GOVERNMENT ADVANCES IN STATISTICAL PROGRAMMING (GASP) 2023

Virtual Conference

June 14-15, 2023

Organized by the Computational Statistics for the Production of Official Statistics (CSPOS) Interest Group of the Federal Committee on Statistical Methodology

Planning Committee

Co-Chairs: Lisa M. Frehill (Department of Energy) & Peter Meyer (Bureau of Labor Statistics)

Nathan Cruze (NASA Langley Research Center)

Andreea Erciulescu (Westat)

Drake Gibson (Department of Homeland Security)

Kelsey Gray (Westat Insight)

Brandon Kopp (Bureau of Labor Statistics)

Wendy Lynn Martinez (US Census Bureau)

Cecile McWilliams Murray (US Census Bureau)

José “Bayoán Santiago-Calderón (Bureau of Economic Analysis)

Matthew Williams (RTI International)

Special Thanks: Donna LaLonde and the American Statistical Association for hosting the conference!

FINAL PROGRAM

Government Advances in Statistical Programming (GASP)
Wednesday, 14 June – Thursday, 15 June 2023

Wednesday 14 June 2023

12:00 pm Conference Opening Session

Conference Co-Chairs: Lisa M. Frehill (Department of Energy) and Peter Meyer (Bureau of Labor Statistics)

FCSM CSPOS Chair: Nathan Cruze (NASA Langley Research Center)

Introduction of Keynote Speaker: Wendy Martinez (US Census Bureau)

Keynote: "Understanding How Dimension Reduction Tools Work"

Cynthia Rudin, *Earl D. McLean, Jr. Professor of Computer Science and Engineering at Duke University and Director of the Interpretable Machine Learning Lab*

1:00 – 1:50 pm Paper Session 1: Case Studies in Machine Learning

Chair: Andreea Erciulescu (Westat)

Bringing Search to the Economic Census: A NAPCS Classification Tool
Clayton Knappenberger *(US Census Bureau)*

Machine learning is the easy part: recoding write-in responses in the 2021 VIUS
Cecile Murray *(US Census Bureau/Reveal Global Consulting)*

Nowcasting European Production Price Indexes
Gergely Attila Kiss *(Hungarian Central Statistical Office, Central European University)*

1:50 – 2:15 pm Lightning Talks 1: The Next Generation

Chair: Jun Yan (University of Connecticut)

Using CANS scores in the Community Risk Result (CRR) project in R
Jamie Joseph *(Vanderbilt University), Rameela Raman (Vanderbilt University)*

Open Source Software in the Federal Government: An Analysis of Code.Gov
Gizem Korkmaz (Westat), ***Rahul Shrivastava*** *(Westat), Anil Battalahalli (Westat), Ekaterina Levitskaya (Coleridge Initiative), José Bayoán Santiago Calderón (Bureau of Economic Analysis), Ledia Guci (Bureau of Economic Analysis), Carol Robbins (National Science Foundation)*

Timely Examination of Survey Data Quality
Winnie Xu *(Westat)*

2:15 – 2:30 pm Break

2:30 – 3:15 pm Paper Session 2: Natural Language Processing

Chair: Cecile Murray (US Census Bureau)

Text Analysis of Health Equity and Disparities in IRS Form 990 Schedule H
Emily Hadley *(RTI International), Laura Marcial (RTI International), Wes Quattrone (RTI International), Georgiy Bobashev (RTI International)*

Manufacturing Sentiment
Tomaz Cajner (Federal Reserve Board), ***Leland D. Crane*** *(Federal Reserve Board), Christopher Kurz (Federal Reserve Board), Norman Morin (Federal Reserve Board), Paul E. Soto (Federal Reserve Board), Betsy Vrankovich (Federal Reserve Board)*

Automating Text Cluster Naming with Large Language Models
Alexander Preiss (RTI International), *John Bollenbacher (RTI International), Anthony Berghammer (RTI International), John McCarthy (RTI International), Caren Arbeit (RTI International)*

3:15 – 3:40 pm Lightning Talks 2: Evaluations and Models

Chair: Drake Gibson (Department of Homeland Security)

Evaluation of a method for georeferencing farms
Robert Emmet *(USDA, National Agricultural Statistics Service), Kevin Hunt (USDA, National Agricultural Statistics Service), Rachael Jennings (USDA, National Agricultural Statistics Service), Kara Daniel (USDA, National Agricultural Statistics Service), Denise A. Abreu (USDA, National Agricultural Statistics Service)*

treecompareR: an open-source software package to visualize hierarchical chemical classifications
Paul Kruse *(Oak Ridge Institute for Science and Education), Caroline L. Ring (Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency)*

Building a Modeling Platform for DC Government
Hersh Gupta *(DC Office of the Chief Technology Officer), Gautam Chakravarty (DC Office of the Chief Technology Officer)*

3:40 – 5:00 pm Paper Session 3: Record Linkage and Disclosure Advances

Chair: Peter Meyer (Bureau of Labor Statistics)

Is Place of Birth an Appropriate Measure of Childhood Exposure?
John Sullivan *(US Census Bureau), Katie Genadek (US Census Bureau), Carlos Becerra (US Census Bureau)*

Linking the 1980 Enumeration Sample to the Decennial Census
Kelsey Drotning *(US Census Bureau), Katie Genadek (US Census Bureau)*

Harnessing Open-Source Tools to Link Records with PII
M Daniel Brannock *(RTI International), Ed Preble (RTI International), Lynn Langton (RTI International), Marguerite DeLiema (University of Minnesota)*

Identification of Anomalous Data Entries in Repeated Surveys
Luca Sartore *(NISS/NASS), Chen Lu (NISS/NASS), Justin van Wart (NASS), Andrew Dau (NASS), Valbona Bejleri (NASS)*

Refining Disclosure Controls for the Census of Fatal Occupational Injuries
Danny Friel *(US Bureau of Labor Statistics), Alyssa Gillen (US Bureau of Labor Statistics), Julie Krautter (US Bureau of Labor Statistics), Yvangelista Saastamoinen (US Bureau of Labor Statistics)*

5:00 – 5:30 pm Lightning Talks 3: Refine Files or Revise

Chair: Hiroaki Minato (Department of Energy, Energy Information Administration)

Expanding the R Shiny Apps in the growclusters Package
Randall Powers *(Bureau of Labor Statistics), Terrance Savitsky (Bureau of Labor Statistics), Wendy Martinez (US Census Bureau)*

Sampling Frames Simulation with Copulas in R
Eli Kravitz *(New Light Technologies), Daphne Liu (US Census Bureau, Center for Economic Studies), Stephanie Coffey (US Census Bureau, Center for Economic Studies)*

