Linking Together the Entire US Population

Joe Price
joe_price@byu.edu
rll.byu.edu

Two Audacious goals for 2018

- [1] Handwriting recognition + NLP
  - Convert records in archives, libraries, churches and courthouses into usable data.
- [2] Link together the US censuses between 1850-1940
  - Create a shadow tree for the 217 million people who lived in the US during this time period.

Two Audacious goals for 2018

- [1] Handwriting recognition + NLP
  - Convert the records in archives, libraries, churches and courthouses into usable data.
- [2] **Link together the US censuses between 1850-1940**
  - Create a shadow tree for the 217 million people who lived in the US during this time period.

- Family History + Machine Learning
- Family Tree
- Academic Partnerships
- Microtasks

Linking the US Population

	US Population	New additions	New additions (<1906)
1850	23.2	23.2	23.2
1860	31.4	10.9	10.9
1870	38.6	12.9	12.9
1880	50.2	16.4	16.4
1900	76.2	41.4	41.4
1910	92.2	27.5	18.9
1920	106.5	28.8	2.9
1930	123.1	29.3	1.8
1940	132.1	26.8	0.6
Total	673.5	217.2	129.0

- Value of a census-based tree:
  - Identify possible duplicates. Everyone has a place but just one place. Identify twins and siblings with similar name and age.
  - Record-to-record linking of new data collections (many-to-one matches that can be hand-checked).
  - Catch mistakes in the census records (enumerator and indexer).
  - Better experience for new users (particularly using phones)

