INSTITUTE *for* RESEARCH *on* POVERTY
UNIVERSITY OF WISCONSIN–MADISON

Robert M. La Follette School of Public Affairs
UNIVERSITY OF WISCONSIN–MADISON

From Consent to Linkage: New Data Infrastructure Payoffs

Timothy M. (Tim) Smeeding, University of Wisconsin—Madison

for the CNSTAT/NIA Workshop: *Improving Consent and Response in Longitudinal Studies of Aging* September 27-28, 2021

Session– "Looking Ahead: Applying Innovative Strategies to Improve Consent and Response"

Context for this presentation

- ***Improving Consent to Link Data and Improve Response*** **is the general topic here- this talk is about new/exciting linkage opportunities that consent can bring**
- **The biggest linkage key is Title 13 of the United States Code.**

"Title 13 provides the following protections to individuals and businesses: … It is against the law to disclose or publish any private information that identifies an individual or business such, including names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers"

- **There are now innovative new linkages to Census Data alone of two kinds, both of which should be of interest :**

1. **Linkages *outside* of Title 13– after 72 years, *going back* now easy**
2. **Linkages *within* Title 13 – *going forward* is on the way via RDCs**

Title 13 - The "72 year rule"

- In 1952 an agreement with National Archives (later codified into law) restricts public release of Census records for "statistical purposes only" for 72 years.

- So the mostly all linked Census data is coming in two types of datasets:

1. Beyond title 13– 72 years or longer, the 1850-1940 Censuses can and have been linked and are available -- genealogy anyone ?

2. Under title 13 -less than 72 years old , the 1950*- 2020 Censuses with some links already (2000 on, including the ACS) , others in progress (1960-1990), for use in FS RDCS or 'data clouds' only

* The 1950 Census is 72 next year in 2022 -- and can be publicly accessed soon

Beyond the 72-year window

- **New and robust linked data is available from the Census Linking Project** offering researchers the ability to create longitudinal datasets using historical US Census data (1850-1940) via links between each pair of complete-count Censuses using a wide variety of linking algorithms

- **Economists have made it easy to access and with very low false positives -- See seminal Automated Linking of Historical Data, 2021 JEL**

"The recent digitization of complete count census data is an extraordinary opportunity for social scientists to create large longitudinal datasets by linking individuals from one census to another or from other sources to the census"

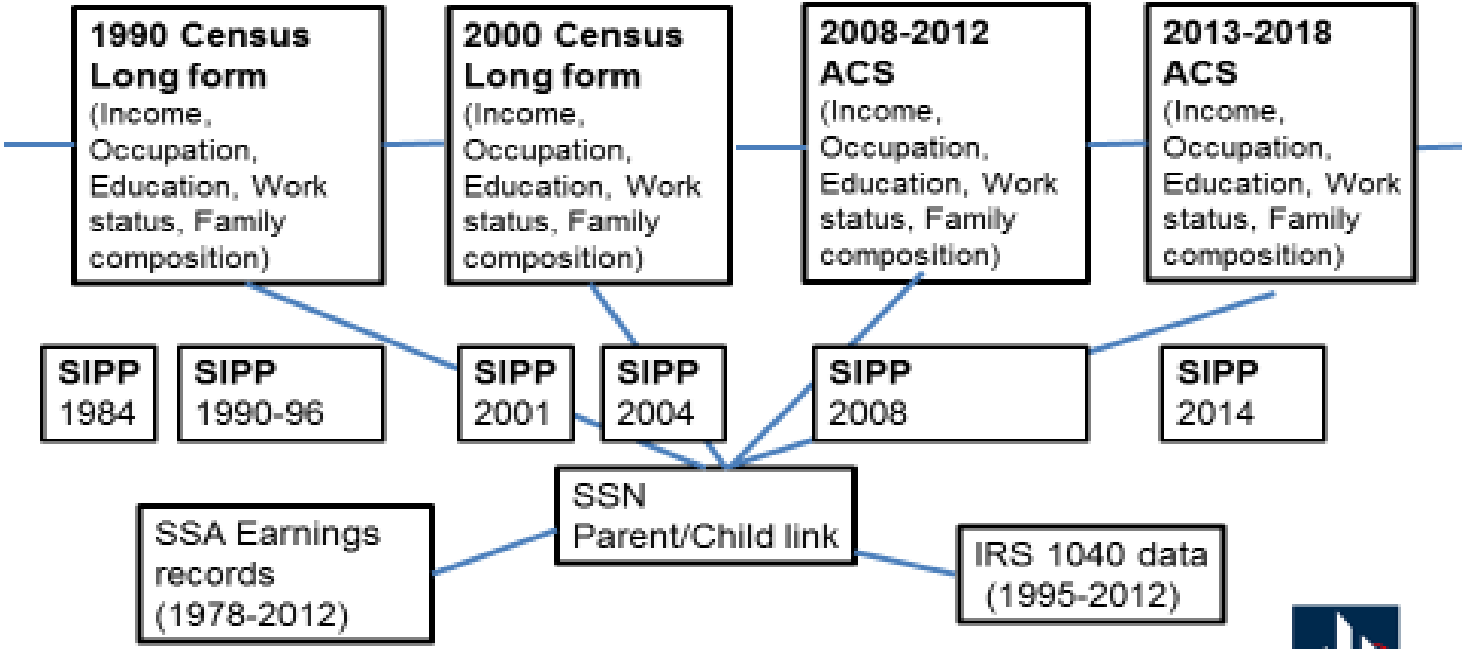
- **Papers are popping up like wildfire , eg intergenerational mobility by race and so on– lots more cited in the seminal 2021 JEL article**

