IARIW – CIGI 2023

IARIW – CIGI 2023

Thursday, November 2 – Friday, November 3

Valuing Data or Collecting Data on Data –Which Priorities?

Michael Wolfson (University of Ottawa)
mwolfson@uottawa.ca

Paper prepared for the Conference on The Valuation of Data November 2 – November 3, 2023
Session 1: Methodologies of Data Valuation: National Accounts Approaches
Time: Thursday, November 2, 2023 [10:00AM-11:30 AM EST]

DRAFT

Valuing Data or Collecting Data on Data –Which Priorities?

Michael Wolfson, mwolfson@uottawa.ca

IARIW-CIGI conference "The Valuation of Data"

November 2-3, 2023, Waterloo Ontario

"The evidence of huge technologically driven change is everywhere in daily life, and almost nowhere in the standard economic statistics. (Coyle, 2021, p138)

Abstract

With the dramatic and rapidly growing role of "data" in contemporary societies, there is increasing interest in how best to reflect this reality in official statistics. One approach is to assign monetary values using the framework and methods of the System of National Accounts. In this paper, however, we argue that such an approach will inevitably rest on arbitrary assumptions and logically flawed underlying theory insofar as "data" are conceived as a form of aggregate capital. Instead, the focus is on the kinds of official statistics providing "data on data" that would be of greater utility. This alternative approach to recognizing the value of data is motivated by two major policy needs, privacy and health, and on two more general "social proprioception" concerns, inflation and entertainment. The conclusion is that official statistics efforts should focus on creating a microdata "portrait" not only of data bases, but also the data flows among them.

Introduction

It is no overstatement to say that there are revolutions underway in the volumes and speeds of computerized data flowing around the planet. A major question is how National Statistical offices (NSOs) should prioritize statistical developments that accurately portray the dramatic economy- and society-wide increases in the roles of "data" in their statistical programs.

One view is that the role of data needs to be better incorporated into national accounting, in order to ensure that GDP growth is appropriately measured, as well as reflected in related statistical measures like productivity.

This approach seeks to value data in monetary terms. It would be accomplished by forcing statistical descriptions of society's stocks, flows, and uses of data into the framework of the System of National Accounts (SNA). Doing so faces innumerable challenges, not least how to put a dollar value on a given set of data, and the meanings attached to aggregations of these dollar values.

An alternative view is that statistical information on data is far too heterogeneous and has far too many potential uses to be reflected primarily in the SNA, the view taken in this paper.

The key question, then, is kinds of collections of data on data most needed, and more specifically for what public good purposes. We focus on two groups of such purposes: to support a range of public policies, and to provide a portrait for the public at large to appreciate and understand these dramatic trends – in a phrase for this latter purpose, to provide the material for social proprioception.

With regard to the first approach, there is a long but generally ignored history of objections to the aggregation involved in constructing the SNA, dating back to Guy Orcutt (1957) who started doing macro econometrics, but then eschewed aggregate economic statistics as he came to appreciate the tremendous heterogeneity of economic agents, especially firms and their investment decisions. Subsequently, Richard and Nacy Ruggles wrote extensively not only on the needs to provide explicit microdata foundations for the aggregates in the SNA (e.g. 1975), but also developed methods and supported specific efforts to do so. The idea of more extensive disaggregation of the SNA was given a significant boost by the Stiglitz, Sen, Fitoussi report (2009) wherein they observed that GDP per capita was a poor indicator of economic well-being (indeed suffering from "construct invalidity", a point well established in the late 1940s and early 1950s in the debates about welfare economics, but largely ignored since), preferring instead median (family-size adjusted) family incomes, along with measures of income distribution and inequality.

Still, the SNA culture is prevailing, e.g. at Statistics Canada where a top-down approach has recently been implemented to provide breakdowns for the household sector (Statistics Canada, DHEA). The data elements of the household sector are divided into income quintiles based on a highly detailed microsimulation model (Statistics Canada, SPSD/M). However, this is nowhere near a fully integrated and articulated micro-macro linkage, as produced decades ago by Adler and Wolfson (1988).

The thesis of this paper is that there are more fundamental and important kinds of data to be collected about data flows and accumulations in modern societies from the perspectives of public policy and social proprioception. As a result, we focus here on these raisons d'etre for collection data on data, leaving

SNA considerations, including assigning a monetary value to the accumulation of data, aside and for others to pursue as they may wish.

Data and data landscapes have become large, complex, and intertwined. Still, our focus is generally on how data on data could or should be collected from the perspectives of national statistical organizations (NSOs), and Statistics Canada more specifically.

A Digression on Neoclassical and Heterodox Economics

Before the main part of this paper, which will focus on why and how to collect data on data, it is useful to review the dominant orthodoxy in the field of economics, in which the SNA is situated, and contrast it with various challenges to this orthodoxy. These challenges are clearly fundamental, logically correct, and/or involve much greater realism. However, they have remained largely on the fringes of mainstream economics, which in many ways has more of the characteristics of a religion (e.g. the priesthood of the "math econ", Leijonhufvud, 1973; see also Lipsey, 2001, on the failures of mainstream economics to confront empirical realities).

One personal example is the way Pierro Sraffa's book, "The Production of Commodities by Means of Commodities" (1960) was taught e in a course on linear economics, covering Leontief and Von Neumann models. The story told was there was this unusual practice in Cambridge, England where all the math was worked out on the side, the book was written, and the foolscap with all the math then tossed into the rubbish. The lectures were then an exegesis of this underlying math, completely devoid of context.