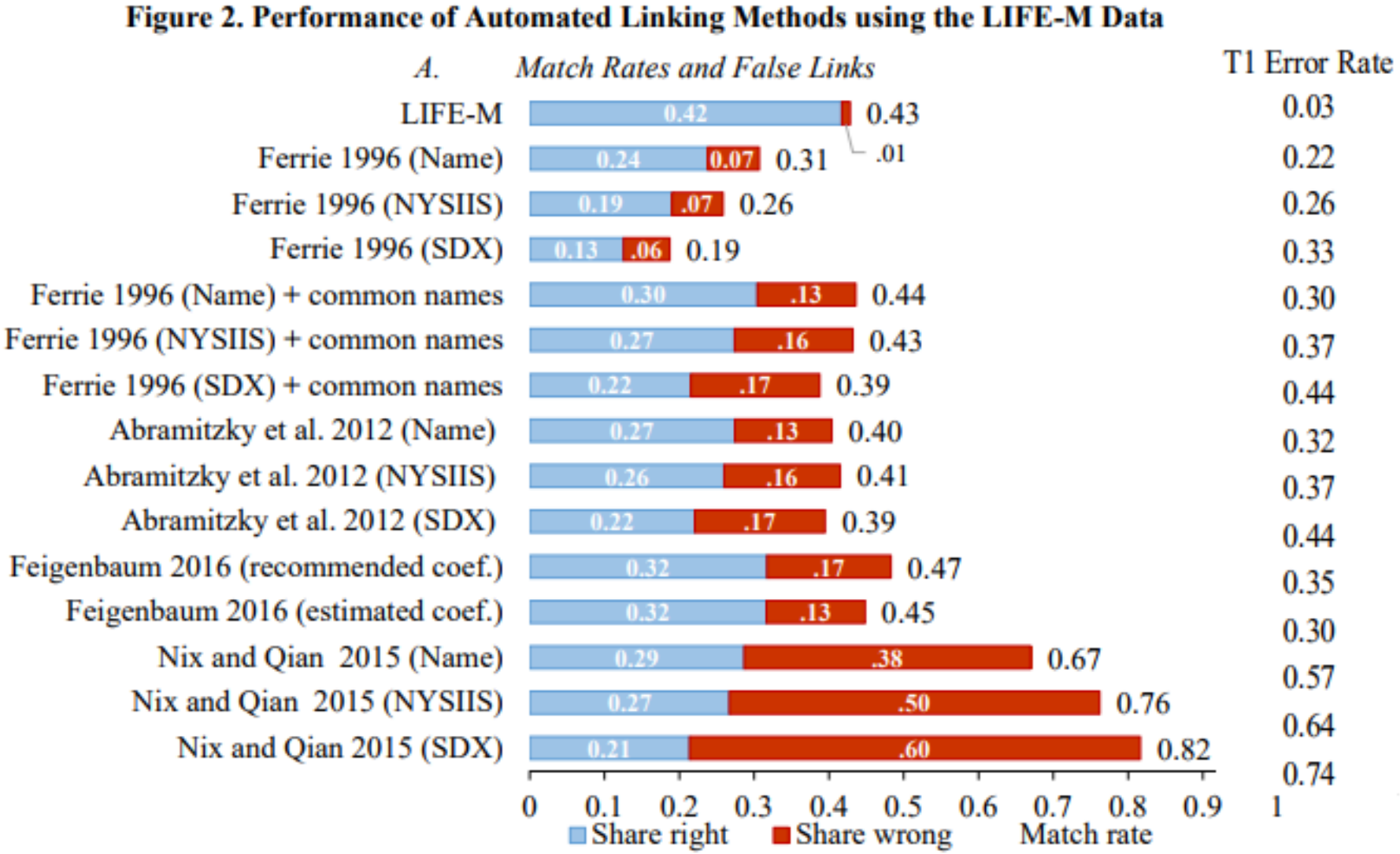
The button question

- “Imagine there was a button you could push and the Family Tree would all magically appear. Would you push the button?”

The button question

- “Imagine there was a button you could push and the Family Tree would all magically appear. Would you push the button?”
- Pros:
  - Community unity (see how we are connected). Valuable tool for research (social sciences and medicine). Allocate resources to most-needed data collections. Everyone would want to use it (more contributors). Less duplication. More focus on stories, photos, and memories (family history).
- Cons:
  - Lose process benefits of family history work (learning by doing). Automation might introduce errors. Mechanical replaces the spiritual.

Full automation isn't working yet.



[1] Family history + Machine learning

- Two important ways that traditional family history can contribute to the machine-learning approach:
- First, observing the decisions make while doing family history can provide training data for machine learning.
- Second, traditional family history tools can help with groups that have lower match rates using machine-learning approaches.
  - Includes women, immigrants, people who migrate a lot or have names that get misspelled.

[2] Family Tree

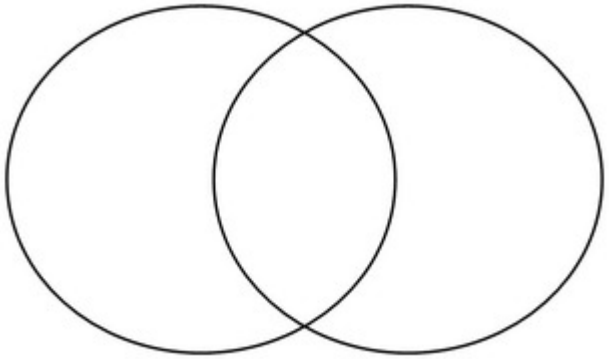
- My great grandfather has 12 sources attached, including 4 census records + M + D.
- This provides a huge training set for machine learning.
- 1900, 1910, and 1920 census: each have about 10 million people attached (12 million pairwise links).
  - 50 million 1900-1910 and 60 million 1910-1920 links should exist.
- As more people use the Family Tree, this will grow even faster.

[3] Academic partners

- There are lots of academics using various matching approaches to link census records together.
- Might be possible to create a central clearinghouse for these linkage efforts.
  - Scholars could send their predicted matches and get feedback on the accuracy of their methods.
  - Those predictions can be incorporated into new models to seek ways to improve (identify features and methods that work well).
- It might be possible to create a training set that can shared with academic partners to help improve their models.

Overlap approach

- There are family trees created by other groups.
- We are working with two done by academics: Utah Population Database and LIFEM.
  - Also: IHC, Ancestry, WikiTree, WeRelate, etc.
- We can overlay the two datasets and create a Venn diagram:
  - Both trees have the info (quality check)
  - New info for your tree
  - New info for the partner



Overlap with Utah Population Database

Table 1. Comparing Links in Family Search and UPD Data

	UPD Attached Only		FS Attached Only		Both Attached		None Attached	
	N	%	N	%	N	%	N	%
1880	541	5.4%	616	6.2%	1798	18.0%	7045	71.4%
1900	2056	20.6%	0	0.0%	7994	79.9%	0	0
1910	924	9.2%	1203	12.0%	3906	39.1%	3967	39.7%
1920	767	7.7%	1346	13.5%	2981	29.8%	4906	49.1%
1930	575	5.8%	1429	14.3%	2356	23.6%	5640	56.4%
1940	517	5.2%	1049	10.5%	1863	18.6%	6571	65.7%

- Combining with UPD would increase the census links on FS by: 22% for 1880, 26% for 1900, 18% for 1910, 18% for 1920, 15% for 1930, and 18% for 1940.