Within the 72-year window-

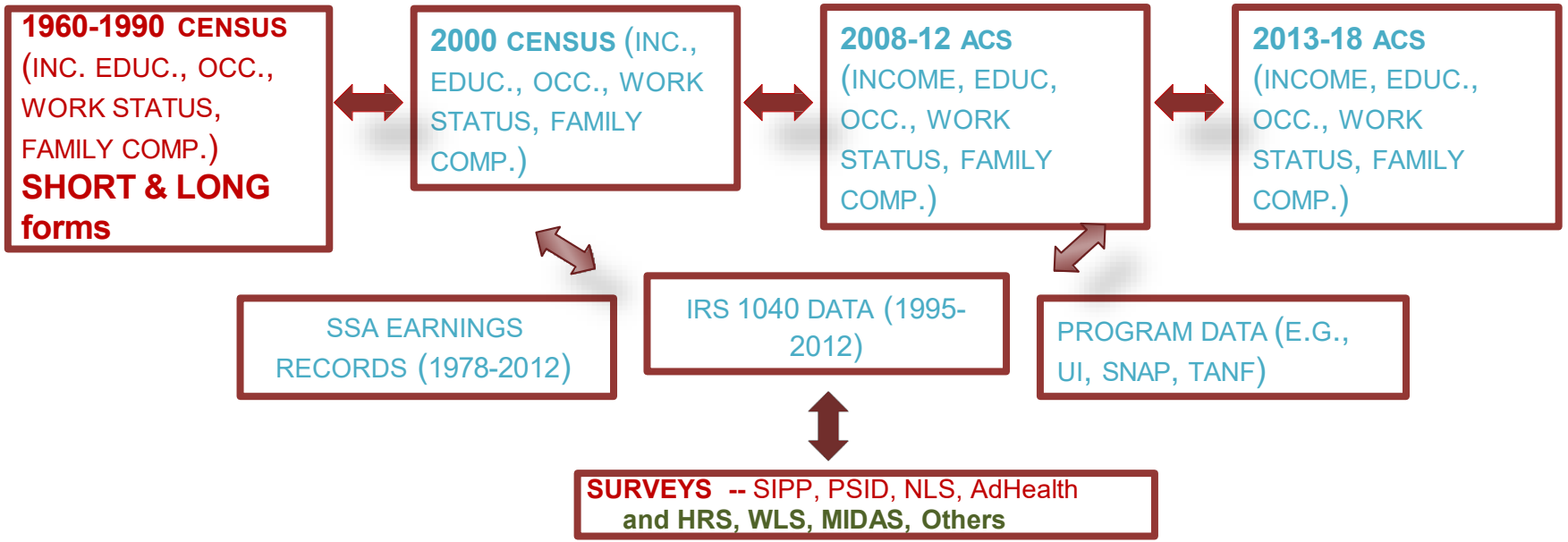
- Thanks to the AOS, the American Opportunity Study, a researcher led vision of a database linking existing data in social surveys to the decennial Census and other government administrative records , began in 2009, report in 2011
- Thanks to several donors organized by Raj Chetty , a full linkage of Census records is underway in 2021 !!!
- The 'Decennial Census Digitization and Linkage Project', DCDL , a collaborative Census project linking 1960-1990 Censuses
- The rest of an older talk about the possibilities that the AOS and DCDL are creating for all to link families across generations in the 10-year annual dataset that we all have(by law)complied with, the Decennial Census , is attached at the end for interested parties
- A very short summary follows

Then, the original SIPP 'Gold Standard' slide that moved us all to the AOS

The Possibilities with Census and ACS data And SIPP



AOS summary slide
"Add the Survey to Other Linkages"

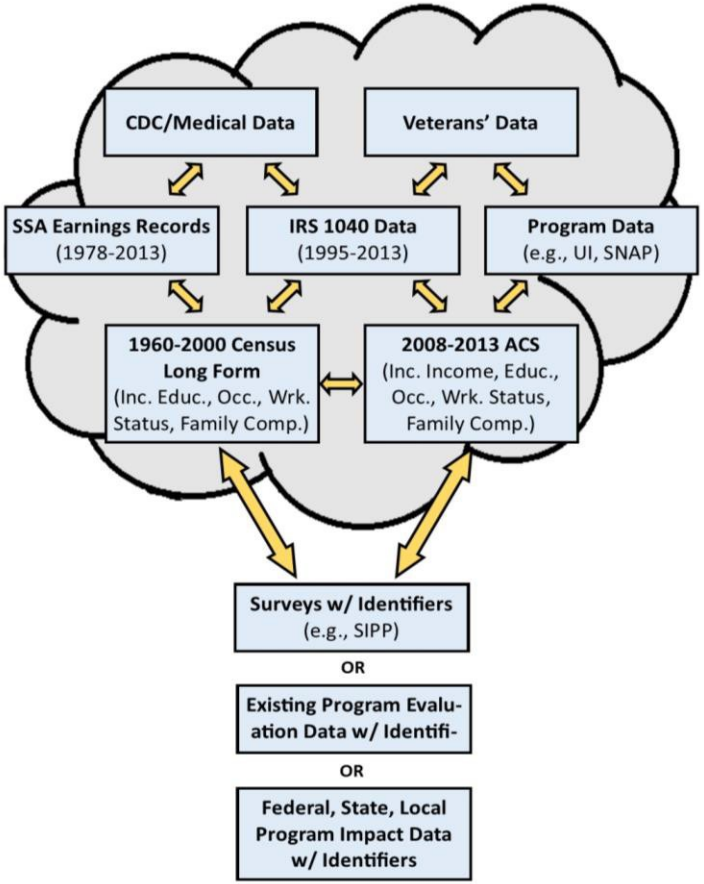


SURVEYS WITH IDENTIFIERS CAN BE added (E.G.,HRS, WLS, SIPP, ECLS, OR PSID, NLS, ADHEALTH, ETC.) -- THE SURVEY NOW BECOMES A " dynamic VALUE-ADDED INSTRUMENT" via linkages to other data

Summary Figure —Schematic design of the AOS --expanded to other uses in near future

Q: After the Committee on Evidence Based Policy (CEBP report) what would The AOS look like ?

