Discussion:
Academic-Statistical Agency Collaboratives to Create Data Infrastructure for Evidence-Building

FCSM Session K-4 November 4, 2021

Daniel L. Goroff

Not necessarily presenting opinions of the Alfred P. Sloan Foundation or any other institutional affiliation.

The Promise of Administrative Data

- Not originally collected for research or statistical purposes.
- E.g. transactions records of retailers, government agencies, social media, etc.
- Potentially more consistent, timely, and granular than surveys.
- Traditional surveys are getting harder to carry out in any case.
- Neither individuals nor companies want to fill out surveys these days.

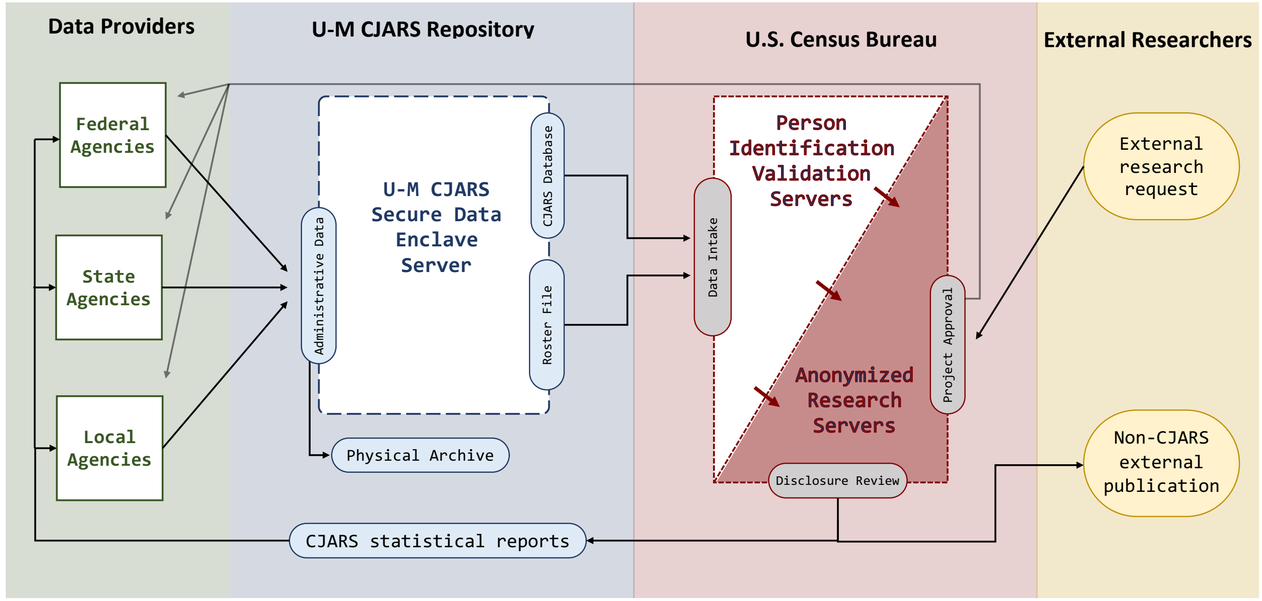
The Perils of Administrative Data

- Not originally collected for research or statistical purposes!
- Little regard for permissions or other privacy protections, for metadata or other vital documentation, for consistency or validity checks, etc.
- Negotiations concerning access terms can be long and frustrating.
- What if sources front run? Manipulate? Change or stop their data flows?
- The data is almost never a representative or probability sample of any population of interest to the federal statistical system.

1. RESET

- Re-engineer economic indictors like CPI for speed, consistency, granularity
- Use item-level transactions data that firms and aggregators already have
- Capture hedonic quality information and compare different indices
- Results are startling and suggest that federal statistics my systematically overestimate inflation by so much that macro textbooks need revising

