

When ML Meets Historical Data: Automated Data Linking and Its Application

Sun Kyoung Lee (University of Michigan)
Census FedCasic Conference: Advances in Data Science

Data Science

- the study of data to extract meaningful insights

Data Science: Life Cycle



Today's main focus

Roadmap

- Machine Learning-based Record Linking
- Specific Examples of Record Linking
- Comparison w/ Existing Record Linking Techniques
- Quality of Matches

Machine Learning Based Record Linkage

- Supervised algorithm: requiring a **training data**
- Key Insights: simultaneously/jointly considering linking criteria
 - This allows algorithms to resolve some cases where deterministic algorithms like Ferrie (1996) and Abramitzky Boustan Eriksson (2012, 2014) cannot
- Examples: Support Vector Machine (IPUMS), XGBoost (ancestry.com), Random Forest Classifier
- **Strength:** accuracy, scalability, ease of tuning

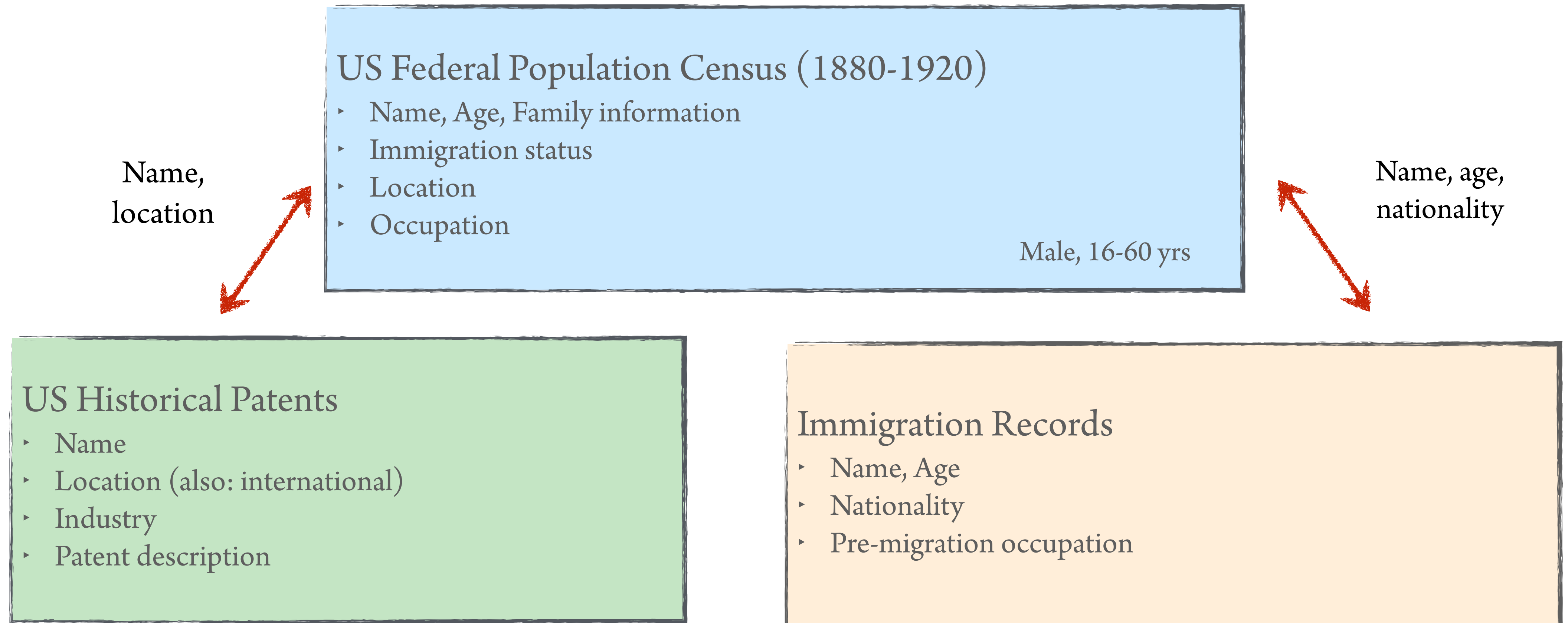
Roadmap

- Machine Learning-based Record Linking
- Specific Examples of Record Linking
- Comparison w/ Existing Record Linking Techniques
- Quality of Matches