A: See right– find all the linkages inside the National Secure Data Service (NSDS) cloud



Two discussion questions

1. How can we interest survey responders to give broad consent to data linkages ?

– eg would it be possible offer respondents a 'personal genealogy' report, *going backward* based on the Census Link Project Data

2. What incentives can help respondents allow you to *go forward* linking their records to those of their offspring ?

Thanks for your time

Feedback welcome

smeeding@wisc.edu

Addenda

2019 FSRDC annual workshop talk follows

“Beyond the American Opportunity Study (AOS),and on to the Decennial Digitization and Linkage (DCDL) Project”

INSTITUTE *for* RESEARCH *on* POVERTY
UNIVERSITY OF WISCONSIN–MADISON

Robert M. La Follette School of Public Affairs
UNIVERSITY OF WISCONSIN–MADISON

Beyond the American Opportunity Study (AOS),and on to the Decennial Digitization and Linkage (DCDL) Project

For the 2019 FSRDC Conference

UW Madison September 6th 2019

Timothy M. (Tim) Smeeding, University of Wisconsin—Madison

with the help of

Katie Rose Genadek(CENSUS/ERD FED)

Jonathan D Fisher, Stanford RDC

David B. Grusky, Stanford University

Michael Hout, New York University

C. Matthew Snipp, Stanford University

Research | *Training* | *Policy* | *Practice*

Preamble

This presentation covers a lot of ground and flies at 30,000 feet.

But the big picture is important as we move from a dream—the AOS, to a firm plan for the key element in the AOS, a full linkage of Census records, under the DCDL, a collaborative Census Bureau proposed project.

The beliefs and hopes and plans summarized here are attributable only in the least to Smeeding. The others at Census, especially Katie Genadek , but also at Stanford (David Grusky, Jonathan Fisher, Matt Snipp), and at Michigan (Trent Alexander) and a host of others are working to make it a reality

Overview : date linkages at the Census and elsewhere

1. Things are moving very fast in some ways but slow in others

-- **Use of administrative data in social science research is exploding in the name of evidence based policy (EBPs)**

- *But linkages and datasets have been patchwork, one-off linkages, with much of the access , at least at the federal level, depending on networks and connections*

-- **The Committee on Evidence Based Policy (CEBP) has made its recommendations and already one or two pursuant bills (eg the *Foundations for Evidence Based Policy Act*) have passed congress and federal agencies are now organizing for data sharing across and within agencies, with more below on other actors like the NRC's Committee on National Statistics and the Bipartisan Policy Center (BPC)**

-- *But data quality and methodological standards are almost non existent*

Overview , continued

2. The research and policy communities will need a centralized capacity to effect linkages and deliver the linked data to safe and secure research sites in an unbiased & transparent way

---The first step --building an *on-demand administrative database*, e.g., as modelled after the National Secure Data Service (NSDS)proposed by the CEBP

--The second step a framework and adequate capacity for ensuring ongoing, full and secure safe access with RDCs clearly in the mx

---------- but in the meantime ------

—The Census Bureau is now in beginning a project to realize the key linkages which will provide a data backbone for approved, RDC based and Census based research on population change with linked records across the 1940-2010 and then 2020 Censuses, the DCDL – see paper , "The DCDL Linkage Project" , by Genadek and Alexander, at https://deepblue.lib.umich.edu/handle/2027.42/150659 and more below

Outline for presentation

1. Summarize some of the important linkages already extant, key findings for evidence based policy analysis and program evaluation , as well as the need for filling in the blanks to identify mechanisms
2. History --how the AOS came about, became a standing committee and pushed issues and technical "proof of concept " study : from name recognition to record linkage for Census merged/linked files
3. The DCDL and its promise
4. The support networks growing and expanding and pushing the science
5. Key issues in need of deeper thought and attention: the statistical science of adjusting for missing matches and related issues, and making it easier to access all types of administrative data

* https://www.nap.edu/catalog/23583/using-linked-census-survey-and-administrative-data-to-assess-longer-term-effects-of- policy

1. Important linkage projects and policy lessons learned already

--National data linkages follow the Scandinavians and register based studies (external validity issues -eg 'NO' kids papers)

--States began and continue: WI on child support and MSPF files (--

--Localities too : Bob Goerge in Illinois/Chicago; Dennis Culhane, Penn; and Amy O'Hara Georgetown

--Education networks now with lots of *within* state linkages: FL, NC, WA ,MI, others—(feds still can help enormously with FSRDC sanctioned linkages to federal data, eg incomes beyond FRPL)

--and much more , on brain science and youngest kids, on exposure in-eutero, and many papers at this conference !

Why do we need more: policy effcts

-- Barker Hypothesis --what's vs. why's and how's ?

- Dozens, hundreds ? of papers now published with exposure event X happening in childhood (Medicaid, SNAP, EITC, child care, etc.) , and outcome Y (schooling, income, earnings , etc.) occurring 15-40 years later
- Falsification test show *something* happened and it made a difference
- Bottom line, "causal" evidence shows that a little money, dose of pre-school, coverage by Medicaid or SNAP or FRPL matters a lot (eg NAS report "Roadmap for Reducing Child Poverty " 2019 report, chapter 3 https://sites.nationalacademies.org/DBASSE/BCYF/Reducing_Child_Poverty/index.htm)

From what's to why's and how's

--But the question remains, why and how ?

-*what* happened to a given observation between the treatment/exposure and the outcome?

-*what* mechanisms were important in producing better outcomes because of the intervention ?

