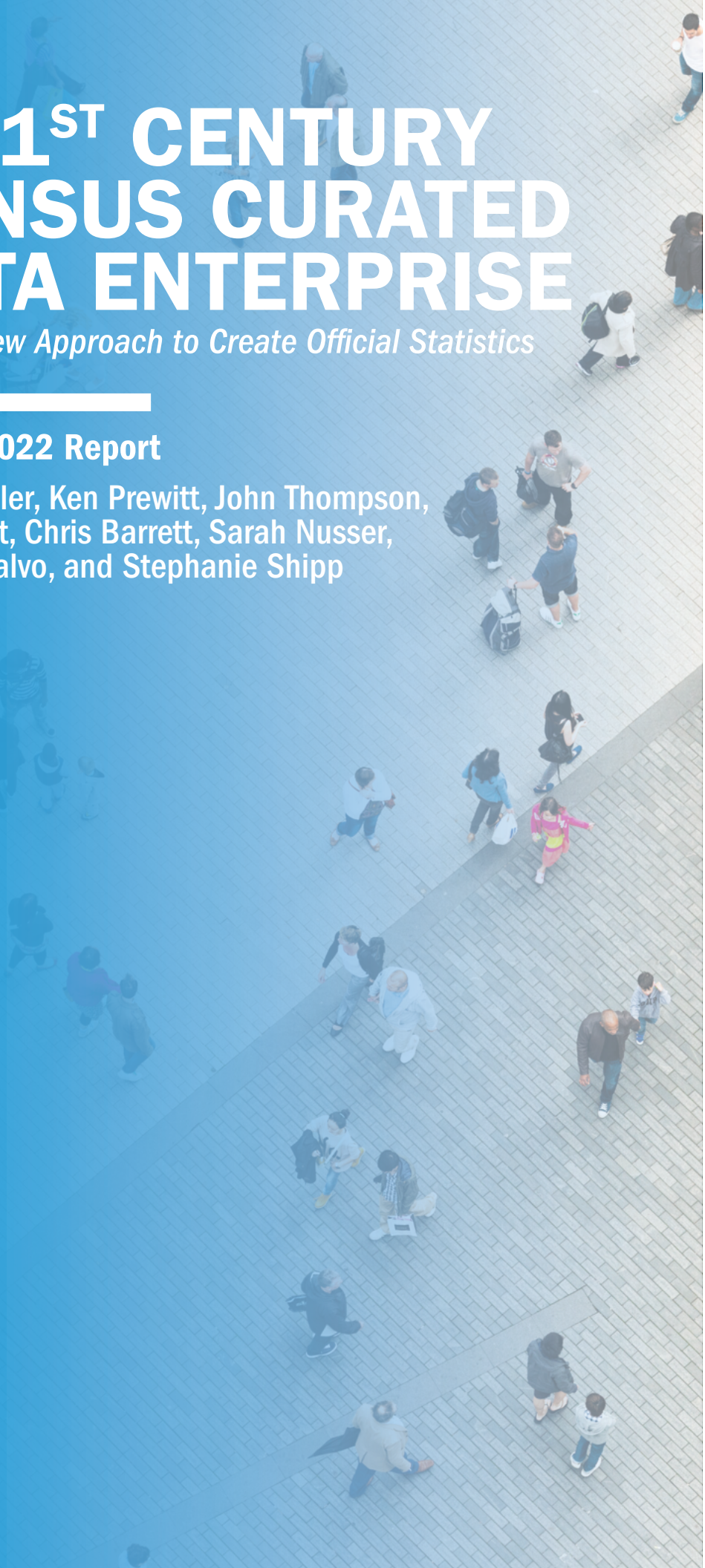
A 21ST CENTURY CENSUS CURATED DATA ENTERPRISE

A Bold New Approach to Create Official Statistics

Spring 2022 Report

Sallie Keller, Ken Prewitt, John Thompson, Steve Jost, Chris Barrett, Sarah Nusser, Joseph Salvo, and Stephanie Shipp



UVA | BIOCOMPLEXITY INSTITUTE

REPORT CONTENTS

Spring 2022

EXECUTIVE SUMMARY

The Framers of our Constitution laid the cornerstone for the federal statistical system by requiring a decennial census as the basis for apportioning the House of Representatives. Over the next 233 years, the U.S. Census Bureau's mission has expanded to serve as the nation's leading provider of comprehensive, quality data about its people and the economy. To achieve this mission, the Census Bureau has made a steady set of innovations to modernize its data collection and dissemination. Here we share a concept and initial steps forward for their next innovation, the creation of a Curated Data Enterprise.

The Curated Data Enterprise is both an infrastructure and a continuous evolving ambition to empower and enable Census Bureau scientists and their data users to progress from a focus on individual data elements or surveys items to one focused on the purpose and use of the information. This can result in new and better measures of America's people, places, and the economy. It is a new vison to exploit multiple data sources across many sample surveys, censuses, tribal, federal, state, and local administrative data, as well as private-sector data, to produce more robust, granular, timelier, and comprehensive measures of demographic changes, social trends, and economic activity.

Today the Census Bureau is innovating its processes to take advantage of new data sources and data science computing innovations and to adapt to declining survey response rates that are challenging the public and private survey world. The Census Bureau is also exploring additional means of producing data to address shortfalls that are becoming increasingly inherent with surveys. It has stood up a team representing a cross-section of its demographic, geographic, and economic programs to break down siloed activities and build a new enterprise infrastructure. One component of this infrastructure is the "Frames Program" that will link four key components of the internal architecture, Geospatial, Business, Jobs, and Demographic frames, into shared enterprise resources that will bring together, serve, and support all Census Bureau surveys and make appropriate use of administrative and third-party data. This resource will form the foundation for the Curated Data Enterprise, creating a scaffold to hold and link massive amounts of public and private sector data.

With support from the Sloan Foundation and the Census Bureau, we have shared the Curated Data Enterprise concepts with a diverse set of Census Bureau stakeholders, including researchers, economic developers, economists and business leaders, advocates, public policy analysts, and applied demographers. Here we report on their expert viewpoints, including support for the concept, concerns, technical gaps, research challenges, key partnerships that will be needed, and unforeseen opportunities.

Our conclusion is that the innovations offered by the Curated Data Enterprise represent a necessary evolution beyond the survey-only model that has reached scientific and practical limits in an era of increasing demand for more data, more often, and more urgently. It holds the promise of producing more timely, robust, and accurate findings and to more fully reflect the diversity of the nation's racial and ethnic composition.

AN EVOLVED MISSION

Our Framers placed a decennial census in the U.S. Constitution, its key purpose being to redistribute the members of the House of Representatives by state every ten years. Because the 13 states grew at different rates and new states would be formed at different times, it was critical that the state-by-state population shifts would be accurately recorded on a decennial basis. The decennial census was an immediate success. Soon congress thought if it could reapportion accurately, why not add additional questions, for example, about economic activity. And so, the questionnaire expanded, census by census, steadily, fairly, and always working to improve quality.

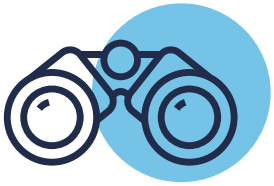
As we know, two and a half centuries later, the Census Bureau is the indispensable source of statistical data and serves as the nation's leading provider of quality data about its people, places, and economy. Census data are a gold standard, foundational to all official statistics. They are an extraordinary public good, striving to make data equally available to all including individuals, businesses, and tribal, federal, state, and local agencies, across topics from economics to education, agriculture to health, housing to transportation. The Census Bureau annually conducts more than 130 surveys and censuses. This two-centuries-plus success story has earned the nation's trust for high quality, consistent, and reliable data products. It has achieved this reputation by adhering to strict scientific standards in its operations.

U.S. CENSUS BUREAU MISSION STATEMENT

The Census Bureau's mission is to serve as the nation's leading provider of quality data about its people and economy.

U.S. CENSUS BUREAU GOAL STATEMENT

The Census Bureau's goal is to provide the best mix of timeliness, relevancy, quality, and cost for the data collected and services provided.



Staffed by some of the nation's most accomplished demographers, economists, statisticians, geographers, and survey methodologists, the U.S. Census Bureau is respected around the world for its scope and scale, its rigorous, reviewable, and repeatable processes, its commitment to minimize survey error, its attention to the validity of geographic, temporal, and demographic comparisons, its documentation of coverage and coherence of estimates from different sources, its effort to assure data availability for all areas and subgroups from the smallest of rural communities to the largest cities, and its attention to demographic changes in the nation's racial and ethnic composition.

"Thanks to Katherine Wallman, the 2000 Census replaced a misleading race/ethnicity question with one that allows "mark one or more." This 2000 option is now revealing the steady growth of multiple race marriages and, of course, multiple race offspring. Demographic diversity tells us that the country is on the move, and so, therefore, must its census."

— **KEN PREWITT**
Census Director
October 1998-January 2001

An Abundance of Data

FEDERAL AGENCIES USE CENSUS DATA TO ALLOCATE OVER

$1.5 Trillion

TO LOCAL, STATE, AND TRIBAL GOVERNMENTS EVERY YEAR AS THEY FUND

Neighborhood Improvements

Public Health

Education

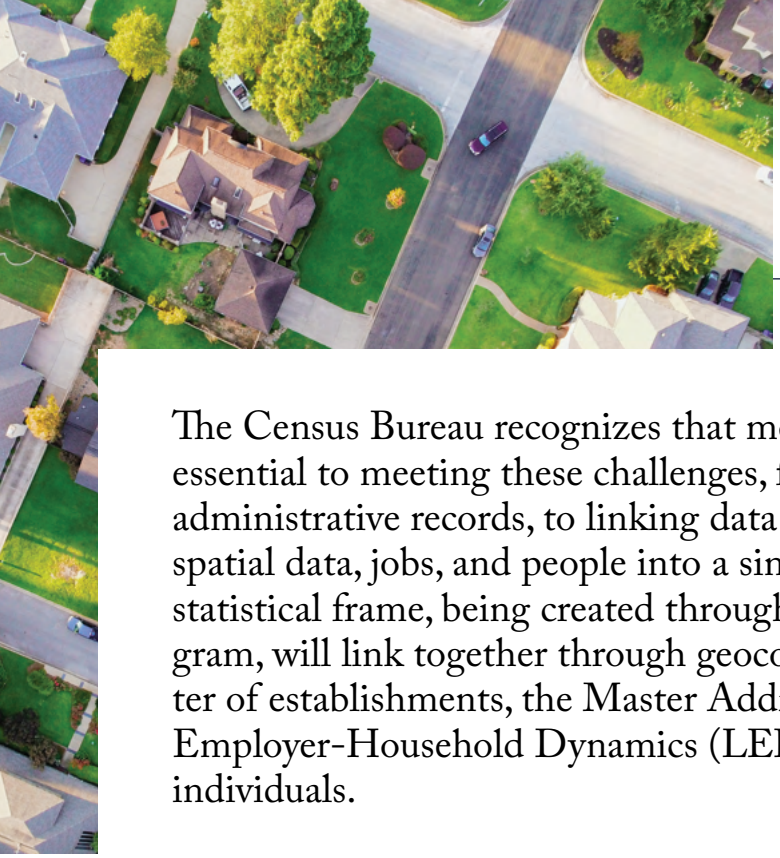
Transportation

AND MUCH MORE!

The Census Bureau's mission leverages an array of surveys. Beyond the familiar decennial census conducted every 10 years, there is the annual American Community Survey (ACS). It is the premier source for information about America's changing population, housing, and workforce. There is also a Census of Governments, which identifies the scope and nature of the nation's state and local government sector including public finance and public employment and classifications; an Economic Census every five years that provides detailed information on employers and businesses, including detailed data by industry and geography; as well as 100 other surveys. In the economic sphere the Census Bureau publishes 18 monthly or quarterly key economic indicators.

Collectively, these data shape local government planning decisions about community services: where to provide services for the elderly, build new roads and schools, locate job training centers, and prepare for the challenges posed by climate change. Federal agencies use census-derived data to allocate more than $1.5 trillion to local, state, and tribal governments each year as they fund neighborhood improvements, public health, education, transportation and much more.

