Combining Humans and AI to Link Historical Records

Joseph Price
Brigham Young University
NBER
IZA

Record Linking LAB

Record Linking LAB

rll.byu.edu

Three things we do:

- Computer vision
- Natural language processing
- Record linking

<u>Our goal:</u>

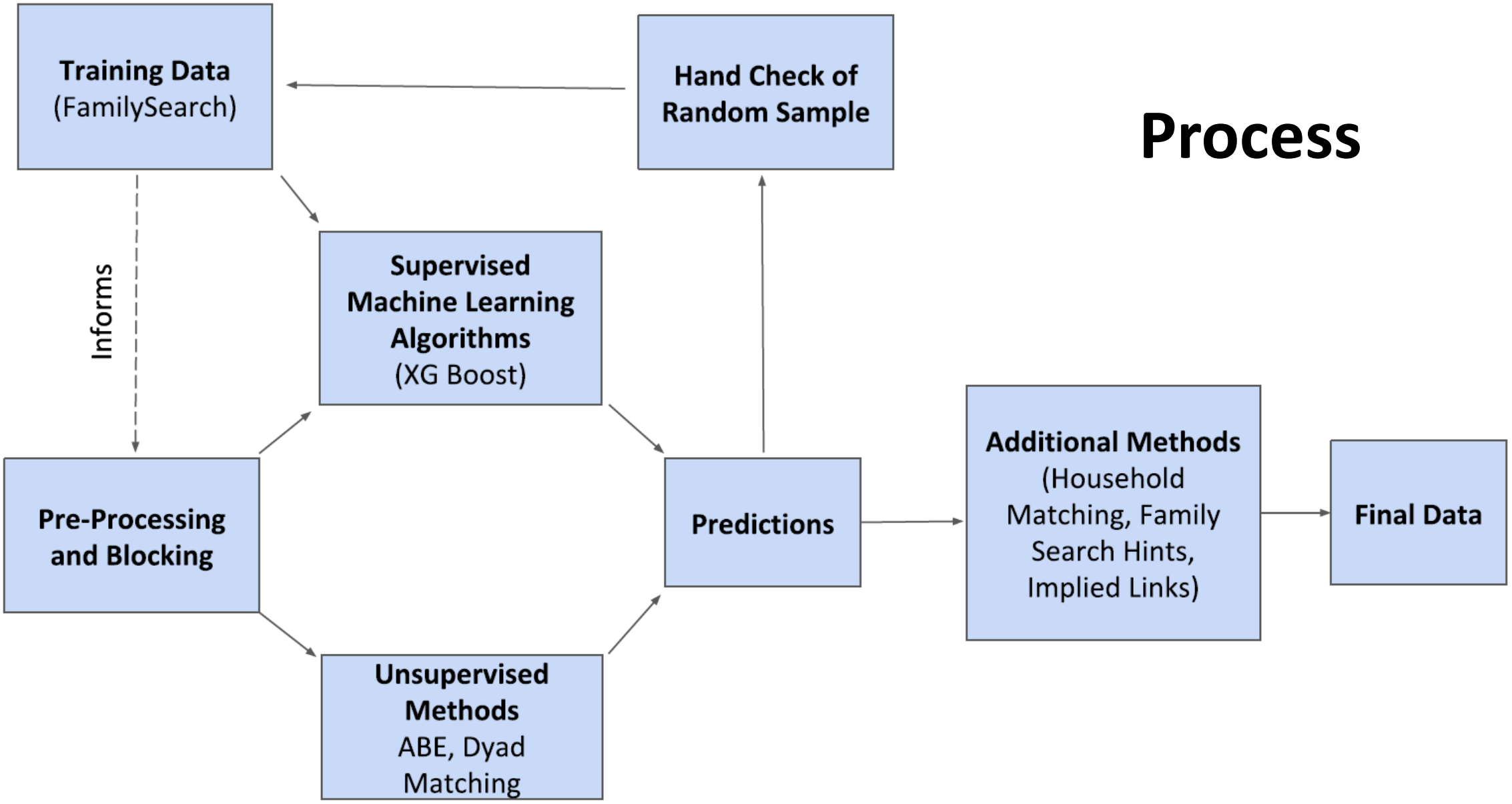
- Convert any historic records into linkable data that can be used for research.

Record Linking Goal: Census Tree

- Link everyone that lived in the US between 1850-1940 across each of the census records they appear in.
- Linked to their parents, siblings, spouse, and children.
- Linked to vital records, school records, draft cards, etc.

