 2. This larger share is due to the fact that capital becomes an increasingly scarce factor of production as data becomes more abundant and output prices fall. Accordingly, a moderate decline in output prices lowers firm profits without decreasing either the demand for capital or the capital rental

rate.[11] Conversely, the sole data seller receives a smaller share of the economic value from free data than they did in figure 2 because lower output prices decrease the demand for data.

Figure 6. Value of Nonspecific Substitutable Data, by Quantity and Funding Method

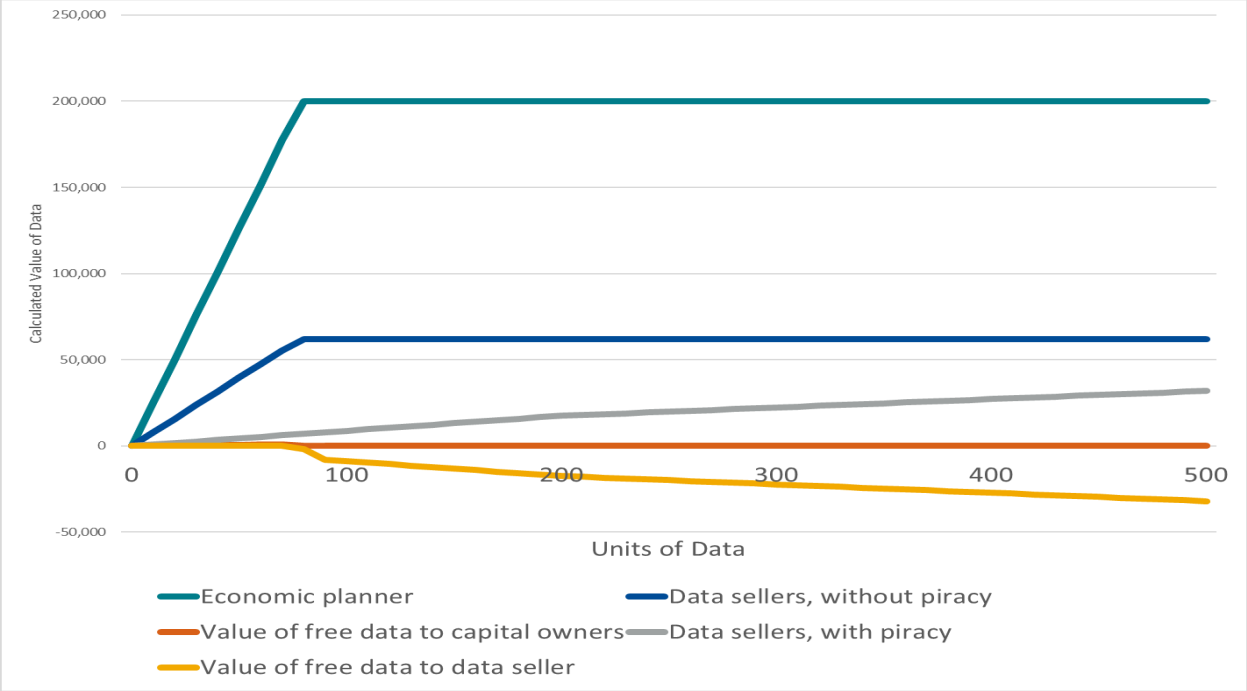


Figure 6 shows that free data has no value for capital owners and a negative value for data sellers. This negative value is easy to explain. Because output prices fall when total output is high, the data seller is indirectly competing with the free data. With a different production function, the data seller may even be directly competing with free data as well. In that case, the value of data will be even more negative. Accordingly, data sellers will certainly not fund the creation of free data and may even expend significant resources to prevent its creation.

11. The completely flat capital revenue shown in figure 5 is due to the assumption of fixed capital rental rates. But results are qualitatively similar if capital rental rates are allowed to vary based on demand.

Figure 7. Value of Moderately Specific Substitutable Data, by Quantity and Funding Method

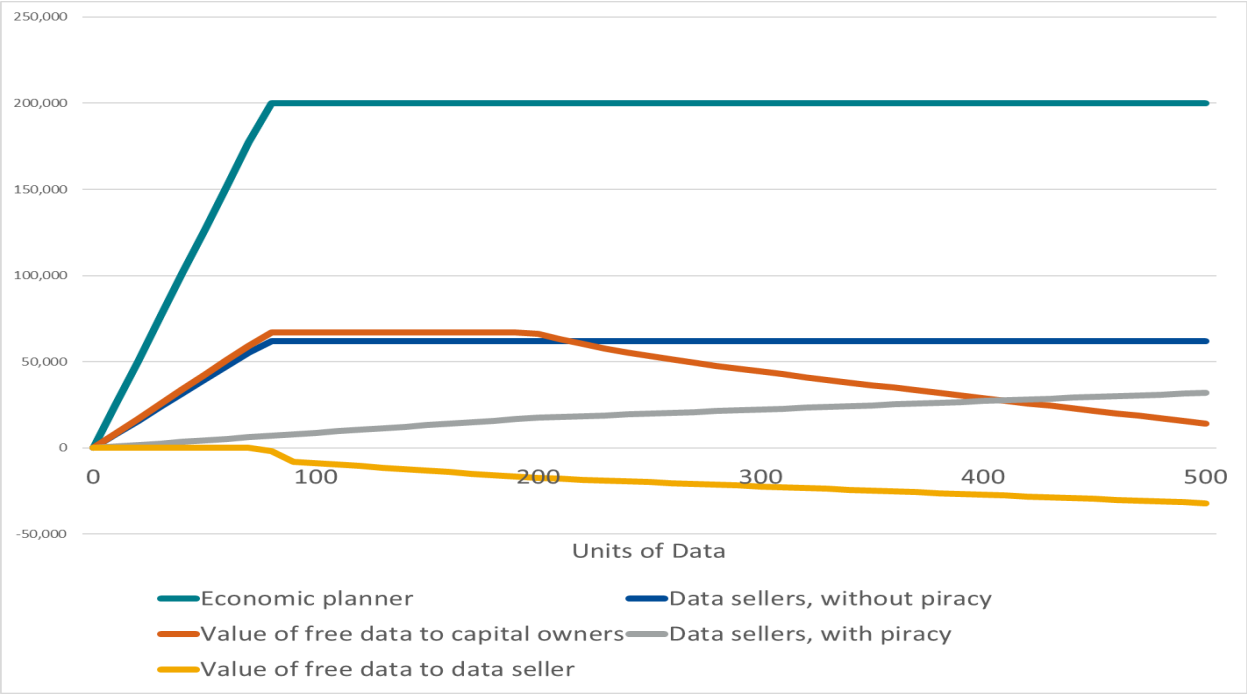


Figure 7 shows qualitatively similar results to figure 4. In other words, specific data are valuable to capital owners both in a market with fixed prices and in a market where output prices depend on total output. However, there is one interesting difference. In figure 4, positive spillovers do not lower output prices—and therefore capital owners prefer to fund free data which offer the largest direct benefits. In figure 7, positive spillovers lower output prices—and therefore capital owners prefer to fund free data which offer reasonably large direct benefits without too many indirect benefits. In other words, capital owners act much more "selflessly" when demand for the output of the n firms is not downward sloping. But even when capital owners act "selfishly," they may still fund free data which offer large indirect benefits if those free data have very large direct benefits.

