Big Data for Economic and Demographic History: The MPC-Ancestry Complete Count American Census Series

Evan Roberts
Minnesota Population Center



Minnesota Population Center

IPUMS



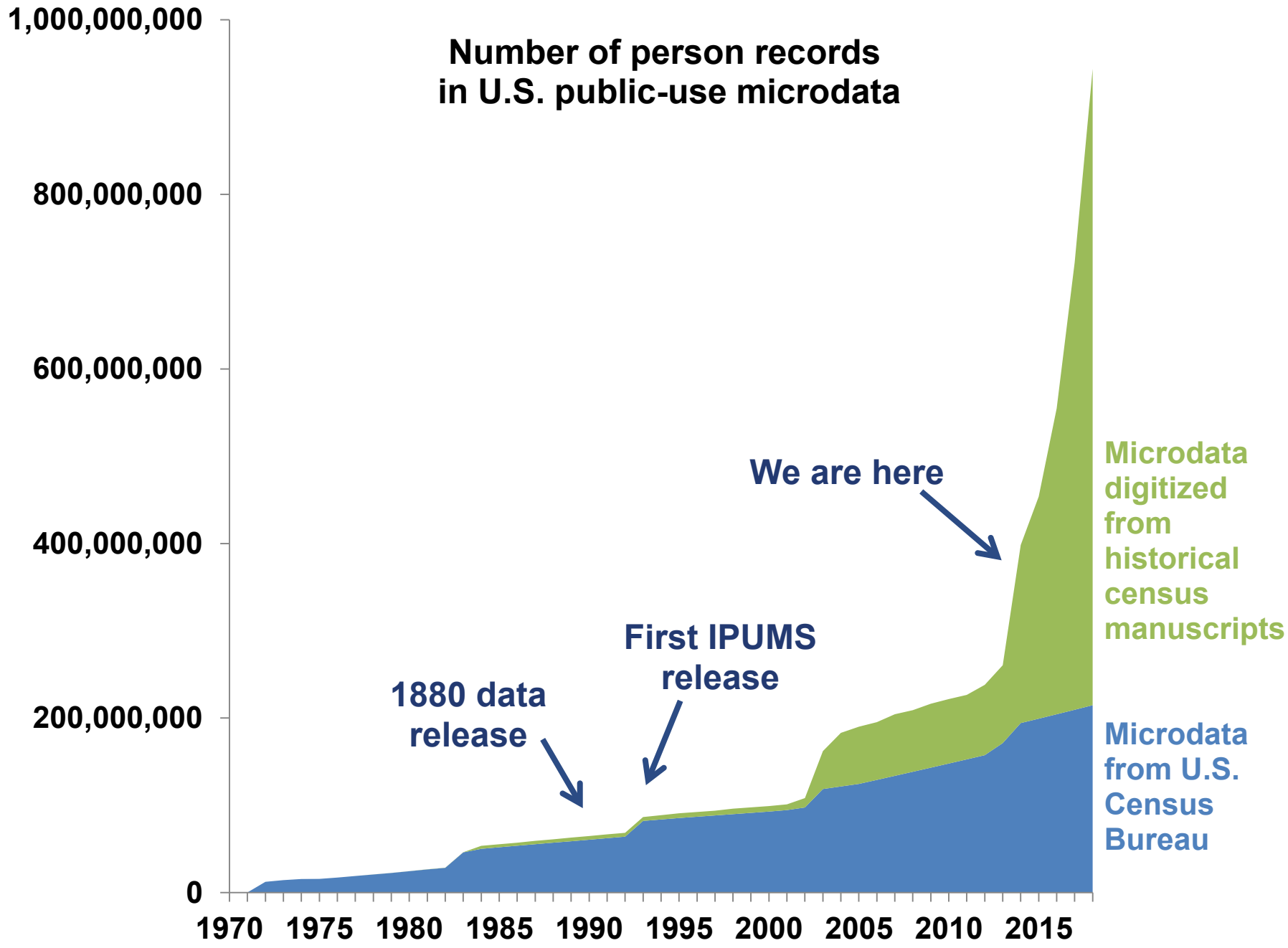




Complete microdata for the U.S., 1790-1940

- 750 million records
- Data-entry required over 10,000 years of effort
- Donated data would cost $520 million to replicate

MPC


Number of person records
in U.S. public-use microdata
1,000,000,000
800,000,000
600,000,000
400,000,000
200,000,000
0
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
We are here
First IPUMS
release
1880 data
release
Microdata
digitized
from
historical
census
manuscripts
Microdata
from U.S.
Census
Bureau

Note A.—The Census Year begins June 1, 1879, and ends May 31, 1880.

Note B.—All persons will be included in the Enumeration who were living on the 1st day of June, 1880. No others will. Children BORN SINCE June 1, 1880, will be OMITTED. Members of Families who have DIED SINCE June 1, 1880, will be

Note C.—Questions Nos. 13, 14, 22 and 23 are not to be asked in respect to persons under 10 years of age.

Received July 22, 1880.

119

st. No. 2

st. No. 96

1st District Princeton Township

E 1.—Inhabitants in Borough of Princeton, in the County of Mercer, State of New Jersey,

enumerated by me on the ____ day of June, 1880.

