Comment on "The Missing Value of Data"

by Ankit Bhutani, Guillermo Ordoñez, and Laura Veldkamp

John Fernald*

INSEAD

The Bhutani, Ordoñez, and Veldkamp (2026) paper is terrific. I learned a great deal from thinking about the issues it raises. It identifies and quantifies a novel channel through which investments in data are missing from GDP. The channel is that better data improve the firm's forecast of its own prospects, which leads to better decisions and fewer costly mistakes, raising profits. This channel is surely not the most important one for missing data to matter for GDP measurement. But it raises thorny measurement challenges. The paper suggests a method for moving forward that might be applicable to other challenges as well.

In this comment, I make three points. First, the revenue-forecasting channel identified in this paper is probably not where most of the action is on data. That said, the channel is novel and the approach is innovative. The paper complements the approach typically taken in the large and growing national accounting literature on intangible investments, including in data.

Second, I am not persuaded at all by the consumer-side arguments in the paper. Yes, data barter is important. But the paper isn't measuring that data barter. Rather, it takes its estimate of the value of improved firm forecast precision and, in its benchmark case, assumes that the consumer gets the same value through data barter. So when the paper says that the aggregate value of data is about 1-1/2 percent of GDP, that is double the number it actually quantifies as missing data investment.

Third, even the direct estimate it finds for missing data investment, about 3/4 percent of GDP, is too high. The paper, loosely speaking, identifies the value of data to a firm by how accurately the firm forecasts its revenues relative to what it would forecast using public data. The public data they use for this comparison are stale relative to the publicly available data

*Prepared as a discussion of Bhutani, Ordoñez, and Veldkamp, "The Missing Value of Data," for the NBER Macroeconomics Annual 2026. We are grateful to the editors and conference organizers for the opportunity, and to the authors for a stimulating paper and for generous and good-humored discussions of it. I especially thank Sabine Bejjani for helpful research assistance and discussion. The views expressed here are my own; any errors are mine as well. Correspondence: john.fernald@insead.edu.

when the firm actually provides guidance about revenue. Hence, their estimates of missing investments in data are overstated. Somewhat subjectively, I will discount their estimates by a third. So in my interpretation, their approach suggests that adding their channel would boost GDP by perhaps 1/2 percent.

Motivation

Let me begin with a motivating quotation from a different Laura Veldkamp talk, one she gave last year at INSEAD (Veldkamp 2025): "Data is one of the most important and highly valued assets in the modern economy...."

I have to stop the quote there, because it already exposes a deep controversy in the literature, a divide on which the authors and I are on opposite sides: are data singular or plural? This paper, like Laura's prior work, treats data as singular. However, I spent decades in the Federal Reserve System, which is as careful about grammar as it is about data. The issue as settled. Data *are* plural. Admittedly, when I raised the point with Claude, it advised me that the plural usage "would probably sound slightly odd to most people under 50." That is a comment on the authors' enviable youth, not their grammatical wisdom.

Where we can certainly agree is that the *challenges* of data are plural—they are many. Laura has written a book around the challenges (Baley and Veldkamp 2025). Guillermo has written prolifically about information, precision, and prediction dynamics—the aspect of data taken up by this paper.

So let me repair and continue Laura's quote: "Data *are* one of the most important and highly valued assets in the modern economy—and also one of the hardest to observe, measure, and put a price on." Observing, measuring, and putting a price on the data economy is precisely what this paper is about.

1. A novel channel, but not where most of the value of data lies

The objective here is to quantify one data channel, which is improved forecast precision regarding a firm's *own* near-term revenue. In their model, better prediction leads to fewer mistakes and higher profits. They find quantitative evidence consistent with that channel. Using their model and empirical estimates, they can impute a GDP value for the missing data investments firms make in improving their precision.

Let me be clear about what the paper is *not* capturing. It is explicitly not capturing data

as a factor of production. Think of Google. The value of data to Google is not mainly that it can better forecast its own revenue or earnings this year. The whole company is about data. Some of Google's data investments are currently included in GDP, though much surely is not. That's separate from the issues in this paper.

Nor are we capturing whether firms with more data erect barriers to entry and charge higher markups; or the effects of data on allocative efficiency or the chances of a financial crisis. From a measurement standpoint, that is fine—those indirect effects, to the extent they raise or lower output, are already in the GDP data we have. This paper (Bhutani, Ordoñez, and Veldkamp 2026) is about something narrower but also novel: investments in a stock of knowledge that improves a firm's forecast precision. It is a narrow channel, but one we did not previously have any way to measure.

