IARIW – CIGI 2023

IARIW – CIGI 2023

Thursday, November 2 – Friday, November 3

On the Value of Data

Dan Ciuriak
(Centre for International Governance Innovation)

Paper prepared for the Conference on The Valuation of Data November 2 – November 3, 2023
Session 3: Methodologies of Data Valuation - II
Time: Thursday, November 2, 2023 [2:00 PM-3:00 PM EST]

On the Value of Data

Dan Ciuriak

Paper for IARIW-CIGI Conference on "The Valuation of Data", November 2-3, 2023, Waterloo, Ontario, Canada

Abstract: Data is often said to be the most valuable commodity of our age. It is a curiosity, therefore, that it remains largely invisible on the balance sheets of companies and largely unmeasured in our national economic accounts. This note seeks to unpack what we mean when we refer to data as the "new oil" or the essential capital of the data-driven economy, how it differs from information in general, how it is transformed into value, and what might be the approximate scale of the value of data in a modern data-driven economy. A key feature of data is that it is captured rather than acquired in a market transaction for which there are invoices and receipts – which undermines normal market frameworks for attributing a price to it. Second, while it can be bought and sold in secondary markets, it cannot be owned. A third critical feature of "big data" is that unlike data that was mobilized for analytical purposes historically, the information is, almost by definition, of a scale beyond what most firms can access, which creates new conditions of information asymmetry that in turn serve as the foundation of a business model based on exploiting of information asymmetry for commercial advantage – which by definition is a market failure. This paper comments on the problems of using cost-based or transactions-based methods to establish value for a nation's data and suggests that the value of data is to be identified in enterprise value. Traditional accounting looks through the firm to its tangible (and certain intangible) assets; that may no longer be feasible in measuring and understanding the data-driven economy.

Keywords: data valuation, national economic accounts, enterprise value, superstar firms, heterogeneous firms, big data, data-driven economy, artificial intelligence, machine learning, machine knowledge capital, information asymmetry, business process optimization, price discrimination, negative externalities, national security

JEL Codes: D82, D83, L15, L16, O31, O32, M4

Disclaimer: I read that more than 100 AI papers are published very day. That's over 30,000 so far this year, of which I may have read a few. Which means I am asymptotically approaching perfect ignorance on this subject as well as most others. As a corollary, I'm in a position to vouch that ignorance is not bliss, at least not asymptotically. With that, let's to it.

1 Introduction

Economists have famously been pilloried for knowing the price of everything and the value of nothing. In the case of data, economists know neither the price nor the value. That is a problem for a market-based economic framework that depends on the discovery of prices through market exchange to determine the value of things. In particular, while data is often said to be the most valuable commodity of our age, the essential capital of the modern data-driven economy, and the source of instruction for artificial intelligence (AI) whose rapid evolution is ushering in the age of machine knowledge capital, it remains largely invisible on the balance sheets of companies and largely unmeasured in our national economic and trade accounts.

Unlike other productive assets that served as the essential capital asset of their age (land in the agrarian age, the machinery of mass production in the industrial age, and traditional intellectual property in the knowledge-based economy), a key feature of data is that it is captured rather than acquired in a market transaction for which there are invoices and receipts. This undermines the market frameworks developed since the marginal revolution for attributing a price to an asset – no marginal cost, no marginal price, no inference as to market value.

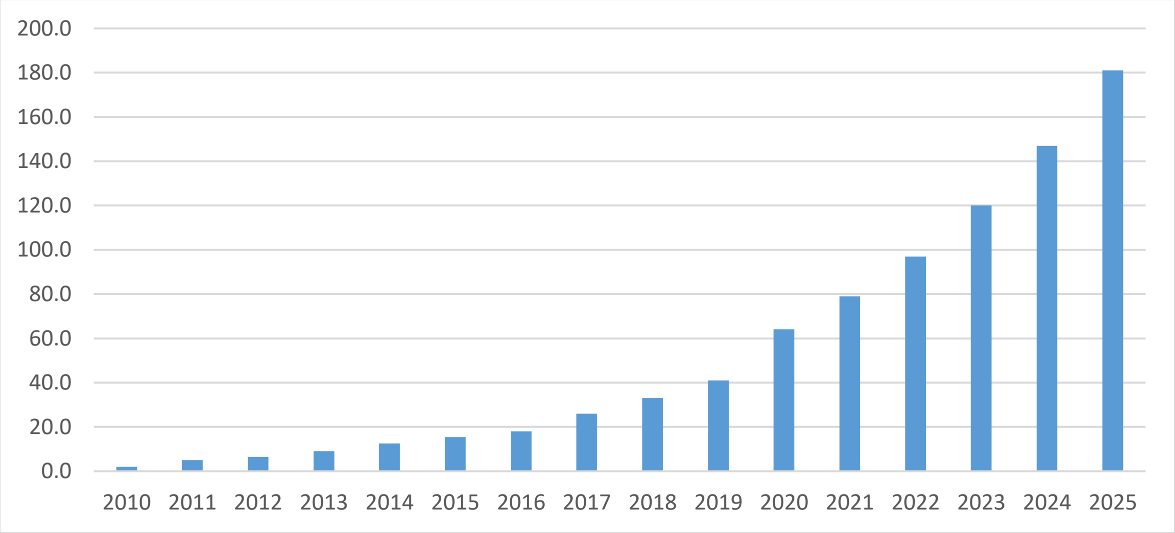
At the same time, while data can be bought and sold in secondary markets (once assembled into databases owned by companies), "ownership" of the data itself is not possible. There is no limit to the number of companies that can stake a claim in the same data and claim a monopoly on their own dataset. Moreover, while some insights into the value of data might be obtained from secondary market transactions in curated databases, the large pools of data that define the data-driven economy (i.e., those assembled by the superstar firms) are not traded. They are akin in this sense to the "dark pools" of capital in equity markets that allow private exchange without influencing market prices through transparent bids.