5. Back-of-the-Envelope Calculation of Free Data Creation

Neither information on data sales revenue nor information on the value received by the owner of a complementary capital asset were located. Instead, the back-of-the-envelope calculations in this section use cost-based proxies to value free data. National accountants often use cost-based proxies, such as the payments for purchased inputs or the wages of employees who create own-account output, when studying difficult to value products (BEA 2022). Hence, the back-of-the-envelope values for free data can be compared with other values for difficult to value products in the published national accounts.

The back-of-the-envelope calculations in this section rely on four case studies of free data. One case study focused on credit reports that record individual loan balances and individual repayment history. That case study used previous academic research to calculate a value of $0.6 trillion for the free individual credit data created in 2017 (Soloveichik 2023). Another case study in that same paper focuses on tax forms that record income and deductions. That case study used official government estimates of tax filing time to calculate a value of $0.4 trillion for the tax forms created in 2017. A third case study focuses on driving data such as licenses, tickets, and insurance claims. That case study used the academic literature and expert judgment to calculate a value of $0.4 trillion for the driving data created in 2017 (Soloveichik 2024). The final case study focused on marketing. That case study used BEA's published input-output (I-O) tables, industry sources, and the Occupational Employment and Wage Survey (OEWS) to estimate total marketing investment of $0.4 trillion in 2017 (Sveikauskas et al. 2023). This paper extrapolates from those four case studies to estimate of total free data creation in 2017.

Tracking free data increases measured GDP, but the GDP increase is lower than the value of free data creation. Free data which are created by households are out-of-scope for the national accounts. For example, an online beach map created by an individual in their leisure time would be included in household production rather than GDP. In addition, free data which are purchased by consumers are already included in personal consumption expenditures (PCE) and measured GDP. For example, destination wedding planners are already included in PCE whether they are helping a couple plan their ceremony or creating an online beach map for the wedding guests. Finally, free data which are complementary to new investment items whose purchase is already included in GDP are currently included in measured investment. For example, purchases of new fishing boats by tour guides are already included in investment whether the boat's value is due to the physical equipment or to the free beach maps created by the boat manufacturer. In practice, these counter examples account for only a minority of free data and the back-of-the-envelope calculations later in this section show a noticeable increase to measured GDP.

Similarly, tracking free data increases the measured asset stock, but the asset stock increase is much lower than the value of the free data stock. To start out, free data which do not add to measured investment also do not add to measured capital stock. In addition, some free data are implicitly included in the market value of natural resources. For example, natural resource exploration raises the market price for non-produced assets like land (Soloveichik 2022). Hence, tracking free data associated with land shifts some portion of land value from the tangible non-produced asset stock to the intangible produced asset stock without changing the total asset stock. Finally, some free data are implicitly included in the market value of used capital assets. For example, vehicle accident reports raise the

market price for used vehicles (Soloveichik 2024). Hence, tracking free data associated with vehicles shifts some portion of vehicle value from the tangible produced asset stock to the intangible produced asset stock without changing the total asset stock. Just like with measured GDP, these counter examples account for only a portion of free data.

Valuing Purchases of Free Data Creation Services

Businesses often sell data creation services to clients. A few industries sell free data creation as their primary product: law firms produce arguments that are presented to the court and therefore implicitly shared with anyone who comes to court, and medical laboratories produce test results that are given to patients and therefore implicitly shared with the patients' entire medical team. Many other industries bundle free data together with their primary product: doctors bundle medical diagnoses and health insurance claims together with disease treatment, schools bundle report cards and reference letters together with teaching, and property insurers bundle claims history together with insurance services. This paper uses BEA's published I-O tables, expert judgment, and other sources to calculate that industries whose primary product is free data creation had a total gross output of $1.5 trillion in 2017 and industries which bundle free data creation with other products had a total gross output[12] of $18.8 trillion in 2017.

Some types of data creation are already tracked in the published National Income and Product Accounts (NIPAs). NIPA table 5.6.5 explicitly tracks three data types: research and development, entertainment originals, and software. In addition, some data related to structures are implicitly included in measures of structures investment.[13] To avoid double counting, this paper excludes industries which sell those already tracked data types. This paper focuses on private data, so it also excludes data purchased by governments. After those exclusions, this paper estimates that industries whose primary product is free data creation services sold $1.4 trillion of gross output to businesses and consumers in 2017. In addition, industries which bundle free data creation services with other products sold $12.3 trillion of gross output to businesses and consumers in 2017.

This paper calculates possible ratios of free data to gross output for industries which bundle free data creation services with other products. The possible ratios are based on the three papers mentioned in the previous subsection that studied specific types of free data. One paper found that consumers

12. Insurance output is measured based on gross premiums rather than premiums after expected payments. As a result, measured output is about $1 trillion higher than reported in BEA's published I-O tables.

13. In particular, surveys for immediate construction are included in the cost of a newly built structure (United Nations 2008, sec 10.51) and mineral exploration is included in mining structures (United Nations 2008, sec 6.231).

implicitly purchased $274 billion of free individual credit data and $27 billion of free tax data from the retail and banking[14] sectors in 2017 (Soloveichik 2023). BEA's I-O tables report that retail and banking sold consumers a total of $1,881 billion of output in 2017. Hence, a ratio of free data to total output of 0.16 can be calculated from that paper.[15] Another paper found that customers implicitly purchased $145 billion of free insurance claims data from the insurance sector in 2017 (Soloveichik 2024). Based on the 2017 Economic Census, this paper calculates that motor vehicles collected gross premiums of $263 billion in 2017.[16] Hence, a ratio of free data to total output of 0.55 is calculated from that paper. The final paper found that the advertising industry before redefinitions (NAICS 5418) supplied $95 billion of free data to their customers (Sveikauskas et al. 2023). BEA's I-O tables report that that same industry had a total output of $144 billion. Hence, a ratio of free data to total output of 0.66 is calculated from that paper. A complete analysis of free data would likely require many separate case studies to cover all data types produced by the remaining industries and their $10 trillion of revenue. Nevertheless, the range of ratios (0.16 to 0.66) suggests that the value of data creation services sold by the remaining industries is likely within the range of $1.6 trillion to $6.6 trillion. This paper uses the weighted average of all three ratios, which is 0.23, to calculate that the remaining industries sold around $2.3 trillion of data creation services. The extrapolated $2.3 trillion of data creation services as a secondary product for the remaining industries can be added to the $0.6 trillion of data creation services as a secondary product that were measured in the case studies and the $1.4 trillion of data creation services sold as a primary product. In total, private businesses sold a total of $4.3 trillion of data creation services in 2017.