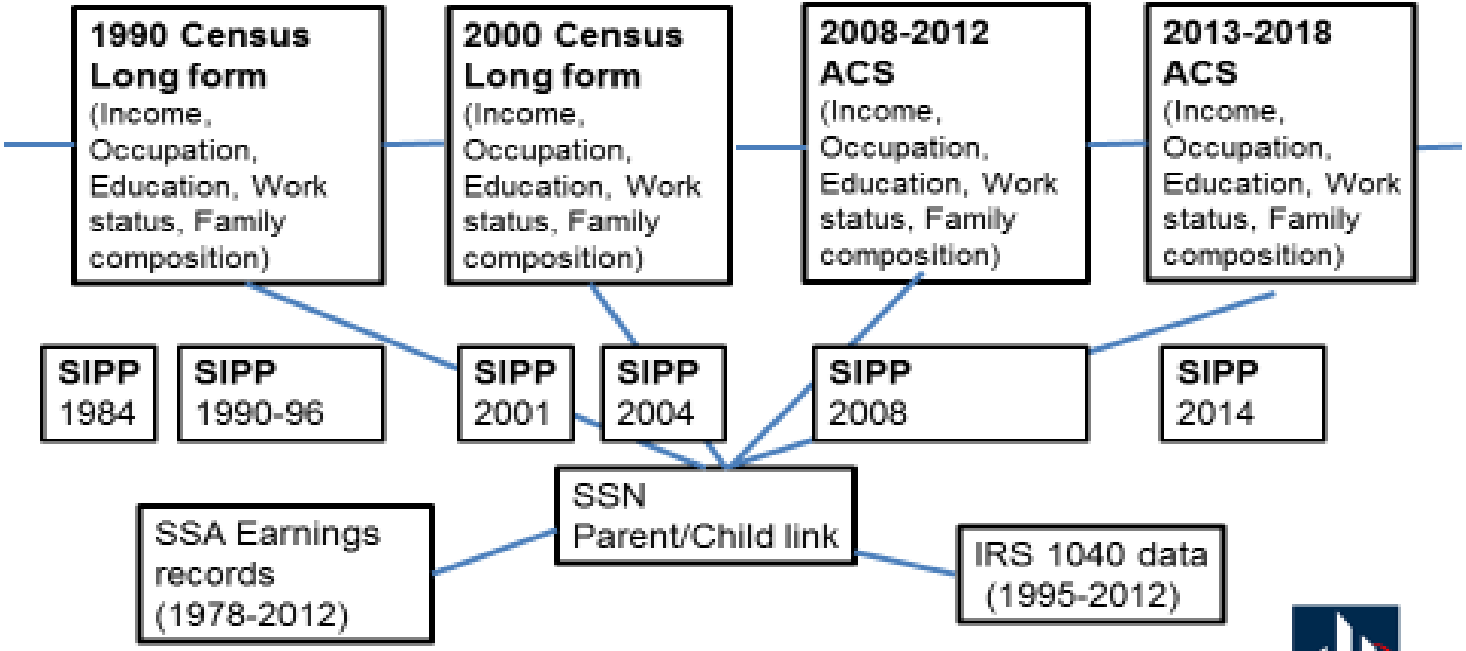
- **Way to learn what happened is to follow the child across time and space and family situations using other data to link to the primary study**
- **Enrich other longer term mobility data– add people (families, kids, institutions) to Reardon education data and Chetty/MTO spatial data**
- **Add linkages for surveys like WLS , HRS , PSID and AdHealth to other administrative data on families with name, age, quarter of birth, sex, parental marital status, state or country of birth and so on**

2. History --how the AOS began, became a standing committee and pushed issues

- **The beginning in 2011 : asking what data will allow us to better monitor social mobility ?**
- **Longitudinal surveys help us focus on one cohort , e.g., start with NLSY, PSID *parents as origin* their offspring as they grow up—so always *looking back* and wait for the kids to grow**
- **Lightbulb #1 goes off ---- what can some linked administrative records (inspired by the SIPP Gold Standard model) tell us about income and earnings mobility for recent cohorts, tracing current kids or young adults' experiences as origin back to their parents?**

Then, the original David Johnson slide that moved us all toward AOS

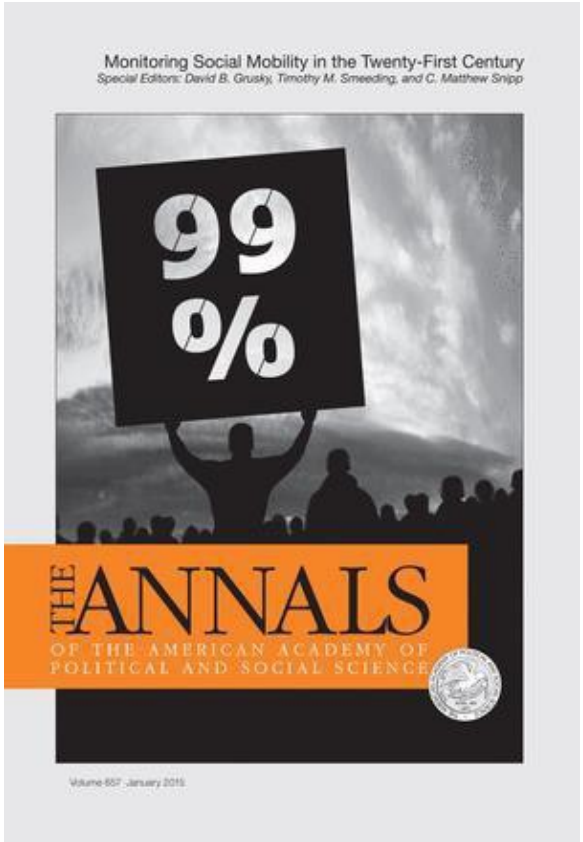
The Possibilities with Census and ACS data And SIPP



