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

Evan Roberts
Minnesota Population Center

MPC

Minnesota Population Center

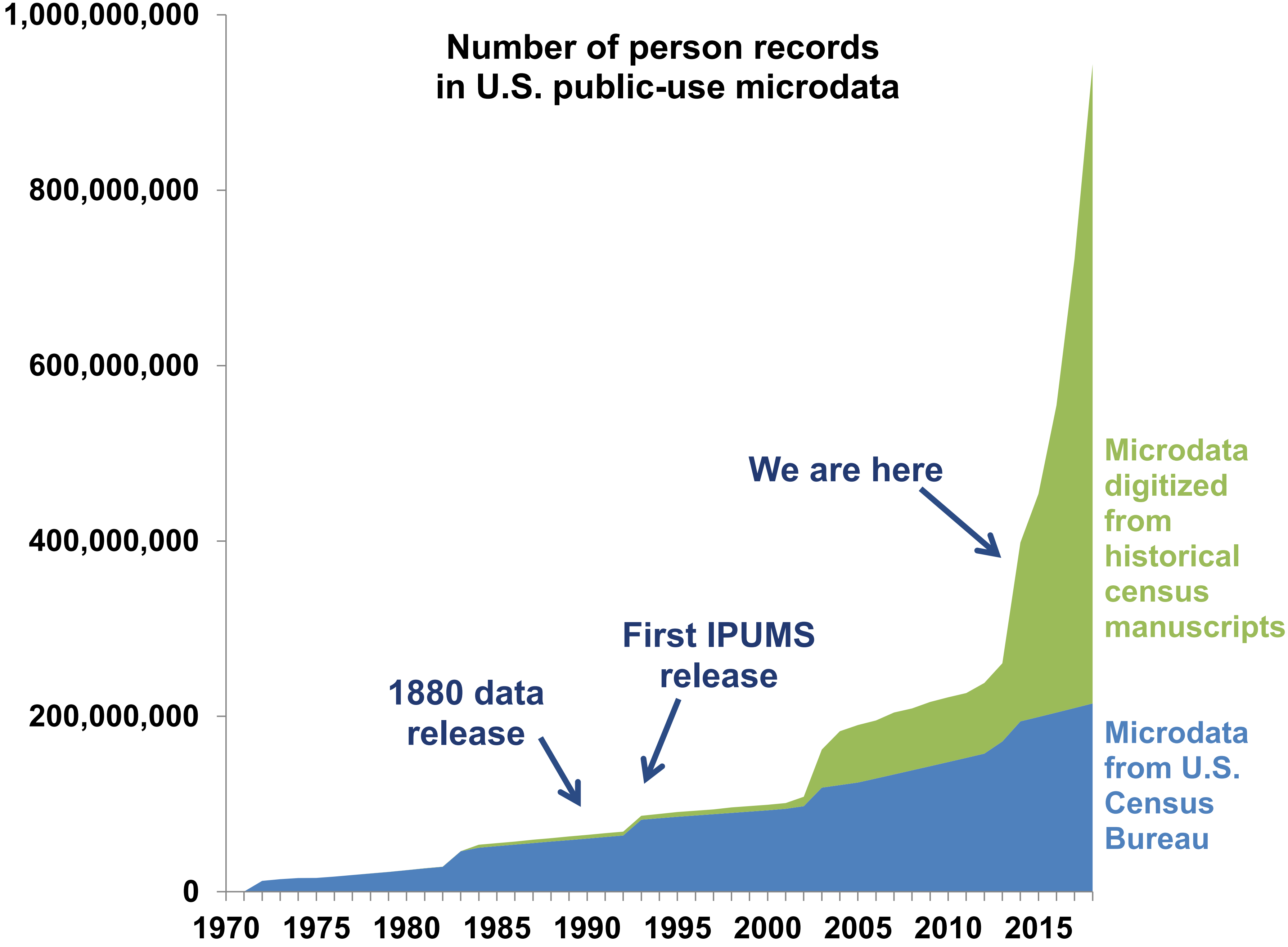
FamilySearch

MPC
Minnesota Population Center

ancestry.com®

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



Number of person records in U.S. public-use microdata

5

st. No. 2

ist. No. 96

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 I

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

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

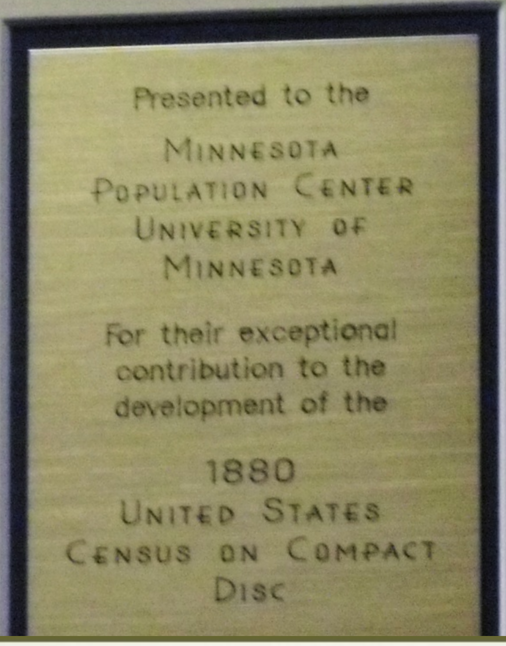
enumerated by me on the ______ day of June, 1880.

Chas O Hudnut

Enumerato

		Personal Description.					Civil Condition.				Occupation.		Health.						Education.			Nativity.		
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.	Color—White, W.; Black, B.; Mulatto, Mu.; Chinese, C.; 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: 3/12.	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.	Single, /.	Married, /.	Widowed, /. Divorced, D.	Married during Census year, /.	Profession, Occupation or Trade of each person, male or female.	Number of months this person has been unemployed during the Census year.	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, /.	Attended school within the Census year, /.	Cannot read, /.	Cannot write, /.	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				/			Laborer	V										New Jersey	New Jersey	New Jersey
	Phoebe	Mu	F	32		Wife		/			Keeping House									/	/	New Jersey	N J	N J
89	Johnson Clara	B	F	51					/		Keeping House									/	/	New Jersey	N J	N J
	Jane	B	F	29		Daughter	/				Servant											New Jersey	Pennsylvania	N J
	Mary E	B	F	21		Daughter	/				At Home											New Jersey	Penn a	N J
	Georgiana	Mu	F	14		Daughter	/				At Home								/			New Jersey	Penn a	N J
	Viola	Mu	F	2		G. Daughter	/															New Jersey	Penn a	N J
	Harry E	Mu	M	10/12		G. Daughter	/															New Jersey	Penn a	N J
90	Dickson James	B	M	48				/			Laborer	3V										New Jersey		
	Lucinda	B	F	48		Wife		/			Keeping House											New York		
91	Golden Ellen	W	F	55					/		Keeping House	V								/	/	Ireland	Ireland	Ireland
	Mary	W	F	17		Daughter	/				Servant	V										New Jersey	Ireland	Ireland
	Philips	W	M	20		Son	/				Laborer	V									/	New Jersey	Ireland	Ireland
92	O'Conner Ellen	W	F	50			/				Servant	V								/	/	Ireland	Ireland	Ireland

