IARIW – CIGI 2023

IARIW – CIGI 2023

Thursday, November 2 – Friday, November 3

Private Funding of "Free" Data: A Theoretical Framework

Rachel Soloveichik (Bureau of Economic Analysis)
Rachel.Soloveichik@bea.gov

Paper prepared for the Conference on The Valuation of Data November 2 – November 3, 2023
Session 4: Methodologies of Data Valuation: Special Topics
Time: Thursday, November 2, 2023 [3:15PM-4:45PM EST]

Private Funding of "Free" Data: A Theoretical Framework

By Rachel Soloveichik[1]

This paper develops a theoretical framework in which privately funded data can either be sold or given for "free" by the owner of a complimentary asset. For example, maps of tropical beaches can either be sold to individuals who have already booked a local hotel room or given for free by a local hotel to everyone who might be interested in a tropical vacation. This paper then solves that theoretical framework and identifies plausible parameter regions at which the data sales revenue is lower than the revenue increase for the complementary asset associated with free data. Data sales revenue is particularly likely to be lower than the revenue increase 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 concludes with a back-of-the-envelope calculation estimating the value of privately funded free data in the United States. To start out, the paper reviews four previous case studies which together studied $1.8 trillion of privately funded free data creation in 2017 (Soloveichik 2023a) (Soloveichik 2023b) (Sveikauskas et al. 2023). This paper then uses the U.S. Bureau of Economic Analysis' published input-output tables and the Occupational Employment and Wage Survey to extrapolate total private creation of free data of $6.6 trillion. In 2017, including free data raises measured gross domestic product by more than 20 percent and raises measured household production by more than 100 percent.

JEL Codes: D12, E01, and G14

---

[1] The views in this paper reflect those of the authors and not necessarily those of the Department of Commerce or Bureau of Economic Analysis. Comments can be sent to Rachel.Soloveichik@bea.gov.

Introduction

The United States economy depends on data. Data are an integral part of many services: 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 use data like job references to evaluate workers and governments use tax forms to determine eligibility for benefits. Finally, households use data to determine which businesses to buy from and which individuals to socialize with. 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).

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. Some previous national accounting papers studying data have 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, a hotel might post maps of local beaches and hiking trails on a website for the world to see or a consumer loan applicant might share their credit score with every bank that offers loans in their area. This paper focuses on those “free” data.

This paper first develops a theoretical framework to study how data are distributed. That theoretical framework is solved and plausible parameters for which the data sales revenue is lower than the value of free data to the owner of a complementary capital asset are identified. To support that theoretical framework, this paper presents a back-of-the-envelope calculation that estimates total private free data creation was \$6.6 trillion in 2017. Including this \$6.6 trillion of free data in the U.S. Bureau of Economic Analysis’ (BEA’s) economic statistics would raise measured gross domestic product (GDP) by more than 20 percent and raise measured household production by more than 100 percent in 2017.

This paper is divided into four sections. The first section discusses the unique features of data. The second section develops a simple theoretical framework with only one capital asset and only one data type. The second section then 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 uses case studies of four separate types of free data and an extrapolation to calculate the total value of all privately funded free data creation in 2017. That section then describes how BEA's published economic statistics might change if all types of privately funded free data were tracked as intangible capital assets. To be clear, the calculations presented in this section are a back-of-the-envelope estimate that is presented only for discussion purposes. Much more work needs to be done before free data could be included in GDP.

Section 1: Data Definitions and Features

Defining Data

This paper defines data as information that can be copied easily. 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 their analysis to include all types of data. Some types of data are simple enough to be stored in a text 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, and are sometimes stored entirely electronically. Modern data are typically recorded digitally, but data can also be formatted recorded on paper, on DNA strands, or even in an oral tradition. The discussion in this paper focuses on data with a useful lifespan of more than one year because those data might be tracked as produced capital assets in future national accounting guidelines (Rassier et al. 2019) (Eurostat 2020). However, the theoretical framework developed in this paper also applies to data with a shorter useful lifespan. Those short-lived data might be tracked as intermediate inputs or inventory.

Data describing an asset are conceptually separate from the asset itself. For example, maps of the wildlife on tropical beaches are conceptually separate from either the wildlife or the tropical beaches. In national accounting terms, maps are produced intangible assets while both wildlife and tropical beaches are non-produced natural resources. Similarly, maps of a hotel are produced intangible assets while the hotel itself is a produced structure. Finally, maps showing restaurants with good chefs are produced intangible assets while the chefs themselves are people who own their own human capital. This paper studies the value of data but not the value of other assets.

Unique Features of Data

The minimal cost of copying is the single most important data feature for this paper. In the extreme, digital data can be copied almost instantly at virtually zero cost. However, other data formats can also be copied relatively quickly at a copying cost far below their original creation cost. Pieces are paper can be Xeroxed, DNA strands can be replicated with cultivation and breeding, and oral traditions can be retold. Because copying costs are already minimal, technologies which make copying 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 Xeroxed to electronic maps that can be instantly copied. Accordingly, this paper does not bother splitting data by the format it is recorded in. Instead, it assumes that the copying costs are negligible and social welfare is maximized when data are widely shared between authorized users (Coyle 2022) (Jones and Tonetti 2020).

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 fun smartphone game to anyone who does so. This paper assumes that the small positive profit earned from some free data and the small negative profit earned from other free data cancel out. As a result, the data user's average price for free data can be assumed to equal zero.

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. The remainder of the discussion calls usage without permission "piracy". A single large firm that has multiple employees using the same data can be just as vulnerable to piracy as a data seller who sells data to multiple firms. Hence, many data sellers face a binary choice of either restricting access (Drolet 2016) or accepting the fact that many data users will pirate the data.

To be clear, acceptance of piracy does not mean zero privacy. Data holders 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 and the set of authorized users includes every entity with a genuine need to know. For example, a doctor might record medical data in a patient's chart. Those medical data are free in the sense that the patient can send their chart to a new doctor for only a small fee (Baker et al. 2015) and doctors treating the patient in an emergency can use the data without payment (Page et al. 2020). However, medical data are protected by the law and doctors cannot sell patient data to curious journalists or give them to nosey family members without authorization from the patient. For simplicity, this paper assumes that data security is always sufficient to protect genuine privacy concerns (Acquisti et al. 2016) (Acemoglu et al. 2022).

Section 2: Simple Theoretical Framework with One Capital Asset and One Data Type

This paper develops a simple economy to study data distribution and usage. That 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 paper assumes that r is a fixed constant which does not necessarily maximize profits for the capital owner. Results are qualitatively similar if the physical capital owner is allowed to pick a profit-maximizing r, but then the equations are more complex to solve. In addition, there is a data owner who rents their intangible data asset at a rental rate of p that is chosen to maximize their profits. Finally, there are n separate firms which buy capital services and data services. These n firms all have constant elasticity of substitution (CES) production functions which combine capital services and data services 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 $s_1 \geq s_2 \geq \ldots \geq s_n$. This model assumes that each firm decides how much capital, k, to rent and how much data, d, to buy. 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 than they would otherwise. Capital is rival, so the total capital supply of K is split proportionally to firm demand in case of a shortage. Data are nonrival, so each firm is given the maximum quantity of data, D, in case of a shortage. For any i between 1 and n, the revenue and profits of firm i can be written as:

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

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

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 has walking maps for tourists. For example, a climate 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 resell those hotel rooms to individual tourists. These n travel agencies differ in their skill selling hotel rooms, but they all sell more hotel rooms when they can use maps to match customers with the room best suited to them.

When capital is in abundant supply, there are no externalities between the n firms which use capital and data. However, there is an externality between the capital owner and the data owner. When the capital rental rate is high, firms demand less data and therefore the data seller earns lower revenue. And when the data rental rate is high, firms demand less capital and therefore the capital owner earns lower revenue. Due to that mutual externality, both the capital owner and the data seller would prefer that the rental rate for the other input be lower. Similarly, both the capital owner and the data seller would prefer that the supply of the other input be high enough that there are no shortages. 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 $-\infty$.

When capital is in short supply, there are 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. Accordingly, each firm would prefer that its rivals not use any data. In a market with flexible prices, a shortage of capital raises the price of capital and therefore creates a positive pecuniary externality for the capital owner. But in this model, the price of capital is fixed and therefore the capital owner is indifferent between demand precisely equal to the supply of capital and demand higher than the supply of capital.

Capital Usage and Data Creation when a Social Planner Controls All Outcomes

This subsection considers a world in which data distribution is done by a social planner whose goal is to maximize total profits. Given the distribution of data, each firm then decides how much capital to rent.[2]

[2] The social planner does not decide capital allocations, and total profits may be lower than the maximum possible if the capital owner sets a rental rate, r, that is too high.

In the tropical island example, the social planner might be a local government which wants to maximize the region's business profits or a trade association of hotels and travel agencies. Regardless of the identity of the social planner, they are assumed to be capable of implementing their preferred decisions.

The social planner's problem is trivial. Data sharing is assumed to have zero marginal cost. Therefore, the social planner distributes the total amount of data, D, to every firm. The data seller needs to receive enough revenue to cover its operating expenses. So, the social planner charges each firm a fixed fee for data that does not vary with the quantity of data used. Given that universal data usage, each firm's profit function can be expressed as:

(3) Profit of firm i = $(\alpha k_i^{\rho}+(1-\alpha)*(D*s_i)^{\rho})^{(1/\rho)} - rk_i -$ fixed 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) $dProfits/dk_i = (\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1/\rho)-1} *\alpha k_i^{\rho-1} - r = 0 \rightarrow (\alpha k_i^{\rho}+(1-\alpha)*(Ds_i)^{\rho})^{(1-\rho)/\rho} *\alpha k_i^{\rho-1} = r \rightarrow$

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

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

We can calculate the total value added for all businesses:

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

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

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

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

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

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

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

$\boxed{k_i = s_iK/(s_1+ s_2+ ...+ s_n)}$

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

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

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

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: r, K, s_1 to s_n, α, ρ, D. Figures 1 and 2 in a later subsection show the social benefits of data for two selected parameter regions. For now, the theoretical framework discusses some general patterns to the solution of equations (5) and (7). Most obviously, the total social value is higher when the n firms have more total skill using data. In addition, the total social value 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 each firm makes decisions to maximize their individual profits without considering either positive or negative externalities. Decisions are made in three sequential steps. First, the data owner sets their data sales price, p. Second, the firms decide how much data to buy. 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 purchase and the rental rate of r. Next, it calculates how much data each of the n firms buys given the amount of capital each expects to rent and the data price of p. Finally, it calculates a data price which maximizes data seller revenue.