A third critical feature of "big data" is that unlike data that was mobilized for business and analytical purposes historically, the information delivered by big data is, almost by definition, beyond what the human mind can access - in a sense, there is an opacity threshold that is passed, which creates pervasive information asymmetry. Information asymmetry is a source of market failure. Exploitation of information asymmetry for commercial advantage is at the heart of the business model of the data-driven economy. In other words, data empowers the development of markets based explicitly on market failure as a business model. This too raises unprecedented conundrums since information asymmetry is something that regulation and markets seek to "correct" – in other words to annihilate. At the same time, it is instrumental in driving enterprise value, which of course no-one wants to annihilate.

Finally, the data-driven economy greatly amplifies the feedback effects of data-driven analysis on the social and economic structure that generated it in the first place in not necessarily good ways. The negative externalities of data are many and significant – this is data as the new plutonium rather than the new oil. This challenges a market-oriented valuation framework that traditionally ignored negative externalities. Whether the risks can continue to be treated mostly as caveats while concrete valuation continues to be on monetary value (e.g., OECD 2013), is an open question.

The estimates of the value of data to date have placed a rather small value on it – on the order of 1-3% of GDP (Sargent and Denniston 2023; Nakamura et al. 2018). This is in prima facie terms incongruous with the astronomical amounts of data captured over the course of the short life of the data-driven economy (Figure 1) and the transformative impact that data is having on our economy and society, not least as the key input into training AI. We are experiencing an earthquake and measuring a tremor. The thesis advanced in this note is that this incongruity is attributable to the conundrums listed above.

Figure 1: Data Generated During the Data-Driven Economy Era (zettabytes/year)



Source: Duarte (2023).

The rest of this paper is organized as follows. Section 2 discusses the conundrums raised by trying to fit a capital asset which does not fit the paradigm around which market economies are organized into a market accounting framework.

Section 3 argues that the value of data in commercial terms is internal to the enterprise, noting the various ways that have been identified in which data improves the functioning of firms, as a complementary productive asset to their other capital assets. These include inter alia: (a) optimization of business processes; (b) capturing consumer surplus; (c) allowing firms to exploit information asymmetry for market advantage; (d) shift of innovation into machine learning space, which in effect permits the industrialization of learning, accelerating the process of innovation, providing a speed advantage to firms that are able to harness this element; and (e) the creation of machine knowledge capital as a new factor of production.

Section 4 discusses the implications of a firm-centred approach to assessing the value of data and by extension value creation in the data-driven economy and draws tentative conclusions.

2 The Conundrums

2.1 Non-reducibility and emergence = no micro foundations to value of data

The Ur problem of data as an economic asset stems from the fact that it is not acquired in a market transaction for which there are invoices and receipts.

Markets are reducible to transactions. Data not so.

To take one analogy of non-reducibility, Georges Seurat's famous pointillist painting, La Grande Jatte, consists of approximately 220,000 dots (Goldstein 2019). Knowing that – and even knowing the distribution of the colours of the dots across the spectrum (Seurat used virtually every one of the 72 colours on Michel-Eugėne Chevruel's colour wheel, the creation of which was an inspiration for his work) – tells us nothing about what we see in the painting or even the colours that we see perceive. Notably, all colours are interpretations that we place on wavelengths of light, but some colours don't even exist in nature as a discrete part of the spectrum, they are to some extent optical "illusions". Similarly, the value of data is rather like the meaning of the collection of dots on a canvas. From the perspective of the dots, it's an emergent property, not an intrinsic one.

In short, an individual datum or observation is not traded, nor does it have economic value absent a context (i.e., a set of correlations with other datums or observations). A collection of such observations – data – if sufficiently large has enormous value but there are no micro foundations to this value – it cannot be traced back to the individual observations for which values are established by markets and aggregated to yield the value it has.

Accordingly, the valuation of data has to be pursued indirectly, which places great importance on which indirect method is chosen.

2.2 The source of value redux: Back to LTV?

The source of value is not a new problem for economics and in grappling with the value of data we are going back in time. The classical economists, including Adam Smith, David Ricardo and Karl Marx, tried to attribute the value of a product to the amount of labour that went into its creation (Marx added "surplus value" as an early recognition of the returns to capital). The Labour Theory of Value (LTV) didn't succeed – although, interestingly, we do use horsepower to measure the power of automobile engines which is somewhat analogous to LTV.

Another train of thought introduced the abstract but seemingly intuitive concept of utility as the unit of value. While in the absolute sense advanced by Jeremy Bentham and John Stuart Mill this didn't succeed either, the marginal revolution led by William Stanley Jevons, Carl Menger, and Léon Walras linked value, in a context of scarcity, to the marginal utility of a product, which led directly to marginal cost and marginal price establishing the value of things in a transactions-based market economy. The result of this evolution gives rise to what might be called the Product Theory of the Value of Labour, standing LTV – or better, the Labour Theory of the Value of Products –

on its head, as the value of labour is determined by the value of the product created rather than vice versa.

In the absence of reducibility of data to transactions, adopting a framework that values data based on the expenditures to acquire it confronts all the issues that caused LTV to fail. This is a problem because, as Sargent and Denniston (2023) report, this is the path being taken by statistical agencies as the Canadian, Dutch and US authorities have all tried to value data-related assets using a cost-based methodology.

Recalling the adage that "you can't manage what you don't measure", it follows that you can't manage well what you don't measure accurately. Applying a version of LTV to data leads potentially to the Soviet Union of the data-driven economy.

2.3 The Data Rush

Back in the day when we used to dig money out of the ground (the metallic standards era), the rapid growth of industrial economies put a premium on expanding the money supply to avoid deflation and all the ills that go with that. And so we had gold rushes. Miners would claim "stakes" and given the lack of supporting paper-based legal infrastructure, would literally split a wooden stake several ways with each person holding a splinter being a "stakeholder", denoting ownership. The key point is there was ownership of the stake, which gave traction to markets.