	US Population	New additions
1850	23.2	23.2
1860	31.4	10.9
1870	38.6	12.9
1880	50.2	16.4
1900	76.2	41.4
1910	92.2	27.5
1920	106.5	28.8
1930	123.1	29.3
1940	132.1	26.8
Total	673.5	217.2

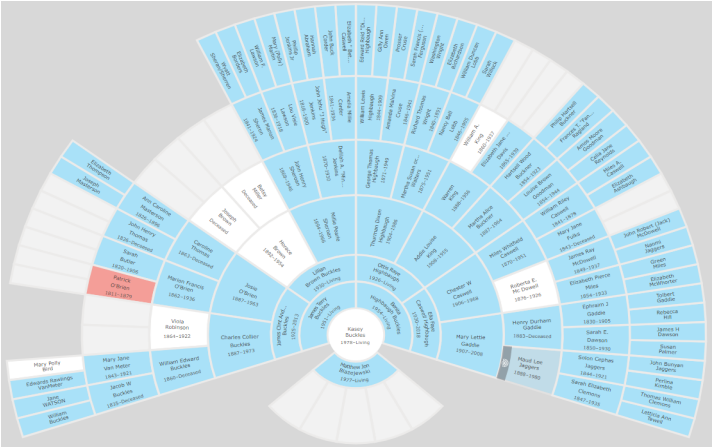
Process

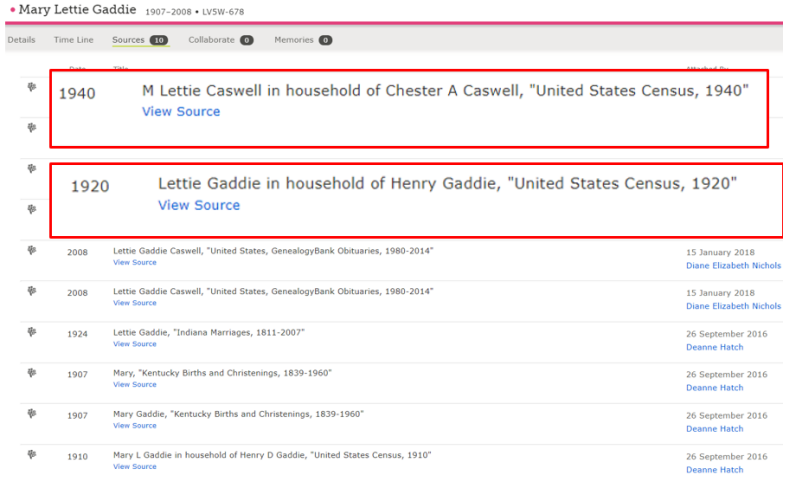

Training Data
(FamilySearch)
Hand Check of
Random Sample
Informs
Supervised
Machine Learning
Algorithms
(XG Boost)
Pre-Processing
and Blocking
Predictions
Additional Methods
(Household
Matching, Family
Search Hints,
Implied Links)
Final Data
Unsupervised
Methods
ABE, Dyad
Matching

FamilySearch Family Tree as Training Data

10.4M pairs among 1900, 1910, & 1920 Censuses

	Women	Men
Only 1900 & 1910	1,275,583	1,356,810
Only 1910 & 1920	1,433,637	1,557,970
Only 1900 & 1920	442,814	536,256
1900 & 1910 & 1920	905,095	1,003,087





- Kaplanis et al. (*Science,* 2018) confirm linkages on a similar site using DNA data, “demonstrate that millions of genealogists can collaborate in order to produce high quality population-scale family trees.”
- Bailey et al. (2017), ABEFP (2019): Genealogy links are the “gold standard.” These papers have used FamilySearch’s user-generated links as benchmark for assessing linking methods.

Table 5. Contributions of the Different Methods Used to Link Records

	1900-1910		1910-1920	
Method	Matches	New Matches	Matches	New Matches
Family Tree data	4,567,392	4,567,392	4,912,838	4,912,838
ABE	17,093,182	14,665,647	21,811,611	19,064,068
XGBoost	26,855,325	14,845,508	29,063,701	19,028,103
Household matching	13,656,233	2,364,018	6,710,830	2,593,163
Dyad matching	12,846,431	548,382	25,085,386	4,214,342
ABE (second time)	961,544	656,070	3,424,294	1,111,118
FamilySearch hints	23,527,806	3,712,348	30,578,568	2,036,346
Implied from 1900-1920	3,087,749	1,335,657	2,579,389	1,847,714
Total		42,695,022		54,807,692
Match rate		67.8%		71.4%

Ways to get closer to matching everyone

Six ways that we can still improve:

[1] Improve the quality of the transcription of the census records.

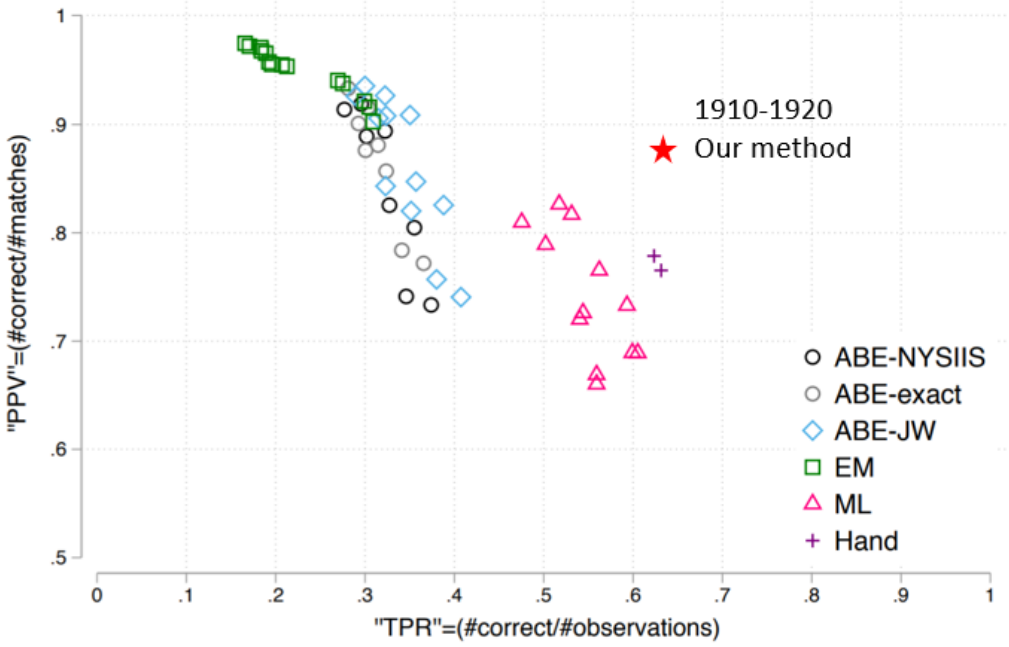
[2] Use census sheet links to identify possible matches and improve precision.

