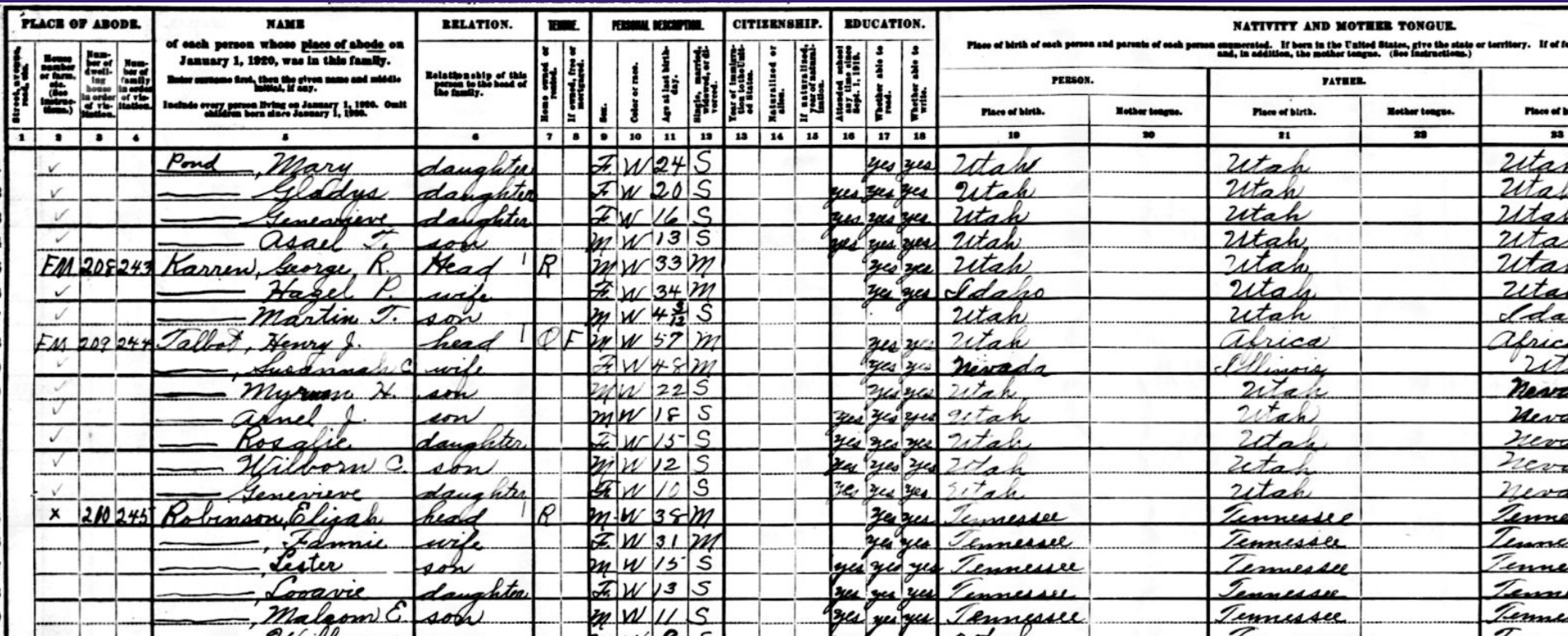
# Applications of Subword Spotting

Brian Davis

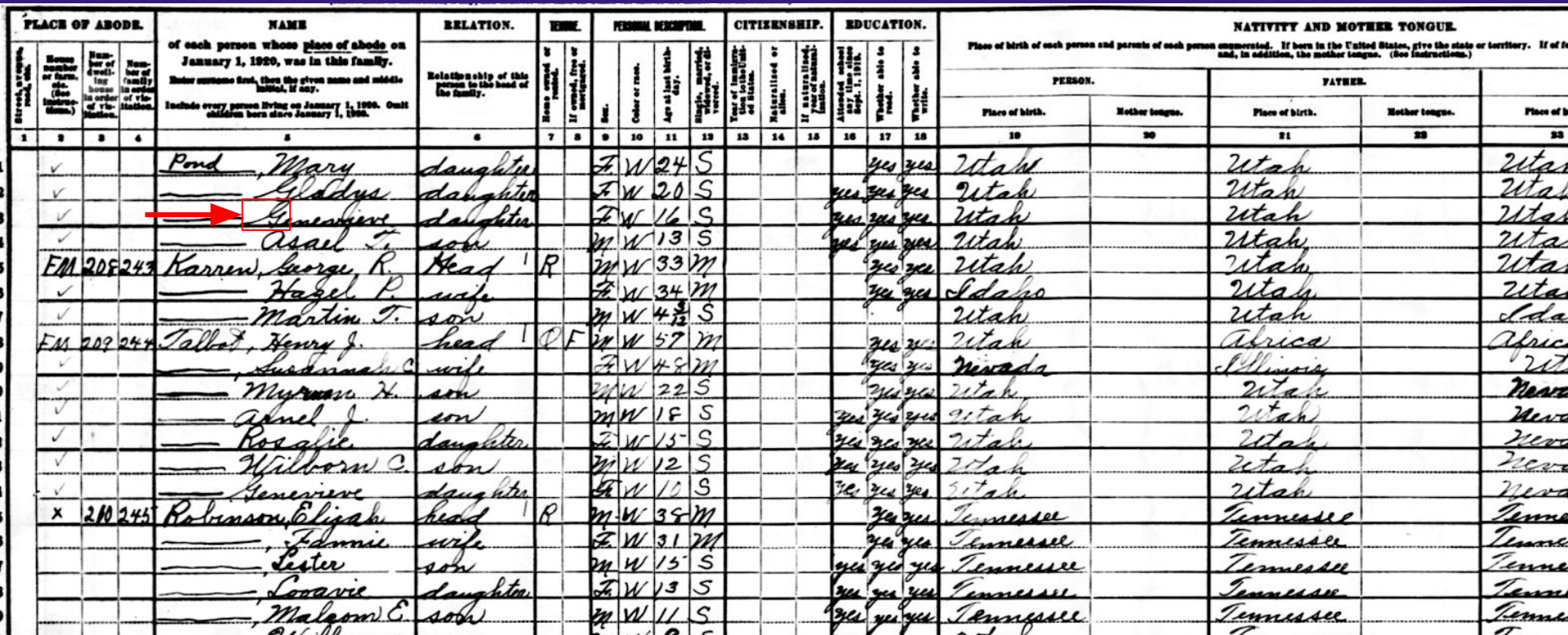
BRIGHAM YOUNG UNIVERSITY
Family History
TECHNOLOGY WORKSHOP