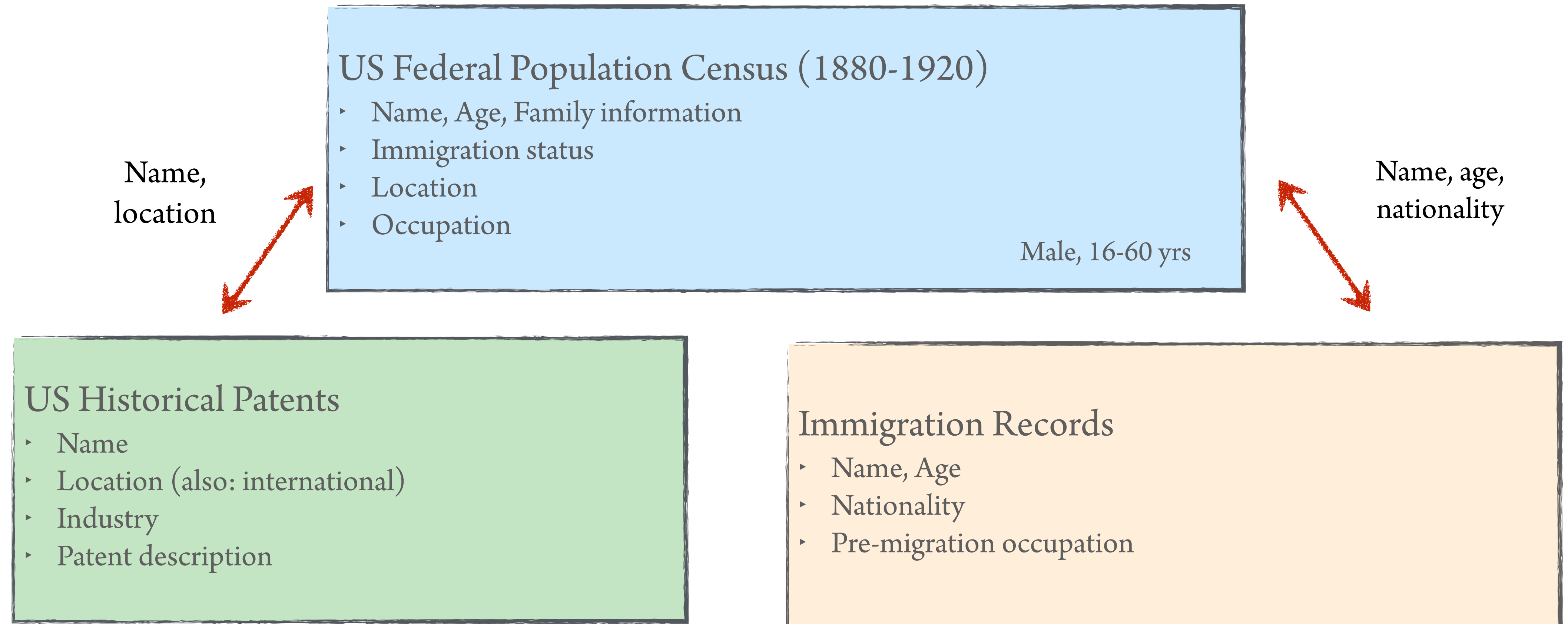
Record Linking: Application

- **Immigration, Innovation, and Urban Hubs:** Theory and Evidence from the Age of Mass Migration, by Costas Arkolakis, Sun Kyoung Lee, Michael Peters, Working Paper, April 2024
- The role of European Immigration on local and aggregate economic growth in the United States
- Big data and machine learning approach to link individual-level data:
 1. US Population census
 2. Universe of US patents
 3. Universe of immigration records

Data Construction: Immigrants & Innovation



Data Construction: Immigrants & Innovation



Final Data Set: 1880-1920

Total Population	Foreign born	Matched Patents	Matched Immigration Records
97,958,009	20,790,886	450,917	633,431

Data Construction: Immigrants & Their Post-Arrival Records

Population Census (1880-1920)

- Name, Age, Family information
- Immigration status
- Location
- Occupation and Industry

Male, 16-60

Name, age,
nationality



Immigration Records

- Name
- Nationality
- Pre-migration occupation

Form 500 B
Department of Commerce and Labor
IMMIGRATION SERVICE

SALOON, CABIN, AND STEERAGE ALIENS MUST BE COMPLETELY MANIFESTED.

LIST OR MANIFEST OF ALIEN PASSENGERS FOR THE UNITED STATES
Required by the regulations of the Secretary of Commerce and Labor of the United States, under Act of Congress approved February 20, 1907, to be delivered

16 S.S. *Thimistolis* sailing from *Piræus* *2 October*, 1909

No. on List.	NAME IN FULL.		Age. Yrs. Mos.	Sex. Married or Single.	Calling or Occupation.	Able to—		Nationality. (Country of which citizen or subject.)	† Race or People.	* Last Permanent Residence.		The name and complete address of nearest relative or friend in country whence alien came.	Final Destination. *(Intended future permanent residence.)	
	Family Name.	Given Name.				Read.	Write.			Country.	City or Town.		State.	City or Town.
1	Seledakis	Constantine	22	m	Bookkeeper	yes		Greek	Greece	Crete	Rethymno	John his father Rethymno	N. J.	New York
2	Geombra	Stergios	19	m	Workman	yes		Greek	Greece	Macedonia	Grevena	John his father Grevena	N. H.	Manchester
3	Seadakis	John	40	m	Workman	yes		Greek	Greece	Crete	Apostolonia	Maria his wife Apostolonia	Utah	Clear Creek
4	Seadakis	Constantine	34	m	Workman	yes		Greek	Greece	Crete	Apostolonia	Vasiliki his wife Apostolonia	Utah	Clear Creek
5	Vezerys	John	22	m	Workman	yes		Turkey	Greece	Thessalia	Chios	Nick his father Chios	N. J.	New York
6	Mufas	George	22	m	Workman	yes		Turkey	Greece	Thessalia	Ypsos	Liss his father Ypsos	N. H.	New York
7	Artilakis	Saras	28	m	Workman	no		Greek	Greece	Crete	Rethymno	Argyna his wife Rethymno	N. J.	New York
8	Trimboucius	Carlos	35	m	Workman	yes		Greek	Greece	Crete	Rethymno	Maggelin his wife Rethymno	N. J.	New York

Potential Concerns:

- Mobility
- Anglicization
- Mistranscription
- Common Names

Record Linking Challenges & Application

- **Who are the Descendants of Enslaved People:** Finding Better or Worse Ways to Implement Reparations by Sun Kyoung Lee, Brendan O'Flaherty, Working Paper, March 2024
- Linking direct victims and particular class of living people:
- Big data and machine learning approach to link historical and (relatively) contemporary population censuses

Our working example:

“Reparations should be paid to descendants of enslaved people in the US”

False Negatives →

← False
Positives

Our working example:

“Reparations should be paid to descendants of enslaved people in the US”

Data Linking

Historical Class

- Ideal class of people who should receive payments if we had perfect information, all of whom are living

