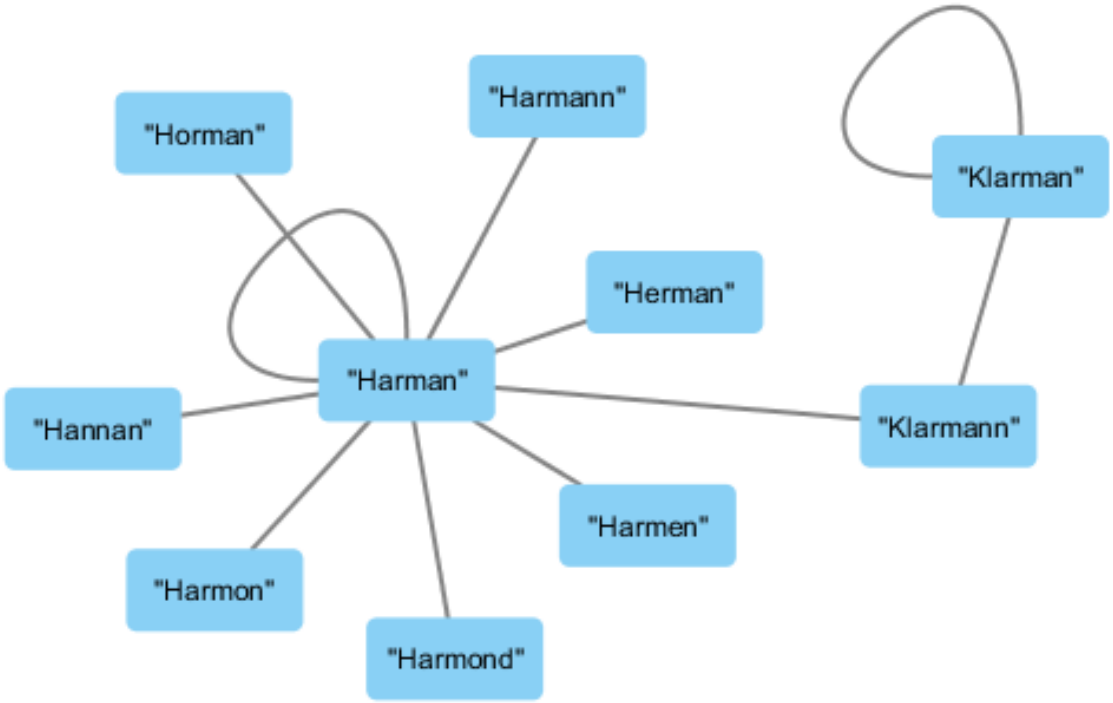
Name Networks

Christopher Cook
cookchr2@byu.edu

What is a Name Network?



Evans
Evens
Evan
Evins
Eveans
Eavens
Eavans
Evons
Eavins
Ivans
Evanes
Ewans
Everland
Evams
Evains
Eans
Evanas
Evanson
Emans
Emans
Eavns
Evrans
Evrans
Erlans
Eubanks
Evean
Evenes

Uses of name networks

- Identify common misspellings made by census takers and indexers
- Vectorize a name

e.g. Smith => [0.75, 1.23, 5.1, 3]

Results

- Name Vectorization improved recall by 3%
- Identified 500,000 names as “Correctly Spelled” (out of 10M unique surnames).
- Linked 1.2 Million misspellings to a correctly spelled name.

Misspelling	Name
Lehmai	Lehman
Dobberten	Dobbertin
Muyingo	Muzingo
Schaepering	Schaefering
Skarbak	Skarbek

Automated Approaches to Census-Linking

Isaac Riley
iriley@byu.edu

Motivation

- Immediate goal: identify the same individual across censuses
- Applications: family tree, historical research
- Problem: How do we know whether a similar person is actually the same person?

Example

	1910	1920
Name:	Marie H Smyth	Mary H Smith
Event Place:	Minneapolis	St. Paul
Birth Year:	1860	1861
Birth Place:	Wisconsin	Wisconsin
Mother's BP:	Prussia	Germany
Father's BP:	Rhode Island	Connecticut
Marital Status:	Married	Married
Race:	White	White
Sex:	Female	Female

Household Matching

Pros:

- Mimics human linking
- Allows us to identify probable deaths
- Narrows down the set of match candidates
- Uses more information

Cons:

- May miss subtle relationships in the data that machine learning (ML) picks up on
- Usually requires some hard boundary; less subtlety than ML
- Key assumption fails for institutional “households” – combinatorial explosion

Household Matching: Method 1

Start from a set of “anchor matches” and examine other pairwise combinations of household members for matches.