# A common scenario...



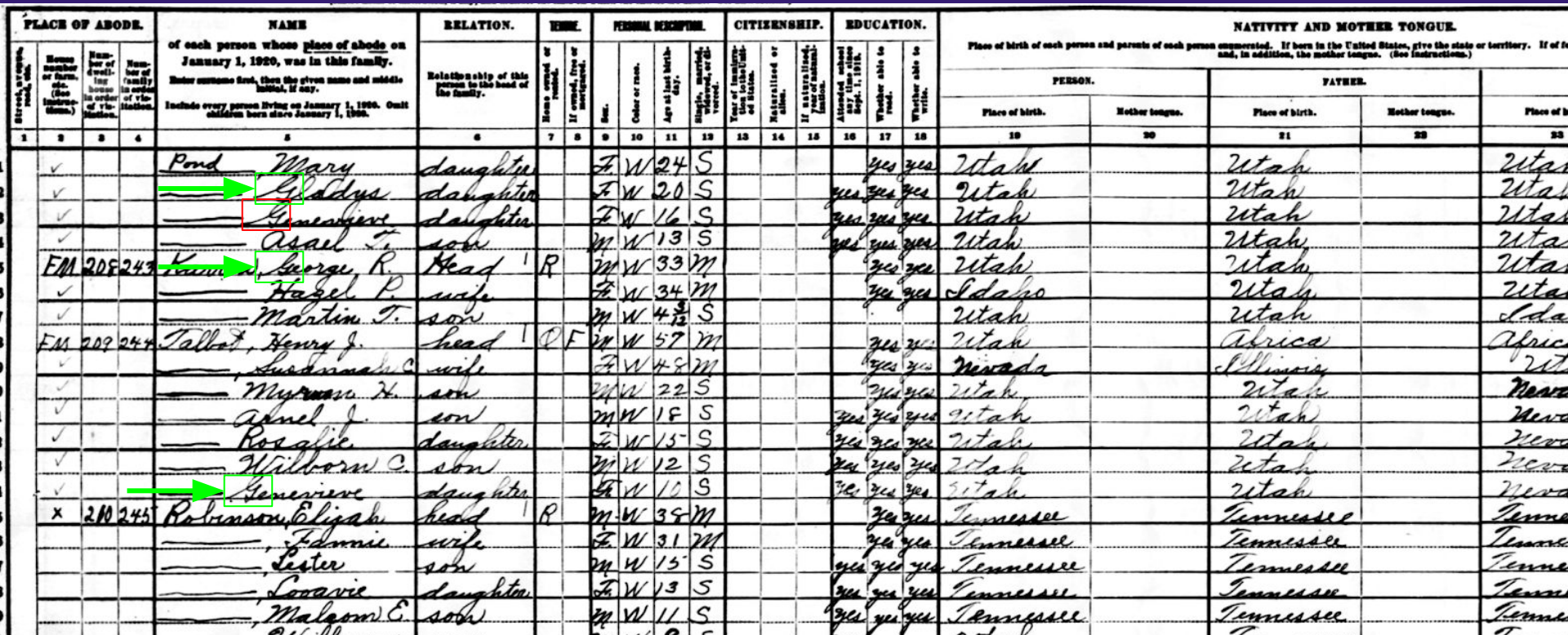
| PLACE OF ABODE. | | | | NAME of each person whose place of abode on January 1, 1920, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on January 1, 1920. Omit children born since January 1, 1920. | RELATION. | TENURE. | | PERSONAL DESCRIPTION. | | | | CITIZENSHIP. | | | EDUCATION. | | | NATIVITY AND MOTHER TONGUE. Place of birth of each person and parents of each person enumerated. If born in the United States, give the state or territory. If of f and, in addition, the mother tongue. (See instructions.) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | | | PERSON. | | FATHER. | | |
| Street, avenue, road, etc. | House number or farm, etc. (See instructions.) | Number of dwelling house in order of visitation. | Number of family in order of visitation. | | Relationship of this person to the head of the family. | Home owned or rented. | If owned, free or mortgaged. | Sex. | Color or race. | Age at last birthday. | Single, married, widowed, or divorced. | Year of immigration to the United States. | Naturalized or alien. | If naturalized, year of naturalization. | Attended school any time since Sept. 1, 1919. | Whether able to read. | Whether able to write. | Place of birth. | Mother tongue. | Place of birth. | Mother tongue. | Place of b |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| | ✓ | | | Pond, Mary | daughter | | | F | W | 24 | S | | | | | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | ——, Gladys | daughter | | | F | W | 20 | S | | | | yes | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | ——, Genevieve | daughter | | | F | W | 16 | S | | | | yes | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | ——, Asael T. | son | | | M | W | 13 | S | | | | yes | yes | yes | Utah | | Utah | | Utah |
| FM | 208 | 243 | | Karren, George, R. | Head | R | | M | W | 33 | M | | | | | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | ——, Hazel P. | wife | | | F | W | 34 | M | | | | | yes | yes | Idaho | | Utah | | Utah |
| | ✓ | | | ——, Martin T. | son | | | M | W | 4 3/12 | S | | | | | | | Utah | | Utah | | Idaho |
| FM | 209 | 244 | | Talbot, Henry J. | head | O | F | M | W | 57 | M | | | | | yes | yes | Utah | | Africa | | Africa |
| | ✓ | | | ——, Susannah C. | wife | | | F | W | 48 | M | | | | | yes | yes | Nevada | | Illinois | | Ut |
| | ✓ | | | ——, Myron H. | son | | | M | W | 22 | S | | | | | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | ——, Ariel J. | son | | | M | W | 18 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | ——, Rosalie | daughter | | | F | W | 15 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | ——, Wilborn C. | son | | | M | W | 12 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | ——, Genevieve | daughter | | | F | W | 10 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | × | 210 | 245 | Robinson, Elijah | head | R | | M | W | 38 | M | | | | | yes | yes | Tennessee | | Tennessee | | Tenne |
| | | | | ——, Fannie | wife | | | F | W | 31 | M | | | | | yes | yes | Tennessee | | Tennessee | | Tenne |
| | | | | ——, Lester | son | | | M | W | 15 | S | | | | yes | yes | yes | Tennessee | | Tennessee | | Tenne |
| | | | | ——, Lovavie | daughter | | | F | W | 13 | S | | | | yes | yes | yes | Tennessee | | Tennessee | | Tenne |
| | | | | ——, Malcom E. | son | | | M | W | 11 | S | | | | yes | yes | yes | Tennessee | | Tennessee | | Tenne |

