Identity in the Census

Finding people in more than one

What is Identity?

- A unique set of identifiers

What is an Identifier?

- Any measurable attribute
- In Census - Name, Age, Sex, Birth State
- AND Household characteristics

Basic Record Linking

- Generalize identifiers to block
- Compare within a block more specifically to match

Why?

- GEDCOM was about exchange – we've abandoned that in favor of linkage
- Local conclusions, remote evidence

Household Characteristics?

- Oldest male
- Oldest female
- Oldest boy
- Oldest girl

Types of Identifiers

- Cultural
- Biological

Cultural Identifiers

- Surname
- Given Name
- Family Role

Biological Identifiers

- Sex
- Age
- Parent / Child roles

Coding Identifiers

- Soundex
- Initials
- Birth year

Why code identifiers?

- Because matching doesn’t work
- Expressions of identifiers in records vary – granularity etc.
- To speed up comparisons by allowing blocking on a matched code

Examples: Carroll Co AR

- 1860 and 1870
- Surnames beginning with K and L

What kind of keys did you use?

- qry1860OldWoman
- Sex (f)
- Initial of Surname
- Initial of coded first name
- Estimated birth year / 5
- Example: 1860 Mary Keelan age 13
- fKM369

1860 Family

- John
- KEYES
- 30
- Hannah
- KEYES
- 27
- Housekey = mKJ366fKH366
- Less granular key = mKJfKH

Easy Match

- Surname Soundex
- First initial
- Birth Year

- Easy List

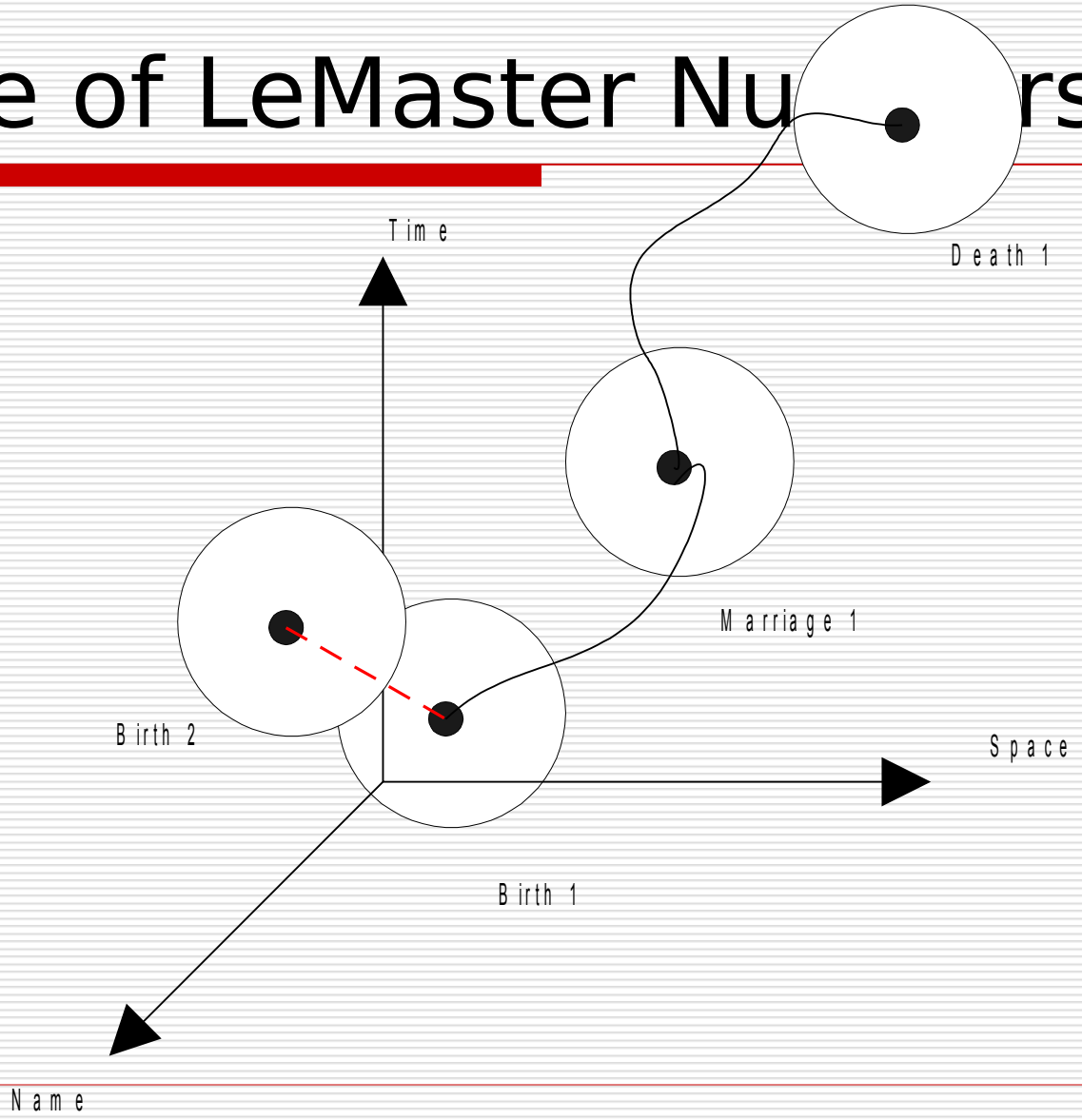
Other Matches

- Universe 778 records
- Key 2 – mKJ - 449 matches
- EasyList – 78 matches
- Key 3 – mKJ382 – 38 matches
- Mom and oldest boy – key3 – 8 matches – 1 right
- Housekey - mLA368fLE368 – 4 matches – 3 right

Work to do

- Measure the effectiveness of different sets of identifiers
- Scale the algorithms to larger data sets
- Abandon linking for a Cartesian Event Space

Example of LeMaster Nu rs



Questions