We're having a "data rush" given the voracious appetite of AI systems for data. However, no-one can claim ownership of the data itself. For example, even when data is generated by market transactions, numerous parties have access to the information content of the transaction, from the purchaser of a product, the vendor of the product, the credit card company that processes the payment, the banks of the purchaser and the vendor, the telecommunications provider and probably any number of apps and government agencies that monitor traffic on the Internet (see e.g., OECD 2013; 12). In the digital age of surveillance capitalism and national security surveillance, there is no expectation of privacy for any communication over a telecommunications network or anything done in public. Moreover, where in the pre-digital age, most information had the half-life of a firefly, the datafication of information means that now it is indefinitely-lived and subject to any amount of analysis and re-analysis by any number of parties, including data brokers who monetize by selling it onward, both in real time as the observations are captured (e.g., by Google with its real-time advertising push) and with an indefinite lag as the data are incorporated into curated databases by others.[1] There can be no presumption of either ownership or control of our datums. By extension, there is no ownership of data.

What is interesting in this context is that any number of stakeholders can claim a stake in more or less the same data. Kevin Kelly (2017) tells an insightful (and now oft-recounted) story of a conversation he had with Larry Page, one of the co-founders of Google, at a Silicon Valley party

[1] On the distinction between "observations" and "data", see Sargent and Denniston (2023; 2). In the age of mobile, it is both possible and lucrative to monetize observations. For example, if I am in Barcelona, searching restaurants at the dinner hour, and my search history shows a proclivity for sushi, Google is able to put that information in front of me in real time, including if I'm on the move in an Uber, using geolocation to identify those restaurants close by.

in 2002. Kelly asked Page how Google was planning to make money off a free search engine. Page answered that they were building an AI. The thing is that Microsoft's OpenAI is almost certainly drawing on much of the same data as Google to train its own AI. Google's Bard read most everything on the internet in creating a model of what language looks like (Pelley 2023). So did ChatGPT. The scholastics used to discuss things like how many angels can fit on the head of pin; we could have the same discussion today about how many stakeholders can hold a stake in the same data.

Until data, forms of capital had characteristics that allowed unique ownership rights to be staked out. Even intangible assets such as traditional intellectual property (IP) and trade secrets allowed legal ownership regimes to be applied, albeit with substantially greater work for legal systems in determining infringement. With data, that's no longer the case. That is unusual and implies potentially different behaviour of the data-driven economy, including raising many issues not directly related to the main issue under discussion here.[2]

2.4 Data, Combinatorial Expansion, and the Matthew Principle

The major breakthroughs that have been made recently in improving AI models, especially large language models (LLMs), came from scaling the power of AI systems: the size of specialized AI computer chips broke through the trillion transistor level, the size of LLMs soared past the trillion parameter level, the power of training methods increased by orders of magnitude, and power consumption of AI chips was improved by orders of magnitude pushing back the limits on scaling.

For deep learning models, recent experience with the improvements of LLMs testifies that the larger the data set, the better the trained AI is in understanding context, interpreting out-of-sample data, capturing outliers, and handling nuances and variations in language, etc. This is not a conventional economies of scale issue. Indeed, it's not an *economy* of scale, it's a *power* of scale. Adding more data points increases the value of the entire dataset beyond the incremental value of each new data point, because it expands the number of potential connections in combinatorial fashion – each new data points creates new possible relationships with all existing data points.

There is a direct connection between this property of data and the superstar firm phenomenon that is characteristic of the data-driven economy. One might think of this as a data Matthew Principle that drives runaway market concentration: the more data the market leader has, the better the AI models that can be trained and the quality of inferences extracted, the stronger the network effects

[2] Ownership of data appears to bear mainly on the question of who owns the library costs of capture, classification and curation – besides the costs associated with archival hosting, in the modern digital context, library costs include the costs associated with access management and use/breach-based liability coverage. The generation of economic value from use of data can be based on proprietary or open data (in the latter case avoiding many of the ownership-related costs, although not necessarily all. The exercise of usage rights creates obligations related to "ownership", although the exact extent is not settled, as there are open questions whether owning a copy and exercising valid, limited use rights should be viewed as only part of an umbrella "ownership" of representations of a common fact (all the data pertaining to the same fact), or whether each representation attaches the whole universe of usage rights, and therefore constitutes independent ownership. Clearly, since a data-driven company can be bought and sold, de facto ownership of data as a capital asset (even if not in its raw form) is established and can be transferred.

in inducing users to come on board, the greater the data edge, and so on – a self-reinforcing loop that continuously widens the gap between entities with extensive data and those without. This includes but is not limited to the two-sided markets with zero prices on one side which are the poster child of this type of economy. Zero-price markets were a curiosity in the pre-digital age; now they are mainstream.

2.5 Summary

The technological revolutions that have culminated in the modern data-driven economy emerged in a market economy in which the productive capital assets had ownership rights with prices established in market transactions, allowing straightforward aggregation of the value of a nation's capital assets in its economic accounts. By extension, gross domestic product could serve as a intermediate target for economic management. While there were sources of potential market failure, they were far from predominant and the mature industrial era economy of the late 20th Century featured competitive market conditions that were ideal for the emergence of a rules-based economic governance system in which governments had little incentive to intervene. The data-driven economy does not fit this mould. The characteristics of data are at the root of the problem and trying to shoehorn data into the economic accounts and governance systems developed for an earlier age is not likely to be feasible or effective. Accordingly, we need to return to first principles.

