This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

Making "Drill Down" Analysis of the Economy a Reality

By

John Haltiwanger*

University of Maryland

Abstract: To address core questions in social science, analysts need to be able to "drill down" from key aggregates (as examples, economic indicators like productivity, job creation, and unemployment) to the person and business level with rich contextual information about person and business characteristics. Such drill down capability need to be accessible to the social science research and policy communities without jeopardizing the privacy and confidentiality of person and business level data. Much progress has already been made but there are significant bureaucratic, legal, and methodological challenges to achieve the vision of drill down analysis.

* This work is licensed under the Creative Commons Attribution-NoDerivs 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

The vision

Here is the vision. A social scientist or policy analyst is investigating the impact of the "great" recession and anemic recovery on businesses and workers. The analyst begins by exploring the latest aggregate data showing economy-wide, sector-level and broad regional-level variation in terms of business productivity, output, capital investment, prices, wages, employment, unemployment and population. The data on employment changes can be decomposed into hiring, quits, layoffs, job postings, job creation and destruction. The data on unemployment can be decomposed into gross worker flows tracking flows into and out of unemployment. The data on workers is linked to measures from household data tracking income, consumption, wealth, consumer finances and household composition. The data are high frequency and timely. The data are available not only for the present time period but historically for several decades permitting analysis of both secular trends and cyclical variation.

Starting at the economy-wide level, the analyst can drill down into the various key indicators by detailed worker and firm characteristics such as gender, age, immigration status, and education of workers and business size and business age for firms. The business data permits identifying firm startups and also permits tracking firm exits. The firm characteristics include measures of intangible capital assets such as investments such as R&D and innovation as well as tracking foreign trade and outsourcing activity. The sources of financing for businesses by type of financing and by type of business are available. The analyst can conduct empirical studies at the economy-wide, broad sectoral and broad regional level with data broken down by all of these dimensions. In addition, the analyst can drill down to the individual and firm level creating a longitudinal matched employer-employee data set with all of this information at the micro level. This permits panel data analysis using rich cross sectional and time variation data tracking the outcomes of businesses, workers and households. These outcomes can be tracked at the very detailed location and characteristics level. The drilled down data aggregates to the national key indicators that receive so much attention.

The analyst can ascertain, for example, is it really the case that it is small, young businesses that normally would be creating jobs given their productivity and profitability who can't get credit that is accounting for the anemic recovery. The analyst could track what type of financing has especially decreased relative to other economic recoveries. The analyst could analyze the impact of policy interventions historically and how they have or have not had influence on different types of businesses and in turn on the workers employed by these businesses.

Beyond analysis of the recession and recovery, the drill down infrastructure would permit a range of analyses of the factors driving economic growth and other important economic outcomes for households and businesses. The role of business startups can be tracked in terms of their contribution to productivity, innovation, and job growth. The origins of business startups can be tracked given the longitudinal matched employer-employee data – e.g., what is the career path of entrepreneurs and the factors that impact career path as well as success and failure of

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

business startups? The drill down data infrastructure also permits rich studies of the outcomes of households and workers. The impact of immigrant flows on native and immigrant labor market outcomes can be tracked and studied. The drill down data infrastructure enables tracking the education experiences and outcomes of individuals so the factors impacting human capital investment can be studied. In principle, the drill down data infrastructure also is integrated with a variety of other measures of experiences – the housing experiences of children and adults, the health experiences of children and adults, the criminal activity as well as the victims of crimes while children and adults. With this added dimensions, a wide range of socio-economic issues can be investigated.

The Reality as of 2010

The above vision is not as far from reality in 2010 as one might first surmise but achieving the above vision faces many different challenges. The good news is that progress has already been made on many of these challenges. The bad news is that many very difficult challenges remain that may take decades to overcome.

First, the good news. The U.S. federal statistical agencies, state agencies, and private sector data developers have been working on various components of this vision for the last decade or more. The rapid increases in computer speed and disk storage has meant that the ability to track every business, household and worker in the U.S. on many of these outcomes is already a reality. State agencies are likewise developing longitudinal databases tracking education experiences and outcomes and in turn to labor market outcomes for their workers. Other major efforts developing the longitudinal micro data on households and businesses include the efforts tracking health and other outcomes for older Americans as well as efforts by the private sector in tracking U.S. businesses.