[3] Create machine learning models for specific groups (African American, Germans, etc.).

[4] Involve humans in helping with the unmatched cases (record hints).

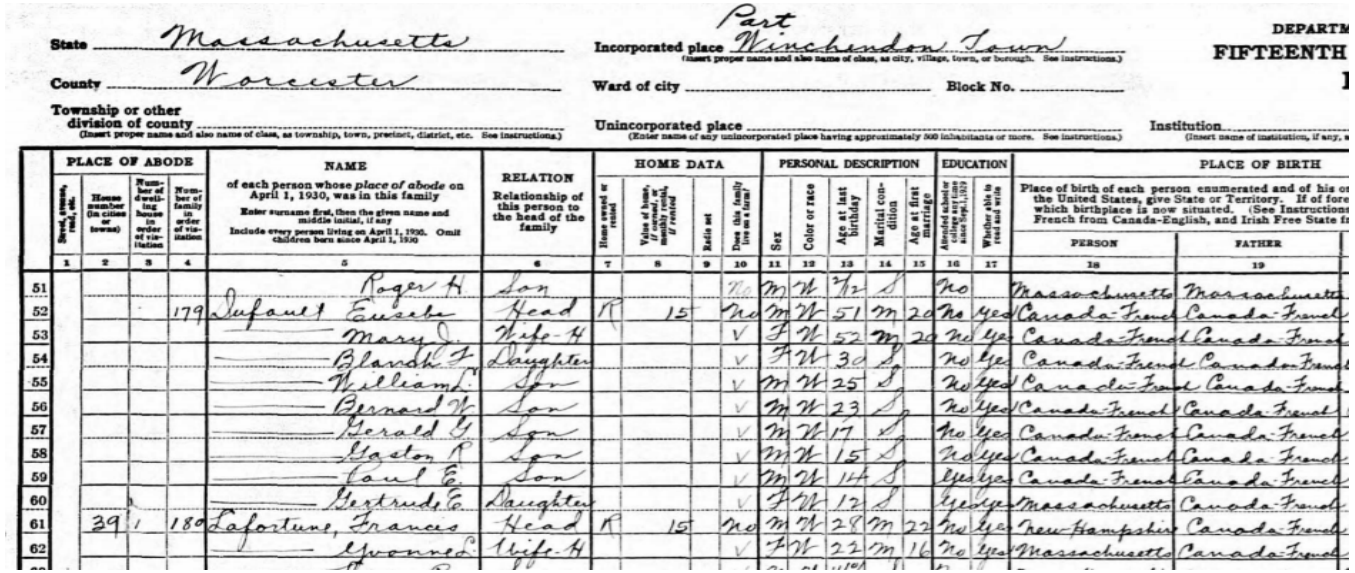
[5] Link to records provide the maiden name of women (BMD).

[6] Use new distance metrics for comparing names and places.