3 Enterprise Value

There are good reasons to believe that there are large economic rents generated by data, which implies that the cost of collecting, cleaning, classifying and curating data falls well short of its value. Meanwhile, although there is some data that is traded in the market in the form of databases or subscriptions thereto, the large proprietary databases of the superstar firms, the "dark pools" which are the signature feature of the data-driven economy, are not traded. This leaves only enterprise value as a viable candidate to estimate the commercial value of data, which includes data rents.

3.1 The Basic Intuition

Businesses have always used data for analytics, decision-making, forecasting, etc. Data is thus central to enterprise value. For example, the business of banking is based on understanding creditworthiness; the business of insurance on understanding risk; the business of restauranting on how much of what foods to buy on what schedule. In this regard, nothing has changed with the data-driven economy – data just got bigger and more powerful.

It is also instructive that there were analogues for the market concentration issues that have flared in the data-driven economy. For example, German insurance regulators require insurers with dominant market positions to share their risk data with their competitors in order to maintain competitive market conditions. This practice recognizes the information asymmetry issue that is inherent to data as a productive asset. And it is also from a context in which the value of data was not recognized separately as a line item in the capital structure of the firm.

This intuition can be sharpened by recalling one of the best known tag lines developed by a company – BASF’s "We don't make a lot of the products you buy. We make a lot of the products you buy better" (Deutsch 2004). Thinking about data in this sense suggests it is a complementary form of capital asset that makes other capital assets that are more conventionally valued on a transactions basis more efficient. This would manifest itself in an increase in profits that implicitly are data rents. The next section briefly runs through some ways in which data performs this function to illustrate the point. This is not intended to be exhaustive.

3.2 Monetizing Data

Numerous ways have been identified in the literature on how data increases the efficiency and profitability of firms.

3.2.1 Optimization of processes

Sector by sector, company by company, big data enables firms to improve business processes, to reduce costs, and increase operating margins. For example, McKinsey Global Institute (2016) estimated that retailers could potentially increase operating margins by as much as 60% through the application of big data. The scope for potential gains varies by sector; however, there are significant improvements, which create the competitive advantage that underpins market dominance. The competitive pressure on firms to use big data to optimize their processes has resulted in widespread adoption of data strategies at the firm level. McKinsey Global Institute estimates that virtually every sector of the United States economy now has more data stored than the Library of Congress.

3.2.2 Capture of Consumer Surplus

Big data on consumer preferences and habits enables companies to apply first degree price discrimination. This form of price discrimination involves a firm charging a different price for every unit consumed, based on the individual consumer’s reservation price. With perfect price discrimination, the firm captures all the consumer surplus. This is the business model for example of Uber.

3.2.3 Exploiting information asymmetry

The information advantage conferred by command of big data can be likened to a sixth sense – but an industrial strength sixth sense. Information asymmetry adds to the sources of potential market failure to the economy built on big data, which include: economies of scale (which are inherent in the investments required to capture, classify and curate); economies of scope (reflected in the increase in value of data when it can be cross-referenced with other types of data); and in many cases network externalities (especially in platform markets). All of these effects tend to promote market concentration (and thus market share capture for the leading firms).

Importantly, market mechanisms emerge to address information asymmetries in the normal course (as in the market for lemons; Akerlof, 1970). But the information asymmetry inherent in big data seems irreducible – there are no market solutions to correct for this information asymmetry. This is the “original sin” of the data-driven economy.

We expect information asymmetry to lead to market failure and indeed we observe it in the emergence of "superstar firms" and rising concentration in the leading data-driven economies. In the most intensively data-driven sectors, we see global near-monopolies. This reflects the exploitation of and thus the monetization of information asymmetry.

3.2.4 Shift of innovation into machine space – acceleration of product development

In the modern innovation-intensive economy, a major source of value in big data is that it enables machine learning (ML). ML is the industrialization of learning. This allows the acceleration of the process of innovation, providing a speed advantage to firms that are able to harness this element. We are at the dawn of the age of machine learning, but this is a source of value of data and the market will reward firms that can make a credible claim to an advantage in this area.

3.2.5 Training AI and the creation of machine knowledge capital

Data is the main feedstock in developing AI systems. The enormous strides in scaling AI systems in the last several years (Ciuriak 2023) inevitably meant there would be major breakthroughs and, of course, when ChatGPT dropped on 30 November 2022, this was viscerally illustrated. Generative AI is now the new buzz word but from an economic perspective, a more insightful term is machine knowledge capital as this points to how it fits into the array of productive capital assets. Machine knowledge capital is to human capital what robots are to physical labour. However, where robots are large, expensive and difficult to deploy, machine knowledge capital can be reproduced at essentially zero marginal cost and distributed globally with frictionless ease. Moreover, machine knowledge capital can be integrated into robots to make them more flexible and greatly extend the tasks they can perform, including in the service industries.

From the perspective of monetization of data at the firm level, the key observation is that the ability to create machine knowledge capital will enable market share capture: this follows from the economics of superstars, whereby even a small quality advantage will lead to dominance in market share and substantial rent capture (Rosen, 1981). From the perspective of national accounting, it implies a rising share of national income flowing to capital as reflected in a rising profit share of national income.

3.3 Some Evidence

The foregoing discussion suggests that the emergence of a major new form of capital that enhances firm profitability should be reflected in the economy in a number of ways.

First, the profit share in national income should be rising. Given the steep asymmetries in the ability to capture data rents, the overall contribution of data would be greater than the observed rise in the aggregate profit share, since some of this would be in effect cannibalized from pre-existing profits of corporations unable to use data.

Second, the share of income flowing to traditional legacy intellectual property (IP) assets should be flattening due to (creative) value destruction in an accelerated innovation context. Meanwhile, the share of IP accounted for trade secrets, the form of IP protection of choice in the data-driven economy, should be rising.