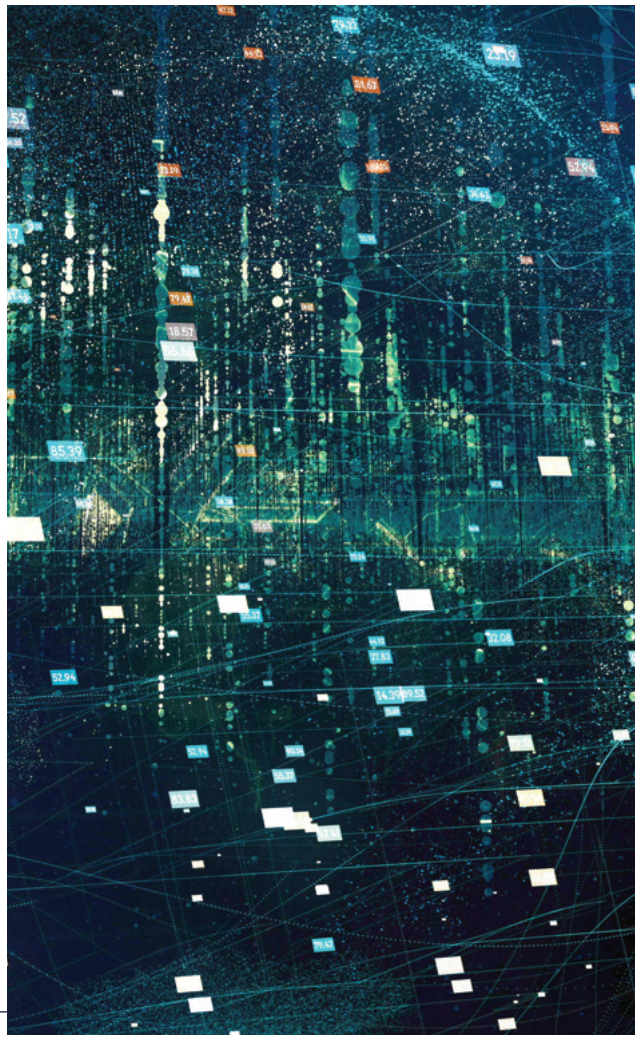
While the Census Bureau has for over a century been a place for scientific innovation in data collection, survey methods, and data dissemination, today it confronts increased demand for more data and more timely data delivery. The Covid-19 pandemic also accelerated the Census Bureau's pace of innovation. There is relentless pressure to make improvements in official surveys. High on that list are the persistent differential undercounts of race and ethnic subgroups, and of young children. Public trust in census statistics is, of course, linked to its confidence that the underlying data accurately and equitably reflect the full diversity of the nation. Fresh attention to evidenced-based policy drives demand for improvements on equitable measures.



The Census Bureau recognizes that modernizing its data collection and processes is essential to meeting these challenges, from expanding the use of third-party data and administrative records, to linking data frames for businesses, housing units and geospatial data, jobs, and people into a single enterprise resource. This new universal statistical frame, being created through the Census Bureau's Enterprise Frames Program, will link together through geocodes and other linking keys the Business Register of establishments, the Master Address File of housing units, the Longitudinal Employer-Household Dynamics (LEHD) jobs frame, and a demographic frame of individuals.

"This is a very exciting time to be working at the Census Bureau and to be part of the transformations further enabling the ways it is achieving its mission as a statistical agency. I am enthusiastic about how the disciplines of survey methodology, demographics, economics, data science, geography, and computation are being brought together to measure the historic changes in our society and economy."

— **JOHN THOMPSON**
Census Director
August 2013 – June 2017



Keeping Pace With New Demands

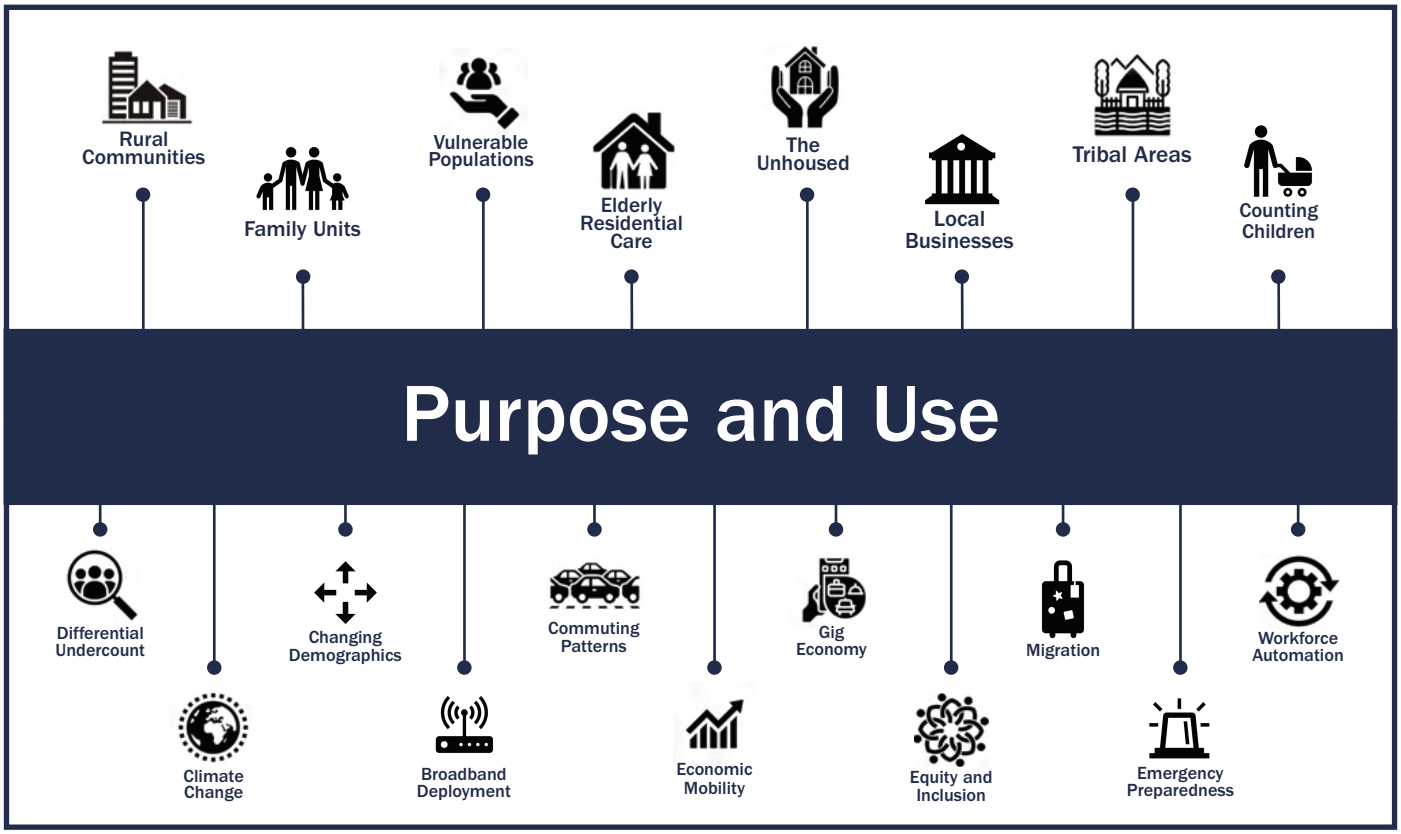
"As we advance into the 21st century, we are experiencing increased demand for our data, struggling with challenges to traditional data collection methods, and exploring rich new data sources and tools that can revolutionize what we do and how we do it. Our success critically depends on our ability to seize the opportunities in front of us to deliver statistical products that address the increasingly complex and diverse needs of our users."

— **RON JARMIN**
Census Deputy Director
Acting Director
2021-2022

Looking ahead, the Census Bureau is moving to increase operational efficiencies, not only exploring how to harvest new data sources but deploying survey innovations such as the Household Pulse and Small Business Pulse Surveys and implementing new statistical methods, all aimed at improving data quality and availability. These greater analytical capabilities and expectations will succeed only if they are responsive to stakeholders, especially in such matters as the use of public and private administrative data and third-party data. Exploiting these additional sources can help deliver official data faster, enhance efficiency, and improve accuracy while reducing respondent burden. The Census Bureau is in the early stages of a comprehensive transformation of all its survey operations.

While always mission oriented, modernization should be especially focused on the purpose and uses of the new data products. This focused approach aims to be nimbler, increasing the ability to address emerging topics, adapt more quickly to external challenges such as the recent pandemic, as well as compelling demands for new data.

Essential to this modernization are reforms designed to integrate or connect data across administrative records, surveys, and other public and private data sources into a Curated Data Enterprise (CDE). The linking of now disconnected survey frames on businesses, housing units, jobs, and people into a unified resource forms the foundation of the CDE concept. To that end, the CDE is best described as both an infrastructure and a continuously evolving ambition to empower and enable Census Bureau scientists and their data users to develop together new measures of the nation's people, places, and economy.



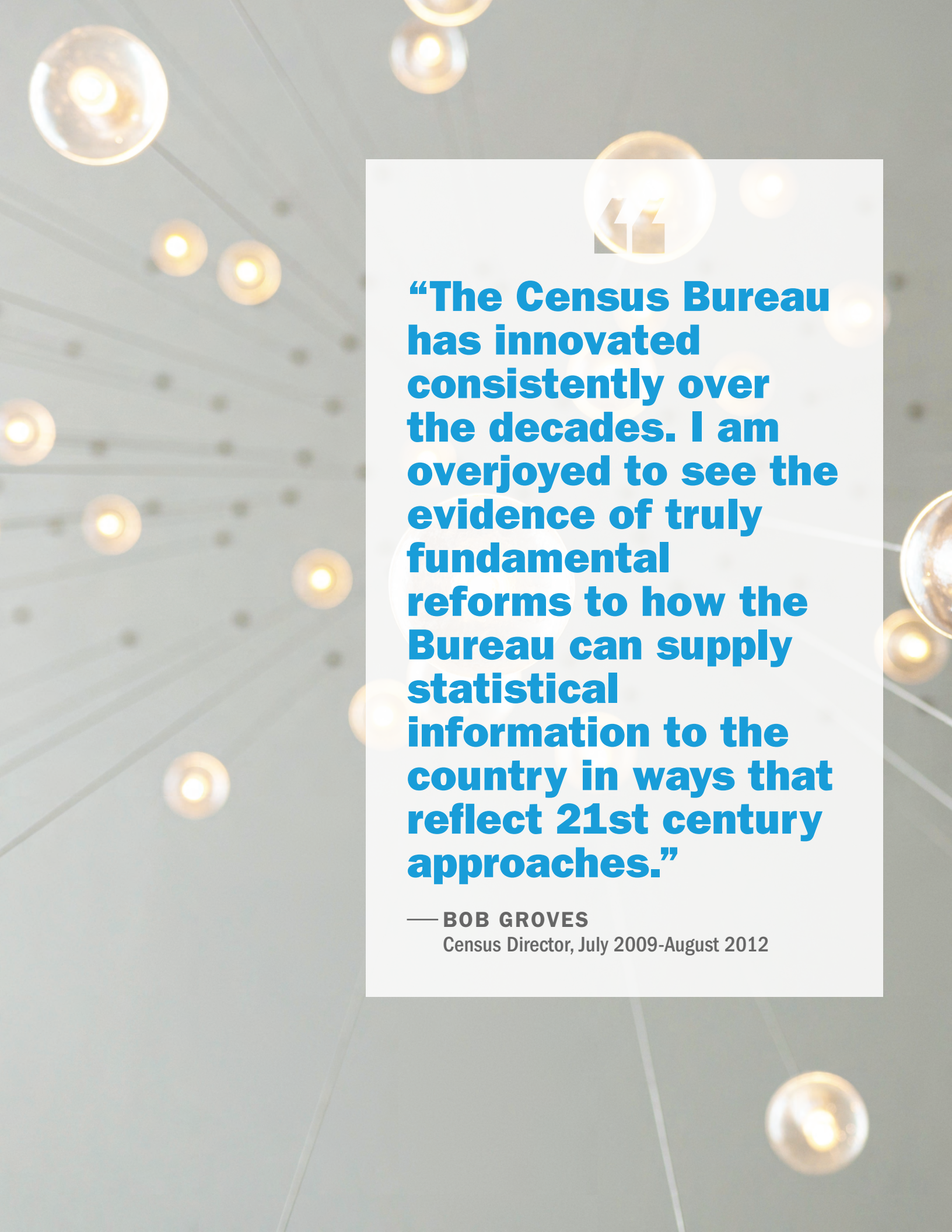
BEYOND A SURVEY-ALONE MODEL

The universal statistical frame that links the Census Bureau's sampling frames for demographic and economic surveys forms a scaffold for the Curated Data Enterprise (CDE) to hold and integrate massive amounts of public and private data. This involves multiple-sourced data products that, taken together, can build a new, more accessible, and reliable information platform for the nation.

Focusing the integration of all data sources on relevant purposes and uses should should provide more timely and more robust data for smaller geographies, rural areas, tribal lands, and subpopulation groups that are less well served by current Census Bureau data products, including decennial census data files.

Linking across time, and surveys, supplemented with third-party and administrative data can help improve the coverage and completeness of Census Bureau data, innovating beyond a survey-alone model that has reached its scientific and practical limits. The CDE's multi-source data innovation offers a better option to help reduce differential undercounts of children as well as race and ethnic populations than a survey or census enumeration alone.

