CHOICES

AAEA
Agricultural & Applied Economics Association

Volume 39. Quarter 1

An Outlook on Evidence-Based Policy Making

Lauren Chenarides, Andrew S. Hanks, Amelia Finaret, George Davis, and Andrea Carlson

JEL Classifications: C81, C82, D04
Keywords: Agricultural policy, Applied economics, Data linkages, Food policy

Introduction

Given the many concurrent challenges both in the United States and abroad—including the COVID-19 pandemic, economic crises, inequality, and climate change—it is becoming increasingly necessary for researchers to rely on multiple datasets to answer their research questions (Wilson et al., 2021). Individual datasets often lack the information and depth needed to address complex societal and economic issues, and no single dataset contains all the relevant variables needed for the most accurate analysis of many of the most pressing food and agricultural policy issues. Linking multiple datasets is a strategy that allows researchers to generate comprehensive, high-quality data for empirical studies. Data linkages enable researchers to better understand the markets they study and evaluate the implications of and potential need for current and suggested policies.

In this article, our objective is threefold: First, we emphasize the need for and value of data linkages for producing policy-relevant research. Second, we explain four barriers that researchers face when linking datasets collected across agencies and organizations. Third, we present current solutions and emerging opportunities that facilitate researcher access to the most state-of-the-art methods available that can overcome the barriers identified. We derive these insights from a 2-day workshop in which leaders in agricultural, labor, and health economics shared their expertise regarding data linkages and emphasized the need for these linkages in producing policy-relevant research.

Why Data Linkages?

The 2018 Foundations for Evidence-Based Policymaking Act (H.R.4147), originally introduced in October 2017 by Representative Paul Ryan (R-WI-1), codified the call to action for evidence-based policy making. This law, coupled with two other notable federally sponsored reports—A Commission on Evidence-Based Policymaking ("Commission") and A Consumer Food Data System for 2030 and Beyond—articulates the importance of investing in data infrastructure designed to produce rigorous evidence and inform policy making (Abraham et al., 2018; National Academies of Sciences, Engineering, and Medicine, 2020). The Commission's report stated, "greater use of existing data is now possible in conjunction with stronger privacy and legal protections, as well as increased transparency and accountability," but current practices "are not currently optimized to support the use of data for evidence building" (p. 1). Advances in technology, storage, privacy protection, and analytical methods have provided researchers with access to more data than ever before, but the infrastructure to harness the power of these advancements is not yet in place.

As more resources become available to researchers, agricultural and applied economists find themselves at an inflection point. Integrating the abundance of data into existing research programs requires building an interoperable network of existing data assets collected by numerous public and private organizations. Such an interoperable network would be a system that connects datasets in a coordinated way with minimal or no effort by the end user (i.e., an application that significantly reduces the costs of linking data). The value of data linkages is highlighted in the National Academies report, which expresses a vision to "build a comprehensive, integrated data system to efficiently deliver credible evidence for informing research and policy" (p. 16). The terms "comprehensive" and "integrated" suggest that datasets should be linked, or have the potential to be linked, even when procured separately by different organizations and for different research purposes (Bailey et al., 2020). Investing in infrastructure that supports data linkages, therefore, provides an opportunity to respond to the recent legislation and calls set forth by both the Commission and the National Academies.

Researchers have defined data linkages as "the bringing together from two or more different sources, data that relate to the same individual, family, place, or event" (Holman et al., 2008, p. 767; Emery and Boyle, 2017, p. 615). We build on this definition and conceptualize data linkages as combining two or more datasets such that separate entities (e.g., agencies, countries, organizations) collect the data and each dataset contains a key variable of interest that is necessary for answering a researcher-specified question. For example, the U.S. Department of Agriculture's (USDA) National Household Food Acquisition and Purchase Survey (FoodAPS) data combines data from survey, interview, and food acquisition recall tools as well as administrative records for participation in programs such as the Supplemental Nutrition Assistance Program (SNAP). Data from the surveys, interview, and food recall tools are all collected by the USDA and are part of the same survey, so these data would not be considered linked. However, combining the USDA-collected household FoodAPS data with SNAP administrative data collected at the state level creates a data linkage.