# A common scenario...

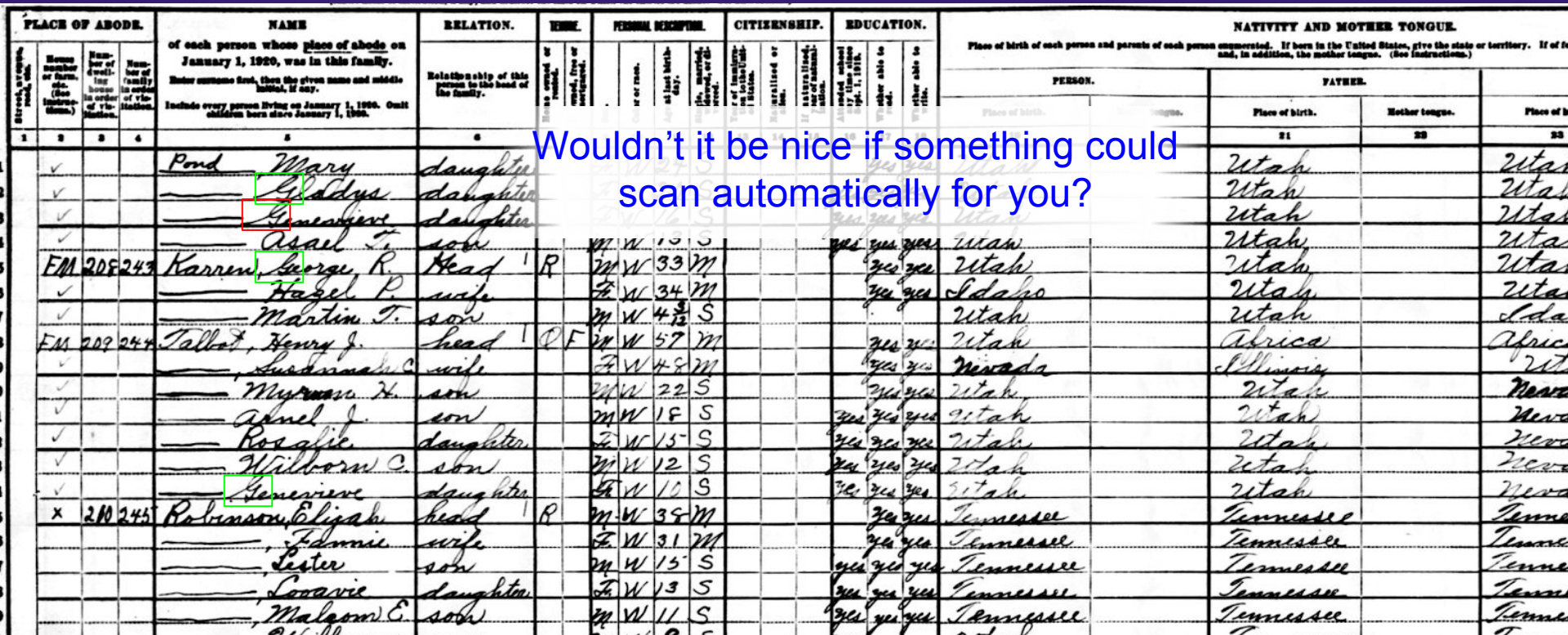


| Street, avenue, road, etc. | House number or farm, etc. (See instructions.) | Number of dwelling house in order of visitation. | Number of family in order of visitation. | NAME of each person whose place of abode on January 1, 1920, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on January 1, 1920. Omit children born since January 1, 1920. | RELATION. Relationship of this person to the head of the family. | HOME. Home owned or rented. | HOME. If owned, free or mortgaged. | PERSONAL DESCRIPTION. Sex. | Color or race. | Age at last birthday. | Single, married, widowed, or divorced. | CITIZENSHIP. Year of immigration to the United States. | Naturalized or alien. | If naturalized, year of naturalization. | EDUCATION. Attended school any time since Sept. 1, 1919. | Whether able to read. | Whether able to write. | NATIVITY AND MOTHER TONGUE. Place of birth of each person and parents of each person enumerated. If born in the United States, give the state or territory. If of foreign birth, the country, and, in addition, the mother tongue. (See instructions.) PERSON. Place of birth. | PERSON. Mother tongue. | FATHER. Place of birth. | FATHER. Mother tongue. | Place of birth |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| | ✓ | | | Pond, Mary | daughter | | | F | W | 24 | S | | | | | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | ——, Gladys | daughter | | | F | W | 20 | S | | | | yes | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | ——, Genevieve | daughter | | | F | W | 16 | S | | | | yes | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | —— Asael T. | son | | | M | W | 13 | S | | | | yes | yes | yes | Utah | | Utah | | Utah |
| FM | 205 | 243 | | Karren, George, R. | Head | R | | M | W | 33 | M | | | | | yes | yes | Utah | | Utah | | Utah |
| | ✓ | | | —— Hazel P. | wife | | | F | W | 34 | M | | | | | yes | yes | Idaho | | Utah | | Utah |
| | ✓ | | | —— Martin T. | son | | | M | W | 4 3/12 | S | | | | | | | Utah | | Utah | | Idaho |
| FM | 207 | 244 | | Talbot, Henry J. | head | O | F | M | W | 57 | M | | | | | yes | yes | Utah | | Africa | | Africa |
| | ✓ | | | ——, Susannah C. | wife | | | F | W | 48 | M | | | | | yes | yes | Nevada | | Illinois | | Utah |
| | ✓ | | | —— Myron H. | son | | | M | W | 22 | S | | | | | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | —— Arnel J. | son | | | M | W | 18 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | —— Rosalie | daughter | | | F | W | 15 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | —— Wilborn C. | son | | | M | W | 12 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | ✓ | | | —— Genevieve | daughter | | | F | W | 10 | S | | | | yes | yes | yes | Utah | | Utah | | Nevada |
| | × | 210 | 245 | Robinson, Elijah | head | R | | M | W | 38 | M | | | | | yes | yes | Tennessee | | Tennessee | | Tennessee |
| | | | | ——, Minnie | wife | | | F | W | 31 | M | | | | | yes | yes | Tennessee | | Tennessee | | Tennessee |
| | | | | ——, Lester | son | | | M | W | 15 | S | | | | yes | yes | yes | Tennessee | | Tennessee | | Tennessee |
| | | | | ——, Lovavie | daughter | | | F | W | 13 | S | | | | yes | yes | yes | Tennessee | | Tennessee | | Tennessee |
| | | | | ——, Malcom E. | son | | | M | W | 11 | S | | | | yes | yes | yes | Tennessee | | Tennessee | | Tennessee |

# A common scenario...



# A common scenario...



Wouldn't it be nice if something could scan automatically for you?