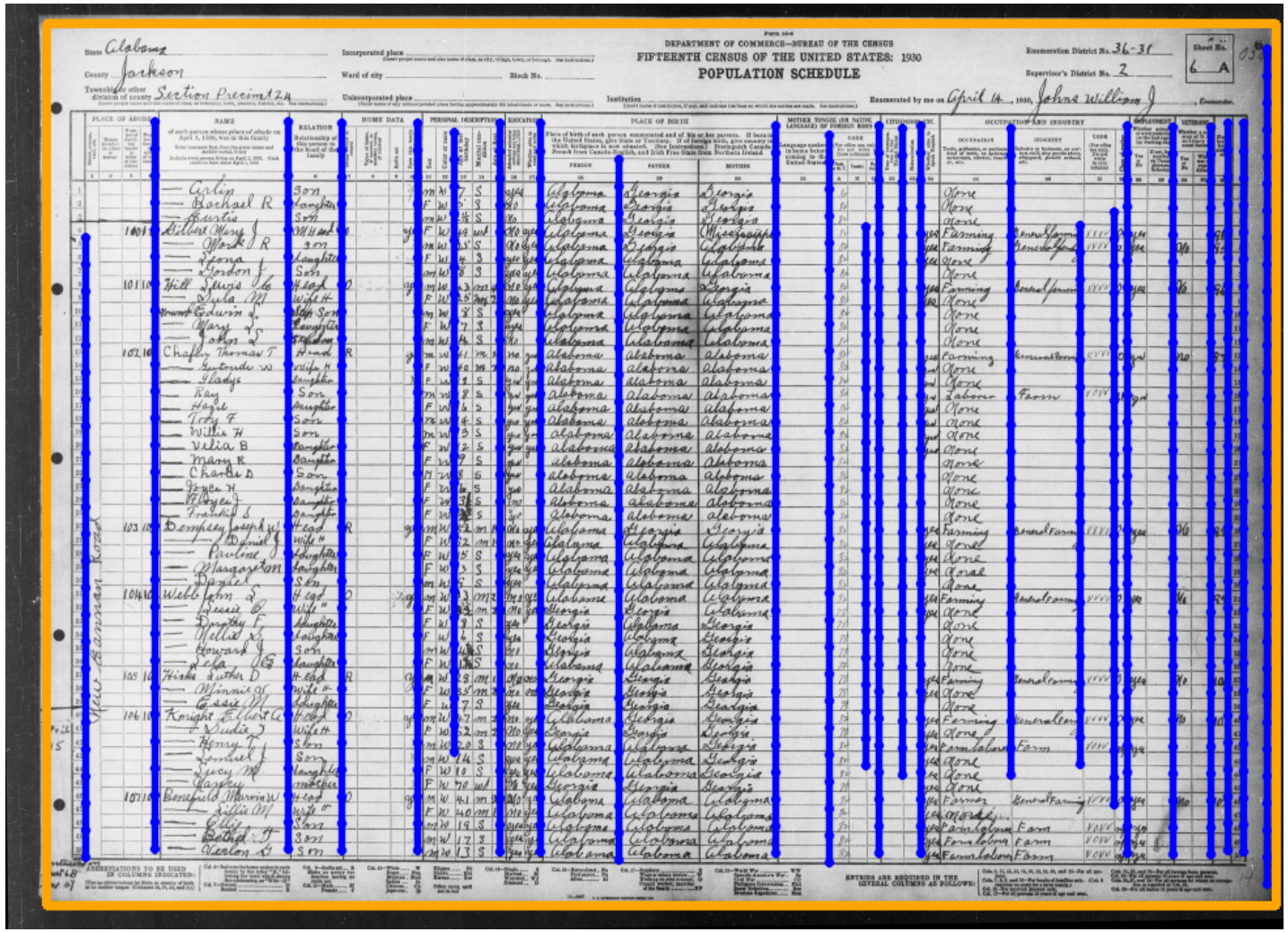
Auto Indexing Historical Records

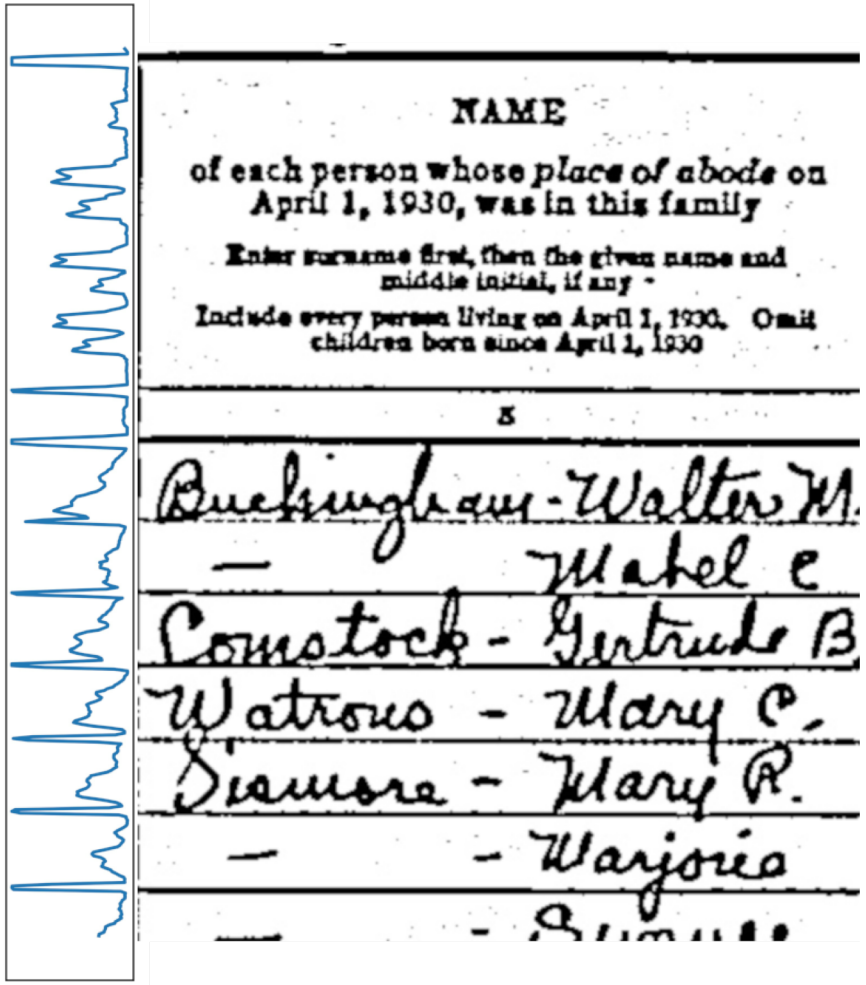
- Five-Step Process:
  - Line segmenting
  - Linking to labeled data
  - Lexicon
  - Hand-writing recognition
  - Labeling by trainers