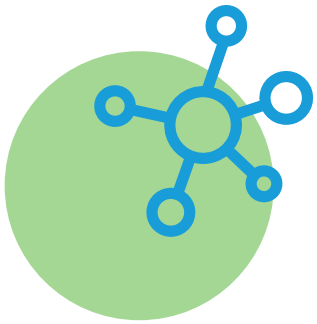
The CDE can be augmented over the decade with government and private sector data sources. For example, starting with the 2020 decennial census records, we can populate the curated platform with rich social and economic data at a scale the country has not previously imagined across a wider array of geographic levels.



“The Census Bureau has innovated consistently over the decades. I am overjoyed to see the evidence of truly fundamental reforms to how the Bureau can supply statistical information to the country in ways that reflect 21st century approaches.”

— **BOB GROVES**
Census Director, July 2009-August 2012

Emergence of the Curated Data Entreprise Concept



The impetus for the Curated Data Enterprise (CDE) started in 2016, focusing on the decennial census and a query that former Census Director John Thompson made to the JASON advisory group[1]. He asked them to re-envision the 2030 Census starting from a blank sheet of paper. In their deliberations, JASON embraced the expectation that by 2030 the Decennial Census and the ACS would be making increased use of federal, state, and local administrative data and other public and private data. JASON recommended an approach that would be a paradigm shift from a housing unit-centric Decennial Census to one focused on collecting information about individuals and then placing them geographically where they live. This suggestion aligned with expected changes in demographics (smaller households) and decreasing mobility (less than one-tenth of people moved in 2019). Their recommendations pointed the Census Bureau beyond decennial to reconsider the American Community Survey (ACS) and a wholistic data ecosystem[2].

One year later, an experienced group of current and former Census Directors and stakeholders came together to share their expertise about the need for significant change across federal statistics that would lead to innovative, practical, and cost-effective processes for the Census 2030 and beyond (See table on the following page). Their concerns focused on the Nation's lack of ability to meet these future objectives. This meeting concluded with a commitment to work together and put forth a vision for the Census Bureau and its stakeholders to embrace as we look toward 2030. The original premise emerging from this group was:

Structure an essential-data only Decennial Census for apportionment and an enhanced American Community Survey (ACS) for redistricting, Voting Rights Act, and related social justice programs, annual allocation now of approximately $1.5 trillion in federal grants to cities and states, benchmarking federal and private-sector surveys and related data products, and such other applications that Decennial Census data have historically facilitated.

[1] https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/final-analysis/alternative-futures-2030-census.html

[2] https://www.science.org/content/article/researchers-think-they-ve-found-much-better-way-conduct-2030-us-census

The initial premise was explored with support from the Alfred P. Sloan Foundation, including determining if constitutional or legal barriers exist to inhibit the proposed shift. None were found, and the group, now formed as an Advisory Panel, moved forward, developing a plan for unrolling this novel approach with stakeholders.

Diverse Expertise to Shape a New Vision

Initial Working Group	
Co-leads	**Sallie Keller**, Distinguished Professor of Biocomplexity, University of Virginia **Ken Prewitt**, Carnegie Professor of Public Affairs, Special Advisor to the President, Columbia University (Census Bureau Director, 1998-2001)
Census Bureau Directors	**Vince Barabba**, Chairman and Co-founder of Market Insight Corporation, (Census Bureau director, 1979-1981) **Robert Groves**, Executive Vice President and Provost, and Gerard J. Campbell, S.J. Professor in the Math and Statistics and Sociology Departments, Georgetown University (Census Bureau director, 2009-2012) **Steve Murdock**, Allyn and Gladys Cline Professor Emeritus of Sociology, Rice University, (Census Bureau director, 2008-2009) **John Thompson**, Distinguished Institute Fellow, Biocomplexity Institute, University of Virginia, (Census Bureau director, 2013-2017) **Ron Jarmin** (ex-officio), Deputy Director and Chief Operating Officer. U.S. Census Bureau
Advocates	**Arturo Vargas**, Executive Director, National Association of Latino Elected and Appointed Officials **Vanita Gupta**, United States Associate Attorney General, U.S. Department of Justice, former President and Chief Executive Officer of the Leadership Conference on Civil and Human Rights
Former Federal Officials	**Cathie Woteki**, Iowa State University and Visiting Distinguished Institute Professor, Biocomplexity Institute, University of Virginia, former Under Secretary for USDA's Research, Education, and Economics **Stephanie Shipp**, Professor, Biocomplexity Institute, University of Virginia, former Senior Executive Service, National Institutes of Standards and Technology

Starting in late 2019, while some planning was interrupted by the 2020 Census activity the initial working group was impressed with and influenced by two substantial innovations by the Census Bureau. First, the Bureau rapidly implemented new data collection through the Household and Business Pulse Surveys to respond to social and economic measurement needs of the Covid-19 pandemic. Second, the Census Bureau launched a set of enterprise-wide activities, including creating a linked universal statistical frame.

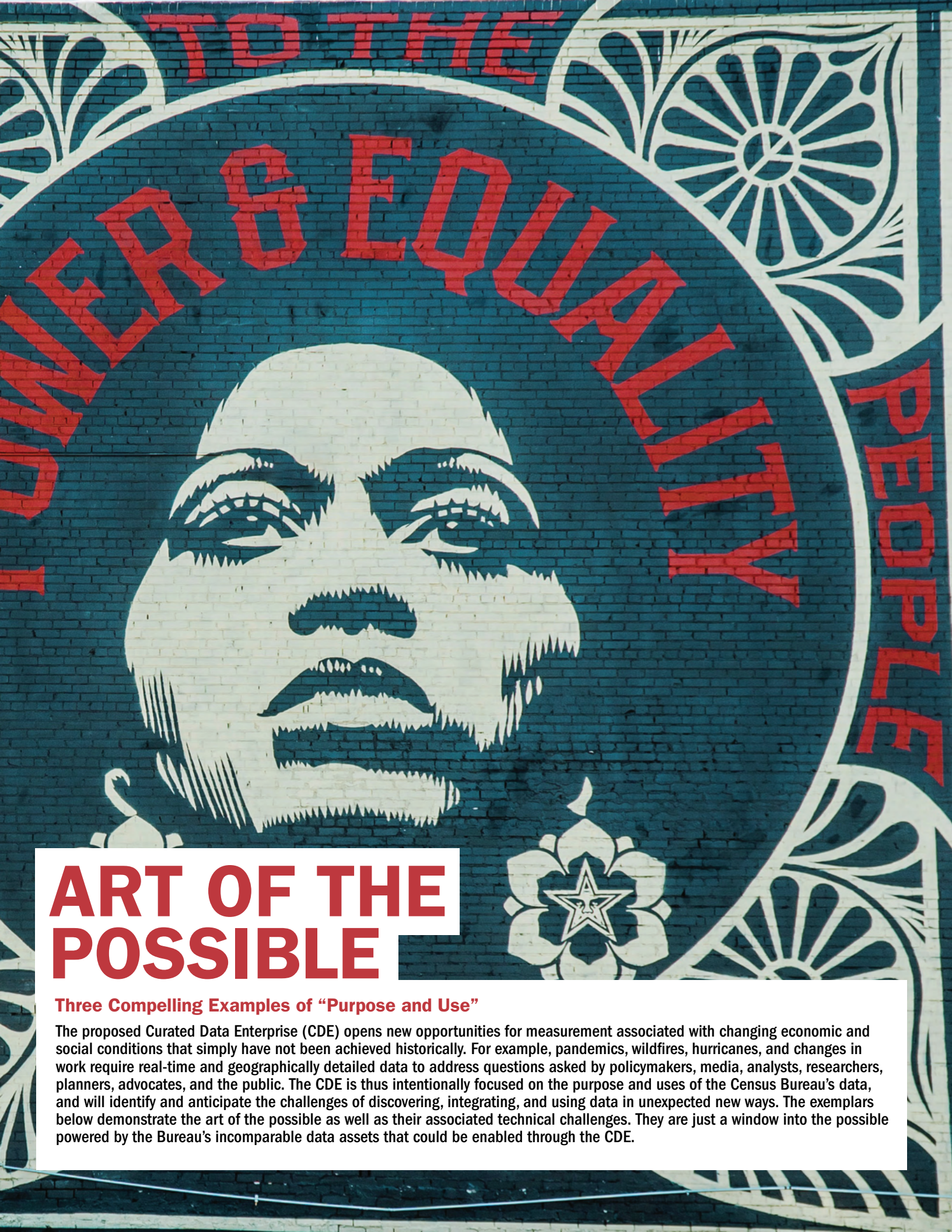
Household & Business Pulse Surveys

Universal Statistical Frame

These innovations changed the shape of the discussion. It was realized that the linked universal statistical frame, coupled with the expanded use of public and private administrative data, current and new survey data, and third-party data, could form the foundation for a new data infrastructure. Thus, the premise moved well beyond its initial scope on the decennial census to a much more expansive and powerful transformation across the agency and its 130 surveys. The CDE, a new proposed infrastructure encompassing the complete set of activities the Census Bureau conducts to develop trusted metrics on people and the economy emerged out of the discussions lead by the University of Virginia team authoring this report. This new approach amplifies the use of multiple data sources, from sample surveys, censuses, and other official administrative and third-party data, with the goal of providing more robust, timely, and comprehensive measures when combined.

In 2021, the Census Bureau created a Memorandum of Understanding with the University of Virginia to identify and characterize the stakeholder communities and elicit their feedback on the challenges and opportunities presented by the CDE. This report summarizes what has been learned thus far.

Based on the findings summarized here and shaped through the extensive solicitation of additional stakeholder input, a research implementation plan will be drafted. This will include areas of opportunity, technical challenges, and barriers to innovation for the Census Bureau over the next decade. The recommended research agenda will allow the Census Bureau to partner internally and externally with researchers and other stakeholders to fully achieve the ambition of the CDE.



ART OF THE POSSIBLE

Three Compelling Examples of "Purpose and Use"

The proposed Curated Data Enterprise (CDE) opens new opportunities for measurement associated with changing economic and social conditions that simply have not been achieved historically. For example, pandemics, wildfires, hurricanes, and changes in work require real-time and geographically detailed data to address questions asked by policymakers, media, analysts, researchers, planners, advocates, and the public. The CDE is thus intentionally focused on the purpose and uses of the Census Bureau's data, and will identify and anticipate the challenges of discovering, integrating, and using data in unexpected new ways. The exemplars below demonstrate the art of the possible as well as their associated technical challenges. They are just a window into the possible powered by the Bureau's incomparable data assets that could be enabled through the CDE.

Exemplar 1: Domestic Migration Patterns – Why, where, and how often do people move?

Most demographers define domestic migrants as movement across "communities of residence," as opposed to movers, who move more locally within those communities. As a practical matter, given the available data, county boundaries are generally used to distinguish these two types of movement. Most moves from one year to the next occur within counties, with migrants (those that move across a county boundary) just a subset of all persons who move. ***Migration is the most volatile component of population change and the most difficult to estimate in real time, especially for small areas and groups.***

The CDE could be an engine for better migration and mobility data, leading to improved planning for labor market changes, emergency response, and state revenue projections. State and local economic planners and demographers need information about the interplay between migration and local labor markets, especially given recent dramatic changes in the nature of work during the pandemic. Yet current measures do not keep pace with these changes.

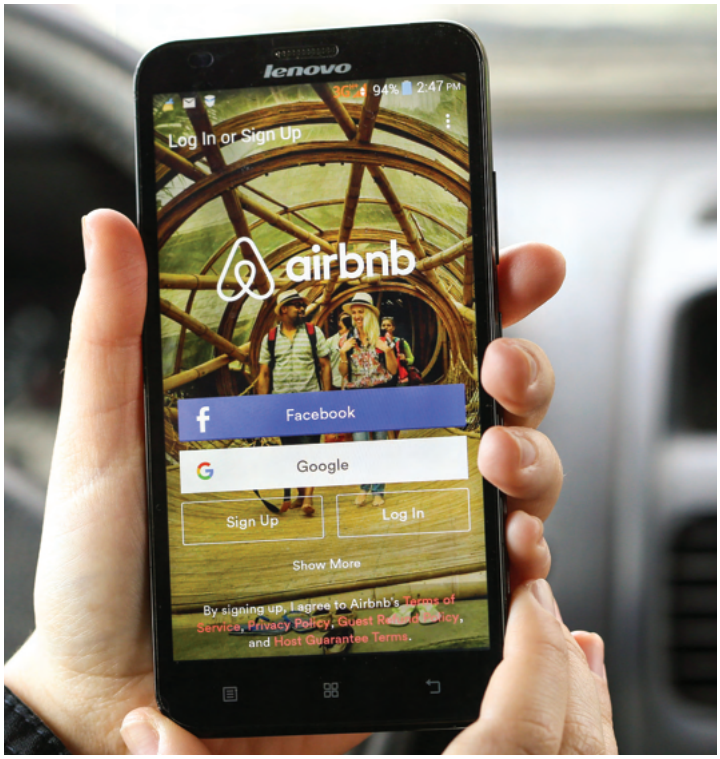
Any area of the country affected by wildfires, extreme weather, flooding, and droughts associated with longer-term watershed issues, experience changes in their patterns of migration. These events precipitate movements that need to be monitored close to real time, so local and state planners can react quickly. When disaster strikes, emergency preparedness plans are needed, informed by reliable data on the size of affected populations, the composition of households, their access to travel or transit, to assess the potential for population displacement. State and local governments require population data for revenue projections, the distribution of resources, the identification of needs, and the establishment of priorities and strategies to address those needs.

