

Disclosure Avoidance and the American Community Survey

Rolando A. Rodríguez, U.S. Census Bureau

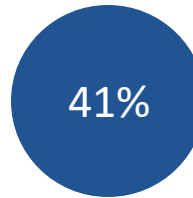
Presented to the 2021 ACS Data Users Conference

May 20, 2021

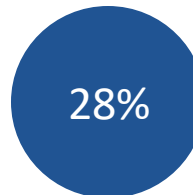
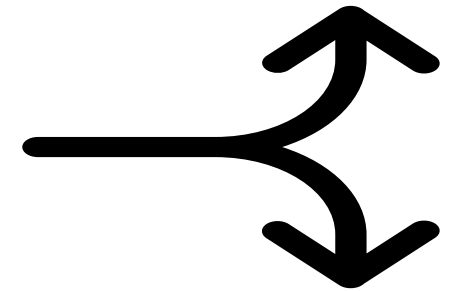
This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Respondents make hard choices

- Respondents weigh risks and rewards of survey participation
- Benefits of participation may not be obvious
- No public disclosures from Census products does not equate to no privacy concerns... or risks



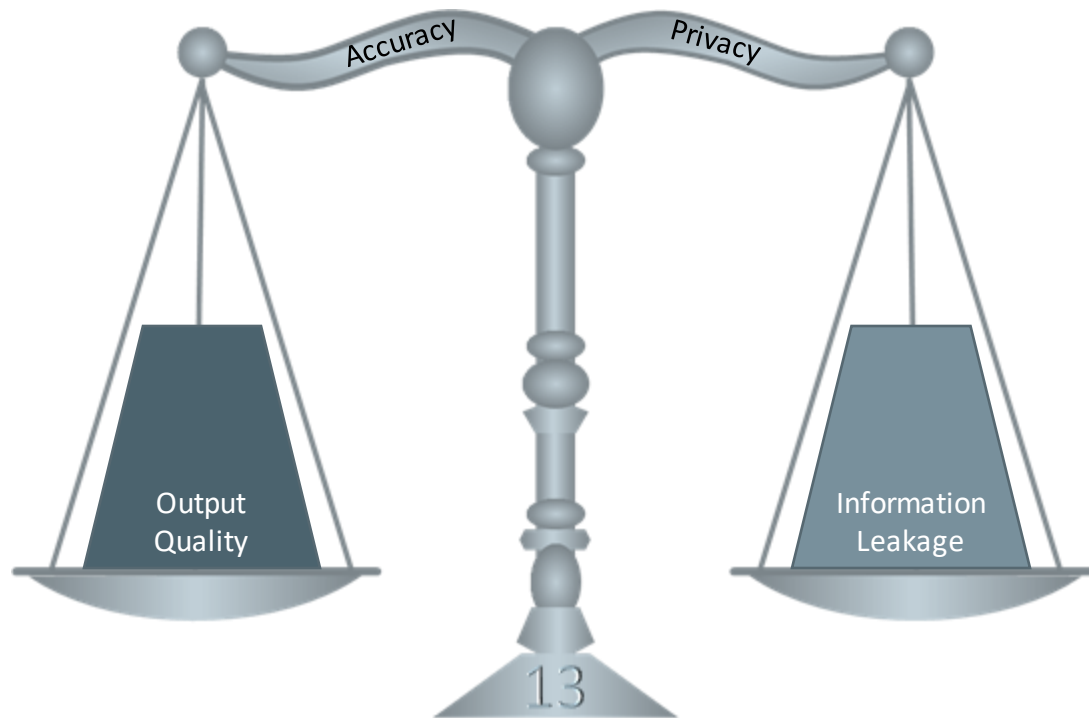
It matters a moderate amount or less if I'm personally counted[†]



I'm very or extremely concerned my answers won't be kept confidential[†]



