Celebration of 10th Anniversary of Data Science Center at RTI International: *A Vision for Modernization in Official Statistics at the U.S. Census Bureau*

Director's remarks as prepared for delivery

April 23, 2024

- Good morning. It's great to be able see old friends, and it's such an honor to join you in my role as your director of the Census Bureau.
- First, let me congratulate RTI on this remarkable 10-year anniversary of the Center for Data Science . . . ten years of great work and amazing growth!
  - *<<Reminisce about connections to RTI folks>> Lisa, Kelly L., Jill Dever, Paul B., PhilK, Joe M., Frank Mierswa, Tim G., Steve Cohen, Rachel Harter, Craig Hill, Jamie Ridenhour, Sara Zuckerbruan, Emily Hadley, Andy Peytchev, and Rachel Casper.*
  - *<<we are community . . . help and support each other via our relationships and networks regardless of competitor status>>.*
- Today, I'll talk to you about our vision for modernizing official statistics at the Census Bureau and throughout the federal statistical system.
- I'll start with a little context.
- The landscape of federal statistics is changing and has been for decades.
- The Census Bureau—and indeed all our fellow statistical agencies—face new and complex challenges.
- Our challenges include such things as decreasing participation in surveys; increasing sophistication of data science in a way that threatens disclosure risks; increasing burden of responding to what were once simple questions, like monthly household income; and the relevance of seemingly well-grounded measures such as sex and race-ethnicity.
- We're now two decades into the 21st century.
- We are a global society that embraced technological advances.
- And these advances have altered how society sees itself . . . like how we consume data, how much we rely on it in daily life, the pace of our data consumption, how we approach problems and their solutions, and even how we think about things like privacy, even when there is not much privacy left in this world.
- By the way, it's concerns over privacy and trust in government that have contributed to steadily decreasing response rates for federal surveys.
- But getting back to the topic of new technology and data, for some of us old fogies like me who used slide rules and computer punch cards, living through and experiencing our technological renaissance has been nothing short of transcendental.
- And as we all know, such change is accompanied by new ways of leveraging technology.
- We now have cloud computing; we have data science, the art of big data analytics, and artificial intelligence.

- We're increasingly blending data from different sources, like surveys and administrative records, to develop new statistical data products.
- Data vizualizations have blossomed, including countless applications involving geospatial and geotemporal data.
- And thanks to new and increasing access to data, we as human beings think of ourselves differently.
- I, for one, have used DNA testing to find my ancestral roots come, in large part, from indigenous northern Mexico and from southern Spain.
- Many others have done similarly and, combined with population migrations, many in our nation now think of their racial and ethnic identities in a more nuanced, richer way.
- That affects the quality of data when trying to capture race-ethnicity.
- And when the pandemic befell the world, our nation, like others, shut down for a while.
- We needed to know what was happening to our nation's population and how our condition was changing day to day.
- Yet, our federal statistical agencies, the Census Bureau included, were tailored to a mission of producing gold standard statistical data, which of course takes time, lots of time, too much time when there's a pandemic and we need contemporaneous information.
- The immediate need for data led to the inception of the biweekly national Household Pulse Survey, which interestingly sacrificed gold standard accuracy for timeliness.
- And as we all can appreciate, timeliness is an important and sometimes ignored element of data quality.
- <<pause>>.
- Now, I wanted to spend the first few minutes laying some context for the modernization of official statistics at the Census Bureau.
- It's really important to do that to understand the challenges that all federal statistical agencies face.
- Our evolving statistical universe comes with new and complex data user needs, deepening data collection challenges, the ever-present demand to do more with less, and to do it faster.
- It comes with the need for improved and more meaningful collaboration with stakeholders and partners.
- And it comes in the face of stronger computing power, and the proliferation of alternative unofficial data products, as well as new technologies.
- We at the Census Bureau are proud of our history of innovation.
- We rely on innovation to helps us achieve our mission to be the leading provider of quality data about our nation's people, places, and economy.
- Historically, we've pursued our mission by collecting data from our censuses and surveys.
- Unfortunately, while still critical in this day and age, it turns out that censuses and surveys alone are insufficient to meet policy maker, researcher and the public's modern need for data.
- We need a better approach, one that combines new and traditional data sources.
- Our modernization work hinges on combining data from different sources . . . and doing so more quickly and efficiently.
- In fact, we see modernization—and the transformation that accompanies it—as a huge opportunity.
- Modernization isn't an end to itself, but the pathway that leads to better data, more relevant data, and more timely data.
- But please note, the modernization of official statistics requires more than implementing the latest technology or the latest statistical methods.