In the tropical island example, the hotel owner starts out with a pre-set room rental rate of r. The map owner then picks a price for their maps that maximizes their data sales 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 rented maps with rented rooms to sell hotel rooms to individual tourists.

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):

(8) Profit of firm i $= (\alpha k_i^{\rho}+(1-\alpha)*(d_i*s_i)^{\rho})^{(1/\rho)} - rk_i - pd_i$

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

(10) Profit of firm i = $(\alpha\{(1-\alpha)^{(1/\rho)}(d_is_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}\}^\rho+(1-\alpha)*(d_i*s_i)^\rho)^{(1/\rho)}$

$- r\{(1-\alpha)^{(1/\rho)}(d_is_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(1/\rho)}\} - pd_i =$

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

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

Note that 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 their data price high enough that the marginal firm is precisely indifferent between buying data or not:

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

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

0 otherwise.

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

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

Equation (11) is long and would be difficult to solve mathematically, but all the terms in the equation are constants and so it is straightforward to solve numerically. 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 social value of data shown in equation (5). In other words, the private value of data is lower than the social value.

Capital allocation when capital is in short supply can be solved using a similar profit function as equation (4) and a similar marginal profit function as equation (6):

(12) $d_is_i*K/(d_1s_1+ d_2s_2+\ldots+ d_ns_n) = k_i$

(13) Profit of firm i = $(\alpha\{(d_i s_i)*K/(d_1s_1+ d_2s_2+\ldots+ d_ns_n)\}^{\rho}+(1-\alpha)*(d_i*s_i)^{\rho})^{(1/\rho)}$

$-\ r(d_i s_i)*K/(d_1s_1+ d_2s_2+\ldots+ d_ns_n)\ -\ pd_i =$

$[\alpha(d_i s_i)^{\rho}* K^{\rho}/(d_1s_1+ d_2s_2+\ldots+ d_ns_n)^{\rho}+(1-\alpha)*(d_i*s_i)^{\rho}]^{(1/\rho)} -\ r(d_i s_i)*K/(d_1s_1+ d_2s_2+\ldots+ d_ns_n) - pd_i=$

$$\boxed{d_i *\{[\alpha s_i^{\rho}* K^{\rho}/(d_1s_1+ d_2s_2+\ldots+ d_ns_n)^{\rho}+(1-\alpha)*s_i^{\rho}]^{(1/\rho)} -\ rs_i*K/(d_1s_1+ d_2s_2+\ldots+ d_ns_n) - p\}}$$

Equation (13) is not quite linear with d_i because the term $K/(d_1s_1+ d_2s_2+\ldots+ d_ns_n)$ depends slightly 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 $K/(d_1s_1+ d_2s_2+\ldots+ d_ns_n)$ is almost constant and firms are very likely to be at a corner solution. This paper solves for equation (13) 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:

(14) Revenue of the data seller = $\sum pd_i$; d_i =D if

$s_i\{(\alpha K^{\rho}/(Ds_1+ Ds_2+\ldots+ Ds_i)^{\rho}+(1-\alpha))^{(1/\rho)} -\ rK/(Ds_1+ Ds_2+\ldots+ Ds_i)\} \geq p$

0 otherwise.

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

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

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

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 hotel sets the room rental rate the same as before and the mapmaker sells the same maps as before. But now every travel agency has a Xerox machine and can resell copies of whatever maps they bought whenever they want. Once those maps are distributed either legally or illegally, 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.

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):

$$(15)\ k_i = (1-\alpha)^{(1/\rho)}(d_i s_i)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$$

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 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 buys the maximum amount of data, D, first at a price p_1, then resells the maximum amount of data, D, to firm 2 at a price p_2, and then firms 3 to n buy D at a zero price. Similar to equation (11) we can solve to get the maximum possible p_2:

$$(16)\ [\alpha(1-\alpha)s_2^{\rho}*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)*s_2^{\rho}]^{(1/\rho)} - r(1-\alpha)^{(-1/\rho)}*s_2*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)} \leq p_2$$

Firm 1 knows that it will earn both its standard revenue from production and also the p_2*D that it gets from reselling pirated data to firm 2. Accordingly, we can solve for p_1 similarly to (11):

$$(17)\ p_1 \leq p_2*s_2[\alpha(1-\alpha)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)]^{(1/\rho)} - s_2*r(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)} +$$
$$s_1[\alpha(1-\alpha)[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)]^{(1/\rho)} - s_1*r(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}$$

$$(18)\ \text{Data seller revenue} = (p_1 + p_2)*D =$$
$$\boxed{(s_1 + s_2)*D*[\alpha(1-\alpha)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{-1} + (1-\alpha)]^{(1/\rho)} - r(1-\alpha)^{(-1/\rho)}[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}}$$

The critical difference between equation (18) 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.

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

$$(19)\ k_i = d_i s_i K/(d_1 s_1 + d_2 s_2 + \ldots + d_n s_n)$$

Just like the earlier piracy scenario, at most two firms buy legal 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, D, first at a price p_1, then resells the $d_2<D$, to firm 2 at a price p_2, and then firms 3 to n buy d_2 at a zero price. Similar to equation (11) we can solve to get the maximum possible p_2:

$$(20)\ [\alpha(d_2 s_2 K/(Ds_1+d_2 s_2+\ldots+d_n s_n))^{\rho} + (d_2 s_2)^{\rho}]^{(1/\rho)} - r(d_2 s_2 K/(Ds_1+d_2 s_2+\ldots+d_n s_n) \leq p_2$$

We can then compare equations (16) and (20) 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, p_2, in both equations. But in equation (20), firm 2 also pays an indirect 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 indirect cost is sufficiently high that firm 2 maximizes its profits by taking the corner solution of buying exactly enough data so that capital demand equals capital supply.[3] We can solve to get that quantity of capital and data:

$$(21)\ k_1 = (1-\alpha)^{(1/\rho)}(Ds_1)*[(r/\alpha)^{\rho/(1-\rho)} - \alpha)]^{(-1/\rho)}\ \&\ d_1 = D;$$

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

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

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

$$p_2 = (\alpha[s_2 K/(D+d_2 s_2+\ldots+d_n s_n)]^{\rho} + (1-\alpha)*s_2^{\rho})^{(1/\rho)} - r[s_2 K/(D+d_2 s_2+\ldots+d_n s_n)]$$

(22) Revenue of data seller =

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

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

The data seller in equation (22) earns strictly lower profits than the data seller in equation (18) because firm 2 only buys part of the data rather than all of the data. In other words, even a data seller who is

[3] 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.

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 its data sales revenue low even when there is no shortage of capital.

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

This subsection considers a world in 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:

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

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

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

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

rK if capital is in short supply =

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

Equation (24) can be solved numerically fairly readily. Figures 1 and 2 show the capital revenue associated with free data for two selected parameter regions. For now, the theoretical framework discusses some general patterns to the solution. Just like a social 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. But unlike the social planner, the capital owner also derives more value from data when capital and data are strong complements and less value from data when capital and data are weak complements.

Numerical Examples of Free Data Funded by the Capital Owner

This section shows the value of data by funding method for two selected parameter regions. All the graphs in this section are calculated using the parameters $\alpha = 0.5$, r = 0.5, K=50, and firms with moderately heterogenous skills, $s_1 = 1$, $s_2 = 2^{-0.4}$, $s_3 = 3^{-0.4}$, …, $s_{100} = 100^{-0.4}$. Figure 1 shows the values when data and capital are strong complements, $\rho = -5$, and figure 2 shows the value when data and capital are weak complements, $\rho = 0.5$. It is straightforward to calculate similar graphs for any particular set of parameter regions.