In dramatic contrast, in Cambridge, England in the 1970s, Sraffa was worshipped by many fellow research students as a god. His little book had laid waste to the logical foundations of neoclassical production functions, as well as the rationale that profits were the just deserts to owners of capital (K) due to its marginal productivity. It is hard to imagine a more fundamental broadside against the ideological foundations of the neoclassical orthodoxy.

The 1970s were also toward the end of the so-called Cambridge Controversies in Capital Theory, pitting Cambridge England against the leading neoclassical growth theorists in Cambridge, Mass. The key metaphor, rather than gobs of "putty" as the essence of the neo-classical abstraction of an aggregate capital K in economic growth theory, was of a very large book where each page was a blueprint for a given production possibility. Critically, the many pages = blueprints in this massive book, representing alternative combinations of inputs and outputs, (not just one K and one L but fuller column vectors)

could not be uniquely ordered according to wage-profit tradeoffs – the famous re-switching possibilities. The fundamental implication is that the logical foundations for the utility of an aggregate capital K index of capital stock are inherently flawed. In turn, the utility of K, including as it is measured in the SNA, and would be extended by the inclusion of a monetary index of capitalized investments in "data", as an indicator used for any analysis of economic growth and production possibilities, necessarily fails in terms of construct validity.

Cambridge England clearly won the logical battle, but lost the war with Cambridge, Mass (and most other universities in the world). The only acknowledgment of the logical victory was the admission, from some of the leading Cambridge, Mass proponents that neoclassical production functions were really only "parables".[1]

Somewhat similarly, Nelson and Winter's wonderful but generally ignored book on evolutionary economics (1982) showed that a microanalytic approach (indeed a computerized microsimulation model) to growth theory could not only reproduce the simple results of the Solow growth model, but also account for the size distribution of firms, all without an aggregate K.

In the 1980s and 1990s, SNA topics dominated the meetings of the IARIW and discussions regarding the future direction for the SNA. One area was debate about the "core" or central framework of the SNA e.g. whether to add in various imputations or extend the range of satellite accounts. It has long been accepted that imputed rental income on owner-occupied housing should be included in the SNA. But somehow including the imputed value of homemaking services has remained beyond the pale (more precisely, the "production boundary").

With the growing appreciation of R&D as an important factor in production and economic growth, a number of neoclassical theorists began putting a variable into their abstract neo-classical differential equation growth models representing "the stock of R&D capital". Eventually, the framers of the SNA agreed to include expenditures on R&D as a stock, subject to depreciation, rather than a current expense flow. However, with closer examination, the methods for capitalizing R&D in the SNA involve made up numbers and highly arbitrary assumptions.

[1] For example, while admittedly not a perfect indictor of the hegemony of the neoclassical perspective decades after the Cambridge Controversies, a Google Scholar search on "Cobb Douglas Production Function" returned 17,800 entries while a search on "CES production function" returned 18,900 entries, in both cases since 2019, on August 21, 2023.

Another fashion over the years in the IARIW was with satellite accounts. Statistics Canada was concerned in the mid-1980s that the excitement that had engaged many economists across the country in the 1950s and 1960s with the building of the SNA had been lost. As a result, an advisory committee suggested that Statistics Canada should consider investing in satellite accounts, and mentioned specifically one for the health area. However, certainly for health, starting from the SNA was a completely wrong approach. Of course, the costs of health care are a major component of the economy, but the SNA fails completely in measuring "outputs", let alone truly important health "outcomes", nor would it be sensible to try to include health outcomes within the even a satellite account of the SNA. "Health is too important to be a mere satellite orbiting the sun of the SNA" (Wolfson, 1991). Following this think piece, Statistics Canada's health statistics program grew tenfold, though not in the SNA; rather this expansion involved a greatly expanded set of health surveys and administrative data collections – generally intended to provide the empirical foundations for understanding what really matters: health outcomes, population health status and its determinants.

One of the challenges with complex, multivariate, longitudinal data sets – the quintessence of individuals' health and health care trajectories – is that they do not yield their most powerful insights with simplistic methods of aggregation, as in the SNA, nor with "parable only" theoretical growth models. Epidemiology has developed increasingly sophisticated statistical methods in this regard, such as various forms of hazard regression.

But even these "one at a time" statistical regularities are too simple when one is trying to uncover a dynamic "web of causality" (Krieger, 1994) where many factors co-evolve, reciprocally affecting each other over time. A far better conceptual framework for the foundation of health statistics is that of complex systems. The data are collected at the individual level, explored with a variety of statistical methods, and then implications (including policy impact projections) are inferred (albeit always imperfectly, e.g. due to omitted variable bias) by embedding the posited and estimated causal story in a sophisticated microsimulation model, and then running various counter-factual scenarios – e.g. what if we changed from age-based breast cancer screening to regimens where screening intensity was conditioned on genetic risk factors (e.g. Wolfson et al., 2021).

Economics would be better off as a discipline if it eschewed the abstraction of general equilibrium, and instead embraced the more difficult notion of "general interaction" models. Why it has not done so can be ascribed to some combination of the mathematical tractability of equilibria and infinitesimally thin

smooth curves like isoquants[2] along with homogeneous representative agents (and intractability otherwise), lack of training in analysis of (real world) microdata (though this has been improving), entrenched economics faculties' investments in neoclassical theory, and an ideological preference for free markets.

Correspondingly, NSOs would be more relevant and useful to the extent that they gave more effort to the collection and cleaning of a wide range of microdata, not only as the foundations for the sectors in the SNA, but more broadly. In the case of valuing data for inclusion in some sort of monetary terms in the SNA, there are both resource and intellectual opportunity costs, which could be better deployed in other ways to advance societies' appreciation of the transforming roles of the "data revolution".

