



# CIVIC DIGITAL FELLOWSHIP



**Generating a Person to Place of Birth Dataset Crosswalk for the Census Numident**

**Francesca Marini**

**Supervised by Katie Genadek & Keith Finlay**

**Economic Reimbursable Surveys Division**

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

# Motivation

## What is the Census Numident?

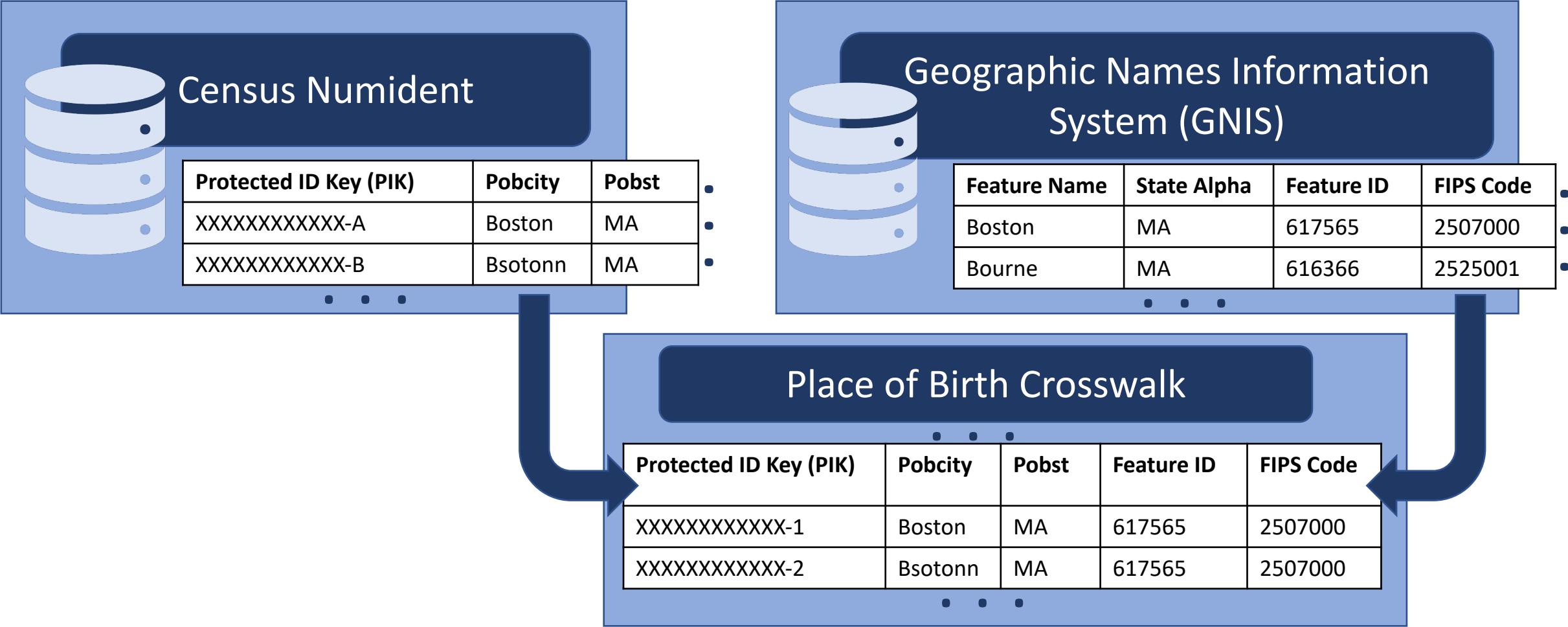
- Based on **Social Security application** data
- Each person assigned **unique anonymous PIK**, generated based on their SSN
- Data contains **place of birth strings**
- Crosswalk will map **PIK and place of birth strings** to numerical **geographical location codes**

## Benefits

- LOTS of people use the Numident for **research**, both in and outside of Census
  - Impact of early life conditions on future socioeconomic / medical outcomes
  - Population projections
- **Better coverage of very young children** in demographic surveys, including decennial censuses
- Potential benefits to **SSA programs**