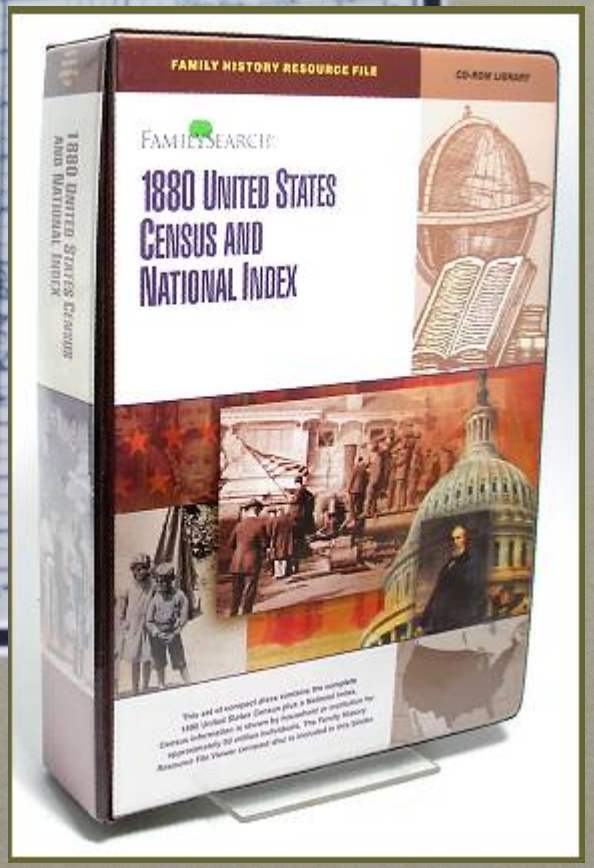
Chas O Hudnut

Enumerato

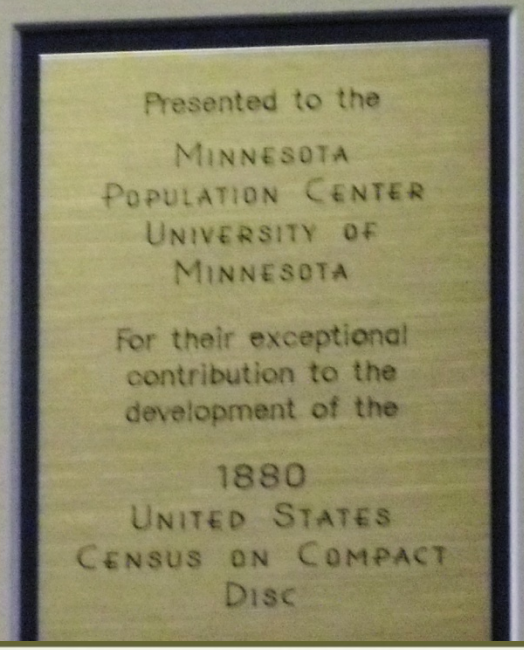
Families numbered in order of visitation.	The Name of each Person whose place of abode, on 1st day of June, 1880, was in this family.	Personal Description. Color—White, W.; Black, B.; Mulatto, Mu.; Chinese, Ch.; Indian, I.	Sex—Male, M.; Female, F.	Age at last birthday prior to June 1, 1880. If under 1 year, give months in fractions, thus: ⅚	If born within the Census year, give the month.	Relationship of each person to the head of this family—whether wife, son, daughter, servant, boarder, or other.	Civil Condition. Single, ✓	Married, ✓	Widowed, ✓	Divorced, D.	Married during Census year, ✓	Occupation. Profession, Occupation or Trade of each person, male or female.	Number of months this person has been unemployed during the Census year.	Health. Is the person [on the day of the Enumerator's visit] sick or temporarily disabled, so as to be unable to attend to ordinary business or duties? If so, what is the sickness or disability?	Blind, ✓	Deaf and Dumb, ✓	Idiotic, ✓	Insane, ✓	Maimed, Crippled, Bedridden, or otherwise disabled, ✓	Education. Attended school within the Census year, ✓	Cannot read, ✓	Cannot write, ✓	Nativity. Place of Birth of this person, naming State or Territory of United States, or the Country, if of foreign birth.	Place of Birth of the Father of this person, naming the State or Territory of United States, or the Country, if of foreign birth.	Place of Birth of the Mother of this person, naming the State or Territory of United States, or the Country, if of foreign birth.
2	3	4	5	6	7	8	9	10	11	12		13	14	15	16	17	18	19	20	21	22	23	24	25	26
88	Schenck Daniel	Mu	M	40				1				Laborer											New Jersey	New Jersey	New Jersey
	Phoebe	Mu	F	32		Wife		1				Keeping House									1	1	New Jersey	Ky	Ky
89	Johnson Clara	B	F	5-1					1			Keeping House									1	1	New Jersey	Va	Va
	Jane	B	F	29		Daughter	1					Servant											New Jersey	Kentucky	Ky
	Mary E	B	F	21		Daughter	1					At Home											New Jersey	Penn	Ky
	Georgiana	Mu	F	14		Daughter	1					at Home								1			New Jersey	Penn	Ky
	Viola	Mu	F	2		G. Daughter	1																New Jersey	Penn	Ky
	Mary E	Mu	F	10/12		G. Daughter	1																New Jersey	Penn	Ky
90	Dickson James	B	M	48				1				Laborer	3										New Jersey		
	Lucinda	B	F	46		Wife		1				Keeping House											New York		
91	Golden Ellen	W	F	55					1			Keeping House									1	1	Ireland	Ireland	Ireland
	Mary	W	F	17		Daughter	1					Servant											New Jersey	Ireland	Ireland
	Philip	W	M	20		Son	1					Laborer										1	New Jersey	Ireland	Ireland
92	O'Connor Ellen	W	F	5-9			1					Servant									1	1	Ireland	Ireland	Ireland

1999: Collaboration with the Latter-Day Saints




FAMILY HISTORY RESOURCE FILE
CD-ROM LIBRARY
FAMILYSEARCH
1880 United States Census and National Index