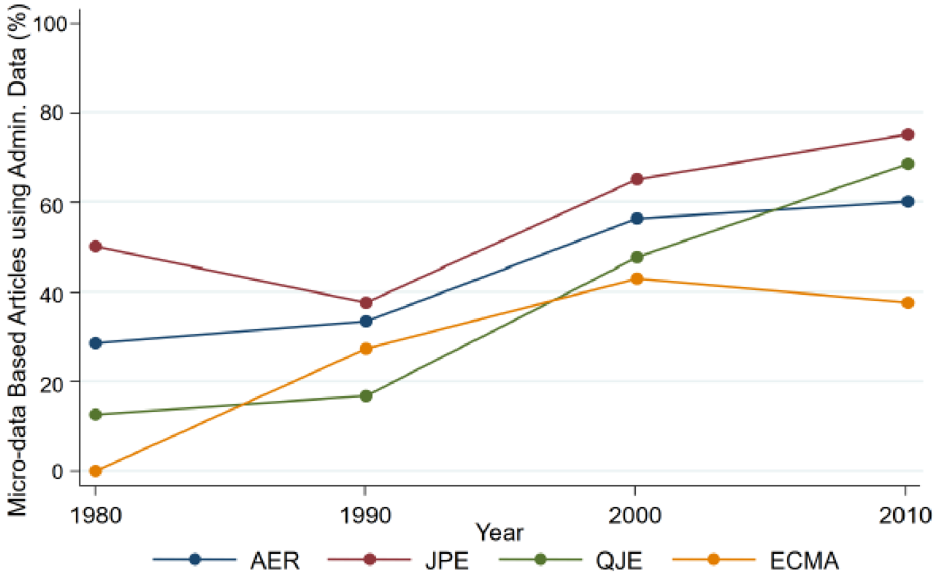
In the workshop we hosted, Dr. Bruce Weinberg suggested that linking data might yield exponential growth in the number of potential questions researchers can answer. To illustrate this point, the researcher's problem is a typical economics problem: How to allocate their resources (inputs) to achieve a desired level of outputs (isoquant)—a production function problem. For simplicity, suppose the production of research is a function of two primary inputs—data and researcher's labor or skill. The data input includes data products used to generate analytical results. The labor input includes time spent to access data as well as to learn and apply the necessary analytical methods (e.g., econometrics, remote sensing, and machine learning). The output would be research publications, grant proposals, or dashboards. Based on these two inputs, we can imagine the researcher choosing a dataset consistent with their skill level to reach a given isoquant (output level). This choice clearly would be affected by their knowledge and familiarity with the dataset. Platforms that lower search costs for data linkages change the research technology and are also likely to be labor-biased technical changes. As a result, labor becomes relatively more efficient and the time cost of reaching the same output (isoquant) is lower.

Simply producing more research outputs is not sufficient for generating policy-relevant research. The quality of a research output depends on the quality of the data and methods used to generate it. As an example, we turn to another presenter's insights on the use of administrative data to produce high-caliber research publications. According to Dr. Timothy Beatty, the use of administrative data in publications in leading economics journals doubled, on average, between 1980 and 2010 (see Figure 1). Administrative data are largely regarded as the highest quality microdata, as they allow for very large sample sizes, have minimal attrition, and track individuals or households over long periods of time. These data are one example of a common pool resource available for research-oriented producers. With advancements in data linkages beyond administrative data, the frontier for high-quality research production will continue to expand.

Barriers to Using Data Linkages

Despite the value and benefit of linking disparate datasets to enhance research and evidence for policy, multiple barriers may inhibit researchers. These roadblocks vary depending on factors such as researcher experience, funding availability, and

Figure 1: Use of Administrative Data in Publications in Leading Journals, 1980-2010



Note: Image from Tim Beatty's presentation: Why should we care about data linkages? (October 1, 2021). AER=American Economic Review; JPE=Journal of Political Economy; QJE=Quarterly Journal of Economics; ECMA=Econometrica.

Source: Chetty (2012).

professional networks. In general, researchers may not know 1) where to find appropriate data, 2) how to access available data, 3) which data are linkable, and 4) how to link data. The following sections discuss each of these barriers.

