THE BROOKINGS INSTITUTION

CAN BIG DATA IMPROVE ECONOMIC MEASUREMENT?
PART OF THE HUTCHINS PRODUCTIVITY MEASUREMENT INITIATIVE

Washington, D.C.
Thursday, March 14, 2019

Introduction:

LOUISE SHEINER
Senior Fellow and Policy Director, Hutchins Center on Fiscal and Monetary Policy
The Brookings Institution

Case Studies in Using Big, Privately Gathered Data:

JEREMY MOULTON
Associate Professor of Public Policy, University of North Carolina, Chapel Hill
“Using Zillow Microdata to Value Housing Services”

CLAUDIA SAHM
Chief of Consumer and Community Development Research,
Federal Reserve Board
“Real-Time, High-Frequency, Geographic Measures of Consumer Spending”

Panel Discussion - Using Big Data: Potential and Obstacles:

DAVID WESSEL, Moderator
Director, The Hutchins Center on Fiscal and Monetary Policy
Senior Fellow, Economic Studies, The Brookings Institution

MICHAEL BROWN
Principal U.S. Economist, Visa

FIONA GREIG
Managing Director, Director of Consumer Research, JPMorgan Chase Institute

ERIC GROSHEN
Visiting Scholar, Cornell University
Former Commissioner, Bureau of Labor Statistics

CRYSTAL KONNY
Consumer Prices Branch Chief, Division of Consumer Prices and Price Indexes
Bureau of Labor Statistics

Closing Remarks on Privacy:

BRIAN HARRIS-KOJETIN
CNSTAT Director
Panel on Improving Federal Statistics for Policy
and Social Science Research Using Multiple Data Sources

* * * * *

PROCEEDINGS

MS. SHEINER: So I'd like to welcome all of you and thank you for coming to this conference on the use of big data to improve economic measurement. My name's Louise Sheiner. I'm the policy director of the Hutchins Center on Fiscal and Monetary Policy here at Brookings where our mission is to improve the quality of fiscal and monetary policy and the public understanding of it.

This conference is part of a larger, ongoing initiative on productivity measurement that we're doing here at Hutchins, which we are undertaking with support from the Sloan Foundation. And Danny Goroff is here, thank you very much; and Antoine van Agtmael, who is also with us today. So thank you both for being here.

As you know, productivity growth in the U.S. has slowed over the past 15 years and there's been this ongoing debate about whether or not that slowdown is real or whether it's because we're really increasingly unable to measure the real economy. So regardless of the answer to this question, that debate has refocused attention on the shortcomings of the official measures and the challenges that the statistical agencies confront in keeping up with a rapidly evolving economy where cellphones are a substitute for digital cameras, services on the web appear to be free, and where increases in wellbeing and GDP and welfare are more likely to be through the quality of goods than the quantity.

So the larger productivity initiative we put together a panel of experts from industry and academia, co-chaired by me; Janet Yellen, who was here, thank you very much; and Jim Stock, who's also here. And we've commissioned about 15 papers on various aspects of productivity measurement that are now in the -- you know, being written, so stay tuned. We hope to see some of those papers possibly in the next year or so.

You know, one issue that came up when we were thinking about this broader initiative is obviously about the potential for big data to address some of these measurement difficulties. And luckily, the Conference on Research and Income and Wealth, the CRIW, was already involved in putting together a large conference on this very topic. And we actually decided we would work with them to highlight some of their findings and sort of in a less technical, more public-facing fashion. So that's what this is today.

So we are highlighting two of the papers from their conference. And if you want to see all the other papers, they have a whole bunch of papers on this topic, you can search for CRIW on the web and they have the full agenda and links to the papers. So if you're interested, please do that.

And we are really grateful for the organizers of that conference, Katherine Abraham, Ron Jarmin, Brian Moyer, and Matt Shapiro, who agreed to work with us and allow us to sort of, you know, work with them to do this conference in conjunction with their conference this weekend, tomorrow and Saturday.

Okay, so let me just briefly tell you about the plans for this afternoon. So our first session is going to highlight two of the papers from that CRIW conference with presentations by Jeremy Moulton from the University of North Carolina Chapel Hill and Claudia Sahm from the Federal Reserve Board. And we'll follow up those presentations, which we'll have time for Q&A after those presentations, with a panel discussion that's really going to focus on the potential for big data to help us in improving economic measurement, but also the challenges and obstacles that need to be confronted in trying to use -- harness big data for official economic measurement.

And finally, we'll close with remarks by Brian Harris-Kojetin from the Committee on National Statistics. Harris-Kojetin? Sorry. I asked him. I forgot. (Laughter) Sorry. The Committee on National Statistics, who will talk about how concerns about privacy affect the prospect of using big data for economic measurement.

So let me just note that this conference is being live streamed. We're posting the video. And thank you again for coming and hope you enjoy this afternoon. And let's start with Jeremy.

MR. MOULTON: Well, thank you for inviting me to share this work with you today. And I just want to say this joint work with Marina Gindelsky and Scott Wentland, who are over here with the BEA. And they'll join me up here for Q&A.

I need to make sure to express that these are our views. These aren't the views of the BEA or the Department of Commerce or even the Zillow Group.

Now the BEA, their main directive is to think about national accounts and how best to sort of measure those, one of those most importantly being GDP. And for today's talk I want to focus on one component of one part of GDP. So consumption is one part of that. We're going to focus on this component called housing services or space rents.

How it's currently done at the BEA is that we think about homes that are rented and we think about all other homes, as well. And homes that are rented we can go out and we can do a survey and sort of figure out how much are people paying in rent and that was done in the Residential Financial Survey back in 2001. And so we take that and we sort of spread it forward using data from the Census Bureau to change sort of housing quantities. Have those increased over time or decreased, but mostly increased? And then how prices and quality have sort of changed using data from the BLS.

Then you take those rental and you think about, okay, well, if you owned a home, how much would you rent your house to yourself for, is essentially a way to think about it. And so using that data they spread it across to sort of all owner-occupied homes.

And so looking down here at sort of the two rows, we have "Rental of tenant-occupied" homes and "Imputed rental of owner-occupied" homes. And if you notice, about three-quarters of that value that's in that space rent service, housing services is actually this imputation of rents onto owner-occupied homes. And we've gotten really interested in that part and also sort of in this rental part, as well.

This is what the data looks like currently. And so going back to about 2001, it's about 2016, we have in blue the PCE housing services number. And it looks pretty linear there. And then we also have what percent is this housing services? So it looks just about 10 to 12 percent, depending on the year.

This is what sort of made us think a little bit about this. If you look at housing prices, so this is the Case-Shiller Index, other indices look very similar. Anybody who's lived through this knows there as this big boom and then there was a bust, and it's sort of come back again. And so we sort of thought is there a way is there a way -- and this whole paper is essentially a proof of concept and it's another thought of is there another way to think about housing services and different ways to sort of measure that?

And the way that we're approaching this is going to take these prices a little more into account directly. And so to start with I want to think about are there pros and cons to the current approach? There are a lot pros and there's a few, couple cons that I want to focus on.

One of the pros is that it's the most common method that's used across the world. And rent is really this direct measure of what we want to think about when we're thinking about what the housing services is, so it's a direct measure of it. And it's a reasonable approximation if you have a pretty thick rental markets. So if there's a lot of rental units within an economy, it makes a lot of sense.

Some of the cons are that not all economies have thick rental markets. Some of them are very thin. The World Bank I think says if you have less than 25 percent you should use this other method I'm going to show you on the next slide. And then this is the real big one. If you've ever lived in a rental and you've lived in an owner-occupied home, they're occasionally different. So I've lived in the same house, it was built in the '60s, for seven years, just bought a house, super different. Very different houses. And so there are potential problems with trying to spread rental rates from rental properties onto owner-occupied properties.

Another one, and this is really relevant I think for this particular one, which is the data series can expire. So they go back to here, the residential financial survey that they're using, 2001 is where it was benchmarked last. They are sort of looking at other alternative data sources to push this forward, but data series can expire. And so you have to sort of rely on other data to help push this forward.

So what we're doing is, in many ways, a very different approach. It's called "user cost." And we're thinking about calling it essentially working from the bottom up. So we calculate a user cost or essential a rental equivalent -- a user cost for every single home in the United States and then we aggregate up to a national level, which is a very different approach.

We think about using capital theory. Essentially, what does it cost you to own that home? And we go through all sorts of different components of that. One of the most important ones being the price, interest rate, how much does it cost you to finance it. There's a risk premium associated with this: depreciation; maintenance, the house might fall apart; your property tax rate. And then do you expect this thing to appreciate or depreciate?

And so the values that we're using here have been established in the literature and we're using those. The method has been established in the literature. Our contribution to this is we're focusing on price and we're focusing on the idea that we can build this bottom-up approach.

And so to focus on that I want to look at how do we actually figure out what the price for every single home in the United States is? We use a hedonic model and any realtor will tell you there's three things that are important for valuing a house. Right? And they're all right up here, right? (Laughter) Location, location, and location. Those are the three most important things, right? They're right there. We've got other stuff, but those are the three important things.

And really it's not just sort of where it's located, but it's also how big is your house? That's the square foot. How big is your lot? And then we've got other stuff, you know, bathrooms, bedrooms, how old is it, and so on and so forth. But really those three things -- the location, how big it is, and how big your lot is -- and we're allowing those thing to vary with location. So in New York an extra square foot is going to be different than Chapel Hill where I live, or something like that. And so this is the model we're using.

And we're using actual sales prices. So we observe sale prices across most of the U.S., and then we're going to spread those sale prices to every single house in the United States. Here's where the data comes from. This is one of our other huge contributions, I think, to this approach is that we're able to actually observe real prices. And most of you have probably been to the Zillow's website. It was just redone recently, so I've got new pictures up here. But if you haven't been, you get pictures; we actually don't have pictures, which is too bad.

I have in orange things we don't have access to. So Zillow makes this data available to researchers, but some of the stuff that they have they actually can't release to us. And so I just want to make sure this is clear. We do not have the list price, which is what's up here at the top. We don't know what the house was listed for, but we do know how big it is: 4 bedrooms, 3 bathrooms, 3,500-ish square feet. And we know exactly where it's located, we know the address. Every home.

We also know what type it is. Is it single family? Is it a condo? Is it a townhome? How big is the lot? What year it was built. I got down here how many stories is it? Does it have a deck? Does it have a pool? So on and so forth.

We also have right now 2015's tax amount. So how much did they pay in property taxes? How much was it assessed for?

We also -- so over here on the right, just to make clear, we know when the house sold. We know the date that it sold, but we don't know listings and we don't know sort of when prices changed and so forth. That would be awesome to have, but we don't have it with the Zillow data. So down here we've got this, the little money sign there telling you you have that price.

And the Zillow data is composed of two parts. One of those parts is the tax assessment data. And that one we think we have the universe essentially of all houses in the United States and we know the characteristics of those homes. So we know how big they are, we know how many bedrooms and bathrooms they are -- they have. And so it's about 200 million parcels in 3,100-ish counties.

This is the dataset that we're going to spread prices across to. And then we'll use the transaction data, which we've merged onto this data, and this is where all the sales prices occur. So we see how much did it sell for? When did it sell for? Did they have a mortgage? Was there a foreclosure? So on and so forth.

One issue with this data is that for the non-price disclosure states we actually don't have their sales price. And so we need to actually -- I'll show you in a minute, we're going to use Census divisions when we do aggregation. We're going to have to assume that homes -- sorry, states that are sort of in the same Census division but release their price data are similar-ish and we can sort of aggregate it from there. Now, there are other datasets we could use to get this price data, but for now, we're using the Zillow data and trying to show that there's proof of concept that this could actually potentially work.

Marina and Scott have done a comparison between Zillow's data and the American Community Survey just to see are they drastically different? Is Zillow giving us something totally different than what Census gives? And for the most part, they look pretty similar, which is good. This makes us feel pretty good.

So here's the method. There's a lot of words here. You're not supposed to do this if you present, but I did it. So there's a lot of words here and essentially what we do is we estimate a hedonic model. We try to figure out a predicted price for every single home in the United States using the sales that actually exist.

We move on and we estimate year-to-year growth in all those prices for every single county. And then we spread, we say, okay, if we have an actual price for your home, let's use that price. And then we're going to spread that price forward using price changes within the county. If we don't ever observe a sale for the house, we use the predicted price from the hedonic model.

Once we've done that, we calculate the user cost for every single house within the area. Then we figure out what's the average user cost within that state for different bedroom types. And that's very similar to what they're currently doing, only they're doing it more on a national scale. So we have five bedrooms, four bedrooms, so on and so forth. We know what the average user cost is in that state in that quarter.

And then we sort of aggregate up using Census divisions, using those values that we get in part 7, and we also use the quantities from the ACS, so we know year to year how home quantities are changing. And then obviously, we aggregate up to the national level.

And so I'm going to spend just a second on this slide just showing you the difference between SFRs, single-family residences, and the non-SFRs. But really I want to focus on these slides. These are the ones where I'm looking at sort of annual changes or -- annual values here.

The blue line here is our user cost method using the Zillow data. And the orange dashed line here is the current PCE housing values that are estimates from BEA. Right? And they're actually very similar for the last several years, but they deviate during the boom and beforehand actually.

We also thought, okay, we've used some of the values that are in the literature. What happens if we change those a little bit? So on this slide right here what we have in blue is our default and our default includes the 10-year Treasury as sort of the value plus some sort of risk premium on top of it. If we instead use the 30-year mortgage rate instead and don't add that risk premium on top of it, we get basically the same thing. So the green line there is basically the same thing.

And then if we say one of the components of the user cost was do we expect the home to appreciate or depreciate based on sort of what happened in the prior year, if we say, okay, the default is to use about 2 percent. A lot of countries use that when they're using the user cost. If we just throw that out we actually get a slightly higher user cost. Same exact sort of trends.