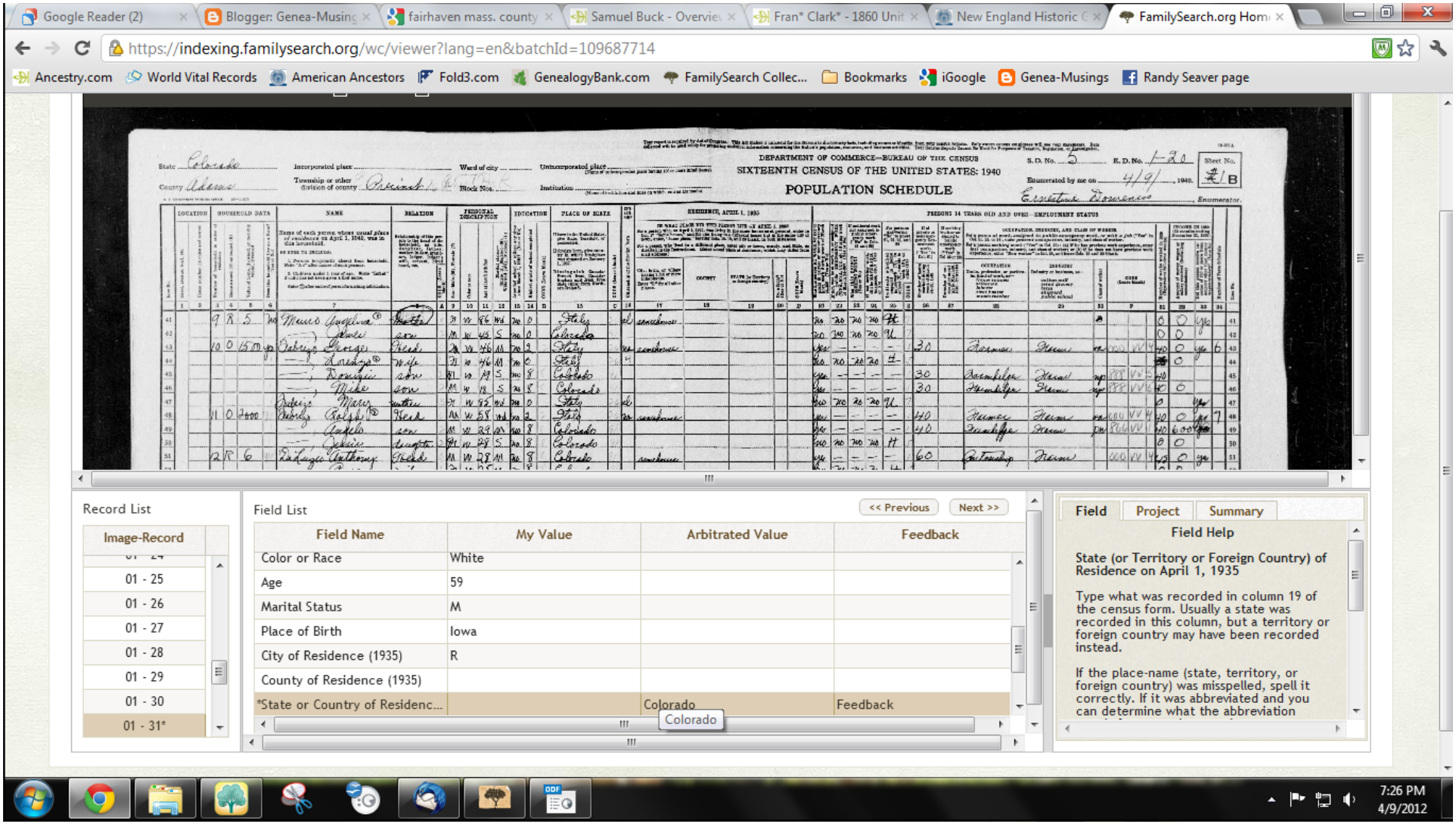
55 CD-ROM Set


Presented to the
Minnesota
Population Center
University of
Minnesota
For their exceptional
contribution to the
development of the
1880
United States
Census on Compact
Disc



FAMILYSEARCH

Crowdsourcing Project 2006-2011



Form 15-6

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

FIFTEENTH CENSUS OF THE UNITED STATES: 1930

POPULATION SCHEDULE

7339

Sheet No. 9 B

State Illinois — Incorporated place Chicago City — Enumeration District No. 16-2772

County Cook — Ward of city 41 — Block No. 58 — Supervisor's District No. 3

Township or other division of county Precinct 0 — Unincorporated place — Institution — Enumerated by me on April 7, 1930, Irma L. Burroughs, Enumerator.

Place of abode: Street, avenue, road, etc.	Place of abode: House number (in cities or towns)	Place of abode: Number of dwelling house in order of visitation	Place of abode: Number of family in order of visitation	NAME 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	RELATION: Relationship of this person to the head of the family	HOME DATA: Home owned or rented	HOME DATA: Value of home, if owned, or monthly rental, if rented	HOME DATA: Radio set	HOME DATA: Does this family have a radio set?	PERSONAL DESCRIPTION: Sex	PERSONAL DESCRIPTION: Color or race	PERSONAL DESCRIPTION: Age at last birthday	PERSONAL DESCRIPTION: Marital condition	PERSONAL DESCRIPTION: Age at first marriage	EDUCATION: Attended school or college at any time since Sept. 1, 1929	EDUCATION: Whether able to read and write	PLACE OF BIRTH: Person	PLACE OF BIRTH: Father	PLACE OF BIRTH: Mother	MOTHER TONGUE (OR NATIVE LANGUAGE) OF FOREIGN BORN: Language spoken in home before coming to the United States	CITIZENSHIP, ETC.: Code (For office use only. Do not write in these columns) A	B	C	CITIZENSHIP, ETC.: Year of immigration into the United States	CITIZENSHIP, ETC.: Naturalization	CITIZENSHIP, ETC.: Whether able to speak English	OCCUPATION AND INDUSTRY: Occupation	Industry	Code (For office use only. Do not write in this column)	Class of worker	EMPLOYMENT: Whether actually at work yesterday (or the last regular working day)	If not, line number on Unemployment Schedule	VETERANS: Whether a veteran of U.S. military or naval forces	What war or expedition	Number of farm schedule	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	A	B	C	22	23	24	25	26	D	27	28	29	30	31	32	
Ford	7217			Stegma, Allen N.	Grandson				No	M	W	6 5/12	S			Yes	Illinois	Missouri	Illinois		61					Yes	None									
Ford	7219	129	129	Kerta, William	Head	O	10,000	R	No	M	W	65	M	25	No	Yes	Germany	Germany	Germany	German	13	13	V	1870	NA	Yes	Trunkmaker	Leather	7740	W	Yes		No			

IPUMS

1940 Census Project



Collaboration of MPC and Ancestry.com

- 7.8 billion keystrokes
- 134 million persons

MPC

State New York Incorporated place Bronx Borough Ward of city 1 Unincorporated place
County Bronx Township or other division of county A.D. 1 Block Nos. A B C & F Institution
(Name of unincorporated place having 100 or more inhabitants) (Name of institution and lines on which entries are made)

Your report is required by Act of Congress. This Act makes it unlawful for the Bureau to disclose any facts, including names or identity, from your answers or records. Only sworn employees will see your statements. Data collected will be used solely for preparing statistical information concerning the Nation's population, resources, and business activities. You

DEPARTMENT OF COMMERCE—BUREAU O
SIXTEENTH CENSUS OF THE UNITE
POPULATION SCHED

I. D. No. 3-102 Sheet No. 1 A
April X, 1940. Enumerato

Line No.	LOCATION: Street, avenue, road, etc.	LOCATION: House number (in cities and towns)	HOUSEHOLD DATA: Number of household in order of visitation	HOUSEHOLD DATA: Number of dwelling house (D) or second (S)	HOUSEHOLD DATA: Number of family, if more than one in dwelling house, or of household, if more than one in apartment	HOUSEHOLD DATA: Number of household line on a farm* (Yes or No)	NAME: Name of each person whose usual place of residence on April 1, 1940, was in this household. BE SURE TO INCLUDE: 1. Persons temporarily absent from household. Write "Ab" after names of such persons. 2. Children under 1 year of age. Write "Infant" if child has not been given a first name.	RELATION: Relationship of this person to the head of the household, as wife, daughter, father, mother-in-law, grandson, lodger, lodger's wife, servant, hired hand, etc.	PERSONAL DESCRIPTION: Sex (M, Male; F, Female)	PERSONAL DESCRIPTION: Color or race (W, White; Negro; Am. Ind., American Indian; Chi., Chinese; Jap., Japanese; Fil., Filipino; Hin., Hindu; Kor., Korean; Mex., Mexican; Other (specify))	PERSONAL DESCRIPTION: Age at last birthday	PERSONAL DESCRIPTION: Marital status (S, Single; M, Married; Wd, Widowed; D, Divorced)	EDUCATION: Attended school or college any time since March 1, 1940 (Yes or No)	EDUCATION: Highest grade of school completed	PLACE OF BIRTH: If born in the United States, give State, Territory, or possession. If foreign born, give country in which birthplace was situated on January 1, 1937. Distinguish Canada-French from Canada-	CITIZENSHIP: If foreign born, is he naturalized? (Yes or No)	RESIDENCE, APRIL 1, 1935: IN WHAT PLACE DID THIS PERSON LIVE ON APRIL 1, 1935? For a person who, on April 1, 1935, was living in the same house as at present, enter in Col. 17 "Same house," and for one living in a different house but in the same city or town, enter "Same place," leaving Cols. 18, 19, and 20 blank, in both instances. For a person who lived in a different place, enter city or town, county, and State, as directed in the Instructions. (Enter actual place of residence, which may differ from mail address.) City, town, or village having 2,500 or more	RESIDENCE, APRIL 1, 1935: STATE (or Territory)	RESIDENCE, APRIL 1, 1935: Farm (Yes or No)	RESIDENCE, APRIL 1, 1935: Same house in 1935 (Yes or No)	RESIDENCE, APRIL 1, 1935: Same place in 1935 (Yes or No)
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	[illegible]	10	1																	YES	—	—	—