Barrier 1: Finding Data

A common obstacle that researchers, especially the most junior researchers, face is the substantial search costs to identify and locate data. The current data "search and discovery" process that governs the beginning phase of the research lifecycle usually occurs by conducting literature reviews, web browsing, or establishing professional networks, but this process can be slow and labor-intensive. Upon identifying an appropriate dataset, there is also a significant learning curve: A researcher must spend many hours familiarizing herself with the structure and contents of the dataset. Data repositories, such as the Inter-university Consortium for Political and Social Research (ICPSR, https://www.icpsr.umich.edu/web/pages/), reduce search costs for researchers, but these services vary by the associated costs of constructing an analytical sample from data housed within a repository (i.e., ease of use) and level of instruction offered to shorten the learning curve (i.e., instructional value).

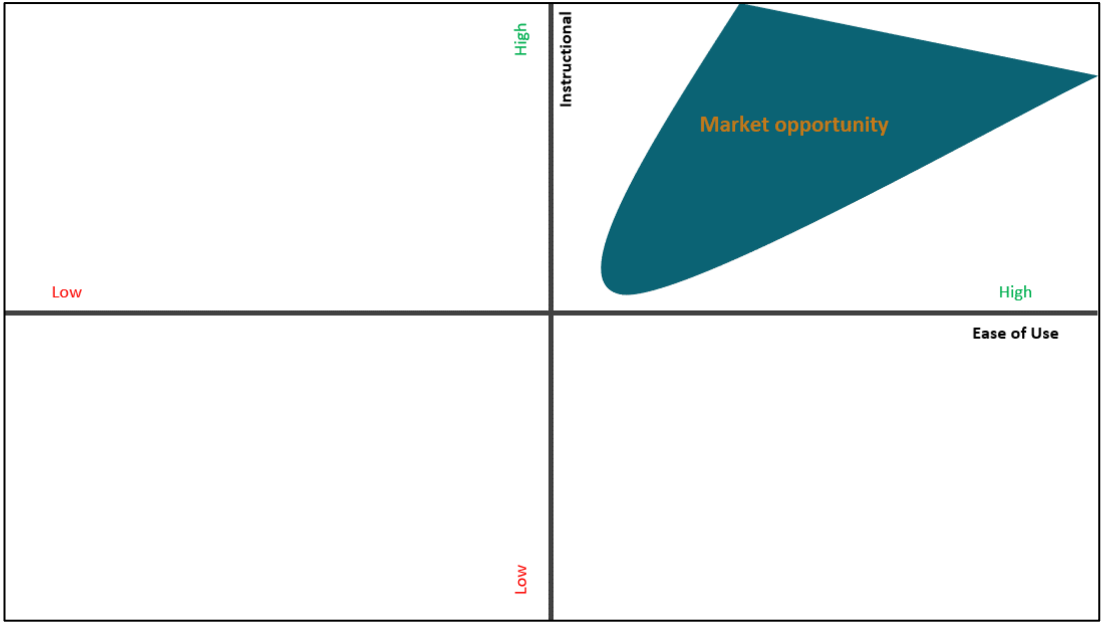
In Figure 2, we conceptualize how data repositories could be perceived based on ease of use and instructional value. We define the ease-of-use value of a data repository (the *x*-axis) as the degree to which the repository facilitates the analytical sample construction process. The value of dataset along the *x*-axis is a function of 1) the details provided about the datasets in a repository that are necessary to assemble an appropriate analytical sample and 2) the content and dimensions of the datasets offered (including whether variables in these datasets can be linked to other data resources). We define the level of instructional value (the *y*-axis) as the extent to which the repository provides users with resources, such as codebooks and data dictionaries. A repository would have high informational value, for example, if it included important details about the data generating process of the datasets, sampling procedures, and how the data can be linked with other datasets. Simply, the *x*-axis is the "what" and the *y*-axis is the "how": What is available, and how can it be used?

In our preliminary review of several data repositories, we note that many repositories offer access to users to construct rich datasets included in the repository, but few offer high instructional value, and, to our knowledge, none offer support on linking repository datasets to external data.