Cross-checking accuracy

Table 2. Agreement Rates between UPD and Family Tree

Census year	# Linked to both UPD and FT	Agreement Rate
1880	1798	94.2%
1900	7994	98.0%
1910	3906	98.0%
1920	2981	97.3%
1930	2356	98.1%
1940	1863	94.4%

- Volunteers can use the discordant pairs to help fix parts of the Family Tree, including records linked to a person's parent or multiple people linked to the same record.

Network Effects

- As the Family Tree grows, the value of using it will grow. This will attract more people to use the tree and create a virtuous cycle.
- Similar to what happened with Wikipedia.
  - Initial resistance from teachers and public but now as more and more academics contribute to it, it has gained legitimacy in the educational community.
  - Everyone uses it now which makes the value of using it even greater and draws more people to contribute.
- To strengthen this process we need to:
  - Measure how complete the tree is now (next slide).
  - Involve more people in contributing to the tree (microtasks + linking with other websites).

	2012	2013-2017	Close Relative	Total
Knox, Kentucky	48%	18%	18%	84%
Plymouth, Iowa	62%	12%	9%	83%
Frederick, Maryland	42%	25%	13%	80%
Bingham, Idaho	62%	14%	2%	79%
Wood, West Virginia	53%	10%	7%	70%
Beadle, South Dakota	36%	16%	11%	63%
Washoe, Nevada	30%	18%	5%	53%
Storey, Nevada	30%	16%	5%	51%
Cumberland, North Carolina	23%	14%	14%	51%
Humbolt, Nevada	26%	11%	5%	43%
Laramie, Wyoming	18%	9%	3%	30%
Ouachita, Louisiana	14%	8%	0%	24%

[4] Microtasks

- Find death date
- Find maiden name
- Find parents
- Find full first name
- Find exact birth date
- Resolve possible duplicates
- Resolve red flags
- Tags/sources for BMD
- Attach census records
- Include alternate names
- Resolve knots

Some of these tasks are easier to do than others.

One of the challenges faced by new users is that they often encounter a hard task early on (e.g. duplicates).

This is especially true new users with lots of family members who have done a lot of work already.

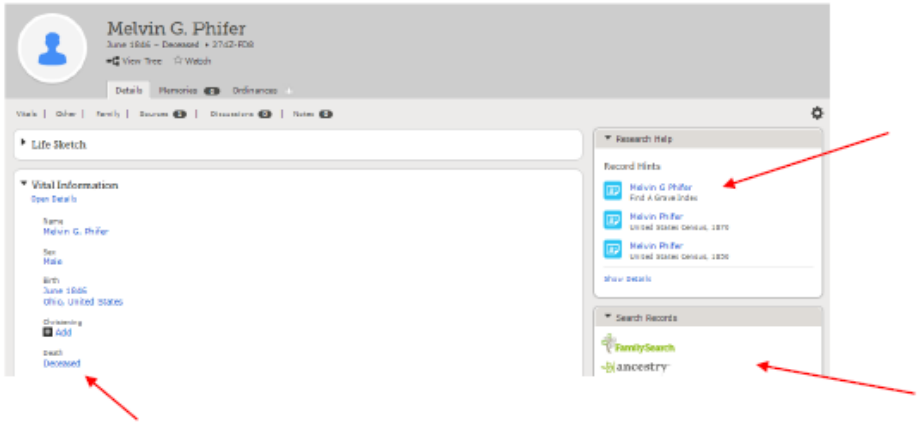
But in order to become an advanced user on Family Tree, you need practice (really important for youth).

Find the Death Date

For each person on the google spreadsheet, find the individual's death date, add that information to the Family Tree, and (when possible) attach a source that confirms that information. There are lots of sources that provide death dates including death certificates, gravestones (Find-a-Grave or billion graves), social security death index, or an obituary.