Line No.	If neither at work nor assigned to public emergency work. ("No" in Cols. 21 and 22): Was this person SEEKING WORK? (Yes or No)	If not seeking work, did he HAVE A JOB, business, etc.? (Yes or No)	For persons answering "No" to quest. 21, 22, 23, and 24: Indicate whether engaged in home housework (H), in school (S), unable to work (U), or other (Ot)	CODE (Leave blank)	If at private or nonemergency Government work. ("Yes" in Col. 21): Number of hours worked during week of March 24–30, 1940	If seeking work or assigned to public emergency work. ("Yes" in Col. 22 or 23): Duration of unemployment up to March 30, 1940—in weeks	OCCUPATION, INDUSTRY, AND CLASS OF WORKER. For a person at work, assigned to public emergency work, or with a job ("Yes" in Col. 21, 22, or 24), enter present occupation, industry, and class of worker. For a person seeking work ("Yes" in Col. 23): (a) If he has previous work experience, enter last occupation, industry, and class of worker; or (b) if he does not have previous work experience, enter "New worker" in Col. 28, and leave Cols. 29 and 30 blank. OCCUPATION: Trade, profession, or particular kind of work, as—frame spinner, salesman, laborer, rivet heater, music teacher	INDUSTRY: Industry or business, as—cotton mill, retail grocery, farm, shipyard, public school	Class of worker	CODE (Leave blank)	Number of weeks worked in 1939 (Equivalent full-time weeks)
	23	24	25	E	26	27	28	29	30	F	31
1	—	—	—	1	40		TIMEKEEPER	CUT STONE YARD	P.W	246 24 1	52
2	—	—	—	1	40	25	HELPER	CUT STONE YARD	P.W	246 24 1	40
3	NO	NO	H	5							0
4											
5	—	—	—	1	48		SUPERINTENDANT LOFT BUILDING	PHILLIP JONES SHIRT COMPANY	P.W	129 81 1	52
6	NO	NO	H	5							

Line No.	For Persons 14 Years Old and Over: Was this person AT WORK for pay or profit in private or nonemergency Govt. work during week of March 24–30? (Yes or No)	If not, was he at work on, or assigned to, public EMERGENCY WORK (WPA, NYA, CCC, etc.) during week of March 24–30? (Yes or No)	CLASS OF WORKER: If at work (or assigned to public emergency work) during week of March 24–30, enter description and at which he is physically at occupation at which he has worked [illegible] to work. Enter also usual industry in Col. 45 and leave Cols. 46 and 47	Was this person at work on a farm? (Yes or No)	Was this person at work in his own home? (Yes or No)	Was this person at work for wages or salary? (Yes or No)	Was this person at work without pay on family farm or in family business? (Yes or No)	Did this person have a job but not at work during week of March 24–30? (Yes or No)	If No, was he on layoff with definite date to return? (Yes or No)	If No, was he on layoff or looking for work? (Yes or No)	Was this person looking for work? (Yes or No)	USUAL OCCUPATION	USUAL INDUSTRY	Usual class of worker	CODE (Leave blank)	Did this person have more than one job? (Yes or No)	Number of weeks worked in 1939	Wages, salary, commissions, or profits in 1939	Was this person a veteran of U.S. military or naval forces? (Yes or No)	If so, what war or expedition? (Civil War, Spanish War, World War, etc.)	If so, what was the date of entrance into service? (Month and year)	If so, what was the date of separation from service? (Month and year)
	21	22	23	G	H	I	J	K	L	M	N	45	46	47	J	48	31	32	33	34		
1	YES	—	—														52	2280	NO			
2																			NO			
3																			NO			
4																						
5	YES	—	—																NO			
6																			NO			
7	YES	—	—																NO			
8																			NO			
9																			NO			
10																			NO			
11																			NO			
12																			NO			
13																			NO			
14																			NO			
15																			NO			
16																			NO			
17																			NO			
18																			NO			
19																			NO			
20																			NO			
21																			NO			
22																			NO			
23																			NO			
24																			NO			
25																			NO			
26																			NO			
27																			NO			
28																			NO			
29																			NO			
30																			NO			

Line No.	PERSONS 14 YEARS OLD AND OVER—EMPLOYMENT STATUS: Was this person AT WORK for pay or profit in private or nonemergency Govt. work during week of March 24–30? (Yes or No)	If not, was he at work on, or assigned to, public EMERGENCY WORK (WPA, NYA, CCC, etc.) during week of March 24–30? (Yes or No)	If neither at work nor assigned to public emergency work. ("No" in Cols. 21 and 22): Was this person SEEKING WORK? (Yes or No)	If not seeking work, did he HAVE A JOB, business, etc.? (Yes or No)	For persons answering "No" to quest. 21, 22, 23, and 24: Indicate whether engaged in home housework (H), in school (S), unable to work (U), or other (Ot)	CODE (Leave blank)	If at private or nonemergency Government work. ("Yes" in Col. 21): Number of hours worked during week of March 24–30, 1940	If seeking work or assigned to public emergency work. ("Yes" in Col. 22 or 23): Duration of unemployment up to March 30, 1940—in weeks	OCCUPATION, INDUSTRY, AND CLASS OF WORKER. For a person at work, assigned to public emergency work, or with a job ("Yes" in Col. 21, 22, or 24), enter present occupation, industry, and class of worker. For a person seeking work ("Yes" in Col. 23): (a) If he has previous work experience, enter last occupation, industry, and class of worker; or (b) if he does not have previous work experience, enter "New worker" in Col. 28, and leave Cols. 29 and 30 blank. OCCUPATION: Trade, profession, or particular kind of work, as—frame spinner, salesman, laborer, rivet heater, music teacher	INDUSTRY: Industry or business, as—cotton mill, retail grocery, farm, shipyard, public school	Class of worker	CODE (Leave blank)	Number of weeks worked in 1939 (Equivalent full-time weeks)
	21	22	23	24	25	E	26	27	28	29	30	F	31
1	YES	—	—	—	—	1	40		TIMEKEEPER	CUT STONE YARD	P.W	246 24 1	52
2	YES	—	—	—	—	1	40	25	HELPER	CUT STONE YARD	P.W	246 24 1	40
3	NO	NO	NO	NO	H	5							0
4													
5	YES	—	—	—	—	1	48		SUPERINTENDANT LOFT BUILDING	PHILLIP JONES SHIRT COMPANY	P.W	129 81 1	52
6	NO	NO	NO	NO	H	5							