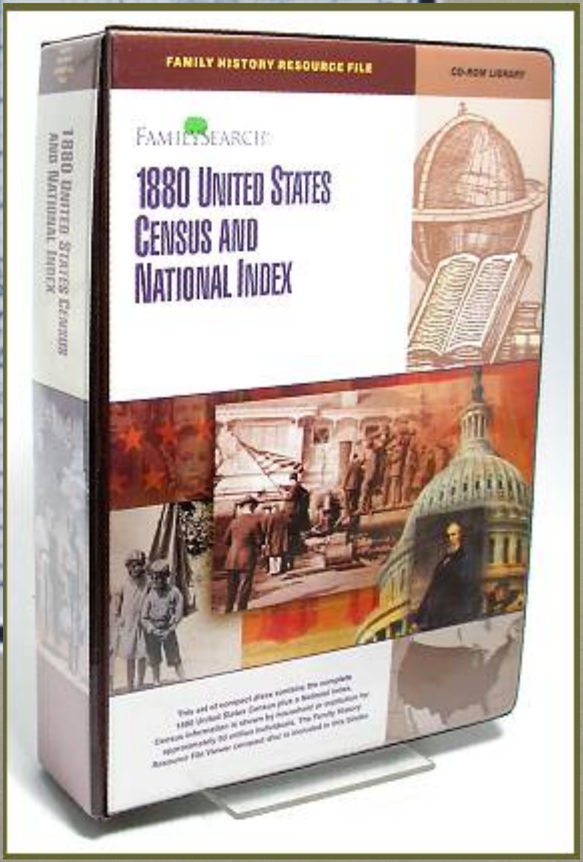
1999: Collaboration with the Latter-Day Saints



Presented to the
MINNESOTA
POPULATION CENTER
UNIVERSITY OF
MINNESOTA

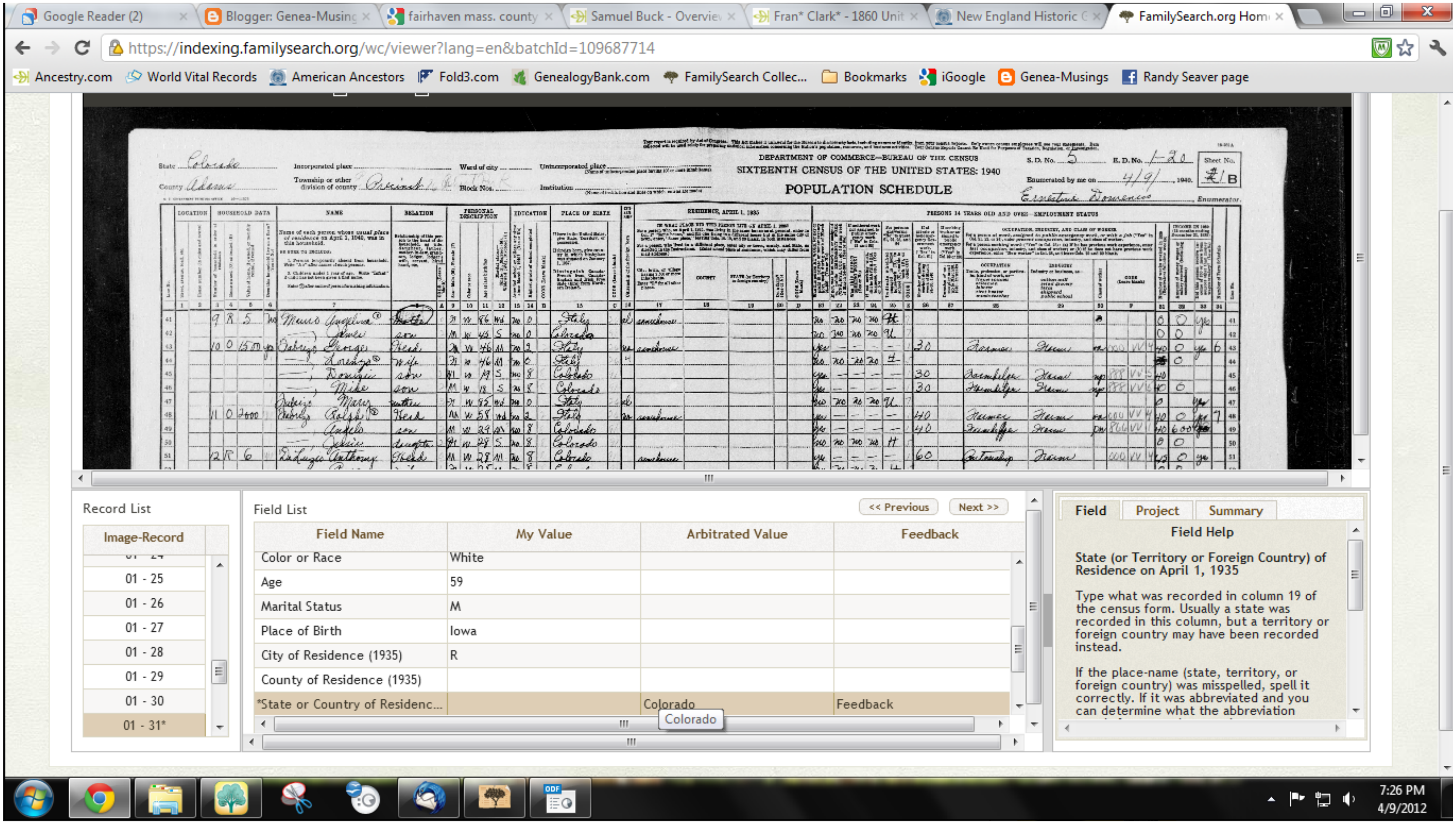
For their exceptional
contribution to the
development of the

1880
UNITED STATES
CENSUS ON COMPACT
DISC


FAMILY HISTORY RESOURCE FILE
CD-ROM LIBRARY
FAMILYSEARCH
1880 UNITED STATES CENSUS AND NATIONAL INDEX

55 CD-ROM Set

Crowdsourcing Project 2006-2011



Record List

Image-Record
01 - 25
01 - 26
01 - 27
01 - 28
01 - 29
01 - 30
01 - 31*

Field List

Field Name	My Value	Arbitrated Value	Feedback
Color or Race	White		
Age	59		
Marital Status	M		
Place of Birth	Iowa		
City of Residence (1935)	R		
County of Residence (1935)			
*State or Country of Residenc...		Colorado	Feedback

Colorado

Field Help

State (or Territory or Foreign Country) of Residence on April 1, 1935

Type what was recorded in column 19 of the census form. Usually a state was recorded in this column, but a territory or foreign country may have been recorded instead.