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

# Objective: Person to Place of Birth Crosswalk



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

# Key Challenges

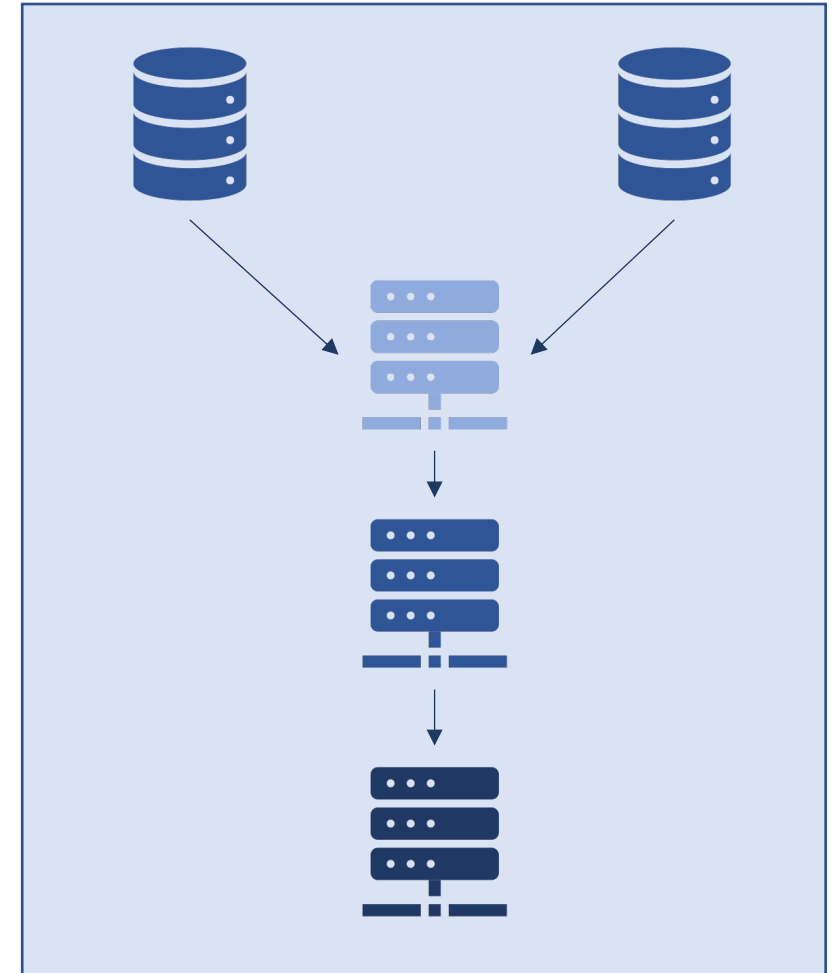
- **String truncation** to 12 characters in the dataset
  - “Charlottesville” becomes “Charlottesvi”
  - Ambiguity between towns and counties of the same name
- **Typos** in the manual entering of place names
  - Need to be able to match towns / counties with spelling mistakes
  - Need to be able to fix cases where the state abbreviation was incorrectly entered
- **Lower confidence** in non-exact matches between strings and geographical codes
  - How to better estimate and improve the quality of string matches?
- Many attempts at crosswalk generation but **no single standard crosswalk** for best use

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

# Crosswalk Generation Process

## Steps to Build Crosswalk:

1. Preprocess Numident and GNIS datasets
2. Exact Matching
3. Duplicate Resolution
4. Fuzzy Matching
5. Manual Corrections
6. Additional Probabilistic Corrections
7. Other Methods (ongoing)



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

- ✓ Numident place of birth string to location code crosswalk
- ✓ Numident person to place of birth location code crosswalk
- ✓ Process for integrating new Numident data into crosswalks
- ✓ Code and crosswalk documentation

Next Steps: further assess the confidence of matches to improve match quality (ongoing)

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