It complements, rather than competes with, the large and growing national-accounting effort to bring intangibles into GDP. The U.S. national accounts added software in 1999, and R&D and other intellectual-property products in 2013.[1] The 2025 System of National Accounts (SNA) proposes to add *own-account* data and databases—the digital records a firm builds with its own employees, valued at input cost, taking care not to double-count what is already inside software or R&D—and recent BEA work by Santiago Calderón (2026) implements exactly this for databases. The broader intangibles program of Corrado, Hulten, and Sichel (2009), and the INTANProd effort that runs in parallel with EU KLEMS, push further into organizational capital, branding, and training. This paper is also about measuring otherwise-missing intangible investments in forecasting accuracy about a firm's own prospects. Accuracy translates into profitability, so it is worthwhile for the firm to invest in accuracy. That is a novel data channel.

There is an important difference relative to the SNA and BEA work on data. That work focuses on digital databases. But nothing in this paper requires that the data be digital. Indeed, the channel in this paper has arguably existed for as long as there have been merchants who placed orders in advance, rather than simply producing products when a customer ordered them. Concretely, Sears-Roebuck a century ago was effectively a data company, even though digital records did not exist at the time. This paper would count their data investments; the SNA 2025 would not.

Comparing this paper's estimates with other estimates of data and intangibles reassures me that the order of magnitude might be about right. Figure 1 compares my estimate of missing data from the paper against other measures of missing—or, in the case of R&D,

[1] R&D was capitalized in Bureau of Economic Analysis's (BEA) 2013 comprehensive revision of the NIPAs (released July 31, 2013), which introduced the "intellectual property products" asset category encompassing software, R&D, and entertainment, literary, and artistic originals.

already included—intangible investment, all as a share of 2022 GDP.

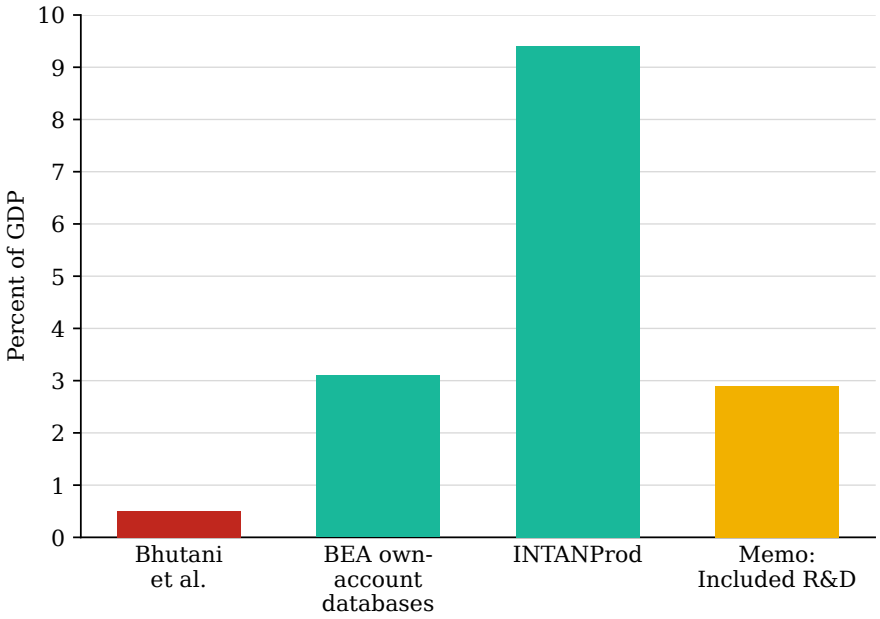

Figure 1: Unmeasured intangible investment, 2022 (percent of GDP). The Bhutani et al. bar is the paper's investment channel, stripping out the barter doubling and downsizing by one-third as described in the text. Sources: Bhutani et al. (2026); BEA own-account databases from Santiago Calderón (2026); INTANProd from `euklems-intanprod-llee.luiss.it`; already-included R&D from BEA.

A word on how I construct the paper's bar. Their headline figure is that we are missing about 1-1/2 percent of GDP. But that includes the 3/4 percent of GDP that they estimate is missing data investment, plus an equal 3/4 percent of GDP that they assume on a priori grounds is missing barter consumption. I strip out the doubling since, as I discuss below, the paper has no actual evidence on that channel; and the logic for that channel is weak. In addition, including the missing consumption would make it less comparable to the other estimates shown in Figure 1, which are about investment, not consumption. As I also discuss below, even the 3/4 percent of GDP estimate is probably overstated. So I drew the bar for this paper as quantifying missing GDP of 1/2 percent.