If the place-name (state, territory, or foreign country) was misspelled, spell it correctly. If it was abbreviated and you can determine what the abbreviation

9339

Form 15-6

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

FIFTEENTH CENSUS OF THE UNITED STATES: 1930

POPULATION SCHEDULE

State Illinois

County Cook

Township or other division of county Precinct 5
(Insert proper name and also name of class, as township, town, precinct, district, etc. See instructions.)

Incorporated place Chicago City
(Insert proper name and also name of class, as city, village, town, or borough. See instructions.)

Ward of city 41 Block No. 58

Unincorporated place
(Enter name of any unincorporated place having approximately 500 inhabitants or more. See instructions.)

Institution
(Insert name of institution, if any, and indicate the lines on which the entries are made. See instructions.)

Enumeration District No. 16-2772

Supervisor's District No. 3

Sheet No. 9 B

Enumerated by me on April 7, 1930, Irma L Burroughs, Enumerator.

PLACE OF ABODE: 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	NAME of each person whose place of abode on April 1, 1930, was in this family	RELATION Relationship of this person to the head of the family	HOME DATA: Home owned or rented	Value of home, if owned, or monthly rental, if rented	Radio set	Does this family live on a farm?	PERSONAL DESCRIPTION: Sex	Color or race	Age at last birthday	Marital condition	Age at first marriage	EDUCATION: Attended school or college any time since Sept. 1, 1929	Whether able to read and write	PLACE OF BIRTH: PERSON	FATHER	MOTHER	MOTHER TONGUE: Language spoken in home before coming to the United States	CODE: State or M. T.	Country	Nativity	CITIZENSHIP: Year of immigration to the United States	Naturalization	Whether able to speak English	OCCUPATION	INDUSTRY	CODE	Class of worker	EMPLOYMENT: Yes or No	If not, line number on Unemployment Schedule	VETERANS: Yes or No	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
Lunt	7217			Stearns, Allen N.	Grandson				No	M	W	5 9/12	S		Yes	No	Illinois	Missouri	Illinois		61					Yes	None								
	7201	129	129	Ketter, 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		
	6903	134	134	Schwiesow, Theodore	Head	O	5,000		No	M	W	76	M	24	No	Yes	Germany	Germany	Germany	German	13	13	V	1873	Na	Yes	None								
				Minnie	Wife-H				V	F	W	69	M	17	No	Yes	Germany	Germany	Germany	German	13	13	V	1872	Na	Yes	None								
	6947	135	135	Schwiesow, William	Head	O	10,000	R	No	M	W	45	M	33	No	Yes	Illinois	Germany	Germany		61	13	0			Yes	Plasterer	Building	04X1	E	No	35	No		
				Rose	Wife-H				V	F	W	38	M	26	No	Yes	Illinois	Bohemian	Bohemian		61	15	0			Yes	None								
	6945	136	136	Schwiesow, David	Head	O	10,000	R	No	M	W	47	M	24	No	Yes	Illinois	Germany	Germany		61	13	0			Yes	Trimmer	Roofer	37X1	W	No	57	No		
				Amanda	Wife-H				V	F	W	40	M	27	No	Yes	Illinois	Illinois	Illinois		61					Yes	None								
	6935	137	137	Olson, Osborne	Head	O	10,000	R	No	M	W	46	M	22	No	Yes	Illinois	Norway	Norway		61	05	0			Yes	Decorator	China	8394	E	Yes		No		
				Augusta	Wife-H				V	F	W	46	M	22	No	Yes	Illinois	Sweden	Germany		61	13	0			Yes	None								
				Harry	Son				V	M	W	22	S		No	Yes	Illinois	Illinois	Illinois		61					Yes	Decorator	China	8394	W	Yes		No		
				Evelyn	Daughter				V	F	W	20	S		No	Yes	Illinois	Illinois	Illinois		61					Yes	Steno.	China	7190	W	Yes				
	6925	138	138	Reiling, Charles	Head	O	4,500	R	No	M	W	64	M	24	No	Yes	Illinois	Illinois	Illinois		61					Yes	None								
				Julia	Wife-H				V	F	W	60	M	20	No	Yes	Germany	Germany	Germany	German	13	13	V	1877	Na	Yes	None								
	6923½		139	Bredfield, Henry C.	Head	R	45.	R	No	M	W	31	M	24	No	Yes	Illinois	Illinois	Norway		61	05	2			Yes	Engineer	Heating	5394	W	Yes		Yes	WW	
				Muriel	Wife-H				V	F	W	25	M	19	No	Yes	Illinois	Missouri	California		61					Yes	None								
				Marilyn	Daughter				V	F	W	3 1/12	M		No		Illinois	Illinois	Illinois		61						None								
	6919	139	140	Blankenhagen, Henry	Head	O	15,000	R	No	M	W	66	M	26	No	Yes	Illinois	Germany	Germany		61	13	0			Yes	Letter carrier	Mail	8876	W	Yes		No		
				Elizabeth	Wife-H				V	F	W	59	M	19	No	Yes	Illinois	Germany	New York		61	13	1			Yes	None								
	6915	140	141	Blankenhagen, Walter	Head	O	12,000	R	No	M	W	38	M	21	No	Yes	Illinois	Illinois	Illinois		61					Yes	Steam fitter	Heating	3028	W	Yes		No		
				Elvelda	Wife-H				V	F	W	38	M	21	No	Yes	Illinois	Norway	Norway		61	05	0			Yes	None								
				Walter	Son				V	M	W	10	S		Yes	Yes	Illinois	Illinois	Illinois		61					Yes	None								
				Elaine	Daughter				V	F	W	7	S		Yes		Illinois	Illinois	Illinois		61						None								
	6907	141	142	Heubach, Henry	Head	O	10,000	R	No	M	W	49	M	32	No	Yes	Illinois	Germany	Germany		61	13	0			Yes	Clerk	Post ofc.	7X76	W	Yes		Yes	WW	
				Lillian E.	Wife-H				V	F	W	47	M	30	No	Yes	Illinois	Illinois	Illinois		61					Yes	None								
				Henry G.	Son				V	M	W	17	S		Yes	Yes	Illinois	Illinois	Illinois		61					Yes	None								
				Robert W.	Son				V	M	W	15	S		Yes	Yes	Illinois	Illinois	Illinois		61					Yes	None								
				Edward W.	Son				V	M	W	12	S		Yes	Yes	Illinois	Illinois	Illinois		61					Yes	None								
Ottawa	6906	142	143	Bretscher, John M.	Head	O	12,000	R	No	M	W	28	S		No	Yes	Wisconsin	Missouri	Michigan		63					Yes	Realtor	Real Estate	8286	E	Yes		No		
				Marie	Mother-H				V	F	W	57	Wd		No	Yes	Michigan	Missouri	Michigan		62					Yes	None								
				Hutchinson, Henry	Bro-in-law				V	M	W	38	M	31	No	Yes	Ohio	Ohio	Virginia		59					Yes	Salesman	Insurance	8885	W	Yes		Yes	WW	
				Margaret	Sister				V	F	W	28	M	27	No	Yes	Wisconsin	Missouri	Michigan		63					Yes	Steno.	Exports	7190	W	Yes				
	6910	143	144	Hoolehan, Carl D.	Head	R	60.	R	No	M	W	40	M	20	No	Yes	Illinois	Illinois	Illinois		61					Yes	Electrician	Wiring	1V81	W	Yes		No		