Valuing Own-Account Creation of Free Data

Almost every industry produces some own-account free data: human resource managers give employees tax forms and job references, communication specialists answer media questions about a company, and websites provide customer information like store hours or product prices. In total, the OEWS reports that employees[17] who likely specialize in the production of free data earned a total of

---

14. These two sectors are combined because they coordinate to process consumer payments.

15. The 0.16 ratio may be a lower bound. Retailers also produce loyalty card data, return history data, and other consumer data which was not studied in the cited paper.

16. BEA's measures of gross output exclude expected payments from insurance industry output (BEA 2022) and therefore record a much lower level of insurance industry output and a higher ratio of data to output.

17. The occupations which specialize in the production of own-account free data are: chief executives; advertising and promotions managers; marketing managers; sales managers; public relations and fundraising managers; financial managers; compensation and benefits managers; human resource managers; training and development managers; property, real estate, and community association managers; emergency management directors; agents and business managers of artists, performers, and athletes; claims adjusters, examiners and investigators; insurance appraisers, auto damage; compliance officers; human resource specialists; farm labor contractors; labor relations specialists; management analysts; meeting, convention, and event planners; training and development

$2.1 trillion in 2017. The paper then excludes government employees because this paper is focused on private data and excludes employees in industries that sell data creation services because their output was counted in the previous subsection. After those exclusions, the paper estimates that employees who likely specialize in the production of private own-account free data earned $860 billion in 2017.[18]

Own-account free data are hard to value because they are never sold in an arms-length transaction. For now, this paper uses specialist employee earnings to proxy for the value of own-account free data. The paper on tax forms and individual credit reports found that private businesses produced $207 billion of own-account free tax data in 2017 (Soloveichik 2023). For the same year, the OEWS shows that private businesses paid $104 billion to employees who likely specialize in the production of own-account free tax data. The $103 billion difference between data values and specialist employee earnings reflect both earnings for non-specialist support staff and non-labor costs like office space. Hence, this paper calculates a ratio of free data to specialist employee earnings of 1.99 from that case study. The paper on advertising found that private businesses produced $56 billion of own-account free advertising data in 2017 (Sveikauskas et al. 2023). In the same year, the OEWS showed that private business paid $84 billion to employees who likely specialize in the production of own-account free advertising data. Hence, this paper calculates a ratio of free data to specialist employee earnings of 0.66 from that case study. The lower ratio is not due to lower non-specialist support staff or lower non-labor costs for free advertising output. Rather, the paper on advertising assumed that only 30 percent of own-account advertising output is long-lived enough to be tracked as an asset in GDP. In contrast those two studies, neither individual credit report data nor insurance risk data have enough own-account data creation to calculate a meaningful ratio of free data to specialist worker earnings (Soloveichik 2023) (Soloveichik 2024). Just like with sold data creation services, the range of ratios (0.66 to 1.99) suggests that the value of own-account free data creation associated with the remaining $672 billion of specialist employee earnings is

specialists; market research analysts and marketing specialists; financial specialists; accountants and auditors; appraisers and assessors of real estate; budget analysts; credit analysts; financial analysts; personal financial advisors; insurance underwriters; financial examiners; credit counselors; tax preparers; web developers; landscape architects; cartographers and photogrammetrists; surveyors; environmental engineers; health and safety engineers, except mining safety engineers and inspectors; nuclear engineers; environmental engineering technicians; surveying and mapping technicians; life, physical, and social science occupations; community and social service occupations; legal occupations; education, training, and library occupations; art directors; merchandise displayers and window trimmers; public relations specialists; media and communication workers, all other; private detectives and investigators; gaming surveillance officers and gaming investigators; security guards; transportation security screeners; bookkeeping, accounting, and auditing clerks; payroll and timekeeping clerks; brokerage clerks; correspondence clerks; credit authorizers, checkers, and clerks; customer service representatives; eligibility interviewers, government programs; interviewers, except eligibility and loan; human resource assistants, except payroll and timekeeping; receptionists and information clerks; and forest and conservation workers.

18. The OEWS is a rolling 3-year panel, so the 2018 wave is used to measure 2017 earnings.

likely within the range of $0.4 trillion to $1.3 trillion. This paper uses the weighted average of both ratios, which is 1.4, to calculate that the remaining occupations produced approximately $0.9 trillion of own-account free data. This $0.9 trillion can be added to the $0.3 trillion of own-account free data studied earlier to get a total of $1.2 trillion.

Measuring household data production is very difficult. One major issue is that household time devoted to data creation is rarely reported as a separate activity on the American Time Use Survey. For example, an individual who answers questions during a traffic stop is involved in a data creation activity. But they are not likely to report the time spent answering questions separately from general driving time. Another major issue is that household data are often mingled with business data. For example, credit reports mingle information on consumer loans, housing loans, and small business loans. Hence, it is not always obvious which sector is creating data or which sector owns the data. This paper uses the ratio of business data creation to total service sector output to proxy for the ratio of household data creation to total household service output. Based on that proxy, the paper calculates that households created $1.1 trillion of free data in 2017. This calculation is very approximate.