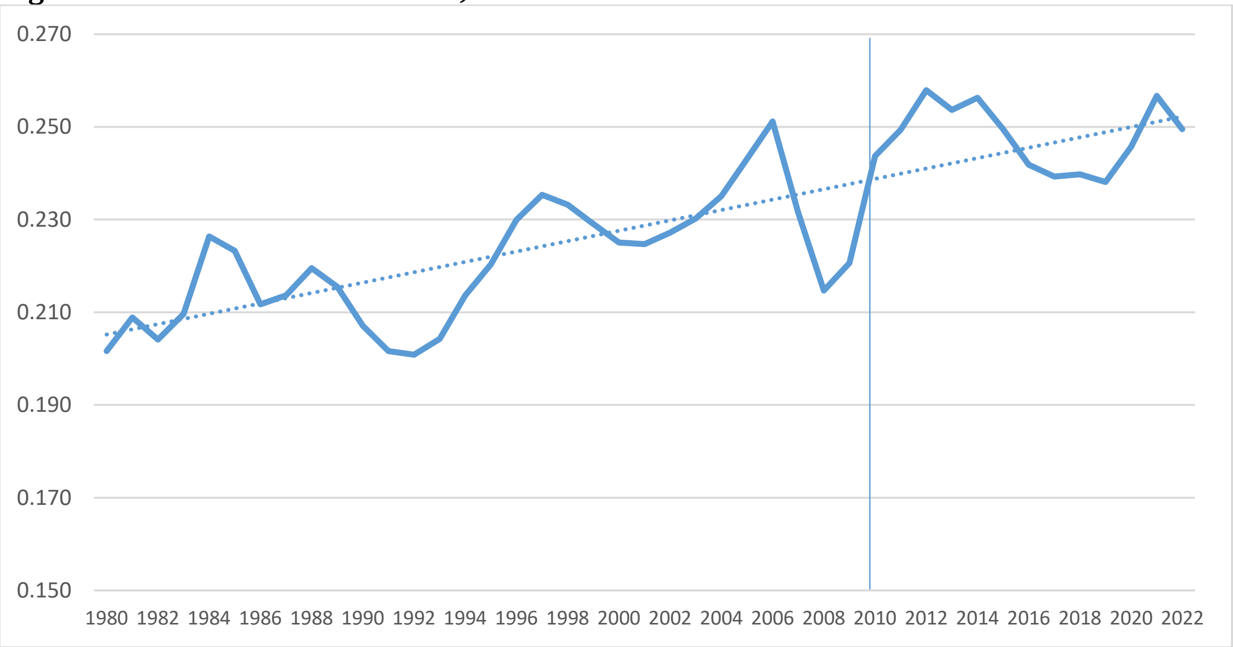
Third, markets would recognize the value by bidding up the market capitalization of data-rich firms and so these firms would rise on the leaderboards. And amongst the leaders, the share of assets accounted for by intangibles would be rising and the share accounted for by physical assets declining.

Fourth, the pace of innovation should be accelerating, and the pace of patenting of the main derivative product of data – AI systems – should be creating the next "hockey stick".

There is evidence for all these effects. It is useful to focus on the United States as the leading data-driven economy.

The US profit share of GDP has risen on trend since 1980, which is the beginning of the knowledge-based economy (KBE) era. That trend continued in the post-2010 era as we transitioned into the data-driven economy (DDE). Recalling that the pre-1980 economy was characterized by the "Kaldor facts" (including a constant labour share of income and constant returns to scale in industry), the rise in the profit share is to be attributed to IP.

Figure 2: US Profit Share of GDP, 1980-2022



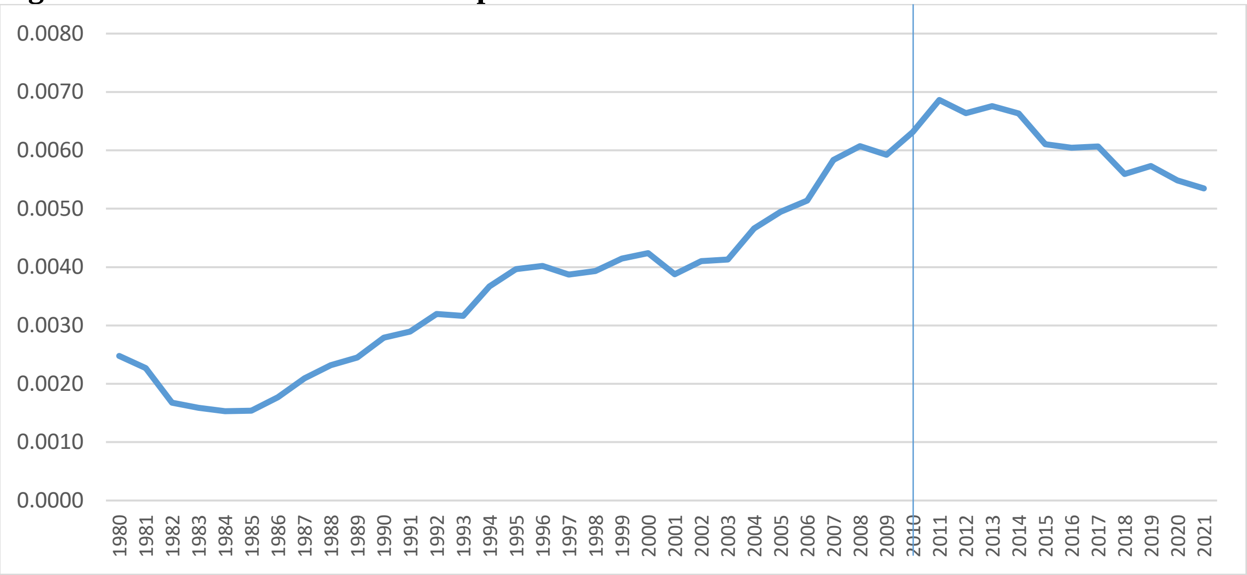
Source: US Bureau of Economic Analysis, National Income and Product Accounts, Net Operating Surplus. https://apps.bea.gov/i; author's calculations.