If we use -- there's an entirely different approach to how we think about appreciation of the home, if we assume that the year-to-year change in the home's appreciation is what they expected to happen for the next year, we get a sort of slightly different user cost trend here. So here's what it is with our default where we're just assuming it's about a 2 percent growth rate. If we do it this way, we get -- essentially it shifts to the right and that's because two things are happening. One is interest rates were low over here and then the appreciation was exploding. And then over here on the right, the opposite is what happened, in that middle part.

Most people don't use this approach. Most countries do not use this approach. And if you were to sort of think of different values for this appreciation, you'd probably have something sort of in between these two.

Lastly, one of the big contributions of using this bottom-up approach is that we can look at different areas of the country and sort of look to see what's happening with user costs. And so what I have here are -- here's all those different Census divisions. And then in pink we've got sort of the Pacific, which includes California, Oregon, and Washington. You sort of expect that's probably going to have some of the higher values for homes and things like that, and so that's why the values are so high there.

When you look at -- these are single-family residences. When we look instead at -- this is sort of outdated, but non-single-family residences, actually the Mid-Atlantic, which includes like New York, actually becomes quite a bit higher, as well. And so it allows us to sort of look at a more micro level at some states even, Census divisions. You can also look at distributions of this rather than just looking at a single value. You could look at what's the distribution of the user cost within a state or within a area?

And so I just want to sort of finish here with some of the advantages that we see to using big data to estimate these sort of national statistics. We're doing this micro to macro approach. And it allow us to do all these sort of different cuts to the data, different disaggregation. The good thing that we think is that we're using sort of actual market transactions. We're not relying on people to tell us what do you think your house is worth?

This is an actual price that actually occurred.

And then lastly is coverage. So we have, for the most part, a good percent of the United States that we can sort of observe and observe all these prices and calculate these statistics for. We also can observe homes that are going to be used as rentals or owner-occupied. We observe them all. And we can also look across the entire price distribution. So some very expensive homes are probably never going to be rented, and so you may never find a rental equivalence for those. And so it allows us to sort of look across the entire distribution of the price.

And so, again, this is just a proof of concept and it’s a potential alternative to be used. We’re not sort of arguing that for sure this is the one that should be used, but we’re just arguing this is a potential that could be used. Thank you. (Applause)

MS. SHEINER: Okay. So let me tell you, and remind me if I don’t do it, when you speak, you have to hit this little mouth thing on the microphone.

So I’m going to open this up for questions in a second, but I have a few of my own and some of it is sort of clarification. Actually a few little questions.

So this was all about nominal housing services. So how do I -- is this all about the deflator or is it about the real? So what do your things do to real GDP?

MR. WENTLAND: Yeah, so we started with nominals. All of our figures are currently in nominal. I think the next step is sort of how we might think about it. But as a component of PCE, you could in theory build this into constructing a different PCE measure if you want to go that route. So, yeah, this is not something that we have redone in our current paper.

MS. SHEINER: You haven’t figured out what the implications will be for real GDP.

MR. WENTLAND: Yeah.

MS. SHEINER: Because I guess the deflators are coming from -- so the deflators would be coming from the BLS still and then you would put whatever. You haven’t gone to that step. Because I had a question about how to think about sort of the user cost and amenities. So an amenity changes and the price changes. Do we think of that as more real services or the price of services? And if you didn't change the price would this be more real services you would be getting?

MR. WENTLAND: Yeah, so that's a good question. So to the extent that amenity values are going to be capitalized in rents that you would sort of see that in the rental data, right? So when Amazon chooses to reside in Northern Virginia, you start to see rentals creeping up because of the convenience of being close to Amazon and in Arlington, Virginia. So that's sort of amenity and other amenity factors. And that, of course, can be bid into prices.

So I would say, yes, that would sort of be directly a factor that I think is real. Yeah.

MS. SHEINER: So when I'm looking at the rental equivalence versus your user cost, how do you think about why they're so different? Is that something about the rental market not -- just how do you think about it? Because in theory, if they were sort of the -- I mean, I understand there are some things that are never rented, but your theoretical construct should be the same, right? Or is that not right?

MR. MOULTON: I mean, I think we talked about this yesterday. There are papers that actually show that rental rates and these PCE measures actually don't track very closely, not super closely. And so, I don't know, I don't know how far I want to go with that, but I think there are some potential issues with how it was spread forward from 2001, when the last RFS was put out. But I'll let you talk.

MR. WENTLAND: Yeah, there's some issues with the data. There's recent literature on computing rental indices for the U.S. and also major markets and they are using different data and different methods. Recent literature has found that the BLS's approach and the BEA's approach with this rental data that people have found actually different time series dynamics with using the different data and different methods. And so whether BEA and BLS is absolute truth, we, you know, have no comment there.

MS. SHEINER: I see. So I shouldn't interrupt the differences as being a difference of the user cost method versus rental method necessarily, but everything that goes into having different dataset, different -- everything like that. That's interesting.

Okay, questions from the audience? Antoine? So please wait for the mic and state your name and tell us where you're from.

MR. VAN AGTMAEL: Antoine van Agtmael, Brookings Trustee. I have two questions.

The first is did I understand you correctly that your method, at least in two of the three, it would seem that it gives you a better sense of how inflation feels because it's used on user data? So that's the first question.

And the second question is, what I thought was very interesting was that in this graph and in how you compute it, you didn't show the impact of the stock market on housing transactions, which we know (inaudible) effect is usually there.

MR. WENTLAND: Yeah, so generally when we show the graph of looking at any of the BEA, BLS's series on housing and using -- relying on this rental data, at least for most of us who have experienced the housing boom and bust in the last 20 years, it doesn't quite feel right, and so -- based on the experience. And so you're in the middle of the 2000s and you're seeing inflation numbers that are pretty low, but yet you're seeing house price inflation, at least measured within the house price indices, exploding. There feels like that tension, that inconsistency, and that's a little bit of what we're sort of getting at.

MR. VAN AGTMAEL: But my question was, does your user data approach actually improve on that, which I thought it would?

MR. WENTLAND: Yes. Yes, it does because sort of the core of our measure is house prices.

MR. MOULTON: That's right.

MR. WENTLAND: Yeah.

MS. SHEINER: Martin?

MR. BAILY: Martin Baily, Brookings. So just help me understand. You've got this big jump. So the BLS measure is pretty flat and you've got this big jump.

All right. So during this period the purchase price of houses went way through the roof and then came down. So is that basically what you're reflecting?

MR. WENTLAND: Essentially, yeah.

MR. BAILY: Because if you're thinking about the user cost of housing, it's true that the purchase price went up, but if you were expecting inflation to continue, then it didn't look like the rental costs -- I mean, the implicit rental cost because you were going to get a 10 percent gain in the value of your house. And that's one reason everybody was rushing to buy houses. It's not because the houses looked cheap. The houses looked very expensive. But people thought they were going to essentially be cheap is the wrong word, but the cost was going to be low because the price was going to be even higher the following year.

So just help me out understand how those things play out in the picture that you described.

MR. MOULTON: So if you looked at I think it was the third picture where we used the alternative approach to looking at appreciation, which actually incorporated year-to-year changes in the prices rather than assuming a 2 percent appreciation, you actually notice that the user cost actually doesn't rise until, when was it, 2000 -- it's much later. And so actually during the sort of escalation of it, you actually do see with that approach the user cost does not explode initially because people expect this sort of 10 percent appreciation on their home price. And that's actually sort of reducing the user cost.

MS. SHEINER: I have a follow-up. So there's this sort of this period where expectations of price changes are going down, but actual prices don't go down? Like how does that --

MR. MOULTON: So, I mean, initially, it was sort of prices were going up and appreciation was going up.

MS. SHEINER: Right, exactly.

MR. MOULTON: And so that's one of the reasons why we push it this way. I think it must flip at some point. But people are still expecting it be higher, but then prices are not doing that. They're doing the opposite, if anything.

MS. SHEINER: Yeah.

MR. KRAY: Hi, Christian Kray. So I was interested how micro does your locations effects model get inside cities where there's a lot of redlining, you know, Chicago, Baltimore? Are you able to do any expansionary work in that kind of level of detail?

MR. MOULTON: Yes. So we have tried to get down as low as possible as far as location is concerned. I think our current one uses Census Tract. And what we've done with the current model is essentially we run one model at the Census Tract level if there are at least 10 sales within the quarter, and then we could move up to the FIPS with the county, and then we move up to the state for those areas that don't have at least 10.

We could go smaller in some areas, I think. We could go in a Census block group, maybe in New York or other areas like that, where you actually have sort of many more sales at the small geographic level. We don't want to go too low if there's not enough sales to sort of calculate these hedonic price models. But we've tried to go as low as possible geographically to get around things, you know, for reasons like this.

MS. SHEINER: Rob?

MR. SEAMANS: Thanks. Rob Seamans from NYU. Fascinating presentation.

You motivated it by talking about how housing matters for calculating GDP. Have you gone through the exercise of sort of recalibrating what GDP should have been, you know, using your method? And I'm just curious --

MR. WENTLAND: That's a good question, yeah.

MR. SEAMANS: So I take it the answer is no, but I would love to see that.

MR. WENTLAND: No. But, yes, we do have some follow-up projects that are doing exactly that. Yes. (Laughter)

MS. SHEINER: Marshall? Or we can go back there because the microphone -- then we'll get to Marshall.

MR. BACHMAN: Thank you. Danny Bachman at Deloitte.

I'm trying to get my mind around that hump because in real terms the housing stock and the actual amount of housing people consumed didn't really change, right? You're living in a house, the house is the same, the value of the house goes up, but you're still living in the same house. And then the value of the house goes down, but you're still consuming the same house. So presumably, you could only use this if you had a proper price index that went along with it, which I guess is the question you were asking at the beginning. Right? This only works if you have a consistent price index that picks that up.

MR. WENTLAND: Yeah, that's probably the biggest challenge for converting this to real sort of -- a real measure. But this is also sort of a fundamental question about economic measurement more broadly, that when we see prices go up, the question is how much of that is purely monetary or inflationary versus how much is a real quality change. And if we see prices going up, is that reflecting some underlying amenities? That was a question earlier. Or is it reflecting something else that may not actually be "real?"

So, yes, that is a fundamental question and something that we should think a bit more about.

MS. SHEINER: Marshall? Giving you a workout, Helen.

MR. REINSDORF: Marshall Reinsdorf, IMF.

A couple of questions. Easy one, age has a nonlinear effect on price. I assume you guys have modeled that and you didn't put in linear.

But the next question I wanted to ask, you know, if you look at the long-term profile of your estimates, it's pretty flat. Right? Not a lot of up for 15 years. And you wonder, you know, my impression is house prices went up a lot in 15 years. And, okay, so why is it so flat? I think, well, okay, interest rates came down. That depresses your user cost, right?

But you might think, I don't know what happened to the stock of housing, it'd be really useful to try to decompose, see if you can come up with a price and quantity components. Because I think most of what you're getting is really prices moving around. But if you have any comments on why is it flat and what's driving the movement it'd be interesting.

MS. GINDELSKY: So one of the things that we wanted to look at in terms of stock of housing is how to weight the Zillow data. So one of the real important challenges with big data is precisely how representative is it? Because in order to do the kind of decomposition that you're describing, which is really important, we need to know what we're truly capturing.

And so one of the things we're doing right now is linking it on an address level to Census data. This link is in progress. And we are at the stage where we're going to start the analysis. And that will give us a sense of what happened to the housing stock and are we actually measuring it, whether that change is reflected in our transactions.

MS. SHEINER: Anybody else? Okay, I'm going to ask you one last question.

So is this something that BEA is thinking about doing? Is this a proof of concept for them or is it just an academic project?

MR. WENTLAND: Yeah, so given the current series and the fact that it was last benchmarked in 2001, we are exploring, and this came up in the last Advisory Committee meeting, that we're exploring different ways in which to improve this measurement. And so a couple different ways.

The thing about it is we have -- you know, you could look at different data, but the same method, which is something that one of our colleagues is looking at using different Census data, but using this rental equivalency method, or you could look at big data like ours or different data in a different method. And so that's one of the things that we're exploring.

And so once we sort of look at it to see what's possible given existing data, we'll then evaluate sort of what are the pros and cons for all of them and make that determination later.

MS. SHEINER: Great. One last question back there? Where's Helen? Here she comes.

MR. REAMER: Hi. Andrew Reamer, George Washington University and a member of the BEA Advisory Committee.

One of the BEA's innovations in the last decade that Marshall was involved with was the regional price parities. It looks at cost of living across the country, and a key component of that is housing. And I believe, Marshall, the source is the American Community Survey. So I'm curious what the potential is for your method for use in regional parities to actually improve the measures of the cost of living relative to the U.S. as a whole.

MR. WENTLAND: Sure. So my understanding of it is one of the key components of that measure is looking at the price-to-rent ratio. And so one of the things that we're working on with linked Census data is potentially to get a richer conception of what that is using this blended data. So that's sort of something we're going to look at and explore in the coming years once we have linked data and we can sort of use both the ACS and the Zillow data together. So that's one of our -- that's where we're headed. Yep.

MS. SHEINER: All right. Well, thank you very much. Please join me in thanking this panel. (Applause)

And I'm going to welcome Claudia Sahm up.

MS. SAHM: All right. Good afternoon. I'm really excited to have this opportunity to share research that I have been doing with a fine set of co-authors at the Board of Governors over the past two and a half years. And what I'm going to share with you today is our efforts to take a very large dataset of card transactions, translate those into a new measure of consumer spending, and then I'm going to spend a lot of my time today showing you an example, just one example, of how we've been using these data to study economic activity.

So I'd already mentioned I have a great co-author team on this project. I also want to point out in their acknowledgements list, which is pretty lengthy, there were many contributors in constructing these data, other colleagues at the Board, and, in addition, staff at both First Data and Palantir. So this is -- I mean, big data is a big effort to put together.

And one last note, I just want to be clear that all of this research and the views I'm going to express today are those of authors on this paper and not anyone -- not necessarily in the Federal Reserve System or other company partners.