Figure 8. Free Data Creation by Data Type

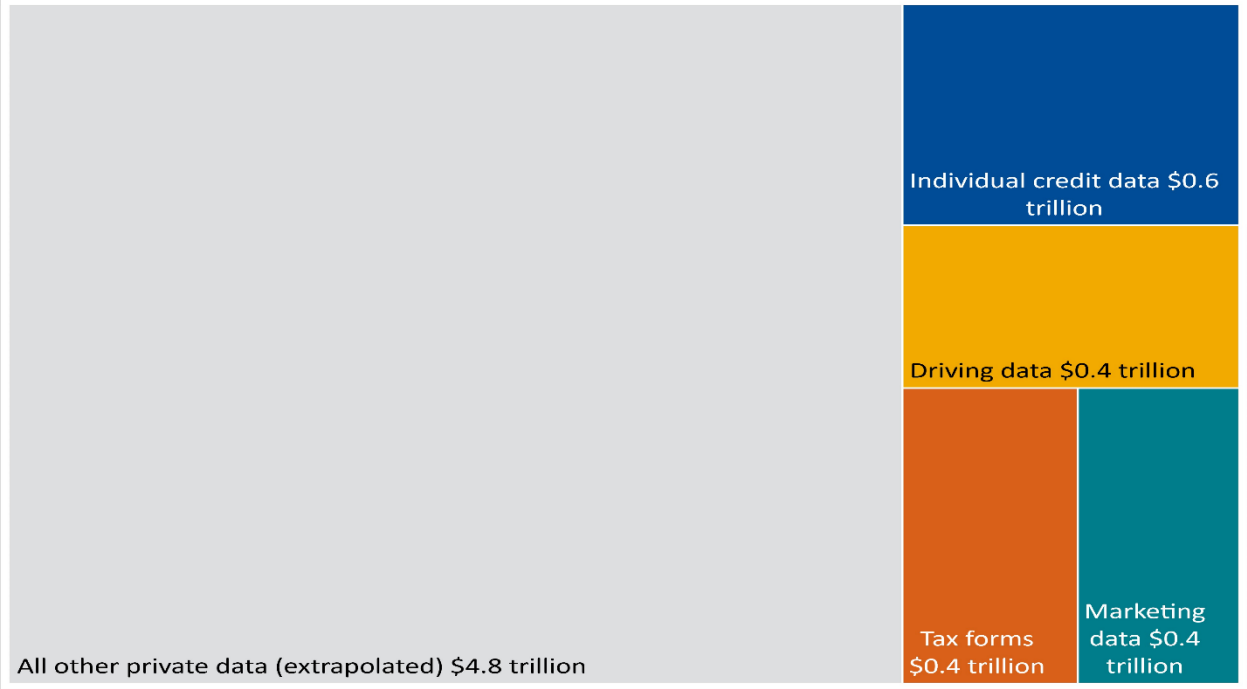


Figure 8 shows that private free data creation totaled approximately $6.6 trillion of free data in 2017. These $6.6 trillion of data can be divided between the $0.6 trillion of individual credit report creation and $0.4 trillion of tax form creation studied in Soloveichik 2023, the $0.4 trillion of driving data creation

studied in Soloveichik 2024, the $0.4 trillion of marketing data creation studied in Sveikauskas et al. 2023, and $4.8 trillion of other free data. Even if national accountants consider the $4.8 trillion of other free data to be too speculative to include in the economic statistics, the $1.8 trillion of free data studied in those four case studies may be large enough to change measured GDP and measured household production noticeably.

Impact of Free Data on the Level of Output in 2017

Free data can be created, funded, and owned in many different ways. This paper examines five separate scenarios: a) data created, funded, and owned by consumers; b) data created by businesses, funded by consumers, and owned by consumers; c) data created by businesses, funded by businesses, and owned by consumers; d) data created by businesses on behalf of the government, funded by businesses, and owned by governments; and e) data created by businesses, funded by businesses, and owned by businesses. The paper describes these five scenarios as: a) household production of consumer durables; b) purchased consumer services reclassified as consumer data purchases; c) non-wage compensation of data; d) in-kind tax of data; and e) business investment of data.

The remainder of this section uses the case studies discussed earlier and expert judgment to split the $6.6 trillion of free data creation between the five scenarios mentioned above. The theoretical frameworks developed in sections 2 through 4 apply to all five data creation scenarios. However, the impact of free data on measured GDP data depends on their creator, their funder, and their owner. This paper briefly describes how each scenario impacts the national accounts.

Scenario a): Consumers often create free data by themselves that they then use in household production. For example, someone looking for romance might write a profile and post it on a dating website. In other words, the model developed in the theoretical framework is tweaked to allow human capital rather than physical capital to be complementary to data. In this tweak, the capital owner is a person who owns themselves rather than businesses which own capital that can be bought and sold. This paper calculates that tracking consumer-created data as consumer durables raises household production by a portion of the $1.1 trillion of newly recognized household data creation.[19] Whether consumer-created data are used for work or leisure, the capital services associated with these data are out-of-scope for GDP but in-scope for a household production account which includes consumer durable

19. Some consumer-created data were previously tracked as other household production (Bridgman et al. 2022).

services together with household labor (Bridgman et al. 2022). This paper calculates that consumer-created data yielded $1.5 trillion of capital services each year.[20]

Scenario b): Consumers also purchase free data creation services from business and then use these purchased data in households. For example, someone looking for a job might hire a job coach to help them write a resume,[21] or someone looking for romance might hire a matchmaker to help them write a profile. Based on the consumer shares reported in BEA's detailed I-O data for 2012, this paper calculates that the consumer sector purchased about forty percent of the $4.3 trillion of sold free data creation services. This paper shifts these $1.8 trillion of purchased data from personal consumption expenditure (PCE) services to PCE durables without changing total PCE. As a result, tracking consumer-funded data does not change measured GDP. However, calculations show that consumer-funded data yielded another $2.7 trillion of capital services each year. These consumer-funded data services may be counted together with consumer-created data services in a household production account.

Scenario c): Businesses sometimes create and fund free data that are then given to consumers. Most obviously, employers are required by law to provide their employees with certain tax forms. In addition, employers often provide references for current employees who need to verify income or former employees who need to verify work experience. Similarly, businesses also give data to self-employed business owners or even non-workers who are somehow associated with the business. This paper tracks these free data as components of both non-cash compensation and PCE. This treatment is similar to the treatment of employer-provided health insurance and other non-cash benefits (BEA 2022). Using BEA's published I-O tables and expert judgment, this paper estimates that approximately one-tenth of the $4.3 trillion of purchased data creation services and approximately one-half of the $1.2 trillion of own-account data creation represents non-cash compensation for either employees or self-employed business owners. Based on those estimates, tracking free data creation is calculated to increase both measured personal income and measured PCE by $1.0 trillion in 2017. In addition, the $1.5 trillion in capital services from business-funded data which are given to consumers may be counted together with other consumer data services in the household production account.

Scenario d): Businesses also create and fund free data that are then given to governments. For example, a court might subpoena a business record related to a legal dispute or a statistical agency might send mandatory surveys to businesses. This paper tracks these free data as taxes in-kind and includes them in

20. This calculation is based on a lifespan of seven years and a real rate of return of 7 percent.

21. One might think that consumer-funded data that is used on-the-job should be a business asset. In fact, PCE includes job-related expenses like vehicles used for commuting and business attire.

measured business output and measured government investment. BEA's standard formula for calculating government output includes a measure of capital services that is based on the consumption of fixed capital for government assets. Accordingly, tracking free data as government assets raises measured GDP twice: once when their creation is counted in business output and again when their depreciation is counted in government output. This paper uses previous research on the share of financial data used by the government sector (Soloveichik 2023) to estimate that 11 percent of the $5.5 trillion in business data creation represents in-kind taxes. Based on that share, free data increase both measured business output and measured government output by $0.6 trillion in 2017.

Scenario e): Finally, the remaining $2.1 trillion of free data are tracked as business investment. Unlike the government sector, measured business output does not depend on measured capital services. Accordingly, tracking free data as a business asset only raises measured GDP once, when its creation is counted in business output. Based on that treatment, this paper calculates that measured business investment rises by $2.1 trillion in 2017.

Figure 9 shows data creation for each of those five scenarios. By assumption, total data creation sums up to the $6.6 trillion shown in Figure 8.