We’re going to do this and other things with *subword spotting*

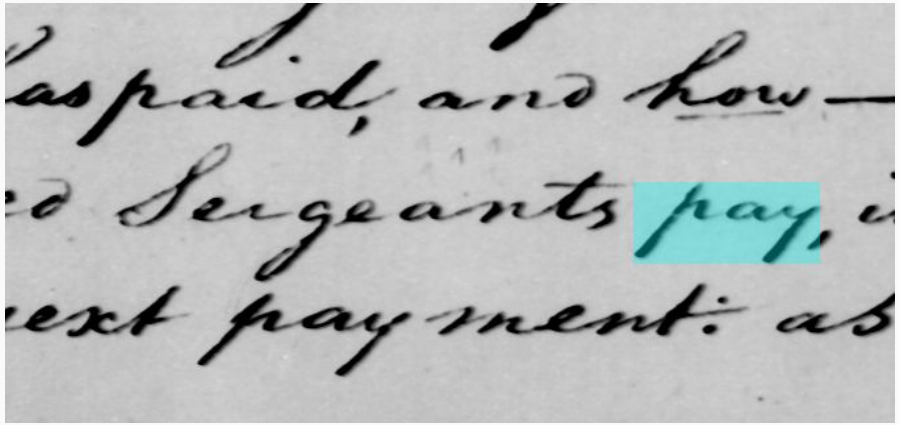
# Outline

- Review word spotting
- Subword spotting
  - Our implementation
  - Performance
- Applications
  - Suffix spotting
  - Transcription assistant demo

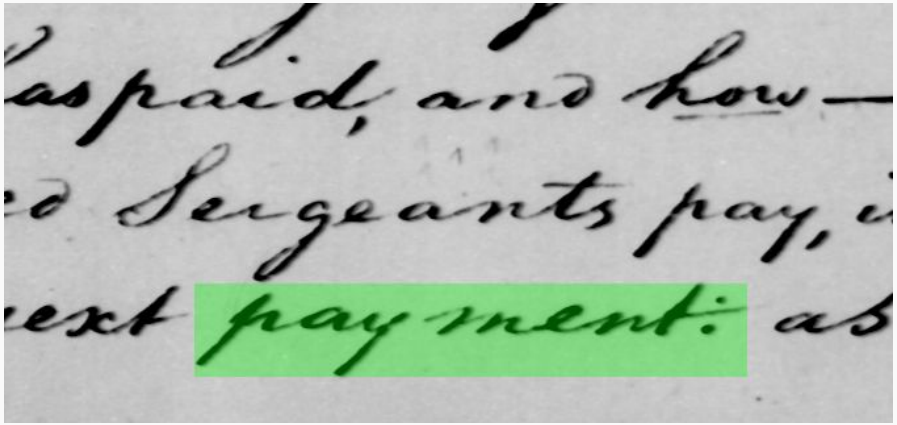
# Word Spotting

- Goal is to search corpus of images directly
- Query-by-string (QbS): search with text
- Query-by-example (QbE): search with an example word image

Search for “**pay**”



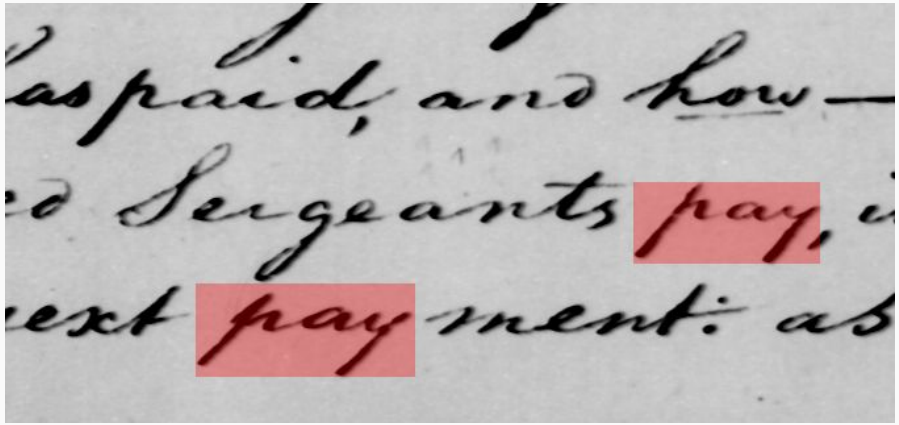
Search for “**payment**”



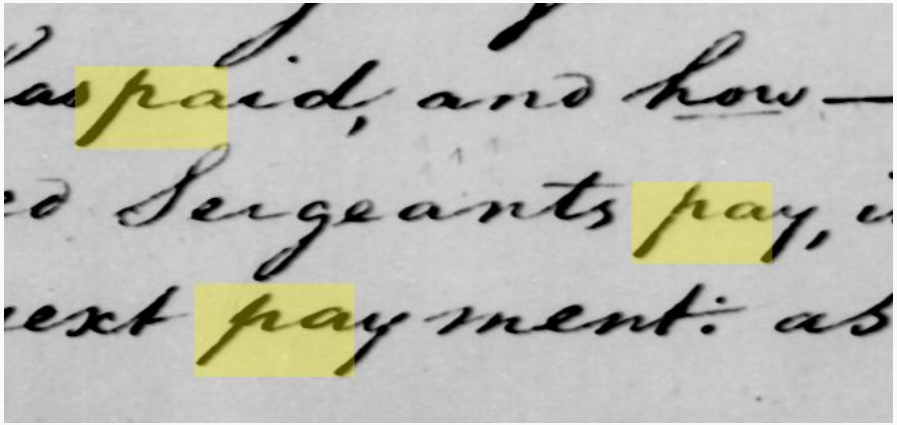
# Subword Spotting

- We now allow spottings within words

Search for “**pay**”

