CIVIC
DIGITAL
FELLOWSHIP

Novel Data Linkage in Support of the Decennial Census Digitization and Linkage Project

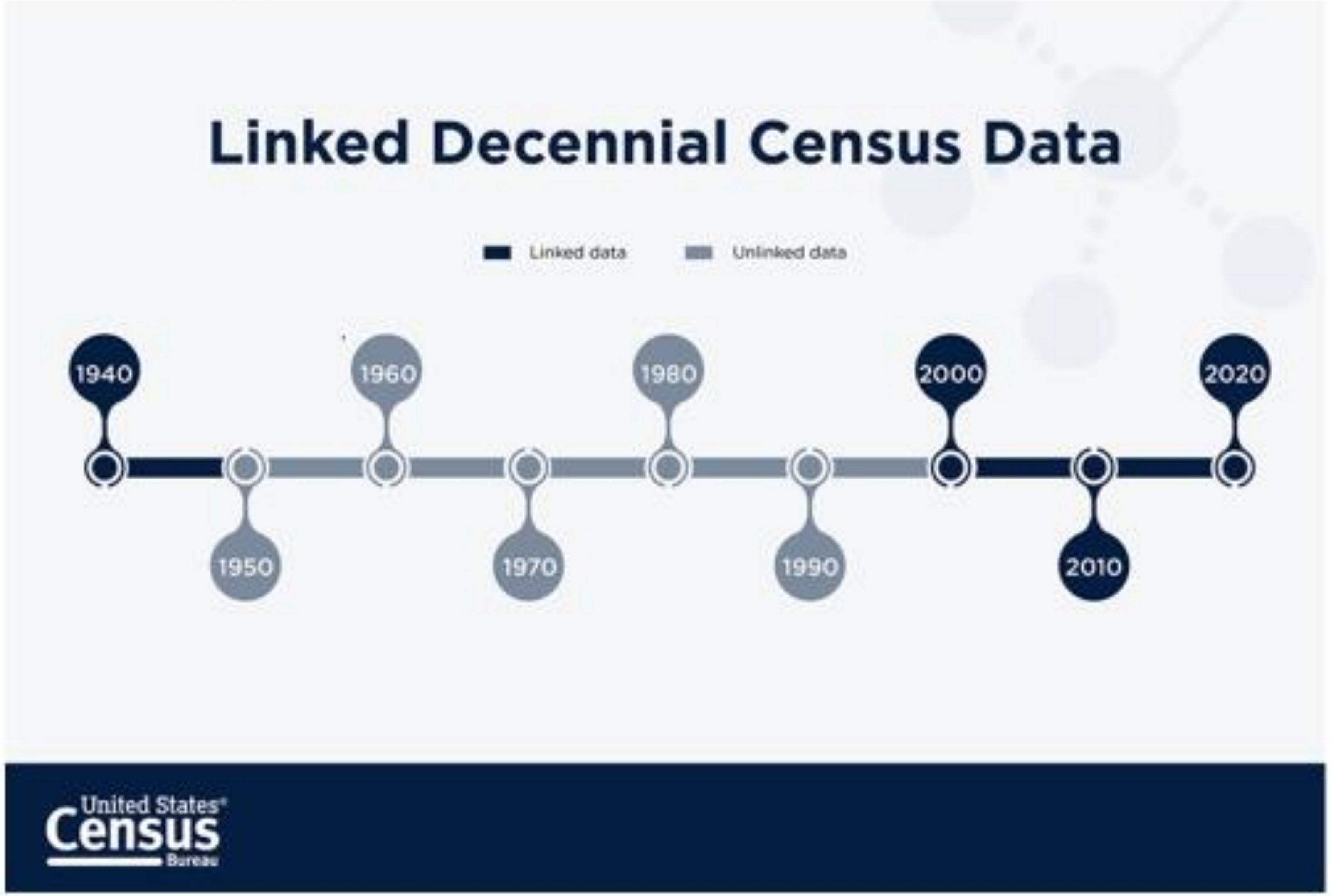
Sohail Kamdar

Supervised by Katie Genadek and John Sullivan

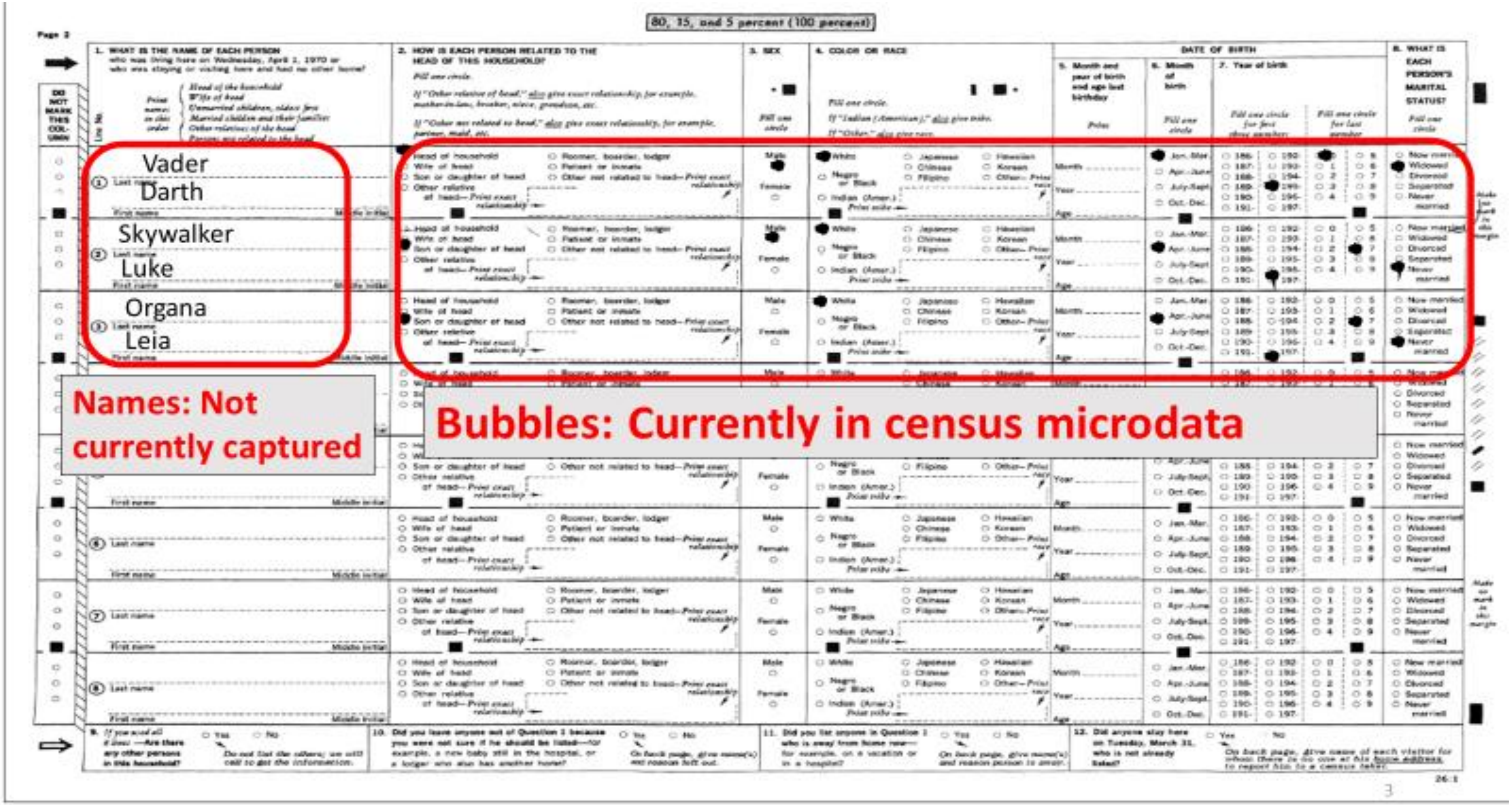
Economic Reimbursable Surveys Division

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

Background and Motivations



Linking person-level census microdata over time will fill a 5-decade gap in longitudinal infrastructure for researchers studying long term social, economic, and health outcomes.



Page 3

80, 15, and 5 percent (100 percent)

DO NOT MARK THIS COLUMN

1. WHAT IS THE NAME OF EACH PERSON who was living here on Wednesday, April 1, 1970 or who was staying or visiting here and had no other home?

Print names in this order: Head of the household; Wife of head; Unmarried children, oldest first; Married children and their families; Other relatives of the head; Persons not related to the head

2. HOW IS EACH PERSON RELATED TO THE HEAD OF THIS HOUSEHOLD?

Fill one circle.

If "Other relative of head," also give exact relationship, for example, mother-in-law, brother, niece, grandson, etc.

If "Other not related to head," also give exact relationship, for example, partner, maid, etc.

3. SEX — *Fill one circle*

