James McAndrews: Challenges and rewards associated with data in the New York Fed

Remarks by Mr James McAndrews, Executive Vice President and Director of Research of the Federal Reserve Bank of New York, at the Fifth Data Management Strategies and Technologies Workshop, Federal Reserve Bank of New York, New York City, 4 February 2014.

* * *

Introduction

Welcome to the Federal Reserve Bank of New York. We are fortunate to have many central bankers and international nongovernment organizations represented here today. Because we are part of the Federal Reserve System, the central bank of the United States, we share many of the same data obstacles you are currently facing. As the head of Research and Statistics at the New York Fed, I'd like to share with you the challenges and rewards associated with data in our organization, particularly the data that we use to analyze policy questions.

The New York Fed is uniquely positioned for several reasons. In addition to being the largest of the twelve Federal Reserve Banks, it has the distinction of including a Markets Group that conducts Open Market Operations for the Federal Reserve System. Additionally, the New York Fed supervises financial institutions, including foreign banks doing business in our district. These special responsibilities contribute to our tremendous reliance on data.

Data is crucial to our core objectives and all aspects of our work.

Data is the foundation for developing international standards in bank capital and liquidity requirements. Data is an integral part of researching economic issues and the labor and financial markets. Data is also central to our jobs of forecasting and identifying trends. Our understanding of interest rates, reserves, and balance sheets is built on data. In our efforts to promote financial stability, data is fundamental to stress testing. In our financial services efforts, payment information and data flows figure critically in our payment systems. The scope of our data consumption and creation spans a wide range. It encompasses local, regional, national, and global coverage and extends from the micro to the most broadly aggregated levels.

Consequently, as an organization we are increasingly dealing with a growing number of data challenges. They include the multitude of legal topics associated with contractual terms and conditions of data use, data sharing, and privacy restrictions. We are constantly challenged to develop and maintain clear and comprehensive data dictionaries that enable us to repurpose data appropriately. Moreover, as we continue to create more data, we are focusing on data governance. This entails oversight and measures to ensure that our own data are accurately posted (with updated revisions and associated documentation) by commercial aggregators. All of this growth requires investments in both technology and personnel to support our ever-expanding data requirements.

Challenges in the use of data

I'll focus my remarks on two important challenges that we face in using data effectively, as well as some models we may draw from to overcome those challenges.

The first challenge is what we might think of as *data innovation*, and the second is *data stewardship*.

With data innovation, the big challenge is to use our imagination and to think hard about what data may be able to say about a particular problem. This includes both novel uses of existing

data as well as the collection of new data. To illustrate what I mean by this, let me describe one effort, among many, here at the New York Fed.

Consumer Credit Panel

I'll describe our ***Consumer Credit Panel,*** which we created in 2010. At the onset of the financial crisis, it became apparent that our knowledge of the liability side of household balance sheets was inadequate. At that time, the main data sources on consumer debt consisted of loan-level data sets on specific categories of loans, such as mortgages, as well as aggregated data on household sector debt from the Board of Governors' Flow of Funds statistical release. While providing a useful aggregate overview of household debt, the Flow of Funds release defines the household sector as a residual sector that also includes nonprofit organizations, such as schools, hospitals and churches, as well as hedge funds and private equity funds. Furthermore, the overview it provides is limited to outstanding balances and lacks important information on the origination, repayment, and delinquency status of household debt. And both loan-level data and aggregates like the Flow of Funds miss an important dimension of borrower behavior that became very important during the crisis and recovery – the relationship between different kinds of credit use by individual borrowers or households.

The creation of the Consumer Credit Panel is an example of the use of pre-existing commercial data in an innovative way that had not previously been pursued. The panel is based on credit report data collected by Equifax (one of the three credit bureaus in the United States) and it contains information on all outstanding loans – including mortgages, auto and student loans, and credit card debt – at the individual consumer level. While credit records are primarily used by lenders to evaluate a potential borrower's creditworthiness and ability to repay, they can also provide a comprehensive picture of outstanding balances and delinquencies and how they interact. Of course, all the data are completely anonymous – the dataset has been stripped of all personally identifiable information.

The CCP was created by our research group using an innovative sampling procedure, based on the randomness of the last four digits of social security numbers. The resulting panel allows one to track individuals over time and across different locations within the nation, starting in the quarter in which they establish their credit report when taking out their first loan. The sampling approach generates the same entry and exit behavior as present in the population, with young individuals and immigrants entering the sample and deceased individuals and emigrants leaving the sample each quarter at the same rate as in the U.S. population. As a result, for each quarter starting in 1999, the panel constitutes a representative cross-section of the population of borrowers.

Moreover, once the random sample of individuals is identified, we then include the credit records of all individuals living at the same address, so the panel can be used to analyze new loan originations, balances, and delinquencies at the household level. Consumer debt traditionally has been measured at the individual level, but most economic decision making is done at the household level. In fact, consumer income and assets are commonly measured and analyzed at the household level. For studying household economic behavior, and in particular for understanding life-cycle household finance, it would similarly be more useful to examine debt and credit aggregated at the household level. It may also be important to consider the distribution of debt within households.