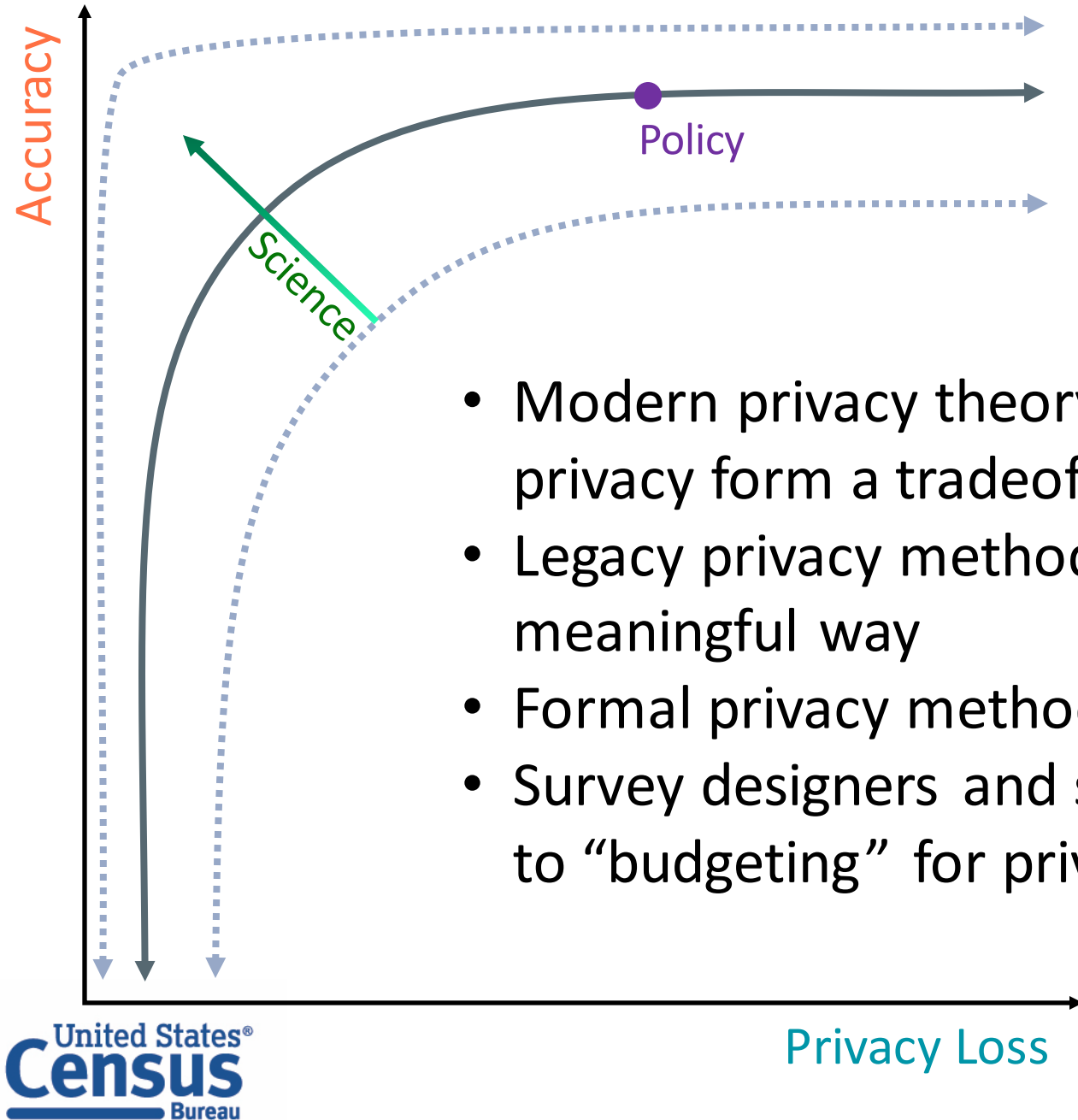
Balancing data release under Title 13



- Census Bureau surveys have competing requirements
 - Release statistics
 - Protect privacy
- Quality of ACS releases are supported by numerous metrics
- Privacy of ACS releases is not currently meaningfully quantified

There is no free lunch

- Modern privacy theory makes clear that accuracy and privacy form a tradeoff
- Legacy privacy methods could not quantify privacy loss in a meaningful way
- Formal privacy methods **can**
- Survey designers and stewards are not accustomed to “budgeting” for privacy



Moving away from illusions of privacy

- Legacy privacy methods tied privacy to arguments made using data available at the time
- But data stewards control little about data sources and uses:
 - Whether a respondent responds
 - What external data is held on respondents
 - Resources of would-be attackers (chance of disclosure)
 - Socio-economic forces that drive data sensitivity (impact of disclosure)
- No control of chance or impact means no control of privacy risk!