Okay, so I want to start by motivating this with a question and really like the purpose of this analysis, and not just this project that we're doing with card transactions. There's other work at the Board under our Economic Measurement Agenda of trying to find ways to use big data from private sector sources to get a better view of economic activity. And the question is, can these big data really help us improve and support macro policy?

And so staff at the Federal Reserve for many years with many official statistics, existing datasets, have spent time trying to characterize current economic conditions and also using that to be able to forecast conditions. And this is one piece of -- important piece of information and kind of supporting the policymaking process.

Now, we're going to hear a lot today about the promises and the challenges of big data. I also wanted to highlight a few pieces of this that I think are particularly important thinking about being in a macro policy setting.

So clearly, there is a huge promise of massive amounts of very detailed data that businesses are producing just by doing their business operations. So this has a real potential if we can harness and tap into that data to fill in gaps in official statistics.

One aspect that is very promising is to get more information on geography, on higher frequencies. In this kind of policy setting it helps us really be confident of what shocks we're identifying. When we see movements in spending, we can very clearly say we think it's because of this weather event or it's because, you know, there was an oil price change and this is an area of the country that's very intensive employment in that industry. So it helps us understand why we're seeing movements in activity.

And then another feature that's very important is these data can be very timely. And that helps in terms of in a policy process you can't wait for all the revisions and years for the data to come out, and so this has the potential to give us some additional information without much delay. So those are a lot of things speaking for this kind of effort.

There are many challenges, as well. And this came up in the earlier conversation, as well. These data are produced in doing business. They are not produced for creating economic statistics. Things like representativeness, like this is a picture of the U.S. economy as a whole, is not guaranteed. And also, in terms of our official statistics there's a lot of theory and rigor that goes into the construction of those statistics. Those same methods can be very hard to point at big data. And particularly in our case, like we ended up developing new methods and a very different approach. And you want to have the same kind of confidence in the big data you're using if you're going to bring it into important policy discussions.

And then a second piece that's important to think about in macro policy is that most of these big data haven't been around that long, so you end up with fairly short time series. Our spending series begins in 2010, so we don't see a full business cycle in our data. And so we think our methods work pretty well in the time period we've seen, but that doesn't mean that they would hold up as well in like sharp changes in economic conditions. And so you have -- it's very difficult to compare your recent estimates with past events. But all that said, we think, you know, there's a lot of promise to this and have pushed forward with this and other efforts.

Okay. So I did also want to point out, and I think it's important because I'm going to show you our series, some case study examples, but this really is very collaborative effort and it takes a lot of cooperation at various points in the process. So private companies that are willing to share their data for economic statistics, and this is an important public good, and it is not a trivial thing to ask of companies to partner with because their data's very important to their business. There are lots of privacy issues. And so those who do, this is really a big opportunity.

We are very thankful to be partnering with First Data, which is a large payment technology company. And so they -- First Data processes about $2 trillion of card transactions annually. In terms of the data this is debit cards, credit cards, any electronic payments using a card. So we're covering a lot of spending, but it's still one company. It's not the full economy.

So that's our data source, but then in the processing of the data we work with -- again, this becomes a multidisciplinary effort because to help ensure the privacy of the data we use an intermediary of Palantir that works with the raw data. All of the series that at the Board we have access to have been anonymized so that we don't have merchant identifiers. We have to be in a geography where there's enough coverage that you couldn't re-identify merchants. And so then those anonymized datasets are what we use.

But it was also very important, even though we don't have access to the raw data, it was very important for us to be active participants in the construction of the indexes. Like we really needed to understand what transformations of the data were being done, what kind of filtering was being done. And so that's where we have and continue to work very closely in those efforts.

And then finally, I just want to be clear this type of work would not have been possible without also relying on official statistics and working with individuals at the statistical agencies. I'm going to show you, we take the step of comparing our new indexes to official statistics. That's part of having confidence in our estimates and being willing to use them in economic analysis. But they also are a part of the method and the building of the indexes. So really we couldn't do this kind of work without the official statistics, as well.

Okay. So I wanted to show you just some of -- a little bit about the data that we're using and to show how important this filtering and aggregation of the data are. So what I'm showing you here is just the 12-month change. So over a year the change in total dollar transactions that we see from -- and this is within spending at retail stores and restaurants. So we focused in our analysis on just a piece of the transactions that First Data would cover. This is very intentionally we chose this group because this is a similar concept that BEA uses when they construct GDP from the Census retail sales, so we're kind of matching up concepts.

And what you see here on the right is before we do anything to the data, the first data has these massive increases in retail spending in this area, so in 2014, like 250 percent increase. That is not what was happening in the U.S. economy in 2014. So what you see is a comparison, the same Census series, is just on a different like magnitude here.

And so what this is, and this was a really important issue for us to address, but this is going to show up in a lot of big data, the changes in this data are -- you can have big changes driven by the business operations. So this was a period of expansion for First Data, not just like adding new merchants, but acquiring other payment processors. So you have really large increases in the amount of dollar transactions that they're capturing. That tells us something about one company. It doesn't tell us about the U.S. economy. So it's important for us to filter a lot of that information out so that it can be usable in analysis.

So I'm not going to go through our filtering process. I'm happy to answer more questions about that, but we -- the goal of our filters is to really focus on the changes in activity that are picked up in the card transactions that we think are economically meaningful, so not just specific to the business operations. And after we go through a process of filtering and aggregating we now see that the First Data series matched up in a similar way. So, again, these are 12-month changes. National data have a very similar pattern to the Census data at the same frequency.

This makes us feel much more confident about using the data and particularly going off and using it for its geographic detail or its, like, daily aspects. So this was an important step in our process.

Okay, so now I want to turn to an example of how we've used these data. So what I'm going to focus on is our estimates of the economic impact of Hurricane Harvey and Irma on retail spending. And what this is showing is just kind of the tracking of the hurricanes. We're going to be using our state-level estimates of the spending effects. This is in kind of a regression setting where we're taking into account the day of the week effects, other seasonal effects.

And so you can see, and I'll show you some more how these estimates look, but one thing that was really notable about this is we received these transaction data within three days of the event. So when we were doing our first estimates of these effects it was as the hurricane had just passed through and there was still disruption, and so it was a very different experience to kind of measure and examine this kind of disruption in real time.

Okay, so with the estimate sitting still so you can watch the pictures, this is using the state-level data. We do an estimate of the impact of the hurricane on national daily spending. Okay. And so what you can see here for Hurricane Irma, shortly after landfall it depressed national retail spending by about 7 percent; with Harvey it was about 2-1/2 percent or about 3 percent. These differences really reflect the fact that the population affected by the two storms was different. So Irma went through Florida, that was the larger population. But you can see the basic pattern of the hurricane effect is there's a disruption soon after the storm hits.

Within a week or so, spending is back about to the level it would have been in the absence of the hurricane. And then what you don't see is shortly after the hurricane a burst of spending to make up that lost spending. Right? So within this timeframe what we're really seeing is a loss in spending that occurred as the storm happened.

Another aspect we can do with these data is look at the very localized effects of the storm. And so here we're using the metro-level data and you can see again that spending is depressed immediately after the storm hits. There are -- but these are huge effects. So when you look at the local data, for the metro areas that are in the direct path of storm you're seeing spending fall by 75 percent, almost 100 percent at the time of the hurricane. But the basic pattern of it falls sharply and recovers, and, again, you don't see kind of this burst of make-up spending. So that's with the detailed data.

And then, again, to go back to the kind of macro setting, we want to take these estimates and think about how these are affecting national GDP, say, at the quarterly frequency. And what we see is that when you go to think about the national effects, what's important is the severity of the storm and, as I mentioned before, the population that's affected. So you think of kind of the consumer spending that's at risk when the event happens.

And unlike other where you kind of look at average effects over past weather events and do an estimate of what this storm would be, we're actually observing this particular storm, this particular hurricane, and seeing what its impact was and then being able to aggregate that up. So it's a different exercise. And the details to matter in terms of the storm itself.

And as I mentioned before, the geographic, the daily data really allow us to say with some confidence that what we're seeing in terms of a disruption we think is happening at the national level is being driven by this weather event as opposed to some other thing that could be happening in the economy in the same month or in the national level where you have the regular data.

And so putting all these pieces together from these two hurricanes we estimate that Harvey and Irma depressed GDP growth in the third quarter of 2017 by a half a percentage point. And so this is just the direct effect from these spending components, but that's not a trivial amount.

And so then in conclusion, as I said before, you know, we are still at the early stages of using these data and other types of big data in a macro policy setting for analysis. This requires the development of new techniques and there is definitely a need to scrutinize the data quality. And that's an ongoing need because the data that are created in business operations, things can change about the business, so you have to constantly be monitoring the quality of the data you're using.

The ability to do high-frequency, localized events allows us to kind of fill some gaps in studies that wouldn't be possible with official statistics. And a couple of examples, we've looked at the delay in the reimbursements -- or the payments for the Earned Income Tax Credit. And also, we've been able to use these data to look at sales tax holidays across various states.

And then finally, one piece that is useful, and this goes back to the idea of we're not trying to replace official retail sales statistics. And, in fact, it could be very powerful to use these in conjunction because it's a very different data source. So you have the ability to have an independent read on the same economic activity, and this can help in real time. Any measure, even official statistics, there's some degree of measurement error, and so if we can use our estimates in conjunction with official statistics it could give us a better picture of the actual underlying activity at that time.

Thank you. (Applause)

MS. SHEINER: Thank you so much. Those graphs are so much fun. I know we all love looking at those things. They're just kind of amazing. Not the moving ones even, but just the pictures and how much it affected spending.

So let me -- I want to sort of ask about how we think about the value of this. And you just said something about like knowing what that hurricane effect is so that I could sort of think of the different pieces. And one of the things is, like, when we're looking at GDP one year versus GDP another and we want to correct for extraordinary things, like the hurricane, it's kind of the only way of knowing, apart from like, you know, it was X percent of the population and we -- right? But so, I think about the hurricane effects. I remember we spent weeks and weeks at the Fed trying to figure out what those were, so this is cool to be able to actually have data.

And David asked me here, ask her about whether or not it was useful for the shutdown and whether or not you would be using that to understand shutdown effects. So what other things have you used it for?

MS. SAHM: Yeah. No, and I think -- and this part of where we're still at the early stages of using the data. Because a very important part of our process was to feel comfortable with the data. And to get to the place where you had a high positive correlation with Census, that took a while.

And part of our validation for the data was using it and testing it on events where we really felt like we knew what we were going to see. So it is nothing new to think about when a hurricane comes through, a large winter storm, that that's going to depress economic activity and it's going to be a temporary depression. You know, it bounces back.

So we started in a space where we felt comfortable where we would kind of see the effects. But even there we did learn things. In terms of a weather effect, the idea that you don't have a rapid make-up and spending afterwards, that we didn't see any evidence of that, that was a piece where we were adding some data to kind of the rules of thumb are different than that So it's a transitory event, it's not -- so I don't think -- we didn't like change kind of the view of how you think about weather effects, but we're bringing additional information.

In terms of how we use these data just kind of in general, the fact that they are very timely, so we receive -- only a few days after the end of the month do we have our first read for the prior month. And we have gone through a lot of effort to feel comfortable with the monthly changes. We do seasonal adjustment on these data. I mean, all of which are nontrivial things to do with the transaction data. But then typically, the advance read for the retail sales comes out about two weeks after the end of the month. So there is a period where we have an early read over the data. I mean, it's two weeks, right?

But in addition to that, the idea we have looked at -- so typically, it's two weeks, right? So we have -- but there's also that advance reading will often revise, so the Census data. And what we've done is looked at our ability to predict revisions to the Census data when we also have our First Data measures. And it does, in fact, help predict revisions. So you can imagine if you have a month, say like December, where there's a really large change in retail sales in the official statistics, our independent data can help us assess that to see do we think that was a real change or is that something we would likely expect to revise away over time?

MS. SHEINER: So all the filtering and stuff that you did, so did you -- so have you done I guess out-of-sample predictions with it after? Or is it, you know, I do all this stuff and I match it, but that's because I did the stuff kind of knowing I was trying to match it? You know what I'm saying.

MS. SAHM: Right. So we weren't -- I wouldn't say we were doing out-of-sample predictions. And it's not like we were testing various methods and waiting until we got the correlation up to a certain amount. I mean, there's very good reasons to expect somewhat different movements over time. But it was more just a quick check on like are we able to see the kind of patterns that we think are economic activity?

MS. SHEINER: And so a little difference between the opposing one and this is that the Fed is not the producer of the data, so you're not thinking about it as an alternative in terms of the data production, you know. But maybe eventually it could be used by the agencies to help with their official data. Right? And I don't know if they -- how you think about that at all.

MS. SAHM: Right. Well, and I would say, like, there are many things that one could do with the transaction data. So this is very detailed data. I mean, it's essentially card swipes at merchants that partner with First Data or one of their affiliates. And we know the industries that those merchants are in. You have very high-frequency data. So the application that we took to this, because we were trying to get these fitted into the macro, looking at time series, that isn't the only application of these data. And you could do much more detailed analysis with it. And, in fact, there's other statistics we have been working with and kind of trying to think about ways to use these kind of data.

MS. SHEINER: David?

MR. WESSEL: David Wessel. Three questions. One, my question about the shutdown was, in real time, when you didn't get Census data because they couldn't produce it, was this alternative source valuable or is it we're not there yet?

Two, do you have to pay First Data for this?

And three, creeping up on the question that Louise asked, do you think that this data could substitute for the retail sales that the Census collects or is it just a supplement?

MS. SAHM: Okay, so I think I'll take them in reverse order. So I want to be -- like we don't see this as a replacement for official statistics. I mean, this is not -- I think our main goal for this was to be able to use it to think about the geographic dimension, the daily frequency. So it's kind of expanding into areas that we can't use the retail sales for. If given a choice -- and even after investing a lot of time and building these series, like if I had to choose between them, I would use the Census retail sales without a question. I mean, it's really hard to deal with representativeness of the data. So this is not a replacement.