State: Massachusetts
County: Worcester
Township or other division of county
Incorporated place: Part Winchendon Town
Ward of city
Block No.
Unincorporated place
Institution
DEPARTM
FIFTEENTH

	PLACE OF ABODE				NAME	RELATION	HOME DATA				PERSONAL DESCRIPTION					EDUCATION		PLACE OF BIRTH	
	Street, avenue, road, etc.	House number (in cities or towns)	Number of dwelling house in order of visitation	Number of family in order of visitation	of each person whose place of abode on April 1, 1930, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on April 1, 1930. Omit children born since April 1, 1930	Relationship of this person to the head of the family	Home owned or rented	Value of home, if owned, or monthly rental, if rented	Radio set	Does this family live on a farm?	Sex	Color or race	Age at last birthday	Marital condition	Age at first marriage	Attended school or college at any time since Sept. 1, 1929	Whether able to read and write	Place of birth of each person enumerated and of his or the United States, give State or Territory. If of foreign which birthplace is now situated. (See Instructions) French from Canada-English, and Irish Free State f	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18 PERSON	19 FATHER
51					Roger H.	Son				No	M	W	7/12	S		No		Massachusetts	Massachusetts
52				179	Dufault Eusebe	Head	R	15		No	M	W	51	M	20	No	yes	Canada-French	Canada-French
53					Mary J.	Wife-H				V	F	W	52	M	20	No	yes	Canada-French	Canada-French
54					Blanch F.	Daughter				V	F	W	30	S		No	yes	Canada-French	Canada-French
55					William J.	Son				V	M	W	25	S		No	yes	Canada-French	Canada-French
56					Bernard N.	Son				V	M	W	23	S		No	yes	Canada-French	Canada-French
57					Gerald G.	Son				V	M	W	17	S		No	yes	Canada-French	Canada-French
58					Gaston R.	Son				V	M	W	15	S		No	yes	Canada-French	Canada-French
59					Paul E.	Son				V	M	W	14	S		yes	yes	Canada-French	Canada-French
60					Gertrude E.	Daughter				V	F	W	12	S		yes	yes	Massachusetts	Canada-French
61		39	1	180	Lafortune, Frances	Head	R	15		No	M	W	28	M	22	No	yes	New Hampshire	Canada-French
62					Yvonne C.	Wife-H				V	F	W	22	M	16	No	yes	Massachusetts	Canada-French