The experience with health data as well as the fundamental critiques of neoclassical economics noted above form an important backdrop to our discussions of collecting data on data, to which we turn next.

Why Collect Data on Data

There need to be clear motivations for collecting data on data, as it will have considerable costs, and especially if these efforts are to swim against the tide of interest in constructing a monetary value of data for inclusion in the SNA. At the outset we noted two main reasons – to support public policy and to provide social proprioception.

For public policy motivations, one of the top contemporary issues with regard to data is privacy. On the one hand, NSOs have too often been constrained in their access to organizations' detailed internal microdata, less so more recently for government data sets in Canada, e.g. in the areas of tax returns and health care records. Still there are powerful vested interests who fear what sophisticated analyses of patients' health care trajectories might reveal – whether "bad apples" among the physician community or underperforming units in hospitals (Wolfson, 2021). As a result, data custodians unnecessarily use "protecting privacy" as an excuse for not sharing the data – in effect creating a pervasive "privacy chill".

Importantly, following many serious problems with data flows related to the recent pandemic, the Public Health Agency of Canada convened an Expert Advisory Group which made strong recommendations to ameliorate the situation (PHAC EAG), and Health Infoway Inc, a joint federal-provincial-territorial crown corporation shortly thereafter published a "roadmap on interoperability" (Health Infoway, 2023). This

[2] E.g. Harcourt, 1986, page 99 notes that Sraffa in 1930 pointed out that "it may be inadmissible in general to draw a schedule because any actual movement along it may alter its position (and those of other schedules."

roadmap makes repeated references to the need to remove "blockages" to bona fide / public good flows of data, including personally identifiable data, not only for high quality patient care but also for improved health sector management and a range of broader health research, including more cost-effective randomized clinical trials.

Private firms are the sources of detailed microdata for NSOs on a range of characteristics – from surveys of retail sales to employment to R&D to financial statements (though in Canada much of these data now come from various tax returns). However, transaction level microdata from private firms remain difficult for Statistics Canada to access, even though these transaction level data are among the largest flows of data in the world, and of great potential value for key economic indicators like the consumer price index (more on this below).

Nevertheless, concerns about "privacy chill" in the context of statistical and research access to detailed microdata have been tremendously overtaken in the opposite direction by largely unfettered and massive privacy invasions, especially by the largest multinational social media and related firms. There are the beginnings of significant legislative constraints, led by the EU, along with some bipartisan investigations in the US Congress. However, legislation is far behind what's needed for informed consent regarding the sharing of personally identifiable and profitable data with and among private firms.

In Canada, there is considerable support for strengthening the powers of the Office of the Privacy Commissioner (OPC), especially with regard to the practices of the private sector. One avenue for this would be to grant the OPC stronger investigatory powers. With such powers, the OPC could compel a firm to disclose details of the ranges and kinds of data it collects and shares. However, Canada's OPC generally operates on a complaints basis; without a specific complaint, it has no power to investigate data behaviours among private firms.

Thus, from a public policy perspective, there is a major conundrum in the area of privacy. On the one hand, there is far too much "privacy chill" in regard to data flows to support major public goods, including most recently and acutely data on infections, vaccinations, hospitalizations, and compliance with various lock-downs associated with the pandemic. On the other hand, there are extremely serious and growing invasions of privacy via the data collections and individual-level linkages occurring in the private sector, especially in the huge social media firms (or VLOPs = very large online platforms).

A second major public policy area where data are central is population health and health care. Progress in automating data collection and analysis, e.g. in the forms of electronic health or medical record (EHRs and EMRs), has been painfully slow, not least due to pervasive privacy chills and powerful vested

interests. However, the potential benefits in terms of population health and more effective management of health care service provision are tremendous. Canada's constitutional division of powers between the federal and provincial governments remains a major stumbling block.

There are decades of reports and studies outlining the kinds of health data needed to achieve these population health and health care benefits, including the recent EAG and Infoway Roadmap reports cited above. In the 2023 federal budget, over $200 billion was budgeted over the coming decade as fiscal transfers to the provinces for health care, including $500 million earmarked for health data (Canada Budget 2023). In this context, it is fundamental to monitor progress toward the intended health data "infostructure".

From the second main perspective, social proprioception, it is illustrative to focus on two major areas for improved data on data: inflation and entertainment. In both cases, one of the fundamental objectives is to shed light on the extents to which (per the Beatles) "things are getting better all the time". In other words, the objective in these cases is provide the general public insights regarding social progress.

Where is the Data Base (DB)?

Much of the discussion of including "data" more fully in the SNA is expressed in terms of data bases (DBs). The concept appears to be that scattered amongst firms and other organizations there is a discrete set of DBs, each characterized by its size (numbers of records, number of fields per record), and the substantive content of the records. One approach would then be to use some method to place a dollar value on the accumulation year by year of new records and data fields in existing DBs, as well as altogether new DBs, using a perpetual inventory method, and then applying some depreciation rate. This is generally the approach involved for the capitalization of R&D in the current SNA.

However, this is an utterly naïve view of contemporary electronic data – of the ways the massive flows of data are currently organized and are evolving. Consider a purchase via credit card in a retail establishment. Details of the transaction flow to both the vendor's and the purchaser's banks (via credit card intermediaries), both of whom add the transaction data to one of their own DBs. The same transaction data likely also flow (somehow) to the vendor's inventory DB so new items can be ordered when stocks on hand get below some threshold. The same data also flow to the vendor's accounting software, and to the tax authorities for the collection of sales taxes or VAT, two further DBs. On the purchaser's side, the transaction data feed not only her monthly credit card statement, but also possibly

other DBs within the bank to support customer relationships including target marketing of other financial services, and beyond the bank or credit card software to credit rating agencies which combine the individual's purchases from all her credit cards, thereby involving several more DBs.