And in constructing our -- one of the ways when we constructed the methods to deal with the representativeness is that we used the Economic Census, so those are every five years. And we used them to benchmark our measures. So we feel comfortable using First Data to get a sense of growth rates in, say, a particular retail industry. But to deal with the fact that the mix of merchants that First Data might have in that industry could be different than the overall economy, it's really great to be able to anchor those growth rates at a point in time with kind of representative levels. So we definitely need the official statistics even to put together what we have.

In terms of pushing more on like the timeliness of the data and the use in the shutdown, I mean, this is one -- I mean, these were data that we had available at that time. And so that does -- well, it increases their usefulness. It also is one where -- these are research experimental series. And so like it does put a lot of pressure on them, too. And we can certainly have -- there's a lot of the business operation in the pipeline, and so things can happen, like, oh, the pipeline wasn't working for two days and so the first estimate ended up getting revised. I mean, these are just business processes. This is a lot of data we're pushing through a system and it can have issues.

So it is the case that having that data continuing, getting early reads is helpful, but it doesn't change the fact that we're eagerly anticipating the Census data when it came out.

MS. SHEINER: Dollars (inaudible).

MS. SAHM: So I'm not going to get into the arrangement with the collaboration.

MS. SHEINER: Back there.

MR. HILL: Thank you very much. Spencer Hill from Goldman Sachs. Thank you very much for this presentation.

I was wondering, first off, if you could, if you're able to just share your sense of whether retail spending did, in fact, decline sharply in December and only partially rebound in January. I think that's a question a lot of us are grappling with.

My actual question was actually about just the other service categories, particularly those in the QSS, the quarterly Census survey. Have you found that your data also may be useful as an early read on service spending in, say, the recreation or information services categories for consumers? This seems like, you know, a potentially useful input in the advance GDP report, for example.

MS. SAHM: Okay. So I'm not going to comment on any specific estimates from December or January from these data. I'm kind of going to stick to the cleared research paper that I was presenting. But those are good questions. (Laughter)

MS. SHEINER: I'm not going to tell you, but I do know.

MS. SAHM: In terms of thinking about this for the QSS, so we very specifically focused in on a subset of retailers and restaurants because, again, this was like a concept in Census that we track and look at a lot in the retail sales report. That does not mean that that's the only piece of data that is well covered by First Data. It's just where we started in terms of our methodology.

So you can think of with these transaction data anything that is paid for predominantly by a card or has a high volume of the spending done by a card are ones that we can cover. And so we certainly have started to look at accommodations, hotels; recreation is a space that we can also look at.

I will say, I mean, there's also, when we go to build these estimates, we do have to take the mapping from -- the merchants that partner with First Data, they use merchant classification codes. We have to map those into the North American industry classification system. So even just kind of setting up and organizing the data isn't a trivial thing. But there is the potential to kind of move into those spaces.

And there you are -- you can be improving timeliness here. We're getting a couple weeks on a monthly series. There you could be getting even more advance on a quarterly series like that.

MR. REAMER: Andrew Reamer, George Washington University.

Last month, the Census Bureau announced it's creating an arrangement with the NPD Group to get a data feed of third-party retail data to actually reduce the burden on response of the monthly retail survey. So I'm curious, has the Fed been in touch with the Census Bureau? So have you advised them on First Data vis-à-vis NPD Group data or are these two operations kind of running independently of one another?

MS. SAHM: Yeah. So, again, I don't want to speak for other agencies and other projects going on. I mean, I think I kind of gave the blanket -- statistical agencies are very interested in these alternative data sources. I mean, you already saw that in the first presentation. And actually, there's quite a bit of work. I would encourage everyone to take a look at the CRIW conference and kind of see some of the work. And I believe there's a paper with the NPD data that's being presented in that conference. So I think that there's a lot of activity.

And the point that you made about reducing respondent burden, I mean, that is an important -- I mean, that is one advantage of using these data.

MR. REAMER: Are there distinct differences between First Data and NPD data in terms of coverage or detail or quality, reliability?

MS. SAHM: Yeah. So I think what's interesting with these big data is the data-generating process or the way these data come about can differ quite a bit. So in our case, what we're seeing is we see card transactions, but only at merchants that work with First Data. Right? So we don't -- like for other datasets it might be where you have the household as the unit of analysis. So you can see their card -- you only see a set of households that work with, say, a financial institution, but you see all of their card payments.

With the NPD data, these are like scanner data. So there, you know, it's like at the point of sale. And there you might be more limited to the types of merchants and the product categories, you know. So you might have very -- so I think this is one of those things that if it's in kind of the sweet spot of that dataset, there's a different answer to the question of like which dataset would be preferable. So it's really going to depend on the question you're asking.

And I will say, I mean, with our dataset, because it's -- we have very merchant-centric data, it does kind of line up with Census retail sales, which is a firm survey with establishments in the retail sector. So that kind of makes a natural pairing for benchmarking. Other big data that would be more household-specific would need the kind of benchmark and compare to household surveys.

MS. SHEINER: Jim.

MR. STOCK: Jim Stock. So this is a big step forward in real-time monitoring. I remember I was involved in trying to monitor the economy during the previous government shutdown and it was tough. And it would have been great to have had data like these.

I guess the one question I have, you say that there's like, I think, a half of a percentage point drop in -- or shortfall in GDP growth in Q3. And I wonder if there's coverage issues there. And I guess what I'm thinking is that sort of some conventional wisdom is there's a decline in consumption and so forth. But then you actually have to do all of this rebuilding, and that rebuilding generates an awful lot of economic activity and might even overcompensate for the drop. And would it be possible that some of that activity might be going outside of, well, your sampling frame or outside of your dataset, which is, I think, essentially consumer or credit card transactions? So I guess it's a question -- this is a question about external validity for longer term effects.

MS. SAHM: Right. What we are sharing, this half a percentage point lower GDP growth, this is purely a direct effect from only this one channel. Right? Because even -- you could imagine an inventory offset or -- like this doesn't have to show through to GDP. So even within that quarter there could be other effects that are happening and certainly rebuilding effects afterwards. So, yeah, this is just one piece of that.

MS. SHEINER: How about Martin up here?

MR. BAILY: How do you adjust for differential use of cards so that over time a lot of people pay cash and now you buy a cup of coffee with a credit card? It's getting to the point you buy it with your phone, although that goes through a credit card. You know, people at different income levels use credit cards differently. And maybe some people want to use cash because it's more anonymous. How do you adjust for that differential use?

MS. SAHM: Right. Well, I would say there are some advantages of having a short-time series. (Laughter) So look for the positive.

No, but it is true. I mean, one piece we're missing, any cash transactions.

And so one thing we have looked at is like with the diary of payment consumer choice, about 30 percent in dollar value of payments that households make are through cards, like card transactions, whereas it's about 8 percent that's made with cash. And, of course, that piece has been trending down over time. So that is -- we don't do any adjustments for that, like that missing piece of spending, but it's true that we don't cover that.

So if you think about -- when I said before that we don't have a full business cycle and, like, feeling confident in these data, that is one where you might think that there could be shifts in the type of card spending. Now, we do have credit and debit, so, I mean, we have kind of cash-like spending in here, as well.

And the other thing that we struggle with in our measurement is the way we had to do our methodology in filtering is that we also miss economic births and deaths because we can't distinguish that from the massive amount of merchant churn, people coming in and out. And so that is another piece where one could worry about if there are changes in the economy and you have many more firms actually exiting. But that would be difficult. That's something we would miss in these data series.

So I just want to be clear, there are a lot of things that are missing in what we're doing and yet it looks like it does --

MS. SHEINER: Do you have online capture? Do you have online sales? Do you have online retailers?

MS. SAHM: So we -- the captures are electronic payments that are done with a card. We don't have -- and those get sorted into the merchant industry. So if it's like a clothing store, it would show up in the clothing industry. Unlike in the retail sales, often this will go into a non-store setting. So we don't organize our data that way. It would go -- but we can capture those.

There's an ability in the data because it's card swipes that we can look at in some cases the bank identification numbers, so we can learn a little bit more about what type of card it is. And for some cards there is a flag for if card is present. So, I mean, that's a scope to do more with the online spending in these data.

MS. SHEINER: All right, last question.

MS. KRISHNAN: I'm Mekala Krishnan from the McKinsey Global Institute. Thank you. That was really interesting. I have two questions.

The first is you've characterized the use cases of this data as high-frequency, localized effects. And I think part of that is because you can make the direct cause-and-effect relationship. But I was curious if you've thought about this for other use cases when maybe that cause-and-effect relationship isn't as apparent.

And then the second thing is I think you said this was merchant data, but to the extent that these data have demographic information, say around income levels, have you thought about, and maybe this goes to the last session around privacy, but have you thought at all about integrating that into some of the findings? I could imagine some of the analysis you've done on the spending, picking back up, that could look quite different across income segments.

MS. SAHM: Right. And so because the data are merchant-centric we can think about geography, so kind of the income level in the surrounding -- I think we probably have the potential to have reasonable statistics maybe down to the county level, so you could think about impacts on higher or lower income areas. There's nothing to these data that would link us back to demographics of the individuals swiping their cards.

In terms of use cases for the data, what I was highlighting are features of these data that we don't have in official statistics. As you've seen, we certainly look at these in the national and monthly data. And if you spend this time building a dataset, you start looking for ways to apply it. (Laughter) So I'm also open to good ideas from others.

MS. SHEINER: All right. Well, thank you so much. Join me in thanking Claudia. (Laughter)

MR. WESSEL: Well, thank you very much, Claudia. I think that actually segues very nicely into the next phase of our program. What we wanted to do was focus a little bit more on the potential for and the obstacles to using big data, both privately gathered and you'll see from Erica, government gathered data. And we also wanted to hear some from the point of view of people who are in companies that are producing these data, what are the ways in which they can and can't share the data with the public and so forth.

So, at the very end, we have Michael Brown who is the Principal U.S. Economist at Visa. Next to him is Fiona Greig who is the Managing Director and Director of Consumer Research at the JPMorgan Chase Institute which is doing a lot of work with the data that JPMorgan Chase have. Erica Groshen is currently a Visiting Senior Scholar at the Industrial Relations School at Cornell University. But relevantly, she was the Commissioner of Labor Statistics from 2013 to the beginning of 2017. And Crystal Konny is the Chief of the Branch of Consumer Prices at the BLS and she also, she and some of her colleagues, Brenda Williams and David Friedman, have a paper at the CRIW which is on the website. Which is interesting because it talks a little bit about how they are using what they call alternative sources of data.

I think that when Louise and I have been thinking about this project and what can we add to the conversation, there are at least three things that we are trying to accomplish. One is we just want to highlight what's actually happening. Because I think in some quarters, there's a great deal of skepticism, particularly in the part of the business community that the statistical agencies are behind the curve and if only they understood how to use big data better and better methods. They could do the whole thing and we could save a lot of money and get better measurements and we'd all learn that the economy is better than we think it is anyways.

Secondly, I think there are really interesting issues within the government and Erica is going to talk to them. And thirdly, there are constraints, they are opportunities but they're also constraints on the private sector. JPMorgan Chase has the luxury of having God knows how many regulatory agencies looking over their shoulders and the Bank Secrecy Act and the anti-money laundering. And then we'll talk, when Brian comes up, a little bit about the privacy issues which are both technical and, I think, we can talk some about there here. But they're also political, how much do the American people want the government to know every time they swipe their credit card and do they believe that it's anonymized and is it just anonymized to the fed. But Palantir knows exactly what I'm buying in there telling the Chinese or something.

No, I think they're real issues and people think that and it's a real issue in the census. People don't believe that the census will be kept confidential, then they're not going to tell the census takers that they're not legal immigrants or they're working off the books. So, these are not theoretical questions.

Let's see. Why don't we start with you, Michael. You've explained to me a little bit about what Visa does with its data which has some similarities to the data that First Data has. What you do with it, what you share, why you can't share more and what you think the pros and cons are of using it as a proxy for what's going on in the economy.

MR. BROWN: Sure. So, just some high-level stats to give you some insight into what we actually are able to see in terms of visibility. Basically, $.20 of every dollar, Visa captures through its eco system on its rails. Last year, that amounted to about $3.7 trillion. That's capturing both consumer as well as consumer and that's roughly just under 69 billion individual transactions that we're able to see.

So, the data richness is quite impressive to say the least. But as Claudia pointed out and her academic paper very thoroughly covers it, the cleaning of this data and the turnaround is quite immense. Just for our team in business and economic insights, we have essentially a team of three individuals besides an entire data science apparatus within Visa, that are required to clean and comb through this data. And so, you know, what are the challenges.

Well, as Claudia pointed out, certainly mapping it to economic data is extremely challenging and we have to go talk to finance groups and our mergers and acquisitions groups and our sales teams to figure out how we onboarded a major new merchant within our eco system that then boosts our payment volumes. So, there's all of these factors that are secondary to the pure economic effects that we're trying to tease out. They're extremely time consuming and it's a very laborious process to get that data to a point where we can actually utilize if for economic analysis.

The types of things that we do currently, we have a product that is know as visit database and this is probably our most high-profile product out of our group at least. Which essentially allows not only countries but states and local communities here in the United States to look at global international travel into and out of their metro area. And so, you can imagine, this is very valuable when you're trying to determine things like where should I market for tourism spending. So, it allows a lot of these countries around the world to not only be sort of more strategic about the way they plan but it also helps local communities in terms of their development and marketing efforts on that regard.

In terms of global data trends, we just completed a study looking at the globalization of cities and more specifically, the growing middle class globally and what could mean as demographics sort of catch up with the consumer spend. So, that's another product. And then one pilot project that we have, currently just a very small number of participants but we have what we call the retail spending monitor. It's a monthly report that takes Visa net data and extrapolates that to account for all forms of payment. So, cash as well as things not cash and are client based. This is similar to the analysis done by First Data or with the First Data information.

And essentially, it's giving us a decent real time pulse of what's going on but there are months that it again is the noise to the signal ratio is far too high for it to be useful as economic analysis. So, I would echo earlier comments that it is another tool in the toolbox, I don’t think it replaces our official government analysis.