Form 15-6

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

FIFTEENTH CENSUS OF THE UNITED STATES: 1930

POPULATION SCHEDULE

ABBREVIATIONS TO BE USED IN COLUMNS INDICATED:
[Use no abbreviations for State or country of birth or for mother tongue (Columns 18, 19, 20, and 21)]

Col. 6—Indicate the home-maker in each family by the letter "H," following the word which shows the relationship, as "Wife—H"

Col. 7—Owned ... O
Rented ... R

Col. 9—Radio set ... R
Make no entry for families having no radio set.

Col. 11—Male ... M
Female ... F

Col. 12—White ... W
Negro ... Neg
Mexican ... Mex
Indian ... In
Chinese ... Ch
Japanese ... Jp
Filipino ... Fil
Hindu ... Hin
Korean ... Kor
Other races, spell out in full

Col. 14—Single ... S
Married ... M
Widowed ... Wd
Divorced ... D

Col. 23—Naturalized ... Na
First papers ... Pa
Alien ... Al

Col. 27—Employer ... E
Wage or salary worker ... W
Working on own account ... O
Unpaid worker, member of the family ... NP

Col. 31—World War ... WW
Spanish-American War ... Sp
Civil War ... Civ
Philippine Insurrection ... Phil
Boxer Rebellion ... Box
Mexican Expedition ... Mex

ENTRIES ARE REQUIRED IN THE SEVERAL COLUMNS AS FOLLOWS:

Cols. 6, 11, 12, 13, 14, 16, 18, 19, 20, and 25—For all persons.
Cols. 7, 8, 9, and 10—For heads of families only. (Col. 8 requires no entry for a farm family.)
Col. 15—For married persons only.
Col. 17—For all persons 10 years of age and over.
Cols. 21, 22, and 23—For all foreign-born persons.
Col. 24—For all persons 10 years of age and over.
Cols. 26, 27, and 28—For all persons for whom an occupation is reported in Col. 25.
Col. 30—For all males 21 years of age and over.

1940 Census Project

Collaboration of MPC and Ancestry.com

- 7.8 billion keystrokes
- 134 million persons

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 census reports. Only sworn census employees will see your statements. Data collected will be used solely for preparing statistical information concerning the Nation's population, resources, and business activities.

16-222 A

DEPARTMENT OF COMMERCE—BUREAU O

SIXTEENTH CENSUS OF THE UNITE

POPULATION SCHED

State New York | Incorporated place ~~New York~~ Bronx Borough | Ward of city + | Unincorporated place .... (Name of unincorporated place having 100 or more inhabitants)

County Bronx | Township or other division of county A.D. 1 | Block Nos. A' B' C' E F | Institution .... (Name of institution and lines on which entries are made)

E. D. No. 3-102 | Sheet No. 1 A

April 2, 1940.

Maynihan Enumerato

U. S. GOVERNMENT PRINTING OFFICE 16—11576

LOCATION		HOUSEHOLD DATA	NAME	RELATION	PERSONAL DESCRIPTION	EDUCATION	PLACE OF BIRTH	CITIZENSHIP	RESIDENCE, APRIL 1, 1935
Street, avenue, road, etc.	House number (in cities and towns)		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.	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.			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-		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.)
1	2								
AVE	10.								

21	22	23	24	25	E		F	31	32	33	34
YES	–	–	–	–	1		266 24 1	52	3280	No	

EDUCATION

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)		For persons answering "No" to quest. 21, 22, 23, and 24		If at private or nonemergency Government work. ("Yes" in Col. 21)	If seeking work or assigned to public emergency work. ("Yes" in Col. 22 or 23)	OCCUPATION, INDUSTRY, AND CLASS OF WORKER						Number of weeks worked in 1939 (Equivalent full-time weeks)
		Was this person SEEKING WORK? (Yes or No)	If not seeking work, did he HAVE A JOB, business, etc.? (Yes or No)	Indicate whether engaged in home housework (H), in school (S), unable to work (U), or other (Ot)	CODE	Number of hours worked during week of March 24–30, 1940	Duration of unemployment up to March 30, 1940—in weeks	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)			
21	22	23	24	25	E	26	27	28	29	30	F			31
YES	–	–	–	–	1	40		TIMEKEEPER	CUT STONE YARD	P.W	266	24	1	52
YES	–	–	–	–	1	40	~~25~~	HELPER	CUT STONE YARD	P.W	496	24	1	40
No	No	No	No	H	5									0
B														
YES	–	–	–	–	1	48		SUPERINTENDANT LOFT BUILDING	PHILLIP JONES SHIRT COMPANY	P.W	128	81	1	52
No	No	No	No	H	5									

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.

SUPPL. QUEST.

FORMER FINISHED MRS. TERRI THER.

SUPPL. QUEST.

INCOME IN 1939 (12 months ending December 31, 1939)

Amount of money wages or salary received (including commissions) | Did this person receive income of $50 or more from sources other than money wages or salary? (Yes or No) | Number of Farm Schedule | Line No.