Then we started to envision a new tool

- A de-facto inter-generational and inter-temporal panel already exists : the decennial Censuses !
  -There is great potential to add depth by linking Censuses to one another and then possibly to administrative data
- Adding surveys or other studies which can be slipped into the structure to answer questions not answered in the structure we envision also appeals – we still need (shorter) surveys !
- Let's use this structure to study mobility—in a broad sense , call it the AOS, and put out a volume on monitoring social mobility and what was needed

We did an ANNALS to kick it off



Lead to The AOS Standing Committee at the NRC (2012-2017) , and Its 3 Major Subcommittees

- **Workshop planning – Proof of Use/Demand--** focus was to build the user community "moving beyond mobility narrowly defined" to other uses of the AOS for evidence based policy
- **Held workshop--**"Using Linked Census, Survey, and Administrative Data to Assess Longer-Term Effects of Policy: Proceedings of a Workshop—in Brief." Washington, D.C. : National Academies Press, 2016
- **Bottom Line--and HOW !! Huge demand for AOS**

The AOS Standing Committee at the NRC, the other 2 Major Subcommittees

- **Matching and record linkage methodology – Proof of Concept** --focus is to understand the record linkage for AOS merged/linked files and their strengths and weaknesses, *especially data linkages across decennial Censuses*
- **Governance – Proof of Operability** -- focus is to look long term at the AOS, how it should be structured to allow access , protect privacy and confidentiality, provide for maintenance.

The AOS was launched and existed for some 3-4 years at NRC

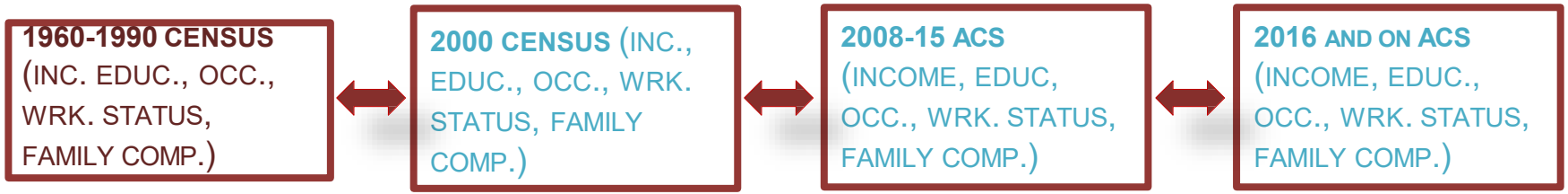
Goals:

- **Fill in the decennial holes between 1960 and 1990 (others are already filling pre-1960—1940 and 1950 via NIH projects—Steve Ruggles, Rob Warren , et. al., etc.)**
- **Provide expertise to improve matching and linkage across files**
- **Assist in identifying key longitudinal research opportunities**
- **Help with governance issues regarding data access, privacy and confidentiality (before the CEBP was formed)**
- **A joint project amongst a large group of stakeholders: AOS NAS Committee, Stanford University, U.S. Census Bureau, others**

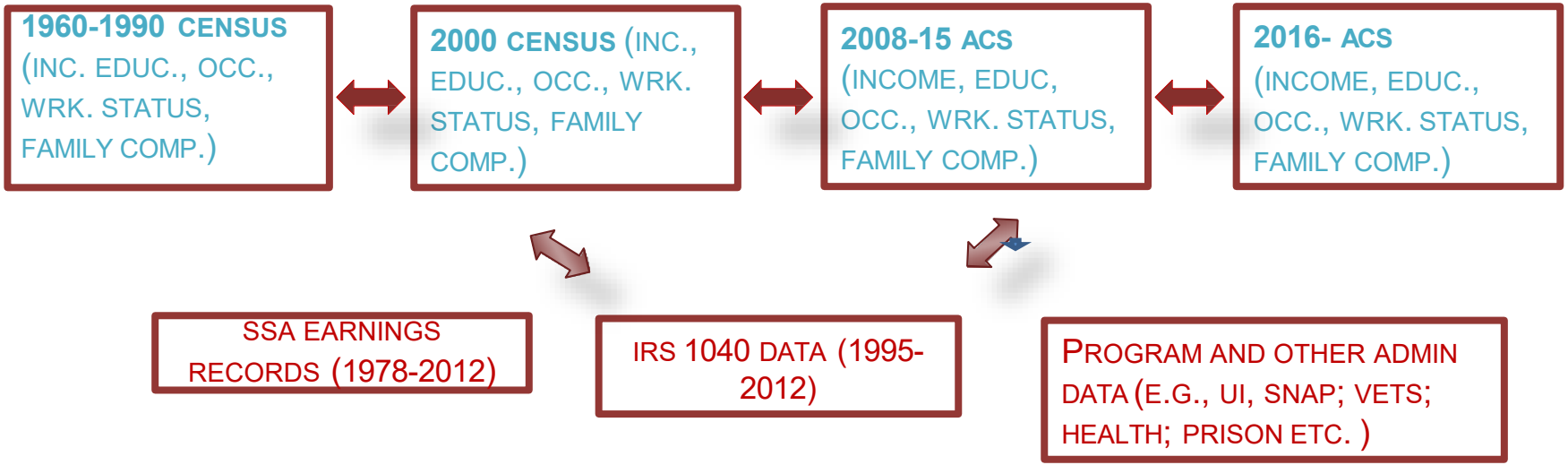
The Technical Hurdles: Proof of Concept

- The digitized 1960-1990 decennial data do not include people's names (other data from the decennials have been captured and digitized, but to reliably link we needed the person's name too) .
- 1990 forms are stored on 120,000 reels of microfilm, and the name is handwritten by the respondent. And the actual names on the 1990 decennial remain restricted-access until 2062 (Title 13)
- Goal#1: accurately assigned a Protected Identification Key (PIK), which allows for linkage to other data.
- Goal #2 : raise serious money to digitize the entire 1990 Decennial *and* then work backwards in time to 1960, adding links to other datasets, once we have shown it can work .