This story becomes even more involved if the purchase is online, via a firm like Amazon. In this case, the online vendor adds the transaction data to its profile of the individual in terms of her favorite products and other tidbits gleaned from cookies on other web sites to which the vendor has access.

For Statistics Canada at present, these myriad transactions are aggregated and arrive from the Canada Revenue Agency as total VAT and total revenue by firm, albeit via (at least) two different DBs housed at the tax authority.

As a result, a single transaction can almost instantaneously appear in myriad DBs. This wide-ranging and virtually instantaneous diffusion of the data from a single transaction has become ubiquitous as the marginal cost and time required for making electronic copies are close to zero.

Thus, in computer science terms, the world of simple one-off DBs is ancient history. Contemporary DB developments and computer science are concerned with the management of truly enormous real-time transaction data flows (Abadi et al., 2022). Handling these data *flows* is at least as important as data *base* architectures.

In sum, it is much more realistic to refer to data bases and data flows = DBDFs rather than sets of disjoint DBs.

Indicators and DBDFs – The 1995 Atlantic cover page headline, "If the economy is up, why is America down", introduced an alternative to GDP, the Genuine Progress Indicator. (ref http://rprogress.org/publications/1995/1995-10_GPI_Atlantic_Monthly.pdf). This article reinforced and abetted a flowering of studies and estimates of summary indicators proposed as more valid alternatives to GDP (and GDP per capita) for assessing social progress. After several years of international meetings convened by the OECD, however, the consensus was that summary indicators were too constraining, and embodied too many implicit but very strong value judgements required to aggregate the diverse sub-indices forming the overall index. Instead, Sen, Stiglitz Fitoussi (2009) recommended moving away from a single indicator (GDP) to a "dashboard" of indicators. Subsequently, the OECD launched just such a dashboard as the centrepiece of its "Beyond GDP" agenda (https://www.oecd-ilibrary.org/sites/9789264307292-en/1/2/1/index.html?itemId=/content/publication/9789264307292-

en&mimeType=text/html&_csp_=9f1c8dfc1a7bb52555bc12e8b8e03fd2&itemIGO=oecd&itemContentType=book).

Coyle (2021, p151) appears rather dismissive of this idea of statistical dashboards, "there is no solid theoretical structure commanding wide consensus… (Note also that dashboards imply drivers…)". However, neither is there a "solid theoretical structure" underlying the many arbitrary decisions embodied, for example, in the definition of the production boundary in the SNA, hence what is included or excluded from GDP, notwithstanding its wide consensus, nor the depreciation rate for capitalized R&D. Nor is anyone likely to fly in an airplane where the only dial in the cockpit is an aggregate index of airspeed, altitude, and fuel remaining.

"Valid" aggregation of sub-indices into some overall aggregate should be based on "principled weights".[3] For (conventional period) life expectancy, mortality rates at different ages are effectively weighted by the steady-state population counts by age; for the CPI, the weights are based on average expenditures by commodity; and for GDP the weights are the money values of market transactions. However, for summary indicators like the GPI, there really are no principled weights.

Still, even without principled weights, imperfect indicators may still be useful. For example, waste water monitoring for COVID virus levels gives only an approximate indication of the prevalence of the disease, but in the absence of more systematic testing of individuals, this indicator is much better than nothing. Similarly, price indices are imperfect indicators of changes in the cost of living, not least because their theoretical foundations are premised on patently unrealistic assumptions. As shown in Wolfson (1999), once account is taken of new goods, increasing returns to scale, disequilibrium trading, income inequalities, and satisficing rather than omniscient utility maximization, price deflators can even go in the wrong direction. Nevertheless, better official statistics on patterns of price and expenditure changes can be "fit for purpose" (see below).

Further, well-meaning individuals too often seek statistical indicators without appreciating the requisite but underlying detailed and expensive data collections required– what I've long called the malaise of "indicatoritis". An obvious example is life expectancy, clearly a fundamental indicator. While the concept is relatively straightforward, constructing a high-quality version of this indicator requires hundreds of millions or billions of dollars for a population census and a vital statistics program that includes complete death registration. But no country would invest in a census or vital statistics program for the sole reason

[3] A phrase used by Dan Usher, a professor at Queens University.

of producing the life expectancy indicator. These two kinds of data collections each serve a multitude of statistical and informational objectives, as well as other areas of administration and public policy. Further, given their microdata foundations, they enable analyses to "drill down" beneath any indicators or substantially aggregated published statistical tables to explore more fully underlying patterns and relationships.

Analogously, a statistical DBDF portrait should be considered as a general purpose statistical activity, designed to meet a wide variety of data, administrative, and policy needs –well beyond the objective of valuing data to form an aggregate sub-index within the framework of the SNA.

What Should NSOs do with DBDFs?

A major challenge, in this context of dynamic, complex, and rapidly expanding DBDFs, is what specific roles NSOs should play. In the following, we consider four areas: the two policy areas of privacy and health, and the two social proprioception areas of inflation and entertainment.

Privacy – Suppose Canada's Office of the Privacy Commissioner (OPC) is granted stronger legislative powers proactively to investigate and act / regulate potential or emerging privacy issues related to DBDFs. Such powers could be analogous to those of the tax authorities who, based on their inventories of tax returns, deploy various algorithms to analyze details of these returns and then select a sample of taxpayers' returns for detailed audit. The essential prerequisite for the tax authority is the inventories of tax returns. Analogously, the OPC would need an inventory of DBDFs.