33
No
No
No
No
No
No
YES
No
No
No
No
No
No
No
No
No
No
No
No
No
No
No
0
No
No
No
No
No
No
No
No
No
No
No

For Persons

SAME HOUSE

Line No.	35	FATHER 36	MOTHER 37	CODE (Leave blank) G	38	CODE (Leave blank) H	39	40	41	CODE (Leave blank) I	42	43	44	USUAL OCCUPATION 45	USUAL INDUSTRY 46	Usual class of worker 47	CODE (Leave blank) J	Has this woman been married more than once? (Yes or No) 48		U	V	W	X	Y	Z
14	IMOR- AUGUSTUS	NEW YORK	NEW YORK		ENGLISH		No	No	No		No	–	-0											2	
29	TENKE VILMA	YUGOSLAVIA	GERMANY	17	ENGLISH		No	No	No		No	–	-0	NONE	STUDENT	S.			0 7 2 1 16 1 8 7		0	00	0	2	

CLASS OF WORKER: occupation and at which he is physically able to work. Enter also usual industry in Col. 45 and leave Cols. 46 and 47

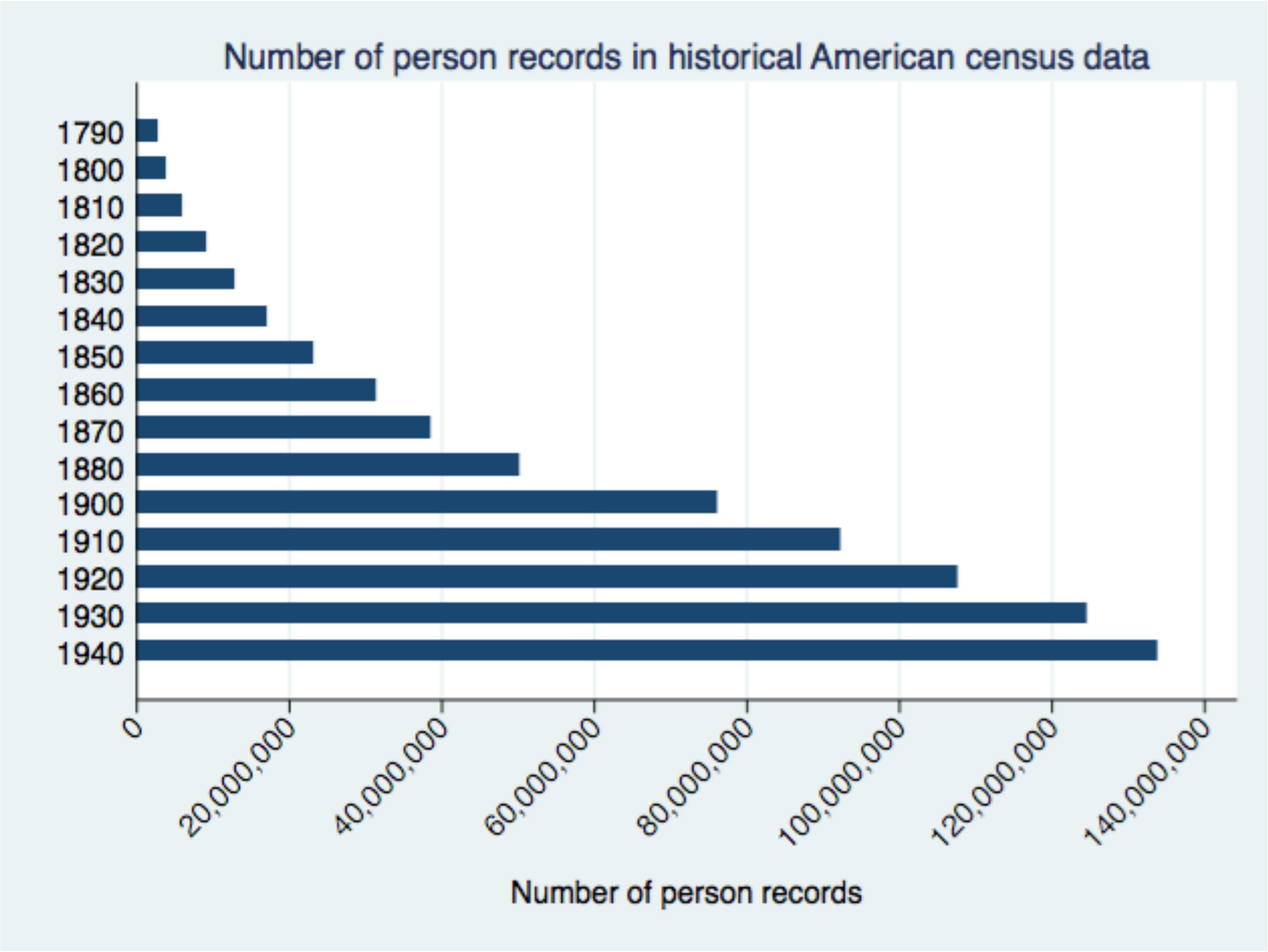
NO	H4	30
NO	H-4	36

Col. 5. VALUE OF HOME, IF OWNED: | 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:

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


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

Data Improvements

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

New Variables – Housing

Housing Characteristics

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

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

Socioeconomic Variables

Wealth variables

- Value of real estate and personal property

Work

- Occupation
- Industry
- Class of worker

Education

- School attendance
- Literacy

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

United States™ Census Bureau

MPC Minnesota Population Center

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

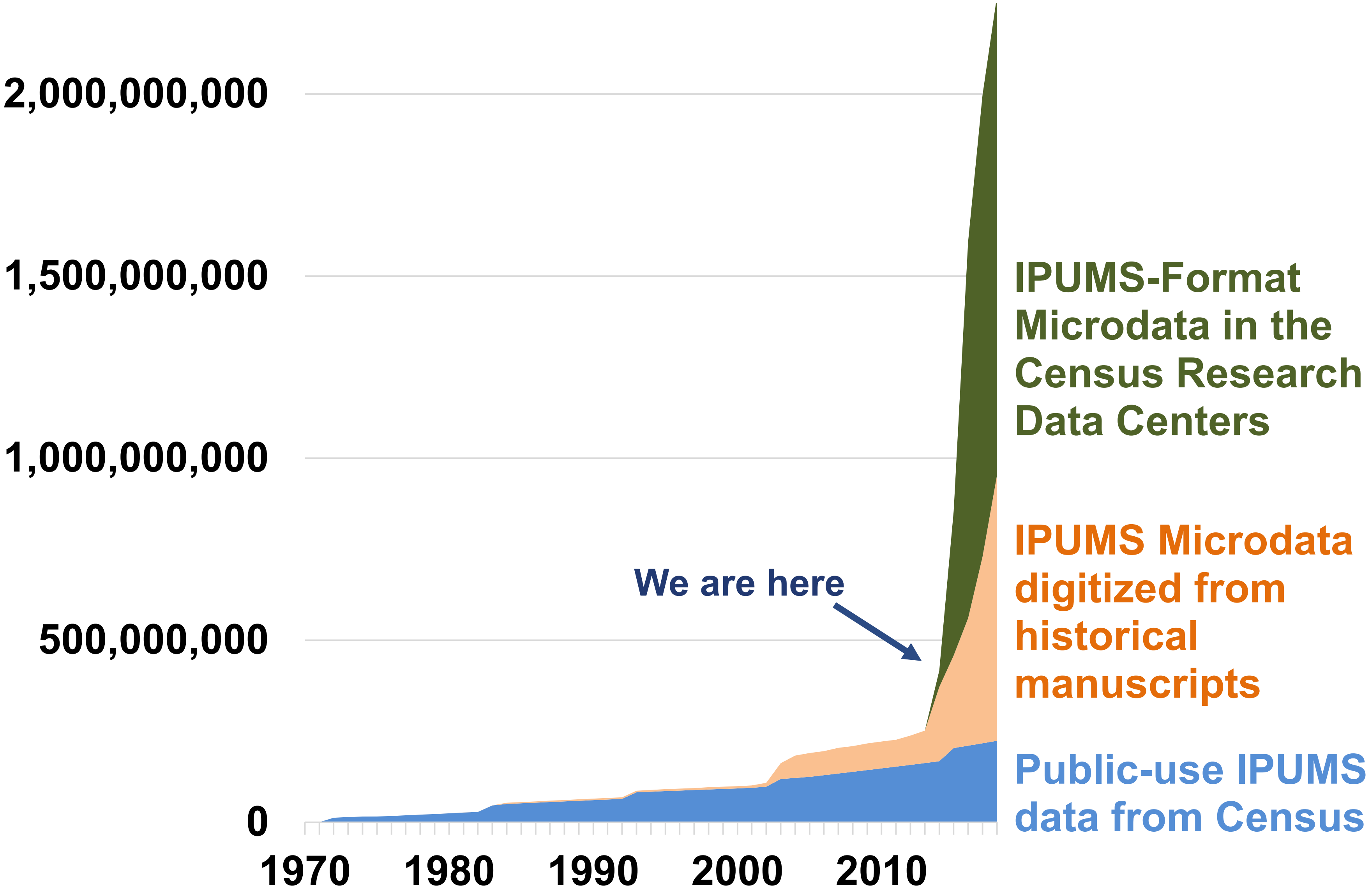
Census Longitudinal Infrastructure Project (CLIP)

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

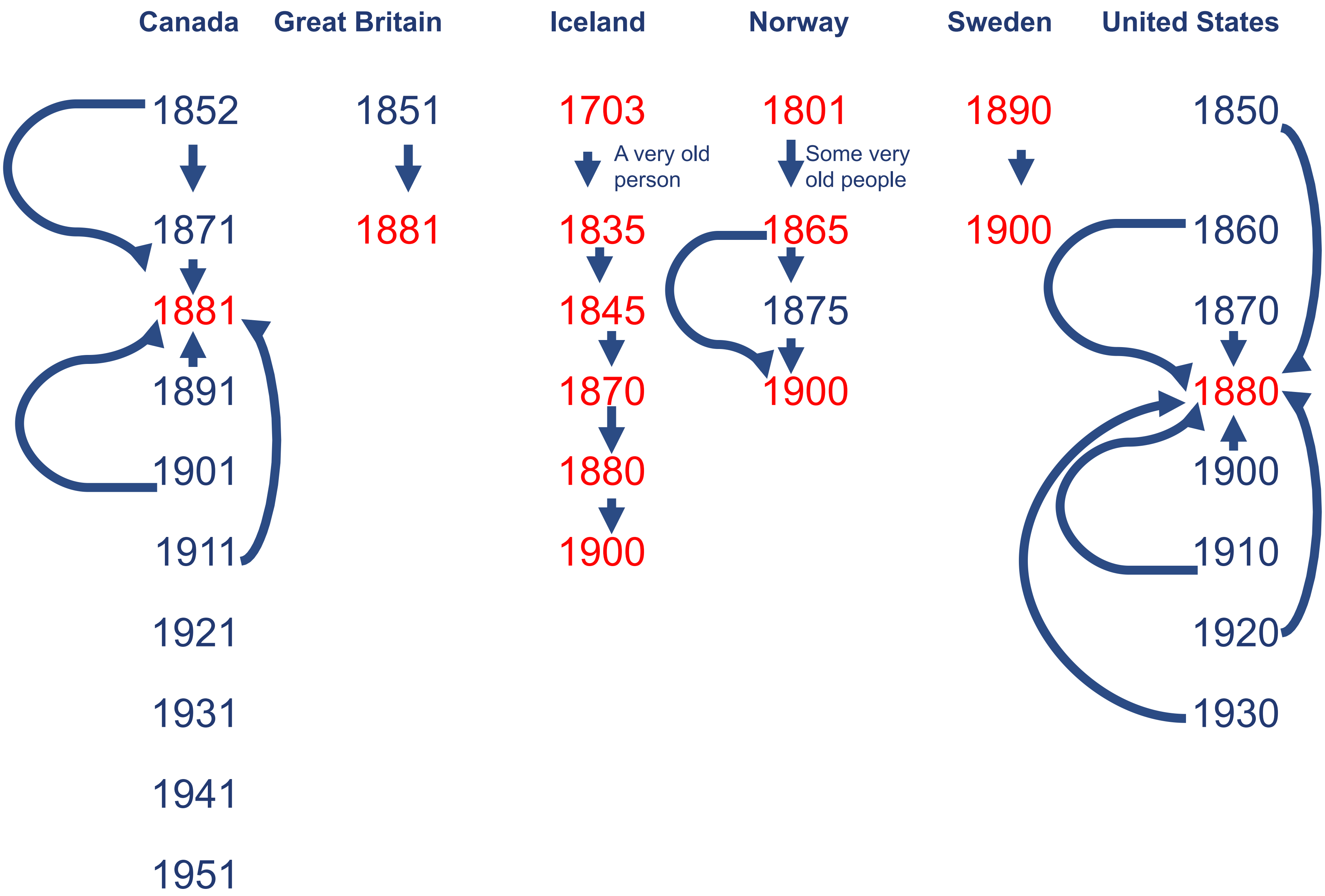
1. Recover and verify of 1960, 1970, 1980, and 1990 long-form and short-form microdata
2. Create an IPUMS-compatible version of internal census microdata files from 1960-present (including ACS)
3. Create new PUMS and summary files for 1960