Step 1: Find the person on FamilySearch using the id on the google spreadsheet.

Step 2: Check record hints first since there might be a death record already there. You can also use the search records options to search for a death record on FamilySearch or Ancestry.com.



Step 3: Add the death date and place to the person's record and attach a source.

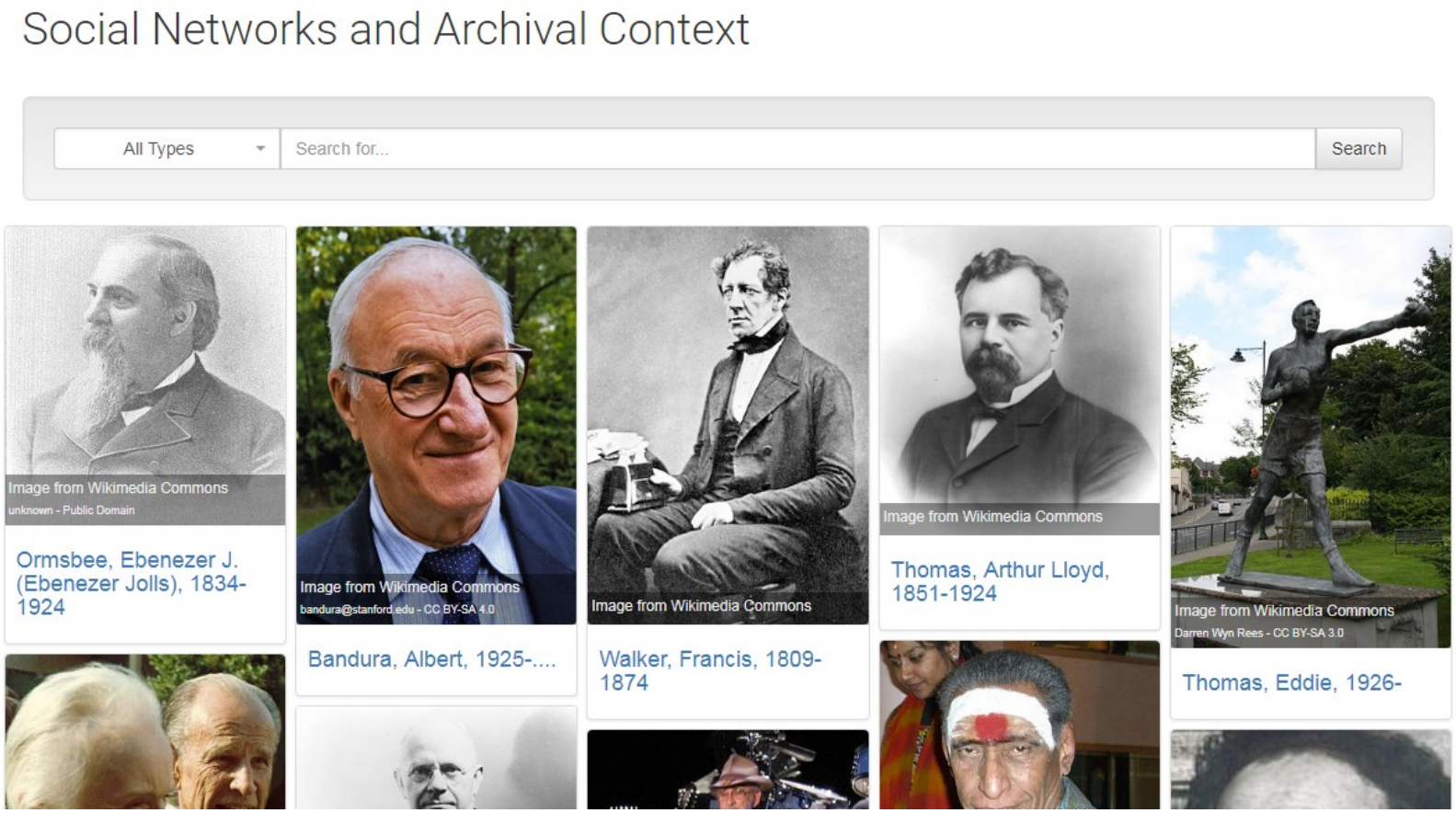
Step 4: Put your name under volunteer when you start to work on a group of people to reserve the part you will be working on and then put a "1" in the result column once you find the death date or a "0" if you are unable to find it. In some cases, the death date will have already been added to the person and just put a "1" in that case.