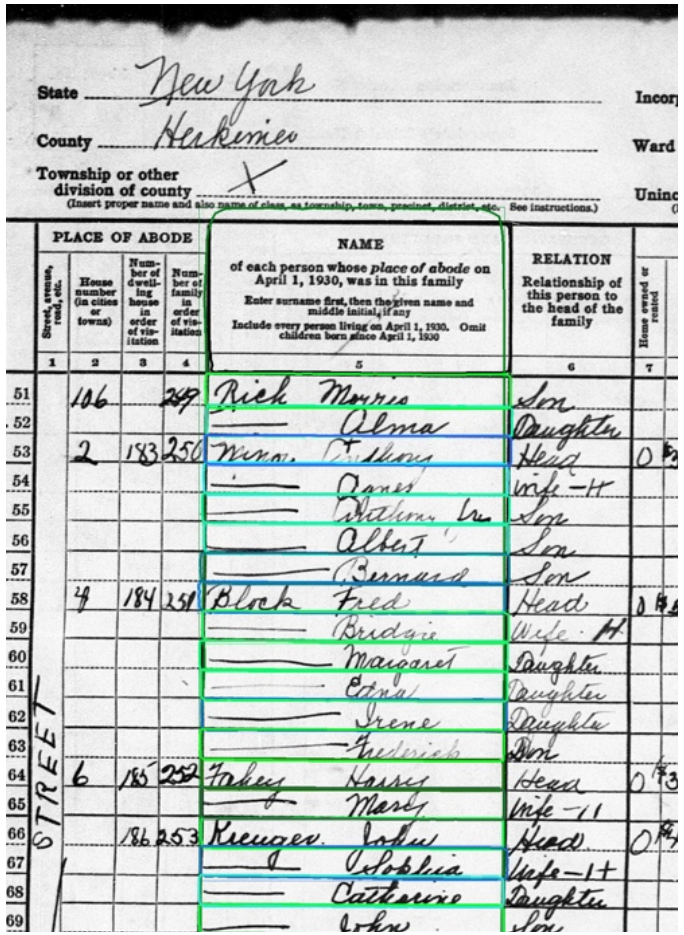
Step #1: Line segmenting

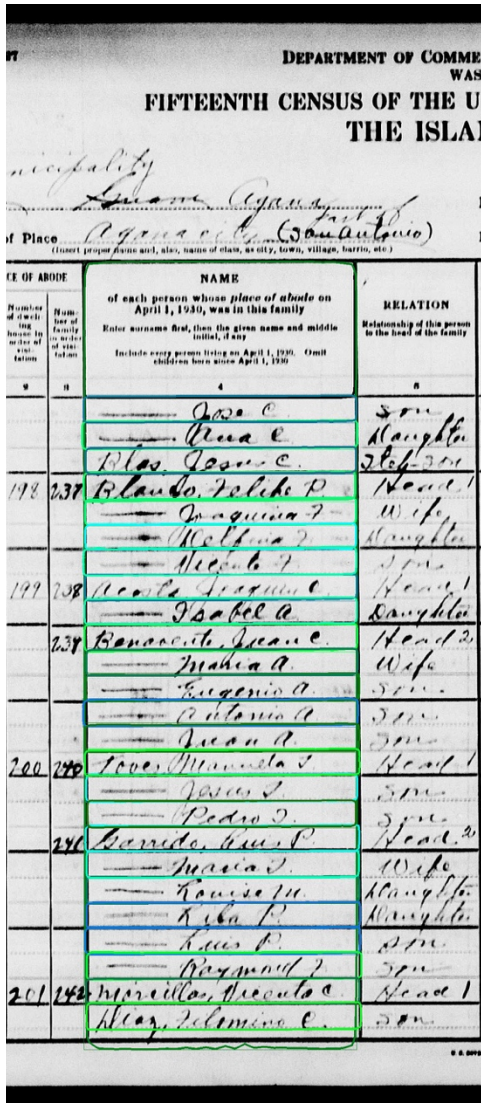




Step #2: Linking to labeled data

- Each little box becomes an image that we can label using the human transcription of the census.
- 1930 census alone provides a training set with 1.2 billion labeled images (10 fields X 122M people)



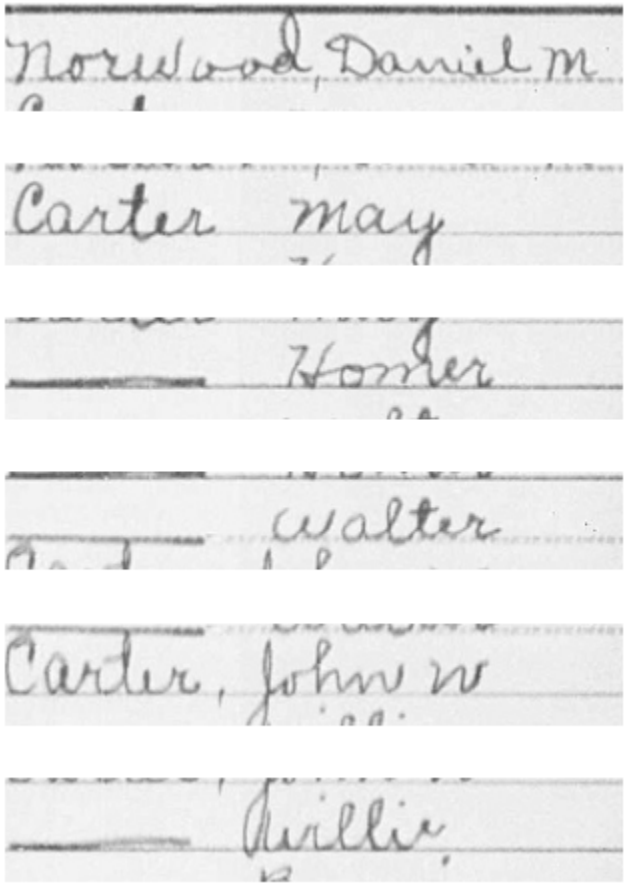


Step #3: Surname lexicon

pr_name_surn	count1900	count1910	count1920	count1930	count1940	count_bill~n	count_numi~t
Altdorfer	3	7	4	6	9	0	2
Altdredge	0	0	0	0	4	0	0
Altdridge	0	4	0	0	0	0	0
Altdringer	1	0	0	0	0	0	0
Altds	0	3	0	0	0	0	0
Altduil	5	0	0	0	0	0	0
Alte	19	33	36	56	58	1	22
Alte Kruse	0	0	2	0	0	0	0
Alte Mueller	0	4	0	0	0	0	0

- There are about 10 million unique strings in the surname field across the 1900-1940 census. Of these, only 500k are correctly spelled.
- We can use a machine learning model to flag the incorrect ones and feed those through our HWR process.