To be most useful, planners need access to methods to measure migration close to real time to discern the difference between temporary versus permanent shifts in population. The CDE could support the integration of relevant data sources including federal surveys such as the Current Population Survey (CPS) and the American Community Survey (ACS), combined with administrative data such as tax data and address data, as well as third-party data, e.g., new home mortgage applications, real estate inquiries/transactions, private company mover records, cellphone detailed records, and utility company accounts. Technical challenges to overcome in the use of these data involve creating linkages between survey data, which have a substantial lag but are still valuable to gauge trends to shorter-term indicators, all for small geographic areas. Locally, migration patterns need to be studied as a way of addressing equity issues, for example involving access to housing, health care, and broadband, and displacement of populations based on race, ethnicity, and income. These observations not only rely on estimates of the total population, but the integral role played by population estimates and projections by age and sex, which undergirds measures of well-being, especially rates related to public health issues and poverty measures.

Exemplar 2: The Role of Gig Employment in the Post-Pandemic World of Work

Definitive estimates of gig employment, also known as alternative employment arrangements, vary widely depending on the definitions and sources used. Estimates range from a low of 15 million to a high of 55 million workers. As these estimates demonstrate, refining the definition of gig employment is challenging. Definitions include some mix of individuals who are paid by a temporary help agency or contract company or are on-call workers or independent contractors. The Bureau f Labor Statistics' Current Population Survey Contingent Worker Supplement (CWS) provides the lower bound for estimates (15 million). These data show little growth in alternative employment arrangements since 2005, likely because the survey's focus is on collecting data about an individual's main job. Other surveys and administrative data however have established that gig work is frequently undertaken to supplement income from a traditional job. As a result, official employment data greatly understate the level of gig employment and its increasing prevalence, especially during the pandemic. This is likely due to the inability of single-source surveys to provide data in anything close to real-time. To achieve a better measure, we must make use of multiple data sets derived from governmental and non-governmental surveys, and administrative data, especially if we are to fully understand the size and patterns of change in gig employment.

In this domain, we do not need to build from scratch, rather it can be informed by multiple studies taking an integrative approach. ***The CDE could facilitate research leading to a common definition and comparable reporting on the gig workforce.*** In this domain, we do not need to build from scratch; rather, it can be informed by multiple studies taking an integrative approach. These studies combine many data sources such as the Current Population Survey, General Social Survey, Internal Revenue Service Administrative Data, and other data sources.



No single data source will suffice to study the rise in gig employment and its many facets. The large number and types of data needed to assess gig employment are ideally suited for the CDE, a data enterprise capable of capturing purposes for data use, as well as the individual data assets themselves.

Exemplar 3: Examining the State of Nursing and Residential Care Facilities for Older Populations

The COVID-19 pandemic highlighted the need for an integrated view of nursing home and residential care to better understand the interactions across residents, workers, and business owners. Multiple stakeholders want integrated data about nursing and residential care facilities. Federal and state legislators and regulators require information about the regulatory, safety, and financial status of care facilities to ensure adequate supply for a rapidly growing and increasingly diverse population. There is also concern about low pay, high turnover of staff, and the challenging work environment of the workers. Finally, older populations and their families want to understand the landscape about the type, costs, and quality of nursing homes and residential care facilities.

The Census Bureau, Bureau of Labor Statistics, and agencies under the Department of Health and Human Services (HHS) collect numerous surveys that measure the characteristics of the nursing homes and residential care facilities from business, employee, and resident perspectives. States regulate and collect administrative data for nursing homes and residential care facilities. One challenge is that these data sources are collected separately. Even within the Census Bureau, these facilities are recorded in two independent databases one supporting the demographic programs, and a second supporting economic programs, using different classification schemes. Yet the characteristics recorded in both are very different and complementary.

The Kaiser Family Foundation has brought together the state nursing home data in one place. There does not appear to be a similar source for the residential care facilities, requiring the collection of these data state by state. This Nursing and Residential Care Facilities example illuminates the complexity of sources and the value of the infrastructure requirements for the CDE. ***There is an abundance of survey, administrative, and procedural data, research expertise to guide the implementation, ethical questions to address throughout, and accessibility challenges to ensure that all stakeholders have access to the data products they need.***

The technical challenges will be to integrate the data sources and use methods that take advantage of current state regulatory data, Center for Medicare and Medicaid Statistics (CMS) nursing home data, and Bureau of Labor Statistics (BLS) workforce data to create timelier estimates. This will require policy and data agreements supporting the integration of these data sources. At the same time, the American Community Survey, Economic Census data, and other Census Bureau surveys provide more geographically detailed data about locations but are often two to five years out of date. In addition, classification schemes in the demographic programs will have to be aligned with those used in economic programs. Creating a holistic data picture of nursing homes and residential care facilities will allow policymakers and regulators to make better decisions to ensure safe and fair environments for residents and workers while helping business owners maintain their profitability.

Curated Data Enterprise Framework

Historically, the Census Bureau has created and implemented rigorous, reviewable, and repeatable processes for each of its surveys to provide data accuracy such that total survey error is minimized. The Curated Data Enterprise (CDE) requires the Bureau to go further and evolve a new data curation model that greatly expands and enables data discovery and retrieval, maintains data quality across multiple and diverse sources of information, adds value by creating new derived variables, and provides for reuse over time for new purposes and time-sensitive questions.

Curation is a flexible term. For instance, curation in the artistic context aims to juxtapose pieces of artwork to create a statement. In the context of census data, the goal of curation is to provide transparency and clear documentation supporting the integration and analysis of data sources in the context of a specific purpose, use, and reuse of the data. For the CDE vision to be workable and have its full impact, the Census Bureau will need to develop processes of curation that (1) evaluate, document, and preserve data and data products for use and future reuse and (2) enable components from data products to be curated in the context of specific purposes (3) allow for the dissemination of curated products on interactive platforms that promote their optimal use by data users at all levels of expertise.

DATA PRODUCT CHARACTERISTICS

Multi-Purpose

Sensitive Data Elements

We anticipate that data products, to include data, will have a few general characteristics that will shape the enterprise curation process. These include:

- Data products will have the potential to be used for multiple distinct purposes, creating a strong return on investment for initial data ingestion and curation efforts.
- Data products will contain sensitive data elements that need to be protected to address privacy, policy, and confidentiality concerns.

Harmonization of the curation processes will be an essential feature of the CDE. Keller et al. (2020, 2021)[3] present a framework whose steps can be adapted to shape the curation processes for the CDE. This framework, represented in Figure 1, has been repeatedly tested in a wide range of applications that seek to help governments, agencies, and companies access and integrate heterogeneous data (public and private) for decision making and policy evaluation (Keller, Lancaster, & Shipp 2017, Keller et al. 2018)[4]. The CDE framework provides a rigorous, transparent, and repeatable structure to build the CDE in the context of purpose and use such as the examples described on the previous page.

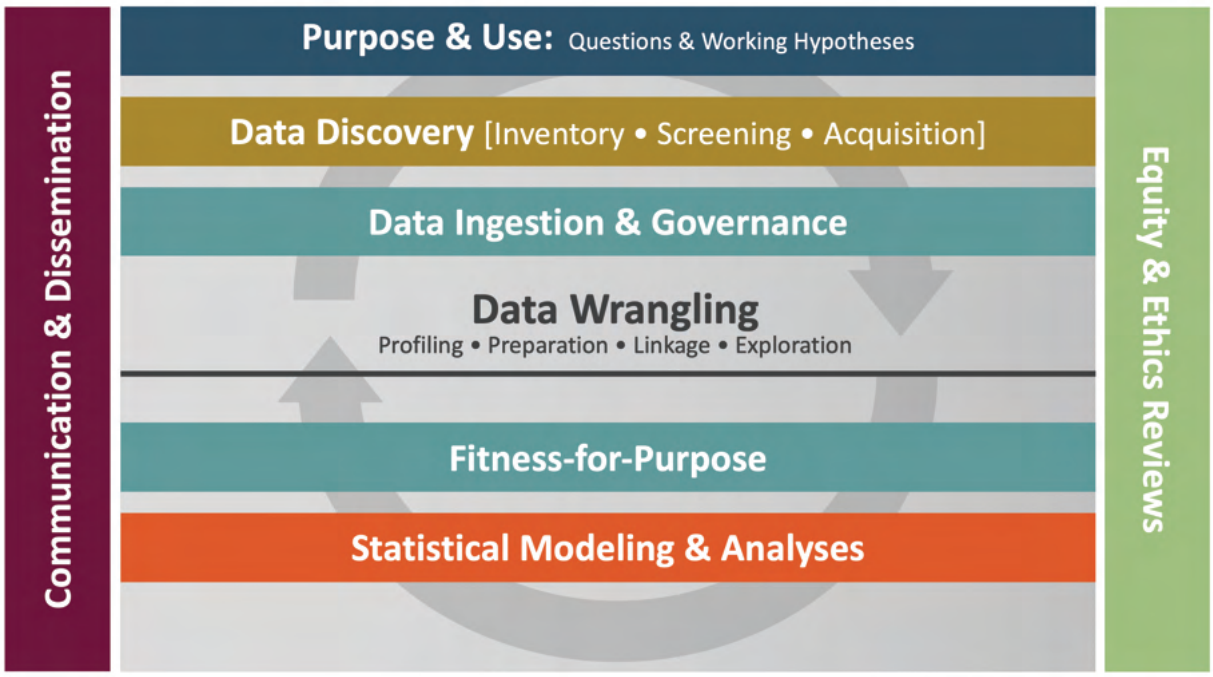
[3]Keller SA, Shipp SS. (2021). Data acumen. Notices of the American Mathematics Society, 66(9):1468-1477. https://doi.org/10.1090/noti2353

Keller SA, Shipp SS, Schroeder AD, Korkmaz G. (2020). Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1). https://-doi.org/10.1162/99608f92.2d83f7f5

[4]Keller SA, Lancaster V, Shipp S. 2017. Building capacity for data-driven governance – creating a new foundation for democracy. Statistics and Public Policy, 4:1-11. https://doi.org/10.1080/2330443X.2017.1374897

Keller SA, Korkmaz G, Robbins C, and Shipp S, (2018). New Opportunities to Observe and Measure Intangible Inputs to Innovation: Definitions, Operationalization, and Examples. Proceedings of the National Academy of Sciences (PNAS), 115 (50):12638-12645. https://-doe.org/10/1073pnas.18000467115

Figure 1. Curated Data Enterprise Framework



For a given purpose, the curation steps in Figure 1 include defining the purpose and use that motivates the curation (specific problem to be addressed); discovery of potential data sources relevant to the purpose and use (inventory, screening, and acquisition); data ingestion and governance; data wrangling to understand and prepare data for evaluation (data profiling, data preparation and linkage, and data exploration); fitness-for-use assessment in relation to the purpose; statistical modeling and analyses to extract metrics from or prepare the data products; communication and dissemination of results with stakeholders; and an ongoing ethics and equity review throughout the entire process (Keller et al. 2021).

To bring the CDE to light, consider its framework in the context of the migration example purpose:

Problem identification, the purpose and use of the data products to be developed, e.g., Domestic Migration Patterns – Why, where, and how often do people move?

Data discovery to identify, inventory, screen, and acquire new and existing data sources to address the purpose and use, e.g., migration measures in surveys (ACS, CPS, AHS, LEHD), administrative data (IRS, US Postal Service), and third-party data (cellphone detail records, utilities, real estate).

I**ngestion and governance to access data that meets restrictions** (i.e., FERPA, HIPPA, Privacy Act, state and local regulations, third-party agreements, and Census Bureau requirements). The migration data discovered and listed above are from multiple entities (Census Bureau, BLS, IRS, USPS, and companies (utility, real estate, moving vans, communication) each with its own rules for access and safeguarding the data. These steps are key to building trust.