Figure 9. Free Data Creation by Scenario

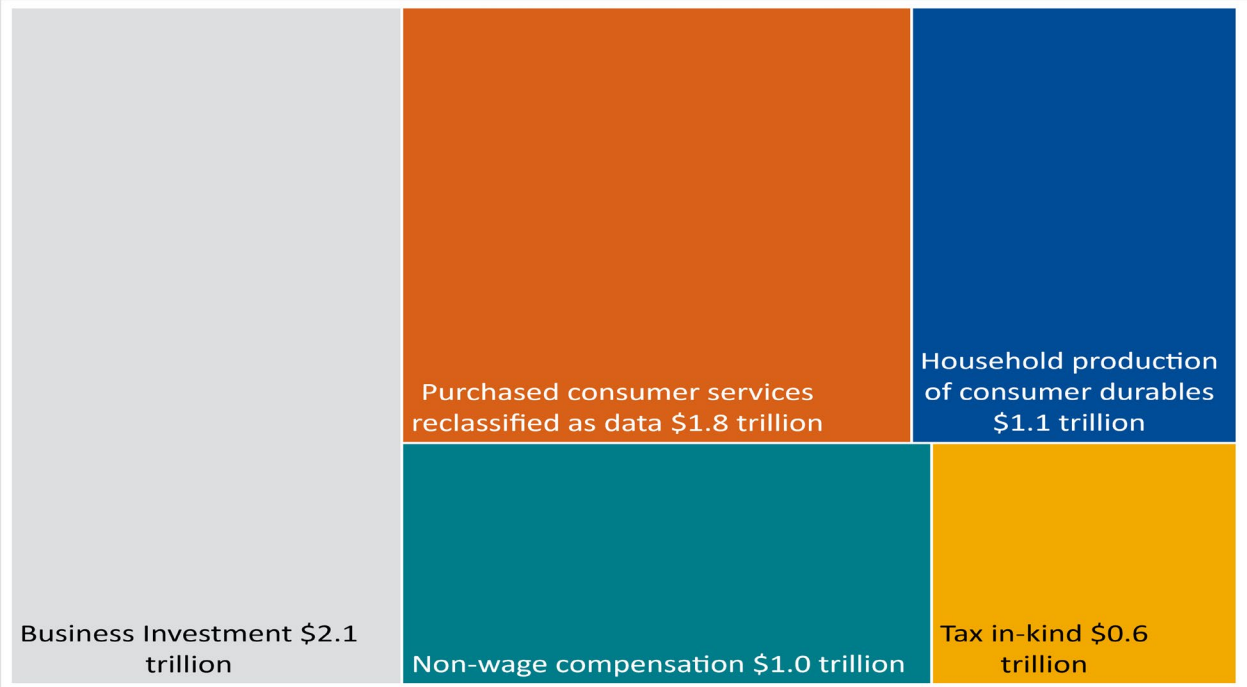


Figure 10. GDP Impact of Tracking Free Data by Scenario

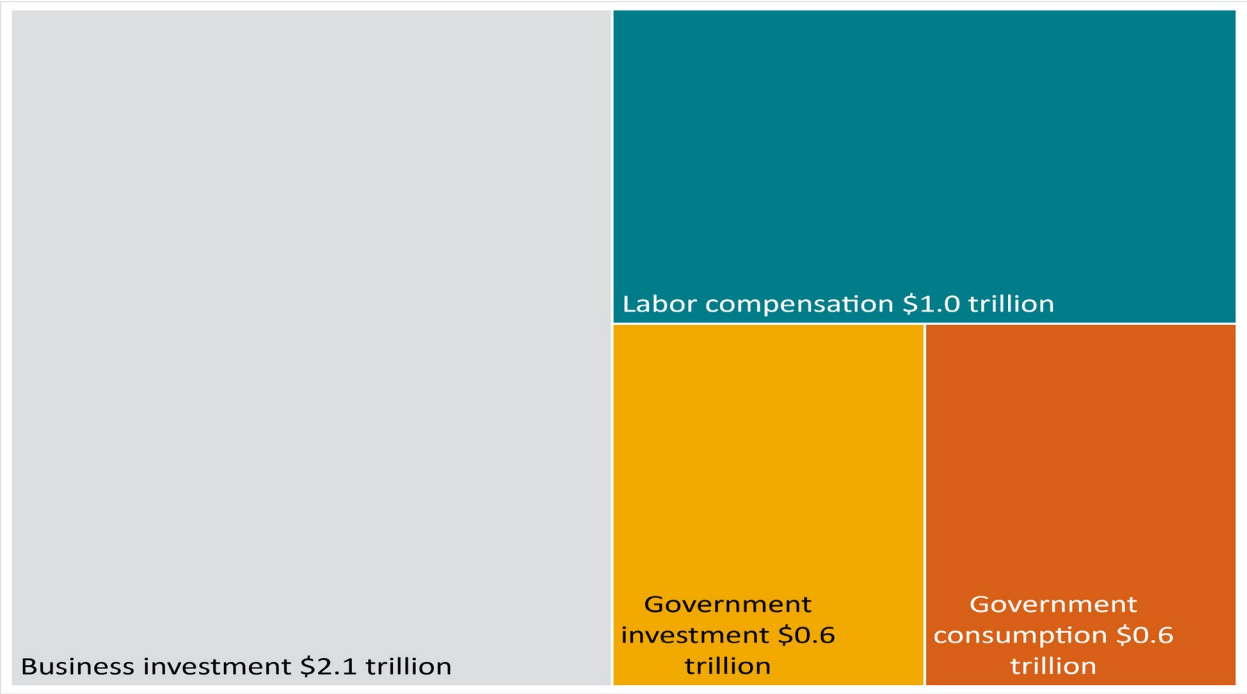
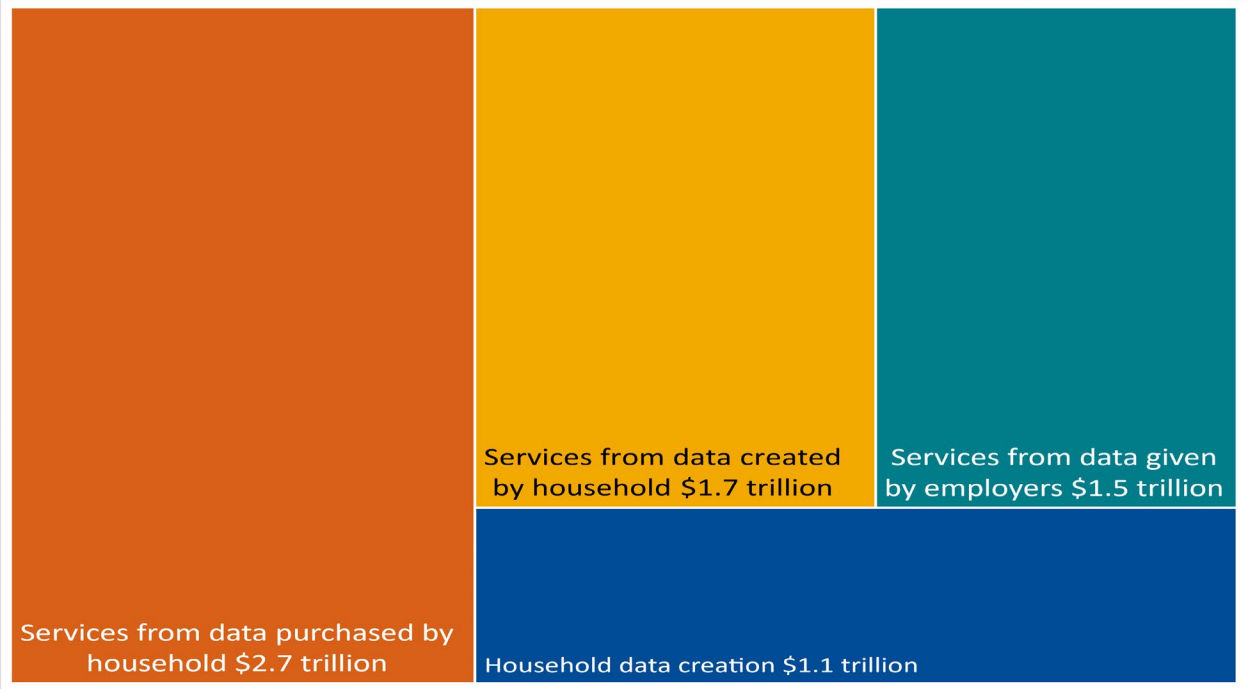