Tiling your Parquet Files for the Greatest Efficiency
Stas Kolenikov *(NORC at the University of Chicago)*

Exploring the Analytical Power of Input-Output Tables in the UK Context
Julen Grube Doiz *(UK Office for National Statistics), Eric Crane (UK Office for National Statistics), Andrew Banks (UK Office for National Statistics)*

5:30 – 5:35 pm Day 1 Announcements

Thursday 15 June 2023

10:30 -- 11:50 pm Workshop: ChatGPT: A First-Look at the Utility and Risks of Large Language Models (LLMs) for Federal Agencies
Benjamin Rogers (NCHS/CDC) and Travis Hoppe (NCHS/CDC)

12:00 – 12:05 pm Day 2 Welcome and Logistics Review

12:05 – 1:25 pm Paper Session 4: Promising Practices

Chair: Brandon Kopp (Bureau of Labor Statistics)

On Principles for Government Open Data Curation
Jun Yan *(University of Connecticut)*

Designing Against Bias: Identifying and Mitigating Bias in Machine Learning & AI
David J Corliss *(Peace-Work)*

Developing Documented, Trustworthy R Packages for Official Survey Statistics
Benjamin Schneider *(Westat)*

Webscraping in Python
Matt Ring *(Westat Insight)*

UI Design Patterns for Code Reuse
Weihuang Wong *(NORC at the University of Chicago),* ***Kiegan Rice*** *(NORC at the University of Chicago)*

1:25 – 2:00 pm Lightning Talks 4: Tools

Chair: Peter Parker (NASA Langley Research Center)

Toward a Tool for Automated Disclosure Risk Review of NCES Data Products
John Riddles *(Westat), Michael Armesto (Sanametrix), Tom Krenzke (Westat), Jennifer Nielsen (National Center for Education Statistics)*

Using NLP to Access to the Unemployment Insurance System
Siobhan Mills De La Rosa *(American Institutes for Research), Meghan Coffee (American Institutes for Research), and Samia Amin (American Institutes for Research)*

Incorporating Survey Weights into Structural Topic Models
Caroline Lancaster (NORC at the University of Chicago), Brandon Sepulvado (NORC at the University of Chicago), Josh Lerner (NORC at the University of Chicago),
Evan Herring-Nathan *(NORC at the University of Chicago), Stas Kolenikov (NORC at the University of Chicago)*

Traffic tracker: A gentle introduction to Python, APIs, and GitHub actions
Emily Mitchell *(AHRQ - Agency for Healthcare Research and Quality)*

2:00 – 2:15 pm Break

2:15 – 3:45 pm Paper Session 5: Methodological Advances

Chair: Matt Williams (RTI International)

Analysis of Crisis Effects via Maximum Entropy
Tucker McElroy *(US Census Bureau)*

Developing a Parameterized and Enterprise Ready Survey Data Review Tool
Andrew J. Dau *(USDA NASS), Michael Jacobsen (USDA NASS), Ryan E. Morton (Morton Analytics LLC), Jared M, Pratt (USDA NASS), Justin P. Van Wart (USDA NASS)*

A Computational Pipeline Applied to a Nested Case-Control Study Simulation
Michelle Mellers *(Uniformed Services University of the Health Sciences and Henry M Jackson Foundation), Celia Byrne (Uniformed Services University of the Health Sciences)*

Deep Neural Network Based Mass Imputation for Data Integration
Sixia Chen *(University of Oklahoma Health Sciences Center), Chao Xu (University of Oklahoma Health Sciences Center)*

Weswgt: A New R tool for Survey Weight Adjustments
Jianru Chen *(Westat), Benjamin Schneider (Westat), Minsun Riddles (Westat)*

3:45 – 4:10 pm Lightning Talks 5: Data Visualization

Chair: Kelsey Gray (Westat Insight)

Python Dash for Visualization: Reusable Dashboards
John Lombardi *(US Census Bureau)*

A Data Viz Makeover: The Evolution of A Visualization Showing State SNAP Data
Brian Knop *(US Census Bureau)*

Developing a Data Quality Scorecard Dashboard to Assess Data's Fitness for Use
John Finamore (NCSES), Elizabeth Mannshardt (NCSES), Lisa B. Mirel (NCSES), Julie Banks (NORC at the University of Chicago), F. Jay Breidt (NORC at the University of Chicago), Benjamin R. Peck (NORC at the University of Chicago), Kiegan Rice (NORC at the University of Chicago), Zachary H. Seeskin (NORC at the University of Chicago), Lance A. Selfa (NORC at the University of Chicago), ***Grace Xie*** *(NORC at the University of Chicago)*

4:10 – 4:55 pm Closing Plenary

Introduction of Plenary Speaker: José Bayoán Santiago Calderón (Bureau of Economic Analysis)

Plenary: "An insider's view of the history of open-source software for data science"

Douglas Bates*, Emeritus Professor of Statistics, University of Wisconsin – Madison*

4:55 – 5:00 pm **Closing Remarks:** Nathan Cruze (NASA Langley Research Center) and FCSM CSPOS

Paper Session 1: Case Studies in Machine Learning	
Bringing Search to the Economic Census: A NAPCS Classification Tool *Clayton Knappenberger (U.S. Census Bureau)*	The North American Product Classification System (NAPCS) was first introduced in the 2017 Economic Census and provides greater detail on the range of products and services offered by businesses than what was previously available with just an industry code. In the 2022 Economic Census, NAPCS consists of 7,234 codes and respondents often find that they are unable to identify correct NAPCS codes for their business. These respondents leave written descriptions of their products and services, and over 1 million of these needed to be reviewed by Census Analysts in the 2017 Economic Census. The Smart Instrument NAPCS Classification Tool (SINCT) offers respondents a low latency search engine to find appropriate NAPCS codes based on a written description of their products and services. SINCT is neural network document embedding model (doc2vec) that embeds respondent searches in a numerical space and then identifies NAPCS codes that are close to this embedding. Attendees who work in survey design will be interested to learn about how machine learning can improve respondents' experience while also reducing the amount of expensive manual processing that is necessary after collection. Data Scientists will benefit from our experience building a machine learning solution that achieves an estimated 71% top-10 accuracy with thousands of possible classes, limited training data, and strict compute and latency requirements.
Machine learning is the easy part: recoding write-in responses in the 2021 VIUS *Cecile Murray (U.S. Census Bureau/Reveal Global Consulting)*	Federal statistical agencies are increasingly adopting machine learning methods to increase data processing efficiency and improve data quality, but applying these methods to real-world survey data often is not straightforward. This talk will describe how we used a machine-learning-based process to assign a large volume of free-text responses to the 2021 Vehicle Inventory and Use Survey (VIUS) to predefined categories, despite wide variation in response validity. Conducted in partnership between the Bureau of Transportation Statistics (BTS), the Federal Highway Administration (FHWA), Department of Energy (DOE), and the U.S. Census Bureau, the VIUS is a survey of the nation's truck population to gather data on the characteristics and use of vehicles on our nation's roads. For respondents who carry goods or products for sale, VIUS asked respondents to estimate the percentage of loaded miles they drove carrying different types of goods using a categorical response format with an open-ended other/write-in option. Many of these other/write-in responses could be classified into the provided categories, but recoding these responses manually would have been prohibitively time-consuming for survey analysts. Instead, we developed a machine learning approach to reduce the need for human review while ensuring quality and consistency, saving significant analyst time and resources. This talk will describe that approach and highlight challenges associated with using machine learning methods when response validity varies widely.