Figure 1: Value of Strongly Complementary Data, by Quantity and Funding Method

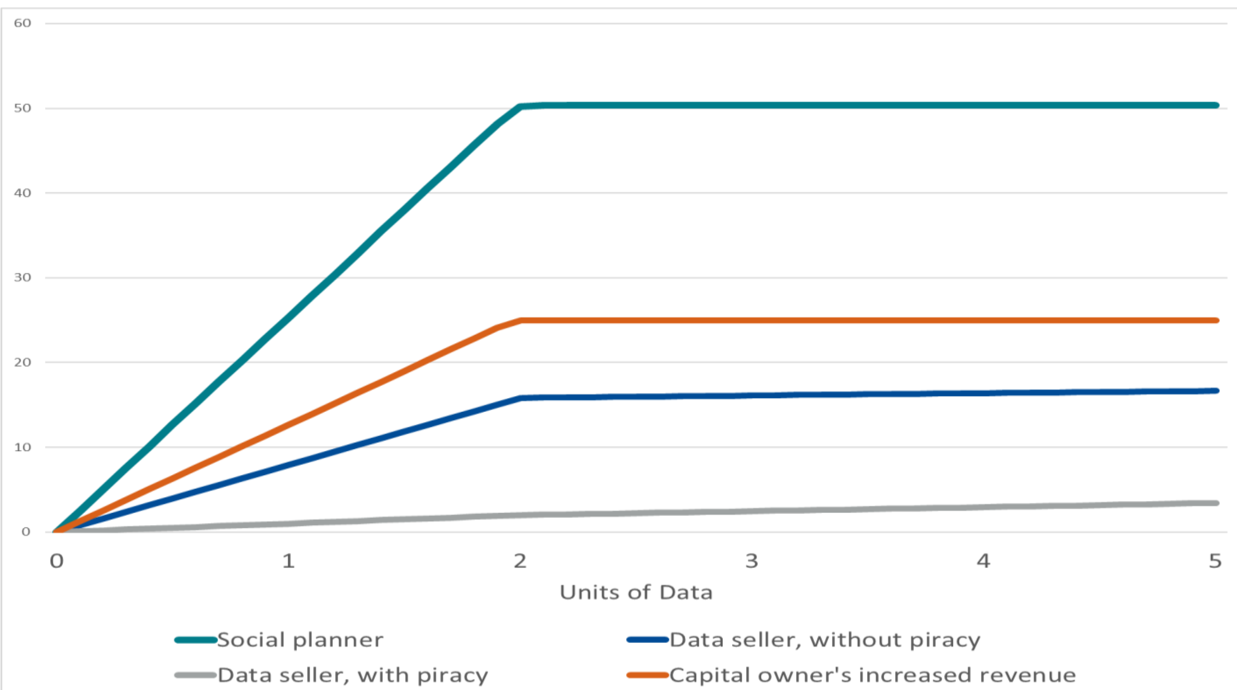
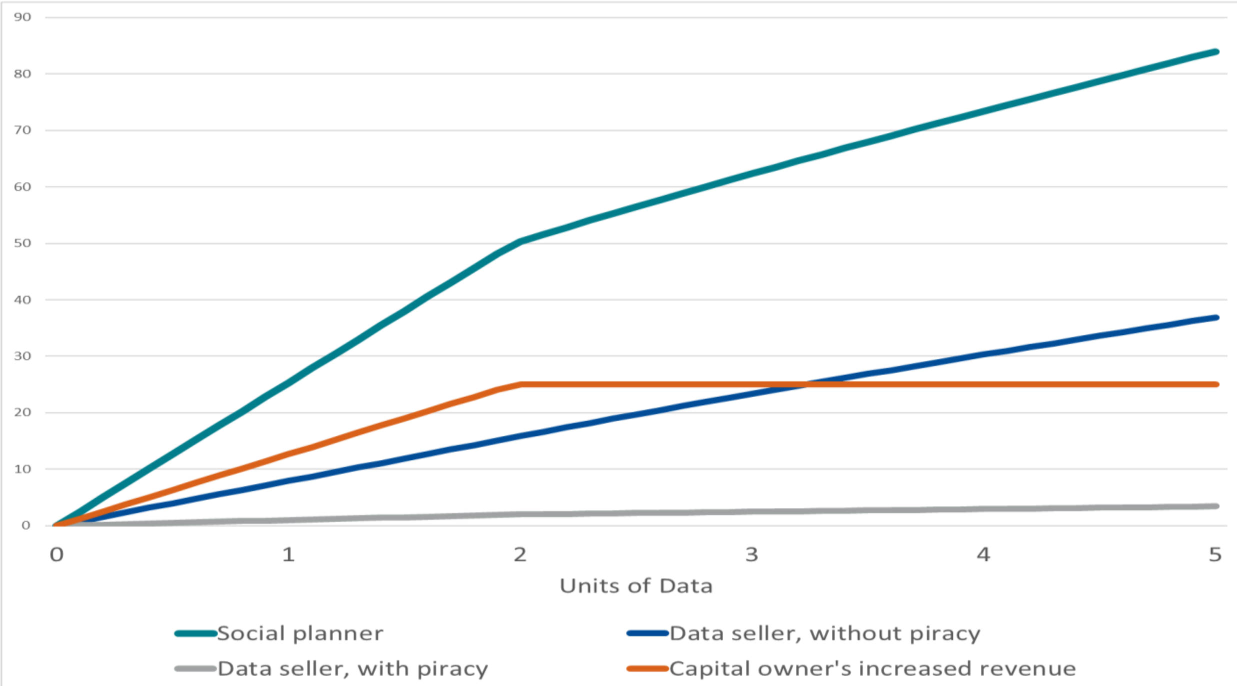


Figure 2 Value of Weakly Complementary Data, by Quantity and Funding Method



Even without piracy, figures 1 and 2 show that the value of free data to the owner of a complementary asset is often higher than the maximum possible data sales revenue. This paper is primarily a theoretical paper and so does not calibrate its parameters to any real world example. Given a range of possible parameter values, it is straightforward to calculate the share of parameter regions where free data dominates data sold without piracy. However, national accountants typically weight parameter regions by the number of data points contained in that parameter region. Information on the number of data points in each parameter region could not be located, and so it is not possible to calculate the weighted share of data that would be distributed free even in a world without piracy. Nevertheless, it is clear that free data often dominates sold data and therefore free data would account for a large share of the total data stock even in a world without piracy.

In practice, figures 1 and 2 both show that data seller revenue with piracy is always very small. The specific types of data which are best protected by intellectual property law are already included in GDP as intangible capital assets (BEA 2022). This paper's back-of-the-envelope calculation excludes those already included data types from consideration. Hence, it is very plausible that all of any data studied in either this paper's back-of-the-envelope calculations or the proposed national accounting guidelines (Rassier et al. 2019) (Eurostat 2020) are very vulnerable to piracy. Accordingly, their owners face a binary choice of keeping them secret or distributing them for free.

Section 3: Extended Theoretical Framework with Multiple Capital Assets and Multiple Data Types

This section extends the theoretical simple economy to study data creation and usage in a more complex world. 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 rent data at a fixed prices per unit of data, p^1 to p^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 nested constant elasticity of substitution (CES) function that first combines the w data types into a single data index and then combines one capital asset with that data index 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, 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.

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

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

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

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

$(r^1k^1_i + r^2k^2_i +\ldots+ r^vk^v_i+ p^1d^1_i + p^2d^2_i +\ldots+ p^wd^w_i)$

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 a Social Planner Controls All Outcomes

This subsection considers a world in which data allocations are made by a social planner whose goal is to maximize total profits. Given the distribution of data, each firm then decides how much of each capital asset to rent. Just like in the simple economy, the social planner's data allocation decision is trivial. Data sharing is assumed to have zero marginal cost. Therefore, the social 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:

$$(27)\ \text{Profit of firm } i = \{\alpha(k^1_i)^\rho+(1-\alpha)*[(\beta^{1,1}D^{1\sigma}+..+ \beta^{1,w}D^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)}$$
$$+ \{\alpha(k^2_i)^\rho+(1-\alpha)*[(\beta^{1,2}D^{1\sigma}+..+ \beta^{w,2}D^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)} + \ldots +$$
$$\{\alpha(k^v_i)^\rho+(1-\alpha)*[(\beta^{1,v}D^{1\sigma}+..+ \beta^{w,v}D^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)} - \{r^1*k^1_i + r^2*k^2_i + \ldots + r^v*k^v_i\} - \text{fixed fee for data}_i$$

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:

$$(28)\ dProfits/dk^1 = \{\alpha(k^1_i)^\rho+(1-\alpha)*[(\beta^1{}_1D^{1\sigma}+..+ \beta^1{}_wD^{w\sigma})^{1/\sigma}*s_i]^\rho\}^{(1/\rho)-1}*\alpha(k^1_i)^{\rho-1} - r^1 = 0 \rightarrow$$

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

(29) Total value added for all businesses =

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

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

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

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

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

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

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

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

(31) Total value added for all businesses =

$$\sum\{(\alpha[K^1*s_i/(s_1+s_2+\ldots s_n)]^\rho+(1-\alpha)*(\beta^1{}_1D^{1\sigma}..+\beta^w{}_1D^{w\sigma})^{1/\sigma}s_i)^\rho\}^{(1/\rho)}+$$

$$\sum\{(\alpha[K^2*s_i/(s_1+s_2+\ldots s_n)]^\rho+(1-\alpha)*(\beta^1{}_2D^{1\sigma}..+\beta^w{}_2D^{w\sigma})^{1/\sigma}s_i)^\rho\}^{(1/\rho)}+\ldots+$$

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

$$(s_1+s_2+\ldots s_n)*\{(\alpha[K^1/(s_1+s_2+\ldots s_n)]^\rho+(1-\alpha)*[(\beta^1{}_1D^{1\sigma}..+\beta^w{}_1D^{w\sigma})^{1/\sigma}]^\rho\}^{(1/\rho)}+$$

$$(s_1+s_2+\ldots s_n)*\{(\alpha[K^2/(s_1+s_2+\ldots s_n)]^\rho+(1-\alpha)*[(\beta^1{}_2D^{1\sigma}..+\beta^w{}_2D^{w\sigma})^{1/\sigma}]^\rho\}^{(1/\rho)}+..+$$

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

Equations (29) and (31) look long and complicated, but all the parameters are constants and the equations can be readily solved on a computer. Figures 3 to 8 show those total profits for selected parameters. For now, this paper simply notes that the same general factors influence revenue in this extended theoretical framework as in the simple theoretical framework.

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, p^1 to p^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 $p^{\text{hiking map}}$, $p^{\text{fishing map}}$, $p^{\text{climate map}}$, and so on. Third, each travel agency decides how many maps to buy of each type. 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 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 prices p^1 to p^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 allocation for the first capital asset, k^1_i, is shown:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$(p^1d^1{}_i+p^2d^2{}_i+...+p^wd^w{}_i)$$

Equation (33) is can be solved using first order conditions to get an interior solution:

(34) dProfit of firm i/dd$^1{}_i$ = 0 at optimum →

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

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

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

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

Equation (34) 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 price of data type 1, p^1 is lower. This is a standard result of almost any demand function. More important for our analysis, it shows that the optimal quantity of data type 1, $d^1{}_i$, depends on the quantities of data types 2 to w. If data types are strong complements, then a higher price for data type i 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 type i may impose 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 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 social 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 owns 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:

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

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

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

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

Equation (37) 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:

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

(39) 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 owner rather than a capital owner. For example, a fishing map owner might buy the rights to fish recipes and then distribute those recipes for free. The example with multiple separate capital owners can be broken down into three steps: how much data should the data buyer distribute free and how much capital should each firm rent.

The second question is difficult to solve. 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 rights to the other data types, distribute only minimal quantities of those data types for free, and then sell their data at a very high price. However, this option is generally forbidden by antitrust law and therefore is not often done. The remainder of this subsection focuses on data types that are sufficiently complementary that the data buyer chooses to distribute all of the purchased data to every firm. Equation (34) showed that demand for data type i depends not only on p^i but also the other p's. As a result, it is difficult to solve mathematically for the value of free data to a particular data owner. However, the value of free data can be solved numerically.

Numerical Examples of Free Data Funded by Capital or Other Data Owners

The six graphs shown in this subsection are all special cases of figures 1 and 2. Just like those figures, data and capital have similar weights, $\alpha = 0.5$, and the firms are moderately heterogenous, $s_1 = 1$, $s_2 = 2^{-0.4}$, $s_3 = 3^{-0.4}$, …, $s_{100} = 100^{-0.4}$. All six 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 data and capital and the complementary between the different data types. Figures 3 and 4 show a parameter region in which the data types are perfect complements, $\sigma = -\infty$. Because these data types are all perfect complements, they are automatically specific to every single capital asset. Figures 5 and 6 show 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 7 and 8 show a parameter region in which the data types are perfect substitutes, $\sigma \sim 1$, and data types are somewhat specific, $\beta^{1,1} = \beta^{2,2} =..= \beta^{100,100} = 0.67$; & $\ldots\beta^{1,2} = \beta^{2,1} = \beta^{2,100} =..= \beta^{100}{}_{99} = 0.00336$. Like figure 1, figures 3, 5, and 7 show results when data and capital are strong complements, $\rho = -5$, and like figure 2, figures 4, 6, and 8 show results when data and capital are weak complements, $\rho = 0.5$.