Barrier 2: Accessing Data

We characterize dataset accessibility as existing on a continuum from publicly available to strictly confidential. Publicly available data have the lowest accessibility constraints in that there is minimal to no "entry cost" (e.g., creating a user account). These data, however, often lack the resolution of confidential or restricted-use data and exclude important microlevel information, such as location, that matters for policy purposes. Certain restricted-use data are also only accessible by users who are employed by a federal agency or an employee

Figure 2: Illustrative Approach on Market Research of Current Data Repositories



Notes: Examples of data repositories in the agricultural and applied economics space include IPUMS, ICPSR, USDA ERS Food-Related Data Sources. IHSN, FRED Economic Data, Harvard Dataverse, Food Systems Dashboard, USDA Ag Data Commons.

of a U.S.-based university or institution. Accessing restricted-use data requires users to sign a data user agreement (DUA) or nondisclosure agreement (NDA) with the data provider, agreeing not to disclose sensitive information contained in the confidential data. These additional constraints require time, resources, and professional connections.

The Federal Statistical Research Data Centers (FSRDCs) provide access to some of the most detailed, albeit restricted, microlevel records for many datasets collected by federal principal statistical agencies such as the Census Bureau and the Bureau of Labor Statistics. Dr. Mark Prell explained that the U.S. Census Bureau has agreements with states to hold state administrative data for programs such as SNAP and WIC. FSRDCs are the entities through which researchers (with approved projects and Special Sworn Status) access the SNAP or WIC administrative data. Researchers must apply for access to these data; once they have access, they can link these records to census or other data, such as the Longitudinal Employer-Household Dynamics (LEHD) files, to study households over time. While these are excellent opportunities for impactful, policy-relevant research, all analysis must be contained within the FSRDCs, and disclosure review can be a time-consuming process.

To reduce barriers to accessing restricted access data, the federal statistical system recently adopted a standard application process (SAP) to access multiple confidential data assets from federal statistical agencies.[1] As is the case with all restricted data, the researcher will need to describe in writing the proposed research plan and what she intends to do with the data. Researchers will generally need to pay to access a data enclave or travel to an FSRDC. As part of the data use, researchers will also need to submit disclosure requests before exporting results from the secure environment and often allow the issuer to review any work before the researcher shares results beyond the preapproved research team.

In their workshop presentations, Dr. Beatty and Dr. Joseph Cummins both highlighted the unfortunate reality that researchers are less likely to publish manuscripts in top-tier journals when their data are from common, publicly available sources. This relationship between data accessibility and journal rank reveals an important mechanism through which asymmetric access barriers across institutions contribute to variation in publication potential. Researchers at well-funded, well-connected institutions can access these resources at much lower costs relative to researchers at institutions that lack funding to travel to the FSRDCs or access a secure enclave or the infrastructure and legal support required to oversee data access agreements. Dr. Beatty also cautioned that administrative data are not collected for research purposes; due to their complexity, researchers often require considerable institutional knowledge. Moreover, the sensitivity and confidentiality of information contained in administrative data often requires special training in how to handle data subject to the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).

Barrier 3: Identifying Linkable Data

Arguably, the single best dataset for applied economists would contain unified data on production, trade, capital, labor, materials, nutrition, health, expenditures, prices, income, time use, program participation, household composition, education, preferences, and demographics. Unfortunately, no such dataset exists. Individual datasets that contain elements or combinations of these variables, however, are available to researchers. Advances in computing, data storage, and cloud-based secure data enclaves have opened opportunities for researchers to access these data in ways that were previously cost-prohibitive.

While opening access has expanded the ways in which data can be integrated, or linked, researchers often default to using a single dataset because they do not know of possible linkages. This knowledge barrier can serve as an impediment to pursue certain research topics. A common practice among data proprietors is to provide codebooks that described the elements of each dataset, so that users are aware of what information is collected and the process by which the data have been generated. A clever researcher will use codebooks to identify possible linkages between datasets, but this process can be labor-intensive and may not produce desired results. Also, because any two datasets vary by scope (e.g., a region sampled) and resolution (e.g., the unit of observation), it takes empirical researchers time to reshape datasets to establish spatial or temporal relationships and use techniques such as matching, modeling, and spatial joining to merge or combine disparate datasets. Given that linking datasets is becoming more commonplace, providing best practices to researchers to streamline the linking process would eliminate or reduce the cost of adopting new data. Providing systematic support for enabling matching efforts would reduce user error and threats to replicability.