Pros:

- Allows lower match thresholds – good for data with errors
- Correctly matches married women with at least one family member in common between censuses

Cons:

- Misses all households without someone already matched
- Not requiring last name to match results in more false positives

Household Matching: Method 2

For each census, create a dataset containing all pairs of family members. Merge pairs to pairs on some loose criteria (Soundex, approximate birth year).

Pros:

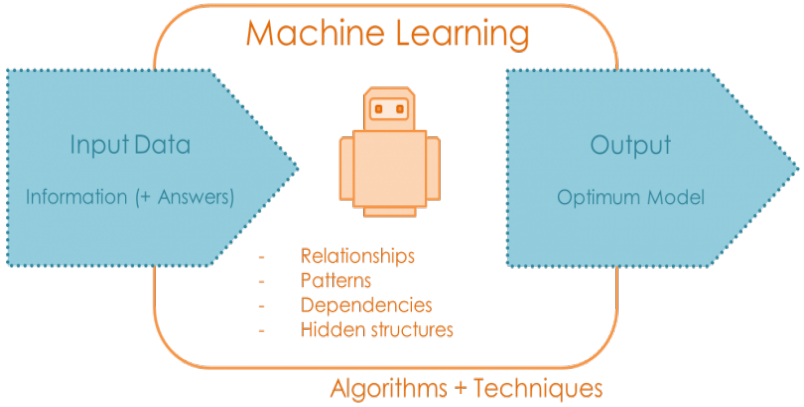
- Does not depend on “anchor pairs”

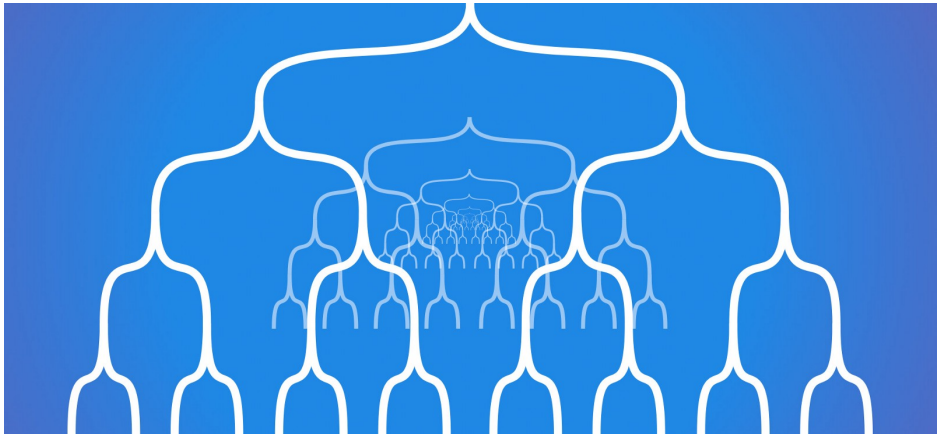
Cons:

- Not able to match married women
- Redundancy

Machine Learning

- Creates a model that takes a set of match “features” and returns a likelihood that the match is correct
- We use the XGBoost algorithm





- Requires a set of known true and known false matches
- Good training data will include “good false” and “less-good true” matches

Training Data

- Data cleaning
- Important derived variables:

  NARA Soundex code – group together similar-sounding names

  Cohort (approximate birth year) – round to nearest multiple of 4 and nearest multiple of 4 + 2 and use both

  Jaro-Winkler – generally the best string similarity measure, but only good for post-blocking applications

first	first_sdxn	mid	last	last_sdxn
Corneileous	C654		Crowley	C640
Cornelius	C654		Carberry	C616
Cornelius	C654	F	Howard	H630
Daniel	D540	S	Dolan	D450
Daniel	D540		Foster	F236
Denis	D520		Cashman	C255
Denis	D520		Curran	C650
Dennis	D520		Durnan	D655
Francis	F652	J	Collins	C452
Francis	F652	S	May	M000
Frank	F652	J	Jehle	J400
Frank	F652	G	Mehling	M452
Franklin	F652	J	Semmelroth	S546