By construction, all of the six 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 3 to 8 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 3: Value of Data that is Perfectly Complementary to Other Data and Strongly Complementary to Capital, by Quantity and Funding Method

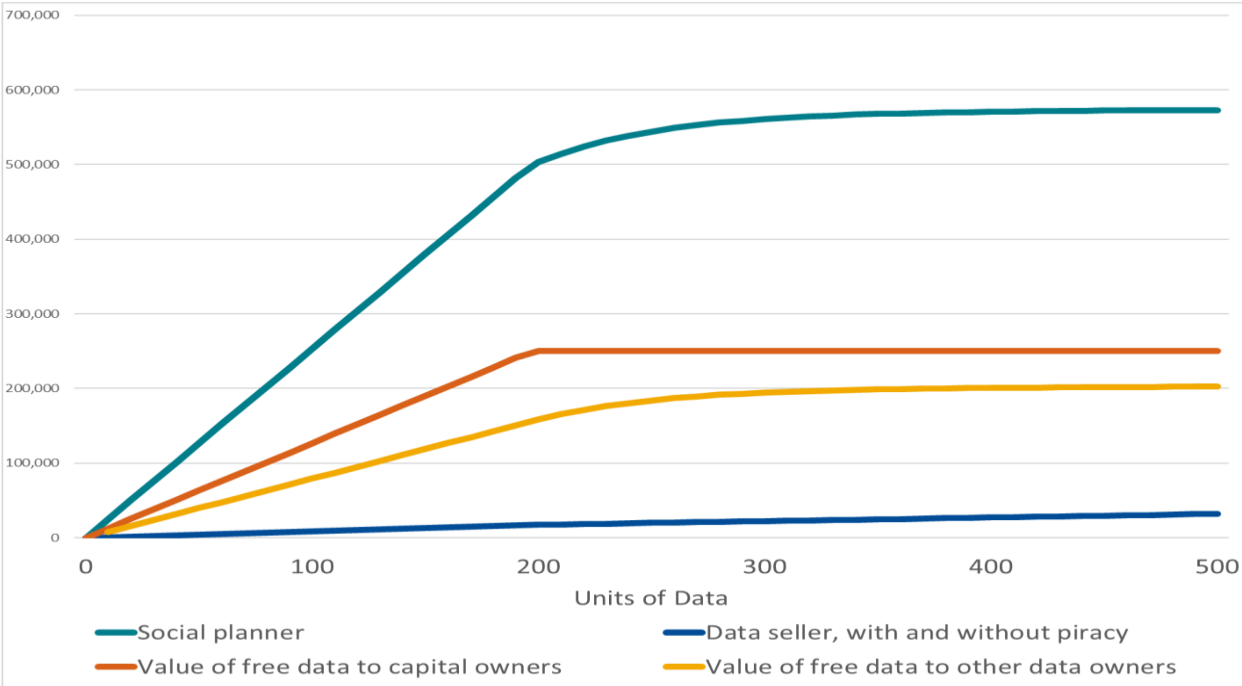
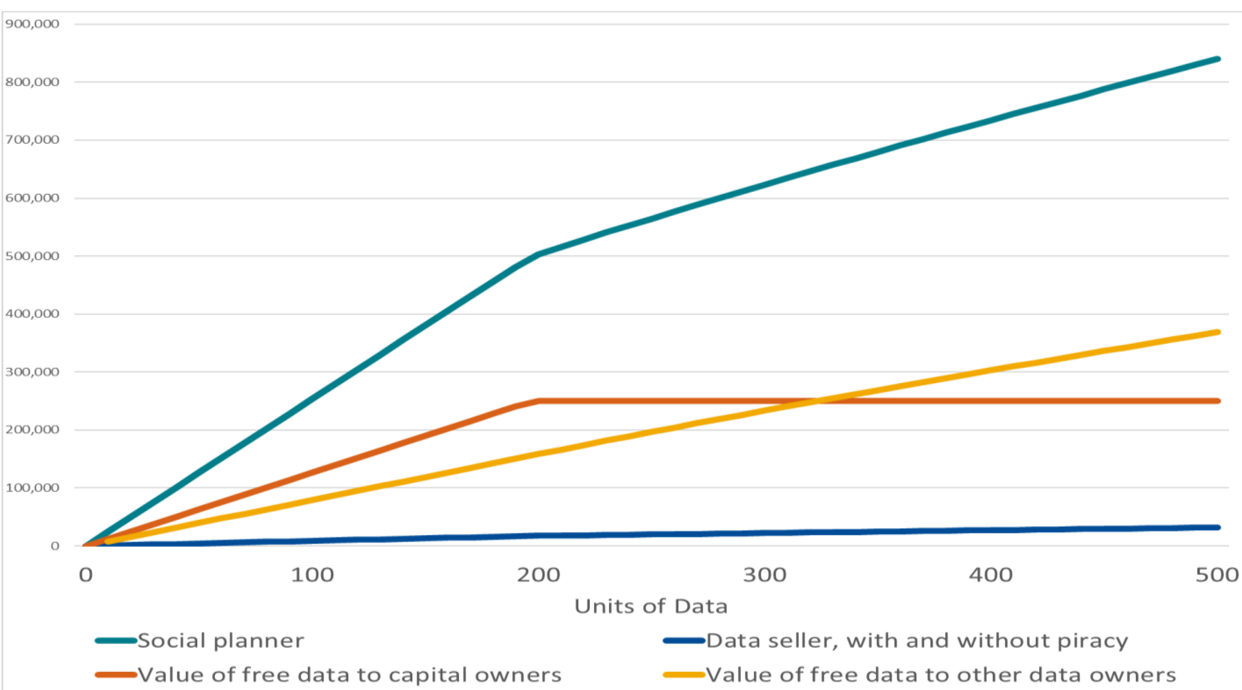


Figure 4: Value of Data that is Perfectly Complementary to Other Data and Weakly Complementary to Capital, by Quantity and Funding Method



Figures 3 and 4 show that data seller revenue is very low 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 maps seller has the power to raise prices very high. For many plausible parameter regions, the only Nash equilibrium between the w data sellers is one where all the data sellers charge such a high price that the only the highest ability firm, firm 1, buys data. At that Nash equilibrium, no firm can raise its prices without reducing demand to zero. As a result, total data sales revenue is sometimes much lower than it is when all data types are sold by a single monopoly data seller.

In contrast, figures 3 and 4 show 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 its capital or data falls to the minimum possible. In other words, collusion is stable because free data owner has a strong incentive to supply whatever data they hold to the market. In the tropical island example, total demand would fall to zero if even one free data owner did not distribute their safety map. This is a harsh enough consequence for withholding data to ensure that every free data owner will choose to distribute data and therefore bartering maps will never result in additional revenue from data. If there are at least as many data types as capital asset types, $w \geq v$, then each capital owner can be assigned specific data types without duplication. 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. 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 5: Value of Data that is Nonspecific, Perfectly Substitutable to Other Data and Strongly Complementary to Capital, by Quantity and Funding Method

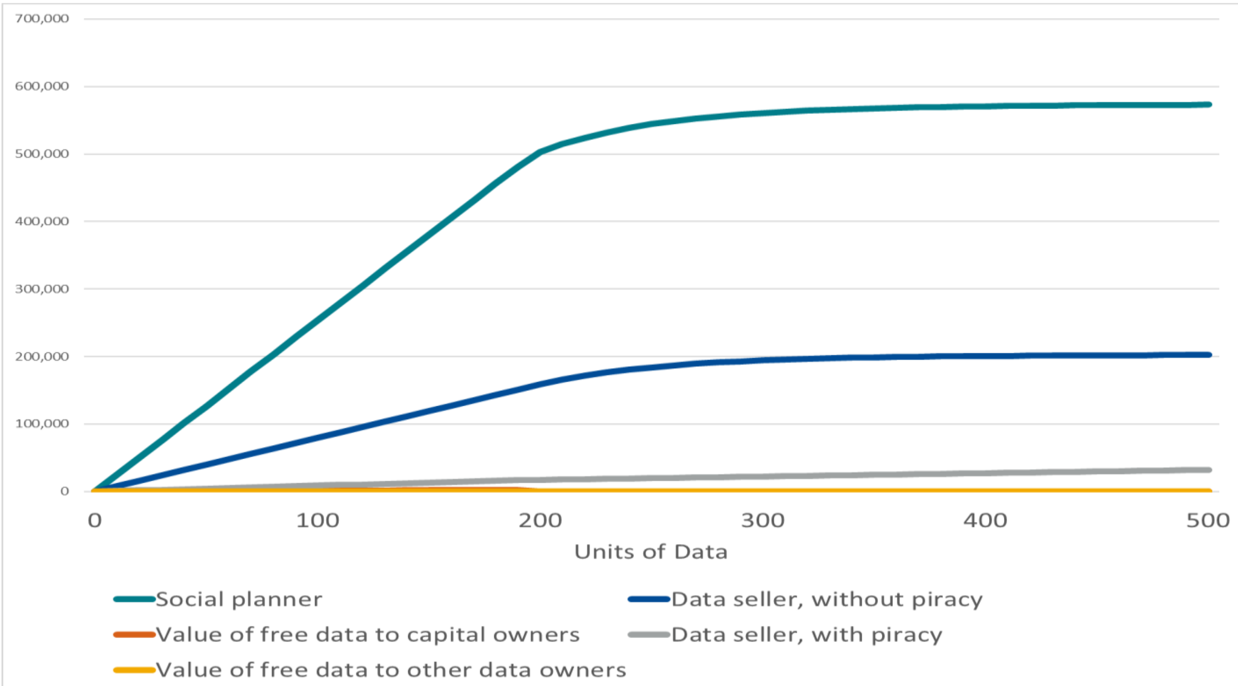
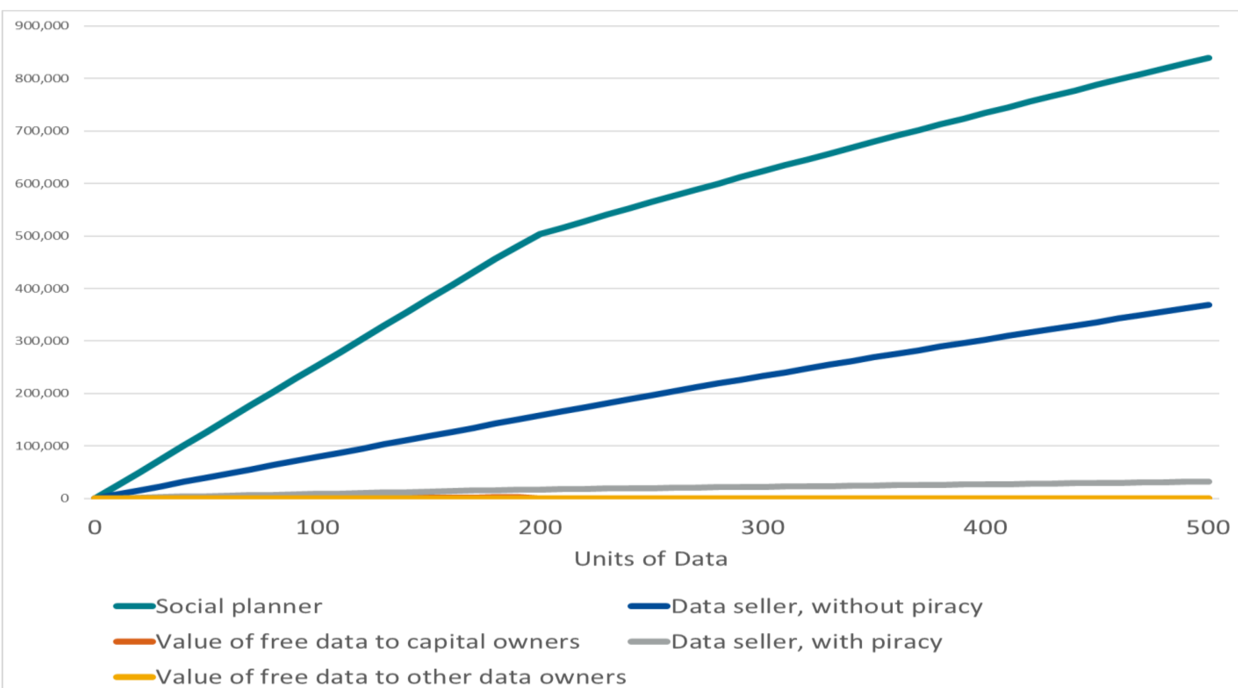