Paper Session 1: Case Studies in Machine Learning	
Nowcasting European Production Price Indexes *Gergely Attila Kiss (Hungarian Central Statistical Office, Central European University)*	This goal of my project is to provide a nowcasting tool to serve the increasing pressure on official statistics to provide timely information on indicators in the present. The tool is used to nowcast monthly Production Prices in Industry(PPI) index for 6 concurrent months in real time for a large set(24) of European countries. My idea is to use standard supervised machine learning models like random forest, ridge, lasso, KNN regressions that use the World Bank Commodity Price Data as explanatory time series to nowcast the Eurozone indices. The algorithm enriches the data with the standard transformations of logging, lagging and differencing the time series to create a large pool of possible explanatory variables. Then, it uses several different feature selection algorithms to create a more concise pool of variables for using in the process. After the pool is complete it starts the hyperparameter tuning using different sample sizes, window types and one-step ahead forecasts for cross validation. By measure of accuracy the average Mean Absolute Percentage Errors (MAPE) for PPI nowcasts are around 1.5 for most of the countries. The future prospects of the algorithm is to generalize for any other time series to be nowcasted based on an eligible set of auxiliary time series considered to be able to predict the target well and to create an appropriate dashboard version of the tool to be explored and used by a large public. The most recent test shows that it can be generalized to other time series as I nowcasted the Production Volume in Industry(PVI) index for a similar set of European countries too.

Paper Session 2: Natural Language Processing	
Text Analysis of Health Equity and Disparities in IRS Form 990 Schedule H *Emily Hadley (RTI International), Laura Marcial (RTI International), Wes Quattrone (RTI International), Georgiy Bobashev (RTI International)*	Many hospitals in the United States are classified as nonprofits and receive 501(c)(3) tax-exempt status partially in exchange for providing benefits to the community. Proof of compliance is collected through tax documentation, including a free-response text section in IRS Form 990 Schedule H known for being long, ambiguous, and unaudited. This research is among the first to use text analysis approaches to evaluate this large text section. The intent is to understand if hospitals are aware of and addressing health equity and disparity issues in community benefits programming. We collaborated with public health stakeholders to identify 29 health equity and disparity themes and 152 related phrases. All themes saw increased usage from 2010 through 2019, with the largest increases in LGBTQ-related terms (1676.6%), social determinants of health (SDOH) (958.4%), and environment (522%). Yet, complementary analysis with Google Trends and semantic search suggest that hospitals may insufficiently meet community needs and priorities. Novel text analysis such as this requires close attention. We discuss rigor in the analysis, uncertainty estimation and interpretation of the results.
Manufacturing Sentiment *Tomaz Cajner (Federal Reserve Board), Leland D. Crane (Federal Reserve Board), Christopher Kurz (Federal Reserve Board), Norman Morin (Federal Reserve Board), Paul E. Soto (Federal Reserve Board), Betsy Vrankovich (Federal Reserve Board)*	This paper examines the link between industrial production and the sentiment expressed in natural language survey responses from U.S. manufacturing firms. We compare several natural language processing (NLP) techniques for classifying sentiment on our manufacturing-specific corpus, ranging from dictionary-based to modern deep learning methods. We find that deep learning models-partially trained on our data-achieve the highest sentiment classification performance on a manually-labeled sample. We assess the extent to which each sentiment measure, aggregated to monthly time series, can forecast industrial production. Our results suggest that the text responses provide information beyond the available numerical data and improve out-of-sample forecasting. We also explore what drives the predictions made by the deep leaning models, and find that a small number of words-associated with very positive/negative sentiment-account for much of the variation in the aggregate sentiment index. We expect the audience to learn which NLP methods work best in this context, and learn how to work with NLP models and Shapley decompositions.
Automating Text Cluster Naming with Large Language Models *Alexander Preiss (RTI International), John Bollenbacher (RTI International), Anthony Berghammer (RTI International), John McCarthy (RTI International), Caren Arbeit (RTI International)*	Text clustering is a way to organize unstructured text data by identifying groups within a set of documents. Government agencies use text clustering for a variety of problems, such as qualitative coding of open-text survey responses. Once clusters are generated, they must be named and described to be useful. This often involves manual review by subject matter experts, which is costly and slow. In this study, we explore the feasibility of using large language models like ChatGPT to automate the cluster naming process. We test a variety of prompting strategies on multiple datasets. We develop an evaluation rubric to assess naming quality and to compare automated names to human-generated names. Preliminary results show great promise even with simple prompting strategies. This presentation will discuss more detailed results, as well as a roadmap for real-world implementation of this technique, including cautions. Attendees will learn how large language models can reduce the cost and time required to unlock insights from unstructured text data.