Identifying the “qualified descendants” via Record Linking

Ideal Contemporary Class

- The living descendants of enslaved people

Correctly Identified Contemporaries

Actual Contemporary Class

- Whatever group of living people we are considering giving reparations to

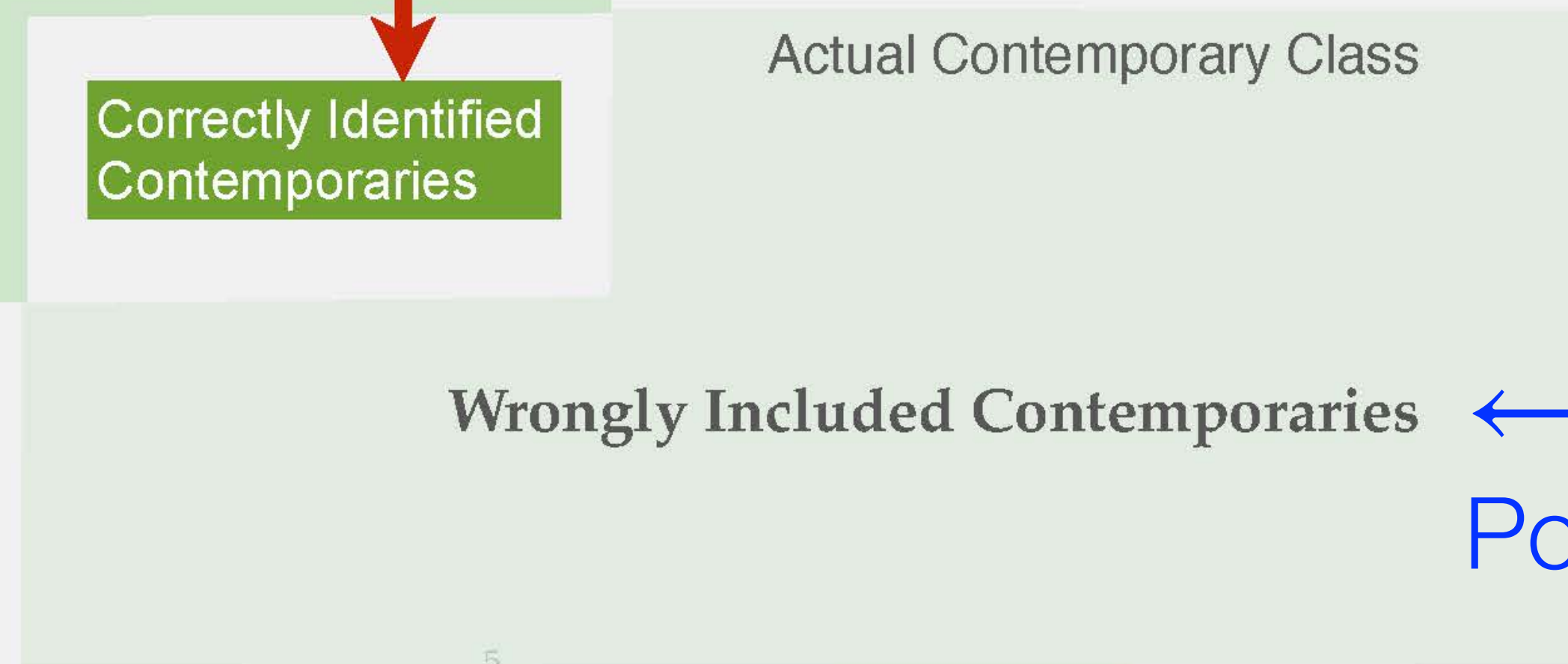
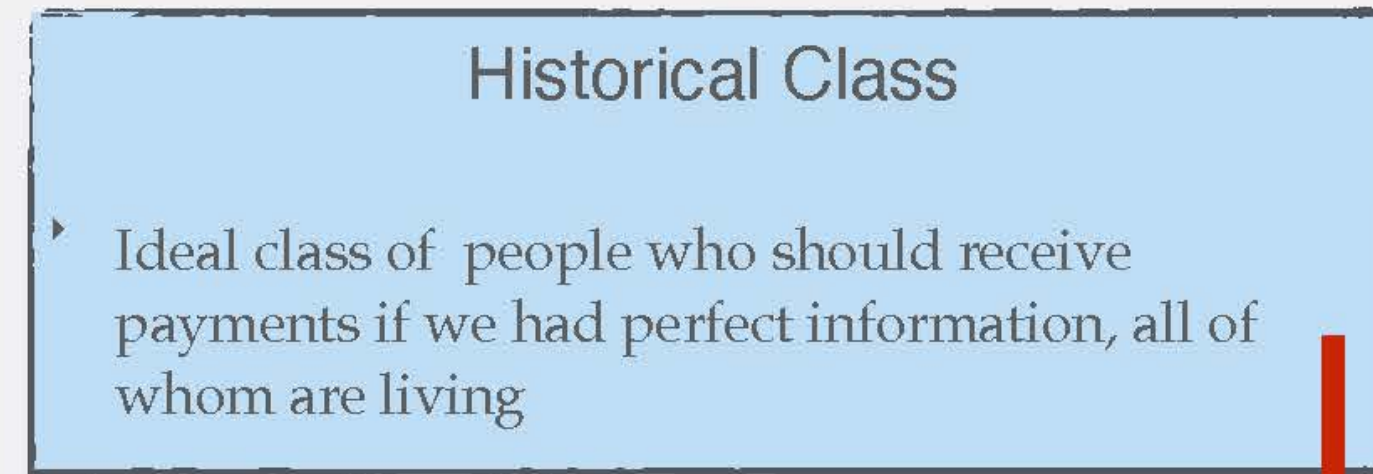
False Positives

False Negatives →

Our working example:

“Reparations should be paid to descendants of enslaved people in the US”

Data Linking



False Negatives →

← False Positives

Hidden Figures:

“The Descendants of Enslaved People”

Hidden Figures: "The Descendants of Enslaved People"