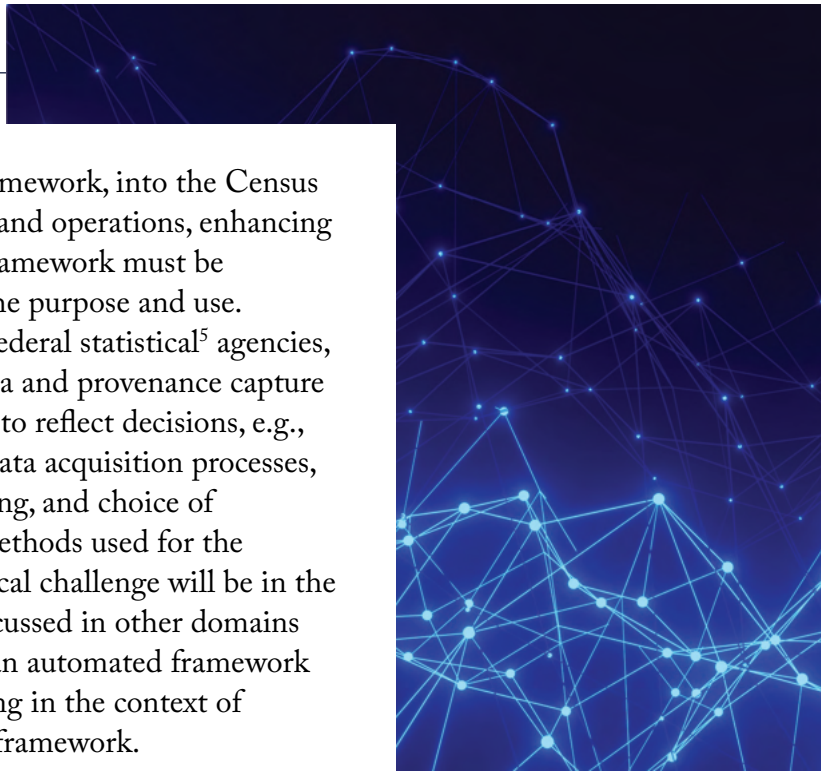
Wrangling starts with data profiling of selected variables to determine the quality of the data and its utility to the proposed purpose (e.g., migration), with respect to the completeness, data field value correctness, and logical consistency between fields and between records. Preparation considers different time frames, geographies, and definitions of variables to assess the quality of the data. Linking integrates the data in multiple ways. Exploration presents the data visually and in tabular form to identify early insights. The migration data presents many of these challenges.

Fitness-for-Use is a function of the modeling, data quality needs of the models, and data coverage (representativeness), and characterization of the information content in the results. There is no one gold standard source of migration data making this validation and assessment step both complicated and critical.

Communication across the team (e.g., subject matter experts, data scientists, and consumers) and with external stakeholders to seek feedback and vet the decisions and methods chosen at each step in the CDE framework. At the Census Bureau, many teams are involved in identifying and organizing data products around migration. Externally, applied demographers, business economists, and local and state economic planners need real-time migration data to inform their decisions. Ongoing communications and documentation of decisions across these groups will be essential.

In parallel, there are continuous discussions and curation about the ethical dimensions of the data, analyses, and product creation. Data choices through data discovery, statistical methods for integrating data across sources, as well as fitness-for-use assessments will inform this discussion. For example, the use of administrative data with survey data allows data users to learn about the demographic characteristics of people moving in and out of areas. One example of discussion around these data must address how the presentation of data and decisions made may favor one group over another or appear to do this.

Dissemination includes providing equitable approaches to access the data, data products, and services by all users, e.g., local government workers can access data through word queries and advanced data users can access the database to explore and statistically analyze the integrated migration data.



The goal is to embed the CDE, and hence curation framework, into the Census Bureau's ethos and ultimately the enterprise processes and operations, enhancing existing work practices. This means every step of the framework must be documented with metadata created in the context of the purpose and use. Consonant with the recent report on transparency in federal statistical[5] agencies, some steps of curation go well beyond current metadata and provenance capture and will require qualitative and subjective information to reflect decisions, e.g., defining the purpose and use, data discovery choices, data acquisition processes, selection of data sources and variables for data wrangling, and choice of imputation or linking methodologies, and statistical methods used for the creation of outputs. An exciting innovation and technical challenge will be in the automation of the curation processes. This is being discussed in other domains such as the work of Pezoulas et al. (2019)[6] to develop an automated framework for data quality assessment that focuses on data cleaning in the context of medical data that mirrors steps proposed in the CDE framework.

The CDE framework would be a new data curation model for the Census Bureau that greatly expands and enables data discovery and retrieval, maintains data quality across multiple and diverse sources of information, adds value by creating new derived variables, and provides for re-use over time through activities including authentication, archiving, metadata creation, digital preservation, and transformation. Consistent with the Bureau's strong tradition for protecting privacy, a key focus of the framework will be to ensure ethical data sharing while adopting procedures to safeguard the confidentiality of respondents' information and at the same time enhance efforts to make published data more trusted and accessible.

[5]National Academies of Sciences, Engineering, and Medicine. 2021. Transparency in Statistical Information for the National Center for Science and Engineering Statistics and all Federal Statistical Agencies. Washington, DC: The National Academies Press.

[6]Pezoulas VC, Kourou KD, Kalatzis F, Exarchos TP, Venetsanopoulou A, Zampeli E, Gandolfo S, Skopouli F, De Vita S, Tzioufas AG, Fotiadis DI. (2019) Medical data quality assessment: on the development of an automated framework for medical data curation. Computers in Biology and Medicine 107:270-283. https://doi.org/10.1016/j.compbiomed.2019.03.001

LISTENING SESSIONS: TESTING THE CDE CONCEPT WITH STAKEHOLDERS

The Curated Data Enterprise (CDE) concept was socialized with researchers and data users through a set of listening sessions. In these sessions, participants were presented and discussed the CDE vision to bring together multiple data sources, enabling accurate measurement of the economy and U.S. population dynamics in real-time, all the time.

The Census Bureau has a large and diverse set of stakeholders that rely on its data products. These are individuals or organizations invested in or affected by the data products produced by the Census Bureau or outcomes that result from an intermediary that uses the data, e.g., to distribute benefits or other outcomes. To guide the selection of participants in the listening session, ensuring full coverage of this broad community, a stakeholder taxonomy was developed. The stakeholder communities and their use of census data products given in Box A.

Box A. Stakeholder Communities and Their Use of Census Data Products

Congress and the Legislative Branch: Congress uses Census Bureau data to apportion the House of Representatives 435 seats every ten years, to allocate funding to states, localities, and individuals, and to formulate policy and evaluate program performance. In addition to Congress, several agencies, such as the Congressional Budget Office (CBO), Congressional Research Service, and General Accountability Office, also use Census Bureau data. For example, CBO uses these data for producing baseline budget projections, economic projections, cost estimates, and reports.

Public Policy Community: The community uses Census data products to support policy creation, program development, and administration. This stakeholder community exists at the federal level, within federal agencies and federal legislatures, and at the tribal, state, and local government levels. These include policy analysts, economic planners and developers, and applied demographers. It also includes organizations that support these groups, such as policy-oriented non-profits and survey research organizations. The uses of the Census data products range from the constitutionally mandated apportionment of seats in Congress to the distribution of approximately $1.5 trillion annually of federal funding to drawing political districts and the placement of roads, schools, and emergency services.

Business and Commerce Community: The business community uses Census data products extensively by repackaging the data to create "value-added" products and services. This community also used the data to derive economic context through analysis to inform strategic decision-making, e.g., a new headquarters location. The stakeholder community ranges from global businesses (e.g., Chief Economists at Amazon, ESRI, Zillow, Google, Claritas) to smaller companies and local start-ups. These include economists, social science researchers, systems engineers, and others involved in advancing data-related research in the private sector.

Non-Government Organizational (NGO) Community: This community uses Census data products to advocate for policies and programs. They use the data to monitor differential impacts across population subgroups and geographies to measure and evaluate whether there is fair and equitable administration and advocate for enforcement of existing policies and propose new ones. It includes special interest groups such as community and cultural associations, religious groups, labor unions, professional associations, and foundations.

Researcher Community: This community uses Census data products as the foundation for vast amounts of economic, social, and policy research. Researchers and educators in the social and economic sciences make up a significant community component. However, today many other areas, including public health, medicine, environmental sciences engineering, and the humanities, use Census data products to support their research and deliver educational programs. This community comes from academia and government agencies (e.g., Census Bureau itself, Bureau of Economic Analysis, USDA Economic Research Service) and other research organizations such as Brookings, Urban Institute, or RAND.

Media: This community uses Census data products to inform the public about the social and economic situation of US residents at local, state, regional, and federal levels, and changes in measurement.

General Public: Although not an organized "stakeholder," this group (which overlaps stakeholder groups) benefits from every program and policy informed by Census products. The public is the primary audience of the constitutionally mandated decennial census, and indirectly by many related data Census Bureau products.

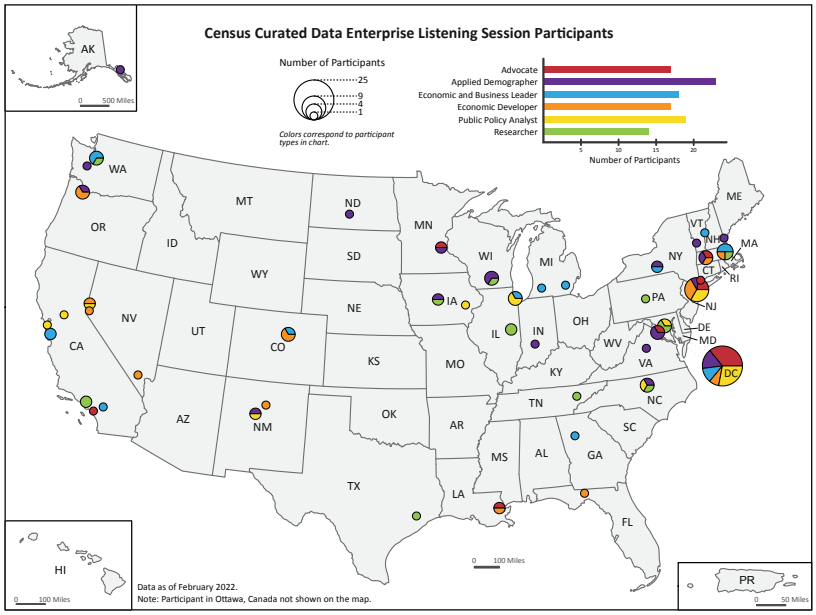
Table B. Listening Session Participant Counts by Stakeholder Community

Stakeholder Community	Date 2021	Count
Research Community		
Researchers	Sep 1	7
Researchers	Sep 2	7
Business and Commerce Community		
Economic Developers	Oct 7	10
Economic Developers	Oct 8	7
Economists & Business Leaders	Oct 14	9
Economists & Business Leaders	Oct 15	9
Non-Government Organizational Community		
Advocates	Nov 4	8
Advocates	Nov 5	9
Public Policy Community		
Public Policy Analysts	Dec 2	8
Public Policy Analysts	Dec 3	12
Applied Demographers	Dec 16	13
Applied Demographers	Dec 17	11
Total	12	**110**

Overall, 185 stakeholders were invited to the listening sessions, and 110 attended the sessions. Many of those who could not attend expressed interest and asked to participate in future sessions or in other ways to contribute to the initiative. Their distribution across the stakeholder communities is given in Table B and the geographic distribution for participation in Figure 2. The list of stakeholders and their affiliations can be found here.

During these 90-minute listening sessions, the CDE was described, information about Census Bureau's Frames Program (linked universal statistical frame) was shared and how it could form the foundation for the CDE was explored. Most of the time was spent in two open discussions around the potential value of the CDE and challenges that will need to be addressed for success.

Figure 2. Geographic Coverage for Listening Session Participants



The map presents the location of the listening session participants by type and number of stakeholders, e.g., Washington DC had the largest number of participants from one area, representing five of the six stakeholder groups. The bar chart shows the total number of participants by stakeholder type, e.g., applied demographers were the largest group and researchers the smallest.

Following the overview of the CDE and its relationship to the foundational linked universal statistical frame, the following questions were used to facilitate the first discussion regarding how the CDE could make a difference with a focus on needs and opportunities for Federal statistics.

What features of the economy and society would you like to see measured that the Federal Statistical System is not measuring (or is measuring inadequately)?

What measurements need the authority of Federal Statistics to inform research and policy as a trusted indicator?

What data innovations are taking place in the private sector that could be mirrored or leveraged by Federal Statistics?

Following the first discussion, one or more of the purposes shared above were used to highlight the art of the possible if the CDE currently existed. This led to a discussion around what the CDE could offer to stakeholders centered around the following questions.