Line No.	For Persons 14 Years Old and Over: Was this person AT WORK for pay or profit in private or nonemergency Govt. work during week of March 24–30? (Yes or No)	If not, was he at work on, or assigned to, public EMERGENCY WORK (WPA, NYA, CCC, etc.) during week of March 24–30? (Yes or No)	CLASS OF WORKER: If at work (or assigned to public emergency work) during week of March 24–30, enter description and at which he is physically at occupation at which he has worked [illegible] to work. Enter also usual industry in Col. 45 and leave Cols. 46 and 47	Was this person at work on a farm? (Yes or No)	Was this person at work in his own home? (Yes or No)	Was this person at work for wages or salary? (Yes or No)	Was this person at work without pay on family farm or in family business? (Yes or No)	Did this person have a job but not at work during week of March 24–30? (Yes or No)	If No, was he on layoff with definite date to return? (Yes or No)	If No, was he on layoff or looking for work? (Yes or No)	Was this person looking for work? (Yes or No)	USUAL OCCUPATION	USUAL INDUSTRY	Usual class of worker	CODE (Leave blank)	Did this person have more than one job? (Yes or No)	Number of weeks worked in 1939	Wages, salary, commissions, or profits in 1939	Was this person a veteran of U.S. military or naval forces? (Yes or No)	If so, what war or expedition? (Civil War, Spanish War, World War, etc.)	If so, what was the date of entrance into service? (Month and year)	If so, what was the date of separation from service? (Month and year)
	21	22	23	G	H	I	J	K	L	M	N	45	46	47	J	48	31	32	33	34		
1	YES	—	—														52	2280	NO			
2																			NO			
3																			NO			
4																						
5	YES	—	—																NO			
6																			NO			
7	YES	—	—																NO			
8																			NO			
9																			NO			
10																			NO			
11																			NO			
12																			NO			
13																			NO			
14																			NO			
15																			NO			
16																			NO			
17																			NO			
18																			NO			
19																			NO			
20																			NO			
21																			NO			
22																			NO			
23																			NO			
24																			NO			
25																			NO			
26																			NO			
27																			NO			
28																			NO			
29																			NO			
30																			NO			

Line No.		FATHER	MOTHER	CODE (Leave blank)		CODE (Le	H	Trouble in speech (Yes or No)	Ware-roll service	CODE (Le blank)	Does this person have a usual occupation?	Work for self (S) or Both (B) (Yes or No)	If No, was on layoff or looking for work? (Yes or No)	USUAL OCCUPATION	USUAL INDUSTRY	Usual class of worker	CODE (Leave blank)	Did this person have more than one job? (Yes or No)	Number of weeks worked in 1939	Wages, salary, commissions, or profits in 1939	Was this person a veteran of U.S. military or naval forces? (Yes or No)	If so, what war or expedition? (Civil War, Spanish War, World War, etc.)	If so, what was the date of entrance into service? (Month and year)	If so, what was the date of separation from service? (Month and year)
	35	36	37	G	38		H	39	40	41	I	42	43	44	45	46	47	J	48	31	32	33	34	
14	IMOR. AUGUSTUS	NEW YORK	NEW YORK		ENGLISH			NO	NO	NO		NO	—	—										
29	TENKE VILMA	NOGO-SLAVIA	GERMANY	17	ENGLISH			NO	NO	NO		NO	—	—	NONE	STUDENT	S.			0	00	NO		

Col. 9. RELATION TO HEAD OF HOUSEHOLD: Col. 10. COLOR OR RACE: Col. 11. AGE AT LAST BIRTHDAY: Col. 14. HIGHEST GRADE OF SCHOOL COMPLETED: Col. 16. CITIZENSHIP OF THE FOREIGN: Col. 21. WAS THIS PERSON AT WORK? Col. 24. DID THIS PERSON HAVE A JOB? Cols. 30 and 47. CLASS OF WORKER: Col. 41. WAR OR MILITARY SERVICE:

EDUCATION

For Persons 14 Years Old and Over: A SAME HOUSE

NO H-4 30
NO H-4 36

10 7 2 1 1/2 1 8 7

IPUMS

Terms of Ancestry License

- **Public data**
  - Numerically-coded data
  - For education and scholarly research

- **Restricted data**
  - Alphabetic strings available through institutional license with data security plan

MPC


Number of person records in historical American census data
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1900
1910
1920
1930
1940
0
20,000,000
40,000,000
60,000,000
80,000,000
100,000,000
120,000,000
140,000,000
Number of person records

IPUMS

Data Improvements

- New variables: 1860 - 1930
- Variable coding
- Data cleaning
- Missing data allocation
- IPUMS constructed variables

MPC

IPUMS

New Variables – Housing

Housing Characteristics

- Farm residence
- Ward
- House number and street
- Homeownership
- Mortgage status
- Value of home/cost of rent

MPC

IPUMS

Demographic and Health Variables

Demographic

- Birth month
- Married in year
- Age at first marriage
- Number of times married and duration of marriage
- Children ever-born and children surviving

Disability

- Deaf, blind
- Sickness

MPC

IPUMS

Socioeconomic Variables

Wealth variables

- Value of real estate and personal property

Work

- Occupation
- Industry
- Class of worker

Education

- School attendance
- Literacy

MPC

IPUMS

Data Availability

Public	Restricted
1790-1840 - soon	1790-1840
1850 – soon	1850
1860 - TBA	1860
1870 - TBA	1870
1880 - now	1880
1900 - TBA	1900
1910 - TBA	1910
1920 – mid 2016	1920
1930 - early 2016	1930
1940 - now	1940

MPC





- Census Longitudinal Infrastructure Project (CLIP)
- National Historical Census Files Project