United States Census, 1910 | Georgia > Chatham > Savannah Ward 3 > ED 63 >

Source Box Attach to Family Tree

Image 25 of 47

Print Download Tools

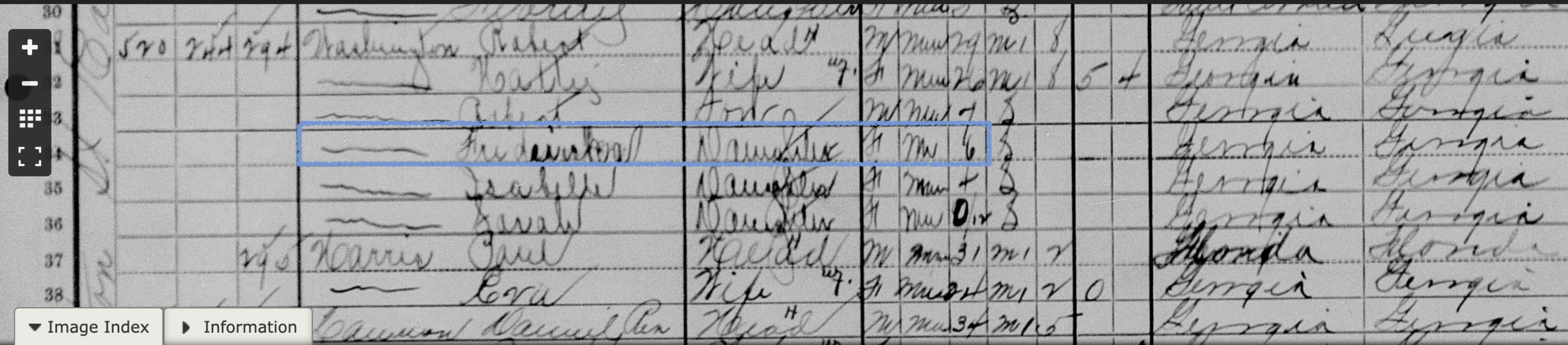
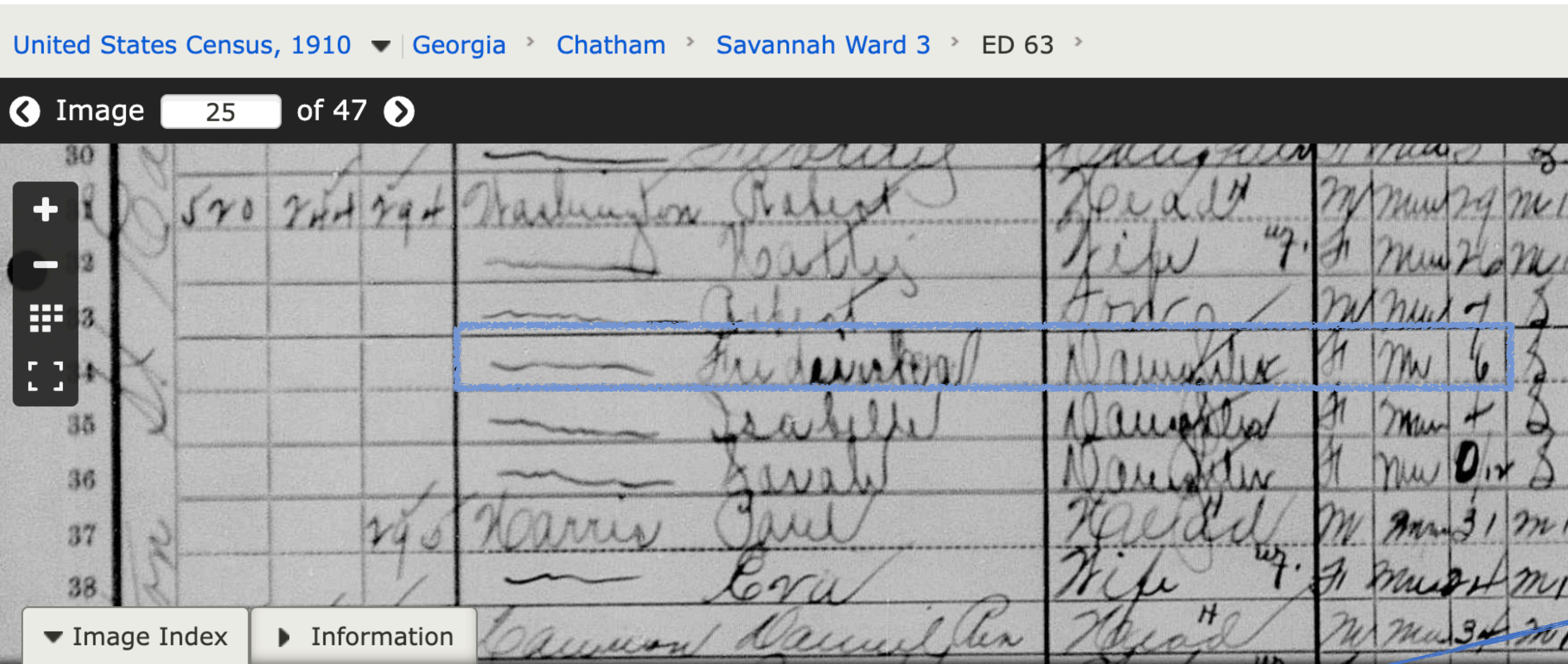


Image Index


Information

	Name	Sex	Age	Birth Year (Estim...	Marital Status	Race	Relationship to H...	Relationship Code	She
Attach	Robert Washington	Male	29	1881	Married	Mulatto	Head	Head	A
Attach	Kitty Washington	Female	26	1884	Married	Mulatto	Wife	Wife	A
Attach	Albert Washington	Male	7	1903	Single	Mulatto	Son	Son	A
Attach ??	Washington	Female	6	1904	Single	Mulatto	Daughter	Daughter	A

Hidden Figures: "The Descendants of Enslaved People"




	Name	Sex	Age	Birth Year (Estim...)	Marital Status	Race	Relationship to H...	Relationship Code	She
Attach	Robert Washington	Male	29	1881	Married	Mulatto	Head	Head	A
Attach	Kitty Washington	Female	26	1884	Married	Mulatto	Wife	Wife	A
Attach	Albert Washington	Male	7	1903	Single	Mulatto	Son	Son	A
Attach ??	Washington	Female	6	1904	Single	Mulatto	Daughter	Daughter	A