MR. WESSEL: This retail monitor, is this something you provide to the public or only to Visa customers or how does that work?

MR. BROWN: Actually, it's not even to all Visa customers, it's in a pilot stage at this point. Because we're just simply testing the waters and soliciting feedback on this product. So, you know, right now I would say that, you know, that's one of the things that we're trying to develop but it's been a very slow process.

MR. WESSEL: And the other things you mentioned, are they things that you just provide as a public service or do you sell to customers or what?

MR. BROWN: So, it varies. Most of them are either some form of payment, you know, to cover the cost of producing the analysis and the high-level aggregate statistics and then Visa also has a program known as VIK or Value in Kind services that we'll provide if they are existing partners with merchants or other entities.

MR. WESSEL: You mean, to give the merchant who's a customer of using your cards information from your --

MR. BROWN: Correct.

MR. WESSEL: And if six smart economists from MIT came to you and said we'd like to find a way to use your data to do something, would your management say, wow this is a great idea, we can improve the usefulness of our data or would the lawyers say, you've got to be kidding?

MR. BROWN: So, without singling out either one of those two groups, I would like to stay at Visa. So, you know, I think the key message here is and you'll get into this later, there are two broad umbrellas of concerns that our executive team and process data users within Visa have. First and foremost is certainly privacy concerns. Privacy concerns amongst not only our consumers in protecting that data as well as our merchants. And Visa has been built on trust and it's very important and the top priority of the firm to maintain that trust and the integrity of the data. So, hence the reason why you don't see as widespread sharing of that information.

And the second is actually more to do with financial securities law. As a publicly traded entity whose revenue stream is dependent upon this sort of real time measure of the consumer, you can imagine it becomes a major concern about front running a quarterly earnings report in particular. So, the question is, who do we in a responsible manner, provide that information but yet not violate the securities law and the apparatus around, you know, insider information and the public reporting of that.

MR. WESSEL: Thank you. Fiona, let me just ask. How many people here know anything about the JPMorgan Chase Institute? Okay, so it's about half. So, maybe you could just start by explaining what it is and the we can get into some of the specifics.

MS. GREIG: Great. So, the JPMorgan Chase Institute is a think tank inside JPMorgan Chase and we use the firm's administrative data to do economic research for the public good. So, everything we do, we put on a public website in aggregate, so not the microdata but the results, the reports et cetera. Our remit in terms of the data that we're accessing has mostly covered the consumer and community bank which includes, you know, relationships with roughly half the households across all of our consumer products. So, you can think about 70 million households roughly that where we might have a checking account relationship or a credit card relationship. You can think about, you know, that would include roughly 30 million mortgages and then auto loans et cetera.

We just accessed a historical student loan portfolio that has been now sold but we still have the data. And then, you know, 2.5 million small business accounts so we can now also start to look at small business checking accounts and credit cards. And then in the investment bank side of the world, we have also started to access trading data, roughly 400 million transactions across 44,000 institutional investors.

So, we're starting to map those data assets to obviously economic concepts, income and spending of households, revenues and the expenditures of business and, you know, and, and, and. And, I think, as I said, everything we do is intended to answer economic questions. A couple of use cases, we also did some work on the impacts of income and spending in the context of hurricane's Harvey and Irma.

But, I think, what's maybe a little bit distinct from these data assets from say what we heard from Claudia or Michael is that it's not just the spending picture with the checking account data, you also see inflows and so you can study income. So, an example of what we've done with that is the work that we've done on the online platform economy where with the virtue of a big, big sample we can narrow down to, you know, 2 million people for whom we observe income coming off of 128 different platforms, you know, like ride sharing, home sharing, et cetera. And so, we can kind of quantify how big is that market.

We, you know, back in 2015 when fuel prices dropped, we looked at the MPC out of lower gas prices. We just released some work on the impacts of tax refunds and tax payments on people's, you know, their spending their debt payments et cetera.

So, it is the large sample sizes that gives you the geographic granularity and the high frequency lens but, I think, one more thing that we've been leveraging to a greater extent is just this ability to look at many financial outcomes for a single household or a business over time. I could go into the constraints, though there are many.

MR. WESSEL: Let me just mention one just one thing you didn't mention which I was impressed by. Sometimes seeing income means seeing receipt of government benefits.

MS. GREIG: Yes.

MR. WESSEL: So, there's a nice paper by Picanon about looking what happens when people get unemployment compensation benefits and when they're about to run out and what happens to their spending. But yeah, talk about the constraints.

MS. GREIG: Yeah, I mean, the constraints are, I think, Claudia did a wonderful job of highlighting some of them. I mean, these data come with no code book. They fall out of the business operations, right? And you have no idea what changes happen under the hood while you weren't looking, right? So, literally this morning we discovered that, you know, a miscellaneous income category went from including ATM cash deposits to no longer including ATM cash deposits but instead including ATM check deposits. I don’t know why somebody made that change but here it is.

So, there's that, that it makes it hard to know exactly how to interpret and importantly, when you see a big spike and there's always this volatility in the data, some of which is real, some of which is just pure when you're looking at daily data, you know, people spend 30 percent more on Fridays and Saturdays, you know, then they do on Monday's and Tuesday's. So, how do you kind of sweep all of the calendar effects out first but then also the noise that also exists within these data.

And, you know, the sampling aspects of this are tricky. Chase, in 2016, released the wildly successful, they ran out of the metal that made the cards, the Chase Sapphire Reserve. Which was targeted to high-income or potential high-income millennials. They did an amazing job. Well sure enough, we have this soaring number of high-income millennials. Spending is increasing dramatically so what of that is real versus just the portfolio changes.

So, we have done a lot of work to reweight our population and try and adjust for the Chase lens on the world. But how to do that over time when those changes could be happening faster than you can reweight is tricky. And then even just having the comparable metrics with which to weight. So, we do know something about the demographics, we know how old people are, that's pretty straightforward.

But income, we don't know what, you know, typically the census tells us, you know, AGI, Annual Gross Income, of that sort. We observe take home income that falls under the checking account. Those are very different numbers and so we've had to spend two years building a sort of machine learning approach to just predict AGI so that we can then benchmark.

So, having all of the tools with which we can then correct for these sampling changes that happen and, you know, the channel question you raised, Martin, is also really relevant here in terms of the usage of cash and checks and the demographic differences certainly by age, by income, by race.

MR. WESSEL: Do you know the race of people or not?

MS. GREIG: So, we are actually working on a big new project right now. The short answer is sort of no. So, I mean, obviously we know -- we can use census to know about racial composition of geographic units. What we're keen to do is have self-reported data which we have through HUMDA. For mortgage applications, we're required to collect that information. So, for a big mortgage sample, roughly 10 million people that we're assembling, we can have self-reported race information on that. Now, of course, people who apply for a mortgage are very different than people who don't apply for a mortgage.

So, one additional data asset that we're building is we're bringing in voter registration data from three states, Florida, Georgia and Louisiana. Those are three states in the southeast where the voter required that they collect race information when they register and those are three states that overlap with our physical branch footprint so we have big sample sizes there. So, we brought that data in behind the privacy curtain. You know, we sit on this side where everything is deidentified but, of course, the data administrators have been doing the match on last name and first name and birthdate and address.

So, we'll bring that in also as another data sort of subsample. Now, those folks are registered to vote, so they're non-felons, they're citizens et cetera so they're also biased but they're biased in different ways that the mortgage applicants. So, with those two pieces, we do hope to have actually what is a very big sample size with self-reported race with which we can then start to look at samples.

MR. WESSEL: And I assume you're not well represented the bottom quintile of the population or the bottom (inaudible).

MS. GREIG: In general. In general, I would say, I mean, I can talk about the -- in general, I would say we feel most confident about the representation of middle-income families. And our ability to say something about middle-income families. The reasons are, at the low end, yes, of course, we miss the 6 or 7 percent of households who are unbanked and the underbanked will cycle in and out of us just like any other financial institution. Also, low-income families use cash a lot, use money orders a lot. And so, our lens on their spending or even on their income is biased by that channel, those channel differences.

At the high end, well administratively our data set does not include so-called private bank, the super fancy person bank. So, the top, you know, whatever 1 percent is not there.

MR. WESSEL: So, you can't see the money going to the soccer coach to get your kid into Harvard?

MS. GREIG: Correct. We didn't see that, we don't see that transaction. I wouldn't be able to identify it if I saw it.

MR. WESSEL: Hopefully if you could, there weren't so many that it would be difficult to identify them.

MS. GREIG: But also, you know, at the very high-end, you worry about people just having quite complicated financial lives and distributing their activities across multiple financial institutions. So, does this checking account, is it comprehensive and really capturing everything that's going on, maybe, maybe not. So, that's why the middle-income.

MR. WESSEL: Great. So, Crystal, tell us a little bit about what you've actually done at the BLS and the Consumer Price Index Program to take advantage of alternative data. And what works well and what are the issues that are not so easy to work with?

MS. KONNY: Okay so let me just say David Friedman and Brenda Williams wrote the paper with me and we tried to survey everything that's going on in the CPI in terms of research and what's moving into production. And what we're looking for in terms of alternative data as well as all of the challenges that we've been facing. And so, going almost last, you've heard all the reasons why we shouldn't put it into production because we face all of the same things that you've heard over and over again. And I can explain those.

I'll give you an example. So, first let me tell you. There are three different types of data that we are trying to put into the consumer price index. One of it is corporate. That's corporately reported data that comes in a data set and it is oftentimes sales and quantities at the individual unique item product level. Which is fabulous for price index calculation.

MR. WESSEL: Because often you don't have price and quantity from the same source.

MS. KONNY: Exactly.

MR. WESSEL: Right.

MS. KONNY: And long term, we're planning on using the quantities. Right now, we're doing a lot with price replacement less so except for sampling with the weights because this is new methodology that we've not done before. So, we have measurement objectives in the CPI that we're trying to stay in line with, at the same time more long term, we're trying to figure out different measurement objectives. That can actually harness the power of the data that we're getting.

So, we've got corporate data, we do we scraping and the API usage. We get agreement from the company, we have to at BLS to be able to use that data. It's not simply plug and play which a lot of people think well, we should just go grab it, it's publicly available and that's not what we do as an agency. And then there's also secondary source data which you purchase, sometimes get free and a lot of talk has been going on about that where you've got a data aggregator that gathers up the data and it's for multiple establishments. So, go on?

MR. WESSEL: Give us the most specific example you have whether it's the Gas Buddy or the hospitals or the new vehicles, something where people get a sense of what you're doing that's different from just the usual survey method.

MS. KONNY: Okay. So, I'm going to talk about a corporation that gives us corporate data that we call Corp X. We are very strict on confidentiality so that's as much as you're going to get out of me about that.

MR. WESSEL: Well, it's a department store. You said that in the paper.

MS. KONNY: It is a department store in the paper. And read more if you want try and figure it out. So, they do give us sales revenue and quantity information for every item that they sell in the store for all of the stores in the areas that we cover in the CPI.

Because we have a geographic sample. So, there's a lot of issues with data. Some data does not have very much descriptive information which makes our methodology of match model and item replacement very difficult.

MR. WESSEL: Match model means you're trying to say how did the price of this same thing change from one period to the next, right?

MS. KONNY: Yes. And then if that same product goes away, how do we get the most comparable item replacement that we can get or quality adjust. And so, a lot of times we don’t have that descriptive data. So, we were getting data and then after several years, we were not using it in production and Corp X we're using a production in the March index is on February 10th. So, it's a major achievement for us.

But we were getting data and then they changed their database structure and they brought each store online at a different time and it totally ruined any continuity and analysis of a history of data and the index calculation. So, we had to start all over. And every single data source you look at, you've really got to look at years of data, depending on the item, depending on the source to make sure that it's making sense.

So, for Corp X, we calculated index. There's a very nice picture in the paper of graphs. We were collecting Corp X online up until this month and then we're going to use the corporate data this month that we used. We replaced the online prices with our new methodology and recalculated the index and it follows very similarly so we trust that it's okay.

MR. WESSEL: And is this lowering the cost of collecting data, improving the quality or coverage of the data or both or neither?

MS. KONNY: Both, both, ideally. We've got to be cost effective. We, as a government agency, we almost always have to make sure that the data -- that whatever we're using is as cheap or cheaper than what we've got. This theoretically reduces the cost of collection because you don't have field staff in the stores reporting the information. But it brings the cost back to Washington office where we have to do kind of -- we have to figure out what we've got, we've got to analyze the data, we've got to see if they've done any more changes to it, try to get descriptive information, try to map it to our item structure.

MR. WESSEL: And do you have a long-term agreement with them so if they get a new general council next month do you have to start all over again?

MS. KONNY: We do not.

MR. WESSEL: Sorry, I didn't mean to cause problems.

MS. KONNY: At any point in time, somebody can say, we're voluntary, that's part of our gig. We're voluntary, anybody can say no at any time. So, at any time, no matter what kind of data we're getting, they can say no.

MR. WESSEL: Okay. Erica, I know you could talk at length about what Crystal did but I want you to save that and talk a little bit about what I know you've been very eloquent about. About the barriers to using data that the government already has within the government.

MS. GROSHEN: Right. So, I'm going to start by just reiterating something everybody here knows that the whole idea of a statistical agency is to provide data for the public good. That's to help businesses, policymakers and families make really important decisions. They also need to maintain the trust of respondents and users so that's, you can't do something to solve one problem that will destroy trust in the agency because that subverts the mission. This means transparency and confidentiality are very important, you can't do things that are not transparent and you can't do things that would violate confidentiality.

And then as Crystal said, you need to be sure that you're getting the very most from every data dollar because those are very constrained and this is often balancing production versus innovation. So, why would you resort to or, you know, push the envelope on using some of these new sources. Well, you want to reduce respondent burden, you want to improve statistics in any number of ways, detail, coverage, timeliness, precision. And you also want to raise efficiency and resilience. And sometimes the fear in adopting one of these sources would be that you would actually lower efficiency because it's so hard to keep it clean or you would lose resilience because you might lose, you know, build an entire system around a data source that could disappear.