Figure 6: Value of Data that is Nonspecific, Perfectly Substitutable to Other Data and Weakly Complementary to Capital, by Quantity and Funding Method



Figures 5 and 6 show that data seller revenue is high when data are perfectly substitutable with other data. The basic issue is simple. 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. As a result, total data sales revenue is generally equal to the data sales revenue that would be earned if all data types were sold by a single monopoly seller.

In contrast, figures 5 and 6 show that the value of free data to either capital owners or other data owners is very low when data are perfectly substitutable and nonspecific. The basic issue is simple. If one capital owner withheld their free data, then total data services would only fall slightly. When capital is not in short supply, this slight fall in data services only slightly lowers revenue for capital owners. And when capital is short supply, this slight fall in data services has no impact on revenue for the capital owner. Similarly, if one data owner withheld their free data, then demand for sold data would not change at all. 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 weather reports only suffers slightly. Similarly, there is little complementarity between weather reports for Tuesday and weather reports for Monday – and so a seller of Tuesday weather reports . As a result, the total value of free data to individual capital owners is much lower than it would be if all free data were distributed by a consortium of capital owners. And the value of free data to other data owners is zero whether the data are distributed by a consortium or not.

The differences between the results shown in figures 3 and 4 and figures 5 and 6 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 different twist on it. If data sellers could collude, then 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 figures 3 and 4 but small penalties for deviation in figures 5 and 6. 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. The combination of high social value and little private value may mean that these data types need government funding (Reiss 2021) to be created.

Figure 7: Value of Data that is Moderately Specific, Perfectly Substitutable to Other Data and Strongly Complementary to Capital, by Quantity and Funding Method

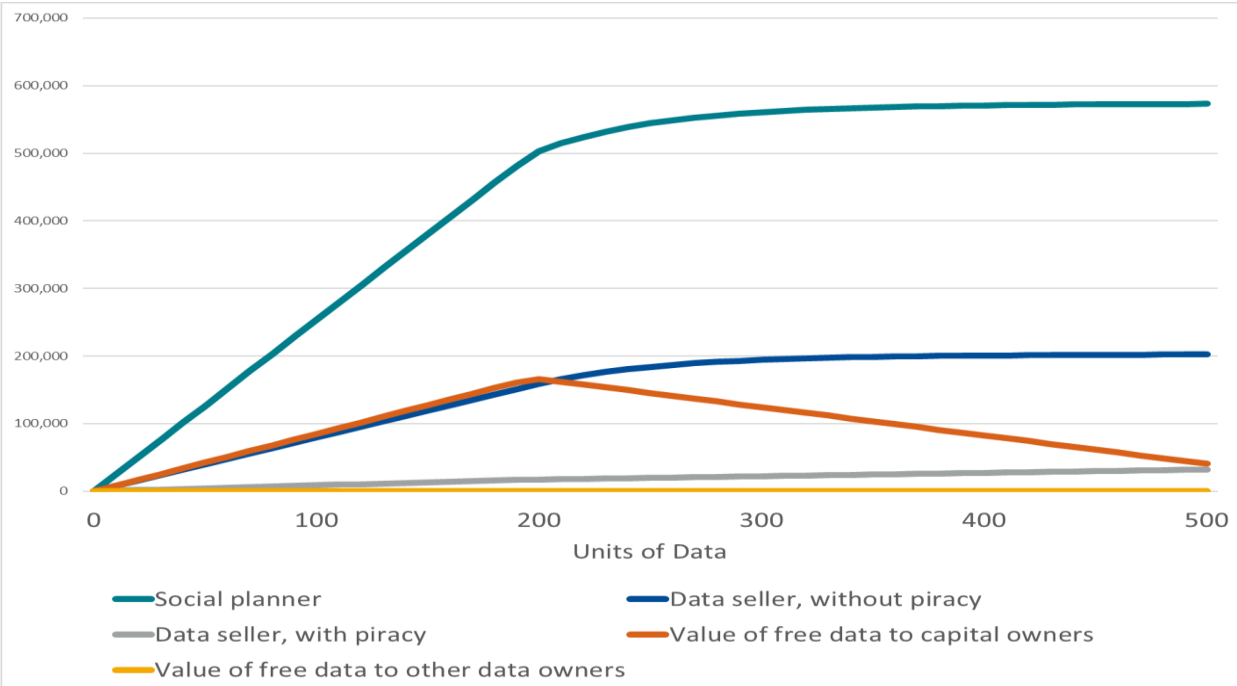
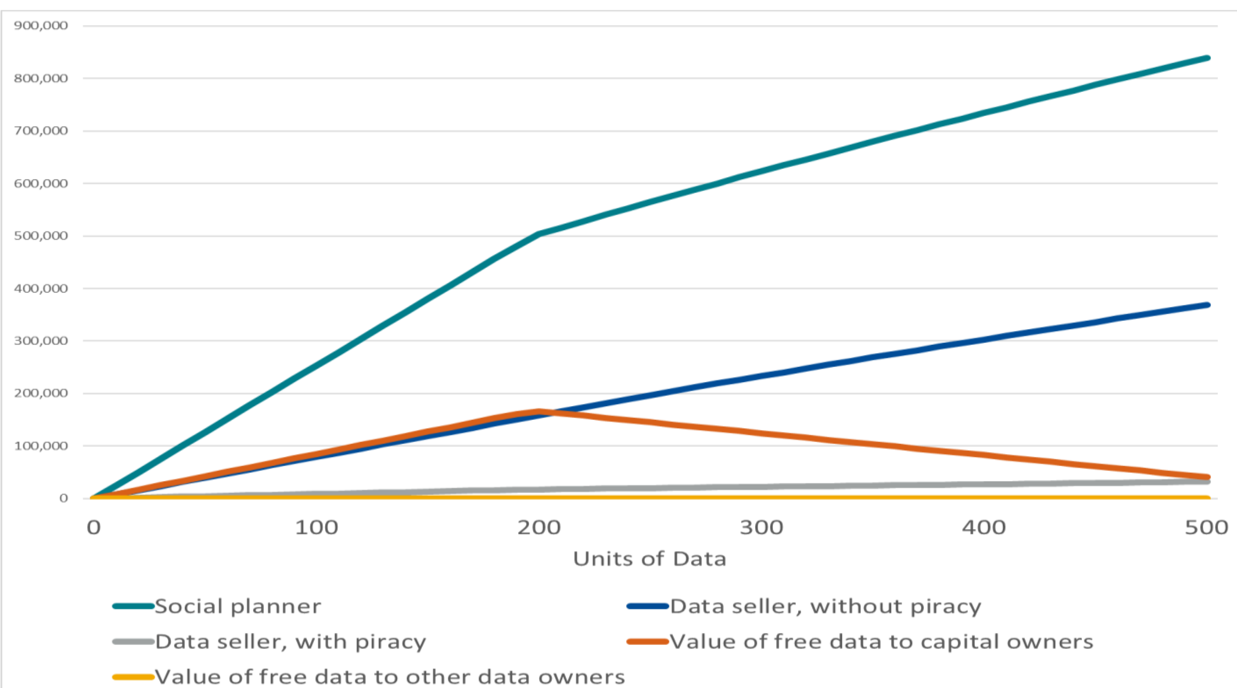


Figure 8: Value of Data that is Moderately Specific, Perfectly Substitutable to Other Data and Weakly Complementary to Capital, by Quantity and Funding Method



Figures 7 and 8 show the exact same data seller revenue as figures 5 and 6. 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 somewhat specific. Similarly, their demand for data is independent of the prices charged for other data types by other data sellers. Just like before, total data sales revenue is generally equal to 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, figures 7 and 8 show that the value of data to capital owners is moderate when data are perfectly substitutable and somewhat specific. This paper assumes that each capital owner holds the data type which is specific to their particular capital asset. If one capital owner withheld their data, then total data services would fall by 67 percent. When capital is not in short supply, this large fall in data services noticeably lowers revenue for the capital owner. But when capital is in short supply, this large fall in data services has minimal impact on 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 if their plane has many empty slots – but not incentive to distribute skydiving rules if their plane is already fully booked for the next year.