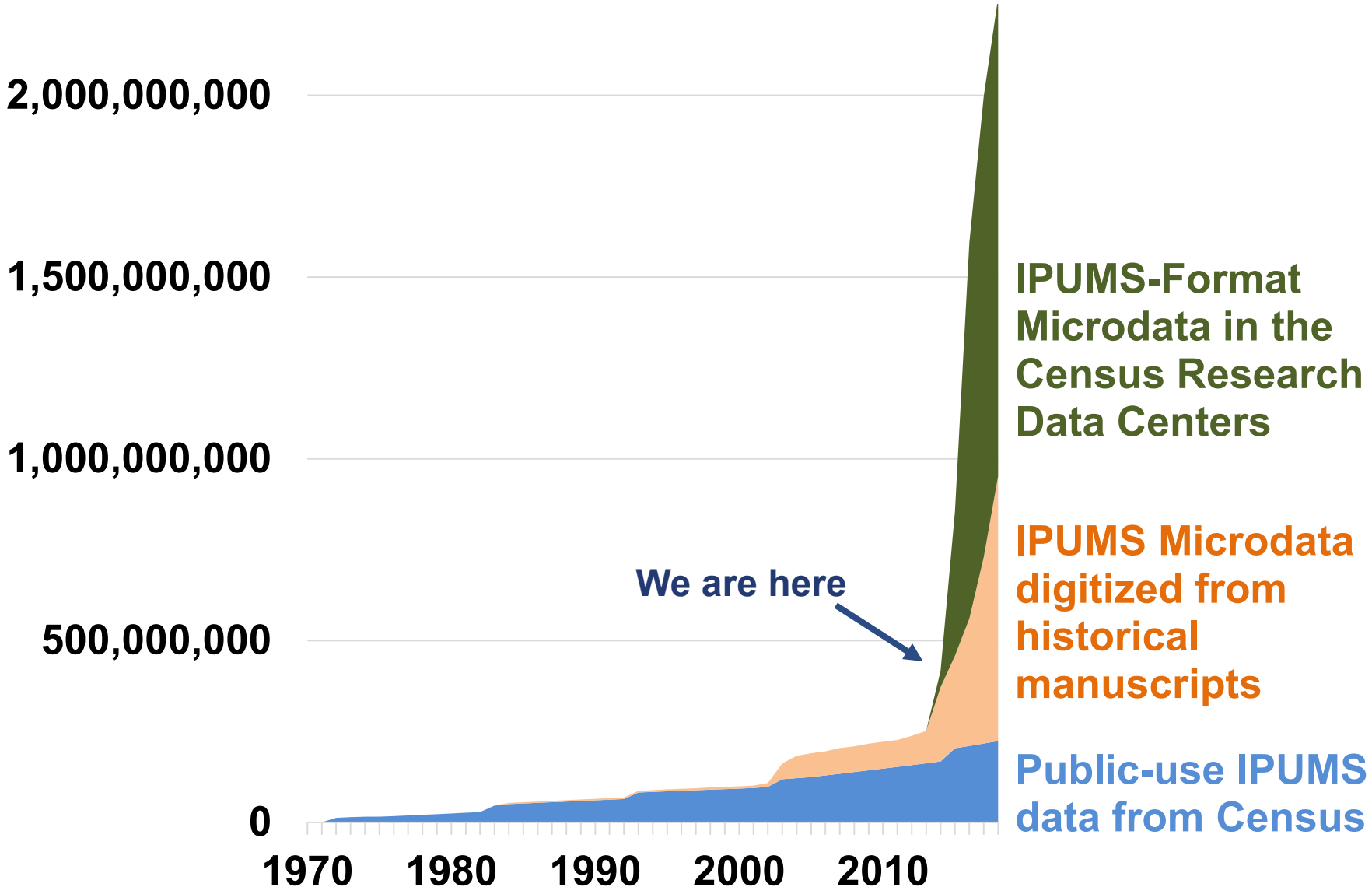
Census Longitudinal Infrastructure Project (CLIP)

1. PIK the 1940 Census (Protected Identification Key)
2. Create a longitudinal resource, linking 1940 census to later censuses, surveys (e.g. CPS, SIPP, NHIS), and administrative records (e.g. Social Security, Medicare, Medicaid)
3. Disseminate CLIP through the RDC network

National Historical Census Files Project

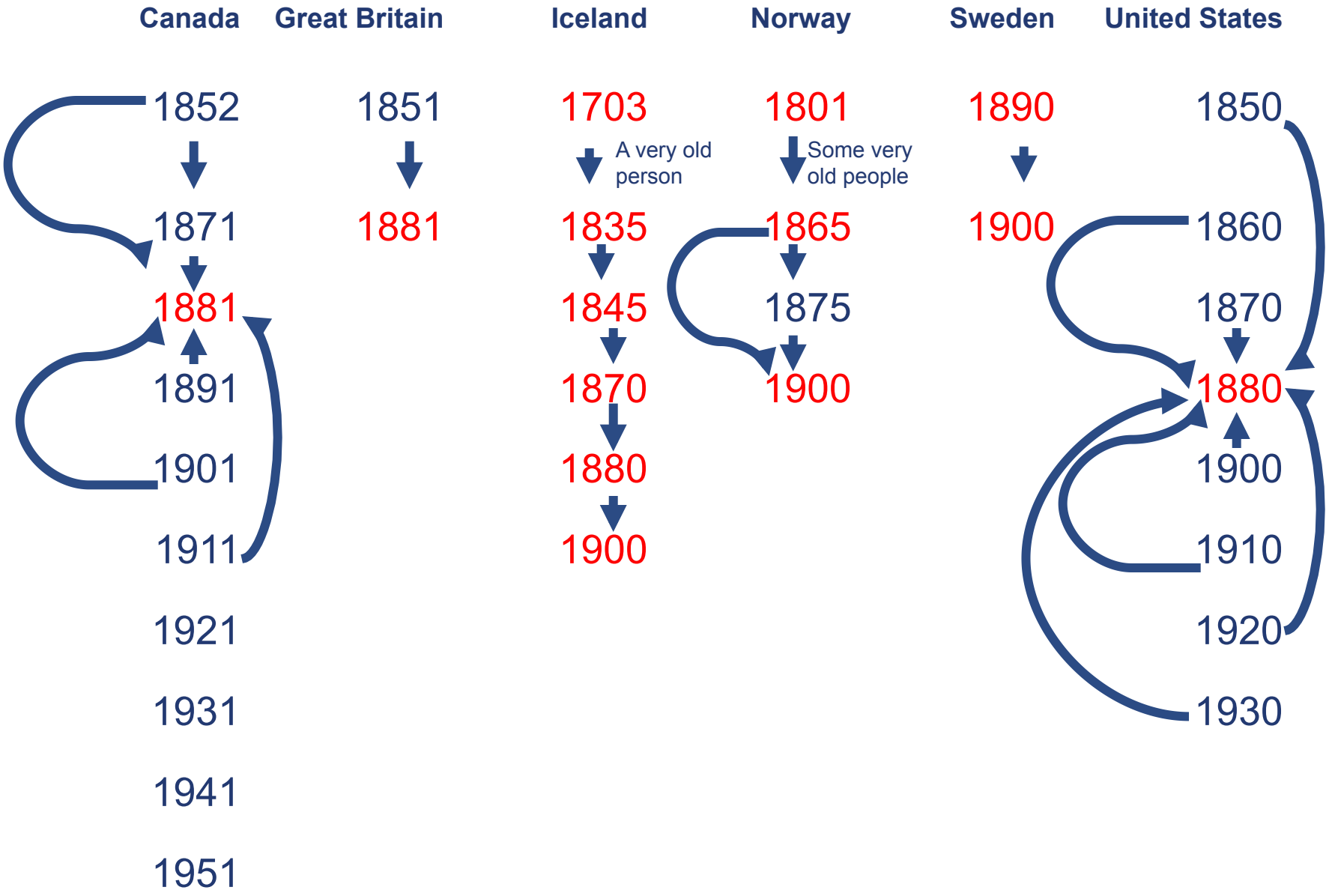
1. Recover and verify of 1960, 1970, 1980, and 1990 long-form and short-form microdata
2. Create an IPUMS-compatible version of internal census microdata files from 1960-present (including ACS)
3. Create new PUMS and summary files for 1960