4. COLOR OR RACE — *Fill one circle. If "Indian (American)," also give tribe. If "Other," also give race.*

DATE OF BIRTH

5. Month and year of birth and age last birthday — *Print*

6. Month of birth — *Fill one circle*

7. Year of birth — *Fill one circle for first three numbers* / *Fill one circle for last number*

8. WHAT IS EACH PERSON'S MARITAL STATUS? — *Fill one circle*

1	Last name: Vader	First name: Darth
2	Last name: Skywalker	First name: Luke
3	Last name: Organa	First name: Leia
4	Last name	First name
5	Last name	First name
6	Last name	First name
7	Last name	First name
8	Last name	First name

Each line: ○ Head of household ○ Wife of head ○ Son or daughter of head ○ Other relative of head—*Print exact relationship* ○ Roomer, boarder, lodger ○ Patient or inmate ○ Other not related to head—*Print exact relationship* | Male ○ Female ○ | ○ White ○ Negro or Black ○ Indian (Amer.) *Print tribe* ○ Japanese ○ Chinese ○ Filipino ○ Hawaiian ○ Korean ○ Other—*Print race* | Month ___ Year ___ Age ___ | ○ Jan.-Mar. ○ Apr.-June ○ July-Sept. ○ Oct.-Dec. | ○ 186- ○ 187- ○ 188- ○ 189- ○ 190- ○ 191- ○ 192- ○ 193- ○ 194- ○ 195- ○ 196- ○ 197- | ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 | ○ Now married ○ Widowed ○ Divorced ○ Separated ○ Never married