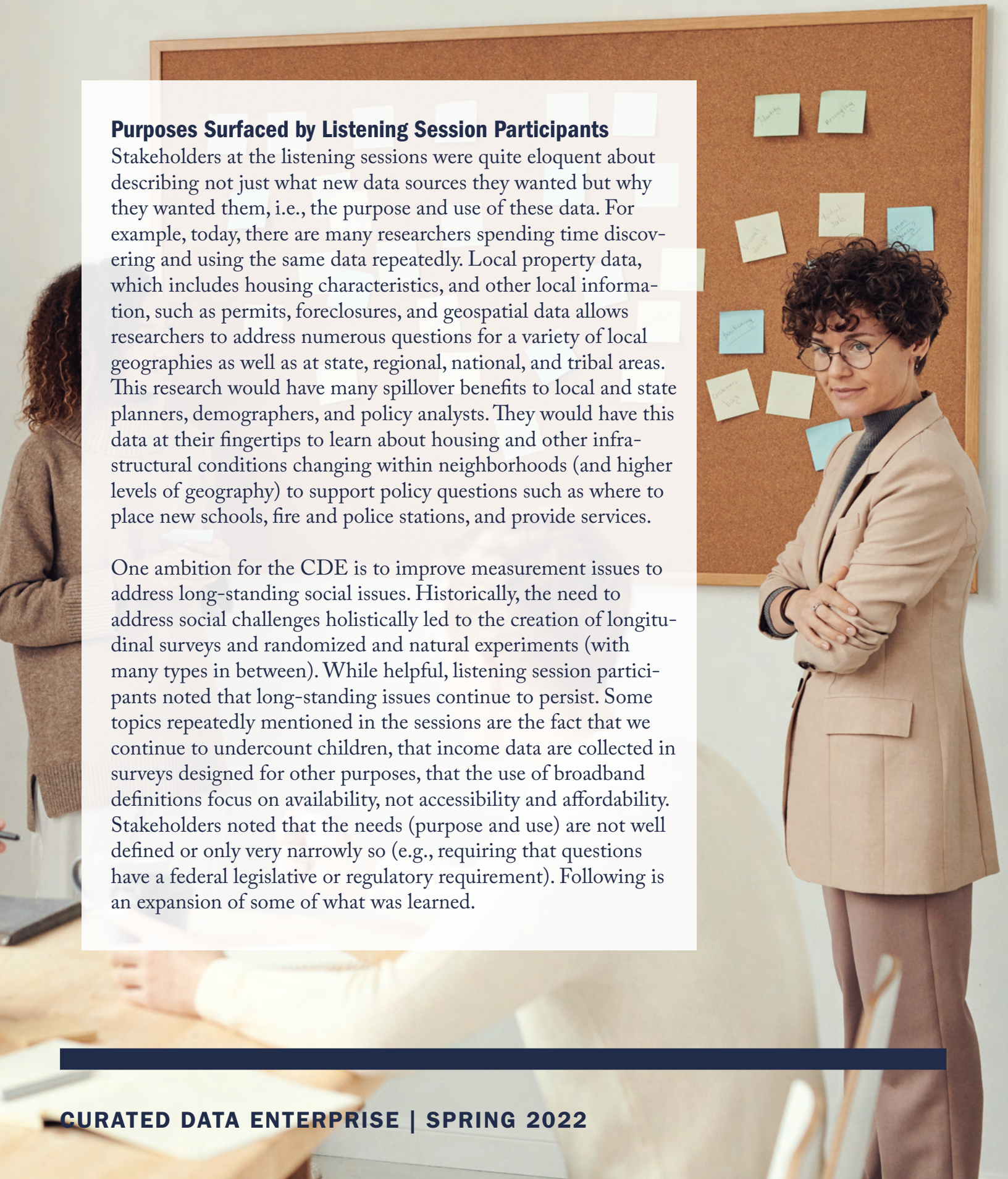
What data do you regularly consume, particularly third-party data? What data are essential to you?

What kinds of CDE outputs would benefit you the most?

How can data from the private and non-profit sectors best be incorporated into the CDE? What partnerships are needed?

What Stakeholders Shared

Across the listening sessions there was overwhelming interest and encouragement to proceed with the concept to create a CDE. Stakeholders noted that partnerships are key with the private sector and other agencies to access new data sources. There were several suggestions about producing data products designed to meet user needs at all levels of expertise as well as information products that directly support other tribal, federal, state, and local agencies, and businesses. The discussions surfaced deep interest in being able to capture and address emerging social and economic issues, frequently highlighting the need for new types of data collection, e.g., pulse type surveys. There were questions and sensitivities about data access, quality, and confidentiality, especially when using administrative data.



Purposes Surfaced by Listening Session Participants

Stakeholders at the listening sessions were quite eloquent about describing not just what new data sources they wanted but why they wanted them, i.e., the purpose and use of these data. For example, today, there are many researchers spending time discovering and using the same data repeatedly. Local property data, which includes housing characteristics, and other local information, such as permits, foreclosures, and geospatial data allows researchers to address numerous questions for a variety of local geographies as well as at state, regional, national, and tribal areas. This research would have many spillover benefits to local and state planners, demographers, and policy analysts. They would have this data at their fingertips to learn about housing and other infrastructural conditions changing within neighborhoods (and higher levels of geography) to support policy questions such as where to place new schools, fire and police stations, and provide services.

One ambition for the CDE is to improve measurement issues to address long-standing social issues. Historically, the need to address social challenges holistically led to the creation of longitudinal surveys and randomized and natural experiments (with many types in between). While helpful, listening session participants noted that long-standing issues continue to persist. Some topics repeatedly mentioned in the sessions are the fact that we continue to undercount children, that income data are collected in surveys designed for other purposes, that the use of broadband definitions focus on availability, not accessibility and affordability. Stakeholders noted that the needs (purpose and use) are not well defined or only very narrowly so (e.g., requiring that questions have a federal legislative or regulatory requirement). Following is an expansion of some of what was learned.

STAKEHOLDER EXAMPLES

1.

An accurate count of young children is complicated to obtain. In 2017, a Census Bureau working group identified three broad recommendations.[7] The first was to improve communications targeted at families and organizations that have contact with young children. Key to this recommendation is understanding the complex challenges of working with these families, such as immigration status, adoption, insufficient income. The second was to form partnerships with organizations that provide services to families with young children or advocate for them. The third recommendation is to improve enumerator training and survey questions. Many of these strategies were implemented for the 2020 Census, yet we learned from the Post-Enumeration Survey (PES) that the undercount of children actually grew compared to 2010.

Clearly, there are limitations on the enumeration or survey methodology for resolving this undercount. This lesson sets the stage for the CDE to bring an additional dimension to this work, in the form of compiling and integrating administrative and third-party data to fill the coverage void of young children in censuses and surveys. Administrative data are often used as the numerator and the decennial census data as the denominator, which will likely underestimate the percent of children receiving benefits. For example, IRS is missing a large share of children in their records, which means that these children are potentially missing out on the child tax credit. Another example is that children born in the first quarter of the decennial census year are missed if administrative data, such as IRS data, are used to fill in gaps. Other administrative data covering births for the first quarter of the decennial year might be available sooner than IRS data and could be used.

2.

Broadband availability and accessibility and affordability are topics recently elevated in the national public policy debate, especially knowing there are multiple data sources at the very moment large national infrastructure programs being implemented. The CDE can help measure the progress and address other questions, such as is the implementation equitable? This is a high-priority topic that cuts across political boundaries and one that integrating multiple sources of data has the potential to improve decision-making across stakeholder groups. The Cooperative Extension System studies broadband because of its potential to advance rural prosperity[8]. Several universities also study broadband to bridge the digital divide in underserved areas[9]. The CDE could support this research.

3.

Title VI of the Civil Rights Act of 1964, 42 U.S.C. 2000d et seq. prohibits discrimination on the basis of race, color, or national origin in any program or activity that receives Federal funds or other Federal financial assistance. Service providers are obligated to provide access to language services. Adherence to this law is often not observed nor tracked. Receipt of services and benefits can be the lifeline to becoming productive citizens. It would be helpful to collect this information to identify what issues immigrants have accessing benefits and other interactions with government (e.g., the court system). How might this be measured? Does their English proficiency affect this? How are their lives affected? These purposes are central to thinking that the CDE should support.

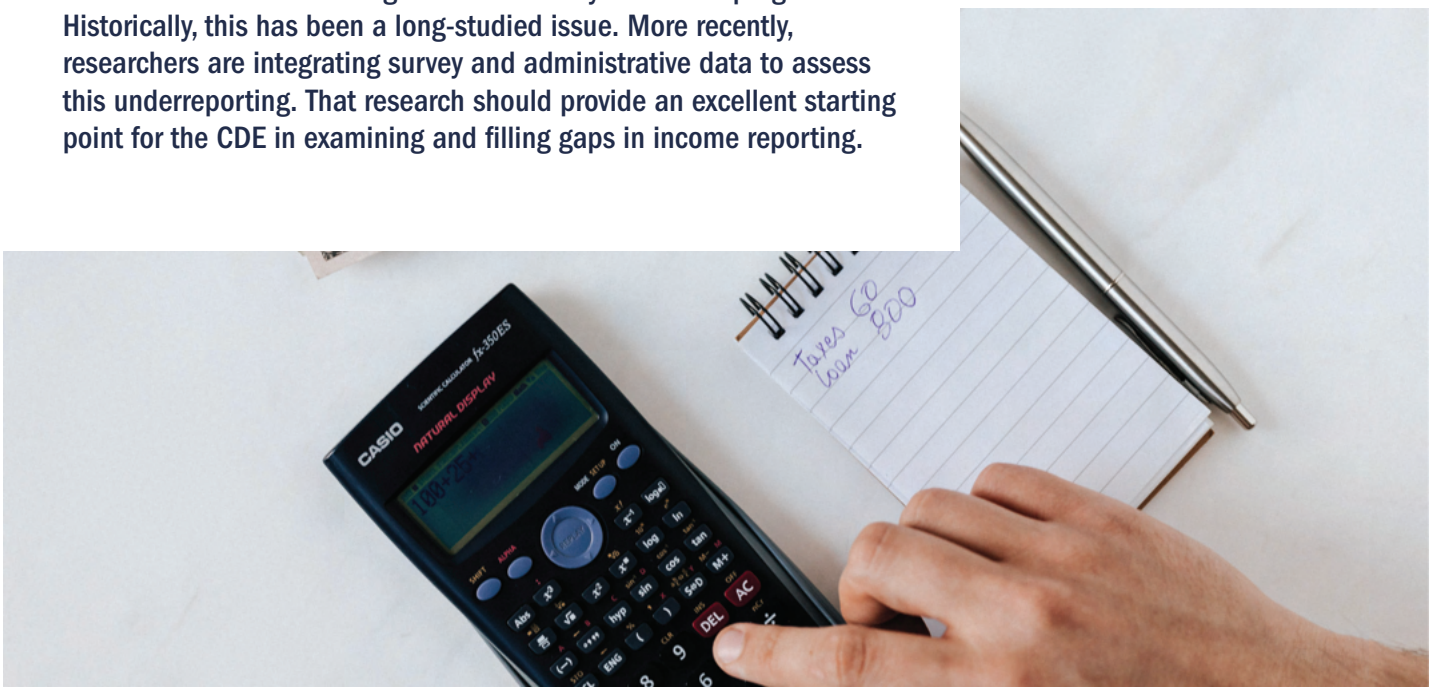
[7]https://www.census.gov/about/cac/nac/wg-undercount-children.html
[8]For one example, see https://www.wiscontext.org/search/content/Broadband
[9]For an example, see https://pcrd.purdue.edu/category/broadband/ and https://uva-bi-sdad.github.io/capital_region/

STAKEHOLDER EXAMPLES CONTINUED

4.

Measuring income accurately is another challenging area that would benefit from curating and integrating multiple data sources. Income responses continue to decline in surveys and affect other data, such as poverty measurement. Many have stated that the middle 80 percent of household income could be captured using survey and administrative data. The challenge is capturing income for the bottom and top 10 percent of the income distribution where survey data is less useful due to more complex income payments. The Federal Reserve Board's Survey of Consumer Finance, conducted every three years, oversamples wealthy households. The Census Bureau's Survey of Income and Program Participation oversamples low-income households. Still, these surveys do not capture all income completely. Many sources of income are only partially captured in these surveys; for example, survey respondents often do not remember less salient forms of income such as Temporary Assistance for Needy Families (TANF) or if they do recall receiving TANF, they often don't remember the amount or timing.[10] Underreporting of social assistance income can exceed more than half of the total government outlays for these programs. Historically, this has been a long-studied issue. More recently, researchers are integrating survey and administrative data to assess this underreporting. That research should provide an excellent starting point for the CDE in examining and filling gaps in income reporting.

