CIVIC
DIGITAL
FELLOWSHIP

# Probabilistic Record Linkage and Imputation for the Commodity Flow Survey

**Merritt Smith**

**Supervised by Emily Wiley and Christian Moscardi**

**Economic Reimbursable Surveys Division**

United States® Census Bureau

**Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.**

## Commodity Flow Survey (CFS)

- Conducted as a joint effort by the Bureau of Transportation Statistics (BTS) and the U.S. Census Bureau
- Used to assess demand for and use of freight transportation in the US
- Can help answer questions like:
  - How are goods being shipped?
  - Where are goods being shipped?
  - What infrastructure is most used and useful?
- Carried out every 5 years (next in 2022)
- 6 million shipments in 2017, 3.6% are exports

## Foreign Trade Export Declarations (FTD)

- Received from International Trade Indicator Area
- Have a Port of Exit (POE) attached to every record

**Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.**

# Problem: Where do goods leave the US?

- In 2017, we used GeoMiler to answer this
- GeoMiler, though, is:
  - Expensive
  - Not scalable
- Task:
  - Replicate GeoMiler's Port of Exit prediction ability for CFS
  - Do this by connecting foreign trade shipment records to CFS data
  - Make it scalable
    - CFS expecting to have up to 100x more records

**Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.**

# Record Linkage: CFS to FTD

**Do these records refer to the same thing?**

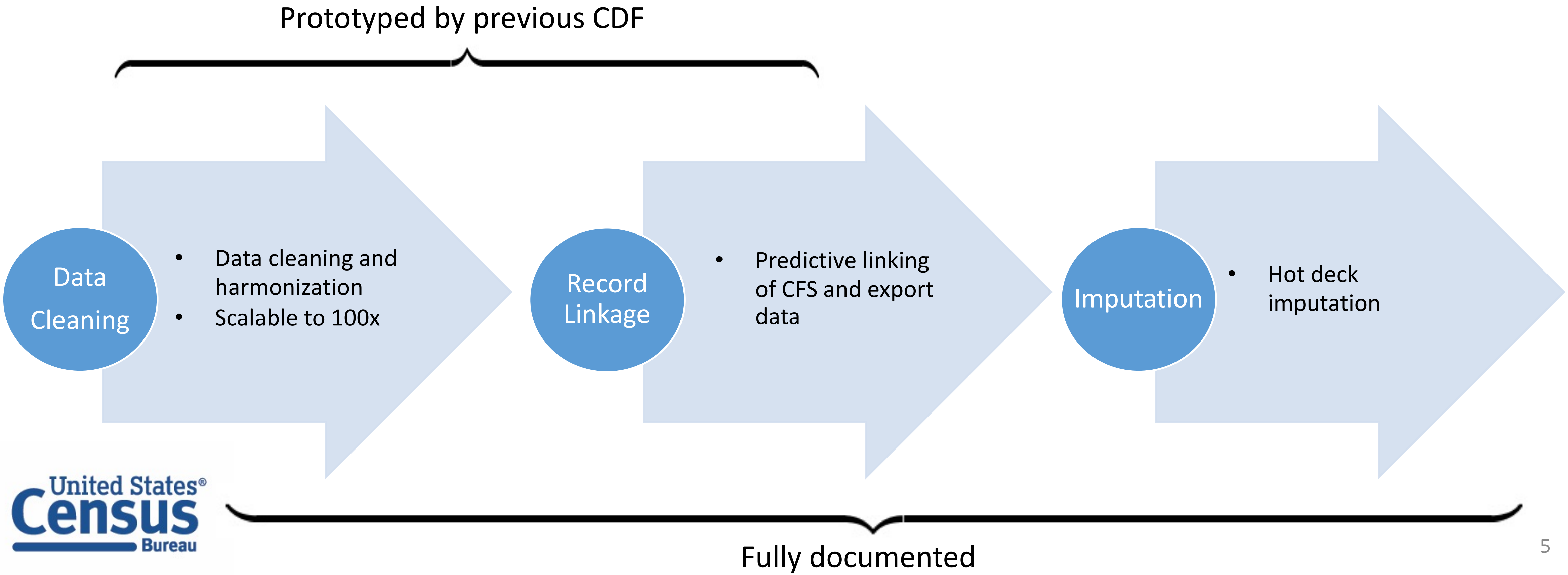
| Feature | CFS Record | FTD Record |
|---|---|---|
| State | IL | IL |
| Country | China | China |
| SCTG | 23 | 23 |
| EIN | 12345678 | 12345678 |
| Shipment ID | 123XYZ | 456789ABC |
| Date of Shipment/Clearance | 11/01 | 11/07 |
| Value | 6560 | 6560 |
| Weight | 2999 | 3000 |

(Mock data)

**Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.**

# Deliverables & Process

- Fully featured, robust, scalable ML pipeline from data to POEs
- Capacity for additional record linkage training through analyst input



Prototyped by previous CDF

**Data Cleaning**

- Data cleaning and harmonization
- Scalable to 100x

**Record Linkage**

- Predictive linking of CFS and export data

**Imputation**

- Hot deck imputation

Fully documented

## Benefits

- Approx. 90% fewer person-hours due to automation
- Reduced cost as a result
- Potentially decreases respondent burden -> more survey responses

## Next Steps

- Integrate pipeline into Census software
- Use pipeline as part of next year's CFS

**Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.**