So, those things, you know, are true with external sources. They're also very true with internal sources, government sources as well. So, let me talk about three categories that are particularly salient, you know, of issues, particularly salient when you think about the government taking more advantage of its own internal administrative data.

The first thing is actually legal restrictions on it. So, there are cases and there's one, a very prominent case, use of IRS data. Where there are very strong restrictions against its use for statistical purposes. In order to change this, new laws would have to be passed and Congress has been very reluctant to do that. This is why BLS and the census both maintain separate business registers. So, that anybody who uses business information from census and BLS will eventually come to understand that these are drawn from two different universes. So, it's bad for the users --

MR. WESSEL: Because the IRS can share with census and not with BLS.

MS. GROSHEN: That's right. The IRS can share with census, census has that built into their business register and the BLS does not get advantage of that information. And that means that people who combine the data face a barrier, it also means that you have redundancy, you know, duplicative activities.

Another example is that even though the federal government pays for the unemployment insurance system across the states, the state's own the administrative data for that. And so, the BLS does not have access, in particular, to wage records from the unemployment insurance system which is everybody's wage records, not just UI recipients but everyone's wage records. And this means that the BLS has to collect more information than it would have otherwise has much less detail than it would and much less timeliness.

So, there are the legal restrictions there are also restrictions that arise from interpretation of law that you might think could be interpreted other ways but some combination of risk aversion on the part of the attorneys and the agencies interfere with this. And this can be about informed consent. So, give you one example. When the census and the BLS both started web scraping information, the BLS, Department of Labor attorneys interpreted informed consent laws as BLS cannot scrape information without informing the company that's posted that information that they were taking this information down. So, there has to be informed consent for web scraped information.

The Department of Commerce attorneys either were not consulted, I'm still not sure, or if they were consulted, they said no, that's fine, it's public, go ahead and use it. So, at least the interpretation is different. And then when you start to trade information around different parts of the government where the responsibility lies for maintaining confidentiality can also be problematic to resolve. So, that's these data sharing restrictions are a problem.

Another problem, of course, is resources. And there's two elements to this. One is, of course, that the operational arms may be under resourced and the statistical agencies surely are. So, that's one element, just under resourcing in general. The second part is who is going to pay for the curation, the storage, the programing, the staff training necessary to convert this operational data into something that's fit for statistical purposes. Is it the producer of the data, is it the statistical agency, how do you work that out, someone's got to pay for it. And so, we heard some examples of where this is happening in the private sector, in the public sector this happens as well.

The third category of barriers -- now all these are manageable, I really believe that but they are barriers, is when you get inconsistent priorities and missions. Nobody has to be evil to have different missions. If you are an operational agency, do you slow down your processes or prioritize giving the statistical agency the information they want over getting your operations entrained, probably not. And then the statistical agency has to figure out how it's going to operate in a world where it may not be able to get the information that it is expecting in order to get its releases out on time.

So, that could affect timeliness, it could affect continuity, we heard about that when programing changes happen. Quality could affect enhancements that the statistical agency would like to add to those data that the operational agency doesn't care about. It could certainly affect documentation as well. It could also affect whether or not the statistical agencies even know about the data because the operational agencies don't have any particular incentive to reveal the existence of potential useful information because that’s not their mission.

And then finally, there is the very human element that no data are perfect and yet it's embarrassing to agencies to reveal that they have imperfect data. And so, there's an understandable reluctance to have somebody else come in and muck around with your data and point out all of its flaws. When you have your own workarounds that have saved you money or saved you grief during that time.

MR. WESSEL: When you say operational agencies do you mean like the FAA?

MS. GROSHEN: Yeah. I think, for example, the unemployment insurance part of the Department of Labor. They are running unemployment insurance programs across the country. But EEOC is another example. All of these agencies, I'm thinking from the part of BLS but they have a programmatic mission and they collect data, they use it for those reasons. Now many of these barriers can be managed. They can be lifted and some of this is underway. The Ryan Murray Commission on evidence-based policymaking that Ron Haskins from Brookings was involved and Catherine Abraham has made recommendations to cut down on some of these barriers. And that led to some recently passed legislation to improve many of these things that we're talking about, including creating a departmental data officer in every agency that is responsible for making sure best practices for maintenance of data are followed and that data is not hidden under a bushel.

It also moves us closer to the presumption that government data, that the default for government data will be shared unless there's a compelling reason otherwise. So, we get closer to that point. And it requires this registration of federally maintained data sets and there are a number of steps in the right direction which will make a difference. But more needs to be done. We still haven't cracked the nut of state ownership of the UI wage records. That's not addressed. We haven't cracked the nut of IRS information restrictions and funding, of course, is a huge issue. That's not addressed at all, if anything, it's getting worse.

I will say though that this administration has proposed moving the Bureau of Labor and Statistics to the Commerce Department for a number of issues that I've just mentioned. That could be very helpful because there's reason to think that restrictions on BLS access to IRS data may be lifted if BLS were within the Commerce Department instead of within the Department of Labor. And so, that would be good, that would increase access to data. It would also reduce --

MR. WESSEL: I assume you're speaking for yourself and not the BLS staff when you advocate that.

MS. GROSHEN: Well, I've talked to a lot of BLS staff and there are reasons why people worry about it. And, I think, you know, those worries are legitimate but I think they can be addressed.

MR. WESSEL: You think it would be a good thing to move it.

MS. GROSHEN: I think it could be done badly but done properly. It would be an extremely good thing.

MR. WESSEL: That was a very good answer.

MS. GROSHEN: And that would take us one step closer to what I think would be the ideal solution of an independent stats USA not in the Commerce Department outside of the cabinet basically.

MR. WESSEL: I have a lot of questions but I want to give you all a chance, if not I'll keep going. Does anybody, yeah, sir.

MR. GRAY: Hi, Christopher Gray. So, I think only person today has actually mentioned machine intelligence or AI. Is that simply because there are so many barriers in the kinds of data sets when talking about basically, we wouldn’t be able to work out all the different legal hurdles and other things in the middle of this?

MS. GROSHEN: So, I didn't get to it. I mean, we were mostly talking about data sources rather than AI but certainly at BLS and I know at census there is an increasing use of AI. Right now, all of the agencies have a fair number of human coders who look at responses and convert text to those codes and AI has been extremely useful for occupation codes, for parts of the body's injured codes, things like that. So, there is very much an increase in use of AI for coding, in particular. Also, for modeling, that's the other place where it's being used. So, yes there's more than you think but less than there could be.

MS. GREIG: Yeah, I did mention one quick application of machine learning in our context which was, you know, we don't observe annual gross income for everybody but we do actually observe it for some people. People who apply for a mortgage, we ask, you know, what is your annual gross income, even people who -- and we have to verify that. For people who have a credit card with us, similar, we ask them for their stated incomes. So, with those truth sets, we have leveraged that through an ML approach to try and predict the annual gross income for everybody else in our checking account universe so that we can then reweight the population to reflect the census. So, that was one application of it.

MS. KONNY: And just real quick, we use machine learning to match the item descriptions to the CPI item structure currently. We're also looking at ways to do some kind of data analysis to try to flag problems in the data without our analysts doing it by hand.

MR. WESSEL: Antoine.

MR. ANTOINE: Erica, could I ask you, I guess you're probably in a better position than anyone to talk about this.

MS. GROSHEN: In the center anyway.

MR. ANTOINE: You know, you have talked about some movement that there has been in terms of legislation which is good. Some movement in terms of the use of big data and AI, that's good. If you, now that you're no longer at the Bureau, would have to make a guess and I realize it would not be more than a professional guess, as to how much time it will take before, let's say the Bureau, uses big data, artificial intelligence in a, what I would call, significant way to replace or add to some of the sampling techniques that are being used now. What would that be in terms of a timeframe?

MS. GROSHEN: So, let me answer what I can. I think you can expect that in general, the statistical agencies will in their operations, always lag behind the private sector. Because the decisions made on those data are so important that for a private data product or an experimental product on the part of the agencies, you can experiment all you want. But when you're talking about the CPI that drives a tenth of a percentage point mistake means the federal government over or under spends a billion dollars on Social Security benefits, you're going to be very careful.

So, that means you need a lot of testing and that means lags. And I think that's what you want from a statistical agency. Now, I think some things will actually be developed in statistical agencies before they get someplace else. A lot of, for instance, seasonal adjustment has really been a statistical agency expertise that the private sector didn't even have for many years. So, there are many things that happened before there.

Then the question is, well how quickly can things that now seem to be quite dependable externally be really embedded in the statistical agencies. Knowing the BLS as I do, I think the biggest barrier there is funding. That in order to make these innovations, you have to run them in parallel for a long time to the current production process before you bring them into production. And that means you have to have the staff, the skills, and the IT capacity to do this in many different ways and many different programs at the same time that you are still producing the monthly and quarterly numbers at the same quality that were being produced before. And right now, BLS has had flat nominal funding for a decade. It's very difficult to be rethinking all of your programs when the real value of your budget has been declining for so long.

MR. WESSEL: Crystal, in your paper you talk a lot about data that you purchased, right? So, I assume there's a budget constraint there, of course. I wonder if you could talk a little bit, I was interested, if you're comfortable, about what you did with physician services and hospitals where you have problems. The paper says there's a very bad response rate and you did something clever there, what we've learned from that.

MS. KONNY: Right, so for physicians and hospital services, it's a 4 percent relative importance in terms of the CPI and the response rate for medical care is the worst in the CPI for a number of reasons because of confidentiality concerns for medical and we can on about that. But the response rate is 48 percent. And it's a very large sample size, it's very difficult to collect, it's very costly to send our field staff and we keep getting a lack of response.

So, for a variety of reasons, we were concerned about -- we had issues with our medical care accuracy, is that too bad?

MR. WESSEL: That's pretty good, we got it.

MS. KONNY: Okay. So, we really wanted to try and figure out a solution for medical care. So, we started, we bought a data set that probably four years ago with our research staff and that was very lagged data for medical claims. So, it's insurance but they've got a lot of information about physicians and hospitals and what the costs are from that. So, it's like a separate use of alternative data. So, we got like what we were considering a truth data set, then we got another data set where we just researched. It was just Chicago, it was just two years of data, it was just a selection, one little sample for physicians and hospitals to try to see if it was feasible to use this for these indexes.

And the results very promising so just this year, we bought more data, we bought it across the country, we bought a bigger sample so we're hopeful that the research at the end of this year, so far it looks good and we're hopeful that this will be a supplement. It's insurance only but it will be a supplement to what we produce for the sample for physicians and hospital services. So, that's supplementing our sample and making that accuracy, the index hopefully much better.

MR. WESSEL: Fiona, could you talk a little bit about how you have integrated academics, PhD students into the JPMorgan Chase project?

MS. GREIG: Yeah. We've done this really since the beginning we have collaborated with a few academics. And I emphasize few because we are often asked by academics to access our data. And the arrangement that we've built has been first starting with some PhD students but now, you know, a few of them have moved on and we've brought on a few other folks, junior, senior faculty. To either have access to the data and write an academic paper on the data or to be a collaborator with somebody else who does have access.

I would say we've waxed and waned in terms of how much we want to expand this program or how much we feel we can expand it just because, you know, the academic publishing timeline is three years. So, the unit costs on a single academic paper from our standpoint are very high. I mean, all of the work that Claudia described in terms of just understanding a data set, leveraging it for economic research, you know, for each and every academic to get up that learning curve and then produce a White paper, take it on the circuit, get it through the publishing process, I mean, it's a big investment.

So, it is not a program that we have felt that we can scale for that reason.

MR. WESSEL: Are there compliance issues?

MS. GREIG: We've worked through compliance -- no, it's not hugely. I mean, there are aspects of our data that we don't allow these academic fellows to access. But, you know, I think one thing that is in addition just to sort of, we don't want to give unfair data access for three years to a single person to write one paper. I mean, that's sort of just a value proposition that doesn't quite work for us in terms of cost versus benefit.

But then underlying, you know, what's going on within the bank, it is the private sector, things change so fast. Whether it's our privacy controls or our governance structure or text doc. I mean, our text doc is completely changing over right now. And so, the code base that these academics wrote two years ago, if they want to revise that paper, guess what they're going to have to refactor the entire code set into a different language in a different processing environment. It's almost like they can't keep up with the rate of change that we experience.

MR. WESSEL: Michael, I wonder if you could talk a little bit about, so what does Visa think. You mentioned one instance where you do provide to government this state and local governments what's happening to international tourism. What, in general, is the stance about giving data. You know, let's say the fed said we'd love to see your data. What are the privacy concerns, the commercial concerns or just in general, what are the constraints that help people understand why this doesn't happen all the time. Why isn't the fed just swooping up all the data from every credit card company in the country.

MR. BROWN: Sure. Well, a lot of it, I mean, one is just technical. So, you know, I'll give you an example. Our CEO actually just met with representatives at the Federal Reserve here last week. We received the request, I believe, it was a week and a half before he came to town here, which was not near enough time to clean our data and turn it around to make it representative of the larger economy.

So, a lot of these technical challenges that Claudia mentioned are probably number one from the economist's perspective. From the lawyers and the shareholders perspective, it really is about data privacy concerns in an environment in which those rules and regulations are changing literally daily some part across the world. You know, Visa is a truly global company so what privacy rules and regs are implemented globally matter for our, at least, IT infrastructure and eco system here in the U.S. as well, so those barriers are erected internally.

And then finally, as I eluded to earlier, it's really about how do we as a public company share payment volume data which is then very representative of what our earnings flow would look like for a specific quarter. From a statistical user's standpoint, I want as much real time data to do analysis on the government shutdown and hurricane effects and so on and so forth. And even myself sitting within Visa, it's extremely restrictive to publish anything or opine on anything or comment on anything until after an earnings release.

MR. WESSEL: Right. So, I have a Visa card. Have I given you permission to use my data for all this stuff or is that how it works?

