The Missing Value of Data

Authors: Ankit Bhutani, Guillermo Ordoñez, Laura Veldkamp

Discussants: John Fernald, Dimitris Papanikolaou.

Gabriel Chodorow-Reich opened by building on a point raised by John Fernald, pressing for more precision on what inputs beyond data itself drive forecast accuracy. He framed it as a production function problem: forecast precision is a joint output of data and its analysis. Since data scientists and analytical software are already captured in GDP as labor income and investment, he asked how the paper isolates the unmeasured residual value of data per se.

Veldkamp acknowledged the point, noting that prior work with Simona Abis had estimated a richer production function for the financial sector, but that the current paper lacks firm-level employment data to replicate that exercise economy-wide. The approach nonetheless remains valid: the aggregate wage bill for data-related workers is netted out of profits before computing data's value. Ordoñez added that the existing literature has also attempted to clean software and database costs from measured profits, meaning the paper targets precisely the portion not yet accounted for elsewhere.

Jennifer La'O raised a conceptual aggregation problem. The paper measures the sensitivity of firm profits to reductions in forecast errors, but she argued that mapping this to GDP requires taking a stand on what those forecast improvements represent. If they capture genuine efficiency gains, product quality improvements, or better consumer-firm matching, lower forecast errors create new value. But if they primarily reflect business stealing rather than new production, aggregate output may be unchanged, weakening the mapping from private firm value to GDP. Veldkamp agreed that this business-stealing is an externality, and emphasized that the paper deliberately focuses first on measuring the private value of data before addressing its broader social value. Ordoñez then conceded that the extrapolation from the sample to the full economy is heroic, as the paper does not differentiate between large and small firms in their data use or account for nonlinearities in how data quantity translates to value.

Valerie Ramey raised two points. First, better measurement of data investment should reduce measured TFP growth, since TFP is a residual and more properly attributed capital means less unexplained growth. Second, she appreciated the matchmaking angle on data value highlighted by Papanikolaou, illustrating it with consumer-learning examples such as Amazon. Ordoñez responded that the matchmaking intuition fits naturally within the paper's framework: better knowledge of customers reduces forecast errors and can alternatively be interpreted as reducing a form of misallocation. Veldkamp added nuance on the TFP question, noting that while better-measured data investment would reduce measured TFP through the capital channel, the paper also identifies an offsetting source of mismeasurement on the output side because consumption bartered for data is omitted from national accounts. In the paper's accounting framework, these two channels offset one another, leaving the overall empirical effect on measured TFP ambiguous.

Veronica Guerrieri then returned to La'O's business-stealing point, arguing that if data merely redistributes customers among firms, measured GDP should be unaffected since those revenues

are already counted in national accounts. The missing component is therefore not the business-stealing itself, but rather the underlying investment in the data asset. Veldkamp agreed that business-stealing effects would already be reflected in GDP, and Ordoñez noted that the sample of roughly one thousand SEC-filing firms likely already embeds substantial business-stealing among large firms, with the critical assumption arising when extrapolating to smaller firms where competitive dynamics may differ.

Gianluca Violante asked how the model accounts for the commercial market in data, citing Equifax and CoreLogic as examples, and whether observed transaction prices could be informative about the precision or quantity of data exchanged. He further suggested that AI-driven demand growth should push commercial data prices upward, which would inform the TFP question. Veldkamp responded that GDP measures newly created value rather than the stock of data, so even if AI raises demand for commercial data, diminishing returns imply that successive units contribute progressively less to measured output, helping explain the relative stability of their Missing GDP estimate over two decades. Ordoñez addressed the market-transaction point using the paper's Visa appendix exercise, which found that the precision of commercially purchased data was roughly one-quarter that of internally generated data, since purchased data is generic rather than specific to the firm's own customer base.

Wendy Edelberg closed by encouraging the authors not to abandon the supply side of the data economy, noting that individuals generate data at near-zero variable cost and asking what that implies, especially given generational shifts in attitudes toward data sharing. Ordoñez replied that while generation costs are low, rapid depreciation limits the value of repeatedly generated data, since firms often learn most of what they need about customers after only a small number of interactions. Veldkamp noted that privacy constitutes a real cost for those who value it, tempering the notion that data generation is truly costless.