Paper Session 3: Record Linkage and Disclosure Advancements	
Is Place of Birth an Appropriate Measure of Childhood Exposure? *John Sullivan (U.S. Census Bureau), Katie Genadek (U.S. Census Bureau), Carlos Becerra (U.S. Census Bureau)*	Information about the location of an individual's birth may be useful for a range of social science studies. For example, place of birth may be used as a measure of childhood exposure to conditions that follow geographic boundaries. However, survey and administrative records sources may provide inaccurate or misleading measures of place of birth. In this paper, we use administrative, decennial census and American Community Survey records, to assess the quality of place of birth information and to examine its utility as a measure of childhood residential location. Specifically, we utilize computationally efficient methods for record linkage to harmonize string reports of place of birth from applications for Social Security Numbers to standard numeric state and county codes. We then investigate multiple sources of potential bias in the use of place of birth as a measure of childhood residence, including discrepancies between reported place of birth and actual residence and selective early childhood migration away from place of birth.
Linking the 1980 Enumeration Sample to the Decennial Census *Kelsey Drotning (U.S. Census Bureau), Katie Genadek (U.S. Census Bureau)*	This paper evaluates the efficacy of methods for linking the 1980 Enumeration Sample to the Decennial Census as a source of truth data for use in a machine learning algorithm as part of the Decennial Census Digitization and Linkage project. Existing record linkage methods rely on either identification keys or fuzzy matching of string variables to evaluate when two records from different datasets are representative of the same person. These techniques have limited use for our record linkage goal because the selected datasets lack identification keys and string variables. We link records from the E Sample to the census using deterministic and probabilistic linkage algorithms. We then compare linkage rates and accuracy across the three methods. In the full paper, we will present results after review and release by the U.S. Census Bureau. Record linkage using limited identifying information presents challenges when using standard linkage techniques.
Harnessing Open-Source Tools to Link Records with PII *M Daniel Brannock (RTI International), Ed Preble (RTI International), Lynn Langton (RTI International), Marguerite DeLiema (University of Minnesota)*	Millions of Americans fall victim to mass marketing scams every year, with many scams delivered through the United States (US) Postal Service. The US Postal Inspection Service seized the "customer" relationship management databases from four criminal mail fraud enterprises, presenting a unique opportunity to gain insights on victimization patterns in mail fraud. However, the names and addresses of individuals represented in the four databases were haphazardly encoded, presenting significant challenges to identifying which transactions correspond to the same victim. The highly personal and identifiable nature of the data, compounded by the population being highly vulnerable, required that all data and computation be performed in a FIPS-certified secure computing environment. Thus, despite there being hundreds of millions of transactions to process, cloud-based record-linkage services could not be leveraged. We used the open-source Python-based text mining tool Dedupe.io, which could be downloaded and packaged to run within the secure environment, to identify individuals that appeared multiple times within and across the four scams. Active learning was applied to train models with as few as 300 labeled examples. Despite significant differences in the availability and reliability of identifiers, we were able to label transactions as belonging to the same individual with an estimated false linkage rate of 0.4% and missing linkage rate of 3.9%. The intended audience for this talk is anyone who may need to link or deduplicate records that lack a reliable, universal identifier. Attendees will learn important considerations, tips, and tools for deduplication efforts.

Paper Session 3: Record Linkage and Disclosure Advancements	
Identification of Anomalous Data Entries in Repeated Surveys *Luca Sartore (NISS/NASS), Chen Lu (NISS/NASS), Justin van Wart (NASS), Andrew Dau (NASS), Valbona Bejleri (NASS)*	Data acquired through repeated surveys can be characterized by a variety of complex relationships making the review and the vetting process time consuming. Therefore, the necessity of maintaining a high data-quality standard becomes more difficult to sustain in relatively short time frames. Thus, the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) has implemented a semi-automated revision process to improve the accuracy of the information acquired through surveys and, consequently, the accuracy of final estimates. Although some level of human intervention will remain necessary for complex cases, an anomaly detection system based on decision rules may not detect anomalies that break linear relationships. As a result, NASS has been investigating alternative approaches to modernize its anomaly detection system with more sophisticated methodologies. This presentation is concerned with a computationally efficient method that identifies four types of anomalous data entries: format-inconsistent, historical, tail, and relational anomalies. Anomalies of each type are detected and scored using a distribution-free method. A fuzzy logic technique is successively used to combine the anomaly scores and to identify the anomalous data entries. NASS survey data for livestock and corn yield are considered for comparing the proposed methodology with existing algorithms.
Refining disclosure controls for the Census of Fatal Occupational Injuries *Danny Friel (U.S. Bureau of Labor Statistics), Alyssa Gillen (U.S. Bureau of Labor Statistics), Julie Krautter (U.S. Bureau of Labor Statistics), Yvangelista Saastamoinen (U.S. Bureau of Labor Statistics)*	The Census of Fatal Occupational Injuries (CFOI) provides a complete count of workplace fatal injuries in the United States. Selecting a disclosure control method for the CFOI is challenging for several reasons. First, because CFOI is a census, sampling cannot add uncertainty to the identity of reporting units. Second, many CFOI cell counts are small, so confidentiality rules are frequently violated. Finally, the large number of companion cells provide many paths to backing out information via table differencing. The CFOI hypercube is a novel disclosure control algorithm that applies suppressions to cells defined by all combinations of levels of CFOI variables, across up to four variables at a time. There are 1,820 such combinations of four variables, and a single combination can yield over one billion cells. The hypercube has improved the confidentiality protections of CFOI data significantly: within tables, it flags up to 85% of cells for suppression. The current work proposes three refinements to the hypercube that could reconfigure and drastically reduce the number of suppressions without compromising confidentiality: 1. Publishing some or all zero-count cells. 2. Strategically limiting the number of cross-table secondary suppressions. 3. Using ranges to provide partial information about cells flagged for suppression. We demonstrate that up to 90% of cells flagged by the hypercube may be publishable after refinements are made. We discuss lessons learned from the testing and implementation of the refinements and identify considerations for other researchers who are refining custom techniques for disclosure control.

Workshop	
Chat GPT: A first-look at the utility and risks of LLMs for federal agencies *Benjamin Rogers (NCHS/CDC) and Travis Hoppe (NCHS/CDC)*	Chat GPT, the conversational AI by Open AI, was launched in November 2022. Since the launch, it has become immensely popular with over 100 million users globally. Chat GPT is based on the GPT-3.5 and GPT-4 families of large language models trained by Open AI and has proven to be effective on tasks such as text generation and programming assistance. This workshop will provide an overview of Chat GPT, discuss current National Center for Health Statistics guidance for using cloud based large language models such as Chat GPT, provide examples of Chat GPT uses that demonstrate both potential benefits and risks, discuss inherent biases within Chat GPT, and identify potential privacy concerns with Chat GPT. This workshop is focused on the federal statistical system and will be overviewing the use of Chat GPT via both the Chat GPT website and API but experience with either is not required. Attendees will learn about the potential risks of Conversational AI, how Chat GPT can be leveraged effectively and safely by users, and what current guidance is being used by federal agencies for employees to follow when using Chat GPT as well as other considerations as the conversational AI space continues to grow and expand.