Integrated U.S. microdata available for research, 1970-2018
(number of person records)

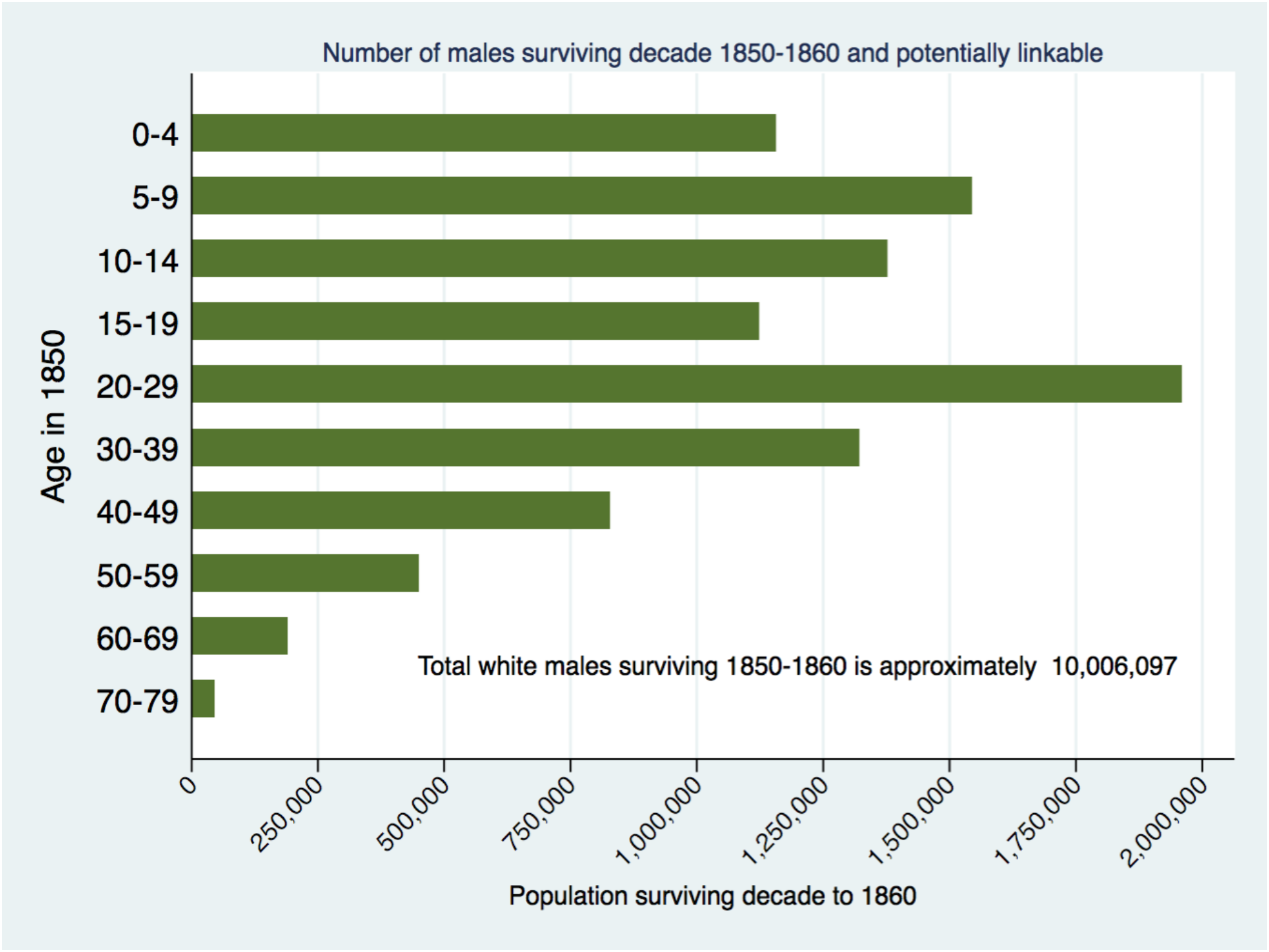


Creating linked samples

2003 onwards – linking samples to complete count datasets




Number of person records in historical American census data
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1900
1910
1920
1930
1940
0
20,000,000
40,000,000
60,000,000
80,000,000
100,000,000
120,000,000
140,000,000
Number of person records


Number of males surviving decade 1850-1860 and potentially linkable
Age in 1850
0-4
5-9
10-14
15-19
20-29
30-39
40-49
50-59
60-69
70-79
Total white males surviving 1850-1860 is approximately 10,006,097
0
250,000
500,000
750,000
1,000,000
1,250,000
1,500,000
1,750,000
2,000,000
Population surviving decade to 1860