Real world examples of free data often belong in the parameter region shown in figures 7 and 8. For example, advertisers often give data like product price, locations to buy the product, uses for a product, and so on. These types of data are specific enough to have obvious benefits to a firm which is considering buying that particular product. But they also have some spill-over benefits to firms which are considering buying similar products 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 have obvious benefit for a firm that is considering hiring that particular worker. But they also have spill-over benefits to firms which are considering hiring similar workers or firms which need a small amount of information but don't 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, the externalities associated with free data are excluded from this paper's analysis.

Section 4: Valuing Total Free Data Creation: Estimates Based on Back-of-the-Envelope Calculation

The theoretical frameworks in this paper value 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. If free data are complementary to a business asset, then the value that the owner derives may be additional product purchases, higher product prices, lower input prices, higher quality employees, or other benefits. For example, a restaurant which posts a five-star Yelp review in its window might attract hungry people who happen to be walking by. If free data are complementary to human capital, then the value that the owner derives may be more job offers, higher future wages, lower consumer prices, better targeted consumer purchases, or other benefits. For example, a graduate student who posts a research paper on their website might attract more job offers. In practice, neither information on data sales revenue nor information on the values 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 such as payments for purchased data creation services or time spent on own-account data creation to value free data. Proxies like these are often used by national accountants to value products which are difficult to value directly (BEA 2022). Hence, the back-of-the-envelope values for free data can be compared with estimates of intangible capital creation 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 2023a). 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 historical claims. That case study used the academic literature and expert judgment to calculate a value of $0.4 trillion for the risk factor data created in 2017 (Soloveichik 2023b). The final case study focused on marketing. That case study used BEA's published input-output (I-O) tables, industry sources, and data from the Occupational

Employment and Wage Survey (OEWS) to estimate a total marketing investment of $0.4 trillion in 2017 (Sveikauskas et al. 2023). This paper extrapolates from those four case studies to get a plausible estimate of total free data creation.

Valuing Purchases of Free Data Creation Services

Many industries create free data on behalf of their clients. These free data are generally bundled together with the industry's 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. In addition, 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 which are shared with patients, their doctors, and any other authorized user. This paper uses expert judgment to identify both the private industries which bundle free data together with their primary product and the private industries for which free data creation are the primary product. This paper then uses BEA's published I-O tables and other sources to calculate that industries which bundle free data creation with other products had a total gross output[4] of $18.8 trillion in 2017 and industries whose primary product is free data creation had a total gross output of $1.5 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. In particular, surveys for immediate construction are included in the cost of a newly built structure (United Nations 2008, sec 10.51). Furthermore, mineral exploration is a component of mining structures and therefore included in the value of a new oil well (United Nations 2008, sec. 6.231). The back-of-the-envelope section of this paper excludes those already tracked data types and focuses on data types which are not currently tracked in the NIPAs. To avoid double counting,

[4] 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.

this paper excludes industries which sell already tracked data. This paper is focused on privately funded free data, so it also excludes data purchased by governments. After those exclusions, this paper calculates that private industries which bundle free data creation service with other products sold $12.3 trillion of gross output to businesses and consumers in 2017 and industries whose primary products are free data creation services sold $1.4 trillion of gross output to businesses and consumers in 2017.

This paper calculates that this $12.3 trillion of bundled output can be split between $2.9 trillion of free data which and $9.4 trillion of non-data output. This split is relatively approximate; precise information on the ratio of free data creation to total output for each industry was not located. For now, this paper uses the three papers mentioned in the previous subsection studying free data to calculate a plausible ratio. One paper found that consumers implicitly purchased $274 billion of free individual credit data and $27 billion of free tax data from the retail and banking[5] sectors in 2017 (Soloveichik 2023a). 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 is calculated from that paper. Another paper found that customers implicitly purchased $145 billion of free insurance claims data from the insurance sector in 2017 (Soloveichik 2023b). Based on the 2017 Economic Census, this paper calculates that motor vehicles collected gross premiums of $263 billion in 2017.[6] 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 before redefinitions. Hence, a ratio of free data to total output of 0.66 is calculated from that paper. This paper uses the weighted average of those three papers, which is 0.23, as its multiplier to calculate bundle free data. When that 0.23 ratio is applied to all industries which bundle free data with other output, this paper calculates those businesses sold a total of $2.9 trillion of bundled free data creation services to businesses and consumers. This bundled output can be added to the $1.4 trillion of data sold as a primary product to businesses and consumers.

[5] These two sectors are combined because they coordinate to process consumer payments.
[6] 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.

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[7] who likely specialize in the production of free data earned a total of $2.1trillion in 2017. This paper then excludes government employees and employees in industries that sell data creation services to calculate that private sector employees who likely specialize in the production of own-account free data creation earned $860 billion in 2017.[8] Own-account free data are difficult 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 2023a) and the paper on advertising found that private businesses produced $56 billion of own-account free advertising data in 2017 (Sveikauskas et al. 2023). In contrast, the paper on insurance risk factors found that private businesses produced minimal amounts of own-account claims data (Soloveichik 2023b). 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, $84 billion to employees who likely specialize in the production of own-account free advertising data, and

---

[7] 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 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; forest and conservation workers.

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

minimal amounts to employees who likely specialize in the produce of own-account insurance claims data. Hence, it seems plausible that the ratio of own-account free data creation to specialist employee wages is around 1.4 [(207+56)/(104+83)]. Based on that ratio, those private businesses created $1.2 trillion of own-account free data.

Measuring household data production is 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 is polite to the police officer who questions them over is not only preventing a dangerous situation in the short-term but is also avoiding an arrest that would tarnish their police record and increase their life insurance premiums long-term. But they are not likely to report politely answering questions as investment in a clean police record. Another major issue is that data often describe a single person who participates in both the household sector and the business sector simultaneously. For example, credit reports mingle information on consumer loans, housing loans, and small business loans. Hence, it is not 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 9: Free Data Creation by Data Type

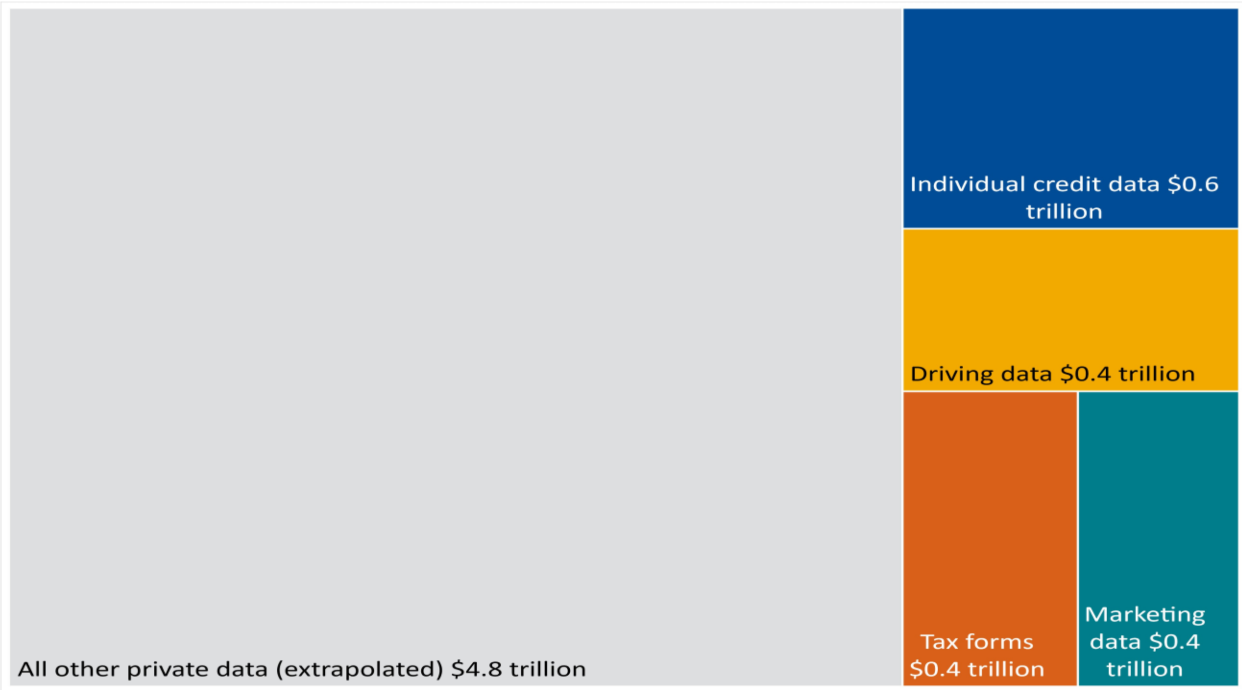
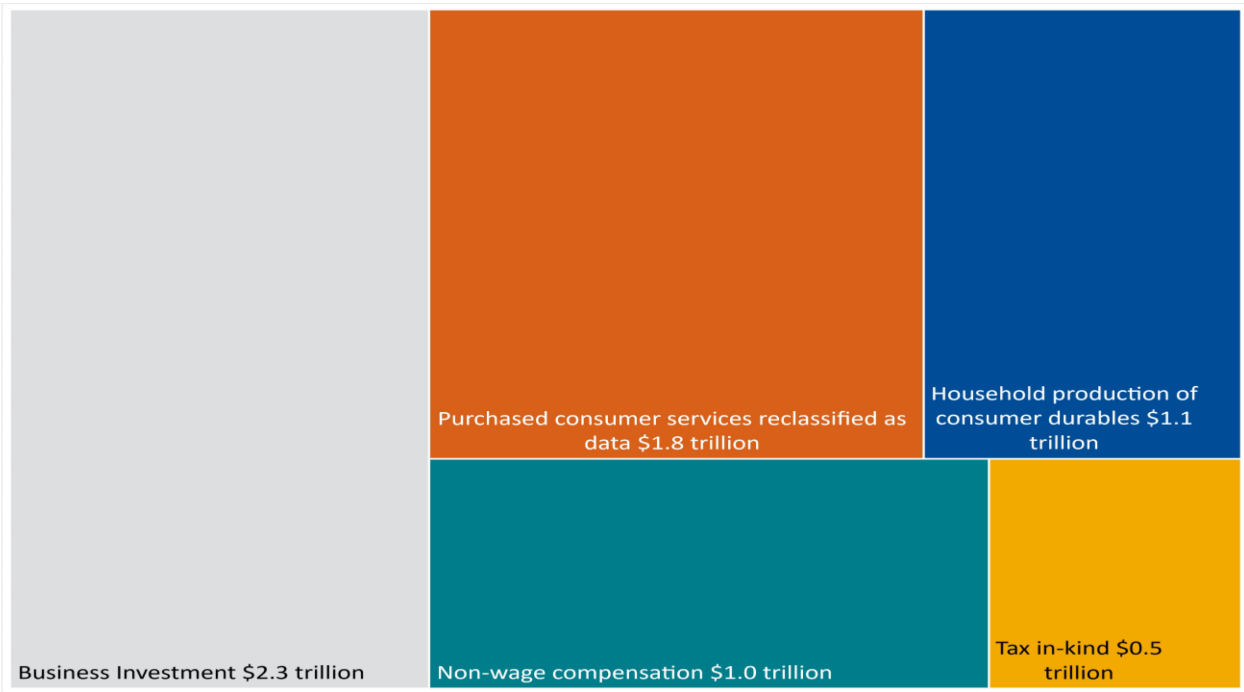