Step #4: Hand-writing recognition



Norwood, Daniel M

Carter, May

---, Homer

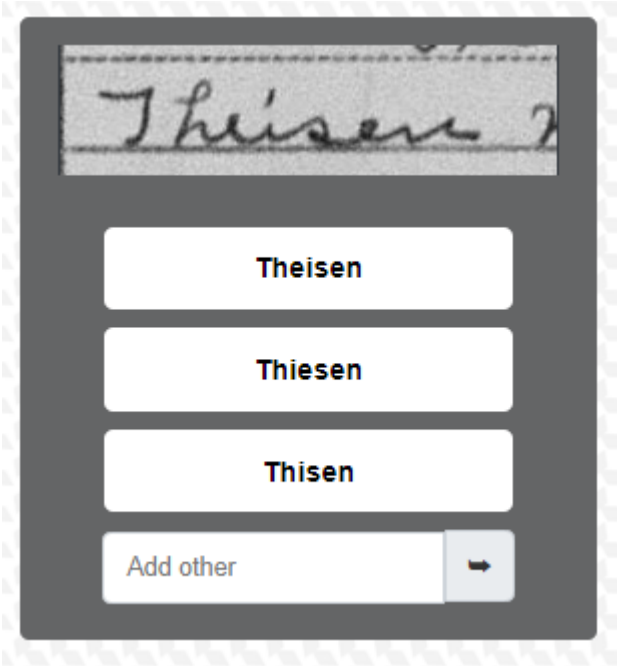
---, Walter

Carter, John W

---, Willie

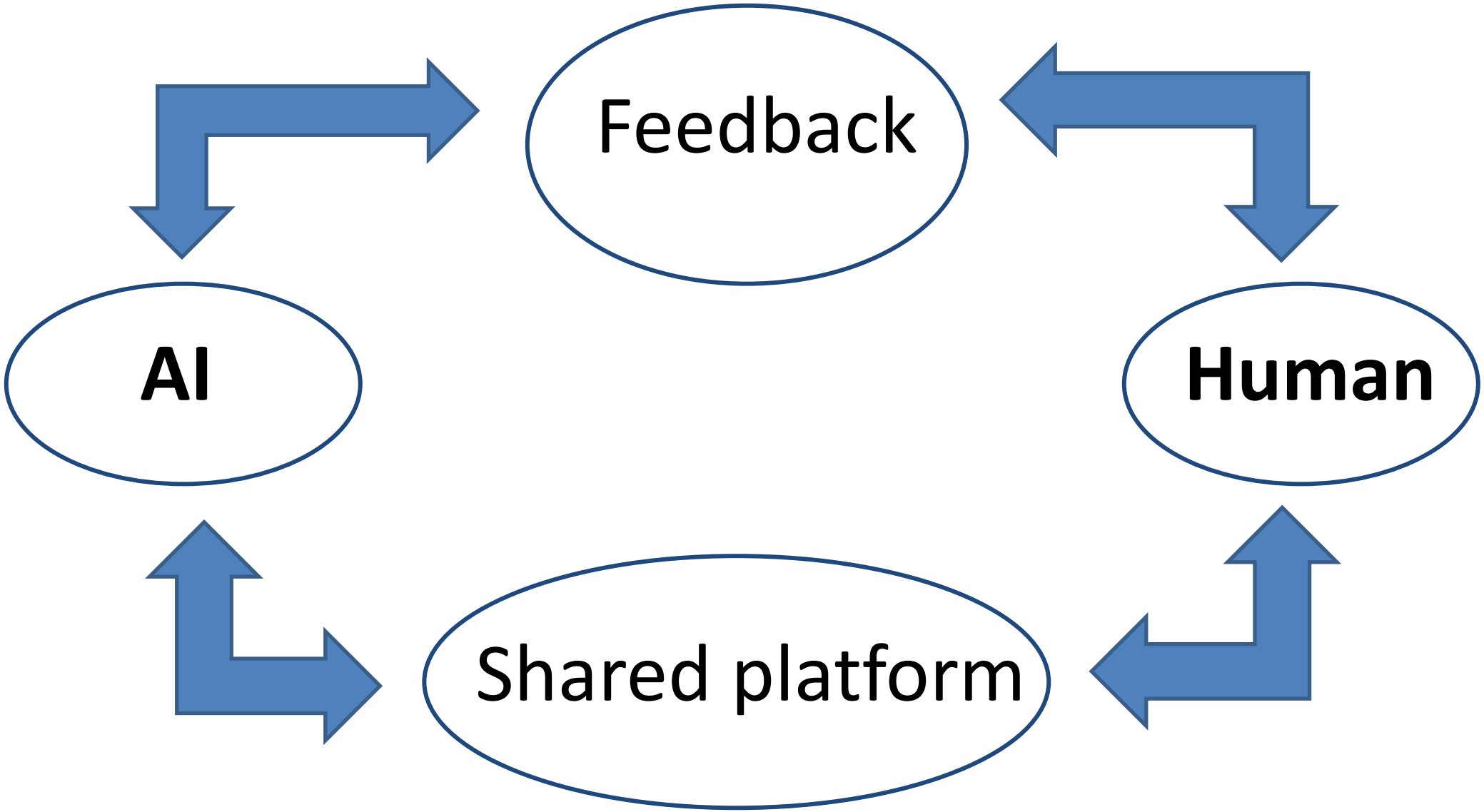
- This will provide multiple possibilities for what the surname should be: the original index, the HWR output, the closest surnames in the correct surname lexicon.
- You can try to match on any of these to improve your recall.

Step #5: Labeling by humans



indexing.fhtl.byu.edu

bit.ly/rll-index


Feedback
AI
Human
Shared platform

African American Pilot Study

- There is a huge gap in White and African American coverage rates on FamilySearch.
- Over 70% of White Americans from the 1900 Census have FamilySearch profiles, but only 4% of African Americans
- Low coverage rates makes genealogy work difficult, especially for new converts.