Figure 11. Household Production Impact of Tracking Free Data by Scenario



To summarize figures 10 and 11, tracking free data changes both GDP and household production noticeably. Broadening GDP to include data investment by businesses, data-related worker compensation, data created by businesses and given to the government as a tax-in-kind, and depreciation on government-owned data increases measured GDP from $19.5 trillion to $23.8 trillion (19.5+1.0+0.6+0.6+2.1). Broadening household production to include both household data creation and consumer durable data services increases measured household production from $4.6 trillion to $11.6 trillion (4.6+1.1+1.7+2.7+1.5) in 2017. These revisions may be large enough to change national accountants' understanding of both GDP and the household sector. To illustrate the potential impacts of these revisions, this paper presents the impact on nominal GDP, real GDP quantities, nominal household production, and real household production of the four data types studied in previous research (Soloveichik 2023) (Soloveichik 2024) (Sveikauskas et al. 2023).

Figure 12. Revision to Nominal GDP Levels in Case Studies from Tracking Data

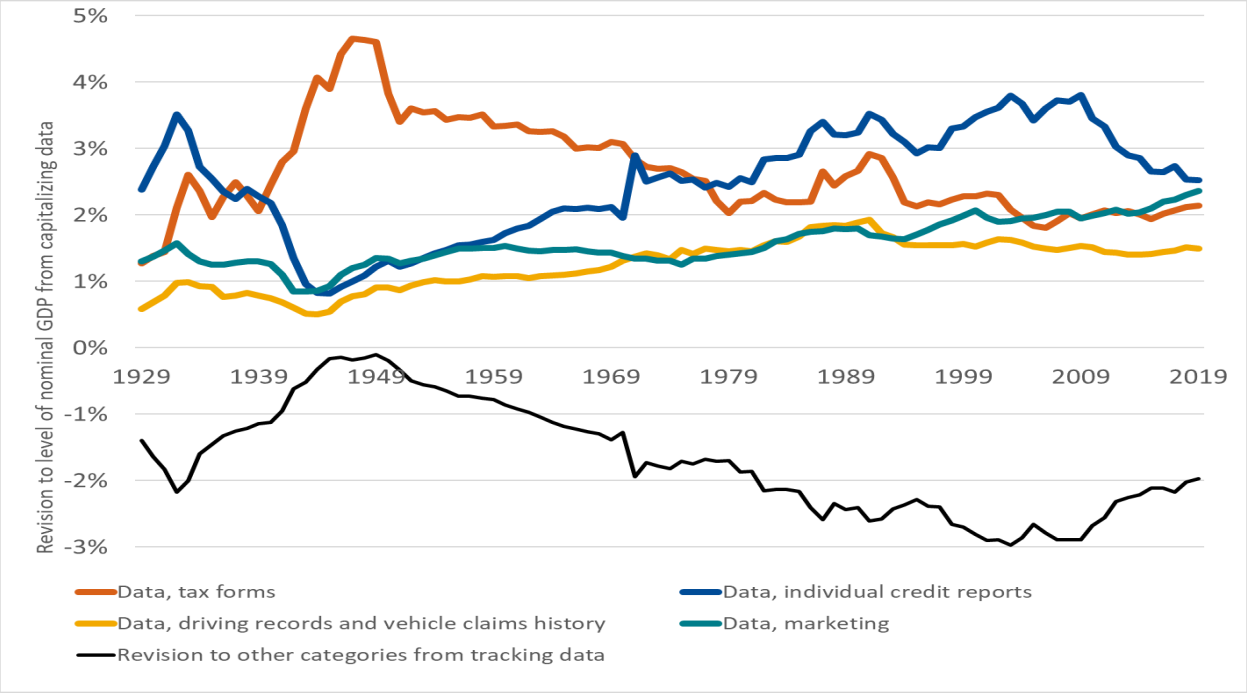


Figure 13. Revision to GDP Quantity Indexes in Case Studies from Tracking Data

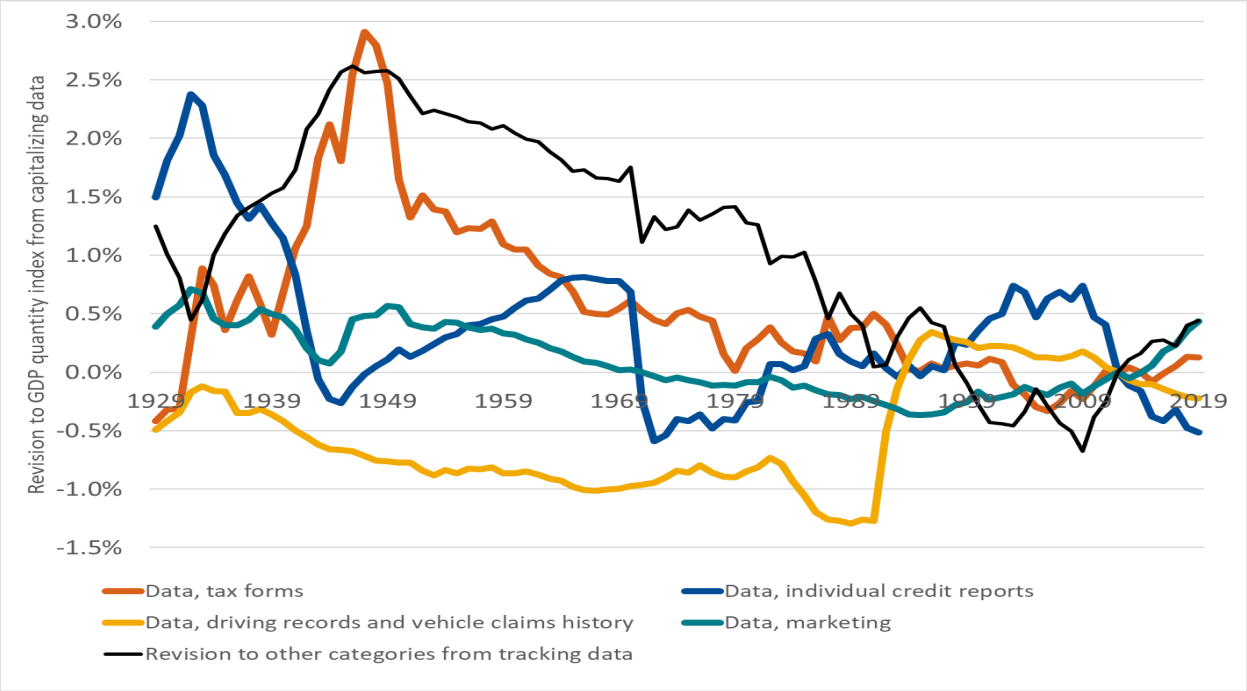