Feel free to add or edit the information on the tree based on the sources that you find. The ultimate goal of this project is to help complete the US portion to the tree so any information you add to the tree will be helpful

1	id	name	birthdate	sex	volunteer	result	notes
486	9X95-5ZH	Carrie Redman	February 1870	Female	Kathryn Arnett	0	
487	9X95-5ZM	Paul Blackburn	March 1863	Male	Kathryn Arnett	0	
488	9XGG-2RX	Maud E. Oswalt	Jul 1885	Female	Kathryn Arnett	1	
489	9XKX-2J4	Bernice Trott	11 Feb 1881	Female	Kathryn Arnett	0	
490	9XNL-1GL	Anna A. Loney	1867	Female	Kathryn Arnett	0	
491	9XPG-YJ9	Francis Albert Jackson	21 November 1887	Male	Kathryn Arnett	0	
492	9XTG-JD7	Elva Florence Allen	29 May 1889	Female	Kathryn Arnett	0	
493	9XXT-CYT	Charles Plumb Workm	14 June 1880	Male	Kathryn Arnett	0	
494	9Z3Z-P3V	Ellen Klotz	May 1836	Female	Kathryn Arnett	1	
495	9Z7R-GYT	Florence B. Higgins	Apr 1880	Female	Jacob Van Leeu	1	
496	9Z9J-478	Ross Silvester Simpson	14 Apr 1890	Male	Jacob Van Leeu	1	
497	9Z9J-4YQ	Archie Major Taylor	8 August 1885	Male	Jacob Van Leeu	1	
498	9ZBW-8HM	Pearl M. Taylor	1895	Female	Jacob Van Leeu	1	
499	9ZCG-H22	Lena Linn Vernon	21 March 1891	Female	Jacob Van Leeu	0	
500	9ZCG-H27	Avarilla Levering	3 Feb 1891	Female	Jacob Van Leeu	1	
501	9ZCG-NNH	Lulu Inez Murphy	23 Feb 1872	Female	Jacob Van Leeu	1	
502	9ZVB-16X	George W Osborn	1842	Male	Jacob Van Leeu	1	
503	H83L-2DN	Ada O Ayer Hills	November 1857	Female	Jacob Van Leeu	1	
504	K14M-3MZ	Alice B. Blakely	September 1865	Female	Jacob Van Leeu	1	
505	K14M-S1S	Charley S. Pealer	April 1861	Male			
506	K1FQ-1G1	John Holleck Dowds	1863	Male			
507	K25P-SB3	Oma McCarron	9 Oct 1876	Female			
508	K25P-SB7	Frederick McCarron	September 1841	Male			
509	K284-561	Harrison Stephens	1825	Male			
510	K2BM-GHV	Rachel Durbin	September 1822	Female			
511	K2D1-K6S	David C Graham	October 1839	Male			
512	K2D1-KD1	Rebecca Clark	January 1843	Female			