Beyond the ability to link individual consumer and household-level loans over time, the panel improves on typical loan-level datasets by linking multiple loans to the same consumer. Thus, at each point in time, one can see whether the individual holds multiple first mortgages or a combination of first and second mortgages while at the same time holding various types of non-mortgage debt.

As you can imagine, the panel has been a great resource, accessible by staff at all Federal Reserve banks through a fast and easy-to-use interface. It has already been used in several

Federal Reserve Board and Treasury reports, Federal Open Market Committee presentations, speeches, and congressional testimonies, and in about 100 research papers and policy analyses. It also has value for our community outreach, as we can profile specific geographical regions.

Particularly impactful have been our releases of the *Quarterly Report on Household Debt and Credit*, containing a comprehensive summary of households' access to and use of credit. Several of our research findings have influenced policy debates. For example, an important question regarding recent household deleveraging has been to what extent the decline in aggregate household debt was attributable to delinquent debt charge-offs as opposed to active debt paydowns by consumers. Our research found that while charge-offs played an important role, a huge change in consumer behavior had taken place. While consumers extracted home equity and took on more debt during 2007, they reverted to actively paying down debt during 2009, creating a remarkable $480 billion reversal in cash flow available for consumption in just two years. In our most recent report we are seeing signs suggesting that the deleveraging period has reached its end. The availability of borrower-level data has been key to our ability to analyze consumer behavior.

Somewhat unexpectedly, it has also turned out that the CCP provides a comprehensive national picture of total outstanding student loan balances and delinquency rates that heretofore did not exist. It was in a Quarterly Report, for example, that we first reported that student loan debt exceeded credit card debt, in early 2010. Our findings helped bring the surge in student debt and delinquencies to the attention of the public. Our research also highlighted its potential implications for consumption growth, showing a strong association between the increase in student loan debt and the decline in mortgage and auto loan origination rates over the past few years.

Coming back to the broader issue of data innovation, it is interesting to note that dataset merges have been behind a growing number of important academic studies and research findings in recent years. While showing how the value of datasets can be greatly enhanced by linking them, these studies underscore the need for creative thinking to maintain anonymity and representativeness of the matched sample. We are currently pursuing various ways to enhance the value of the Consumer Credit Panel further by linking to property deed records, employer payroll records, small business credit data, and student college records.

The CCP shows how existing data – in this case credit reports – can be re-purposed to serve new public policy and research uses. But sometimes existing data sources just can't provide the answers we need, and in those cases data innovation requires creating something *completely* new.

Survey of Consumer Expectations

So a second important example of data innovation at the New York Fed is the ***Survey of Consumer Expectations***, which we introduced to the public just last month. Consumer expectations of future economic outcomes are crucial inputs to the policy process, but we felt that our standard measures of these expectations were quite limited. The primary goal of this new national survey is to collect timely, rich, and high-quality information on consumer expectations about a wide range of household-level and aggregate economic and financial conditions. The survey covers expectations about inflation, price changes for specific goods, home price changes, future wage growth, future quits and layoffs, and residential mobility, as well as expectations of household income, spending, taxes, and credit access.

The SCE has two advantages over existing surveys of expectations: It collects information about consumers' expectations and decisions on a broader range of economic and financial topics, and it does so in a way that captures respondents' beliefs more fully. The objective is to measure an individual's beliefs about the *likelihood* of future outcomes, thereby capturing how certain or uncertain the person is about future events. In some cases, we do this by eliciting a "density forecast," where respondents are asked to assign a percent chance to

different values (or intervals) for the outcome of interest. In addition to providing a measure of individual forecast uncertainty, these density forecasts yield richer information about how expectations vary across people and over time than do single-value forecasts, or "point predictions."

The SCE is the outcome of a five-year collaborative ***research project*** involving economists, psychologists, and survey design experts. During this period, we conducted a large number of experimental surveys to assess the information elicited with existing survey questions and to develop new questions, including those for measuring forecast uncertainty. This preliminary work and the probabilistic format of the new survey questions represent an important innovation that builds on academic work over the past 20 years by economist Charles Manski, as well as some early pioneering survey work conducted by the Bank of Italy.

Creating the CCP and SCE has yielded some important lessons. First, data innovation requires an ability to think outside the box, and it benefits from the collaboration of people with diverse skills and backgrounds – academic economists, but also representatives of other academic disciplines and individuals with business and public policy experience. Skills in identifying valuable but underutilized data, the creation of data linkages, the development of appropriate sampling procedures, and survey methodology and questionnaire design experience are all particularly valuable. In this respect, our experiences mirror those at the many new interdisciplinary data and social research centers and numerous "Big Data" university centers being created across the country. A key impetus behind these centers is to bring together a group of individuals with different skills, interests, and experiences in an environment that encourages independent and innovative thinking. These individuals bring their unique questions and knowledge to the task of finding new ways of using the enormous amount of electronic data that exists today, as we did with the CCP. In addition, they can collaborate to identify efficient ways of creating *new* data, as we did with the SCE.