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

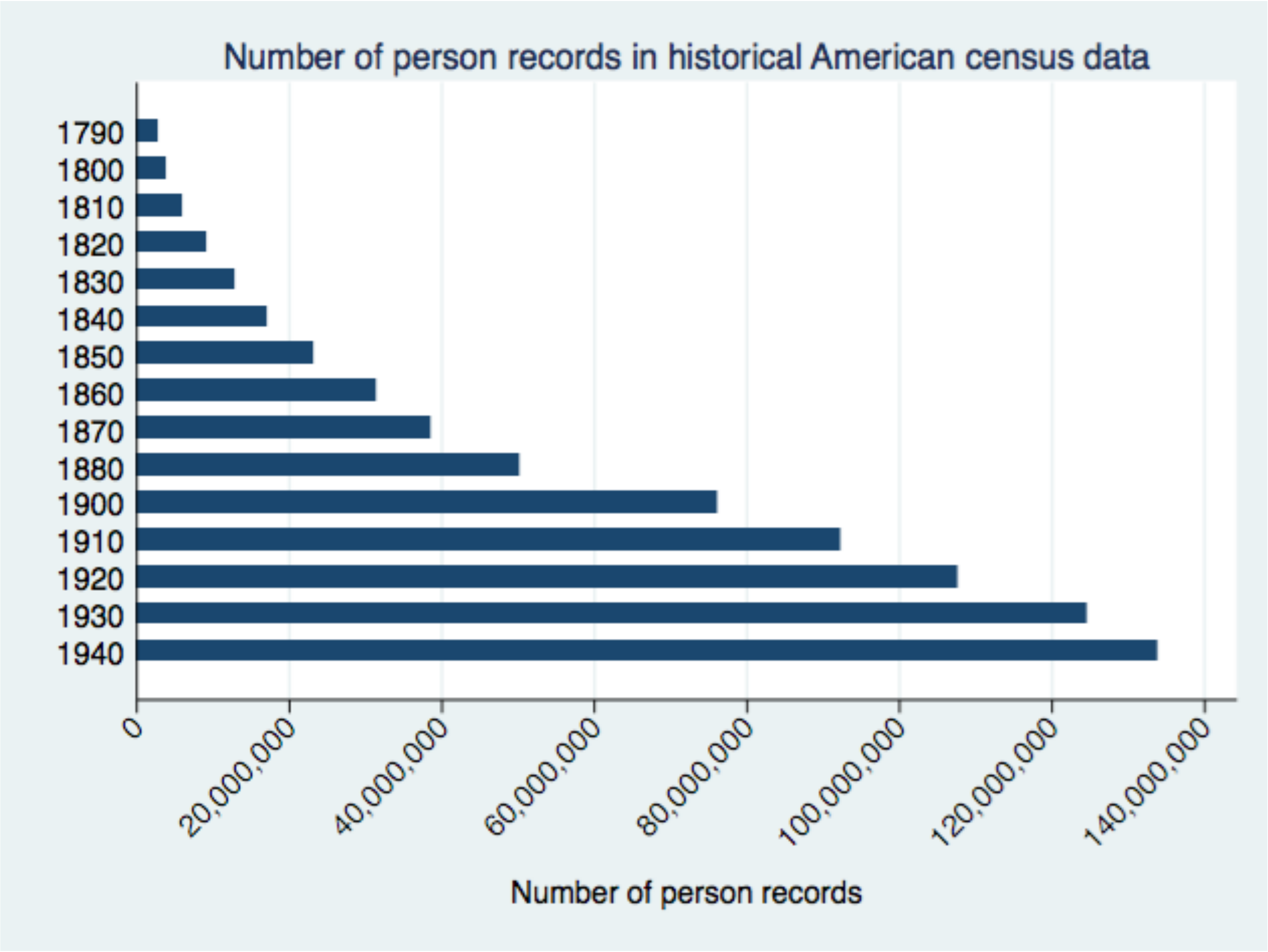

We are here

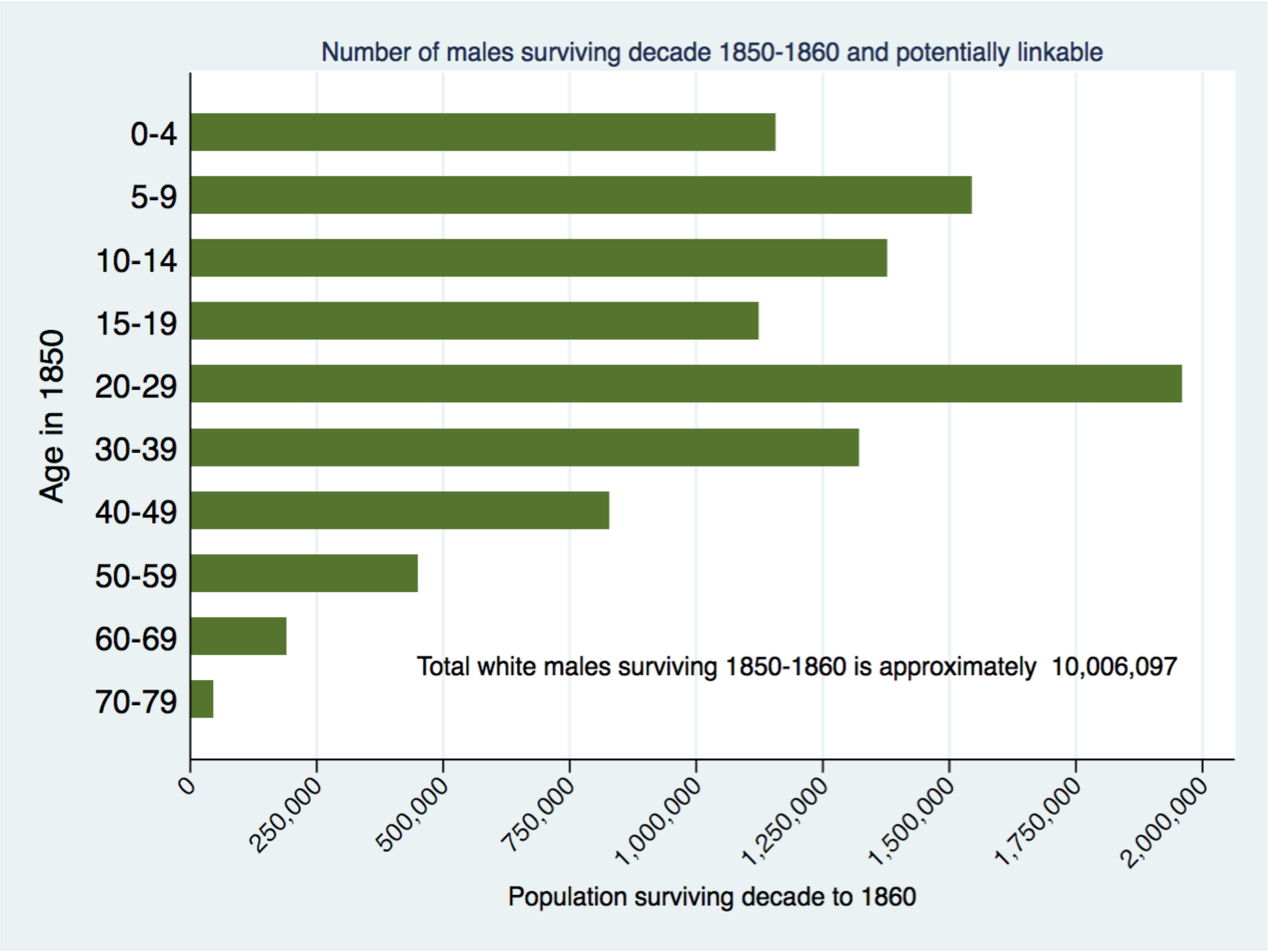
Creating linked samples

2003 onwards – linking samples to complete count datasets



Canada	Great Britain	Iceland	Norway	Sweden	United States
1852	1851	1703	1801	1890	1850
		A very old person	Some very old people		
1871	1881	1835	1865	1900	1860
1881		1845	1875		1870
1891		1870	1900		1880
1901		1880			1900
1911		1900			1910
1921					1920
1931					1930
1941					
1951					


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

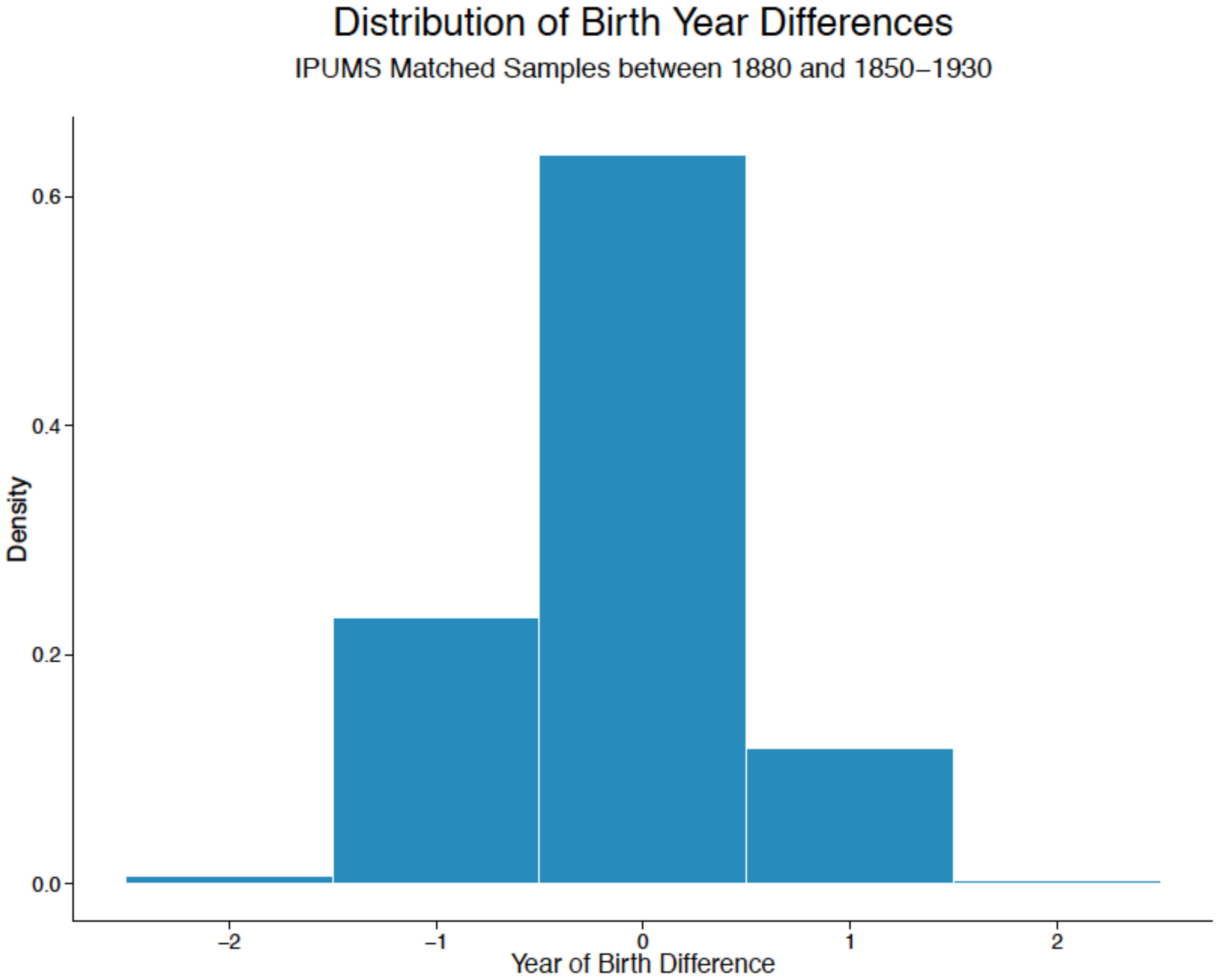
- Sex of individual or dyad
- Race (U.S.)
- Birthplaces
  - US / Canadian data
    - State or province for native born
    - Foreign country for migrants
  - Scandinavian / British
    - Parish of birth [human memory is poor] for natives
    - Foreign country for migrants
  - For all countries: some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

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

$$\mathrm{sim}_{\mathrm{winkler}}(s_1, s_2) = \mathrm{sim}_{\mathrm{jaro}}(s_1, s_2) + (1.0 - \mathrm{sim}_{\mathrm{jaro}}(s_1, s_2))\frac{p}{10},$$

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)