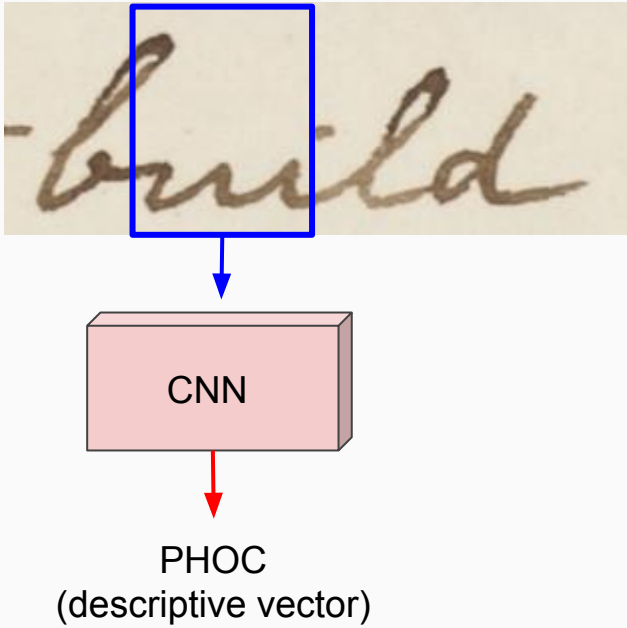


Search for “**pa**”



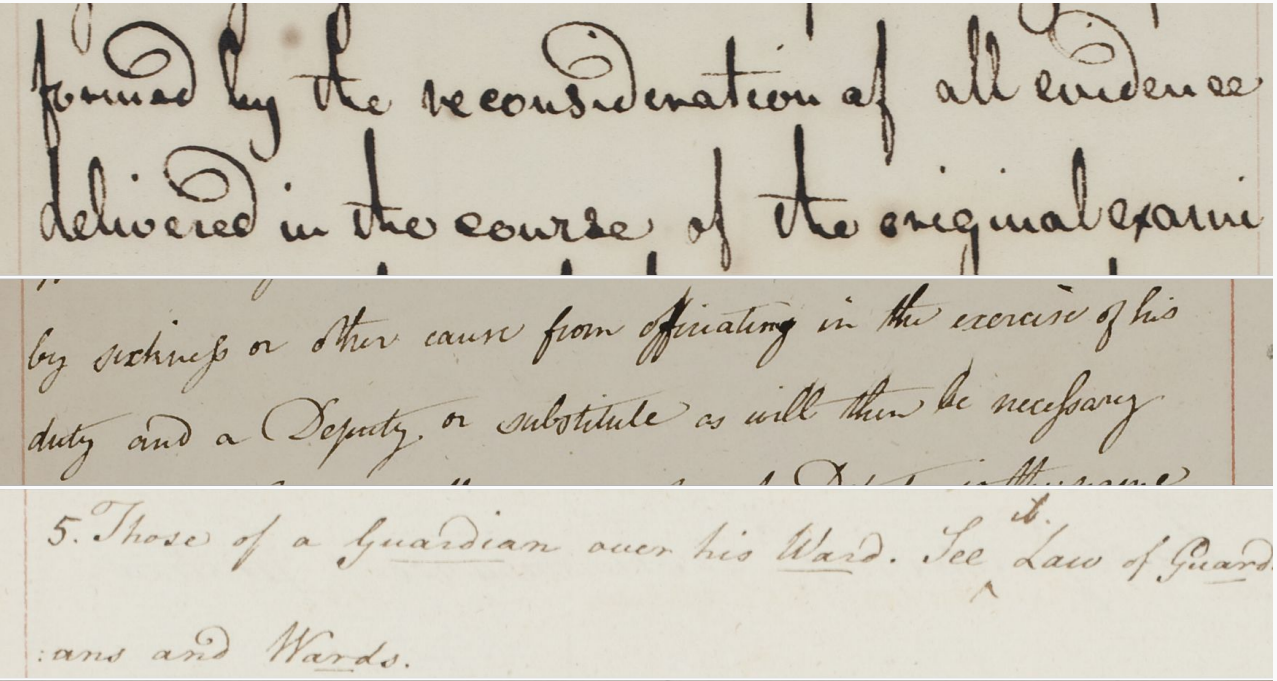
# Subword Spotting Implementation

- Converted Sudholt et al's word spotting method PHOCNet to perform sliding window
- Changed PHOC used and comparison method for better QbS results
  - Less resolution for PHOC
  - Similarity based on cross-entropy instead cosine distance
  - Original PHOC and cosine similarity better for QbE
- Found optimal window width for each subword of interest

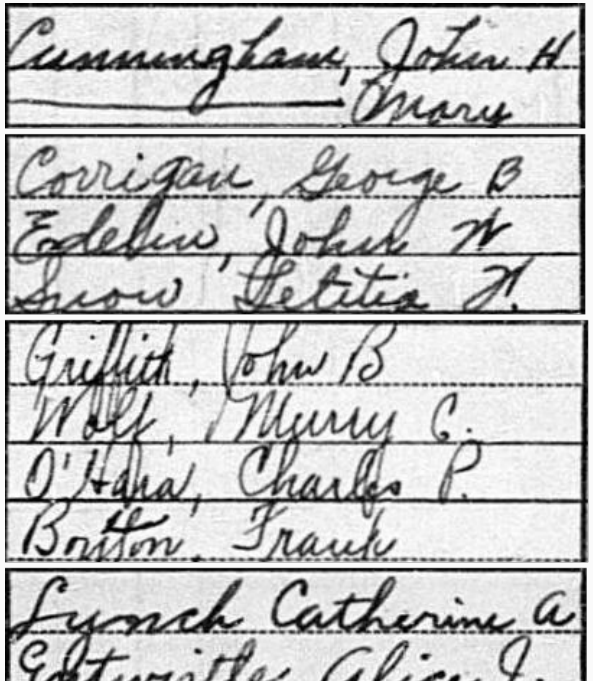


# Datasets

Bentham




US 1930 Census Names



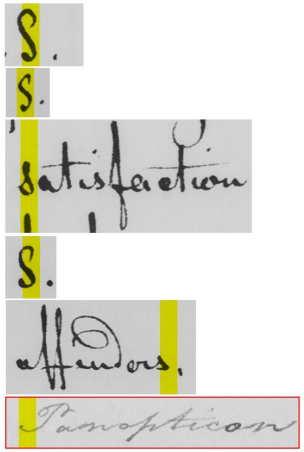
# Subord Spotting Results

- Unigrams: all letters of alphabet
- Bigrams: 100 most frequent in English
- Trigrams: 300 most frequent in English