[10]Celhay, Pablo A. and Meyer, Bruce and Mittag, Nikolas, What Leads to Measurement Errors? Evidence from Reports of Program Participation in Three Surveys (January 20, 2022). University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2022-11. http://dx.doi.org/10.2139/ssrn.4013845



STAKEHOLDERS' HIGH PRIORITY WISH LIST

There were also many themes that came up at each of the listening sessions. This section focuses primarily on data and data products that stakeholders identified as candidates if the CDE is to serve a priority of purposes and uses. Some of the themes came up in multiple sessions across stakeholders, such as the need for better migration data. Others were unique to a stakeholder group but identified as issues that would benefit all stakeholders, for example refining the race categories to be more inclusive, expanding resolution in age pyramids to younger and older ages, or aligning measurement to geographies different from the Census Bureau geographic hierarchy. Below and on the following pages are a selection of issues heard at listening sessions grouped by the six sets of stakeholder community participants in Table B.

Researchers

- What is needed to make better decisions about Housing and Urban Development HUD-subsidized housing programs? What private sector financial innovation data would enrich existing HUD data?
- How might immigration selection criteria be improved? Researchers need data that link immigrants to their classes of admission and their non-immigrant visa statuses. A missing piece of the puzzle is access to data about how long immigrants stay in the country.
- How do prior experiences shape current situations, such as poverty? The recommendation is to connect data across family members, parents, and grandparents to understand generational changes.
- What effect does the underreporting of social benefits data have on policy? For example, a purpose and use would be to examine access to benefits across states to identify high-mobility vulnerable populations.

Economic Developers

- The pandemic highlighted the need for better commuting data and understanding the working population in an area – their characteristics, industries, and businesses they work in, and days of the week.
- Data on effects of robotics, AI, and automation on workforce skills and workforce training needs.
- Timelier data, e.g., 2019 ACS data are old in 2021 as we emerge from the pandemic.
- More geographically detailed data and coverage, e.g., access to local and state data; while census tracts and Census Designated Places are currently used to disseminate census data, it may be helpful to enable data dissemination by neighborhoods or grids to align with specific populations of interest more accurately; in tribal and rural areas, low response rates on decennial censuses and surveys means that the data are not representative of the population.
- Consistency in land use data definitions across agencies and organizations, e.g., BEA, BLM, NGOs, water resource authorities, and other types of organizations that maintain these records.
- Linking and integrating data across many sources, for example:
    - Export data to trade data.
    - O*NET occupation data with SOC codes ("so I don't have to do this every time").
    - Supply chain, related workforce data, and ports data; and
    - Psychology data, e.g., attitudinal surveys and behavioral data.

Economists and Business Leaders

- Emerging areas that bring together data would inform issues such as the effects on rents when private equity firms are buying up rental housing or aggregating opinion data to learn about trends.
- More information related to labor markets, for example,
    - Occupation classification and wage distribution within firms and by size and type of firm.
    - Job mobility (e.g., job-to-job transitions or job-to-quit transitions–what happens during and after these transitions?).
    - Worker characteristics (especially in LEHD).
    - In many states, vendors own the unemployment insurance data making it challenging to obtain; data requires a lot of preparation to make is useful for analysis.
- Retail sales data allow you to track daily events during the pandemic and the business cycle. The top 100 retailers cover more than two-thirds of overall sales in the US. These data could be collected easily at a daily frequency with very low latency (see National Federation of Retailers https://nrf.com/tag/big-data).
- Improved recording of firm entry and exits during crisis events such as the pandemic or during business cycles.
- Tracking the impact of the stimulus funding could be done by combining geospatial data merged with census data to give them some insight by income deciles

Advocates

- New ways to measure economic growth and wealth are needed that reflect the quality of life as well as monetary growth. Gross Domestic Product is only one way to do this. Wealth is a more important indicator of resilience than income. We need frequent and disaggregated measures of wealth.
- Administrative data may have similar problems as survey data, e.g., underserved populations are not in the data.
- The Shadow economy is not measured well in government statistics. The Shadow Economy is an underground, informal, or parallel economy, and includes not only illegal activities but also unreported income from the production of legal goods and services, either from monetary or barter transactions. Since people are unlikely to report such activities, a variety of surrogates need to be developed to capture this underground economy.
- Unemployment is only measuring those who are out of work and actively looking for work; it is not measuring those who have dropped out of the labor force but might return to the workforce with the right incentives. “Unemployment is a woefully outdated measure and gets a ton of headline attention. It should be modernized.”
- Race categories - Concern over accurately capturing Asian American communities in administrative records. They tend to be less present in administrative programs. In surveys, Asian Americans are often grouped into Other. There is also concern that the Arab American communities are not represented in the data. Agencies say they cannot add this category because it is not part of the OMB definition. OMB says that their definition is a minimum standard, not a maximum, and agencies could add this category. (This category is sometimes referred to as MENA - Middle Eastern or North African[11]).
- Capturing consistent and more detailed data about informal housing arrangements and the homeless; demographic information about incarcerated populations across states; people with developmental disabilities; and the LGBTQ population.
- Improved data on voter registration and civic engagement information would be informative about level of people’s involvement by different levels of geography.

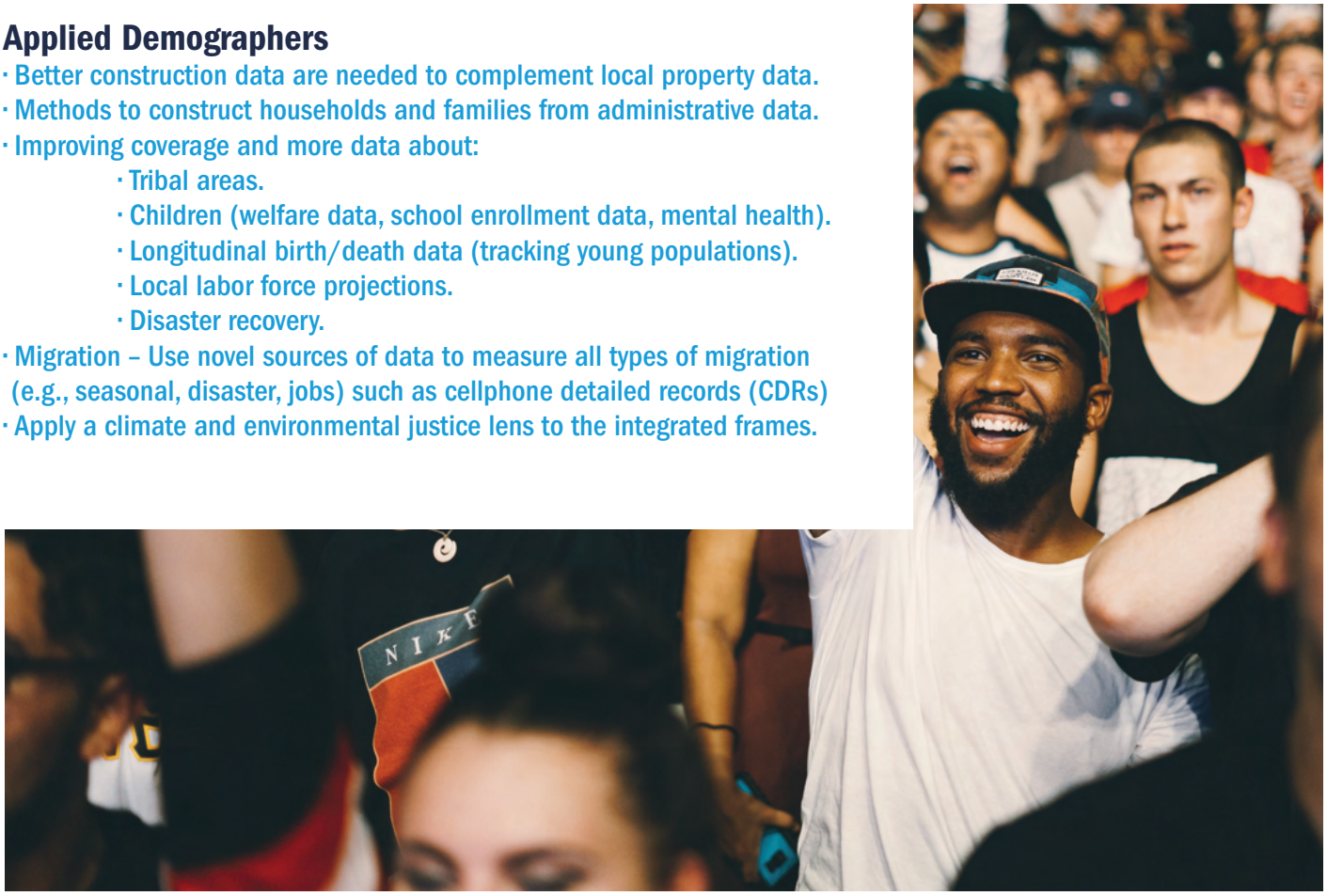
[11]https://www.npr.org/2018/01/29/581541111/no-middle-eastern-or-north-african-category-on-2020-census-bureau-says

Public Policy Analysts

- Counting the homeless using administrative data, e.g., 80% of the homeless get SNAP or Medicaid, is one way to reach this group.
- Workforce skills – Need data that defines the quickly changing landscape. "The Occupational Information Network, O*NET, by definition cannot include job categories that have been around for less than three years, so they're almost by design out of date."
- Integration of social determinants of health and demographic characteristics data.
- Household occupancy and vacancy rates, e.g., many vacation homes became full-time residences during the pandemic.
- Access to school, health data, and local data, e.g., to learn about the pandemic effects of K-12 education on digital literacy and long-term economic and workforce outcomes.
- Connect industry and job data in LEHD in originating and destination states to learn about effects on the industry, e.g., what is the industry mix of workers coming from Nevada to California (such as service and manufacturing), and what industries are they taking jobs in?

Applied Demographers

- Better construction data are needed to complement local property data.
- Methods to construct households and families from administrative data.
- Improving coverage and more data about:
    - Tribal areas.
    - Children (welfare data, school enrollment data, mental health).
    - Longitudinal birth/death data (tracking young populations).
    - Local labor force projections.
    - Disaster recovery.
- Migration – Use novel sources of data to measure all types of migration (e.g., seasonal, disaster, jobs) such as cellphone detailed records (CDRs)
- Apply a climate and environmental justice lens to the integrated frames.



Technical Gaps, Challenges, and Questions that Need to be Addressed

Stakeholders identified many technical challenges, sometimes in the context of specific purposes or uses and sometimes as stand-alone topics. Figure 3 is a word cloud created from the combined transcripts across the 12 listening sessions. We group the gaps, challenges, and questions into five categories and briefly describe them below. These topics will be explored in more depth as a research agenda and implementation plan are developed.

Figure 3. Word Cloud of Listening Session Transcripts

Data Integration and Linkage

Linking and integrating census, survey, administrative, third-party, tribal, local, and state data provides a rich set of data that allows research questions to be answered which cut across program areas and disciplines. This can take many forms including probabilistic approaches that match a handful of variables or data that may be integrated into models to answer specific questions. Linking and integration includes mapping data to a classification scheme, e.g., O*Net to Standard Occupational Classification (SOC) codes. It can involve multiple variable types, such as linking people to household addresses, people to households or families, or imports to exports. Implicit in integrating and linking data is understanding the quality and metadata around each data source to understand where and how to use the data.

Grounding data integration and linking processes in the Curated Data Enterprise (CDE) framework requires that these be done in the context of some purpose or use. From there, the approach might be to link the Census Bureau surveys first, assess the quality, and then identify federal administrative data to fill gaps. Then one might look for local and state data, and finally, third-party data. Or, this process could be reversed. One measure may not need to be designated as a single gold standard for a data attribute. Measures can be created from multiple data sources (such as self-reported IRS data versus survey data on income) could be compared and contrasted. The goal would be to identify the best data for any individual data element needed to support the purpose. Researchers may want to choose the best measures for the tasks at hand. In this context, the CDE framework becomes critical for documenting findings and decisions, creating relevant metadata at each step.

Data integration and linking may be one of the most powerful features of the CDE. The current survey-only model for estimates on the economy and population limit the kind of research questions that can be posed, sometimes waiting years for adding a question(s) to a survey. The virtue of the curated data integration platform allows for more timely research, and for testing alternative constructs to address the same question.