As such an inventory of DBDFs would have many uses beyond supporting the (potentially expanded) privacy mandate of the OPC, it would be far more efficient, to complement any increased powers for the OPC, for Statistics Canada (and other countries' NSOs more generally) to build and maintain an evergreen (and likely rapidly growing) "portrait" of DBDFs in Canada. Indeed, this portrait, essentially a DB of DBDFs, would form the keystone for much of what is needed for an effective and comprehensive program of collecting data on data, and meeting the specific policy and social proprioception objectives which are the focus.

Statistics Canada already has a very broad sample frame as a starting point: essentially any organization in Canada that pays sales tax or pays employees (hence administers income tax source withholding) or has individual or corporate income must file at least one kind of tax return at least annually. These tax

data subsequently flow routinely to Statistics Canada where they are used to construct and maintain the "business register" (n.b. including public sector and non-profit entities as well as private firms), hence providing a near universal sample frame of organizational entities.

In turn, this sample frame is used to elicit data using a variety of focused surveys, ranging from retail trade to R&D. In principle, therefore, it would be possible for Statistics Canada to create a new "DBDF portrait survey" asking these entities to provide basic data on all their DBDFs.

Of course, there are important complexities in designing and implementing such a survey, including:

- providing workable definitions of a DB and a DF,
- having an adequate profile of the entity being surveyed to ensure that the survey itself is sent to an individual within the firm or organization with the knowledge to complete the survey, and
- ensuring that all data flows into and from the entity include adequate pointers to all the other entities party to the data flows.

A further major challenge is international entities that may have no "footprint" in Canada. With the internet, it is easy and very common to be able to interact with foreign entities, e.g. google or google maps searches, where the web site is collecting data on the individual, but has no formal presence in Canada. A particularly invasive and egregious example is Google's timeline (https://www.compunet.ca/blog/google-timeline-the-good-the-bad-and-the-ugly/). Without any (obvious) permissions, and without any explicit link between a smartphone and an individual's laptop, this timeline can by default display on her laptop all the geographic locations where she has stopped.

In cases like this, strong federal legislation will likely be required to compel such international organizations to provide data on their DBDFs insofar as they involve Canadian residents. In the first instance, the reason would be to support any strengthened mandate for the OPC.

It will also be important for any such legislation to be clear regarding the respective roles and mandates of the OPC vis a vis Statistics Canada, as the data on DBDFs thereby generated would play a foundational role for official statistics. In particular, such an evergreen DBDF portrait could serve as a sample frame for a range of more focused data programs, including those described next.

Health – Canada has the potential to be a world leader in managing in the most cost-effective manner its health care sector, in health research, and in rapidly responding to unforeseen events like the recent pandemic. The simple reason is that each province in effect is a single-payer for a wide range of health

care services, so in principle it could manage these services by creating a fully integrated patient-level DBDF. Further, from a pan-Canadian perspective, if these provincial DBDFs used standardized concepts and definitions and were interoperable across provincial boundaries, Canada could rival the likes of the UK's NHS in terms of providing a population-based laboratory for clinical research including more cost-effective randomized clinical trials, health technology assessment, growing appreciation of the power of "real world evidence", and linkages to major population health and related surveys (like the UK Biobank ref). Unfortunately, this potential is far from being realized. A key reason is the many blockages to the appropriate flows of health and health-related data.

As emphasized in the EAG report (PHAC EAG), there have been decades of reports and studies outlining what is needed in the area of health data. The challenge is overcoming the privacy chill and vested interest blockages (Wolfson, 2018). A recent Infoway survey paints a gloomy picture of the ability of patients even to access their own already existing electronic health data (https://www.cihi.ca/en/taking-the-pulse-a-snapshot-of-canadian-health-care-2023/better-access-to-electronic-health and https://insights.infoway-inforoute.ca/data_table_2022).

Since 2005, the Government of Canada has invested billions of dollars in Health Infoway Inc., a crown corporation. Much of this funding was directed toward incentives for provincial governments to prevent each from reinventing the wheel, i.e. rather than each developing (more accurately paying a private vendor for) their own software systems, to share the costs of developing a modern EMR / EHR and related software. Unfortunately, these incentives have largely failed to achieve the objectives of standardized interoperable patient level DBDFs. They have also failed to achieve the nationwide monopsony purchasing objective of lowered software costs from private vendors, who instead have powerful incentives to lock in each province or health authority to their unique software product by inhibiting interoperability.

Still, over the almost two decades since its creation, Infoway's staff have developed a very good understanding of the current landscape of provincial health-related DBDFs, and have produced very detailed architectures (ref). Most recently, Infoway has been charged with developing and leading a "Roadmap" on interoperability (Health Infoway, 2023) which in effect constitutes a DBDF portrait in the area of health care. It is not only assembling this portrait, but also endeavoring to ensure that, e.g. a diagnosis of diabetes or a third line chemo treatment for cancer or the make, model and software version of an MRI machine (say) is coded in a common standard (if not identically) in all the places where such a data field exists.

This is a massive and long overdue undertaking. But individual patients' lives depend on the interoperability of these data, as does cost-effective management of health care services. Unfortunately, though, there is no publicly accessible portrait of Canada's health and health-related DBDFs. Furthermore, there are many entities whose DBDFs are not in scope for Infoway, but are relevant for health policy and understanding the drivers of population health.