Alfred Washington
1856... • MW5N-C11

Select




Sarah Johnson
1860-D... • MW5N-ZM

Select




Robert Thomas Washington
1881-1964 • GMHC-Y6H

Select



Harriet "Hattie" Walker Ward
1884-1915 • GMHC-PS9

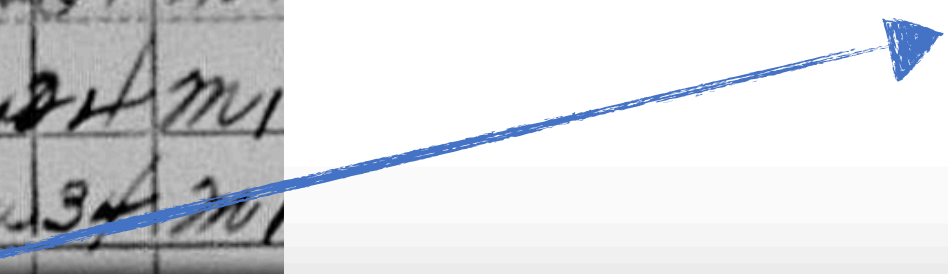
Select



Fredericka Carolyn "Fredi" Washington
1903-1994 • GMHC-R7P

Select

??



Roadmap

- Machine Learning-based Record Linking
- Specific Examples of Record Linking
- Comparison w/ Existing Record Linking Techniques
- Quality of Matches

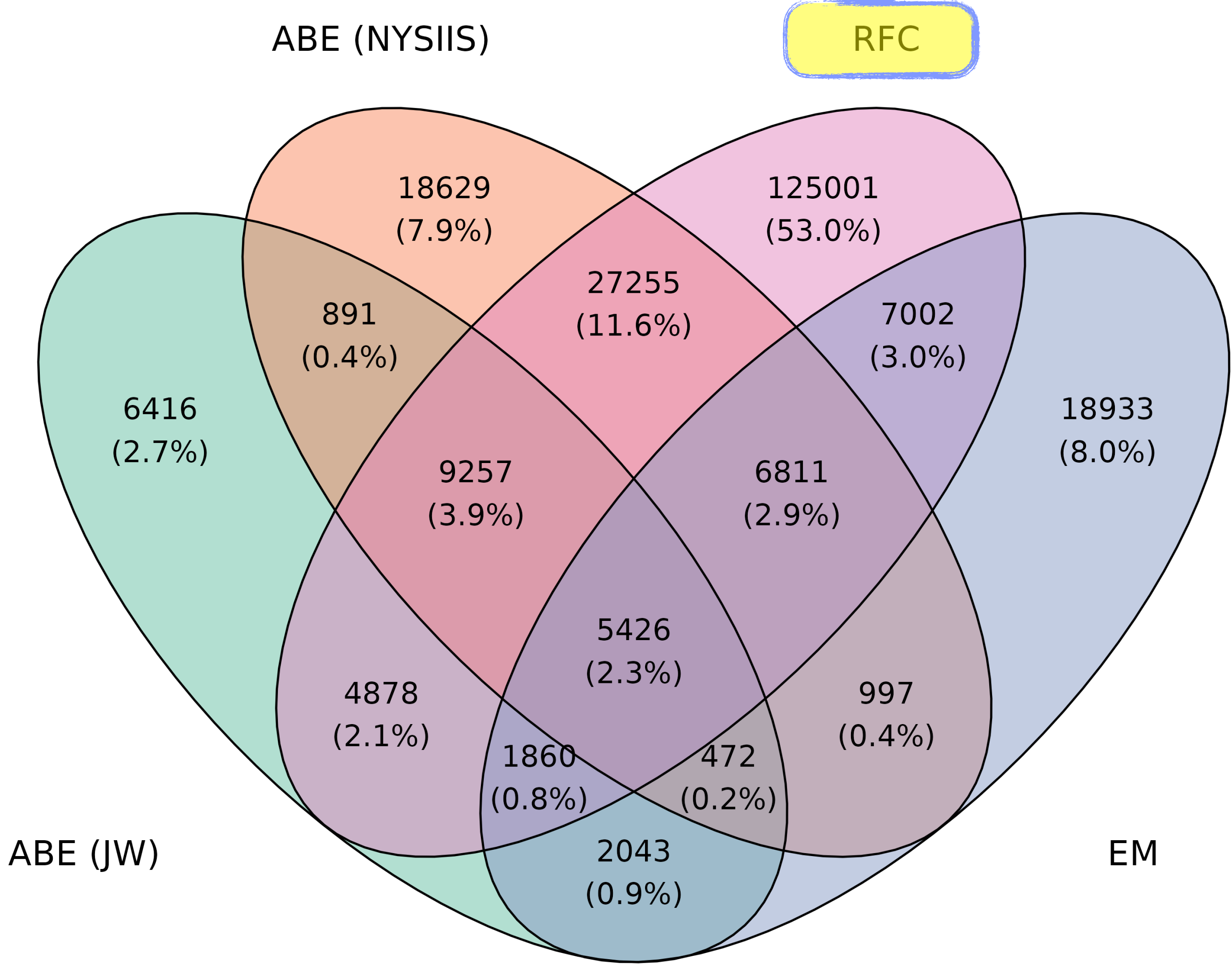
Compared Existing Method I: Iterative Record Linking

- Pioneered by Ferrie (1996), augmented by Abramitzky Boustan Eriksson (2012,2014)
- Steps:
 1. Constrain matching to only individuals in dataset A with a combination of **given name, surname, age, and place of birth that is unique in the dataset**
 2. For the unique individuals identified in step (1), find individuals in dataset B with matching characteristics. Preserve or discard matches by the following criterion:
 - (a) If the unique individual in A is a potential match with multiple individuals in B, discard these matches.
 - (b) If the unique individual in A matches only one individual in B, preserve the match.
 - (c) For the remaining unmatched individuals, repeat step (2) with a tolerance of one year of age difference, then a tolerance of two years.
 3. Repeat steps 1 and 2, reversing A and B.
 4. Return the intersection of matches from A matching B and B matching A.
- **Low recall rate and the record match heavily relies on the matching structure**