Privacy and sampling: It's complicated

- Sampling has been called a privacy technique
- But those in the sample have an increased effect upon survey outputs
- Adjustments to weights can increase this individual impact
- Surveys also tend to collect more sensitive data than censuses
- Interaction between formal privacy methods and surveys is an active research area^{†‡}



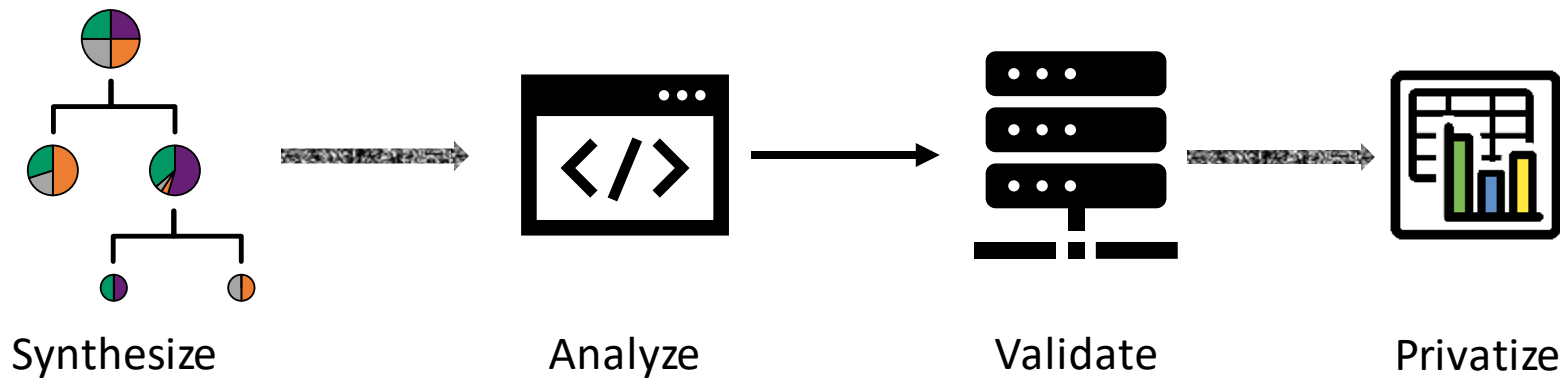
When privacy needs lead privacy science

- Re-identification of respondent information is a Census Bureau issue[†]
- Respondents deserve better protection across products now
- Other paradigms, such as synthetic data, can help fill the gap between swapping and formal privacy



Trust but verify

- Making synthetic data to satisfy all use cases is impossible
- Current research uses classification trees to create synthetic data
- Results of trees are assessed on a generalized marginal metric
- Users will be able to validate synthetic output against internal data