While the profit share of income continued to increase, US international receipts on traditional IP flattened out after 2010 in value terms and fell as a share of GDP. There is a reasonable inference that this was mirrored in overall returns to traditional IP in the US economy and that the continued rise in the profit share was due to data and algorithms.

The rising importance of trade secrets in firms' IP strategies (the form of IP for data and algorithms) is evidenced by the adoption of substantially elaborated and strengthened trade secrets laws around the world around 2016 as awareness of the value of data and the data-driven economy started to dawn.

The acceleration in the pace of innovation is starkly illustrated by the success of Google's machine-learning model AlphaFold in predicting the 3D structure of a protein from a given amino-acid sequence, facilitating the design of molecules for pharmaceutical development (Nourmohammad et al. 2022).

Figure 3: US International IP receipts as % of GDP



Source: World Bank Indicators, Charges for the use of intellectual property, receipts (BoP, current US$). https://data.worldbank.org/indicator/BX.GSR.ROYL.CD

The share of intangible assets in the S&P500 has risen from 16% in 1976 to as much as 90% or $32.3 trillion. Five of the six most valuable companies on the S&P 500 (by ETF ranking) are data-rich companies; their current market capitalization is about $8.7 trillion. Their tangible assets have been estimated at only around 5% of their total value.

Table 1: Market Capitalization of Leading Data-driven Firms

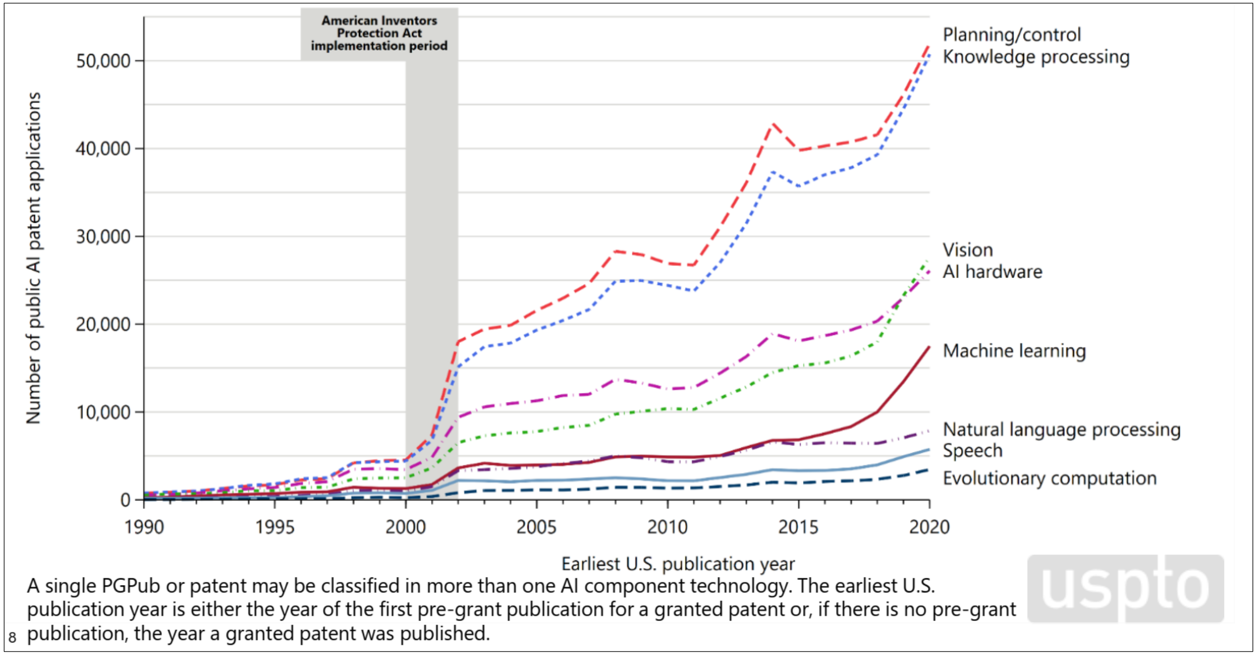
Company	Market Capitalization, 29 October 2023, USD billions
Apple	2,631
Microsoft	2,451
Alphabet	1,544
Amazon	1,318
Facebook	763
Total	**8,707**
Memo: S&P500	35,940
Of which intangibles	32,346

Source: https://ycharts.com/companies/MSFT; accessed 29 October 2023; intangible share: Ocean Tomo (2020)

The major share of the market value of these firms is comprised of IP and data – with data arguably comprising a dominant share as these companies became superstars in the age of data. The sixth member of the top group is Nvidia, whose rise has been jet-fueled by its leading position in making computer chips for AI use.

Finally, there was a clear upturn in the pace of AI patent applications in the United States after 2010. There has been a further steep upturn since around 2019, reflecting the extraordinary advances in the power of AI systems since then.

Figure 4: US AI Patent Applications Issued in the US, 1990-2020



A single PGPub or patent may be classified in more than one AI component technology. The earliest U.S. publication year is either the year of the first pre-grant publication for a granted patent or, if there is no pre-grant publication, the year a granted patent was published.

Source: Pairolero (2022), USPTO.