For Infoway, the foci include hospital and physician encounters, lab tests, diagnostic imaging, vaccinations, and prescription drugs. However, there are many critical kinds of data outside Infoway's scope, including characteristics of the health human resources involved (physicians, nurses, personal care workers – their training, work patterns), vital statistics (e.g. causes of death), over-the-counter drugs, home care and nursing home ownership, operations, staffing patterns, etc., and the kinds of data needed to place health care within the context of the broader social determinants of health.

Further, there are already many players in the health data area beyond Health Infoway and Statistics Canada, including the Public Health Agency of Canada (a federal department), the Canadian Institute for Health Information (ref, https://www.cihi.ca/en), Canada's Drug and Health Technology Agency (ref https://www.cadth.ca/about-cadth) and some provincial counterparts, various provincial government agencies including health ministries themselves, workers compensation boards, and health quality councils (ref e.g. https://www.saskhealthquality.ca/), academic health research organizations (e.g. https://www.ices.on.ca/, https://www.bornontario.ca/en/data/data.aspx, and https://www.popdata.bc.ca/).

Important DBDFs are also held by private sector firms, from pharmacies to lab testing firms to primary care physicians' businesses typically structured as private corporations, to large insurance companies.

Comprising about one-tenth of Canada's economy, it is not surprising that there are myriad entities holding health and health-related DBDFs. Based only on the ad hoc and fragmentary information available, it is clear that these DBDFs are largely uncoordinated, unstandardized, not interoperable from an individual patient's perspective, and more often than not useless for contemporary kinds of probing statistical analyses, which involve large highly multivariate samples of individuals' data.

Having regularly updated and readily accessible data on the state of health DBDFs would provide the general public as well as journalists, policy analysts and decision-makers, an essential moving snapshot of where the most serious gaps in functioning health-related DBDFs were. While it is unlikely to be decisive, such accessible information would further aid in forcing some accountability on the actors whose support and effort are needed to achieve the desired state of health DBDFs in Canada.

Inflation – The economies of many countries in 2023 are suffering from both inflation and the impacts of increased interest rates as monetary policy is deployed in an effort to reign in the inflation. Importantly, there are fairly frequent reports in the popular media where individuals are claiming they are facing much higher inflation than is being reported in the official statistics. There are also longer standing concerns in Canada that the official consumer price index (CPI) does not reflect the inflation faced by particular population groups including the poor and the elderly. While detailed studies have not supported this particular claim (Stat Can, unpublished), it remains an open question how much heterogeneity would be found in inflation rates across individuals and households with varying patterns of expenditures. Hence, NSOs could regularly publish inflation data disaggregated not only by commodity and geographic region, but also by socio-economic group. Meeting these needs, especially the latter, requires significantly better data on expenditure patterns.

At the practical level of implementing the CPI, Statistics Canada is facing growing difficulties collecting data from the Survey of Household Spending (Statistics Canada, SHS) which is used to determine these expenditure patterns, i.e. the basket of goods and services = principled weights underlying the CPI.

Another major concern with price indices, including the CPI, is the role of "new goods". The US Senate-appointed Boskin Commission (1996) argued that the US CPI was over-stated by about one percentage point, where half of this overstatement was attributable to the failure to account properly for the appearance of new goods, such as digital cameras, cell phones, and new drugs. (Streaming music had not yet become widely available.) This new goods problem arises because the volume of sales of an item in question only becomes large enough for it to be included in the price index's basket of goods and services well after the largest declines in its price have already occurred, hence the inflation rate is arguably over-stated.

There is also a widespread recognition that there are ongoing major quality improvements, initially most notably in computers, but also in cars, household appliances, and streaming video services. As a result, NSOs have deployed hedonic regression methods to adjust some commodities' valuations in price index construction to take into account such quality changes. But due to its practical difficulties, hedonic adjustments for quality changes are applied only for a few commodities. As a result, price indices, including the CPI, are missing much of the improvements in quality actually occurring.

Most recently, there has been a dramatic growth in "free" goods, such as online search and videos. These are completely missing from the CPI.

Given all these factors, and as noted earlier, it can reasonably be argued that the official CPI may be seriously biased, but in ways that are presently unknowable. Further, it is unknown the extent to which inflation, measured taking account of the biases just noted (and to the extent feasible), has important distributional consequences, for example varying systematically across different socio-economic groups.

In this context, a fresh start is warranted in conceptualizing and then measuring statistically households' "progress" in terms of consumption, conceptualized more broadly than simply price inflation.

A critical first step in this reconceptualization is incorporating time use patterns. There has been a growth in the deployment of time use surveys by NSOs, most notably in the US (US BLS). While Statistics Canada was an early leader in fielding time use surveys, it has not moved beyond a quinquennial focus in its General Social Survey. But time use patterns are critical for obtaining data on the consumption of "free" goods on the internet. These surveys can also provide the basis for moving from *expenditures*, such as on consumer durables like household appliances to their use as *consumption*. Time use patterns are also essential for understanding consumption of entertainment such as radio, TV, and recorded music, especially to the extent the surveys ask about activities where consumption of "entertainment" is joint with other activities like household chores and childcare. These time use patterns can be combined with questions on the subjective valuations respondents attach to the various activities.

Another critical step is broadening the data flows used to construct the consumption basket, especially given the declining response rates to the household surveys that have provided this basis for many decades. With the dramatic growth of electronic rather than cash payments for goods and services, as well as the use of bar coding for differentiating commodities, there already exist myriad DBDFs with potentially useful data – specifically data on expenditures that are more fine-grained in terms of commodity detail, and are linkable to individuals' and households' socio-economic status.