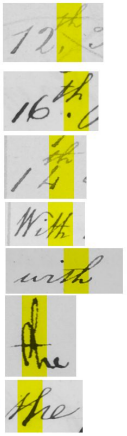
Reported as Mean Average Precision

| | Bentham | | | US 1930 Census Names | | |
|---|---|---|---|---|---|---|
| | Unigrams | Bigrams | Trigrams | Unigrams | Bigrams | Trigrams |
| QbS | 67.7% | 68.2% | 70.5% | 49.7% | 40.2% | 36.3% |
| QbE | 51.1% | 56.9% | 57.1% | 34.0% | 29.5% | 28.5% |

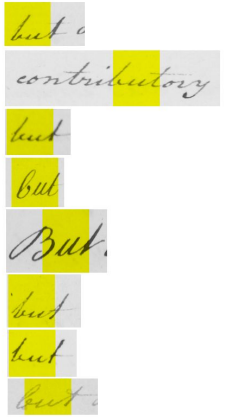
s



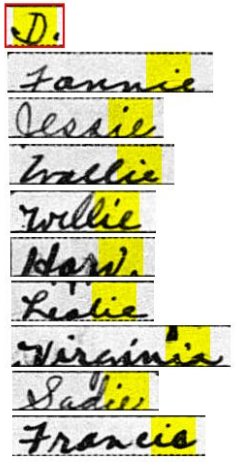
th



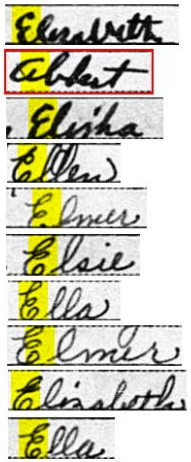
but



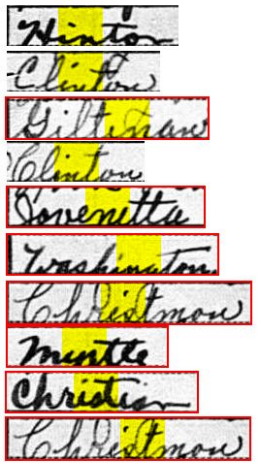
i



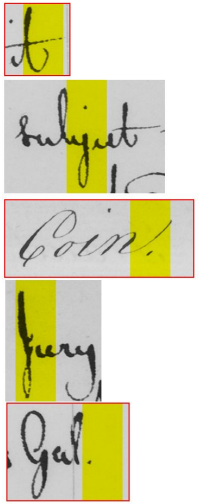
el



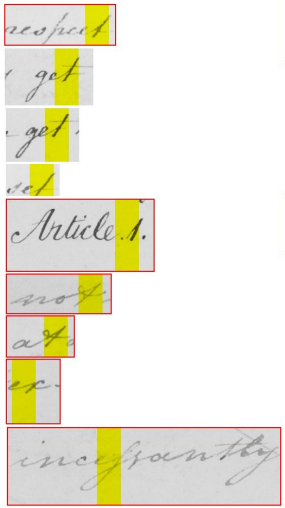
int



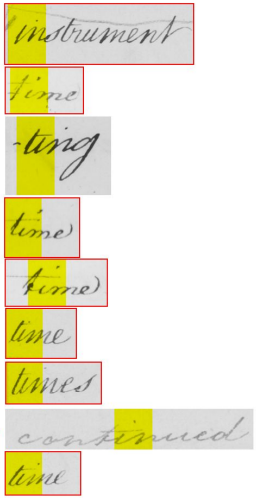
j



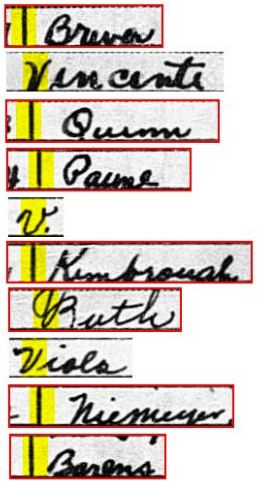
et



tin

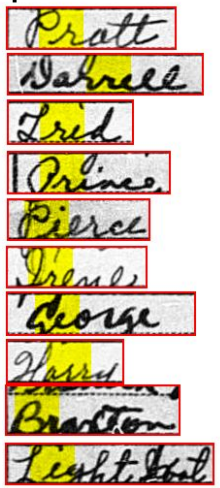


v



ts

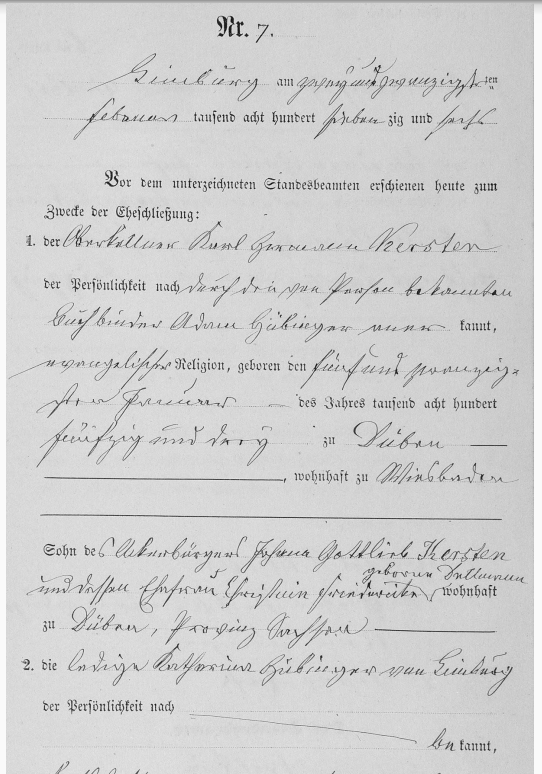
pre



# A searching task

What if I wanted to find all the towns in a set of German documents?

- What about automatically finding all words ending in “-burg”?



Nr. 7.