Figure 12 shows that nominal GDP growth does not change much when the four data types studied in detail are capitalized. However, this null result is likely due to the fact that these four data types are associated with industries whose nominal growth rate has approximately tracked the nominal GDP growth rate. BEA's published industry accounts show that industries which sell free data creation services have generally experienced faster nominal growth than the overall economy. In addition, the OEWS shows that occupations which likely specialize in the creation of own-account free data have generally experienced faster nominal earnings growth than other workers. Hence, it seems plausible that nominal GDP growth might increase if all types of free data were capitalized. This conclusion is very speculative and would require much more research to verify.

In contrast to figure 12, figure 13 shows that each data type studied has its own impact on GDP quantity growth. The major driver of the different growth rates is variation in price growth across data types. Historical prices for each data type depend on regulation (Soloveichik 2023) and data sharing infrastructure (Soloveichik 2024) as well as input costs. A full analysis of the real GDP impact of including all free data would likely require separate case studies for every single data type studied.

Figure 14. Revision to Nominal Household Production Levels in Case Studies from Tracking Data

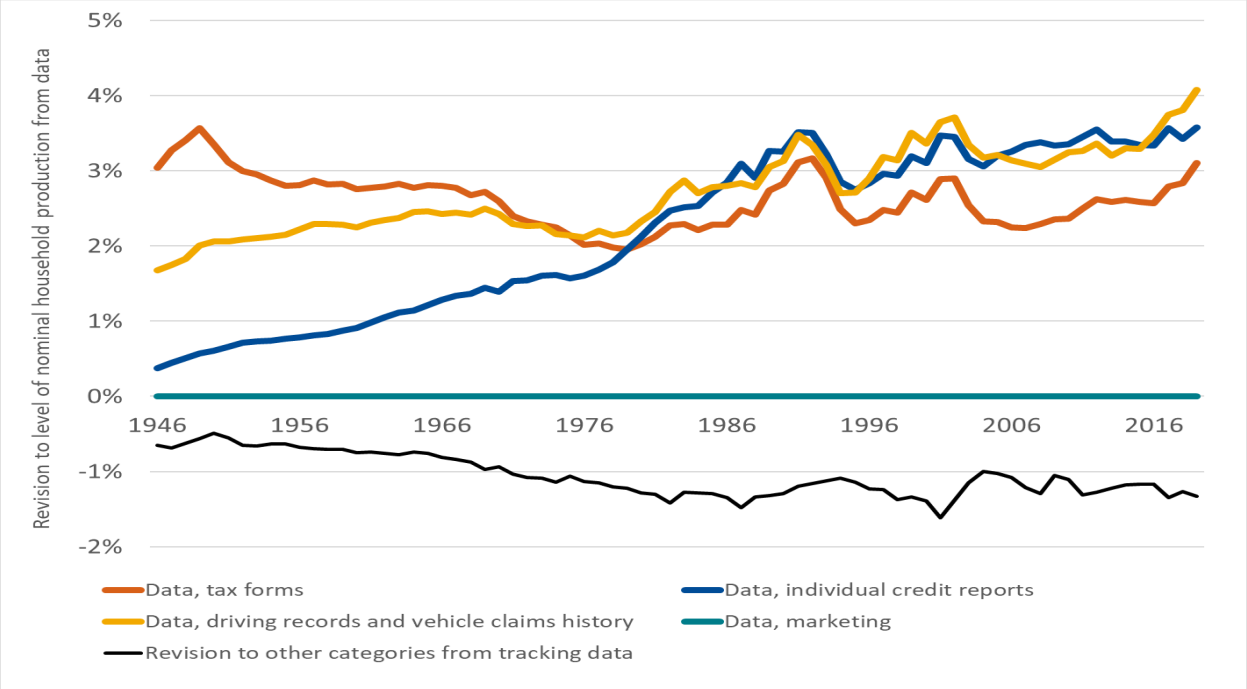
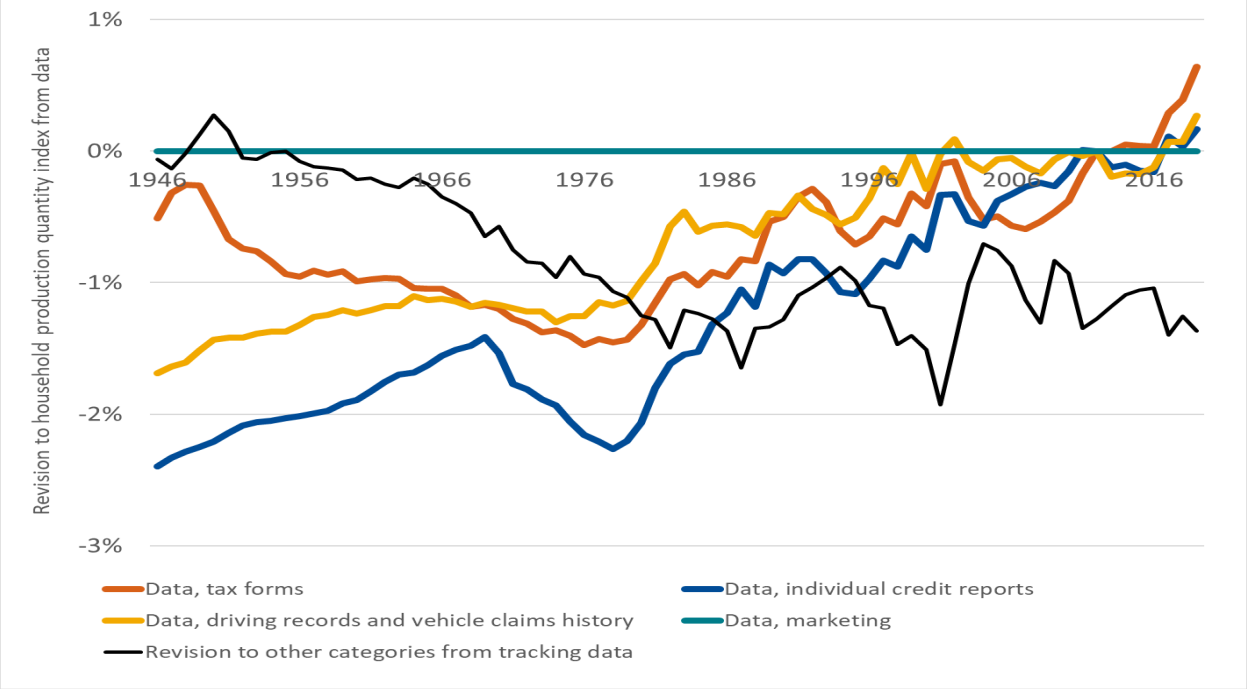