By comparison, R&D investments already included in GDP amount to about 3 percent of U.S. 2022 GDP. Santiago Calderón (2026) estimates that not-currently-included investment in digital databases amounts to an additional 3 percent of GDP. The INTANProd estimates of broad, not included intangibles would add about 9-1/2 percent. Compared with these estimates, 1/2 percent of GDP for the forecasting channel in this paper seems plausible. The BEA databases number covers *any* use of data captured in a database, so a single aspect of data ought to come in below it, and it does. On the other hand, one has to be careful,

because the BEA figure is *databases*, whereas this paper's precision channel might involve digital databases or might not. As noted, the precision channel existed prior to the first creation of a digital database. That said, it would help quantification to know where the precision comes from, to be sure we are not already counting that investment somewhere else.

As a proof of concept, then, the exercise clears the bar. A missing-investment channel of roughly 1/2 percent of GDP—more or less—is plausible.

2. The barter/consumption channel should be treated separately

Before discussing their investment-channel estimates, let me comment on their assumption of barter-consumption equivalence. They assume that whatever they find for missing data investment was paid for by the firm by giving the consumer something of equal value for free, outside of our measures.

But we can think of many examples where the unobserved barter component is not present. Suppose Quaker Oats is deciding whether to launch some new chocolatey chocolate-chip flavor of oatmeal. It runs focus groups to learn whether parents still think of it as healthy oatmeal even as the kids love the chocolate and sugar. Suppose Quaker pays each focus-group participant $100 in cash. Households use the cash to buy what they want and the national accounts measure that. There's no barter component at all. There's no missing consumption, in real or nominal terms.

As another example, suppose CVS gives me a discount—buy one bottle of shampoo, get a second one free—if I use my discount card. Suppose the normal (non-discount) price of each bottle of shampoo is $10. What do the national accounts see? They see I bought two bottles of shampoo for $10, or $5 each. Both of those are correct. That is, CVS gets my data, and the BEA correctly sees that I bought two bottles of shampoo for $5 each. There's no missing consumption, in real or nominal terms.

The point is that the barter equivalence is shaky. It's not just markups that make it shaky but the equivalence itself. Firms pay for data, but some of those payments are recorded correctly in the accounts. Some might not be. Without more evidence, we don't know.

For this reason, my strong preference is to keep the focus on the investment piece, which is what the paper actually measures. The barter channel is, at best, directionally right.[2]

Of course, we do know from a range of work that there are large missing consumer benefits,

[2]There is a further asymmetry: in 2020 forecast precision collapsed on an unforecastable shock, yet households went on receiving free digital benefits—so the investment and consumption values can move in opposite directions, which a fixed ex post one-to-one mapping cannot represent.

including from data barter. A notable channel is the free things one gets on the internet (Google Maps, YouTube videos, and so forth). For example, Brynjolfsson and co-authors estimate these at several percent of GDP using online choice experiments and a "GDP-B" framework (Brynjolfsson et al. 2025); Nakamura, Samuels, and Soloveichik (2017) value free, advertising-supported digital content within a national-accounts production framework; and Soloveichik's (2024) more recent accounting puts the figure far higher still. That literature is important and the direction is clear. Quantitatively, missing consumption probably swamps the investment channel in this paper. But that literature handles missing consumption explicitly and carefully, rather than through an ad hoc multiplicative factor.

So I would set the barter doubling aside and read the paper's contribution as the investment channel alone. The paper finds that the missing investment in precision is about 3/4 percent of GDP. The next section argues that even this is too high.

3. Even 3/4 percent is too high: the public prior is stale

The model, and the way the authors map it to the data, is a clever contribution. Let me highlight a few features. Firms forecast an unknown fundamental state, which has a persistent component they can learn about from the past plus a transitory shock they cannot foresee. Each firm carries a stock of precision, Ω, the total precision of its beliefs about the state. Part of that precision, ϕ, is *public*—what the firm could predict from history alone, with no investment in data at all—so the difference $\Omega - \phi$ is the precision that comes from the firm's own data. That difference is the quantity the paper capitalizes, so it is where all the action is.

Empirically, the authors recover Ω from Compustat and I/B/E/S. For a subset of public firms, I/B/E/S provides the company's own annual sales guidance, issued early in the year, and Compustat provides realized sales. From these the authors compute each firm's revenue forecast error and invert it into a data stock. A firm that forecasts well has a high stock, and a firm that forecasts badly has a low one. For the public component ϕ, they form an AR(1) forecast of this year's revenue from *last year's* realized revenue and compute its forecast error the same way. The calibrated model then maps the gap $\Omega - \phi$ into dollars.