Statistics Canada has the legislative authority to collect such data from banks and retailers, but it does not yet have the "social license" to do so, as revealed in a recent controversy (Wolfson, 2022). In this case, a more measured and gradual approach would be more likely to succeed. It would start with the construction of the "portrait" of DBDFs already discussed. Next, there could be an exploratory pilot study with a very small sample of individual records to ascertain not only the levels of detail available from various kinds of electronic transactions (i.e. credit cards, point of sale bar codes), but also more information on the kinds of software and DBDF architectures the various entities were using to handle and store these data.

It would also be critical for the NSO to have their staff engage personally with the relevant decision-makers in these entities to understand both their sensitivities regarding the disclosure of these very detailed data to the NSO, and also the kinds of response burdens collecting a sample of these data from various types and sizes of organizational entities would impose.

Entertainment – There is no question that there has been an explosion in the availability and consumption of a range of kinds of entertainment, albeit all essentially electronic. These include recorded music, streaming videos, sharing photos with friends, sharing hobby interests with individuals around the world (e.g. in Facebook groups), and computer gaming. As recently as a few decades ago, the idea of a "500 channel universe" was still a dream. Today, we are well beyond 500 channels.

Much of this consumption is "free", without any monetary payments. Much else has essentially zero marginal monetary cost once a subscription has been paid. At the same time, as it is electronic, it all now involves the flows of digital data, and is often coupled with data collection on the viewing or usage patterns of each user.

From the context of social proprioception, and understanding societies' progress, any statistical series based only on monetized market transactions is bound to be seriously biased, most likely understating actual progress. NSOs should be endeavoring to provide their societies valid and engaging statistical information on how these major aspects of our lives are changing.

The "portrait" of DBDFs already described, along with time use surveys just mentioned, provide the foundations for such a new statistical program. The content of such a program will be sufficiently diverse that a family (dashboard) of statistical indicators would be needed, along with "drill down" access to the underlying microdata for more in depth analyses.

As a thought experiment, we can imagine the table of contents for the first publication from this new statistical program on electronic entertainment. At the highest level, it could divide the activities into sectors or domains, analogous to standard industrial classifications, e.g. music, videos (both longer like movies, sports events, and TV shows, and shorter like TikTok), computer games (both solo and multi-player), hobbies, and "friends" (conversing, sharing photos).

In each of these domains, among the key statistics would be how much time individuals were spending engaged in the activity, when during the day or week the activity most often occurred, whether or not it involved real-time interaction with other individuals, and how it was paid for. Further, all of these data

elements would be disaggregated by users' various socio-economic characteristics, not least age, sex, educational attainment, household income group, and geography. As importantly, the trends over time would (eventually) be provided. It is most likely that such a statistical publication would generate considerable headline news.

Beyond its value in terms of social proprioception, other features of the underlying data would be important for various areas of public policy. For example, there are the privacy implications of the data on viewers and game players themselves being collected by the vendors of these electronic entertainment services, the possible implications of corporate concentration of these vendors for competition policy, and in Canada the longstanding policies involved in encouraging Canadian cultural content.

Concluding Thoughts

Even though the metaphor that "data is the new oil" is somewhat strained, there is no question that data bases and data flows have not only grown dramatically, but are also reflected in major changes in the ways we spend our time and money, hence the economy, and the ways we interact socially. As a result, it should be incumbent on NSOs to adapt their statistical programs to encompass and reflect these new realities.

One option is to extend the SNA to incorporate a monetary valuation of "data", as a form of (intangible) capital stock. While the Cambridge controversies in capital theory are largely ignored or unknown in economics at present, the logic is correct, so an aggregate capital K index is fundamentally flawed; it lacks construct validity. It can serve as the basis for parables, but for official statistics if cannot be trusted to tell an unbiased story of economic growth, productivity, or other stories of social progress. It is far more useful, valid, and practical to build such stories using data collections that are more disaggregated, that directly pertain to real phenomena, that do not embody patently unrealistic or arbitrary assumptions, and that do reflect myriad real world heterogeneities.

Yes, some aggregate indicators can be valid, like life expectancy which has obvious "principled weights" for combining age-specific mortality rates. But it is far more useful for an indicator like life expectancy to reside at the top of a coherent system of statistics, with "drill down" capacity to disaggregate by age, cause of death, socio-economic status, geography, and other key covariates. Further, these underlying data should support modern kinds of statistical inference, such as multivariate hazard regressions and

microsimulation modeling, in order to provide insights on the factors affecting (in this case) life expectancy.[4]

Analogously, we have proposed that at the centre of NSOs' adaptation to the dramatic growth of "data", they should focus on the micro foundations – collecting data not only on discrete data bases (DBs), but on data bases and their associated data flows (DBDFs). The core should be an evergreen micro statistical "portrait" of the country's DBDFs. In essence, this portrait would be a census of individual DBs plus a census of all the DFs including both the substance of the data elements flowing and the pointers indicating the source and destination DBs for these data flows.

The reasons to build and maintain the DBDF portrait include both major policy areas such as privacy and health, and key areas of social proprioception – areas where there is general interest in understanding how society is evolving. In this paper, two such areas have been discussed: inflation and entertainment.

Further, to provide essential context, the DBDF portrait should be complemented by more extensive and coordinated statistical data on time use patterns, hence time use surveys of adequate frequency, with sufficient detail, including content on the satisfaction derived from various activities, and using concepts and definitions concorded with the DBDF portrait.