Paper Session 4: Promising Practices	
On Principles for Government Open Data Curation *Jun Yan (University of Connecticut)*	With the recent open data movement and the launch of open-data government initiatives, many datasets collected by the government are increasingly being made open to the public. Such data play an important role in providing insights into civic engagement as well as business opportunities. Nonetheless, some of the open data are released with issues such as inconsistencies, redundancies, and inefficiencies. Using the New York 311 requests data as an example, we propose a few principles on the curation of the government collected open data. The principles help minimizing inconsistencies, redundancies, and inefficiencies.
Designing Against Bias: Identifying and Mitigating Bias in Machine Learning & AI *David J Corliss (Peace-Work)*	Bias in machine learning algorithms is one of the most important ethical and operational issues in statistical practice today. This talk describes common sources of bias and how to develop study designs to measure and minimize it. Analysis of disparate impact is used to quantify bias in existing and new applications. New analytic tools such as the Fairlearn package facilitate measurement of bias, supporting the development of algorithms that minimize bias. These design strategies are described in detail with examples.
Developing Documented, Trustworthy R Packages for Official Survey Statistics *Benjamin Schneider (Westat)*	When developing software that can affect public policy, it is particularly important to ensure that software is trustworthy and clearly documented. However, in R packages for official survey statistics, there are a number of particular challenges. R packages typically rely on other, rapidly-changing packages that must be continuously tested, and there are often insufficient resources available to turn proof-of-concept code into well-documented and tested software. In addition, R packages are frequently developed by statisticians skilled in using R for analysis but who are not trained in software development methods. In this talk, I outline these and other common challenges encountered in developing and maintaining R packages for official survey statistics. I provide an overview of R package development infrastructure and processes that have proven useful for addressing these challenges and identify potential solutions that have been applied in other domains of open-source software.

Webscraping in Python *Matt Ring (Westat Insight)*	This presentation will cover data collection techniques using web crawling and webscraping in Python using BeautifulSoup and Selenium. This presentation will demonstrate the benefits of automating searches, downloading data, and extracting text from websites for analysis. Methods discussed will include opening webpages and extracting text from HTML, PDF, and Word documents using Python. The presentation will provide a step-by-step demonstration of webscraping procedures. Attendees will learn what web crawling and scraping are, limitations, uses, and implementation in Python.
UI design patterns for code reuse *Weihuang Wong (NORC), Kiegan Rice (NORC)*	Code reuse is crucial if analysts want to implement a data processing and analytic pipeline in a consistent way across datasets, or if analysts want to implement a particular feature (e.g., small cell suppression) in an analytic product while leaving all other elements unchanged. Recent discussions of code reuse, such as a CNSTAT report on transparency in federal statistics, focus with good reason on tools that support code portability and sharing like version control systems or software that reproduce computational environments. Our talk, by contrast, focuses on a more prosaic aspect of code reuse: how should code itself be written to facilitate reuse? How can we develop a program so that other analysts know how to configure to get their desired output? We propose that programmers developing programs for reuse should see their programs less as scripts, and more as applications that complete a specific task. In some cases, programs can in fact be developed as simple applications, such as a Shiny app. In other cases, employing an app framework as a lens can help programmers to think about the usability of their programs and bring in UI design patterns to facilitate reapplication of code by others. We discuss various solutions to facilitate code reuse at the program-level, such as parameterized scripts, configuration files, settings panes, dialog boxes, and setup wizards.

Paper Session 5: Methodological Advances	
Analysis of Crisis Effects via Maximum Entropy *Tucker McElroy (US Census Bureau)*	Crises, such as the Covid-19 epidemic, can affect economic time series by distorting historical trend and seasonal patterns with sustained streams of extreme values in the data. A current area of research is the identification and adjustment of such extremes; this is especially important for the production of seasonally adjusted data, since most seasonal adjustment procedures rely upon a "linearized" time series with the extreme values removed. However, an ongoing challenge is the identification of a crisis' end, where the time series returns to more typical pre-crisis dynamics. One way to address this challenge is through the statistical comparison of diverse extreme specifications. A secondary challenge is to model and analyze time series data that has missing values in addition to extremes. In this paper I extend the maximum entropy framework to a generalized class of extreme values, including level shifts, temporary changes, and seasonal outliers. These are described as a particular type of stochastic process that is latent, or unobserved, such that its removal increases the time series entropy (i.e., make the data more closely resemble a Gaussian process). The proposed methods allow one to model and fit time series data using conventional tools in the presence of specified streams of extreme values, as well as missing values. Extreme value adjustment, with a quantification of mean squared error, can then be obtained along with the seasonal adjustment; there is also a test statistic to directly compare two specifications of extremes. The techniques are illustrated on monthly retail data.
Developing a Parameterized and Enterprise Ready Survey Data Review Tool *Andrew J. Dau (USDA NASS), Michael Jacobsen (USDA NASS), Ryan E. Morton (Morton Analytics LLC), Jared M, Pratt (USDA NASS), Justin P. Van Wart (USDA NASS)*	Enterprise level, cloud ready, modern solutions are generally very desired by organizations to help accomplish a variety of goals. In this paper, we focus on the modernization of internal analysis tools for survey data at the United States Department of Agriculture's National Agriculture Statistics Service (USDA NASS). Leveraging new technology, including R Shiny and Plumber APIs, USDA NASS has built a survey analysis system that replaces 26 independent legacy tools, while providing several enhancements to the analysis practices previously used. Furthermore, machine learning approaches for anomaly detection have been researched and implemented in the new tool. Lastly, throughout the entire process of development, principles of Human Computer Interaction have been addressed. From design to deployment, best practices and lessons learned are explored and discussed further.
A Computational Pipeline applied to a Nested Case-Control Study Simulation *Michelle Mellers (Uniformed Services University of the Health Sciences and Henry M Jackson Foundation), Celia Byrne (Uniformed Services University of the Health Sciences)*	Through FAIR principles NIH encourages Findability, Accessibility, Interoperability, and Reusability of data. In planning a study, creating a formal pipeline and testing with simulated data helps to implement these principles for a study. A formalized pipeline describes the flow of data processing and analytic steps for a project, and helps to organize the required computational tools. Creating simulations of varying complexities can help to improve understanding of a study's data structure, analytic plan and potential interpretations. Formal data structure description facilitates early involvement of collaborators. However, formalized pipelines with simulated data are rarely implemented. We illustrate the use of a pipeline and a simulation of a nested case-control study of the impact of chemical exposures on breast cancer rates. To address the scientific hypothesis our pipeline identified all required steps to implement the simulation study. This process included developing file structure organization, writing code files, and testing the simulation for errors. Outcome data and related demographic information were extracted from medical health records. To study exposure data, we simulated the correlated chemicals, analyzing them separately, and together in a mixtures model. The computational component of our pipeline applied reproducible computing techniques to this simulation study. Formalizing the steps of pipelines and testing with simulated data contributes both to a study's reproducibility and the FAIR principles. The views expressed are those of the authors and do not reflect the official views of the Uniformed Services University, HJF or the Department of Defense.

