The Challenges and Opportunities of Record Linkage and AI

Carla Medalia

5/2/24

CNSTAT AI Day for Federal Statistics

Washington DC

United States® Census Bureau

Note: Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau.

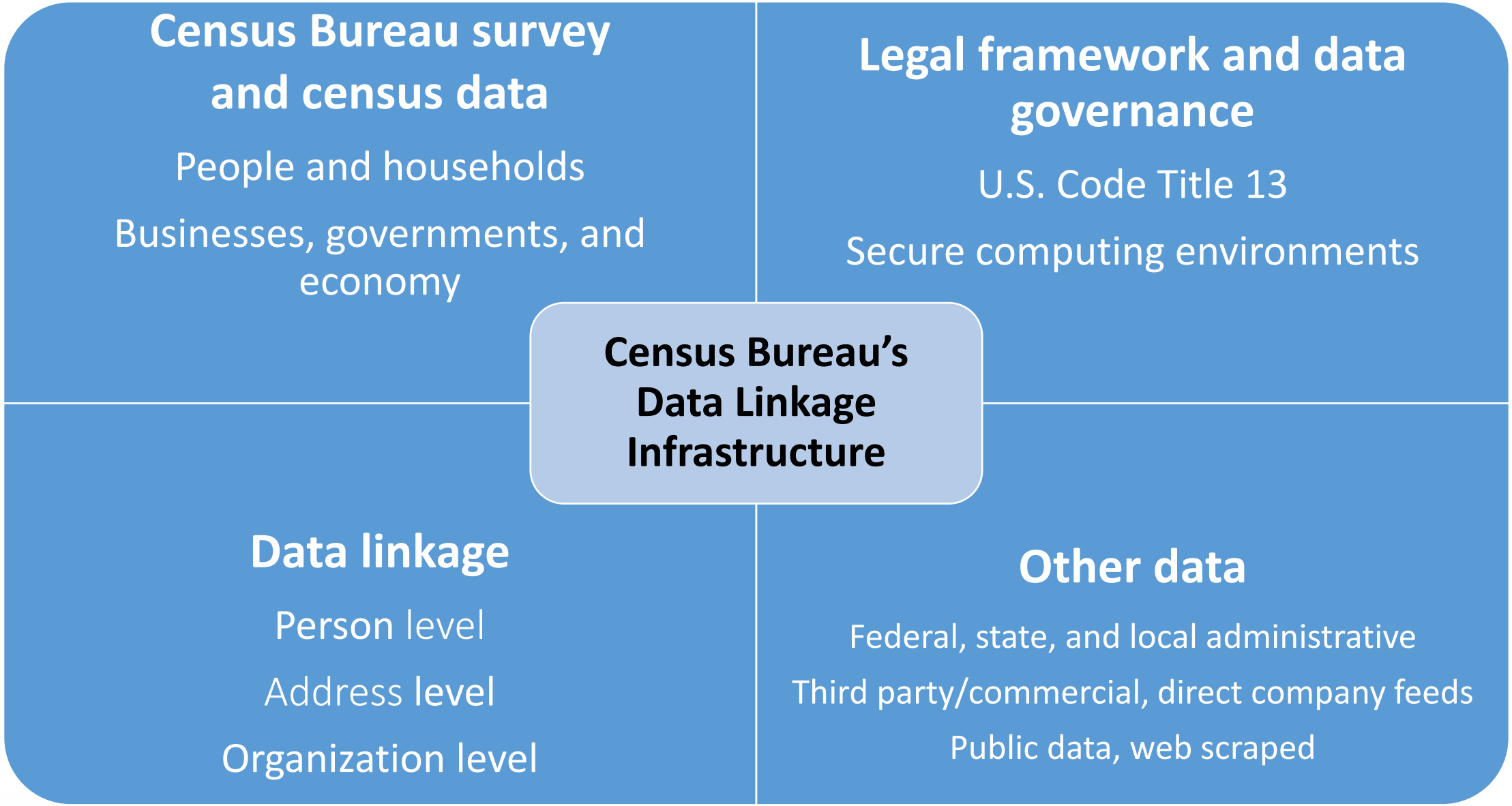
Today's talk

- Why does the Census Bureau need record linkage?
- Why does the Census Bureau need AI to enable record linkage?
- High-level summary of various ML/AI record linkage projects
- Common theme: challenges and opportunities
- Shout out to Krista Park and many others at the Census Bureau!

United States® Census Bureau

Why does the Census Bureau need record linkage?

United States® Census Bureau



United States® Census Bureau

Census.gov/datalinkage

Why does the Census Bureau need AI to enable record linkage?