Vor dem unterzeichneten Standesbeamten erschienen heute zum Zwecke der Eheschließung:

1. der

der Persönlichkeit nach

kannt,

Religion, geboren den

des Jahres tausend acht hundert

zu

wohnhaft zu

Sohn des

geborene

wohnhaft zu

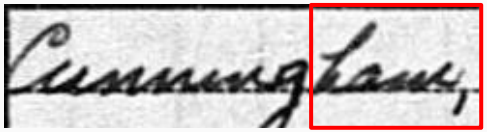
2. die

der Persönlichkeit nach

kannt,

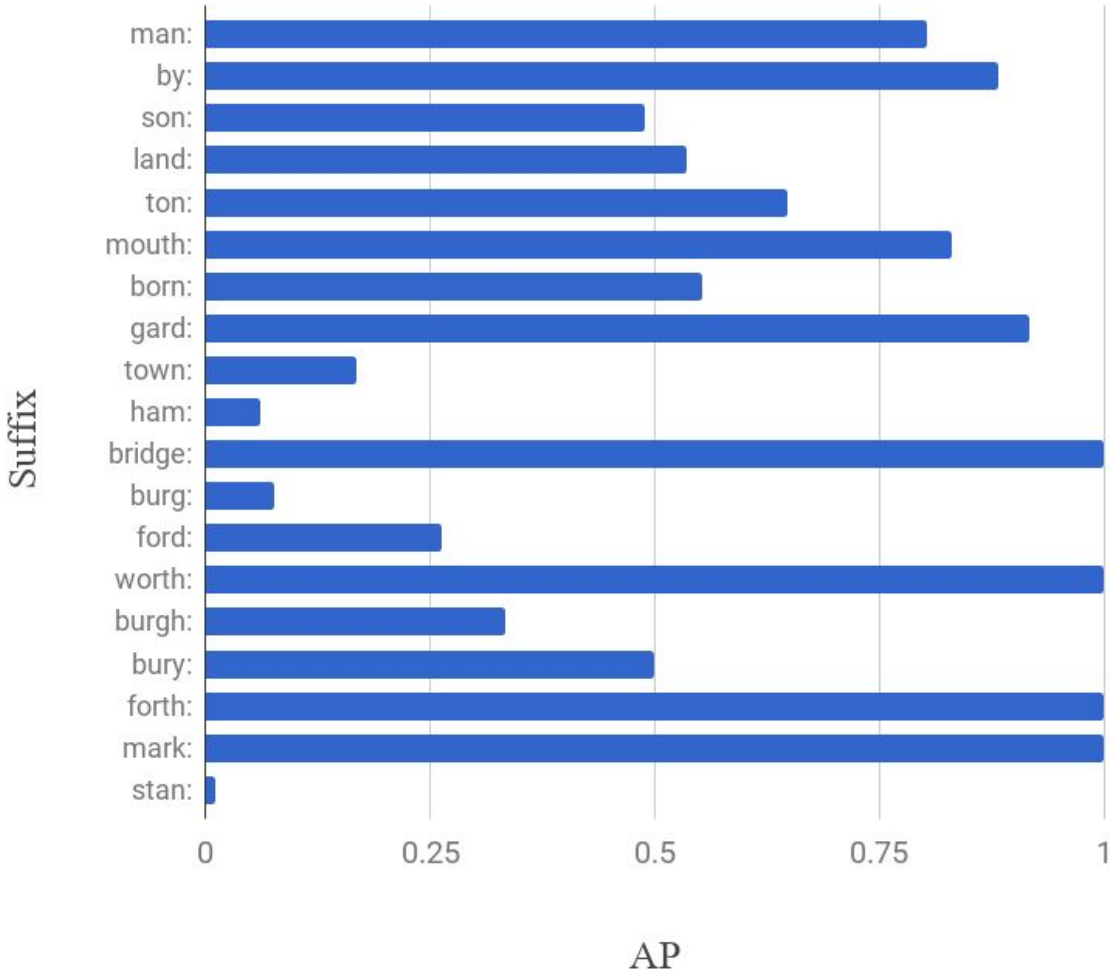
# Suffix Spotting

- Find all words with a given suffix
- Constraint on original subword spotting problem
- Could be extended to handle regular-expression-like queries