Challenges for African American Genealogy

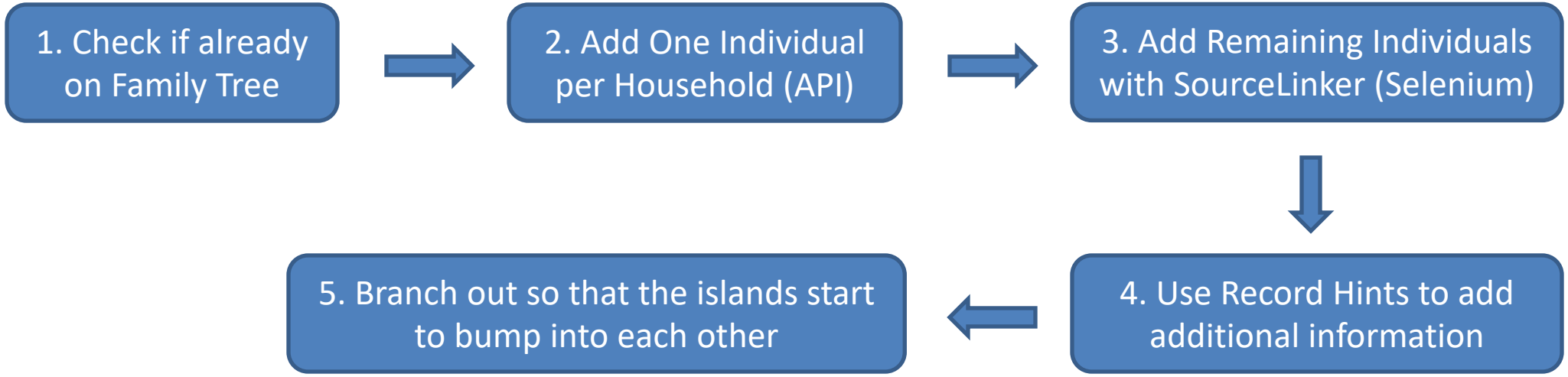
Many factors make working with African American records more difficult, including:

- Erroneous and incomplete records
- Unknown or uncertain birth dates
- May be missed by census takers

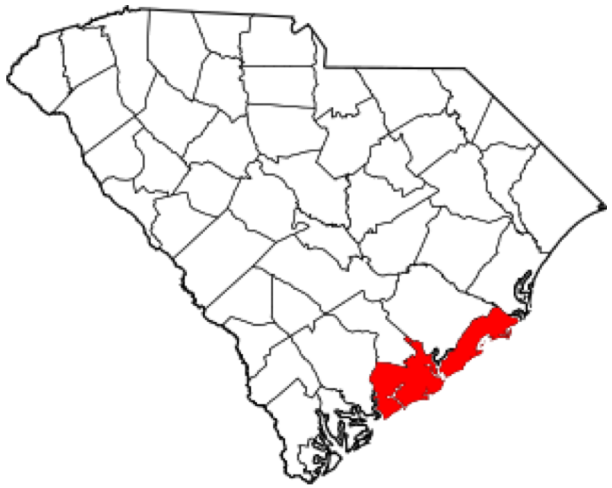
Pilot Process to Increase Coverage

Selected 14 Counties with large African American Populations:

Goal: Add every African American family from 1900 census to FamilySearch



Progress



Charleston County, South Carolina

African American Population (1900): 46,671*

Date	Sept 12, 2019	Oct 3, 2019	Feb 3, 2020
Number on Tree	1,666	44,808	46,138
Percent on Tree	3.57%	96.01%	98.86%

*Only counts individuals in nuclear families

State	County	In Tree (Approx. 9/1/2019)	Number on Tree (2/1/2020)	Total on Census	Percent Coverage
South Carolina	Colleton	1,078	19,478	22,458	86.73%
South Carolina	Beaufort	1,545	27,195	32,181	84.51%
Alabama	Dallas	2,203	38,351	45,903	83.55%
Alabama	Lowndes	1,494	25,174	31,116	80.90%
Mississippi	Washington	2,126	34,346	44,299	77.53%
Mississippi	Bolivar	1,509	24,266	31,435	77.19%
Alabama	Montgomery	2,508	40,321	52,247	77.17%
South Carolina	Charleston	2,898	46,138	60,373	76.42%
South Carolina	Orangeburg	1,994	31,264	41,538	75.27%
Louisiana	Tensas	857	12,775	17,857	71.54%
Georgia	Chatham	1,975	27,670	41,143	67.25%
Tennessee	Davidson	2,086	28,173	43,450	64.84%
Kentucky	Jefferson	2,027	25,146	42,237	59.54%
Tennessee	Shelby	4,042	29,665	84,211	35.23%
14 Counties Southern States	**Total**	**28,342**	→ **409,962**	**590,448**	**72.69%**

Czech Republic Project

- Auto-index the historical censuses.
- Link them together into a Census Tree
- Reconstruct the historical population on the Family Tree
- Invite everyone to use it

Combining Humans and AI to Link Historical Records

Joseph Price
Brigham Young University
NBER
IZA

Record Linking LAB