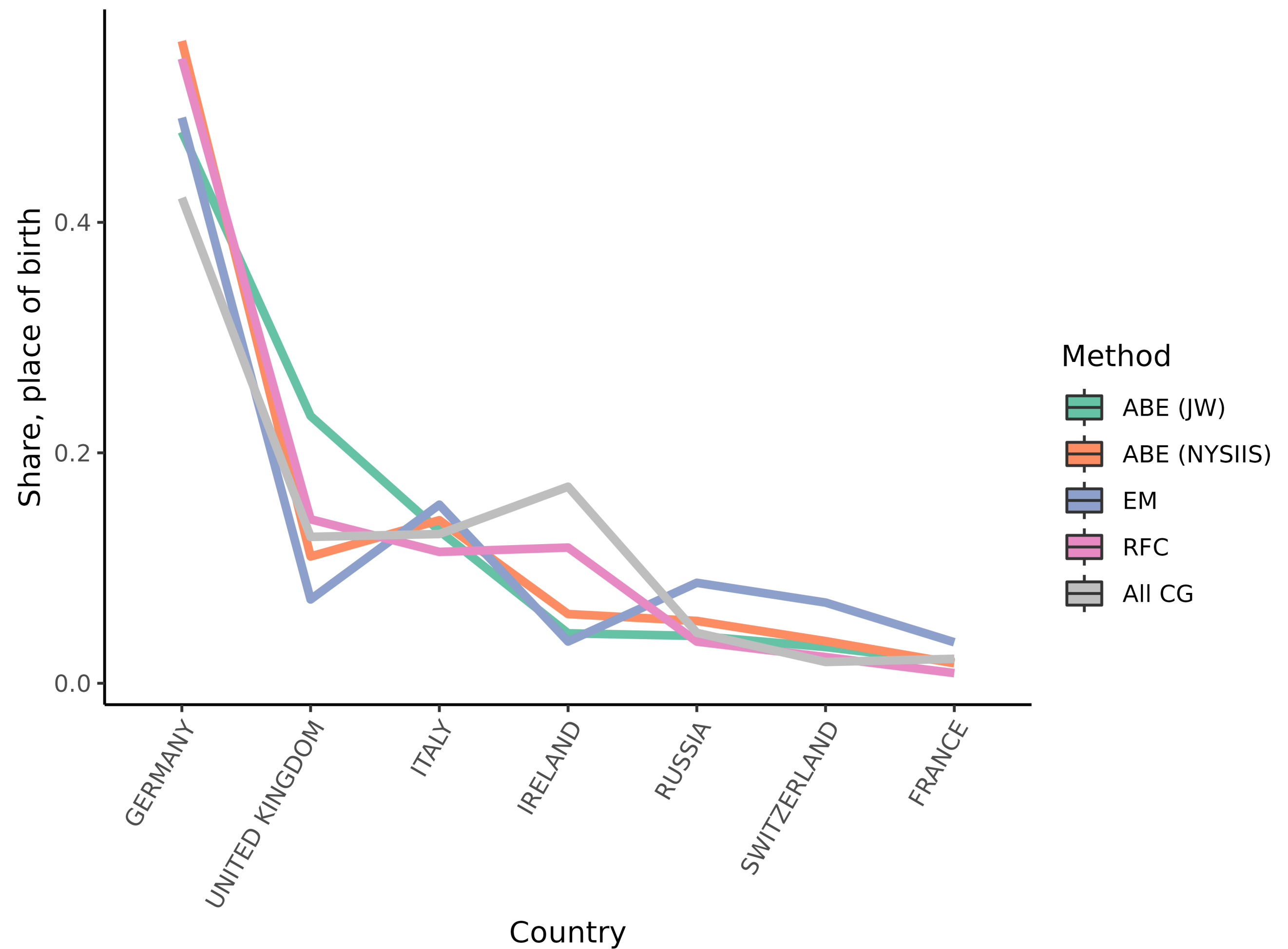
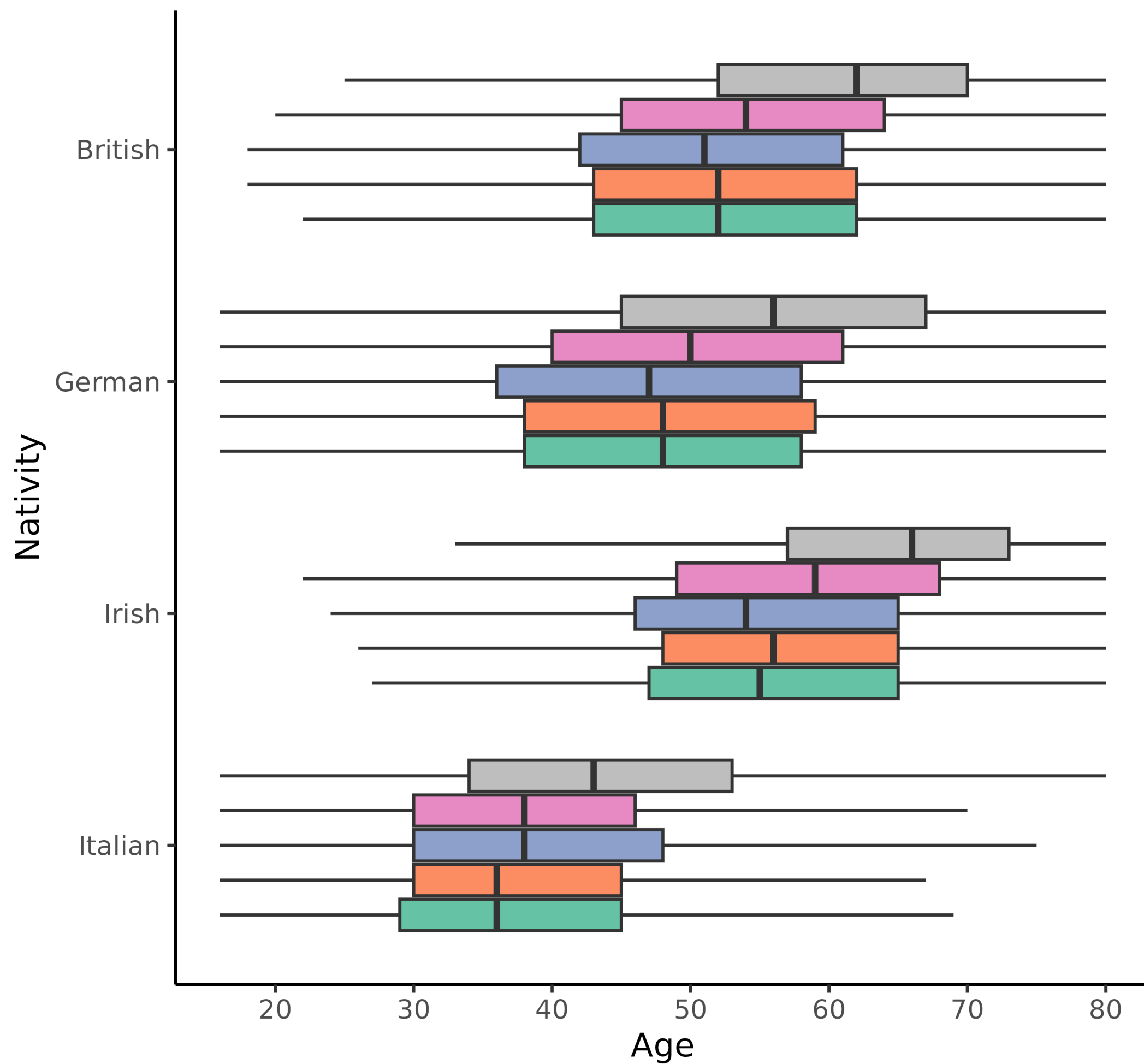
Compared Existing Method II: Expectation Maximization