Linking with other websites

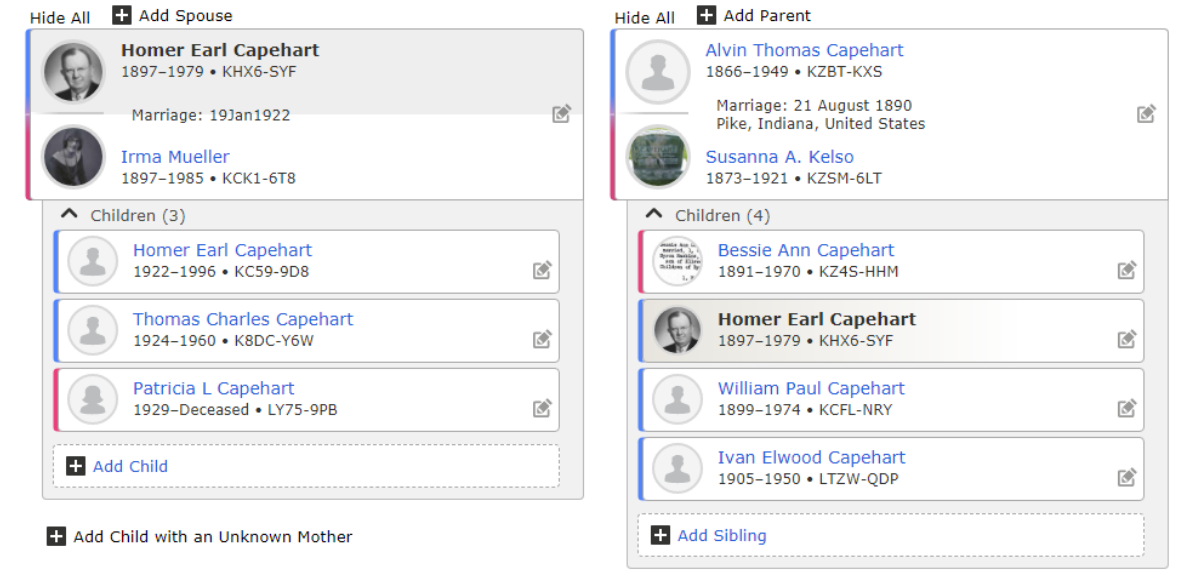
- 2.5 million people with profiles

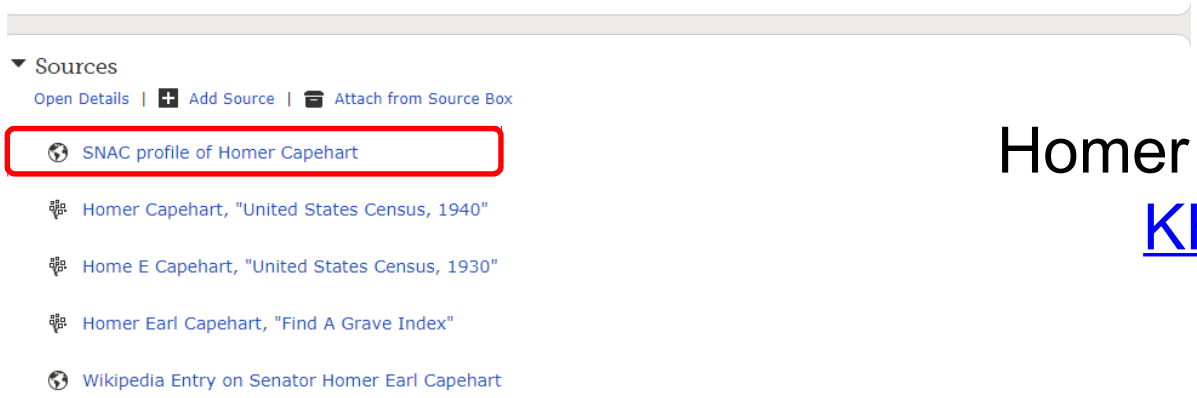


Create links back to Family Search

- The RLL is creating links between SNAC and FS. This will allow people doing family history to access materials in archives and historians to access family history materials about the person.