One such case where linking services are offered to users is in an FSRDC, where researchers can link census data, such as the American Community Survey (ACS), to employment data files, such as the Longitudinal Employer Household Dynamics (LEHD), using protected identification keys (PIKs). The census generates these PIKs based on several common identifiers, such as name and birth year, and estimates the likelihood of a match. Another example in food research are the Food Security Supplement and the

[1] For more information on the Standard Application Process, please visit https://ncses.nsf.gov/about/standard-application-process.

American Time Use Survey (ATUS), both of which are supplements to the Current Population Survey. Public data centers, such as Integrated Public Use Microdata Series (IPUMS), help generate datasets that link individual microdata files for researchers, reducing adoption costs.

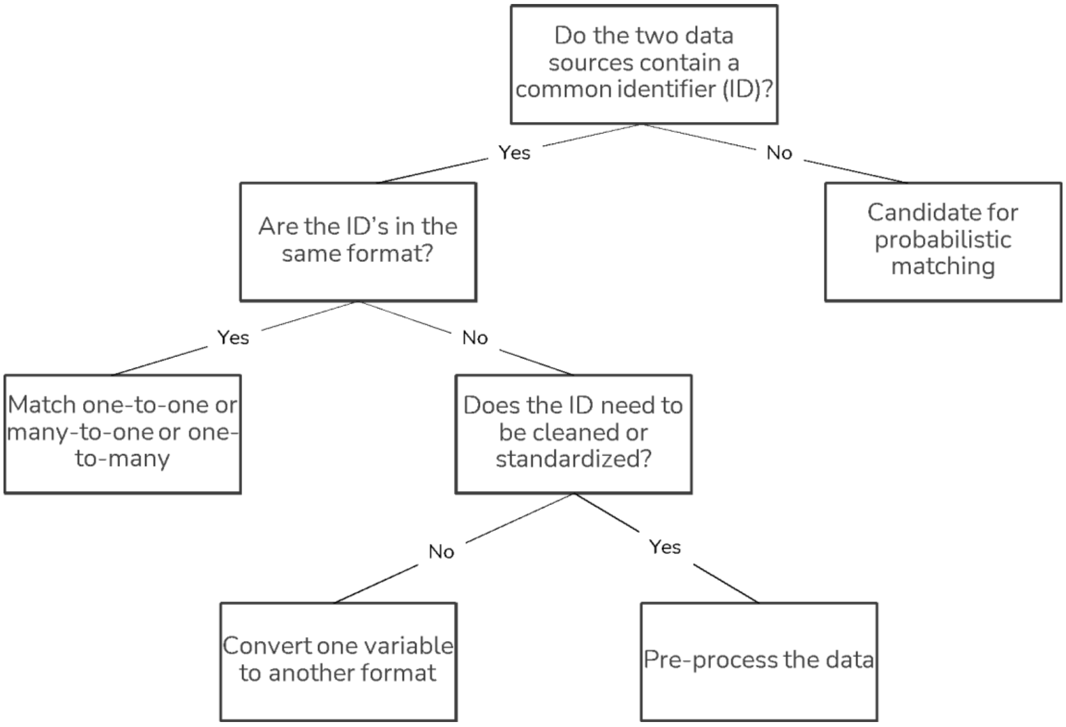
Barrier 4: Linking Data and Data Quality Considerations

For many applied economists, not only agricultural economists, the technical skills needed to link multiple datasets may be a barrier. While various methods to link data exist, we broadly classify these methods as either deterministic or probabilistic (see Figure 3 for a general framework of when to use deterministic or probabilistic matching). Deterministic matching is performed when two datasets share a common identifier. This identifier must be associated with the observational unit on which the researcher intends to match. Deterministic matching is only possible when datasets share a common identifier. Datasets collected by different agencies most likely do not have a common identifier. Unless a user wishes to make a deterministic match on a geographic variable, such as county FIPS code, or a time-based metric, such as year or quarter, deterministic matching may not be possible. Cases may exist where, although the identifiers are not exactly the same, datasets share common observational units, such as abbreviated addresses and names, or approximate geographical locations. However, the researcher would need to standardize variables across datasets and work with the variations and inconsistencies in address and other common linking variables. With some data cleaning, deterministic matching is possible.