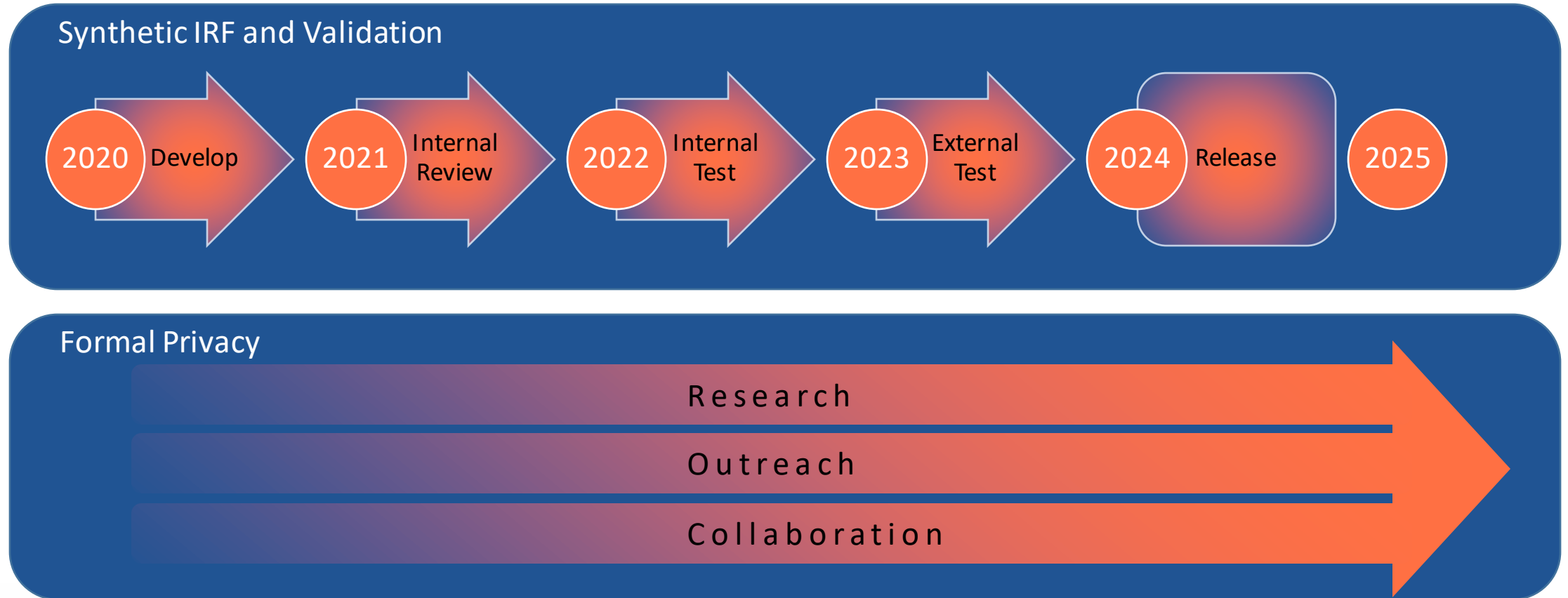


Making public microdata synthetically

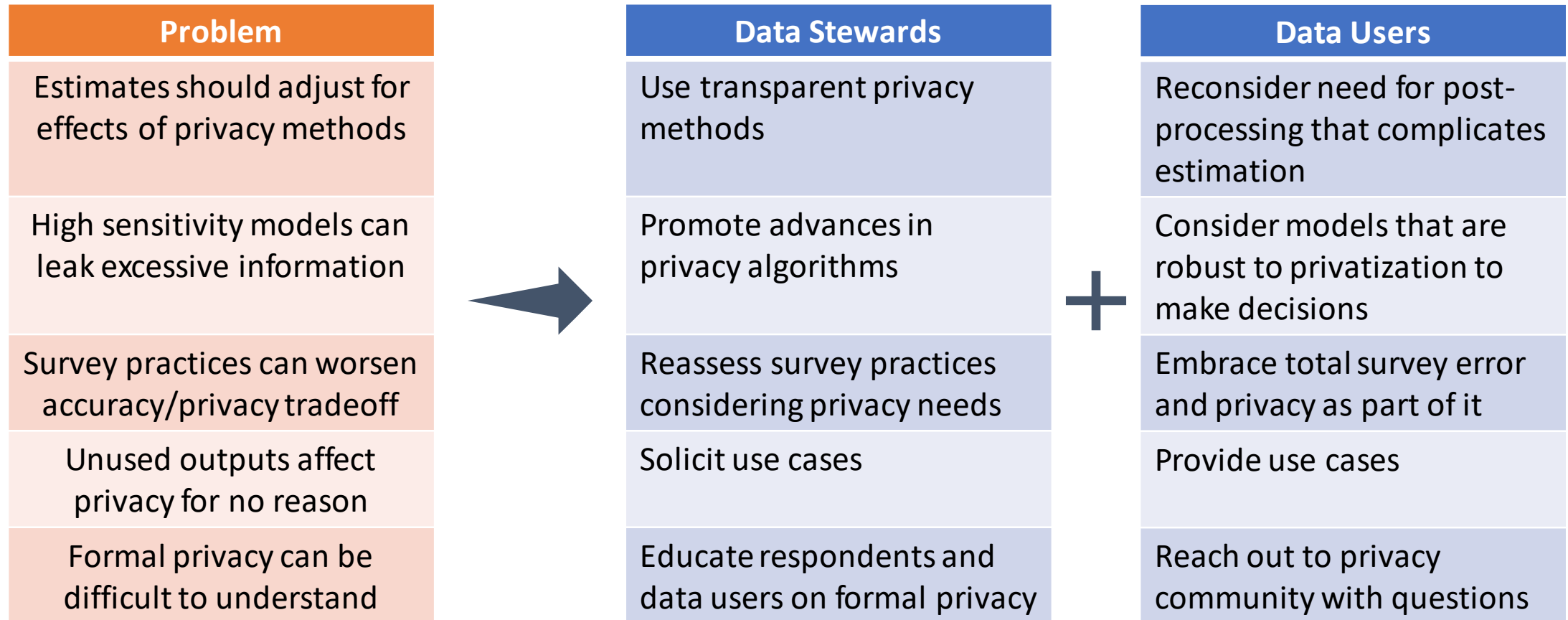
- Synthetic data methods are well-known to stats community
 - In use for certain surveys (SIPP[†], ACS)
 - Software packages already available (e.g., synthpop in R)
- ACS will research a new fully-synthetic data product
- Data will be based on an Internal Reference File (IRF), with no other privacy methods applied



ACS Privacy Modernization Timeline



Working together for responsible data



Contact us

acsprivacy@census.gov