Paper Session 5: Methodological Advances	
Deep Neural Network based mass imputation for data integration *Sixia Chen (University of Oklahoma Health Sciences Center), Chao Xu (University of Oklahoma Health Sciences Center)*	Although probability samples have been regarded as the gold standard to collect information for population-based study, non-probability samples have been used frequently in practice due to low cost, convenience, and the lack of the sampling frame for the survey. Naïve estimates based on non-probability samples without any adjustments may be misleading due to selection bias. Recently, a valid data integration approach that includes mass imputation, propensity score weighting, and calibration has been used to improve the representativeness of non-probability samples. The effectiveness of the mass imputation approach depends on the underlying model assumptions. In this paper, we propose a modified deep learning for mass imputation and compare it with several modern machine learning-based mass imputation approaches, including generalized additive modeling, regression tree, random forest, and XG-boosting. In the simulation study, deep learning-based approaches have been shown to be more robust and effective than other mass imputation approaches against the failure of underlying model assumptions under non-linearity scenarios. We further evaluate our proposed method by using a real application. The target audience is researchers, students, and faculties in survey sampling field. We are mainly using R for the computation. We expect attendees to learn some machine learning based mass imputation methods.
Weswgt: A new R tool for Survey Weight Adjustments *Jianru Chen (Westat), Benjamin Schneider (Westat), Minsun Riddles (Westat)*	Government surveys provide critical data spanning a wide range of topics that both government agencies and the public need to make informed decisions. Weighting adjustments are often required to account for the complex sample designs of these surveys and adjust for non-sampling errors such as nonresponse or coverage. We introduce our new R package, Weswgt, which conducts weighting class adjustments, calibration to known population totals, and weight trimming. Weswgt is user-friendly and customizable, allowing users to specify various weighting parameters. We demonstrate the capabilities of this package with an example of a two-step weighting adjustment procedure: in the first step we illustrate a weighting class adjustment to account for nonresponse, where we have collapsed weighting class cells after evaluating the adjustment factors and sample sizes, and in the second step we rake to population totals to ensure that estimated totals agree with the known totals in multiple dimensions, while trimming weights concurrently. We also discuss the potential applications of this package in various survey fields.

Lightning Talks 1: The Next Generation	
Using CANS scores in the Community Risk Result (CRR) project in R *Jamie Joseph (Vanderbilt University), Rameela Raman (Vanderbilt University)*	The Juvenile Justice Reform Act of 2018 mandates a validated risk and needs assessment for making decisions and recommendations for programming/treatment. We are in the process of examining the predictive validity of the risk and needs assessment tool for youth re-arrest within 90 days of initial arrest in Madison County. Particularly, we utilize the Juvenile Justice Child and Adolescent Needs and Strengths (JJ CANS) assessment to inform service recommendations through an analysis in R. We begin by visualizing and exploring the data on re-arrests and scores broken down by various demographics and predictors of re-entry. We then use factor analysis and logistic regression in order to quantify risk factors and variables that contribute to recidivism.
Open Source Software in the Federal Government: An Analysis of Code.Gov *Gizem Korkmaz (Westat), Rahul Shrivastava (Westat), Anil Battalahalli (Westat), Ekaterina Levitskaya (Coleridge Initiative), José Bayoán Santiago Calderón (Bureau of Economic Analysis), Ledia Guci (Bureau of Economic Analysis), Carol Robbins (National Science Foundation)*	Open source software (OSS) is ubiquitous, serving as specialized applications nurtured by devoted user communities, and as digital infrastructure underlying platforms used by millions of people. OSS is developed, maintained, and extended through the contribution of independent developers as well as people from businesses, universities, government research institutions, and nonprofits. Despite its prevalence, the scope and impact of OSS are not currently well-measured. Recent policies of the U.S. Federal Government promote sharing of software code developed by or for the Federal Government. While the policy to promote reusing and sharing of software created with public funding is relatively new, public funding plays an important and not fully accounted role in the creation of OSS. This paper aims to measure the scope and value of OSS development in the U.S. Federal Government. We collect data from Code.gov, the government's platform for sharing OSS projects, and study contributions of agencies. The dataset contains 17K repositories from 21 agencies, with the majority of contributions originating from the DOE, NASA and GSA. In addition, we collect data on development activity (e.g., lines of code, contributors) of the repositories on GitHub, the largest hosting facility worldwide. Adopting a cost estimation model from software engineering, we generate estimates of investment in OSS that are consistent with the U.S. national accounting methods used for measuring software investment. Finally, we generate and analyze collaboration network resulting from cross-agency contributions to repositories and explore the centrality of agencies in the network.
Timely Examination of Survey Data Quality *Winnie Xu (Westat)*	During the COVID-19 pandemic, some survey programs with in-person data collection offered telephone as a response mode to continue data collection. Consequently, a need to monitor the impact of the new response mode on data quality arose. In this talk, we share our recent experience in building a dashboard to monitor item response rates and compare them by data collection factors, including response mode and respondent location, to examine the effect of changes to data collection methods on data quality in a timely manner while the data is still being collected. The talk covers how the underlying data were constructed and how the dashboard's contents and its format were determined to address the question at hand. We illustrate the construction and use of the dashboard with the Public Use Microdata of the American Community Survey.

Lightning Talks 2: Evaluations and Models	
Evaluation of a method for georeferencing farms *Robert Emmet (USDA, National Agricultural Statistics Service), Kevin Hunt (USDA, National Agricultural Statistics Service), Rachael Jennings (USDA, National Agricultural Statistics Service), Kara Daniel (USDA, National Agricultural Statistics Service), Denise A. Abreu (USDA, National Agricultural Statistics Service)*	The USDA National Agricultural Statistics Service (NASS) has developed a process to georeference the NASS list frame (a list of all known US farms) using USDA Farm Service Agency (FSA) administrative data. Each year FSA collects georeferenced administrative data on US farms that participate in at least one USDA program. Although there is extensive overlap in the NASS list frame and the FSA administrative data, each has farms not covered by the other. Thus, to fully georeference the NASS list frame, non-FSA farms require a separate georeferencing process. Non-FSA areas of possible agricultural activity are identified using maps of FSA farms and cultivated areas. Potential non-FSA farms are identified by linking these agricultural areas to georeferenced county assessor's parcel land ownership records. Agricultural classification surveys are sent to the potential non-FSA farms not on the list frame to determine whether or not they satisfy the definition of a farm, which is any operation that produces and sells or has the potential to sell at least $1000 of agricultural products in a year. In this paper, the non-FSA georeferencing process for identifying farms is evaluated for 11 Midwestern states. The results found 3,933 operators that were not on the list frame and in scope, out of 20,634 operators that were not on the list frame and responded to surveys; 81% of these 3,933 operators were in metropolitan or micropolitan statistical areas, and many were in areas with known underrepresented populations, such as Amish communities. Implications for using this approach to increase the coverage of the NASS list frame are discussed.
treecompareR: an open-source software package to visualize hierarchical chemical classifications *Paul Kruse (Oak Ridge Institute for Science and Education), Caroline L. Ring (Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency)*	We introduce the open-source R package "treecompareR", which offers a framework to visualize and explore chemical data with ClassyFire chemical classifications. ClassyFire is a web tool that uses chemical structure to classify chemicals automatically in a hierarchical "tree of life" ontology (ChemOnt). The ClassyFire chemical classification framework is useful for evaluating the chemical landscape of chemical data relevant to toxicity, exposure, and risk, as part of a cheminformatics approach in new approach methodologies for rapid chemical risk characterization. The treecompareR package allows the user to: • Visualize the ChemOnt ontology as a tree diagram • "Prune" the tree: select and visualize branches of interest in more detail • Assess & visualize similarity of ClassyFire classifications between two lists of chemicals • Highlight tree branches according to list presence, prune tree to represent one list and highlight branches to indicate presence in the other list • Identify portions of data sets that are highly similar or dissimilar to each other by leveraging hierarchical clustering in conjunction with heatmaps and a variety of similarity measures • Annotate tree diagrams with additional chemical-specific data to explore relationships with chemical class • Physical-chemical property data, key exposure descriptors, hazard descriptors The target audience includes those with an interest in creating visualizations to display and explore chemical data. The audience should take away a few examples of the functionality provided by treecompareR. The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.