However, in spite of enormous progress, the bad news is that the U.S. statistical system remains incredibly balkanized. Legal issues still block the major U.S. federal agencies from sharing their data (hence the duplication between BLS and Census – which not only leads to unnecessary duplication but also to significant limitations in the quality of key economic indicators like U.S. productivity and GDP. Legal issues block the federal and state agencies from working collaboratively. Beyond the legal issues, many technical challenges for the cyber infrastructure needed for a drill down data infrastructure remain. We turn to discussing such challenges in the next section.

The Challenges

The challenges are many. At a broad level, the challenge is that attaining this vision requires integration of a vast array of administrative and survey data from a variety of sources with different objectives and legal requirements for using and protecting the confidentiality of the data. There are many, many details – a non-exhaustive list is as follows:

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

1. Attaining this vision requires a change in the way data are collected and processed. The statistical system needs to be smart in terms of using scarce resources to avoid duplication but also to collect the many different components in a fashion so they can be integrated. Even within statistical agencies, there is a "silo" approach towards data collection. Of course, the problem is much worse across federal and state agencies. The efforts discussed above to integrate administrative and survey data have been severely hampered by the disparate nature of the data. The development of standards and the agreed upon use of common identifiers would greatly facilitate data integration. Of course, the use of common identifiers raises many privacy concerns which we discuss below.

2. Overcoming the legal challenges is a significant obstacle in its own right. This vision can likely only be achieved with the intensive use of administrative data that are collected for other purposes. There needs to be a mandate that administrative data from all of the types of sources discussed above can be used for this type of statistical analysis.

3. Privacy advocates rightfully express concerns that the above vision creates "big brother". There are many issues in dealing with privacy concerns which I will not deal with adequately here. One critical issue is insuring a secure environment for developing and accessing this type of data infrastructure. One working model so far is to create secure enclaves (the Census-NSF research data centers). Secure enclaves have shown that they can provide access to many valuable projects from the research community without endangering the privacy and confidentiality of respondents. While this represents great progress relative to 20 years ago, this likely is not the long run solution. Still even if this is not the long run solution, if over the next decade or so the Census-NSF data centers can become Federal-State data center enclaves where a host of federal and state data can be housed this would represent major progress. This has already happened to some degree with agencies such as NCHS now housing their data at the Census RDCs.

4. The long run solution is to create a cyber-infrastructure environment that creates the data infrastructure, is securely protected from hostile attacks as well as inadvertent disclosure, is accessible to the user community from many locations and generates statistically valid inferences from the above data infrastructure without the user ever being able to observe enough details to identify any business or person. This requires real-time disclosure protection on an interactive basis.

5. Further advances in this drill down data infrastructure will require further advances in both hardware and software. Vast amounts of data will need to be stored and processed in a timely fashion. In addition, data integration methods need further development and refinement. Currently, the data matching programs are capable of developing data matches using a variety of criteria with measures of the quality of matches. But we often need more than this – we need an ability to deal with differences in units of observation

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

and frequency. We need smart systems that can integrate weekly, monthly, quarterly and annual data readily. The smart systems need to be able to take data using a variety of different units of observation and integrate. The smart systems need to use the insights and continuing developments from statisticians on missing data imputations with statistical software packages that can adjust standard errors appropriately based upon sampling variation as well as associated imputation. Smart systems are also needed for the collection of data. Many administrative data sources are in principle available on almost a real time basis (often with filings in the last week, month or quarter) but only become available to the statistical system with a considerable lag.

6. The private sector is already in many ways further along in attaining this vision for profit-making purposes using data mining techniques. This competition with the public sector in the provision of statistics is not necessarily a bad thing. However, the ability of the scientific community to develop the methodology and standards for data integration and associated analysis is limited for most private sector developments. Of related concern, many in the social science community are already beginning to conduct studies using these private sector data sources. Again, this is not a bad thing per se but often the representativeness and statistical properties are not known and the access is not regulated in a manner that permits replication – a critical feature in the advancement of science. A very bad outcome would be that in a few decades the private sector developments have superseded the public sector developments so much that social scientists are essentially forced to use private sector data without adequate quality and methodological standards. The private sector data mining developments will not go away and nor should they go away. Finding a way to create public-private partnerships and to help set standards for methodology, access and privacy protection that have some commonality in the public and private sector developments would be valuable.