Alternatively, probabilistic linking is possible when observational units in the data share a common identifying feature or features but not necessarily a unique identifier. In this case, conditional on observable features, a researcher estimates the probability that two or more observational units relate to each other. The result that is generated is a match score. Probabilistic linking utilizes statistical algorithms to determine the likelihood of a match, and the researcher creates a rule that two records are a match based on a certain threshold, or value, for match scores. In the FSRDC PIK process, Dr. Nichole Szembrot explained that PIKS are probabilistic matches, not one-to-one matches. Researchers can leverage resources available within software tools like Stata, R, and SAS to implement either of these data linkage techniques.

The process of linking datasets through either deterministic or probabilistic techniques offers the advantage of significantly increasing the number of research questions that a researcher can answer. However, the process of matching introduces questions about data quality. As noted about administrative data, these data were not collected with the intention of academic research. Similarly, data collected by two different statistical agencies were not designed with the intention of being linked. It is possible, therefore, that not all observations in one dataset can be matched with all

Figure 3: Decision Tree for Determining Appropriate Matching Method


Do the two data sources contain a common identifier (ID)?
Yes
No
Are the ID's in the same format?
Candidate for probabilistic matching
Yes
No
Match one-to-one or many-to-one or one-to-many
Does the ID need to be cleaned or standardized?
No
Yes
Convert one variable to another format
Pre-process the data

observations in a second dataset. Consider this simple example: If a researcher is performing a deterministic match, linking a dataset that includes all counties within the contiguous United States (dataset 1) with a dataset that includes only counties within the state of California (dataset 2) where the common identifier is county FIPS code, all states other than California would be dropped from the sample. At this point, a critical trade-off in performing this data linkage must be assessed, which is the trade-off between having missing data (i.e., only using the state of California to evaluate the research question at hand) or not including the state-level information altogether (i.e., using only the national data from the first dataset). When data linking yields a dataset with a reduced scope, such as in this example, one must assess whether the missingness produce an analytical sample adequate to address the research question.

On the other hand, and often a consideration for probabilistic matching, false links can occur, where two records are erroneously matched due to similarities in certain variables or a matching rate threshold that is not high enough. Understanding this trade-off is essential to interpreting results accurately. Mismatches raise concerns about the power of inference and the reliability of conclusions drawn from the linked data. Careful consideration of limitations and potential biases is crucial when analyzing linked datasets.

In addition to missing and false links, researchers must be aware of potential data quality issues that may emerge, such as potential differences in sampling methods, data collection at different time points, and measurement differences for key outcome variables. Each data product has its own unique traits, and these include sampling and nonsampling errors. Researchers must be cognizant of the biases that might arise when data linkages merge the sampling and nonsampling errors from two datasets into a single data product.

Conclusion

In an age when data are more available than ever before, the chance to produce policy-relevant research has never been better. The discussion highlights the need for tools that can help reduce inequities associated with data access and search costs that researchers face when trying to find relevant data and up-to-date methods for research. One platform DemocratizingData.ai (https://democratizingdata.ai/) is the product of a research initiative that works with several government agencies, including the USDA, to provide more equitable data access to users. Their platform identifies and locates datasets within publications and offers agency and researcher access to data information through versatile methods like APIs, Jupyter Notebooks, and researcher dashboards. Another initiative, Data Integration in Food and Agriculture (https://www.dataifa.org/; under development), aims to simplify data access and the data linkage process by providing a user-friendly search repository and key metadata that harmonizes and streamlines the data documentation process. These are just two examples that have the potential to work in tandem to enhance data accessibility, contributing to an ecosystem where researchers can locate, access, and work with data, fostering a more efficient and collaborative research environment in the field of agricultural and applied economics.

For More Information

Abraham, K., R. Haskins, S. Glied, R. Groves, R. Hahn, H. Hoynes, and K. Wallin. 2018. *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking*. Washington, DC: Commission on Evidence-Based Policymaking.

Bailey, M.J., C. Cole, M. Henderson, and C. Massey. 2020. “How Well Do Automated Linking Methods Perform? Lessons from Us Historical Data.” *Journal of Economic Literature* 58(4):997–1044.