Lightning Talks 2: Evaluations and Models	
Building a Modeling Platform for DC Government *Hersh Gupta (DC Office of the Chief Technology Officer), Gautam Chakravarty (DC Office of the Chief Technology Officer)*	Statistical modeling and machine learning (ML) is used to address challenges in the public sector using prediction- and inference-based models. From predicting structural fire risk and detecting lead pipes to explaining the factors behind extended lengths of stay in homeless shelters and the likelihood of chronic absenteeism, agencies increasingly look to adopt these techniques to support program operations. With these models, great care is essential to ensure that models are secure, valid, and their outputs are trustworthy and reproducible. In government, much of the risk of adopting Commercial off-the-shelf software (COTS) ML models stems from their "black box" qualities: there is little transparency into how these models work, how valid their predictions are, or whether they contribute to inequities in public service. To enable the use of ML while mitigating potential risks, Office of the Chief Technology Officer (OCTO) in the District of Columbia is developing an ML platform for agency use. This platform uses Free and Open Source Software (FOSS) that follows industry best practices to support the machine learning operations lifecycle. Learn how the data team at OCTO is building a modeling platform that integrates the following technologies: a Hadoop data lake for storage, Spark and H2O to build, interpret, validate, and serve models, AirFlow to orchestrate data flows, MLFlow as a model registry, and Kubernetes to integrate, manage, and scale these applications. OCTO is piloting using this platform for developing and deploying models for partner agencies. We will review the tradeoffs associated with building a modeling platform and lessons learned along the way.

Lightning Talks 3: Refine Files or Revise	
Expanding the R Shiny Apps in the growclusters Package *Randall Powers (Bureau of Labor Statistics), Terrance Savitsky (Bureau of Labor Statistics), Wendy Martinez (US Census Bureau)*	Growclusters for R is a package currently in development that estimates a partition or grouping structure for multivariate data. It does this by implementing a hierarchical version of k-means clustering, which accounts for possible dependencies in a collection of data sets, where the elements of the collection correspond to known sub-groups. Each component data set in the collection draws its cluster mean from a single, global partition that we seek to estimate with growclusters. This talk follows a 2022 GASP talk that focused on the creation of two R Shiny apps to be included in the growclusters package. This talk focuses on a third R-Shiny app that implements novel ways of visualizing the results of the clustering across the collection of data sets. Data obtained from a collection of 2000-2013 journal articles from the Bureau of Labor Statistics (BLS) Monthly Labor Review (MLR) will be used to illustrate the R-Shiny app. In this case, the known sub-grouping corresponds to the year of publication.
Sampling Frames Simulation with Copulas in R *Eli Kravitz (New Light Technologies), Daphne Liu (US Census Bureau, Center for Economic Studies), Stephanie Coffey (US Census Bureau, Center for Economic Studies)*	Copulas are a popular approach for modeling multivariate joint densities. Their flexibility allows analysts to model the univariate marginal distributions separately from the correlation structure. We present a real-world problem of generating a synthetic sampling frame with correlated covariates. We use Gaussian copulas to simulate random variables with an interpretable correlation structure. We demonstrate this process using R and provide modular, reusable code to facilitate wider adoption.
Tiling your Parquet files for the greatest efficiency *Stas Kolenikov (NORC at the University of Chicago)*	Arrow Parquet is a storage-efficient format for large data bases frequently used in technological applications. It is a smart, metadata rich format, in the sense that the data set is aware of its content (e.g., minimum and maximum values of numeric variables). The particular strength of the Parquet format is the multi-file storage, where the files can be grouped according to the specific variables that are commonly filtered or tabulated. We demonstrate the efficiency gains working with a version of the USPS Computerized Delivery Sequence File (CDS or CDSF) licensed from a vendor, the data set commonly used at NORC to draw the general population samples which contains all addresses that receive mail in the United States during a given month. We present the different ways to break down the data set (e.g., by states, by counties, by ZIP codes) and compare performance benchmarks in typical CDSF-based operations (e.g., loading all residential addresses in a state to take a sample).

Lightning Talks 3: Refine Files or Revise	
Exploring the analytical power of input-output tables in the UK context *Julen Grube Doiz (UK Office for National Statistics), Eric Crane (UK Office for National Statistics), Andrew Banks (UK Office for National Statistics)*	The Input-Output tables (IOTs) produced by the Office for National Statistics (ONS) allow analysts to estimate the full impacts on the UK economy of changes in final demand. The IOTs are used by analysts and economists to understand the linkages within the UK economy. To make these insights more easily accessible, the Supply and Use team at the ONS have created an innovative IOT dashboard in partnership with the Data Science Campus. The presentation will focus on this tool and is aimed at analysts, economists and policymakers. The dashboard provides easy to use analytical tools to enable quick and efficient insights to address common questions from policymakers about the UK's supply chains and widens awareness of what input-output tables can do. There are three tools within the dashboard. The first allows users to determine the exposure of industries to products in their production process. The second tool allows users to determine the key product inputs to produce a product output. The third tool allows users to simulate simple demand shocks to the economy and measure the impact on different key economic variables. The IOT data was integrated into the dashboard in two steps. First, the initial data processing and formatting was undertaken in Python using the pandas data analysis package. Second, the data was converted from the traditional set of matrices to a long table in a tidy data frame format. We hope the dashboard will stimulate greater awareness of the value and uses of input-output tables in answering key economic questions. Additionally, we hope the presentation provides conference attendees with insights into the analytical potential of our dashboard.