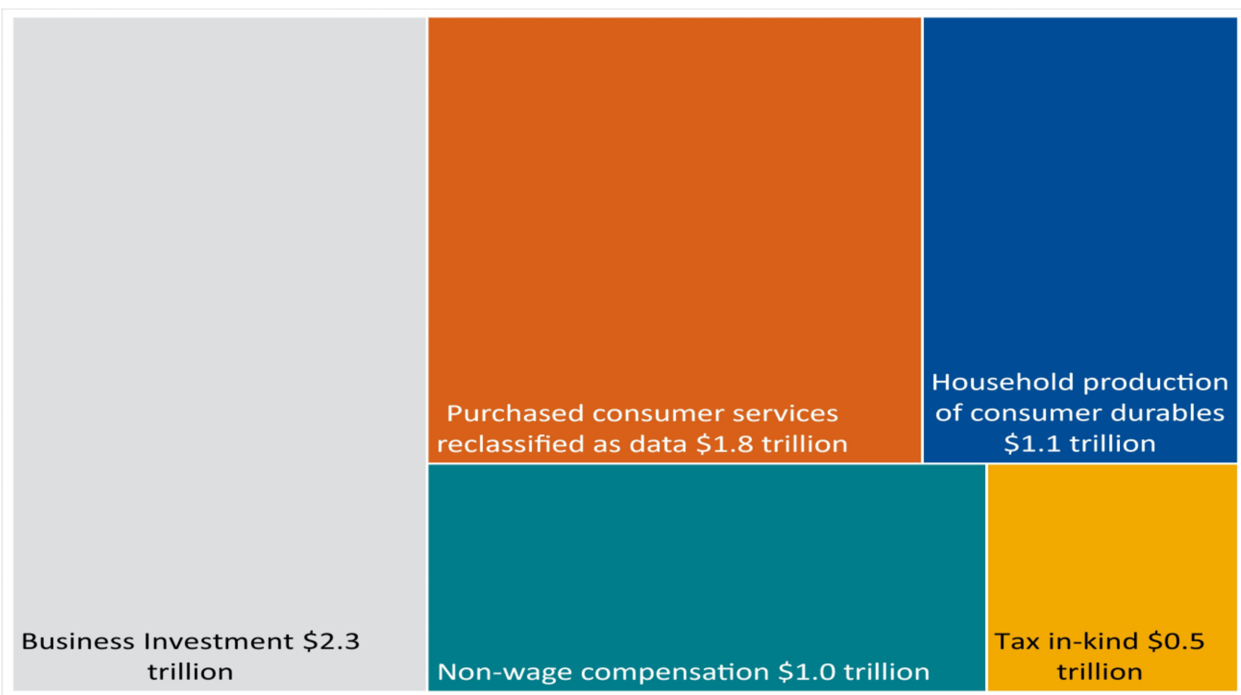
Figure 9 shows that privately funded free data creation totaled approximately $6.6 trillion of free data in 2017. These $6.6 trillion of data can be divided between the $0.6 trillion of individual credit report creation and $0.4 trillion of tax form creation studied in Soloveichik 2023a, the $0.4 trillion of driving data creation studied in Soloveichik 2023b, the $0.4 trillion of marketing data creation studied in Sveikauskas et al. 2023, and $4.6 trillion of other free data. Even if national accountants consider the $4.1 trillion of other free data to be too speculative to include in GDP, privately funded free data would still 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 creation is funded 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 employees of the business; 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 data purchases; c) non-wage compensation of data; d) tax in-kind of data; and e) business investment of data. Figure 10 below shows the approximate magnitude of each category.

Figure 10: Free Data Creation by Funding Mechanism





The theoretical framework developed in sections 2 and 3 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. The remainder of this section describes how each scenario impacts the national accounts.

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.[9] Whether consumer-created data are used by businesses or other consumers, the capital services associated with these data are out-of-scope for the NIPAs but in-scope for a household production account which includes consumer durable 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.[10]

Consumers also purchase free data creation services from business and then use these purchased data in household. For example, someone looking for a job might hire a job coach to help them write a resume,[11] 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 $4.1 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.

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

[9] Some consumer-created data were previously tracked as other household production (Bridgman et al. 2022). Accordingly, other household production falls slightly when consumer-created data are tracked as a durable.
[10] This calculation is based on a lifespan of seven years and a real rate of return of 7 percent.
[11] 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.

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. 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 $4.1 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 $1.4 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.

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 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 $4.1 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 $0.5 trillion in 2017.

Finally, the remaining $2.3 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 $2.3 trillion in 2017.

Figure 11: GDP Impact of Tracking Free Data

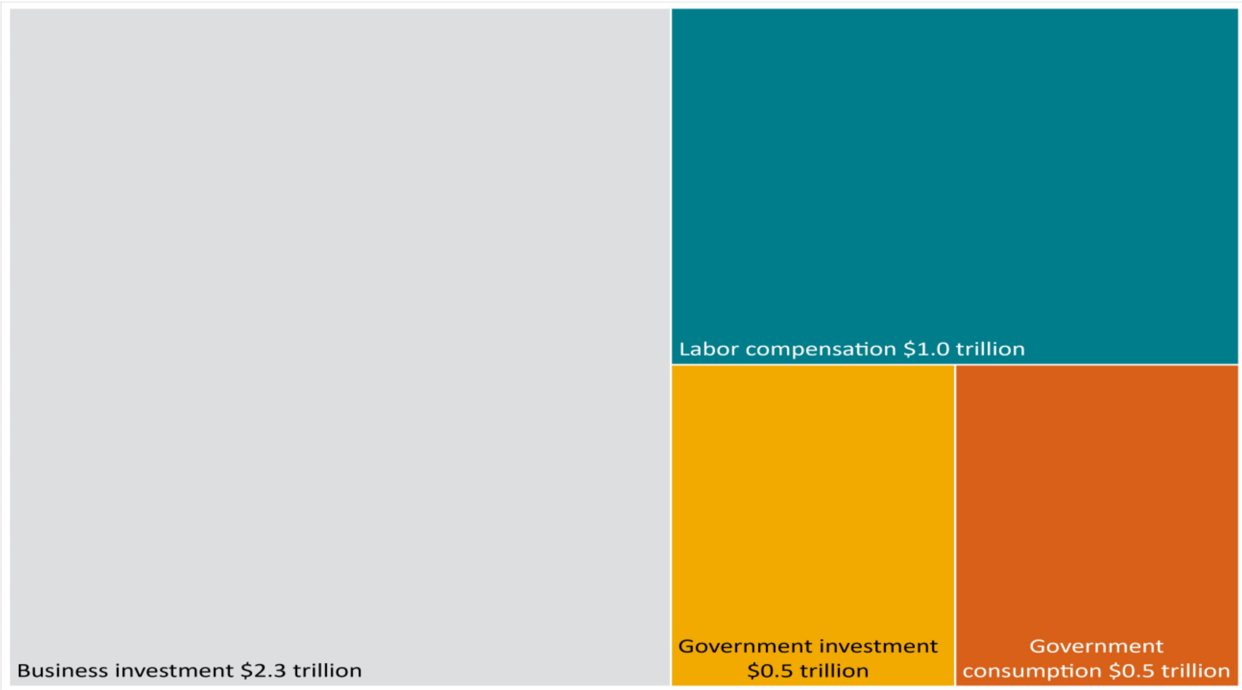
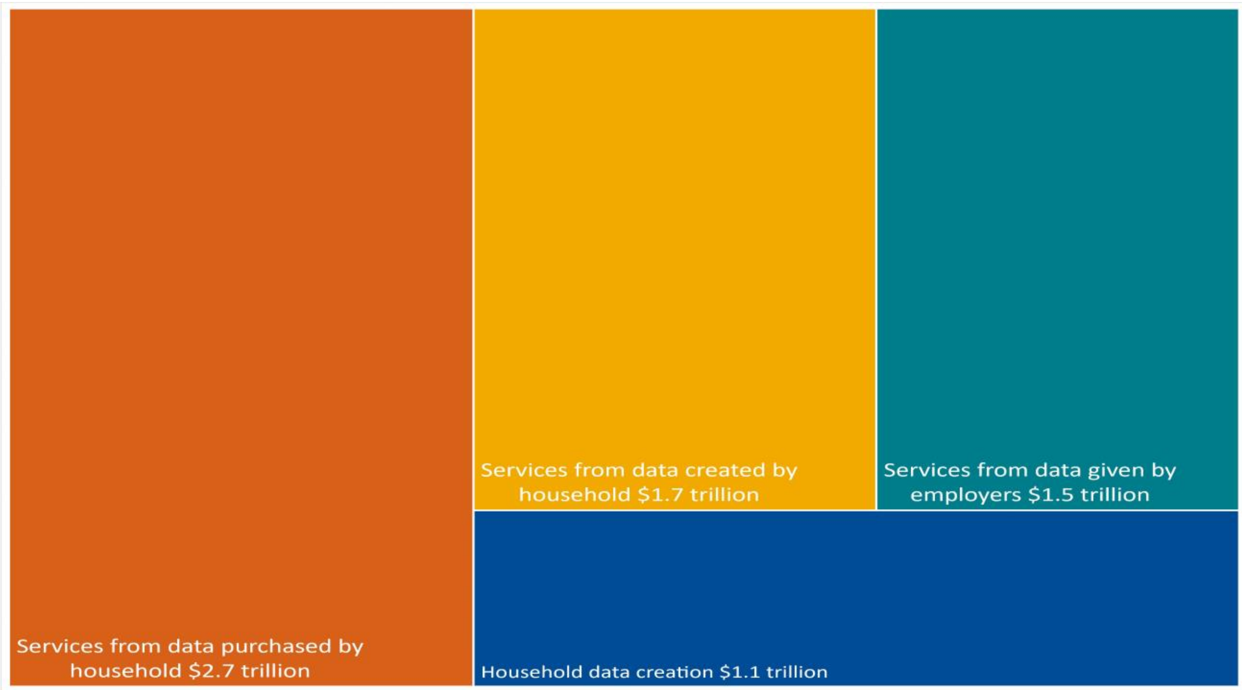


Figure 12: Household Production Impact of Tracking Free Data



To summarize figures 11 and 12, tracking free data changes both GDP and household production noticeably. Broadening GDP to include both data investment and data-related worker compensation increases measured GDP from $19.5 trillion to $23.7 trillion (19.5+1.0+0.5+0.5+2.3). 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 the market sector, the government sector, and the household sector. To illustrate the potential impacts of these revisions, this paper presents the impact on nominal GDP and real GDP quantities for the four case studies described.