Access and Dissemination Models for Data and Data Products

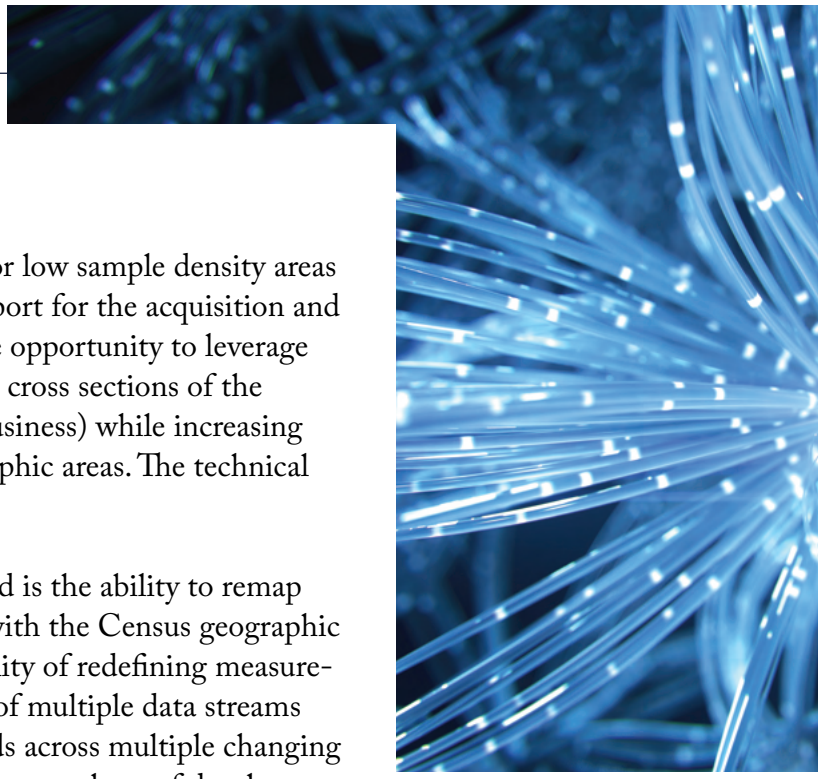
Dissemination and accessibility of data, data products, and services was a common theme heard at each listening session. It was shared that a database (or data enterprise) is built on four basic principles, in that it is

- *A simple data model, noting that the integrated frames are in this format.*
- *Transactionally updated, as new surveys or administrative data come in; the database is constantly updated.*
- *Minimizes redundancy*
- *Multiple products are dynamically generatable based on the information in the data enterprise.*

From an industry perspective, new state-of-the-art access and dissemination methods are available that can provide data users with what they need at various levels. Examples include https://schema.org/ and https://www.datacommons.org/.

Users want to access data, data products, and services that go beyond the Federal Statistical Research Data Center (FSRDC) model. They also would like flexibility to redefine categories, such as age groups and geographies. Approaches may also include new dimensions that go beyond the person, household, or business.

It was suggested that one approach to transforming access and dissemination is to provide new services. For example, data are streamed as maps or analytic services directly to end-users rather than through data downloads to the application environment. This approach provides information products directly from the Census Bureau. The goal is to pipeline these directly into federal agencies, the academy, and the private sector applications. Another example would be for the Census Bureau to have the ability for rapid prototyping for developing data products and/or new data collection related to a query or new purpose.



Technical challenges for filling data gaps

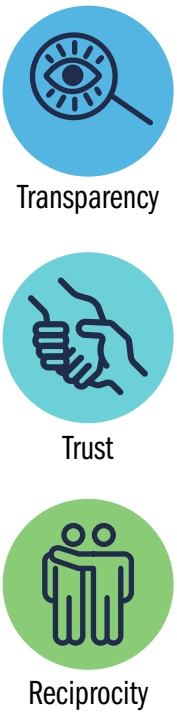
As noted earlier, data gaps are pervasive, particularly for low sample density areas (geographies or demographic groups). The CDE's support for the acquisition and integrating of multiple data sources should provide the opportunity to leverage administrative and third-party data to measure various cross sections of the relevant populations (geographic, demographic, and business) while increasing survey sample size for small and/or low density geographic areas. The technical details for implementation need to be developed.

A technical gap that stakeholders would like prioritized is the ability to remap data and measures to geographies that may not align with the Census geographic hierarchy. In addition, data users would like the flexibility of redefining measurement scales like age pyramids. Finally, the integrating of multiple data streams will create technical gaps in the ability to capture trends across multiple changing data sequences. These issues will all be examined in the next phase of development of the CDE.

Multi-Way Partnerships and Incentives to Participate

The CDE initiative requires partnerships with other statistical agencies and private and public sector companies and organizations. A partnership program includes incentives and governance agreements to discover and acquire data to address a specific purpose. The data quality assessment and implications of the continued availability of the sources should be explored. Alternative data sources are necessary to ensure backups and continuity of data sources or replacements if data sources obtained through partnerships no longer exist and to provide additional insights and perspectives. Examples of third-party and local, state, federal, and tribal agency data and partnerships can be found here.

TRUST-BASED APPROACH



[12]Erlich Y, Williams JB, Glazer D, Yocum K, Farahany N, et al. (2014). Redefining genomic privacy: trust and empowerment. PLOS Biol. 12(11):e1001983. https://doi.org/10.1371/journal.pbio.1001983

World Economic Forum, 2013. Unlocking the value of personal data: From collection to usage. Geneva: World Economic Forum. https://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf

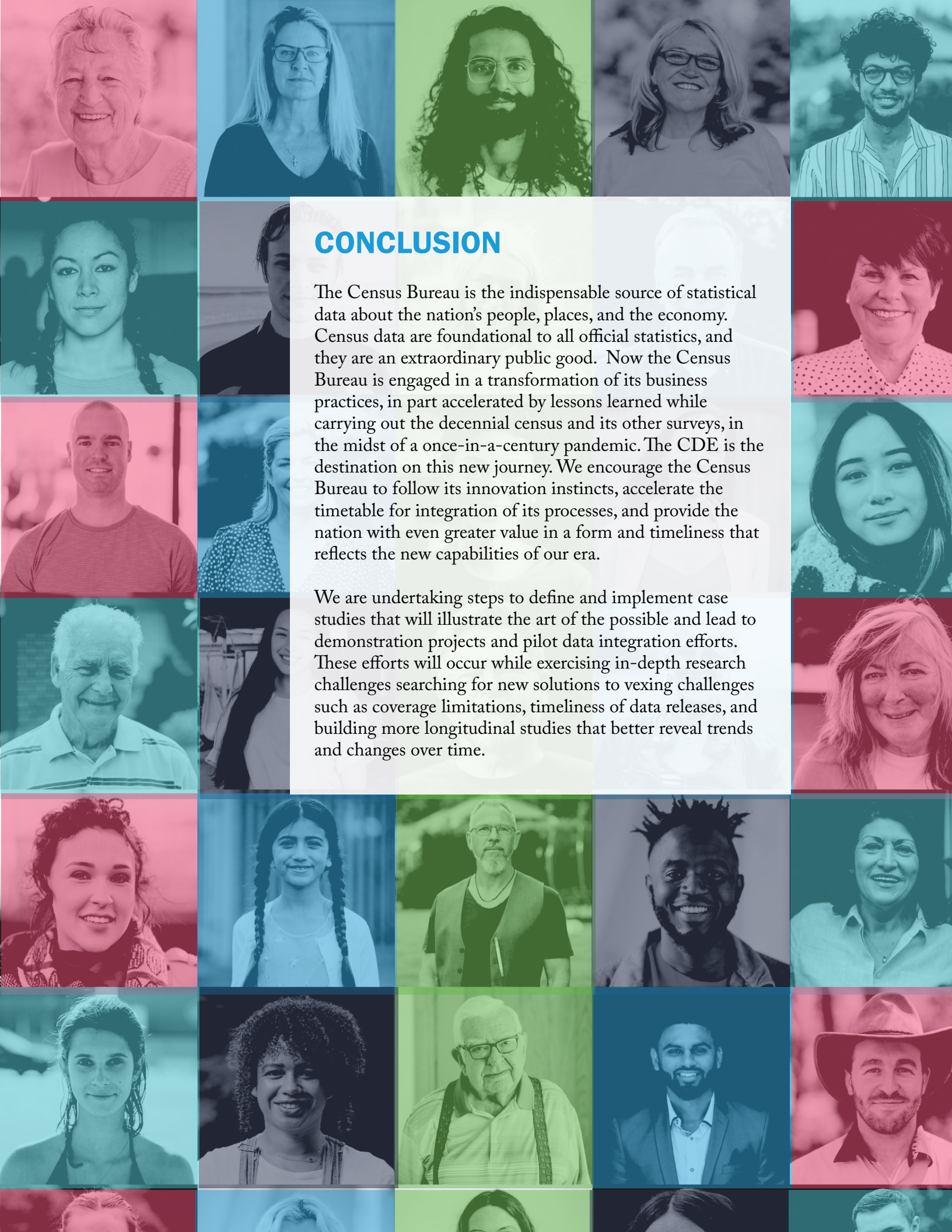
President's Council of Advisors on Science and Technology (US), 2014. Report to the President, Big Data and Privacy: A Technology Perspective. Executive Office of the President, President's Council of Advisors on Science and Technology. https://obamawhitehouse.archives.gov/blog/2014/05/01/p-cast-releases-report-big-data-and-privacy

[13]Keller, S. A., Shipp, S., & Schroeder, A. (2016). Does big data change the privacy landscape? A review of the issues. Annual Review of Statistics and Its Application, 3, 161-180. https://doi.org/10.1146/annurev-statistics-041715-033453

Privacy, Confidentiality, and Trust

The privacy landscape is changing quickly. Although Title 13 is the bedrock for the Census Bureau, participants noted there may be the need and opportunity to reinterpret aspects of Title 13 in the context of the CDE. Concerns that will need to be addressed focus on the perceived lack of transparency about third-party data and the potential for low coverage of underserved populations. It was suggested that advances in other fields be examined, such as the area of genomic research where masking data makes it unusable and other trust centric models around use are being developed.[12] To create a trust-based approach where both researchers and participants benefit, principles of transparency, trust, and reciprocity must be implemented to support good behavior, discourage malicious behavior, and punish noncompliance.[13]

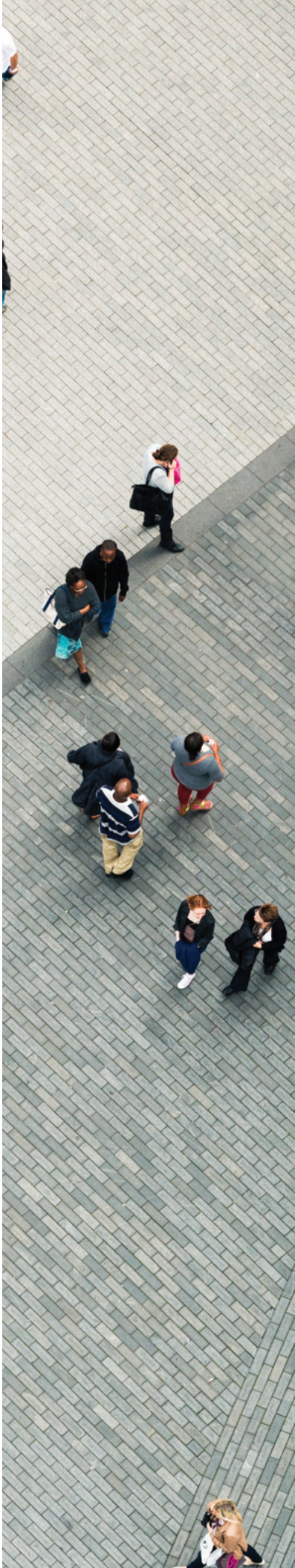
To achieve this state requires clear communications and definitions and equity considerations that take into account sensitivities around the use of administrative and third-party data. For example, an area of research to consider is protecting the data based on purpose and use versus individual data elements as traditionally done.



CONCLUSION

The Census Bureau is the indispensable source of statistical data about the nation's people, places, and the economy. Census data are foundational to all official statistics, and they are an extraordinary public good. Now the Census Bureau is engaged in a transformation of its business practices, in part accelerated by lessons learned while carrying out the decennial census and its other surveys, in the midst of a once-in-a-century pandemic. The CDE is the destination on this new journey. We encourage the Census Bureau to follow its innovation instincts, accelerate the timetable for integration of its processes, and provide the nation with even greater value in a form and timeliness that reflects the new capabilities of our era.

We are undertaking steps to define and implement case studies that will illustrate the art of the possible and lead to demonstration projects and pilot data integration efforts. These efforts will occur while exercising in-depth research challenges searching for new solutions to vexing challenges such as coverage limitations, timeliness of data releases, and building more longitudinal studies that better reveal trends and changes over time.



UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Keller S, Prewitt K, Thompson J, Jost S, Barrett C, Nusser S, Salvo J, Shipp S. A 21st Century Census Curated Data Enterprise. A Bold New Approach to Create Official Statistics, Technical Report BI-2022-1115. [Print]. Proceedings of the Biocomplexity Institute, University of Virginia; 2022.
https://doi.org/10.18130/r174-yk24