Lightning Talks 4: Tools	
Toward a Tool for Automated Disclosure Risk Review of NCES Data Products *John Riddles (Westat), Michael Armesto (Sanametrix), Tom Krenzke (Westat), Jennifer Nielsen (National Center for Education Statistics)*	Data products produced by NCES, including NCES-authored papers and public reports, must go through a rigorous process to identify potential disclosure risks before they can be released. A number of standards have been developed for this process, but it is still largely a manual and time-consuming task. To both save effort and to potentially flag otherwise unidentified disclosure risks, a Shiny app and underlying analysis tool are being developed to analyze submitted documents for disclosure risks. We will present an overview of this tool, its interface, and the types of document review procedures it currently performs.
Using NLP to Access to the Unemployment Insurance System *Siobhan Mills De La Rosa (American Institutes for Research), Meghan Coffee (American Institutes for Research), and Samia Amin (American Institutes for Research)*	This lightening talk will describe AIR's use of Natural Language Processing to enhance claimant understanding and access to the unemployment insurance (UI) system. The presentation will highlight practical tools our team is using to assess the readability scores of current UI documents, prioritize documents for revision, and generate practical revisions that will make documents easier to understand across a more diverse set of claimants. The presentation will also describe how NLP and behavioral science principles can be married to improve the experiences of individuals using government services.
Incorporating Survey Weights into Structural Topic Models *Caroline Lancaster (NORC at the University of Chicago), Brandon Sepulvado (NORC at the University of Chicago), Josh Lerner (NORC at the University of Chicago), Evan Herring Nathan (NORC at the University of Chicago), Stas Kolenikov (NORC at the University of Chicago)*	Surveys often include open-ended questions that elicit text responses. When there are few responses, researchers can manually review them, but manual review becomes infeasible as the number of responses increases, for example, into the thousands or tens of thousands. When researchers want to know the thematic composition of such text responses, topic modeling—an unsupervised natural language processing (NLP) method that identifies topics from a set of texts—offers a potential solution. However, current topic modeling software does not allow the incorporation of survey weights, which can bias resulting statistics based upon the open-ended response text. This presentation will describe our team's efforts to incorporate survey weights into the popular R implementation of structural topic models (STMs). STMs identify topics within a set of texts and allow researchers to estimate how the prevalence and content of topics differ by covariates, such as gender and race/ethnicity. We will present a case study comparing analyses of open-ended survey items using traditional (unweighted) STMs and the weighted STMs that we propose.
Traffic tracker: A gentle introduction to Python, APIs, and GitHub actions *Emily Mitchell (AHRQ - Agency for Healthcare Research and Quality)*	This talk explores a practical example of how Python and GitHub Actions can be used to automate the downloading and storage of web traffic data for GitHub repositories. Measuring and analyzing user traffic is valuable for federal statistical agencies that want to provide high-quality code to their users. However, unlike internally-hosted sites that that can leverage tools like Google Analytics to track website traffic, GitHub repositories only provide user traffic data for the past two weeks. To address this limitation, I implemented a solution that calls the GitHub API from Python to programmatically capture traffic statistics, and then used GitHub Actions to automate this process on a weekly basis. Using this process, we can capture and store web traffic for GitHub repositories beyond the default time frame, thus providing a long-term perspective of the impact of our repositories.

Lightning talks 5: Data Visualization	
Python Dash for Visualization: Reusable Dashboards *John Lombardi (US Census Bureau)*	For the Economic Census, various metrics and visualizations are needed to track things like the progression of the survey lifecycle, or general summaries of data as it comes in from respondents. Python Dash provides a rich ecosystem for developing custom interactive visualizations to understand data in real time. Data scientists at the U.S. Census Bureau have developed code to create reusable and extensible components for dashboarding, enabling new features or new dashboards to be implemented with ease. We will demo both our system for developing with Dash, as well as an overview of some of the dashboards we have created for the Economic Census.
A Data Viz Makeover: The Evolution of A Visualization Showing State SNAP Data *Brian Knop (Census Bureau)*	In 2017, the first interactive data visualization was published to Census.gov, showing county-level SNAP eligibility and access for New York state. In the time since then, the Bureau has developed internal guidelines for creating data visualizations and Tableau training for visual developers, resulting in over 100 data visualizations that can be found in our online interactive gallery. In 2021, we updated the New York SNAP visualization to include data from 15 additional states (the data includes SNAP administrative records from participating states) and improved the design of that initial visualization. Now, five years after the publication of the initial SNAP visualization and two years after publishing a multi-state update, we have created another update with 23 states, a mobile version, and a newly redesigned interface. In this presentation, I share the features of our SNAP data visualization and talk about some of the lessons learned during the process of redesigning our data visualization for a better user experience. I highlight insights about SNAP eligibility and usage that the data visualization offers, as well as best practices that we applied to the redesign efforts. The visualization represents the joint efforts of the US Census Bureau, the USDA's Food and Nutrition Service, the USDA's Economic Research Service, and our state partners to increase understanding of current SNAP program access and inform future SNAP program outreach. It uses American Community Survey data linked to state administrative records to model estimates of SNAP eligibility and access rates at the state and county levels and for a wide range of characteristics.
Developing a Data Quality Scorecard Dashboard to Assess Data's Fitness for Use *John Finamore (NCSES), Elizabeth Mannshardt (NCSES), Lisa B. Mirel (NCSES), Julie Banks (NORC at the University of Chicago), F. Jay Breidt (NORC at the University of Chicago), Benjamin R. Peck (NORC at the University of Chicago), Kiegan Rice (NORC at the University of Chicago), Zachary H. Seeskin (NORC at the University of Chicago), Lance A. Selfa (NORC at the University of Chicago), Grace Xie (NORC at the University of Chicago)*	Assessing data quality is critical to determine a data source's fitness for use. The US Federal Committee on Statistical Methodology (FCSM) has developed a data quality framework that provides a common language for federal agencies and researchers to make decisions about data quality. Recently, the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation and NORC at the University of Chicago collaboratively developed an approach to generate scorecards based on the eleven dimensions outlined in the FCSM Data Quality Framework. Our "Federal Data Quality Assessment Framework" (FDQAF) scorecard generates a score in each of the eleven FCSM data quality dimensions for a specific use case via binary questions about the use case, its data sources, and relevant metadata. We illustrate the use of R Shiny and RMarkdown to develop an interactive interface for applying FDQAF to assess a data source's fitness for use.