This kind of statistical program will provide the foundations for many derivative analyses and areas for further statistical developments. (These could include capitalized valuations of DBDFs in SNA terms, but this would not be a top priority.)

References

Abadi et al., The Seattle Report on Database Research, Communications of the Association of Computing Machinery, August 2022, Vol 65 No. 8 . https://cacm.acm.org/magazines/2022/8/262905-the-seattle-report-on-database-research/fulltext; also https://db.cs.washington.edu/events/other/2018/Seattle_DBResearch_Report-Full.pdf

Adler HJ, Wolfson M. A Prototype Micro-Macro Link for the Canadian Household Sector. Review of Income and Wealth. 1988 Dec;34(4):371-92.

Atlantic Monthly, October 1995 http://rprogress.org/publications/1995/1995-10_GPI_Atlantic_Monthly.pdf

[4] Similarly, instead of aggregate measures of productivity, it would be far more valuable to make more extensive use of and to expand Statistics Canada's marvelous longitudinal microdata on firms, e.g. to observe births, deaths, mergers and amalgamations, divestitures, and growth, firm by firm, associated with a range of covariates – including the dynamics of firms in relation to their production possibility frontiers. Of course, such analyses are more difficult and time-consuming, typically requiring more careful data preparation, hence are less attractive to researchers given the publish or perish competition in academia at present.

Boskin Commission Report, 1966, US Senate https://www.finance.senate.gov/imo/media/doc/Prt104-72.pdf

Canada Budget 2023, Investing in Public Health Care and Affordable Dental Care https://www.budget.canada.ca/2023/report-rapport/chap2-en.html#a2 (accessed October 12,2023)

CIHI report on Infoway survey https://www.cihi.ca/en/taking-the-pulse-a-snapshot-of-canadian-health-care-2023/better-access-to-electronic-health

Coyle D. Cogs and Monsters: What economics is, and what it should be. Princeton University Press; 2021 Dec 31.

Google time line https://www.compunet.ca/blog/google-timeline-the-good-the-bad-and-the-ugly/

Harcourt GC. On the influence of Piero Sraffa on the contributions of Joan Robinson to economic theory. The Economic Journal. 1986 Dec 1;96(Supplement):96-108.

Infoway survey https://insights.infoway-inforoute.ca/data_table_2022

Infoway: Canada Health Infoway, "Shared Pan-Canadian Interoperability Roadmap", May 2023 https://www.infoway-inforoute.ca/en/component/edocman/6444-connecting-you-to-modern-health-care-shared-pan-canadian-interoperability-roadmap/view-document?Itemid=101

Infoway architecture …

Krieger N. Epidemiology and the web of causation: has anyone seen the spider?. Social science & medicine. 1994 Oct 1;39(7):887-903.

Leijonhufvud A. Life among the Econ. Economic Inquiry. 1973 Sep;11(3):327-37.

Lipsey RG. Successes and failures in the transformation of economics. Journal of Economic Methodology. 2001 Jan 1;8(2):169-201.

Nelson R, Winter S. An Evolutionary Theory of Economic Change, Harvard Unv. Press, USA. 1982.

OECD https://www.oecd-ilibrary.org/sites/9789264307292-en/1/2/1/index.html?itemId=/content/publication/9789264307292-en&mimeType=text/html&_csp_=9f1c8dfc1a7bb52555bc12e8b8e03fd2&itemIGO=oecd&itemContentType=book

Orcutt GH. A new type of socio-economic system. The review of economics and statistics. 1957 May 1;39(2):116-23

Public health Agency of Canada (PHAC EAG), " The pan-Canadian Health Data Strategy: Expert Advisory Group Reports and summaries" https://www.canada.ca/en/public-health/corporate/mandate/about-agency/external-advisory-bodies/list/pan-canadian-health-data-strategy-reports-summaries.html (accessed October 12, 2023)

Ruggles R, Ruggles ND. The role of microdata in the national economic and social accounts. Review of Income and Wealth. 1975 Jun;21(2):203-16.

Sraffa P. Production of commodities by means of commodities. Cambridge: Cambridge University Press; 1960.

Statistics Canada, DHEA = Distribution of Household Economic Accounts https://www150.statcan.gc.ca/n1/pub/13-607-x/2016001/938-eng.htm (accessed October 12, 2023)

Statistics Canada, SPSD/M =Social Policy Simulation Database and Model https://www150.statcan.gc.ca/n1/en/catalogue/89F0002X (accessed October 12, 2023)

Statistics Canada, SHS …

Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2009). Report by the commission on the measurement of economic performance and social progress.

UK Biobank …

US BLS, American Time Use Survey https://www.bls.gov/tus/ (accessed October 12,2023)

Wolfson MC. A system of health statistics: toward a new conceptual framework for integrating health data. Review of Income and Wealth. 1991 Mar;37(1):81-104.

Wolfson MC. New goods and the measurement of real economic growth. The Canadian Journal of Economics/Revue canadienne d'Economique. 1999 Apr 1;32(2):447-70.

Wolfson MC "What's preventing Canada from creating a robust health data infrastructure", 2021, IRPP https://policyoptions.irpp.org/magazines/may-2021/whats-preventing-canada-from-creating-a-robust-health-data-infrastructure/.

Wolfson M, Gribble S, Pashayan N, Easton DF, Antoniou AC, Lee A, van Katwyk S, Simard J. Potential of polygenic risk scores for improving population estimates of women's breast cancer genetic risks. Genetics in Medicine. 2021 Nov;23(11):2114-21.

Wolfson M (2018), https://www.theglobeandmail.com/opinion/article-why-statscan-should-have-access-to-our-banking-data/