- Unsupervised Machine Learning by Abramitzky Mill Perez (2019)
- Steps:
 1. Constrain matching for each record in dataset A to records in B with the **same place of birth, matching first and last initials, and an absolute age difference of two years or less**
 2. Compute the JW distance between the given name and surnames of each record
 3. Apply the expectation maximization procedure by estimating the probability of a true match given name and age distances between all pairs of records.
 4. Filter potential matches to those that are both:
 - A. Sufficiently probable, meaning match probability is greater than a researcher-provided parameter (p_m)
 - B. Sufficiently more probable than the next match, meaning that the probability score of the next best match must be less than a researcher-provided parameter (L)
- **Does NOT require a training dataset, but extremely costly in terms of computational time**

Match Overlap Across Linking Methods



Match Comparison Across Linking Methods: via **Distribution**



Roadmap

- Machine Learning-based Record Linking
- Specific Examples of Record Linking
- Comparison w/ Existing Record Linking Techniques
- Quality of Matches

Match Accuracy

- We would like to show:
 1. **Accuracy** of our record linking algorithm
 2. **Compare** our algorithm to existing approaches
- We also demonstrate:
 3. RFC does NOT introduce a significant bias by overrepresenting or underrepresenting certain groups
 4. There are NO systematic differences between RFC-based linked data and other matching algorithms

Match Accuracy (and Comparison) Across Linking Methods

Criteria:

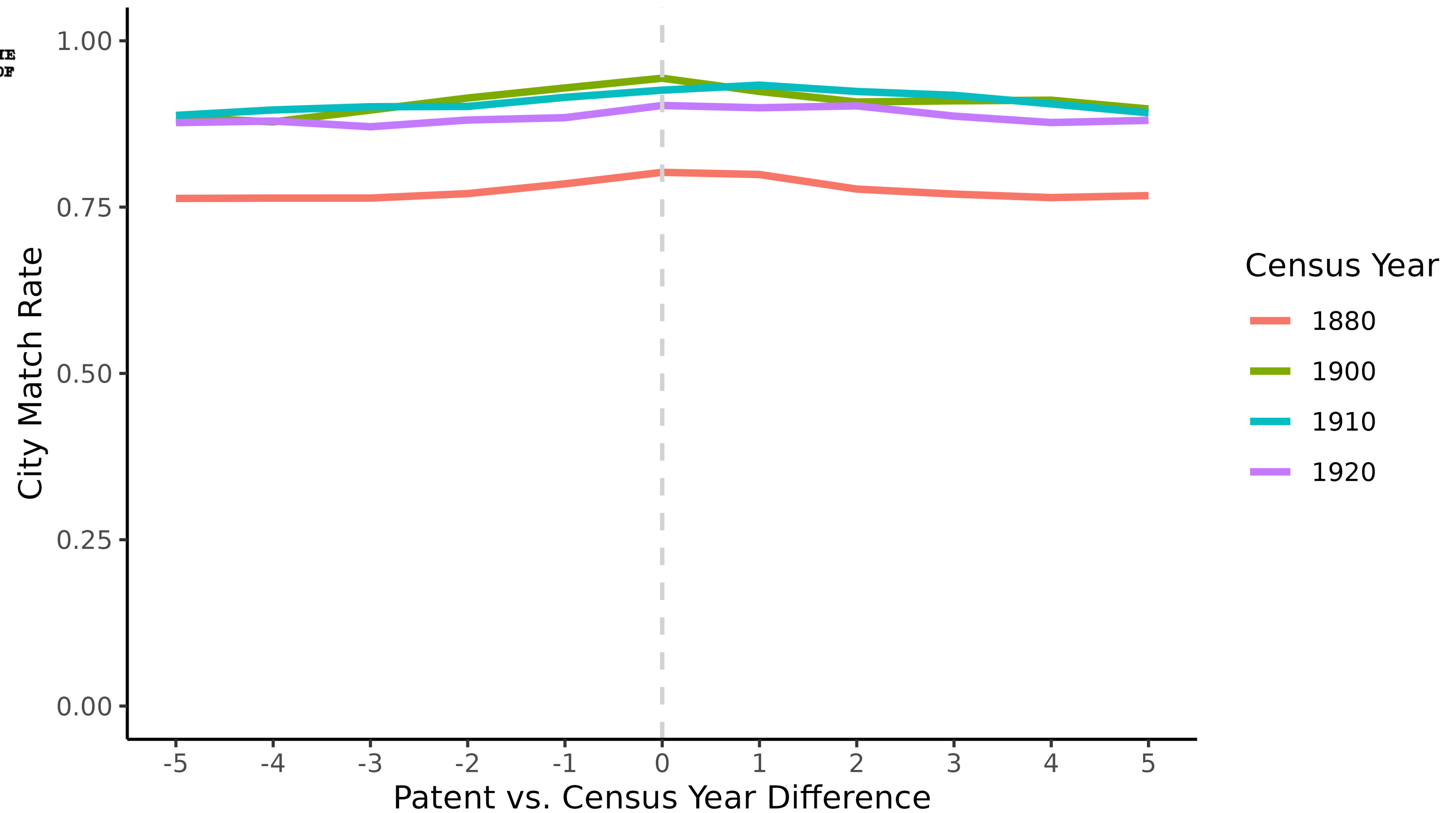
- Precision
- Recall

Match Method	Port Match (2+)	Port Match (3+)	Total Match Count
RFC	79.2	86.8	149,590
RFC (unique)	80.3	87.7	120,726
ABE (JW)	80.9	86.6	28,806
ABE(NYSIIS)	79.9	87.3	70,108
EM	80.4	86.1	53,160

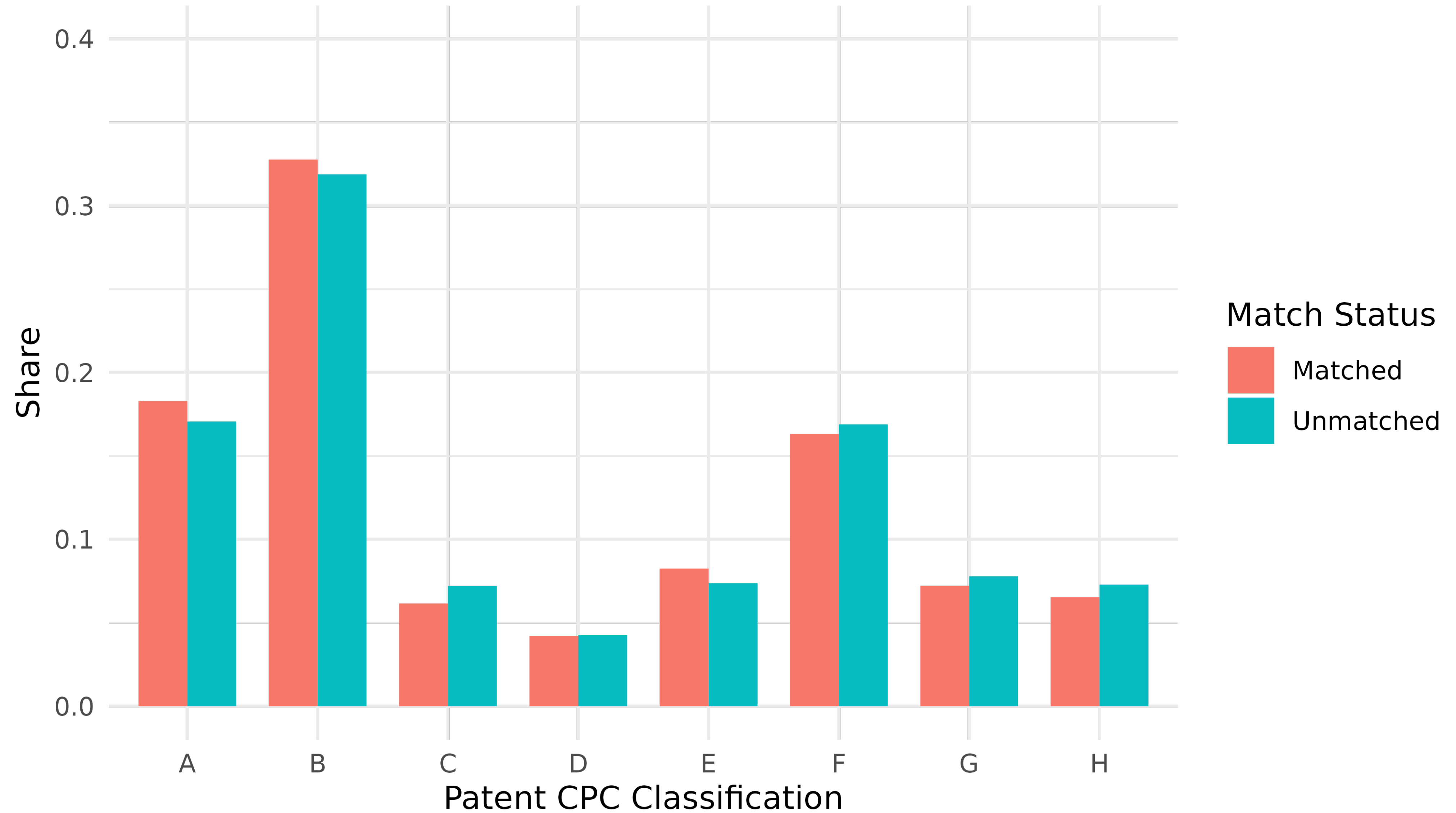
Match Accuracy: Patent and Census Matches

UNITED STATES PATENT OFFICE

FLOYD FIRESTONE, OF ANN ARBOR, MICHIGAN, ASSIGNOR TO THE REGENTS OF THE UNIVERSITY OF MICHIGAN, OF ANN ARBOR, MICHIGAN, A CORPORATION OF MICHIGAN



Match Representativeness: Patent & Census Matches



Taking Stock

- Unlike Bailey et al (2017) “How Well Do Automated Linking Methods Perform”, automated methods perform fairly well
- Some methods and design could introduce bias or high false positive rates more than others
- Examining (almost) every aspect of linking process is a necessary step in well-execute record linking
- Whenever you can, **find ways to validate the accuracy** of the matches