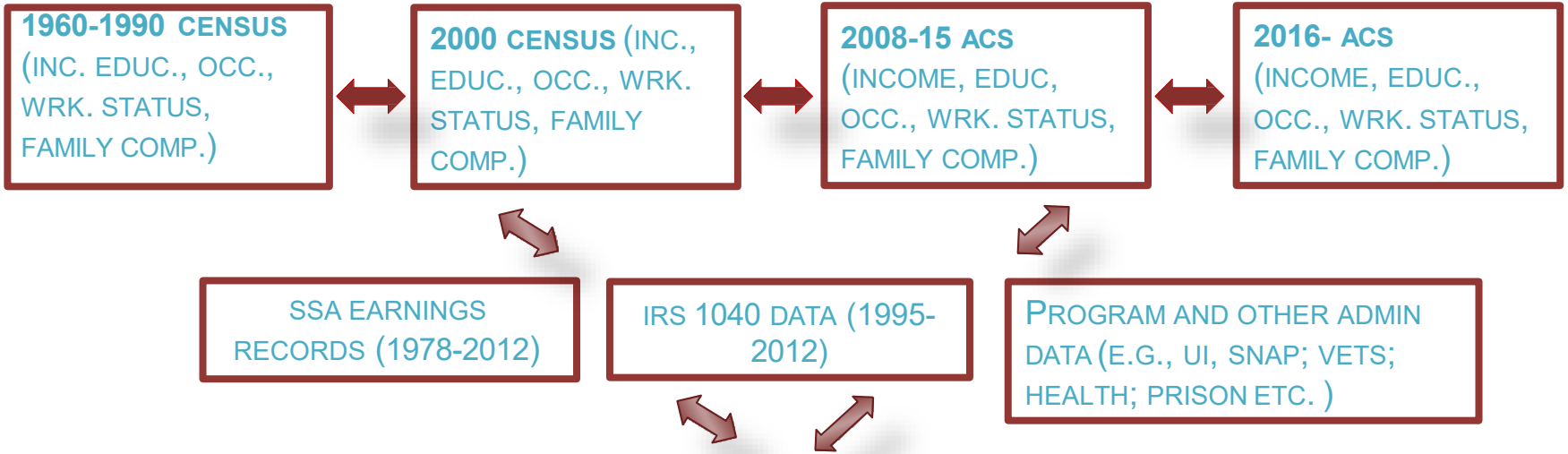
Step #1: linking records from ACS and Census across many years (short and long forms !)



Step #2: adding in links to administrative data



Step #3: linking children with parents

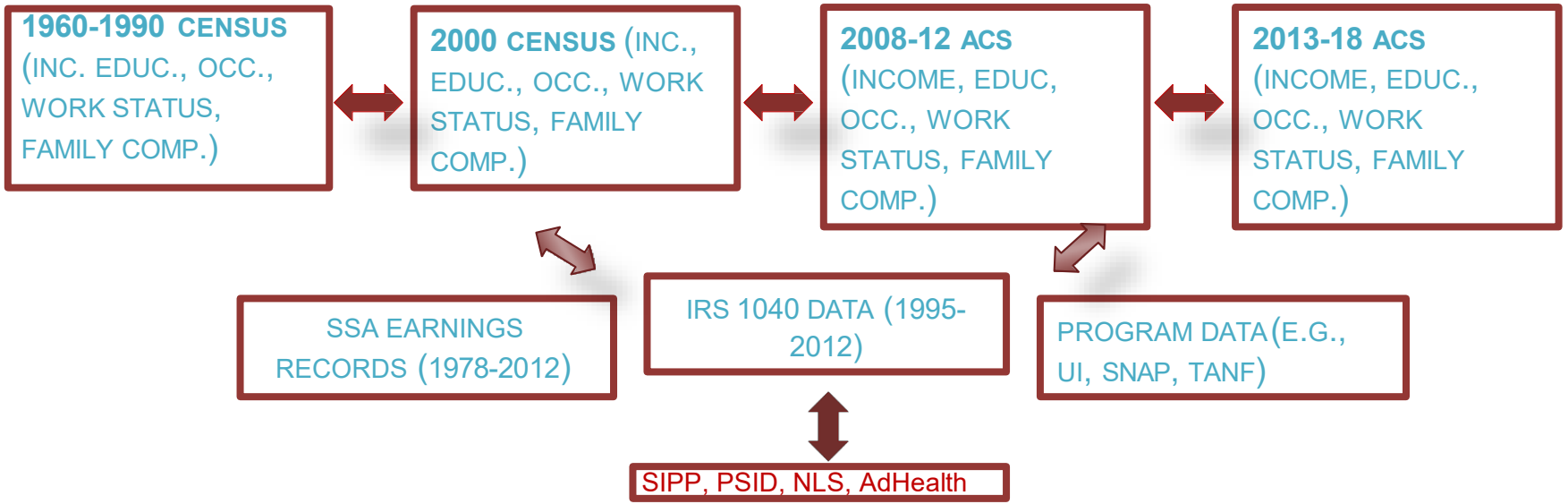


KIDLINK FILES (FORM SS-5*)