The key estimates are in the paper's Figures 2 and 3. The paper's Figure 2 plots the prior and posterior variances—the forecast-error variance from the public AR(1) against the variance from firms' own, better-informed forecasts—and, unsurprisingly, firms' own forecasts are more precise than an AR(1) on last year's published revenue. The paper's Figure 3 converts this into a data stock in precision units. The height of each bar is the precision of firms' own forecasts; whatever the public component does not explain is attributed to private data; and the model splits that private piece into newly acquired and old, undepreciated

data.

One thing jumps out—old data contribute remarkably little; this is barely a capital stock at all. The implied depreciation rate is 75 to 80 percent. The model does require it to be a capital good (depreciation below 100 percent), but as calibrated it sits close to an intermediate input, something firms produce and largely use up within the year.

Still, my main concern is that the public prior and the firms' own forecasts are dated differently, and the mismatch inflates the private component. The AR(1) prior sees only the previous year's annual revenue, whereas firms issue their guidance with far more current public information already in hand. If we understate how much is publicly known at the moment the firm forecasts, we understate ϕ and mechanically overstate the private contribution $\Omega - \phi$—which, in turn, becomes "missing data investment." The better the public prior, the less is left to call private, and the smaller the capitalized number.

The year 2008 makes the problem concrete. In the paper's Figure 2, firms' own guidance forecasts the recession year reasonably well, so Ω is high. At the same time, an AR(1) on 2007 revenue does badly, so ϕ is low. The *gap* $\Omega - \phi$—the private component—therefore *rises* in the Great Recession.

What I think is going on here is a timing artifact. For the typical firm, 2007 was a good and fairly normal year, so an AR(1) built on it badly over-predicts a weak 2008. Firms, by contrast, issued their 2008 guidance early in 2008, by which time the deterioration was already widely visible in public information that the 2007 revenue figure does not contain.

To see this more clearly, consider when 2008 revenue guidance was given in I/B/E/S. Figure 2 shows that the single most common month is February 2008, with the overwhelming bulk issued in the first quarter.

By the first quarter of 2008, the public information set was vastly richer than "2007 annual revenue." Consider what Janet Yellen, then president of the San Francisco Fed, was saying in real time at FOMC meetings.[3] At the October 2007 meeting, "developments since we met [in September]... generally have been favorable and the risks to the outlook for growth have eased somewhat"—late 2007 looked not terrible, which is roughly what the AR(1) prior embeds. But by the January 2008 meeting she said, "The severe and prolonged housing downturn and financial shock have put the economy at, if not beyond, the brink of recession... My contacts have turned decidedly negative in the past six to eight weeks." And by March 2008 she noted "an increase in the frequency and intensity of pretty dire comments I am hearing from my contacts."

[3]Janet Yellen, President of the Federal Reserve Bank of San Francisco, in Federal Open Market Committee meeting transcripts: meeting of October 30–31, 2007, p. 37; meeting of January 29–30, 2008, pp. 51 and 53; and meeting of March 18, 2008, p. 30. Transcripts released by the Board of Governors of the Federal Reserve System.

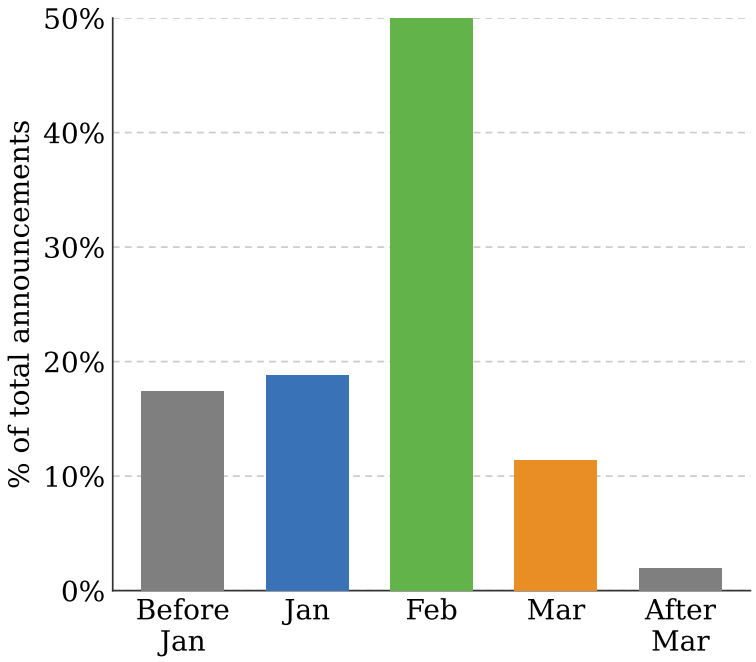

Figure 2: Timing of 2008 revenue guidance: share of I/B/E/S annual guidance announcements by month of issue. Most 2008 guidance was issued in the first quarter, after substantial public information about the emerging downturn had already arrived.