Figure 15. Revision to Real Household Production Quantities in Case Studies from Tracking Data



Figures 14 and 15 shows that data accounts for a growing share of household production. A portion of this relative growth can be explained by an increase in the average amoπunt of time spent on data creation. For example, the number of speeding tickets issued per capita increased between 1946 and 2019. However, the lion's share of this relative growth can be explained by a decline in the share of time spent on other household production (Aguiar and Hurst 2007). As a result, a steady amount of time spent on data creation accounts for an ever growing share of household production. By themselves, the upward revisions to growth shown in figures 14 and 15 are not enough to change the qualitative trends in household production reported in BEA's satellite account (Bridgman et al. 2022). However, the upward revisions are large enough that the qualitative trends would change if other types of free data had the same relative impact as the types of free data shown in figures 14 and 15.

Conclusion

This paper developed a theoretical framework in which data can either be sold or given for free. It then solved that theoretical framework to identify plausible parameters where the maximum possible sales revenue from data is lower than the capital revenue increase associated with free data. Free data are particularly dominant when data are complementary to other data (Coyle 2022) or when piracy reduces the potential revenue from data sales.

The paper then argued that that private free data are a very large asset type. First, the paper presented four previous case studies (Soloveichik 2023) (Soloveichik 2024) (Sveikauskas et al. 2023) which studied a total of $1.8 trillion of private free data creation in 2017. Based on the ratio of free data creation to industry output reported in those case studies, the paper extrapolated that the total private creation of free data may have been $6.6 trillion in 2017. To be clear, the $6.6 trillion value is only a back-of-the-envelope estimate that is presented only for discussion purposes. Many more case studies need to be done before free data could be measured precisely enough to be included in BEA's published economic statistics. Nevertheless, the back-of-the-envelope estimates demonstrated that free data are large enough to change measured GDP and measured household production noticeably. These results suggest that free data deserve more research attention going forward.

Bibliography

Acquisti, A., C. Taylor, and L. Wagman (2016) "The Economics of Privacy," *Journal of Economic Literature* 54(2), pages 442–92.

Aguiar, M., and E. Hurst (2007) "Measuring Leisure: The Allocation of Time over Five Decades," *Quarterly Journal of Economics* 122(3), pages 969–1006.

Baker, L., Bundorf, M., and Kessler, D. (2015) "Expanding Patients' Property Rights in Their Medical Records," *American Journal of Health Economics* 1(1), pages 82–100.

Bridgman, B., A. Craig, and D. Kanal (2022) "Accounting for Household Production in the National Accounts: An Update 1965–2020," *Survey of Current Business* 2022 (2).

Bureau of Economic Analysis (2022) "Concepts and Methods of the U.S. National Income and Production Accounts," https://www.bea.gov/resources/methodologies/nipa-handbook/pdf/all-chapters.pdf, accessed January 17, 2023.

Calderon, J.B., and D. Rassier (2022) "Valuing Stocks and Flows of Data Assets for the U.S. Business Sector," Presentation to the BEA Advisory Committee Meeting on May 13.

Coyle, D. (2022) "Socializing Data," *Daedalus* 151 (2), pages 348–359.

Coyle, D., and Li, W. (2021) "The Data Economy: Market Size and Global Trade," *ESCoE Discussion Paper* No. 2021–09.

De Groot, J. (2022) "What is Data Encryption? Definition, Best Practices & More," *DataInsider*, Digital Guardian, posted November 7, 2022 and accessed February 2023.

Drolet, M. (2016) "3 Ways to Protect Data and Control Access to It," *CSO Online*, posted May 10, 2016 and accessed February 2023.

Eurostat (2020) "Recording and Valuation of Data in National Accounts," Working paper 3.4 https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.20/2020/mtg1/3.4_Recording_of_Data_in_NA_Eurostat_June_2020_after_SG_comments.pdf.

Franklin, B. (1735) *Poor Richard's Almanack* self-published.

Jones, C., and C. Tonetti (2020) "Nonrivalry and the Economics of Data," *American Economic Review* 110(9), pages 2819–58.

Leppamaki, M. and M. Mustonen (2009) "Skill Signaling with Product Market Externality," *The Economic Journal* 119 (539), pages 1130–1142.

Liu, G. and Fraumeni, B. (2020) "A Brief Introduction to Human Capital Measures," *NBER Working Paper* 27561.

Mitchell, J., M. Lesher, and M. Barberis (2022) "Going Digital Toolkit Note: Measuring the Economic Value of Data," OECD Directorate for Science, Technology and Innovation.

Page, Wolfberg & Wirth (2020) "An Imaginary Barrier: How HIPAA Promotes Bidirectional Patient Data Exchange with Emergency Medical Services," posted at Nemsis.org and accessed February 2023.

Rassier, D., Kornfeld, R., and Strassner, E. (2019) "Treatment of Data in National Accounts," *Paper prepared for BEA Advisory Committee*, posted in May 2019 and accessed in February 2023.

Reiss, J. (2021) "Public Goods," *Stanford Encyclopedia of Philosophy*, accessed in February 2023.

Soloveichik, R. (2022) "Natural Resource Exploration as Intangible Investment," *BEA Working Paper* 2022–9.

Soloveichik, R. (2023) "Capitalizing Data: Case Studies of Tax Forms and Individual Credit Reports," *BEA Working Paper* 2023–7.

Soloveichik, R. (2024) "Capitalizing Data: Case Studies of Driving Records and Vehicle Insurance Claims," unpublished manuscript available upon request.

Statistics Canada (2019) "The Value of Data in Canada: Experimental Estimates," posted on July 10, 2019, at https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm and accessed in October 2022.

Sveikauskaus, L., C. Garner, M. Russell, R. Soloveichik, J. Bessen, P. Meyer (2023) “Marketing, Other Intangibles, and Output Growth in 61 United States Industries,” *BEA Working Paper* 2023–11.

United Nations Statistics Division. (2008). *Updated System of National Accounts 2008*. Accessed October 2022. http://unstats.un.org/unsd/nationalaccount/sna2008.asp.