2. CJARS

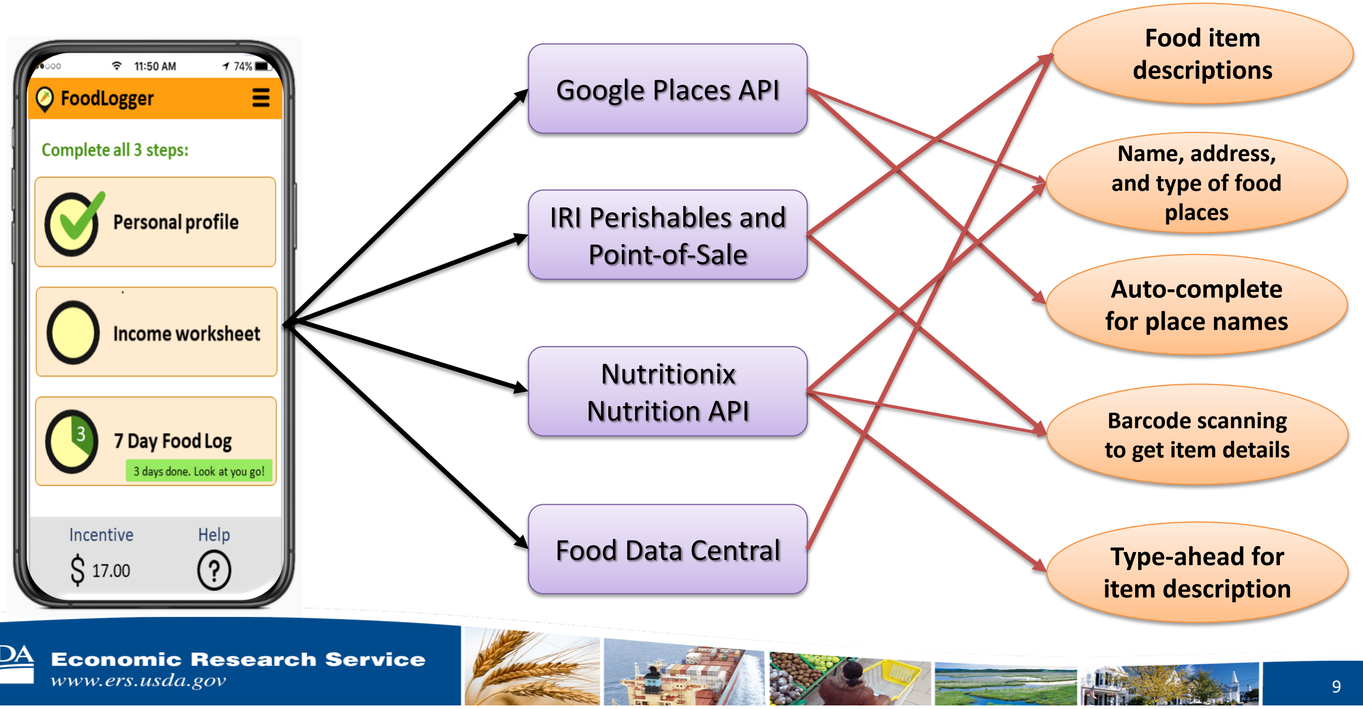


- Incomplete criminal justice data limits research questions, admits inconsistent definitions of concepts like recidivism, hinders search for effective policies
- CJARS collects, harmonizes, and links records to track individuals and cases across jurisdictions and over time: 2b records; 175m events; 36m individuals
- Limited economies of scale and limited sustainability based on private funders

3. Decennial Linkages

- Mircodata files for 1960-1990 exist but do not include names
- Names were handwritten on census forms
- Census forms are stored on 250,000 microfilm reels
- Forms and data are highly restricted to protect respondent confidentiality
- OCR gets 95% of ages but about 82% of last names

4. Agricultural & Food Data



- FoodAPS is first comprehensive & national data on household food purchases
- 35 active researchers, 12 active institutions, 25 USDA grans, 22 papers
- Economic Research Service also studies food production industries
- Cooperative Businesses are particularly important but challenging to identify