A second lesson is that it is important to offer clear incentives for data innovation. Giving researchers the flexibility and the resources to use new unique data in their research and policy analyses is a strong motivating factor. In fact, we have found the availability of unique data to be an important recruitment and retention tool for talent in research.

A final lesson is that data innovation can be costly, both in terms of required staff hours and effort and in direct cost, especially when it involves data purchased from vendors or data collected through surveys. The gains should therefore be carefully weighed against the overall costs.

Data stewardship

Data innovation as I've described it here comes with some costs. In an environment with rapidly growing numbers of large and complex datasets, efficient data utilization requires an ability to easily find, trust, use, and share the data. Without those prerequisites, there is a significant risk that innovative and valuable datasets would be underutilized because their availability is not widely known, their reliability is inadequately evaluated and documented, and their use proves too costly, requiring large time investments by each user. In my view, effective data utilization therefore cannot be achieved without adequate recognition, funding, and support of a data steward role.

Subject matter experts serving as data stewards have a number of important data management responsibilities. A primary one is the production of detailed documentation of the data sources and content. Such documentation is vital for preserving institutional memory, which can be lost when key personnel leave, and it needs to be updated regularly. This documentation also includes the creation of higher-level metadata that can help potential users determine the applicability of the dataset to their specific needs. Data stewards play an important role not only in documenting the provenance of the data but also

in evaluating the quality of the data, to determine whether it can be trusted for different purposes.

The availability of rich documentation on data lineage and usage, combined with high-quality user support, will reduce learning costs for users, lighten the question-and-answer workload for data providers, and promote the efficient allocation of resources. Adequate documentation of data quality and the tracking of known data issues will help ensure that the data are properly used and will reduce the likelihood of user misunderstanding of the data, a risk that can lead to inappropriate data usage and erroneous results.

Our limited experience to date with data stewards who have been explicitly hired or assigned to this role has been very positive. This investment has helped accelerate data usage, ensure analytical quality, and maintain vendor contract compliance. Importantly, it has significantly reduced the burden on senior staff in responding to data questions and requests for data or analyses.

Successful implementation of this model requires not only a formal recognition of the data steward role, but an appropriate level of resources to support data stewardship, including some training in the foundational disciplines of data management. Incentives may be required to induce subject matter experts to take a formal data steward role. To provide these incentives requires thinking through career concerns of people, and may require a whole new career track for data stewardship, one that provides sufficient autonomy to provide the steward time and incentives to conduct data analysis and research. These investments are very worthwhile, because they unlock the potential in innovative datasets and make progress possible in answering crucial policy questions.

Learning from others

Before I close, I'd like to note that data innovation is challenging, but there is an effective and widely available way to achieve it, and that is to learn from others.

Over the past few years we at the New York Fed have hosted and visited colleagues at the Bank of England, Bank of Korea, Bank of Japan, and the Bank of Mexico to exchange ideas and discuss some of our data projects. Much can be learned from sharing experiences in creating and managing new data products. The information uncovered in some countries' well-designed data collections can help other countries to direct their own data-gathering efforts. For example, other countries are very interested in what can be learned about the household sector from credit bureau data like the CCP. We have taken suggestions for particular questions from other countries' surveys.

We can also benefit tremendously from learning through direct collaboration. In a new research project with the Dutch Central Bank, we are investigating how the interview mode affects the survey response rate and response accuracy when asking people about financial issues. More specifically, we are evaluating whether soliciting personal financial information through the Internet elicits more accurate and complete responses than securing the information through face-to-face interviews – a pattern observed in the case of other types of sensitive information. Given the increasing prevalence and popularity of web-based surveys, which are considerably less costly and provide more flexibility, the results are likely to be of relevance to any central bank considering conducting or already implementing a survey on household finances.

Of course, in my brief tour, I have not even touched on the extensive regulatory data collections we undertake. An exciting development in that area has been the creation of the Financial Stability Board Data Hub at the Bank for International Settlements, the result of extensive collaboration among central banks and banking supervisors in many countries.

Conclusion

To conclude: I've focused on two challenges that we face in making the best use of data to address pressing policy questions. The first challenge is to bring innovative ideas to data. Although progress has been made in this area, more imagination and thought are needed in identifying and creating useful data. Once a data product is imagined, the right investment is crucial to deliver good data to analysts and policymakers. The second challenge is to provide effective data stewardship. Good management of data is needed locally, often in the person of data stewards who make information accessible in a world of imperfect data, nonintegrated technology platforms, and statistical packages. Significantly, learning from one another in conferences such as this one is vital to meeting the innovation challenge and to improving data around the globe.