- It requires a transformation in our thinking . . . how we think about data, its relevance in a rapidly evolving society, and how the needs of the public can be met.
- That's why we at the Census Bureau are engaged in an enterprise-level transformation and modernization initiative.
- It's a multiyear, enterprise-wide effort to evaluate and improve current processes, infrastructures, and mindsets.
- We recognize it's not enough to modernize our technology, our equipment, and our processes.
- Our critical thinking needs to be modernized too.
- What's interesting is that the pandemic—a tragedy of global proportions—helped ignite the realization that we not only could be innovative, but we could be nimble.
- It helped us see that timeliness can be critical to providing crucial, relevant data to the public.
- It also helped us appreciate so much more the value of stakeholders, partners and the public more generally.
- It was so clear to us that the degree of success we achieved in the 2020 Census very much benefitted from community-based efforts to get out the count.
- And this led to a transformation insight that's been integrated into our modernization efforts.
- And it's this—We need continuous, rich, meaningful outreach and engagements with diverse stakeholders to achieve our mission.
- In fact, we like to think of stakeholders and the public more generally as our "additional player" on the Census Bureau team.
- We see continuous external engagement as an enterprise-wide, community-of-the-whole approach.
- It promotes the value of our statistical data to communities as well as state and local government.
- And by getting communities to see and benefit from the data, we nurture and strengthen trusted messenger networks across the country, especially among communities with historically undercounted populations.
- So, as we go about our work, it's critical that we at the U.S. Census Bureau reach out to our partners, our stakeholders, data users, policymakers, and the public.
- We need to understand their needs and concerns from their perspectives.
- And then we need to respond to those needs with more relevant, accessible statistical data products.
- That's why a principal priority of mine as the director of the Census Bureau is to seek out, listen to and converse with the multitude of diverse voices across our nation.
- This is the pathway for "modernized official statistics."
- It's the way to understand and respond to the ever-evolving needs of our data users.
- Y'know, we recognize the value and importance of different perspectives of our data users.
- They help us to be innovative and creative, and ultimately produce more accurate, relevant and useful data on our nation's people, places and economy.
- Toward that end, I continue to make a concerted effort to engage stakeholders, partners and local communities across America.
- I've met with hundreds if not thousands of users of Census Bureau data across the nation.
- They include our partners . . . they include state, local, and tribal government officials, as well as community groups and businesses.
- Hey, they've even included nondata users. They provide really interesting feedback once they realize the value and the power of our data for things like community needs assessments, economic development, and public health and disaster planning (add your examples).