Make no mark in this margin

9. If you used all 8 lines —Are there any other persons in this household? ○ Yes ○ No — *Do not list the others; we will call to get the information.*

10. Did you leave anyone out of Question 1 because you were not sure if he should be listed—for example, a new baby still in the hospital, or a lodger who also has another home? ○ Yes ○ No — *On back page, give name(s) and reason left out.*

11. Did you list anyone in Question 1 who is away from home now—for example, on a vacation or in a hospital? ○ Yes ○ No — *On back page, give name(s) and reason person is away.*

12. Did anyone stay here on Tuesday, March 31, who is not already listed? ○ Yes ○ No — *On back page, give name of each visitor for whom there is no one at his home address to report him to a census taker.*

26:1

3

Names: Not currently captured

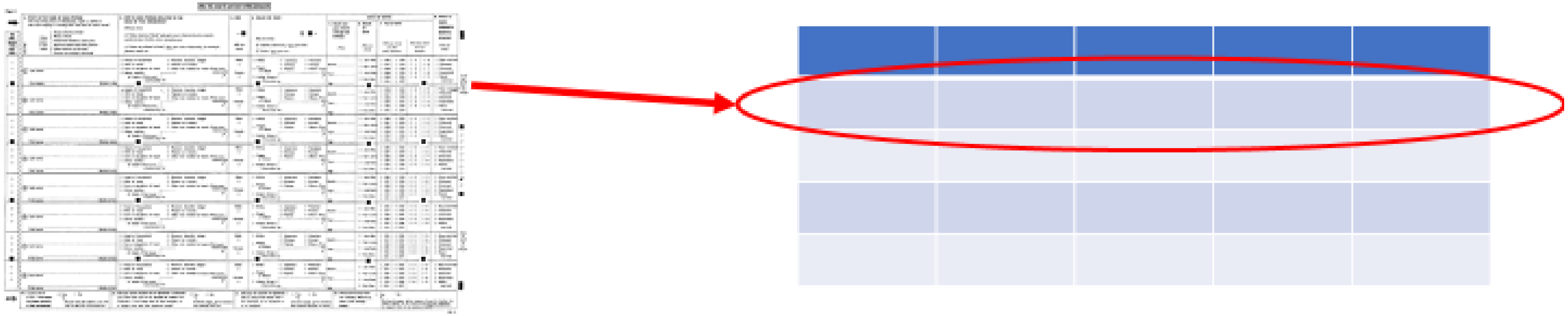
Bubbles: Currently in census microdata

United States® Census Bureau

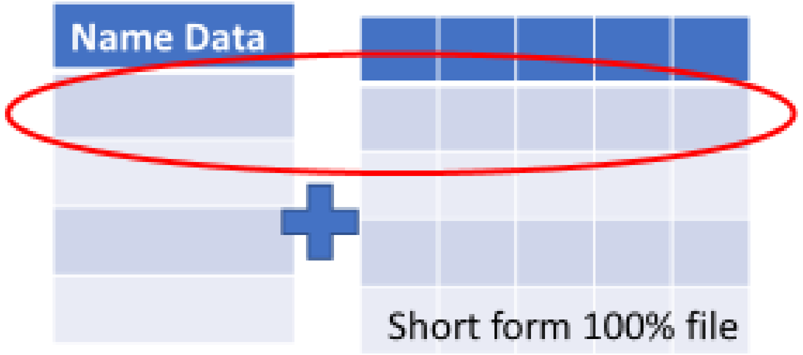
Objectives

Linking keyed microfilm data to short form census microdata

1. Survey forms on microfilm converted to keyed datasets



2. Keyed image data and census micro data are linked



Linkage process:

Name	Age	Number in Household	Sex	Relationship	County
Darth Vader	45	3	0	0	0001
Luke Skywalker	25	3	0	2	0001
Leia Organa	26	3	1	2	0001
(Dataset from images)					

1. Clean microdata and name data such that records are comparable
2. Match household data within geographic areas
3. Match person level data using demographic information
4. Use fuzzy matching to deal with data errors.

Age	Number in Household	Sex	Relationship	County
45	3	0	0	0001
25	3	0	2	0001
25	3	1	2	0001
(Census dataset)				

Additional Challenges and Potential solutions

Challenges	**Solutions**
Lack of string identifiers	Block data into useful variables
Opaque legacy documentation	Use only commonly coded variables
Geographic inconsistencies	Use alternative markers of geography to merge datasets

Preliminary Results and Next Steps

- Match rates for keyed data and microdata are between 80-90% for sample geographies
- Errors and inconsistencies between datasets are documented

Next steps:

- Resolving challenges with missing data in rural areas
- Creating scalable solution to deal with differences within regions