GET CO-RESIDENCY IN ACS, CENSUS LONG/SHORT FORM

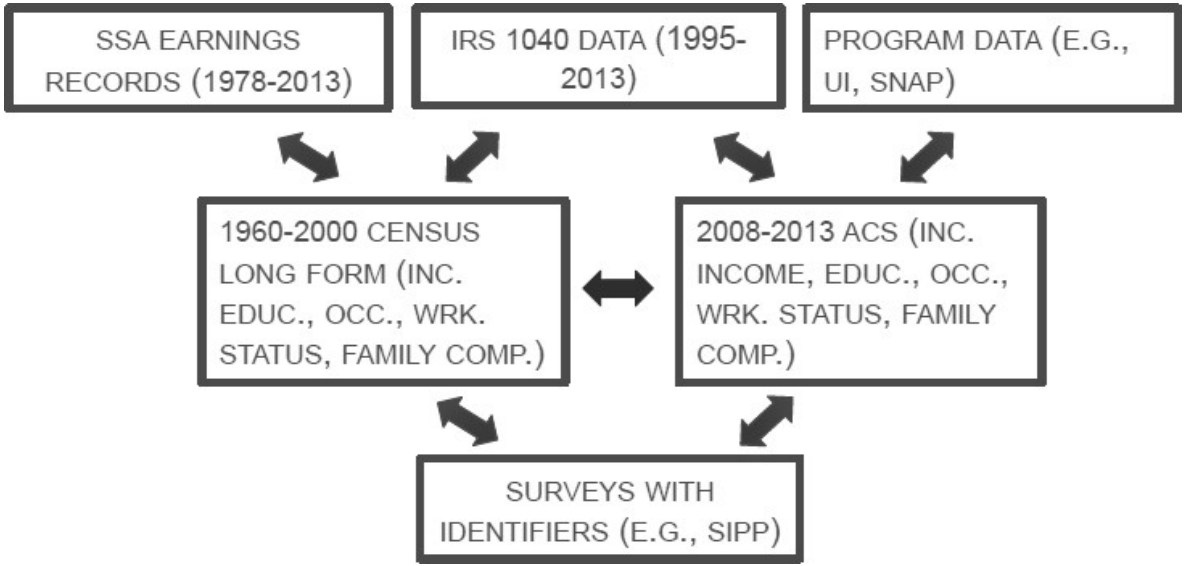
***PARENTAL REPORTS OF CHILDREN'S SSN TO IRS SINCE Mid 1980'S**

Step #4: Slipping in the survey (–or other study)



SURVEYS WITH IDENTIFIERS CAN BE "SLIPPED IN" (E.G., SIPP,ECLS, OR PSID, NLS, ADHEALTH, ETC.) THE SURVEY NOW BECOMES A " dynamic VALUE-ADDED INSTRUMENT"

Summary Figure 1.
The Schematic Design of the AOS



Another lightbulb: from 'proof of demand' work: program impacts and evaluation

--Beyond social mobility narrowly defined- lots of uses from workshop-"What could you do with AOS"?

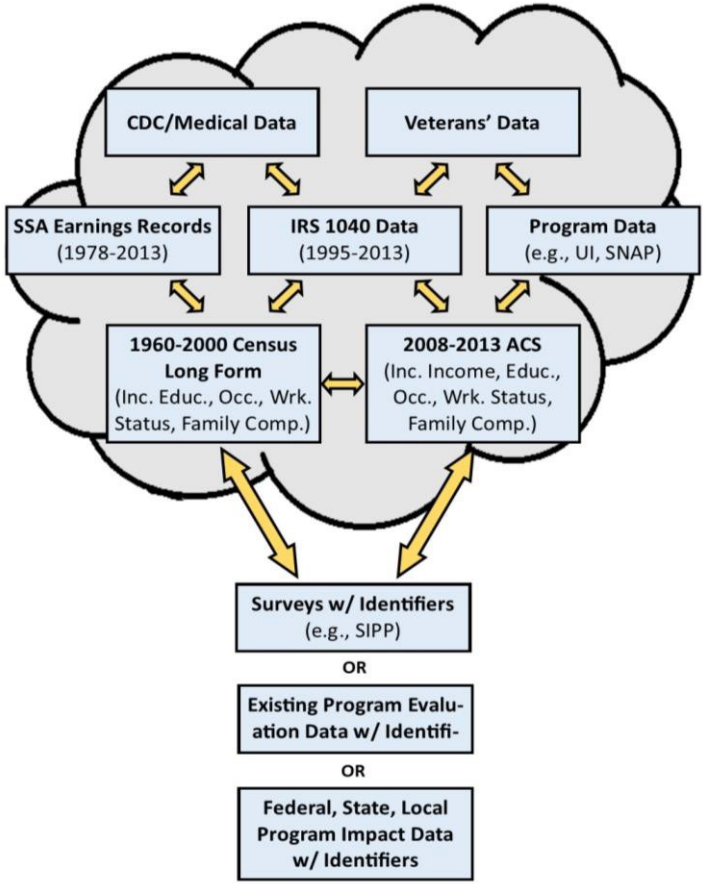
-- LOTS! --we cannot adequately evaluate the long-term effects of key social programs on mobility and other outcomes without use of administrative data (eg MTO, SIME-DIME, SNAP, Medicaid, more)

--We can fill in the evidence deficit on long run effects of policy on mobility but also on a broader range of human outcomes using this same scaffolding

Figure 2—Schematic design of the American Opportunity Study expanded to other uses

After the CEBP report what would The AOS look like ?