- In all this, our goal is to organize and engage a widening diversity of talent and stakeholders to achieving our transformation and modernization goals strategically, purposefully and with a commitment to inclusiveness.
- In a real sense, we're using the principals of community based participatory research to motivate innovation in our statistical data products.
- We're combining technological advances and rigorous research with innovation and creativity and using that to guide our modernized official statistics.
- <<pause>>.
- So, what does our transformation and modernization approach look like?
- Let's talk about that.
- We're thinking about statistical data products in new ways.
- We envision the Census Bureau as being a statistical data producer operating as a data-centric, consolidated business enterprise.
- Under this transformation, our operation encapsulates three basic processes: data ingestion, data collection and processing, and dissemination of statistical products.
- Now, historically we've been organized around a set of programmatic directorates.
- Each essentially produced independent data products that pretty much didn't much connect with each other.
- Also, the siloed programmatic areas focused on a solicitation framework for generating data products.
- We conduct a census or survey by soliciting information from people or businesses.
- Then we process the collected data.
- And then we tabulate it and disseminate it, and sometimes we create a microdata product featuring the survey responses.
- Well, our new vision of a modernized statistical agency flips this framework on its head
- Instead, we focus on pooling and linking data from all sources into a data lake—surveys, censuses, administrative data, whatever.
- This approach improves both efficiency, capacity, relevance and utility of the data we can offer the public.
- In fact, by virtue of linking the data in the data lake, this new approach allows us to address important policy questions well beyond what a single survey or census could address.
- And let's not forget external engagement.
- You add to this approach the diverse community perspectives and diverse data needs, and that leads to new statistical data products that are more relevant to the public.
- Now, this new vision of a 21st century statistical agency requires the full use of newest technology and statistical methods.
- We are in the middle of retooling our factory, so to speak, and I'll talk about those new systems in a bit.
- But first, let me give you a taste of two modernized official statistics that have been developed under this new approach.
- The first is the story of our new populations estimates methodology which started with our Vintage 2021 Estimates.
- OK, so the Census Bureau produces annual population and housing estimates using a target date of July 1 for each year.
- We do this to satisfy a legal mandate, but they're used heavily in both research and policy implementation.

- They're produced for over 80,000 geographic areas in the United States and Puerto Rico.
- This includes population estimates by age, sex, race and Hispanic origin down to the county level.
- In fact, RTI and other research organizations rely on these annual estimates to calibrate your population survey weights via post stratification.
- We calculate our annual population estimates anew each year as a time series.
- They start with the most recent decennial census, in this case 2020, and extend up to and including the vintage year.
- So each year we create a new series.
- Now, the decennial census counts have typically served as the base—or starting point—for our population estimates.
- And what we do for each year hence, is use vital records and other administrative data to add births, subtract deaths and fold in net migration to create data points for each year after the census.
- But when it was time to develop the Vintage 2021 population estimates, we faced a big problem.
- Because of delays caused by the pandemic, the 2020 Census counts that we needed were simply unavailable.
- So, our demographers and statisticians got together to develop an alternative base.
- We developed what is now called a "blended base" approach.
- In place of the usual decennial counts, we combined data from three sources to create the 2020 base population.
- The sources were: One, the Vintage 2020 estimates series which started with 2010 Census.
- Two, some limited data from the 2020 redistricting file which used some actual 2020 census counts.
- And three, the age and sex distributions from our independent 2020 Demographic Analysis.
- Now here's the really interesting part.
- We know from our Post Enumeration Survey that there were undercounts of specific groups of people.
- Well, we found that by using this blended approach, the undercount of young children ages 0 to 4 was somewhat mitigated.
- This new approach, borne out of necessity, revealed a pathway to exploring the other adjustments to the base which might improve known undercounts or other issues with census data.
- As a result of this incredible challenge, data equity had been better served.
- Remember, simply using decennial counts as the base will bake into all population estimates the error profiles of that census.
- Now, we're really proud of the job we did on the 2020 decennial in the midst of a pandemic.
- And we all know that no census or survey is ever perfect.
- So to develop a new approach that helped addressed known limitations of the data was a rather big deal.
- In fact, our research is continuing to explore other ways to improve the base population count for this program.
- <<pause>>.
- Now let's turn to the second example . . . our new Business Formation Statistics Series.
- These economic statistics measure the formation of new businesses using applications for employer identification numbers called EINs.
- They're used to create projections of downstream businesses with paid employees.