Create links from FS to SNAC





Homer Earl Capehart

KHX6-SYF

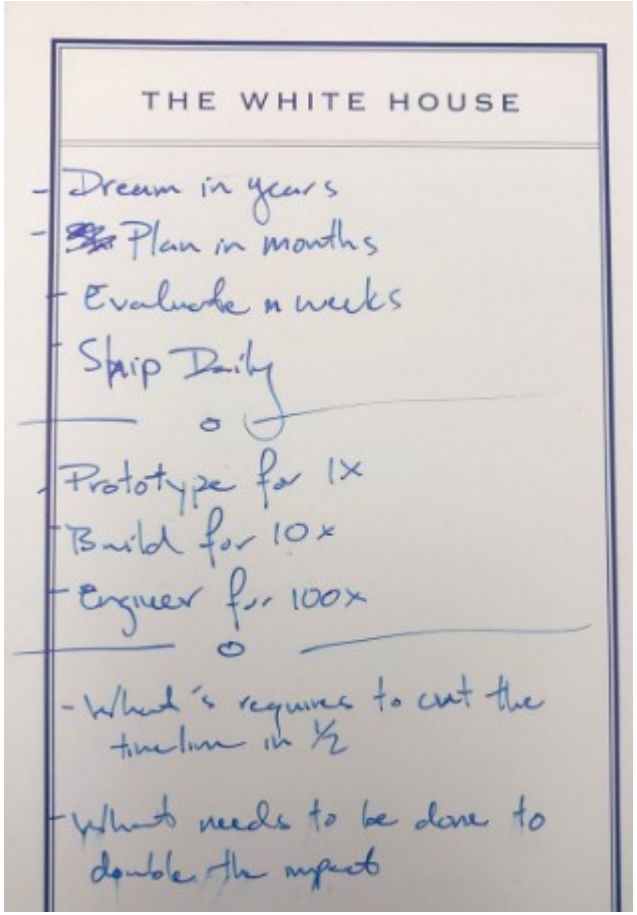
SNAC

Needles and Haystacks

- How many knew about SNAC already?
- How many have a relative in SNAC?
- It is hard to know which haystacks to look in.
- Another approach is to use machine learning to process entire haystacks and pin the needles to the tree.
- It will give people creative ideas of which haystacks to look in.
- Consider all the records that you show up in. Similar records exist for people in the past.

Urgent Work

- DJ Patil- White House note.
- "What is required to cut the timeline in half?"
- "What needs to be done to double the impact?"
- This is not a zero-sum game. We can collaborate and it can be a win-win.
  - Ancestry, FamilySearch, Academics
  - Who can I partner with to cut the timeline in half or double the impact?
- The work we are doing can have a real impact.



Final note – WWI project

ILLINOIS

KILLED IN ACTION

Majors

HILL, Henry Root, Quincy.
LANGWILL, William G., Aurora.
RIVET, James D., Oak Park.

Captains

BALDWIN, William W., Chicago.
DAVIS, Harold Wyman, Sycamore.
DEILEY, Paul C., Chicago.
HOOPES, Harold C., Ipava.
LOWEN, Jesse, Chicago.
MOSELEY, Arthur Francis, Freeport.
PETTIT, William S., Chicago.
SERCOMB, Albert A., Chicago.

Lieutenants

AARVIG, Truman, Pontiac.
BARTON, Lester C., Chicago.
BELLOWS, Franklin B., Wilmette.
BETTS, Elden S., Alton.
BLANCHARD, Merrill, Evanston.
BLANKENSHIP, Frederick Otto, Richview.
BLUM, Herbert C., Chicago.
BROTHERTON, William E., Guthrie.
BURES, George E., Chicago.
BURTIS, Darrel D., Waukegan.
CARPENTER, Jay I., Rochelle.
CARTER, Arthur R., Carbondale.
CLENDENEN, Paul M., Cairo.
COLTRA, Isaac V., Blue Mound.
COWAN, John W., Chicago.
COX, Paul G., Chicago.
CROWLEY, Sydney L., Oak Park.
CROWTHER, Orlando C., Canton.
CUNNINGHAM, Oliver B., Chicago.

Lieutenants—Continued

WOOD, Franklin, Chicago.
WRIGHT, Gustave, Oak Park.

Second Lieutenant

BUSBY, William H., Catlin.

Bat. Sergeant Major

McCOLLUM, Lawrence S., Benton.

Sergeants

ANDERSON, Lee, Elgin.
ANDERSON, Oskar, Chicago.
BACKSTROM, Robert E., Chicago.
BAILEY, Alfred, Chicago.
BECK, Eldon, Barnhill
BEEBE, Harold V., Woodstock.
BERG, Robert A., Chicago.
BISCHOFF, Elmer Joy, Oak Park.
BLASYK, John, Chicago.
BRADSHAW, Albert J., Peoria.
BRADSHAW, John, Pontiac.
BUESCH, Alfred Andres, Balleville.
BUSH, Ivory, Vandalia.
BUTLER, Guy P., Rock Island.
CARDWELL, Sheridan, Thompsonville.
CARROLL, William H., Peoria.
CHERRY, Claude E., Joliet.
CONWAY, Peter, Chicago.
COTTER, Martin, Chicago.
CRAIN, Wilford E., Whittington.
CRAY, Glenn H., Loraine
DE HAVEN, Walter, Chicago.
DELKER, Ferdinand, Teutopolis.
DLUZAK, Zygmond, Kankakee.
DOBRY, Michael J., Chicago.
DULMAGE, Ralph F., Zion City.

Linking Together the Entire US Population

Joe Price
joe_price@byu.edu
rll.byu.edu