If we were to hazard a back-of-the-envelope guess as to the order of magnitude of the value of data to the US economy based on the premises outlined above, we could make the following calculation:

Table 2: Back-of-the-envelope estimate of the value of data in the United States, 2022

Item	Value/%
Increase in the trend profit share of US GDP, 2010-2022	1.46%
Estimated Profit Increment in 2022 (USD billions)	371
Take account of declining share of traditional IP (USD billions)	371
Total Returns to Data in 2022 (USD billions)	741
Share of GDP (percent)	2.91%
Assumed ROI on S&P500 (percent)	10%
Asset value of data (USD billions)	7,411

Source: Calculations by the author

US net worth is estimated by the Federal Reserve to be on the order of $135 trillion (Simko and smith 2023). The value of US domestic businesses is estimated at $55 trillion while IP and “other” similar assets are treated as a separate line item with a value estimated at $2.3 trillion. If we consider the value of data to be part of the value of US enterprises, the figure of $7.4 trillion would represent a mere 13.5% of total enterprise value. If we express it as a share of the intangibles in the S&P500, it is 22%. If we consider the value of the top 5 data-intensive corporations and assume that consistent with an 80-20 rule that they account for 80% of the total data returns in the US economy, these figures seem hardly unreasonable as an estimate of the private value of data in the United States and indeed likely on the low side.

4 Discussion and Conclusions

Traditionally, the accounting of value creation in an economy has looked through the firm to its constituent elements (labour, capital, IP) and ignored externalities (with the exception of special purpose accounts such as "green GDP"). This may not be tenable any longer with the advent of a data-driven economy.

4.1 Looking at firms not through them

Studies using micro data on populations of firms have established that firms that trade tend to be larger, more productive, more innovative, and pay higher wages than domestically-oriented firms; and that firms that become multinationals dominate trading firms the way trading firms dominate domestic firms. For disciplines such as international trade, a firm-centred analysis is critical to understanding the implications of trade or investment liberalization. This is also true of economic development, where the so-called "missing firms" problem is associated with under-development and phenomena such as the ”middle income trap”. Meanwhile economic development is not associated with a specialization or narrowing of activities at the national economy level but a Cambrian explosion of firm types and an increase in firm size and technological capability.

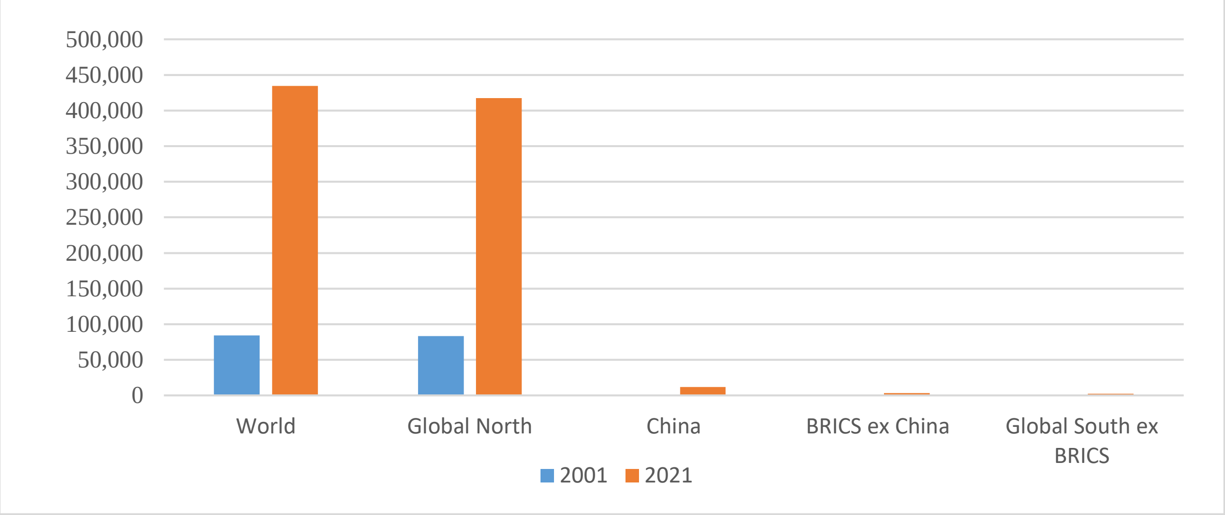
Productivity is not an abstract concept identified as a residual in a production function that looks through the firm – it is intrinsic to firm size and technological (and hence data) capability. Larger firms are able to subdivide tasks more finely, increasing the specialization and efficiency of individual workers (recall Adam Smith and the pin factory; see Ciuriak 2016). As the number of defined tasks increases, these can then be spun off as separate firms that in turn are able to scale up and repeat the process. Trading firms gain access to global scale economies and global value chains.

To be useful, technology must be inside a firm and intrinsic to its products and production processes. Here it is instructive to look at where technologically sophisticated firms are located (Figure 5). They are almost all in the Global North. That's what defines the divide: technology. China is heading north in terms of traditional knowledge production.

If we had a similar chart for data, we would almost certainly see China much larger and close to if not ahead of the Global North: Ericsson reports China's mobile data traffic = 26 exabytes/month compared to 6 in North America (Ciuriak 2023).

This skewed distribution of technological capability is reflected in firm counts: the vast majority of Global 500 firms are in the Global North and China (China ranks first in the 2023 rankings with the United States second). Similarly in Unicorn counts, the Global North (with the United States first by a country mile) and China dominate.

Figure 5: International IP Receipts, Selected Groups, 2001 and 2021, USD Millions



Source: World Bank Indicators, Receipts for Intellectual Property; calculations by the author

The advent of the data-driven era changed the conditions of competition between nations: in the past decade, China filed 389,571 patents in the area of AI, or 74.7% of the world total (WIPO 2021). This reflects Chinese firms' technological capabilities and equally importantly access to vast quantities of data. Baidu reports that there are some 8 million developers using its AI development platform, PaddlePaddle. This almost certainly dwarfs the number of AI developers in the Global North combined. On the pre-2010 technological terrain, China trailed by a large margin. On the new one, defined by data, it does not. When we look at the world through the lens of the firm and technology, we can immediately see why we are in a global technology war. The value of data seen through this lens is far greater than its pure commercial value. But importantly, this value remains inside the firm.