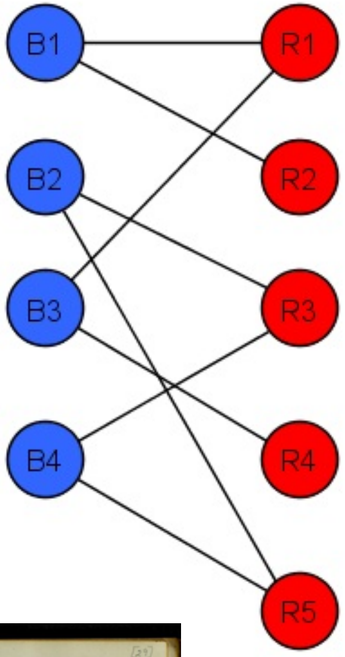
Some Sloan-Supported Data Projects

4. Agricultural & Food Data Related

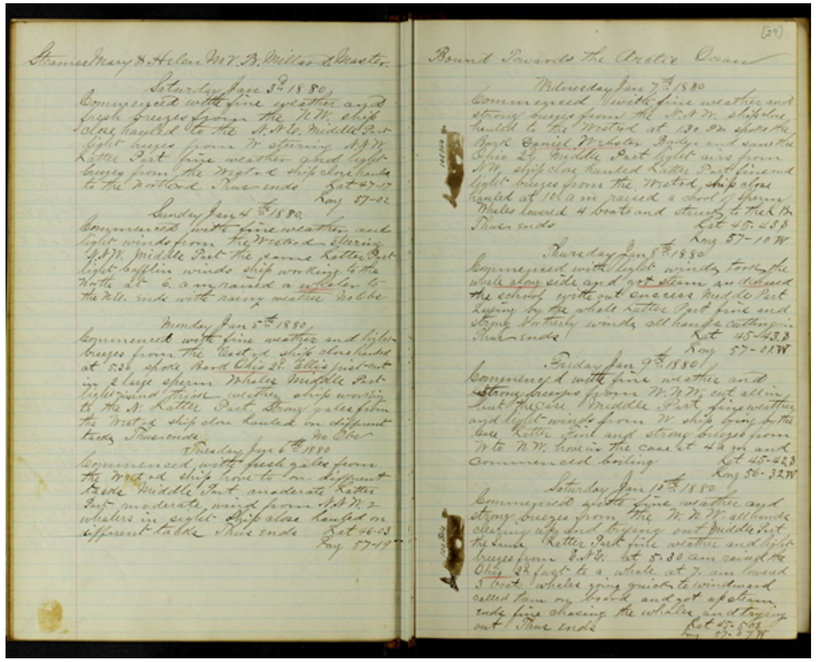
- Global Legal Entity Identifier Foundation

- Sampling Frame Projects

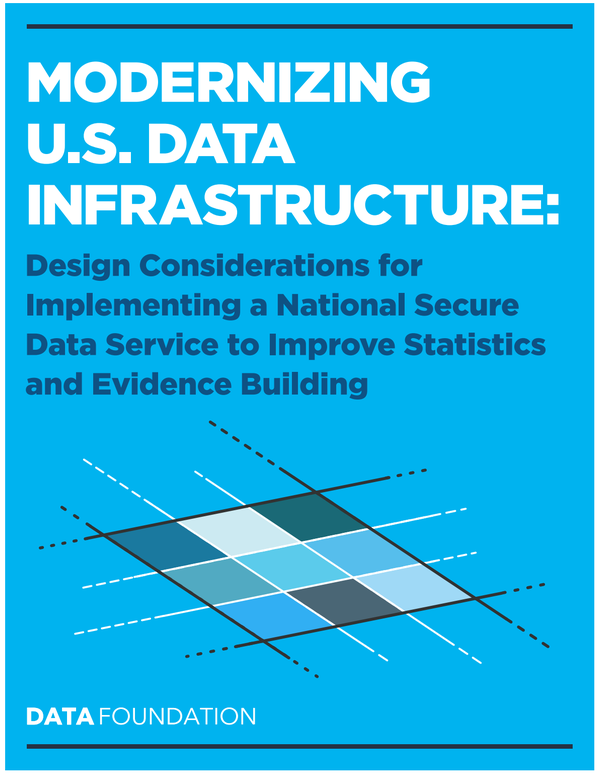
3. Decennial Linkage Related



- Bayesian Entity Resolution and Linking
- Old Weather



2. CJARS Related



- Evidence-Based Policymaking Act Implementation
- Equitable Data Working Group

THE WHITE HOUSE

Sec. 9. Establishing an Equitable Data Working Group. Many Federal datasets are not disaggregated by race, ethnicity, gender, disability, income, veteran status, or other key demographic variables. This lack of data has cascading effects and impedes efforts to measure and advance equity. A first step to promoting equity in Government action is to gather the data necessary to inform that effort.

1. RESET Related

- Administrative Data Research Intermediaries
  (rather than negotiating one by one unsustainably)
  e.g. Institute for Research on Innovation and Science

- Privacy-Protecting Research Protocols
  (since anonymization is not really possible)
  e.g. Differential Privacy and other methods