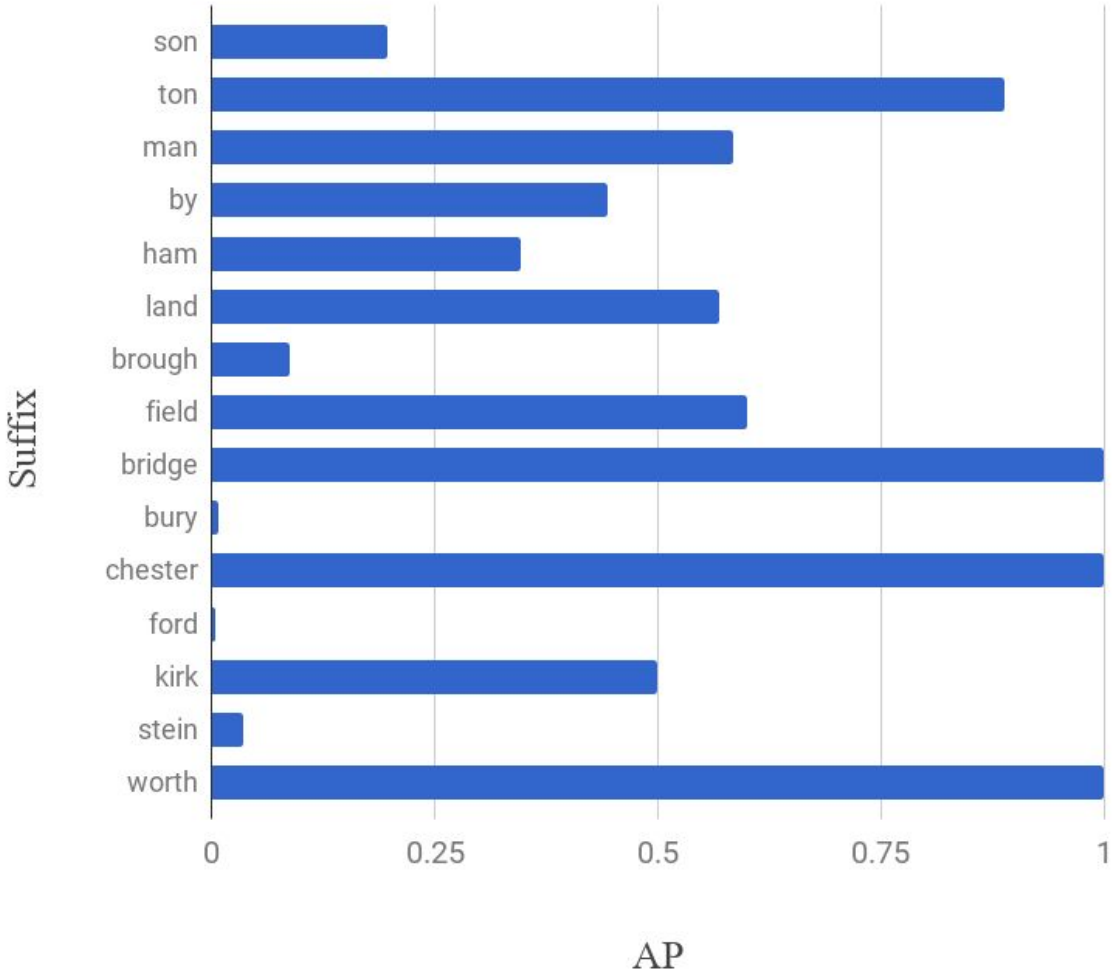
# Suffix Spotting

- IAM results



# Suffix Spotting

- Census Names results



# Transcription Assistant Demo

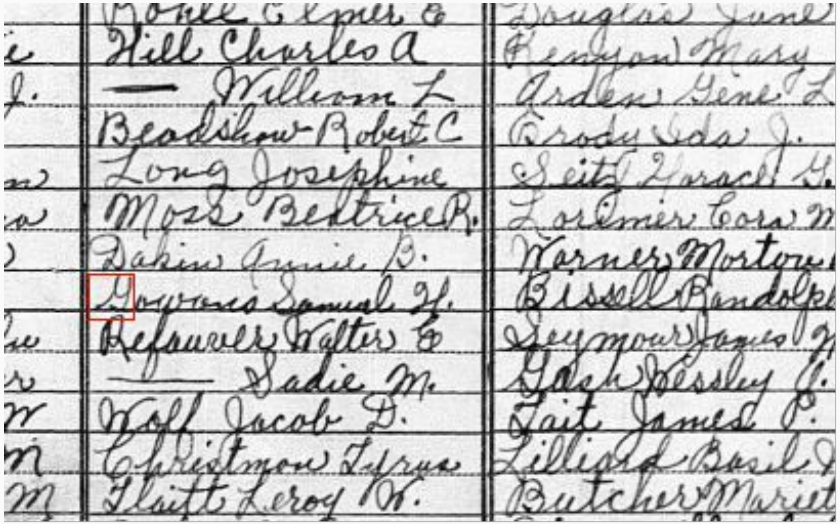
- Using ground truth word segmentations
- Both embedding and PHOCs for windows are precomputed.
  - Selection snapped to closest window
  - ~10 second delay to compute PHOC for all windows of a single size, depending on size, using GPU

# Thank you

Questions?

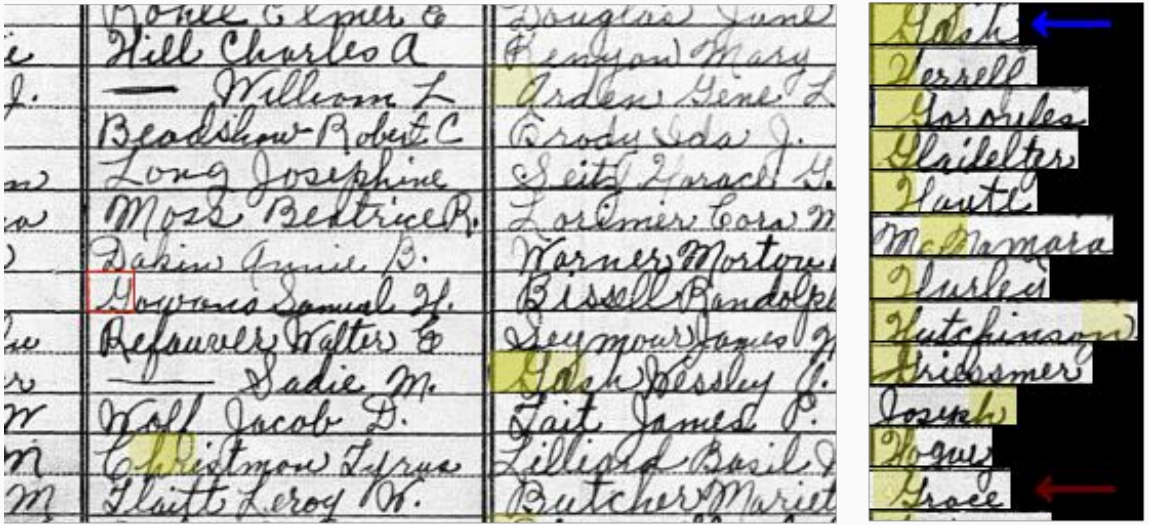
# Transcription Assistant Demo

(Images in case of technical difficulties)



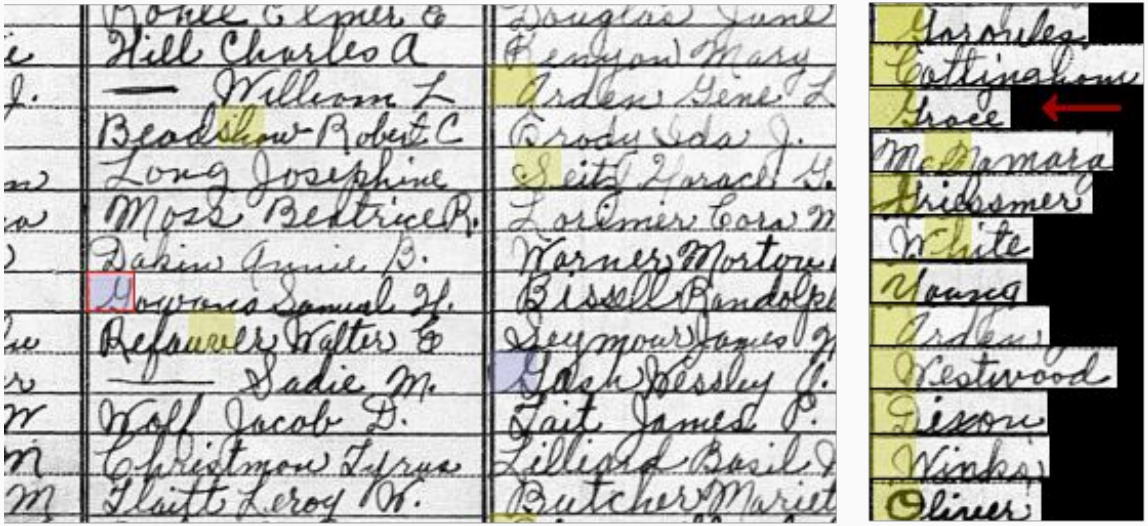
# Transcription Assistant Demo

(Images in case of technical difficulties)



# Transcription Assistant Demo

(Images in case of technical difficulties)



# Subord Spotting Implementation

- Network architecture