

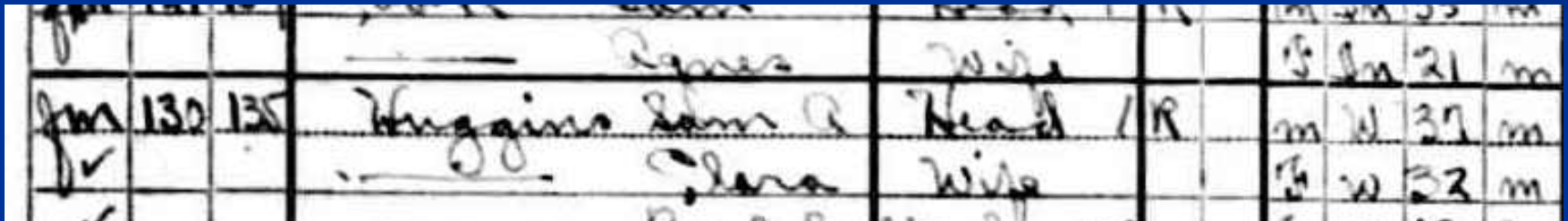
Census Index Merge

Jim Wray

Ancestry.com

Motivation

- What are the resources we possess?
 - Source Data (Image)
 - Interpretation of that data (Keyed digital index)



A photograph of a handwritten document, likely a census form, showing several rows of data. The text is written in cursive and is somewhat difficult to read. The visible text includes names like 'Agnes', 'Wife', 'Huggins Sam A', 'Head / R', 'Slava', and 'Wife'. There are also some numbers and other markings, such as '130 135' and 'm w 37 m'.

Index	Surname	Given	Gender	Race	Age
Head of Household	Wiggins	Sam A	Male	White	37
Every Name	Huggins	Sam A	Male	Indian	37

Complexity

- Head of Household index (heads plus surnames)
 - Approximately 36 million names
 - Keyed in 2001
- Full (every name) index
 - Approximately 107 million names
 - Keyed in 2005
- Brute Force Algorithm $O(n^2)$
 - Almost four quadrillion compares

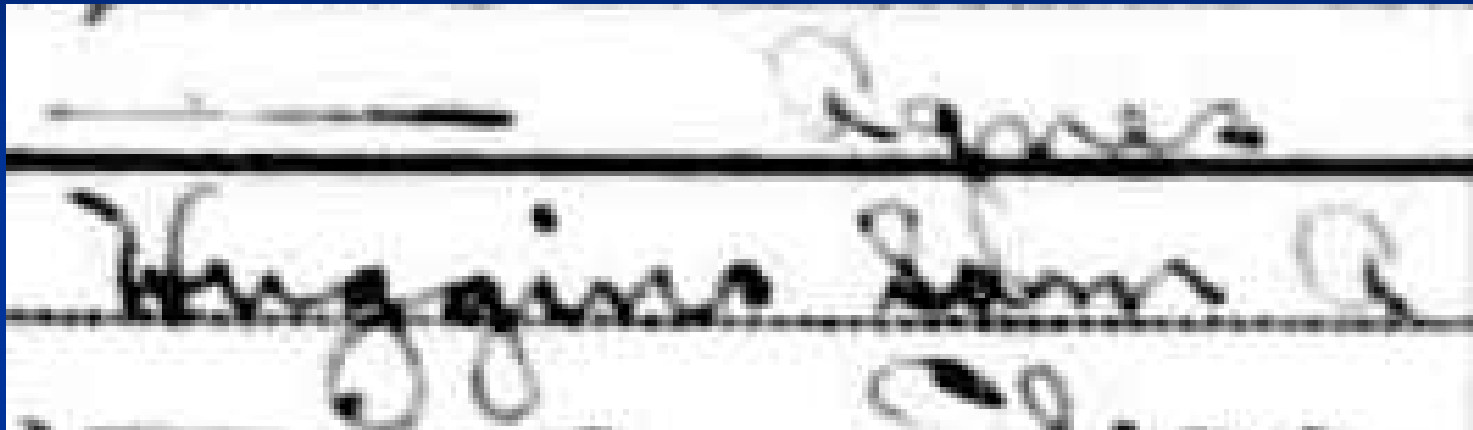
$$107,000,000 \times 36,000,000 = 3.852 \times 10^{15}$$

Complexity (cont.)

- Data is naturally partitioned, how well can we use that to our advantage?
 - Habitation: State, County, Township
 - Census metadata: Districts, Pages
 - Filming organization: Rolls
- Worst case algorithm using partitioning:

$$52,000 \times 17,000 \times 2,076 = 1.835 \times 10^{12}$$

Example



Index	Surname	Given	Gender	Race	Age
Head of Household	Wiggins	Sam A	Male	White	37
Every Name	Huggins	Sam A	Male	Indian	37

- *Wiggins or Huggins?* (Audience Poll)

Method Overview

- Using the partitions:
 - We iterate through the current search space one page at a time.
 - Compare each record by comparing same-field values in a weighted fashion to determine a match confidence score between 0 and 1.
 - Keep track of best match so far and best page for all matches (helps to determine if we need to widen the scope to a larger partition)


Method Specifics

- “Levenshtein” or “Edit-distance” metrics
 - Distance is **shortest sequence of edit commands** that transforms s to t . (s and t are the two strings)
 - Edit commands are copy, delete, insert, and substitute.
- In the example the Edit-distance from “Wiggins” to “Huggins” is 2. (two substitutions)
- How about comparing ages or other numbers?

Index	Surname	Given	Gender	Race	Age
Head of Household	Wiggins	Sam A	Male	White	37
Every Name	Huggins	Sam A	Male	Indian	37

Results – List View

The triangle indicates there is an alternate race, which is displayed in the expanded view.

View Record	Name	Home in 1920 (City, County, State)	Estimated Birth Year	Birthplace	Race	View Image
View Record	Sam A Huggins [Sam A Wiggins] ⚠	Ferdinand, Idaho, Idaho	abt 1883	Wisconsin	Indian ⚠	

This is the initial result or result set that users see after initiating a search on the 1920 census page. Then, if the user clicks on the “View Record” link, they are taken to the individual or detailed view. (next slide)

Results – Individual View

1920 United States Federal Census Record ⓘ

about Sam A Huggins

Name:	Sam A Huggins [Sam A Wiggins] ⚠
Age:	37 years
Estimated birth year:	abt 1883
Birthplace:	Wisconsin
Race:	Indian [White] ⚠
Home in 1920:	Ferdinand, Idaho, Idaho
Home owned:	Rent
Sex:	Male
Marital status:	Married
Relation to Head of House:	Head
Able to read:	Yes
Able to Write:	Yes
Mother's Birth Place:	Canada
Father's Birth Place:	Canada
Image:	870



 [View original image](#)

 [View blank form](#)

Conclusions

- Using multiple indexes (interpretations of the source data) combined into one increases the likelihood that an individual can be found.
- We can use natural partitions in the data to compartmentalize the search space, increasing speed and accuracy
- We used Levenshtein distance metrics for strings to effectively match records between indexes.