birth_year	cohort1	cohort2	first1900	first1910	first_jw	last1900	last1910	last_jw
1850	1848	1852	John	John	1.000	Hamilton	Hamilton	1.000
1854	1852	1856	Richard	Richard	1.000	Silvey	Slivey	0.950
1857	1856	1856	Edward	Edward	1.000	George	Gorge	0.950
1859	1856	1860	Richmond	Richomnd	0.975	Jones	Jones	1.000
1863	1860	1864	Daniel	Danny	0.790	Foster	Foster	1.000
1864	1864	1864	George	George	1.000	Malet	Mallett	0.933
1865	1864	1864	Michael	Mickey	0.822	Daly	Daley	0.953
1866	1864	1868	Joseph	Jacob	0.620	Lacy	Lacey	0.953
1867	1864	1868	Johan	John	0.953	Olsen	Olsen	1.000
1868	1868	1868	Thomas	Tom	0.850	Benjamin	Buford	0.488
1870	1868	1872	Edward	Elias	0.620	Mehan	Hall	0.000
1871	1868	1872	Edward	Edvvard	0.879	O'Donnelly	ODonnelly	0.970
1873	1872	1872	August	August	1.000	Faulkner	Falkner	0.967
1874	1872	1876	Elias	Eli	0.907	Hall	Hall	1.000
1875	1872	1876	Joseph	Joe	0.867	Boniface	Boniface	1.000
1876	1876	1876	John		0.000	Burke	Burk	0.960
1877	1876	1876	Henry	Hank	0.670	Hummer	Hughes	0.733
1882	1880	1884	Charles	Karl	0.726	Fetting	Fetting	1.000
1883	1880	1884	Erick	Erik	0.953	Pundberg	Poundberg	0.967
1884	1884	1884	John	Johnny	0.933	Percival	Pinkerton	0.571

Machine Learning – Blocking Problem

- How to decide which individuals to compare with each other?
- Simple Cartesian Product not an option:
  80,000,000 x 80,000,000 = 6,400,000,000,000,000
- Perfect matches account for just a fraction of individuals
- Solution: merge the censuses on a variety of variable lists
- Challenge: keep size down while including as many of the correct matches as possible

Machine Learning

Pros:

- Complexity: able to pick up on subtleties that rule-based or even human linking methods would fail to pick up on
- Generality: run any candidate pair through the model and get a reasonable prediction; works for one-person households and fuzzy matches

Cons:

- Overfitting always a concern
- Does not make full use of household members
- Creating good training and prediction datasets can be challenging: data size vs. recall

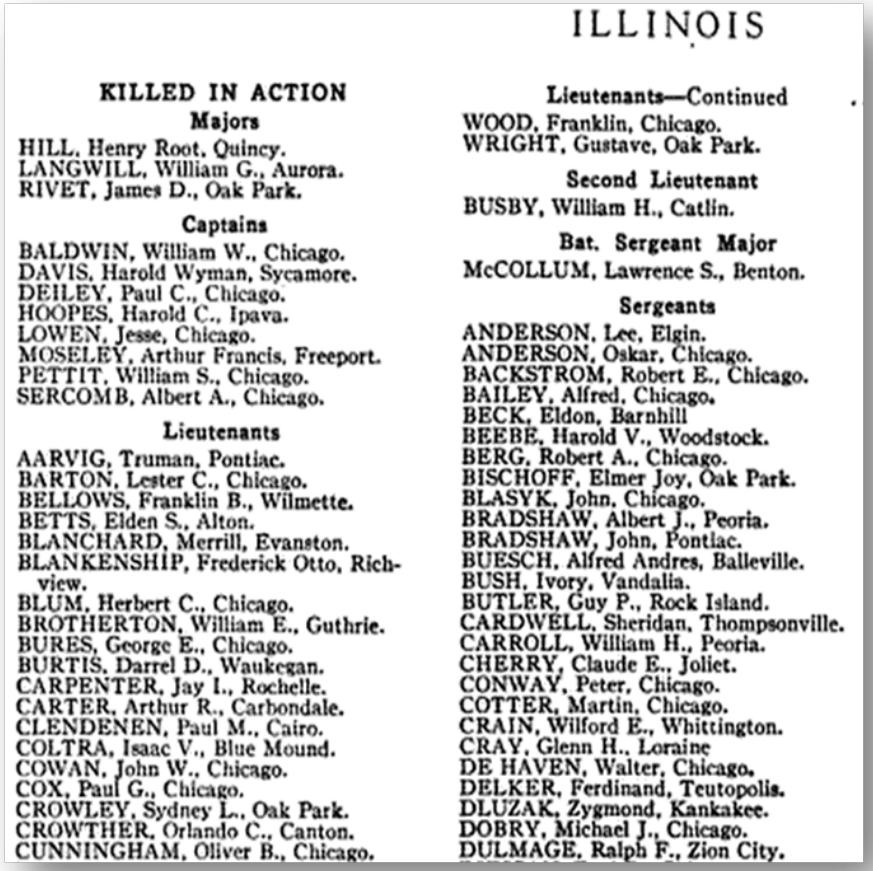
Final Note

- There is a trade-off with any linking method.
- Fortunately, we can combine our methods to improve performance.
- In the Record Linking Lab, we aim to find the ideal combination of hand-linking, rule-based, and machine learning approaches.