-- inside the National Secure Data Service (NSDS) cloud



AOS terminus paper and beyond

- Final AOS paper "The American Opportunity Study: A New Infrastructure for Monitoring Outcomes, Evaluating Policy, and Advancing Basic Science" The Russell Sage Foundation Journal of the Social Sciences March 2019, 5 (2) 20-39; DOI: https://doi.org/10.7758/RSF.2019.5.2.02

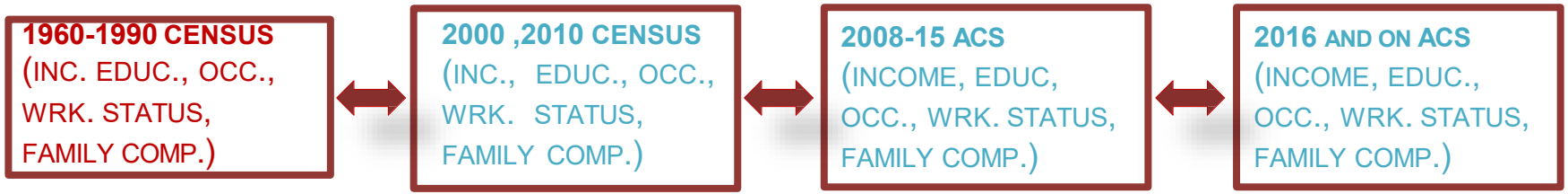
- Others labored on –Grusky (projects using census matched occupation data) ; Snipp (member CNSTAT) ; Hout (from head AOS to head DBASSE) and Fisher (name recognition CARRA projects) combined now with Genadek and Alexander (Census) and others ,working with and in census and vendors on signature recognition software for digitizing census records and linkages

3. The DCDL and its promise

- Concurrent with the National Academies AOS , its teams collaborated with a Census Bureau team led by Trent Alexander to define techniques and costs for recovering names from the 1990 Census.
- With funding from the Carnegie Foundation through the National Academies AOS committee, the 1990 Name Recovery Pilot (NRP) project established Proof of Concept
- Further efforts have been supported by robust partnerships between universities, foundations, federal agencies, and the FSRDC network
- Now we have the DCDL project, still needing funding , but poised to complete a data series including almost the entire population between 1940 and the present, the first step in the AOS vision (next slide)

Back to AOS step #1: linking records from ACS and Census across 80 years (short & long forms)

1940 and 1950 Censuses linked by Census and Ruggles, inc. by 2020
2000-onward Censuses already linked ; 2020 by 2023
Key is middle years 1960,70,80,90 –and this is where DCDL comes in



DCDL plans and objectives

1. ***Digitize and recover respondent names from microfilmed decennial census manuscripts of 1960, 1970, 1980, and 1990***, using Optical Character Recognition (OCR) processes to create machine-readable respondent names from these images for 500 million records .
2. ***Attach recovered names to existing microdata files***. Census Bureau has microdata files from the 1960 through 1990 censuses include all variables other than respondent names, this step appends names to the correct record in the existing microdata
3. ***Link the 1960-1990 censuses together and to the 1940, 1950, 2000, 2010, and 2020 Censuses***, *using* well-established methods that have been used to build an infrastructure that already contains censuses, surveys, and administrative data before and after this period.

DCDL Outcomes

- Individual names are replaced with unique identifiers that permit researchers to trace individuals over time.
- The resulting restricted-use linked files will form the core of a statistical infrastructure documenting most of the U.S. population since 1940.
- DCDL-based data will provide a multi-purpose data infrastructure that will be maintained by the Census Bureau to improve the measurement of the U.S. population, and the data will be made accessible to researchers through the FSRDC network
- Research conducted with the data resulting from DCDL is intended to transform our understanding of intergenerational, social, and geographic mobility, as well as life-course transitions and trajectories
- The resulting data series will serve both academic social sciences and the evaluation of government programs and policies.

4. The support networks are growing and expanding and pushing the science

- **NAS –CNSTAT**

1. New projects (workshops, seminars, studies) abound on the science of data linkages, the new LINK lecture and much more, eg NAS "Data Linkage Day" was October 18th 2019

2. Already Public Seminar on Linkages Among Federal , State and Local Data on May 19th, 2019

- **Bipartisan Policy Center**

***Evidence Works*, a compendium of 20 cases of Federal, State and local evidence use for policy evaluation**

More Networks & Funding

- **WT Grant Foundation**

1. National Network of Education Research-Practice Partnerships
2. Improving the Use of Research Evidence program
3. Adam Gamoran, WTG president, in *Science*, " Evidence lights the way

- **RSF Computational Social Science program**
- **Raj Chetty via the Opportunity Insights Network**

Others who I am failing to mention—sorry

5. FINALLY--Key issues in need of deeper thought and attention

--Even with administrative data, no matches are perfect– how does one adjust for non-matching or over-matching or missing data ?

--Is its analogous to sampling and non-sampling errors in surveys? --eg item non-response ?

-- We can use long running panel surveys like PSID to look at the quality of our generational linkages (and in turn , we can help panel surveys understand their attrition issues)

Last, Proof of Operability

-How can structured access can protect privacy and confidentiality, lead to dataset maintenance and improvement?

- CEBP spoke and here is what we heard :

-- lets set up a NSDS at Commerce,

-- lets put the expanded AOS into a data cloud, only take what you need from NSDS, links then disappear

-- lets have a process of application, quality assurance, and safe, secure access via FSRDCs

Lets also make it easier to use and prepare data for linkages

--More uniform MOUs, DSAs and DUAs

-- Let's rethink Titles 13, 26 and other barriers,with the goal of privacy protection and transparency addressed

-- and with lots more to ease research use in FSRDCs

-We have already begun, HR 4174 , Foundations for Evidence-Based Policymaking Act of 2018 (draft plan)

--Bipartisan Policy Center has more in the works

-- Planning group for NSDS is starting

but with some humility as this will be major undertaking

Conclude – AOS ex-ante vision moves forward

- Policies to affect mobility rely on assessing longitudinal processes and long-run effects of programs, treatments and life experiences
- Move beyond one-off studies to systematically allow limited, orderly, safe, secure access to better data
- Build a new infrastructure and evidence base that makes accurate assessments possible, safe (in terms of privacy and confidentiality) , comprehensive, and cost effective