- To enable record linkage with imperfect information
- To link large data files and a lot of data files
- To reduce resources required to link records
- To improve data linkage quality
- To link data sources that are not subset of a reference file
- To reduce linkage bias
- … and more

United States® Census Bureau

Fun facts

- ML-based record linkage at the Census Bureau dates back to the 1980s
- The Census Bureau currently uses or is researching the following methods of record linkage
  - Deterministic linkage
  - Probabilistic linkage
    - Supervised ML
    - Unsupervised ML
    - AI: large language models
    - Generative AI

United States® Census Bureau

Person-level linkages

United States® Census Bureau

Person Identification Validation System (PVS)

- Goal: assign unique identifiers to person-level records using the Person Identification Validation System (PVS)
- Role of ML: determine optimal blocking/search strategy, conduct searches and make matches, and deduplication
- Data sources: administrative data, commercial data, Census Bureau survey and census data
- Challenges: data sources to link are getting more challenging to match and may be less structured
- Opportunity: use Natural Language Processing and deep learning
- Contacts: damon.r.smith@census.gov and matthew.e.bouch@census.gov

United States® Census Bureau

Person-level linkage: cooperative agreements

- Goal: Leverage Cooperate Agreement authority to improve record linkage algorithms, methods, and technologies
  - University of Arkansas Little Rock, University of Connecticut, University of Michigan, University of Washington
- Role of ML: data cleaning; tune large language model (generative AI), feature selection, historical record linkage, pipelines to modularize linkage
- Data: simulated/synthetic data, administrative data, Census bureau data
- Challenges: data sources to link are getting more challenging to match and may be less structured
- Opportunities: apply record linkage methods from the mail marketing, credit, and biotech industries to address Census Bureau linkage problems
- Contact: krista.park@census.gov

United States® Census Bureau

Address-level linkages

United States® Census Bureau

Address

- Goal: Accurately geocode all addresses; detect changes in landscape; identify and resolve address and feature coverage discrepancies
- Role of ML: automated imagery-based change detection
- Data sources: commercial data (satellite imagery and parcel data), administrative data, Master Address File
- Challenges: keep pace with changes to the landscape and reduce human sources of bias
- Opportunities: Can we develop ML models to reduce the need for human validation and learn from previous work?
- Contacts: daniel.l.keefe@census.gov, dolly.v.garcia@census.gov, seth.schowalter@census.gov, elvis.a.martinez@census.gov, gustavo.maldonado.davila@census.gov

United States® Census Bureau

Organization-level linkages

Businesses and Governments

United States® Census Bureau

Multiple Algorithm Matching for Better Analytics (MAMBA)

- Goal: develop a scalable software program to enable business linkages; scaffolding should be agnostic to algorithms
- Role of ML: assesses multiple string comparators to determine optimal approach to generate precise matches
- Data sources: administrative data, commercial data, Census Bureau survey and Economic Census data
- Challenges: how to disambiguate duplicate records across files?
- Opportunities: use clustering algorithms
- Contact: nathan.goldschlag@census.gov and john.cuffe@census.gov

United States® Census Bureau

Production business record linkage

- Goal: create a pipeline to clean data, send to MAMBA, then output matched records
- Role of ML: use training data to develop matching models; experiment with different truth decks
- Data sources: reference file is Business Register; match commercial, web-scraped, and administrative data
- Challenge: businesses have both physical and mailing addresses; commercial data sources may have varying quality BII
- Opportunity: Triangulate with other reference files; role of subject matter expert and human-machine interaction
- Contact: jessica.l.wellwood@census.gov and rebecca.j.hutchinson@census.gov

United States® Census Bureau

Organizations: justice agencies and facilities

- Goal: link different lists of justice agencies and facilities to the Master Address File (MAF) and Governments Master Address File (GMAF)
- Role of ML: triage/prioritize potential matches to have SMEs manually examine to ensure data quality
- Data sources: administrative data, web scraping, Census Bureau data
- Challenge: identify new data sources and parse unstructured text to validate matches
- Opportunity: incorporate spatial, web scraped, and population data into models
- Contact: keith.ferguson.finlay@census.gov

United States® Census Bureau

Other developing research

- Research and experiment with new methodologies to improve and optimize record linkage across multiple software platforms
  - Adjusting the Statistical Analysis on Integrated Data
  - Entity Resolution and Merging Noisy Databases
  - Record-Linkage Support for the Decennial Census
  - Contacts: emanuel.ben.david@census.gov, yves.thibaudeau@census.gov, daniel.weinberg@census.gov, rebecca.carter.steorts@census.gov, and others
- Other research presented at CNSTAT AI Day today!
  - A Semi-Supervised Active Learning Approach for Block-Status Classification
  - Explainable Artificial Intelligence for Bias Identification and Mitigation in Demographic Models
  - Contact: atul.rawal@census.gov

United States® Census Bureau

Challenges and Opportunities

- Reference files may miss records → Seek out new sources of data to incorporate into reference files
- Linkage quality depends on PII and BII → Leverage unstructured text and develop large language models to analyze data
- Machines are great, but we still need humans → Capture human-validated efforts and train future models
- Linkage bias → Proactively identify sources of bias and develop models to address
- How to determine inflection point for when models no longer work? → feedback loop to continually reevaluate models, fine tuning models for new sources of data

United States® Census Bureau

Thank you!

For more information, please reach out to me or any of the experts identified throughout the presentation

Carla.Medalia@census.gov

United States® Census Bureau