If one dislikes anecdotes, the hard data say the same thing. In the first quarter of 2008, when firms were issuing their 2008 guidance, publicly available retail sales, new orders, and employment were already shrinking, and published sentiment was souring.

So the prior built off 2007 annual revenue is far too optimistic for 2008, and therefore far too imprecise. The model closes the resulting gap by assuming there was a lot of investment in private data. My guess is that this investment is spurious.

This problem is fixable. Let the prior incorporate the high-frequency public information actually available at the guidance date, not just the previous year's firm-level annual revenue. And the implication is general, not specific to 2008. Do a better job with the public prior and there is simply less room for firm-specific investment. That is why I would mark the private, capitalizable component down even taking the model as given. In the interest of round numbers, I will subjectively shave off a third, which brings the investment channel to something like 1/2 percent of GDP.

Wrapping up

To wrap up, this is a terrific and thought-provoking paper. It identifies a new channel for missing data investment which is complementary to what national accountants are already doing. Prior to reading this paper, I would not have known how to begin quantifying this channel. The authors have produced a plausible answer. It is a proof of concept—the first word, not the last. There is ample room for improvement here. This is fertile ground for a generation of second-year papers tweaking the prior, the depreciation assumptions, and the mapping to dollars.

My own reading is that the paper aims to measure one narrow slice of the data economy—and not necessarily the most important slice. Within that slice, the authors' investment channel comes to about 3/4 percent of GDP, and my concern about the public prior suggests the truly private, capitalizable part is smaller still—on the order of perhaps 1/2 percent of GDP. But the precise number matters less than the creativity of the method. The one caveat every reader should carry away is that this is only the tip of the iceberg. Data are valuable for many reasons that lie entirely outside this exercise, as the authors' own work reminds us.

References

Baley, Isaac, and Laura L. Veldkamp. 2025. *The Data Economy: Tools and Applications.* Princeton, NJ: Princeton University Press.

Bhutani, Ankit, Guillermo Ordoñez, and Laura Veldkamp. 2026. "The Missing Value of Data." NBER Macroeconomics Annual 2026 (this volume).

Brynjolfsson, Erik, Avinash Collis, W. Erwin Diewert, Felix Eggers, and Kevin J. Fox. 2025. "GDP-B: Accounting for the Value of New and Free Goods." *American Economic Journal: Macroeconomics* 17 (4): 312–344.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55 (3): 661–685.

Federal Open Market Committee. 2007. "Meeting of the Federal Open Market Committee on October 30–31, 2007." Transcript. Board of Governors of the Federal Reserve System. `https://www.federalreserve.gov/monetarypolicy/files/FOMC20071031meeting.pdf`.

Federal Open Market Committee. 2008a. "Meeting of the Federal Open Market Committee on January 29–30, 2008." Transcript. Board of Governors of the Federal Reserve System. `https://www.federalreserve.gov/monetarypolicy/files/FOMC20080130meeting.pdf`.

Federal Open Market Committee. 2008b. "Meeting of the Federal Open Market Committee on March 18, 2008." Transcript. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/monetarypolicy/files/FOMC20080318meeting.pdf.

Nakamura, Leonard I., Jon Samuels, and Rachel H. Soloveichik. 2017. "Measuring the 'Free' Digital Economy within the GDP and Productivity Accounts." Federal Reserve Bank of Philadelphia Working Paper 17-37.

Santiago Calderón, José Bayoán. 2026. "Toward Economic Statistics on Own-Account Data and Database Assets for 1997–2024." BEA Working Paper WP2026-4. Washington, DC: U.S. Bureau of Economic Analysis. https://doi.org/10.66137/KLUT2015.

Soloveichik, Rachel. 2024. "Private Funding of 'Free' Data: A Theoretical Framework." BEA Working Paper WP2024-1b. Washington, DC: U.S. Bureau of Economic Analysis.

Veldkamp, Laura. 2025. "Valuing Data as a New Asset Type." Keynote lecture, INSEAD Finance Symposium, June 2025.