Chetty, R. 2012. “Time Trends in the Use of Administrative Data for Empirical Research.” Available at: https://rajchetty.com/publication/time-trends-in-the-use-of-administrative-data-for-empirical-research/admin_data_trends-pdf/.

Emery, J., and D. Boyle. 2017. “Data Linkage.” *Australian Family Physician* 46(8):615–619.

Holman, C.D.J., J.A. Bass, D.L. Rosman, M.B. Smith, J.B. Semmens, E.J. Glasson, E.L. Brook, B. Trutwein, I.L. Rouse, C.R. Watson, et al. 2008. “A Decade of Data Linkage in Western Australia: Strategic Design, Applications and Benefits of the WA Data Linkage System.” *Australian Health Review* 32(4):766–777.

National Academies of Sciences, Engineering, and Medicine. 2020. *A Consumer Food Data System for 2030 and Beyond*. Washington, DC: National Academies Press.

Wilson, N., L. Chenarides, J. Kolodinsky, and K. Boys. 2021. “To Ensure That All People Have Safe, Affordable, Accessible, and Acceptable Food for Leading a Healthy and Active Life.” *Agricultural & Applied Economics Association Grand Challenge*.

About the Authors: Corresponding Author: Lauren Chenarides (Lauren.Chenarides@colostate.edu) is Assistant Professor with the Department of Agricultural and Resource Economics at Colorado State University.
Andrew S. Hanks (hanks.46@osu.edu) is Associate Professor with the Department of Human Sciences at The Ohio State University. Amelia Finaret (afinaret@allegheny.edu) is Associate Professor with the Department of Global Health Studies at Allegheny College and Honorary Lecturer with Global Academy of Agriculture and Food Systems at the University of Edinburgh. George Davis (georgedavis@vt.edu) is Professor with the Department of Agricultural and Applied Economics at Virginia Tech. Andrea Carlson (andrea.carlson@usda.gov) is a Senior Economist with the Economic Research Service at the U.S. Department of Agriculture.

Acknowledgments: Funding from the Agricultural and Food Research Initiative (National Institute for Food and Agriculture, USDA) grant no. 2021-67023-34850 and USDA Economic Research Service Cooperative Agreement no. 58-4000-1-0103 are gratefully acknowledged. The findings and conclusions in this publication are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy. We direct those interested in presentations and recordings from the workshop to this link: https://dataifa.github.io/difa-project/index.html.
All those mentioned in the article have granted us permission to disclose their full names. The authors thank the following individuals:

- Bruce Weinberg, PhD, is the Eric Byron Fix-Monda Endowed Professor of Economics in the Department of Economics at The Ohio State University, a research associate at the National Bureau of Economic Research (NBER), and a research fellow at the Bonn Institute for the Study of Labor
- Timothy Beatty, PhD, is a Professor of Agricultural Economics in the Department of Agricultural and Resource Economics at the University of California-Davis. (IZA). Dr. Weinberg co-founded the Institute for Research on Innovation and Science.
- Mark Prell, PhD, is a Senior Economist in the Food Assistance Branch of the USDA Economic Research Service's Food Economics Division.
- Joseph Cummins, PhD, is an Assistant Professor of Economics in the Department of Economics at the University of California-Riverside.
- Nichole Szembrot, PhD, is the Cornell FSRDC administrator.
- Julie Lane, PhD, is a Professor at the NYU Wagner Graduate School of Public Service and an NYU Provostial Fellow for Innovation Analytics. Dr. Lane cofounded the Coleridge Initiative, an entity designed to transform the way governments access and use data for policy making purposes. She also cofounded the Institute for Research on Innovation and Science at the University of Michigan and currently serves on the Advisory Committee on Data for Evidence Building and the National AI Research Resources Task Force.
- Jason Owen-Smith, PhD, is a Professor of Sociology in the Department of Sociology at the University of Michigan. He is currently the Executive Director in the Office of Research Analysis and Data Integration and the Executive Director of the Institute for Research on Innovation and Science.
- Brian McGreal, for his support in coordinating and contributing to the success of our two-day workshop.

The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. This work was supported in part by the U.S. Department of Agriculture, Economic Research Service.