Nominal GDP growth would likely increase if all data types were capitalized. BEA's published industry accounts show that industries which sell free data creation services have 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 experienced faster nominal earnings growth than other workers. However, this conclusion is very speculative and would require much more research to verify.

Figure 13: Revision to Nominal GDP Levels in Case Studies from Tracking Data

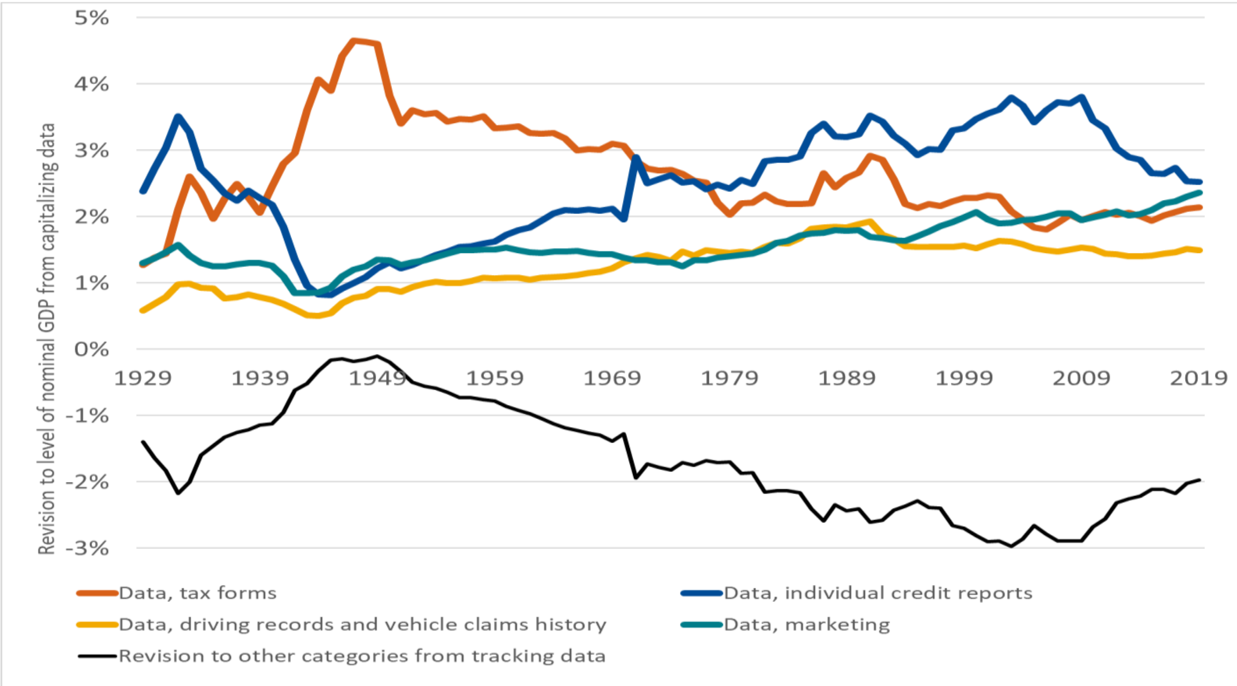
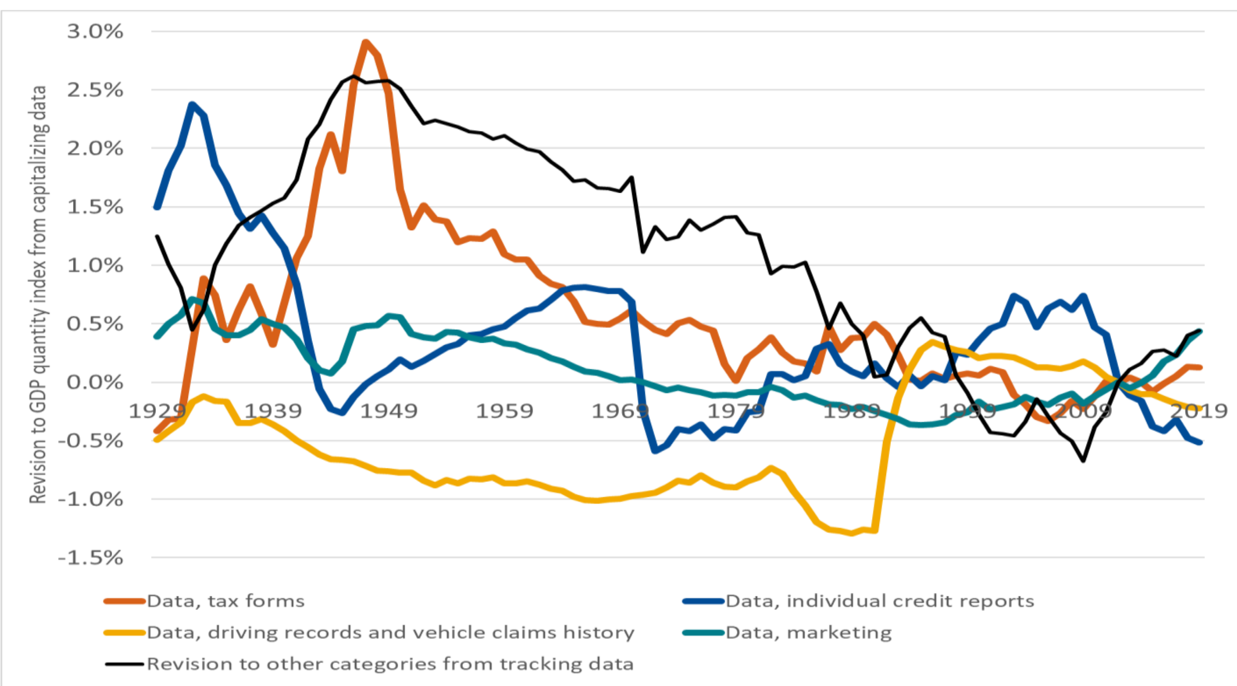


Figure 14: Revision to GDP Quantity Indexes in Case Studies from Tracking Data



Figures 14 shows that each data types 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 2023a) and data sharing infrastructure (Soloveichik 2023b) as well as input costs. A full analysis of the real GDP impact of including data might require separate case studies for each individual data type. Accordingly, this paper will not discuss how real growth might change if free data were included in GDP.

Impact of Free Data on Wealth by Sector in 2017

The stock of business assets does not rise when free data are tracked. It may be true that the measured stock of data are higher when free data are tracked together with sold data. However, the free data are implicitly included in the market value of complementary capital assets. For example, natural resource exploration raises the market price for non-produced assets like land (Soloveichik 2022). Hence, simply tracking free data associated with land shifts some portion of land value from the tangible asset stock to the intangible asset stock without changing the total asset stock. Of course, BEA's current fixed asset accounts do not include every single category of capital. As a result, tracking data that are complementary to an untracked capital category raises the measured stock of business assets.

On the other hand, the stock of consumer durables does rise when free data are tracked. Most consumer data relates to employment or consumption and are therefore not included in the market price for the purchased durable goods tracked in BEA's fixed asset table 1.1.[12] If consumer data last for seven years, then the total stock of free consumer data might be worth more than $25 trillion. This is more than quadruple the value of consumer durables reported in BEA's fixed asset table 1.1.

To be clear, some consumer free data are implicitly included in human capital. Consumers often use their free data to find better jobs or start businesses. For example, a worker might use their resume and references when they apply for a new job, or a potential business owner might use their personal credit score to lease a storefront. The higher earnings associated with these uses of free data are likely to be captured in any stock of human capital which is based on discounted future earnings (Liu and Fraumeni 2020). On the other hand, consumers also use free data in household production, consumption, or

[12]Some free data are complementary to consumer durables. For example, Carfax has vehicle accident data that are used by both automobile dealers and insurers to value used vehicles (Soloveichik 2023b).

leisure. For example, a parent might share their child's activity schedule with potential carpool partners, a patient might share their medical information with a new doctor, or a lonely person might share their dating profile with potential romantic partners. The higher utility associated with these uses of free data are unlikely to be captured in current measures of human capital stock. Information splitting consumer data by user sector could not be located. Therefore, this paper does not attempt to estimate the share of free data which is implicitly included in human capital.

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 demonstrated that privately funded free data are a very large asset type. First, the paper presented four previous case studies (Soloveichik 2023a) (Soloveichik 2023b) (Sveikauskas et al. 2023) which studied a total of $1.8 trillion of private free data creation in 2017. The paper then extrapolated total private creation of free data of $6.6 trillion. In 2017, including free data raises measured GDP by more than 20 percent and raises measured household production by more than 100 percent.

Bibliography

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2022) "Too Much Data: Prices and Inefficiencies in Data Markets," *American Economic Journal: Microeconomics*, 14(4), pages 218-256.

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

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,”

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. (2023a) “Capitalizing Data: Case Studies of Tax Forms and Individual Credit Reports,” *BEA Working Paper* 2023-7.

Soloveichik, R. (2023b) “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) “Advertising and Other Intangibles in 61 United States Industries,” forthcoming in *Review of Income and Wealth*.

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