MR. BROWN: So, the agreement is not only with the cardholder but also with the merchant. Both of those handshakes have to occur for us to have visibility into the data. And then as some of you may know, beginning next year in California, you will have the right to opt out of us having visibility to that data as well as part of those new provisions. That's an example at the state level and we see a number of other states also moving in the direction of greater privacy. Which then gets to the more fundamental question about big data is, is it truly representative and that representation of the data is going to dramatically change over time as tastes and preferences about individual privacy use evolve.

MR. WESSEL: Okay interesting. With that, please join me in thanking the panel for a fascinating discussion. We ended on just the right note. Brian Harris-Kojetin is going to talk about privacy. Brian is the director of the Committee on National Statistics at the National Academies of Science. And importantly for this purpose, I learned that he majored both in psychology and religious studies in college. So, we're going to get both the left and right brain thing. He also has a PhD in social psychology and he knows a lot about this stuff. So, Brian is going to talk some and then we'll do some questions on that. Thank you very much.

MR. HARRIS-KOJETIN: Thank you, David. So, it's always dangerous to be asked to talk about privacy, especially at the end of an exciting set of presentations and discussion. I had planned to use slides but I only thought I probably needed one, just something that says, sorry, no. We, in fact, have at the National Academy is doing a lot of different studies that the Committee on National Statistics and other places that involve privacy. We've tried to recruit privacy experts to be on different panels and I can tell you I've been told, I don't want to be the person who has to tell everybody no.

So, the advantage to being last is that I think all of the really good points have already been made by people who can express them more eloquently than I can. Let me just kind of recap a little bit here. So, it's clear that there are really exciting possibilities with using big data, with using multiple data sources. This is something that is every national statistical office in the world is interested right now. I can tell you I've been to three or four international conferences in the past year and this is the topic of how do every country is trying to figure out, how can they better utilize administrative data, private sector data, sensor data, other kinds of digital exhaust organic data, whatever you want to call it and harness this information and figure out how to, as you can see, deal with all the challenges. Clean it, filter it, weight it and use it and most often to figure out how to either supplement official statistics or enhance existing federal statistics.

The replace word, the substitute, as you heard from several people here today is like no we're nowhere near that realistically. But sometimes we can produce much more detailed, much more timely data, data on other topics that we currently can't measure well at all or that we're not measuring. So, there's a lot of potential here but there are a lot of costs. There's a lot of research and a lot of work that has to go into doing this.

And as Erica pointed out, which I think we're all very well aware, it's also something that agencies feel is really necessary to do although they're not being given the resources to do it. But the costs of their data collections, their budgets have been flat or declining in real terms and yet pressures for more data from users are rising but their data collection costs are going up. Respondents are less likely to want to respond to voluntary or even mandatory surveys that are going on. And so, and some of their concerns of respondents are tied to privacy and confidentiality. We can't tell you exactly what percentage of that is, it certainly doesn't help that they're hearing on the news that the Census Bureau is collecting citizenship data. But these are long-term trends. We've been seeing this happening for many years for decades but it seems to be accelerating in recent years. I haven't seen real recent data, I don't know if it's been accelerating in recent months or not.

And, I think, we all recognize that there are real fundamental societal concerns here and issues involving privacy that arise regularly in the news and in national conversations that implicitly and explicitly are about privacy. How often is it between announcements in the news about different data breaches or some other kind of thing like a data broker or data aggregator which we knew about but lots of people in the public didn't know about. It's like wait a minute, there are people out there who do nothing but vacuum up data or Experian now has all this information on me and now it's gone.

And then there's the bug that's a feature, the hidden microphone in your thermostat and other sources of surveillance, technology, cameras, sensors, cell phones. Add to this, the obtuse privacy policies that someone referred to just a little while ago in the terms of service. You know, for every app that you're doing, that you want to utilize something that, you know, I don't read on a tiny screen. You accept because you want to use this nice, cool, wonderful gadget and you just gave them permission to collect all kinds of information on you.

So, this larger context here is certainly coloring the perception of the public as well as some of us. But it colors perceptions of the things that we're trying to do for the public good and there's a not a lot we can do about that. So, company "X" sharing data with the government, you know, as you heard first-hand and better than I can express, perceptions and concerns about that and different arrangements for that.

Likewise, the government, you know, if it was revealed that the government was sharing data with company "X", wait a minute, what's going on here and who owns stock in this company. So, all of this is being very much viewed in an atmosphere of mistrust and in the context of all these breaches and other things going on. And I'm sorry to say that if we walk in and say, trust us we're researchers we're only doing things, that's probably not going to help a lot, unfortunately.

You've seen a little bit here and I'm sure there's a lot more examples in the conference to come. But are different arrangements for data being shared that how this is being done and some of it wasn't as explicit. This is happening under a variety of different circumstances. So, sometimes it's microdata but often it's not, it's often more aggregated data that is actually being shared. So, BLS and census and other agencies start trying to get more companies to say give us lots of data and you have longstanding relationships with many establishments and are trying to, you know, continue to cultivate those relationships. But it's a relationship with each company that you're doing that with. But there are also other arrangements.

It was mentioned Palantir is kind of the intermediary here for first data and there are other similar kind of arrangements where you're not getting access to, BEA is not getting access to the microdata and being able to match it at establishment level or anything like that. And so, there are different ways of dealing with this.

And you can see here that some of the issues that were especially clear in the last panel. Even when parties are interested and motivated, there is a whole raft of complex legal and policy issues that have to be addressed and that take a lot of time to negotiate these things. BEA just didn't call up first data and say, hey or you know, how about sharing some data with us. Next week good for you guys, we'll get together and to this. Even BLS and census, BEA and census and other agencies sharing what they're able to share, legally able to share the rule of thumb used to be about 18 months to negotiate some of these memoranda of understanding between agencies for getting access to some of the administrative data that Erica referred to as well as other kinds of data. So, these are kinds of things take a -- even when folks want to do them and believe they're doing it for the right reasons and for the right purposes take a lot of time.

If you start talking about the private sector, the company has to address some of the issues that you heard raised here. Liability in case the data are illegitimately accessed. What is the scope of access, are they complying with privacy laws, are they complying with security and exchange laws which I hadn't thought of before so thanks for sharing that. What are the impact of freedom of information laws? Is sharing violating terms of service and then there's simply a perception of well, you know, the company is hanging onto their data and they're protecting it, they believe really well, why would we let it out. Can the government protect it as well as we do, what happens if somebody else has access to it and they breach it? What are our responsibilities, what are our liabilities and what are the perception issues then?

There's a fair amount of talk about trying to do these agency, kind of the public/private partnerships. These kinds of situations, and I think it's worth pointing out that sometimes our federal agencies aren't the greatest partners in the world. We are very proud of our confidentiality laws and especially Title 13 but they also really put a limit on what the agencies can do. So, we think it's great that yes, we can take the data in to the agency and don’t worry, it can't come out. Well, that's kind of a lousy deal for a partner so what's the partnership here? So, I give you my data, you get to use it and what do I get in return?

So, in terms of figuring out kind of how better we can work, the agencies can work with private sector firms to make this a win-win situation. So, it isn't just well, we'll buy the data from you. So, figuring out how to make this work in an atmosphere and under a situation that respects the confidentiality and protects the data of both the private sector firm and the statistical agency.

I can share a few observations of what we've seen by doing a few studies at the National Academies. I just want to point you to a panel that many of you may have heard of. We had a panel chaired by Bob Gross looking at multiple data sources for federal statistics. Shameless self-promotion, that panel produced two reports. You can get these off the National Academies Press, NAP.edu. You can download them for free, just look up federal statistics. This panel was specifically charged with looking at how to utilize multiple data, new data sources to enhance federal statistics. Out of the two volumes, three chapters are on privacy, more than any other topic. Recognizing that this is a fundamental issue that has to be addressed, it's not only an issue with the survey data that agencies produce and disseminate now but it becomes even more of a critical issue as you try and bring in whether it's administrative data sources or private sector or other kinds of data sources.

Similarly, the Commission on Evidence Based Policy Making has been mentioned several times. Privacy was fundamental to kind of -- what they talked about as well. They had privacy experts on that panel, on the Commission, I should say, and that figured prominently in their solution. Both on the Commission and our panel, recommended going in the direction of some kind of intermediary, some kind of trusted third party. Many of you know, the Commission referred to it as a national secure data service, not a big huge data warehouse that would have a nice big target painted on it. But a service that is brought together and linked these data and provided access and then let them disappear. Which researchers go, you let the data go away? They figure that was the only way that they could get buy in for something like that.

Our panel similarly talked about -- they did not make anywhere near as specific a recommendation but talked about the various models that there might be for doing this and gave some different examples and talked about the strengths and weaknesses of each one in terms of the different locations whether it should be an agency or it should be a federally funded research data center or it should be some kind of university partnership or something like that. Where it would be, what kinds of functions it would serve and those kinds of issues.

Trying to address kind of the two major types of privacy risk which have been talked about a little bit indirectly. One that is kind of the obvious one is the security breach but the other is the inferential breach. Is the ability to even when as many of you know, even when you're just putting out aggregate level of statistics, if you put out enough aggregate statistics, too many of them too accurately the fundamental law of information recovery says from cryptography and computer science shows that you can reconstruct a data set. And Census Bureau has done some cutting-edge research in this regard and has recently talked about actually you reidentify some of these people. Census was able to reidentify some of these people because they knew the truth. Other people could reconstruct the data set but wouldn't be able to know if the record that they reconstructed was an actual person and who that person was.

So, these kinds of issues are what we're dealing with. The more proliferation of these kinds of data that are out there, the more risks and threats there are to the business that we're trying to do in producing information for the public good. And so, the panel talked about kind of really recognized we're right now in a period of transition. We have all of these data sources out there and we're aware of the weaknesses of our current methods for kind of protecting them, at least from an inferential basis and it's rather obvious that there are weaknesses in protecting security data sets more generally.

But the feasibility of actually implementing some of these technologies, such as differential privacy, hasn't been clearly demonstrated yet. We need more case studies, we need more research and development to show how this can be done and what the costs and benefits of implementation of these methods are. We're going to need a lot more work across collaborators and stakeholders to move forward, to provide good privacy protection while getting access to more data sets. And we need robust discussions of the implications of this for all stakeholders and for users. What does this mean for users. These discussions really need to be informed by some concrete examples that will help people understand what the implications are of these technologies.

And in my final act of shameless self-promotion, we're putting on a workshop on this June 6th and 7th at the National Academies on challenges and approaches for protecting privacy and federal statistical programs, trying to highlight some of these issues. Looking at statistical disclosure limitation methods, looking at differential privacy, talking about the policy issues, talking about the implications for data users and what agencies are struggling with right now. And so, if you're interested, that will be on our committee on National Statistics website. We'll have a flyer posted in about a week.

We really hope to kind of help have a dialogue here with a broader statistical and economic community to deal with these privacy issues. They're pervasive, they're not going away and they fundamentally effect what we're all trying to do, so thanks.

MR. WESSEL: I want to ask you a couple of questions and then we can -- I want to offer the audience a chance to ask questions of anybody, any of the speakers who are still here, we can be open. First of all, what is differential privacy?

MR. HARRIS-KOJETIN: Good question. Is Danny still here?

MR. WESSEL: He's in the back.

MR. HARRIS-KOJETIN: Danny, do you want to answer that? He's writing a paper on this. Differential privacy is an approach, it's a property actually of an algorithm that infuses noise in statistical results. So, making it more difficult to be able to reconstruct the underlying data. And this is something that comes out of cryptography in computer science. There is a lot of theoretical development on it. It's relatively very recent being applied to federal statistical agencies but a number of agencies around the world are looking at this. The UK has been looking into it, the Census Bureau has been looking into it, John Abowd has been leading an effort there.

Trying to deal with the shortfalls in our current methods of statistical disclosure limitation methods, things that you're all familiar with doing some kinds of pergerbation (?) doing data swapping, doing top coating. All of those things used to work in the world a couple of decades ago, they're not sufficient anymore.

MR. WESSEL: So, basically the question is, what we're worried about is if the Census Bureau says, in this census track there are three black people, it's pretty easy to figure out which one is male and which one is female. And with all sorts of other things, you can pretty much figure out who they are and so then there's no privacy there. And the old system is, you pretend that there's only two black people there put one of them somewhere else or something? And this is a cryptographic way to accomplish the same goal but not to distort the data so much or have I done violence to this?

MR. HARRIS-KOJETIN: So, under the old system, they'd tell you that they swapped some percentage of records and, of course, the exact percentage that they swapped is --

MR. WESSEL: Swapped from one potential census track to another.

MR. HARRIS-KOJETIN: Yes. And so, even if you find that it's like, ah ha, I found this record and I bet this is you and, you know, your age and race and your household. But there's uncertainty with that. It's like, well it could be you but it could also be somebody in the next county.

MR. WESSEL: I see. This is an alternative to that.

MR. HARRIS-KOJETIN: Yes.

MR. WESSEL: Okay. So, I want to step back for a minute and ask kind of a big picture question. So, there's one view that it's just a matter of time before we figure out how to take advantage of all this big data and the machine learning and AI techniques we have and the speed with which we can collect it. And we'll just have an increasingly ever clearer picture of what's going on in the world, we'll have a much richer sense of academics, we'll have a better sense of what's really going on and the cost of gathering data will be less and the statistical agencies will be able to tell the Federal Reserve to the minute how the economy is doing. So, that's kind of like one caricature.

And the other is, you referred to this and so did some of the panelists, are this is at about the limit. Because the angst about A, the security of the data, the hacking and B, the people wanting to protect their privacy, the Europeans, the Californian's. Means that like the more attention this gets the more there's going to be reaction and that the first case scenario is just going to basically forever be a raisin that dries in the sun. Because the public response, this urge for privacy, this reaction against big brother, you know, everybody jokes that if I ever forget my password, I'll just call the Chinese and they'll give it to me. That politically, that's gathering steam. I wondered, first of all, what do you think of my characterization and do you agree that somehow the privacy fear of hacking thing is growing and it's going to constrain us from the ideal.