Linking processes

Blocking strategy

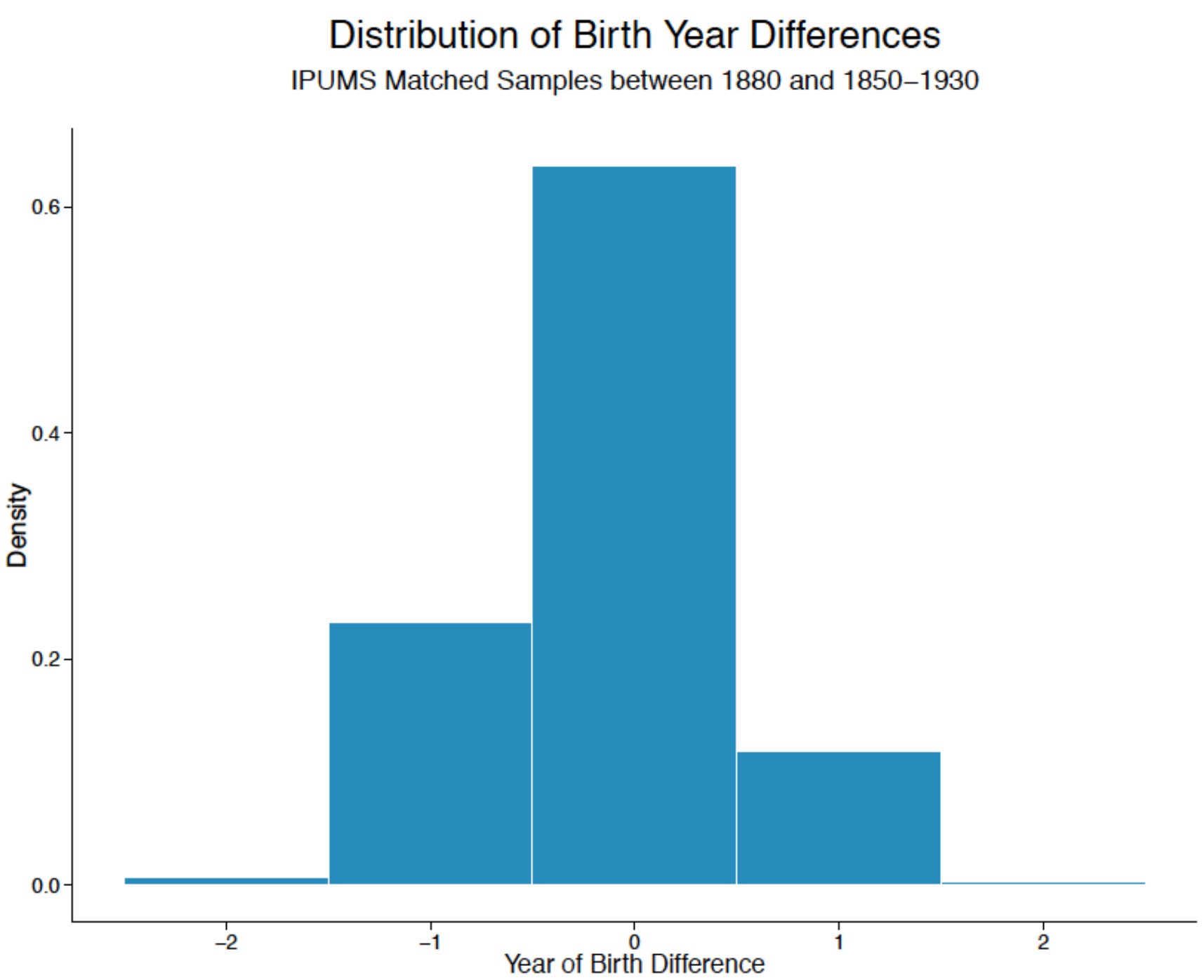
- Sex of individual or dyad
- Race (U.S.)
- Birthplaces
  - US / Canadian data
    - State or province for native born
    - Foreign country for migrants
  - Scandinavian / British
    - Parish of birth [human memory is poor] for natives
    - Foreign country for migrants
  - For all countries: some random error with more or less precision than required by enumeration

Name cleaning

- Remove stray characters (non-alpha), titles, middle initials
- Parse into separate fields for first and last names
- Some standardization of names for common names with variants and abbreviations [practice has varied over time …]
  - Tom / Thomas
  - Wm. / William
  - Edward / Theodore
- Original name retained as separate variable

Age windows

- Integer age at census transformation into year of birth OR expected age at next census has inherent error
  - Month of enumeration varies
  - Age heaping from numeracy / enumeration practices
- Potential links restricted to age window of +/- 4 to 5 years
  - Have experimented with up to 7


Distribution of Birth Year Differences
IPUMS Matched Samples between 1880 and 1850–1930
Density
0.6
0.4
0.2
0.0
−2
−1
0
1
2
Year of Birth Difference

Comparison of names

- Jaro-Winkler scores calculated for
  - First name pairs
  - Standardized first name pairs
  - Last name pairs

Jaro-Winkler string comparator

$$\mathrm{sim}_{\mathrm{jaro}}(s_1, s_2) = \frac{1}{3|}\left(\frac{c}{|s_1|} + \frac{c}{|s_2|} + \frac{c-t}{c}\right).$$

Where s1 and s2 are the two strings

e.g. Knutsen Haugen and Knudtson

c : Number of characters that are the same in half the length of the longer string

t: Number of transpositions of adjacent characters

Jaro-Winkler string comparator

$$\mathrm{sim}_{\mathrm{winkler}}(s_1, s_2) = \mathrm{sim}_{\mathrm{jaro}}(s_1, s_2) + (l.0 - \mathrm{sim}_{\mathrm{jaro}}(s_1, s_2))\frac{p}{10},$$

Winkler modification increases the score (similarity) if beginning is similar and differences appear later.

p: Number of agreeing characters at start,

$0 \leq p \leq 4$

Composite score

- Jaro-Winkler scores summed to create composite score
- Pairs of records ≥ threshold value written out
  - Decision: What is threshold

Training data

- Clerical review of pairs of links to mark records as true or false links
- Use household information to confirm which links are true

Additional indicator variables

- NYSIIS match
- Last name matches on double metaphone
- Is first name match a pair of initials?
- Middle initial present in both records

Age scores

- age70pct

  $1 - |age_1 - age_2| / [\max\{age_1, age_2\}/0.71]$

  Intuition: Deviations at younger ages more important
- mrecage_plus_norm
  Classifies age from earlier year
  Floor{ (age/5) + 1} / 100
- agediff_norm_abs
  $1 / (|age_1 - age_2| + 1)$

Support vector machine

- Training data provides set of true links with associated characteristics on variables previously described
- SVM selects best combination of data from potential links
- Select positive confidence records (we think they're a match) that don't conflict with matches to another record

Steps in process

Step	Format	Software
Edit and format input data Blocks Expected ages	Starts as IPUMS files Fixed width Coded	Statistics package Perl tools on ASCII files (on MPC servers)
Name score algorithm	Delimited files	Python (on MSI computer)
Re-attach additional variables	Delimited files	Python (on MSI computer)
Run data through classifier	Delimited files	Libsvm (MSI)
Evaluate links		Statistics package
Post-linking filtering		Statistics package
Format dataset		Statistics package
Make weights		Statistics package
Make documentation		Text editor / Word

Acknowledgements

National Institutes of Health
National Science Foundation
Ancestry.com
Family Search
U.S. Census Bureau

MPC

Minnesota Population Center

(NIH/NICHD R24HD041023)