- Now, for years we'd been receiving weekly data from IRS to update our business register—literally a frame of all businesses.
- We used the IRS data to create quarterly Business Formation Statistics.
- Using these data, we developed prediction models for which applications for EINs would go on to become businesses.
- Obviously, not all EIN applications turn into businesses.
- In any case, then we took this approach to the next level of timeliness.
- Our staff started producing weekly and monthly formation statistics during the pandemic, instead of just quarterly.
- If you recall, during the pandemic, it was really important from an economic policy perspective to know what was going on in the business world.
- These new weekly and monthly estimates turned out to be leading indicators of economic well-being.
- Again, they were developed at a time when our nation needed more contemporaneous data to understand and deal with the crisis of the pandemic.
- As a result, we now have a new, more frequent standard product that provides the earliest indications of entrepreneurism, of employment levels, and of jobs.
- Our Business Formation Statistics are now published monthly and include estimates at the national, regional, and state level.
- And we are looking to expand that to other geographies.
- Plus, the National data is also available by industry.
- This program has proven so popular with the business community that we're working with OMB to establish the monthly series as a Principal Federal Economic Indicator.
- That would be a big deal.
- <<pause>>.
- OK, so let's turn to the architecture of our single-enterprise business ecosystem.
- We're working on series of four foundational systems under this transformation and modernization effort.
- The first system is our Enterprise Data Lake (EDL).
- It's the repository that houses all of our data, including but not limited to all of our master frames like the business registry, the housing unit address file, all our geospatial data, the administrative records we receive from IRS, SSA and others, and of course our decennial data.
- It's built in the cloud to allow both scalability and use of modern processing tools.
- The data lake will provide both operational and analytical processing to create a more seamless flow between research and ongoing operations.
- The second system ingests and collects data from all sources . . . be they censuses, surveys, administrative records, or whatever.
- We call it the Data Ingest and Collection for the Enterprise system, fondly referred to by its acronym, DICE.
- DICE is the vehicle by which data flow into the data lake.
- But it also provides a platform for one-stop shopping for census and survey collection applications in any mode.
- It will be able to accommodate our business or population censuses just as well as well as an American Community Survey or any of our 130 other surveys.
- In fact, it's already being used for our Annual Integrated Economic Survey of businesses.
- The third system is a platform for data dissemination.

- Of course, it has a name and acronym.
- The name is the Center for Enterprise Dissemination Services and Consumer Innovation.
- We call it by its acronym, CEDSCI.
- This dissemination platform will provide products much more quickly than our current siloed processes.
- It will also facilitate the development of new data products based on the linkages formed across data sets in the data lake.
- The last system in our modernized Census Bureau is called Frames.
- Frames represents the foundational datasets that will underpin the Census Bureau's future work.
- This system brings together our Geospatial Frame, Business Frame, Job Frame, and Demographic Frame.
- Our vision is to create enterprise-wide frames that are linkable in nature, agile in structure, and accessible for production or research.
- Each frame will be linkable to other frames or datasets.
  - For example, both businesses and people will be linkable to jobs.
  - And all linked data will be linkable to the corresponding geospatial data.
- The linkable Frames will enormously expand our ability to explore changes and trends affecting the nation's population, our economy, and local communities.
- Moreover, from a data equity lens, the linked frames data may be helpful in reducing data gaps for the hard-to-enumerate population.
- Naturally, this is no "silver bullet," but it may help to mitigate the differential undercount as well as address item missing data.
- The ultimate promise of Frames is to maximize the use of existing linked data to address user needs, thereby reducing the need and burden of data solicitation.
- It will allow us to focus on where the gaps in data are and lead to a more effective and efficient targeting of data collection to where we need the data most.
- This means more resources focused on historically undercounted and under-represented populations and communities.
- So, those are the four systems that make up our transformation and modernization effort.
- Keep in mind we're not undertaking transformation and modernization for their own sake.
- We are looking to rethink, reinvigorate, reimagine what it means to live our mission… to be a leading provider of data on our nation's people, places and economy.
- And that means we need to transcend the traditional process of soliciting data, processing it, tabulating it, and releasing it.
- We want to start with the public . . . with our data users . . . with our stakeholders, and even with our potential data users who don't yet realize the value of our data.
- They, including YOU, have information needs that often can't be met with a dump of data from a given survey or census.
- But if we can bring together economic, demographic, geospatial, health, sociodemographic, education, and other data sources, that combination offers much potential to address needs that up to now have not been addressed.
- Our vision of modernized official statistics involves new data products, new statistics, more timely statistics, new and easier ways of accessibility . . . all driven by user needs.
- Ultimately, Modernization is putting people like you as the central of official statistical agencies to generate culturally, timely statistical data that will benefit all of us in our everyday lives and businesses for years to come.
- Thank you.