MR. HARRIS-KOJETIN: I'm not sure the ideal is ideal.

MR. WESSEL: Right, extreme, the extreme.

MR. HARRIS-KOJETIN: I think there are legitimate concerns about privacy and I think, as the Commission talked about as well as our panel, you know, the idea of a national data center came up in the 1960's. And it blew up and ended up inspiring the privacy act. So, I think folks are trying to be very, very careful this time and figure out how we can do this right and how we can really build in these protections so that we can do some very important work and get valuable insights. So, I think it's critical that we do it right. Which way it's going to go, I have no idea.

MR. WESSEL: Right. I guess there are two competing issues. One is lack of trust in institutions in general and lack of trust in experts, that's one thing. On the other hand, people seem to be willing to give up an extraordinarily amount of data about themselves to Facebook or Amazon or Netflix. So, it's not clear to me -- they say they're worried about privacy but they don’t act on those concerns.

MR. HARRIS-KOJETIN: Yeah, somebody can probably answer that better than I can. It doesn't mean that -- so part of the issue here is that no person can -- they're not reading those terms of service. They're sacrificing something for an immediate, for the coolest, latest thing but then they're absolutely appalled when they find out that Facebook has done this or this app has been collecting this information or there's a microphone in my thermostat that I didn't know about.

So, we've reached a point where nobody can possibly understand these terms of service. No company, no reasonable company can tell them exactly how their data are going to be used under all circumstances or how they're sharing it with other folks. So, some people say that we kind of need a whole different conversation here and Europe has obviously gone one direction here about what does privacy mean, what kind of society do we want to live in and how do we deal with this and I have no idea where that's going to go.

MR. WESSEL: Sir. Can you bring the mic down here? I want to emphasize, you can pepper Brian questions but if any of the other speakers are here and either want to weigh in on this privacy thing or you want to ask a question, please go ahead.

MR. BRODSKY: My name is Mark Brodsky. I think all of this is hopeless in privacy. You haven't got to the real identifier which is DNA related information. And once that gets up there and you then can track who put up the DNA and where and relate it to everything else. Why haven't I heard DNA mentioned in this whole discussion today?

MR. WESSEL: Well, I think the short answer is that so far, very few people are using DNA to improve the measurement of the economy. So, I emphasize so far. Although I think you illustrate exactly the thing I noticed. So, you read these incredible stories about people who have voluntarily had their DNA tested, made it available to other people who have similar DNA and all of the sudden you discover your father isn't your father. So, and I think also the issue comes up quite a bit in health insurance. You know, once they start knowing everything about your genome, it's going to destroy the very concept of insurance. But, I think, the short answer is the one I gave before. Gentleman in the back.

QUESTIONER: A question for Brian and then a quick mention of a conference coming up that is going to have many panels like the one we just -- the great one we saw today. Differential privacy. John Abowd and Steve Ruggles from University of Minnesota are having like a traveling debate society where they debate each other regarding -- so, John is the chief scientist at Census, and he's really the advocate and really the evangelist for differential privacy. And maintains that differential privacy will not really harm the ability of economists to do research. That the underlying statistics will still be accessible and Steve Ruggles takes exception to that.

So, really a question for you, do you have an opinion or does CNSTAT have an opinion about the potential negative implications of differential privacy for economic research. And then just very quick, National Association for Business Economics, NABE, next November 3-5th in Seattle is having a conference in which Erica, Crystal were at the last one, John Stevens. Talking about this kind of thing, big data held by the private sector and the role of the government in getting access to them in way that is useful for economics and statistics.

MR. HARRIS-KOJETIN: So, no way am I getting between John and Steve. Very dangerous place. So, I think part of the issue here is we don't have concrete examples of what this really looks like in practice and we don't know what it would mean. I mean, users are scared to death that their microdata are going to be taken away and there will be no more agency microdata ever again. And so, that's driving some fears and we don't know or at least I certainly don't know what does it mean to have epsilon of one or two or ten. So, I think what I believe is needed or is a first step is much more dialogue and back and forth rather than this is the way we have to go and by God, no way are you going there which is a little bit of what's happening now. That's a gross oversimplification but I'm.

MR. WESSEL: The gentleman here in the sweater.

MR. SADOWSKI: Thank you. George Sadowski, ex-Brookings. You mentioned Europe. And recently the European Union has stepped up enforcement with penalties of its GDPR, General Data Protection Regulations. To what extent are they going to be the leader in this privacy tool? What extent are they going to set by restriction, the policy for data sharing of private firms?

MR. WESSEL: Anybody want to answer that?

MR. HARRIS-KOJETIN: So, the U.S. has obviously had a very different system that's much more sectoral. So, healthcare data are covered under HIPAA, education data under FERPA, some data are privacy acts. So, we've taken a very divided approach here and they've taken a much more unified approach. The more globalization is occurring, you know, they're impacting some of the big technology companies like Google and Apple and forcing changes in what they do. To the extent that that then spills over here. I think they could end up having a large impact.

MR. WESSEL: George, you wanted to add something?

MR. SADOWSKI: Visa must have European --

MR. WESSEL: Yeah, I was going to ask Michael. George, you're the first person I ever met who was ex-Brookings. I thought it was like the Marines, you are never ex.

MR. BROWN: All right, so hard to follow that. So, essentially there's a few things that Visa is doing, obviously, a number of things that we've had to do to conform to GDPR regulations in Europe. The sort of nuance with our firm in particular is that Visa Europe for the longest time was really operated as a separate business. And it was only in recent times has that been folded into our universe and that's why we're starting to see it sort of transform the data flow that we have internally.

So, from our perspective, we are seeing a migration of the IT systems into a unified system. And as part of that, obviously we need to build in the same barriers, backstops and protections that we have in Europe to ensure that both the U.S. and Europe data are not mixing. But also, that those provisions for data that is allowed to mix are separate and comply with GDPR regulations.

MR. WESSEL: And this is before China lets you in, right?

MR. BROWN: Right.

MR. WESSEL: The gentleman in the back in the red sweatshirt.

MR. SLOAN: Mike Sloan. A word that has not been mentioned is X12 EDI from ANSI. It has a pretty mature standard for HIPAA and privacy and a few other things. And it can be prevented from being aggregated from disaggregated so you can find that so for privacy. They have a continuing effort and it's also structured so it can interface with big data and AI tools. It doesn't do all of the psychosocial type things but it certainly would answer the question of what happens when you have a hurricane and what's happening to everything. IRS, some of the bigger companies are using it as a reporting standard from companies like John Deere and Granger to report to the IRS. I just wonder if that's come across your purview?

MR. WESSEL: Do any of the people here have a view on that? We'll note that. Over here, did you have a question?

MS. GONZOLSKI: Marie Gonzolski, BEA. My question is both as an academic researcher and as a government economist. Do you think there should be different standards of -- well, not different standards but different treatments of differential privacy for statistical agencies versus academic research? Because one of our primary challenges as academic researchers is to correct for biases that we see from things like added noise when we get this sort of data. And as government economists, we spend so much time trying to get our statistics exactly right that, of course, you know, the addition of noise to data scares us. So, I was wondering if you think there should be a different sort of treatment for those things?

MR. HARRIS-KOJETIN: I have no thoughts about that. I have not heard anyone talk about that before either. One of the advantages that is touted about differential privacy is that it actually gives us better insights into the properties of the data or the impact of disclosure protection on the data then we have right now. Right now, you have no idea what BLS or Census or anyone else has done to the data, how that actually affects the result. With differential privacy, you actually can model and understand the noise that was added to that. So, there may be some better insights possible that way.

MR. WESSEL: I think there's also, as I understand it as an observer here, there's a bit of question about differential access to some. Some researchers get to look at the unnoised data and others only get to look at the noise and that creates a little bit of tension among the researchers. Danny, do you want to say anything about differential privacy since I know you live and breathe this?

SPEAKER: Well, you twisted my arm.

MR. WESSEL: Didn't take much twisting.

SPEAKER: So, first of all, the name is unfortunate because it does suggest that some people get a different amount of privacy than other people. And, in fact, the techniques don't even have to do so much with privacy per se. They really have to do with trying to answer questions so that you get answers that depend on the distribution and not on the sample. And if you think about it that way, then it's actually what you would want to do even if you were talking about data that comes from rocks and stars rather than from people.

So, it's also not an algorithm, it's not adding noise to the data, it's adding noise to the answers. And you need to do that because of federal law and because of theorems. Not because somebody said, oh this is a little bit better than swapping but because the federal law says that census, for example, is not allow to reveal any kind of information that comes from -- that can be tied back to an individual. And the Census Bureau has shown in its experiments that you can implement some of the theorems that have been proven over the last few years and from the release tables, the aggregate tables that are routinely put out from the census, you can reidentify individuals.

And if you combine it with commercially available data, you can get their names and addresses and if you pay a few pennies, you can get thousands of fields of information about them. So, that's unfortunate if you don't want people to know something about ethnicity and race and other things that aren't readily available in commercial data sets. But now if you imagine that there's a citizenship question, then that's a very, very, very serious matter if you can reidentify from published tables, people's citizenship.

So, you have to do something about that, that's theorems and federal law. And the theorem is the definition, differential privacy is a definition, it's a concept. It says that it's virtually impossible to find out anything about an individual by answering questions in a differentially private manner. So, a mechanism for answering questions is either differentially private or not.

Again, you're not adding noise to the data because in order to do that, you need a model. And once you've assumed a model, you'll never really get anything else out of the data. So, you don't add noise to the data, you add noise to the answers. And then you can prove a theorem that says that you can actually protect privacy and you can't even tell whether a given individual was in the data set or not. And if you can't tell whether I was in the data set or not, then obviously you can't tell anything about me. And so, that's the theorem, that's what they're trying to implement at the Census Bureau and it's hard to know any other concept that would apply given the laws and given the theorems.

MR. WESSEL: Thank you. Antoine and then Jim.

SPEAKER: First of all, thank you for what I thought was a fantastic, if I can when I'm talking about productivity even exciting. You mentioned something as a risk that is perhaps the longer we wait, the more difficult it is to get some of these data as opposed to the other side. And so, I had a thought. I heard Erica talk about the funding which is obviously a very big problem. Although, in the big picture it's, of course, not a big problem, I mean in the overall government budget.

But you mentioned something that struck me and you said an 0.1 percent difference in the measurement of the CPI is a serious matter because it may cost a billion plus dollars. So, I had a thought. We know that if we measure the productivity improvement right, inflation will go down. So, we will save the government money so this is a very good investment to save money or am I wrong?

MR. WESSEL: Without prejudicing, whether we're measuring it too high or too low, that's the goal of our project.

SPEAKER: Can I have one more go at your caricature question. If we were to look back from the future of 2030 and if it is correct, that there are going to be lots more computer brains and everything else over this next decade. Are there any countries which have a model which is, you know, just worth looking at? I mean, a couple of ideas. Some people say that Lichtenstein's blockchain is -- I'm not talking about minting coins. And some people would look at India and say well, they put (inaudible) in the middle of creating a billion-person identity, maybe that is the sort of future of census and surveys at least for India. So, I was just wondering if there are any cases that people look at and think, well at least those are good for those countries or those are worth looking at.

MR. WESSEL: I was just going to say, if all you care about is getting information on individuals to a great extent and you don't have to worry about privacy, the Chinese seem to be pretty good at that. With all these social scores they're building and the facial recognition, I'm hoping that when we look back from 2030 that the United States does not come into that world and that 1984 looks like amateur hour. Can you imagine what Orwell would have done if he had known about AI and facial recognition?

I mean, it's a good point. I think there obviously are things to learn from other countries. Maybe not the conclusion I just drew from China but when you look at statistical agencies the Canadians have done, what Erica thinks we should have an independent statistical agency. There's a CRIW paper that some work from the Bank of England has done about scarping job, help wanted sites to get information. So, I think that and we've learned in other context that American's have a certain arrogance that we must do this stuff better than everybody else so there's no point in talking to people from other countries and we're trying to press against that prejudice.

MR. HARRIS-KOJETIN: One anecdote is a lot of national statistical offices around the world, again in trying to adapt to this new world. The statistics Netherlands has kind of been on the cutting edge in terms of trying to utilize new alternative data sources and produce official statistics. It has taken a rather different response in terms of really trying to reorient the whole agency in terms of producing information that is, in fact, immediately relevant or mining from what they're doing and being more information oriented that was opposed to, it's not that they're not producing GDP just like they used to but trying to be more responsive to various policy issues. So, it's causing a lot of folks around the world to rethink exactly how to maintain their core principles but be more responsive and relevant.

MR. WESSEL: Right, that's a good point. I think someone from the Netherlands Statistical Agency is speaking at the CRIW conference tomorrow. I want to thank the CRIW for helping us organize this. Ho Nguyen and Anna Dawson and Vivian Lee and Mike King who helped us on our staff, Stephanie Cencula. These conferences always seem like no one had to do any work to make it happen but in my experience that's not the case. So, this one didn't have any flaws so that's even better. I want to thank all the panelists and speakers who came, and, of course, all of you for being so thoughtful and having such good questions.

I have one request. If there's a glass or piece of paper at your feet, if you could take to the back of the room or put it outside, I'd appreciate that. And would you all please join me in thanking everybody who participated.

* * * * *

CERTIFICATE OF NOTARY PUBLIC

I, Carleton J. Anderson, III do hereby certify that the forgoing electronic file when originally transmitted was reduced to text under my direction; that said transcript is a true record of the proceedings therein referenced; that I am neither counsel for, related to, nor employed by any of the parties to the action in which these proceedings were taken; and, furthermore, that I am neither a relative or employee of any attorney or counsel employed by the parties hereto, nor financially or otherwise interested in the outcome of this action.

Carleton J. Anderson, III

(Signature and Seal on File)

Notary Public in and for the Commonwealth of Virginia

Commission No. 351998

Expires: November 30, 2020