4.2 Data externalities

The societal benefits of innovation unlocked by data (e.g., implicitly learning the physics of protein folding) and the AI applications that are now being deployed and developed are at this point beyond our capability to value. However, it is likely that they are substantially greater than the private commercial benefits captured by firms. Support for innovation has long been the key industrial policy in the advanced countries and such support is now fully justified for data-driven innovation. The fact that we do not have adequate measures of the value of data in terms of enterprise value let alone in terms of positive externalities means that we are almost certainly under-investing.

However, the negative externalities are also massive. The digital transformation made data the "new plutonium" when applied in social and political contexts, including ushering in the age of disinformation (Ciuriak 2021). The toxicity of social media is now the daily bread of commentary. The scope for data-driven disruption of societies has been demonstrated at the technical level: the notes published by OpenAI with the release of GPT4 commented on the ease with which algorithms could be tuned to generate hostility and friction between social groups. Similarly, data-driven innovation that enables discovery of new medicines that address disease while minimizing toxicity to humans can be tweaked to maximize toxicity to humans. In one experiment, such a simple tweak allowed the AI to discover VX and other lethal compounds.

And of course there are the incalculable risks associated with increasingly powerful AI that is beyond human capability, especially when it is empowered with agency as is increasingly being done, including in drone warfare.

Accordingly, while ignoring externalities was excusable when externalities were understood to be a marginal knock-on effect to the main economic results, it is not when the externalities might well be much greater than the direct commercial value. Simply put, if the value of data by any calculation turns out to be a few percentage points of GDP, it's significance in policy priorities will for all intents and purposes fall off the radar.

4.3 Concluding remarks

The digital transformation enabled the emergence of a data-driven economy, in which data became the new essential form of capital that could capture economic rents, but did not fit the conventional paradigm of being a traded good with a market price that would allow the aggregation of individual transactions into national accounts. In confronting this conundrum, the thesis in this note is that it forces us to stop looking through the firm in constructing national accounting frameworks and to look at the firms, recognizing their heterogeneity, and their role in value creation, productivity and innovation. Enterprise value is critical to establishing the value of data.

References

Akerlof, George. 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," The Quarterly Journal of Economics 84(3), August: 488-500.

Ciuriak, Dan. 2023. "Optimizing North American Supply Chains in Critical Technologies: The USMCA Digital Advantage," Chapter 6 in Joshua P. Meltzer and Brahima S. Coulibaly (eds) *USMCA Forward 2023*, Brookings: https://www.brookings.edu/essay/usmca-forward-2023-chapter-6-data-flows-and-critical-technologies/.

Ciuriak, Dan. 2021. "The Age of Disinformation: The Role of Market Power in Information Space," Ciuriak Consulting Discussion Paper. https://papers.ssrn.com/abstract=3944863

Ciuriak, Dan. 2019. "Unpacking the Valuation of Data in the Data-Driven Economy," Notes for Remarks at the Guarini Law and Tech Centre conference on Global Data Law, NYU Law, New York, 26-27 April 2019. https://papers.ssrn.com/abstract=3379133

Ciuriak, Dan. 2016. "Productivity and Innovation: The MFP Puzzle Considered Through the Lens of Trade Theory," Ciuriak Consulting Working Paper. http://papers.ssrn.com/abstract=2761414

Deutsch, Claudia H. 2004. "A Campaign for BASF," The New York Times, 26 October. https://www.nytimes.com/2004/10/26/business/media/a-campaign-for-basf.html

Duarte, Fabio. 2023. “Amount of Data Created Daily (2023),” Blog, ExplodingTopics.com, 3 April. https://explodingtopics.com/blog/data-generated-per-day

Goldstein, Joseph L. 2019. “Seurat’s Dots: A Shot Heard ’Round the Art World—Fired by an Artist, Inspired by a Scientist,” *Cell* 179, 19 September: 46-50.

Kelly, Kevin. 2017. “The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future,” Penguin Books.

Knight, Will. 2018. “Google’s Self-Training AI Turns Coders Into Machine-Learning Masters,” MIT Review, 17 January. https://www.technologyreview.com/2018/01/17/146164/googles-self-training-ai-turns-coders-into-machine-learning-masters/

Nakamura, Leonard, Jon Samuels, and Rachel Soloveichik. 2018. “"Free" Internet Content: Web 1.0, Web 2.0, and the Sources of Economic Growth,” Federal Reserve Bank of Philadelphia Working Paper, WP 18-17, May 2018

Nourmohammad, Armita, Michael Pun, and Gian Marco Visani. 2022. “Machine-Learning Model Reveals Protein-Folding Physics,” *Physics* 15, 28 November: 183. https://physics.aps.org/articles/v15/183

OECD. 2013. “Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value,” OECD Digital Economy Papers, No. 220, OECD Publishing, Paris. http://dx.doi.org/10.1787/5k486qtxldmq-en

Pairolero, Nicholas A. 2022. “Artificial Intelligence (AI) trends in U.S. patents,” Powerpoint presentation, Artificial Intelligence and Emerging Technology Inaugural Stakeholder Meeting, 29 June 2022.

Pelley, Scott. 2023. “Is artificial intelligence advancing too quickly? What AI leaders at Google say,” CBS News, 16 April. https://www.cbsnews.com/news/google-artificial-intelligence-future-60-minutes-transcript-2023-04-16/

Rosen, Sherwin. 1981. “The Economics of Superstars.” *American Economic Review* 71(5): 845-858.

Simko Paul J. and Richard P. Smith. 2023. “U.S. Net Wealth Is Over $135 Trillion. Here’s Where That Money Resides,” Barrons, 13 September. https://www.barrons.com/articles/net-wealth-is-over-135-trillion-where-that-money-resides-d722c9c6

WIPO. 2021, “China Leads the World in AI Related Patent Filing,” Conference report, 28 September. https://www.wipo.int/about-wipo/en/offices/china/news/2021/news_0037.html