Research Applications of the Census Tree

Jacob Van Leeuwen
jacobrvl@byu.edu

Anti-German Discrimination During World War I



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, Belleville.
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.

- *Soldiers of the Great War* contains name and hometown of 70,000 soldiers who died in WWI
- Use this information to find soldiers in FamilySearch
- We can link soldiers to 1900 and 1910 US Censuses

Anti-German Discrimination During WWI

- Use soldier's census hometown to create treatment and control groups
- Census birthplace information allows us to find 1st and 2nd generation Germans
- We can use geocoordinates to link census towns from 1910 to 1920 census

pr_name_gn	pr_name_surn	pr_bir_place	pr_fthr_bi~e	pr_mthr_bi~e
Frederic	Martens	Delaware	Germany	Germany
Robert	Carson	Delaware	Delaware	Delaware
Bessie M	Scully	Delaware	Delaware	Delaware
Rebecca	Golstein	Germany	Germany	Germany
Arthur R	Veit	Pennsylvania	Germany	Germany

city1910	state1910	germans1910	city1920	state1920	germans1920
Pittsford	Iowa	58	Pittsford	Iowa	73
Ripley	Iowa	105	Ripley	Iowa	86
Shell Rock	Iowa	54	Shell Rock	Iowa	39
Washington	Iowa	135	Washington	Iowa	32
West Point	Iowa	166	West Point	Iowa	115

Anti-German Discrimination During World War I

Table 1. Effect of WWI death on percent change of ethnic Germans from 1910 to 1920

VARIABLES	(1)	(2)	(3)	(4)	(5)
Had Soldier Death	-0.168***	-0.071***	-0.062***	-0.053***	-0.056***
	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)
Population Controls	No	Yes	Yes	Yes	Yes
Race Controls	No	No	Yes	Yes	Yes
Gender Controls	No	No	No	Yes	Yes
Marital Status Controls	No	No	No	No	Yes
Observations	30,028	30,028	30,028	30,028	30,028
R-squared	0.005	0.261	0.266	0.270	0.272

Notes: Population controls include population for 1910 and 1920, as wells as the square of the population for 1910 and 1920. Race controls include Percent White and Percent Black for 1910 and 1920. Gender controls includes Percent Female, and marital status controls includes Percent Married. Cities in the sample were restricted to have a 1910 population of 55,000 or less.

- We find the number of individuals reporting birthplace as Germany decreases from 2.5 million in 1910 to 1.7 million in 1920
- We find a 5.6% decrease in German individuals living in a census town in 1920 which had a soldier die in World War I

Determinants of City Size

- 62 of the 500 largest cities in the 1900 census decreased in size in the 1940 census
- We want to examine the demographic factors that are related to cities decreasing in size
- We can exploit census place data from the census tree to link cities across censuses
- Census tree allows individual's location to be tracked across censuses

City Linking Process

city1910	county1910	state1910	city1920	county1920	state1920	pop
Winston	Forsyth	North Carolina	Winston-Salem	Forsyth	North Carolina	5549
Goldmine	Franklin	North Carolina	Gold Mine	Franklin	North Carolina	422
Bennetts Bayou	Fulton	Arkansas	Bennett Bayou	Fulton	Arkansas	169
Fort Atkinson	Jefferson	Wisconsin	Koshkonong	Jefferson	Wisconsin	1157
Tripoli	Bremer	Iowa	Frederika	Bremer	Iowa	144
Plattsburg	Clinton	Missouri	Concord	Clinton	Missouri	450

- We find individuals linked in consecutive censuses (i.e. 1910 and 1920)
- For each location in the first census, we find the most frequent location in the second census
- We link the most frequent location in the second census to the first census

City Linking Process

- Advantages
  - Can match cities that would not be matched with a string-matching algorithm
    - Ex: Winston, NC in 1910 becomes Winston-Salem in 1920
  - Matches cities with minor misspelling between censuses
  - Matches difficult to match census places
    - Ex: Township 1 in 1910 matches to Montecito in 1920

city1910	county1910	state1910	city1920	county1920	state1920	pop
Township 1	Santa Barbara	California	Montecito	Santa Barbara	California	194
Township 3	Santa Barbara	California	Goleta	Santa Barbara	California	128
Township 4	Santa Barbara	California	Santa Ynez	Santa Barbara	California	96
Township 5	Santa Barbara	California	Lompoc	Santa Barbara	California	364
Township 6	Santa Barbara	California	Los Alamos	Santa Barbara	California	52

Final Note

- The Census Tree has unique features that can be exploited in research
- Census records contain additional information that can be used in research
  - Education, income, employment, occupation
- The Census Tree can be used to link families through time in intergenerational research