What can we do with millions of records?

Paper presented at:
'FreeReg and FreeCen @ Twenty': FreeUKGenealogy Conference, 2019
King's Manor, University of York, 29/9/19

Twitter: #FreeUKG2019

Dr Oliver Duke-Williams
o.duke-williams@ucl.ac.uk
@oliver_dw
www.ucl.ac.uk/dis/people/oliverdukewilliams

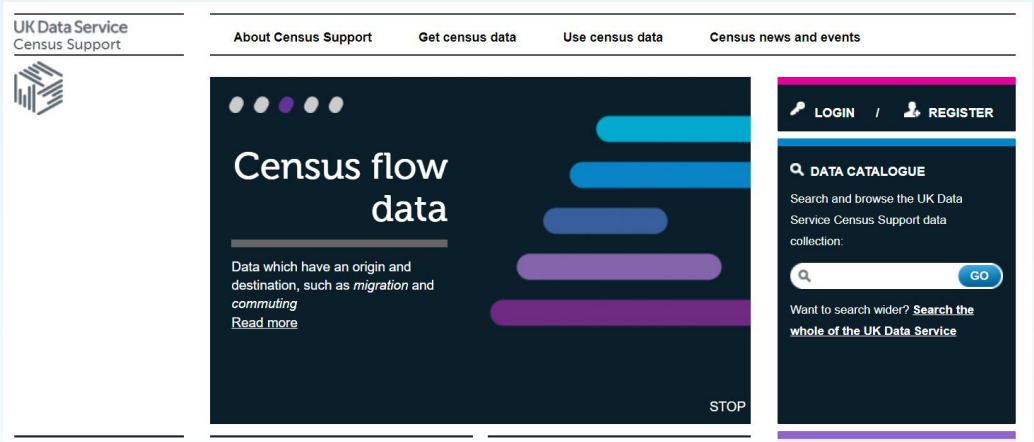
UCL Information Studies
Celebrating 100 years

37,433,884

My background

UCL CENTRE FOR DIGITAL HUMANITIES

www.ucl.ac.uk/digital-humanities/



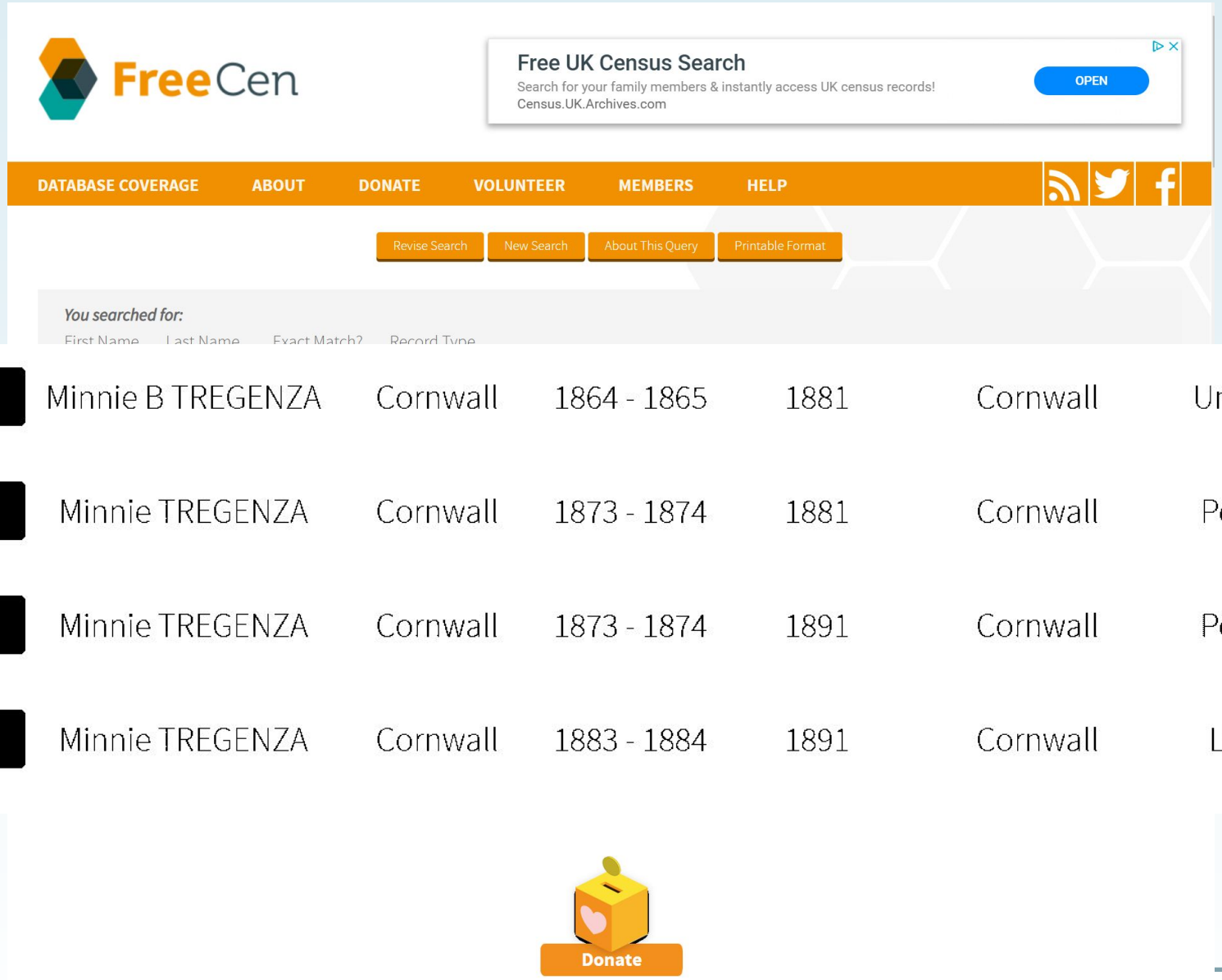
census.ukdataservice.ac.uk/

CeLSIUS

Centre for Longitudinal Study Information & User Support

ucl.ac.uk/celsius

Typical FreeCen use case



Source: freecen.org.uk

Finding more information about Minnie

FreeBMD

New Query | Revise Query | Save Search | Download

Key: B M D — Change of shade means entries are not contiguous — An entry in **Bold** means it has been double keyed — A district in *Italic* means it is suspect (click on it for details) — Scan, Comment, Postem, New entry, System Entry, System Link

Surname	First name(s)	Age	District	Vol	Page	
Births Mar 1874 (>99%)						
Tregenza	**Minnie**		Penzance	5c	294	Info
Births Jun 1884 (>99%)						
Tregenza	Minnie		Liskeard	5c	48	Info ?
Marriages Dec 1898 (>99%)						
Tregenza	**Minnie**		Penzance	5c	451	Info
Births Jun 1902 (>99%)						
Tregenza	**Minnie Isabel**		Halifax	9a	370	Info
Marriages Sep 1904 (>99%)						
TREGENZA	Minnie		Redruth	5c	305	Info
Marriages Jun 1906 (>99%)						
Tregenza	**Minnie**		Liskeard	5c	95	Info

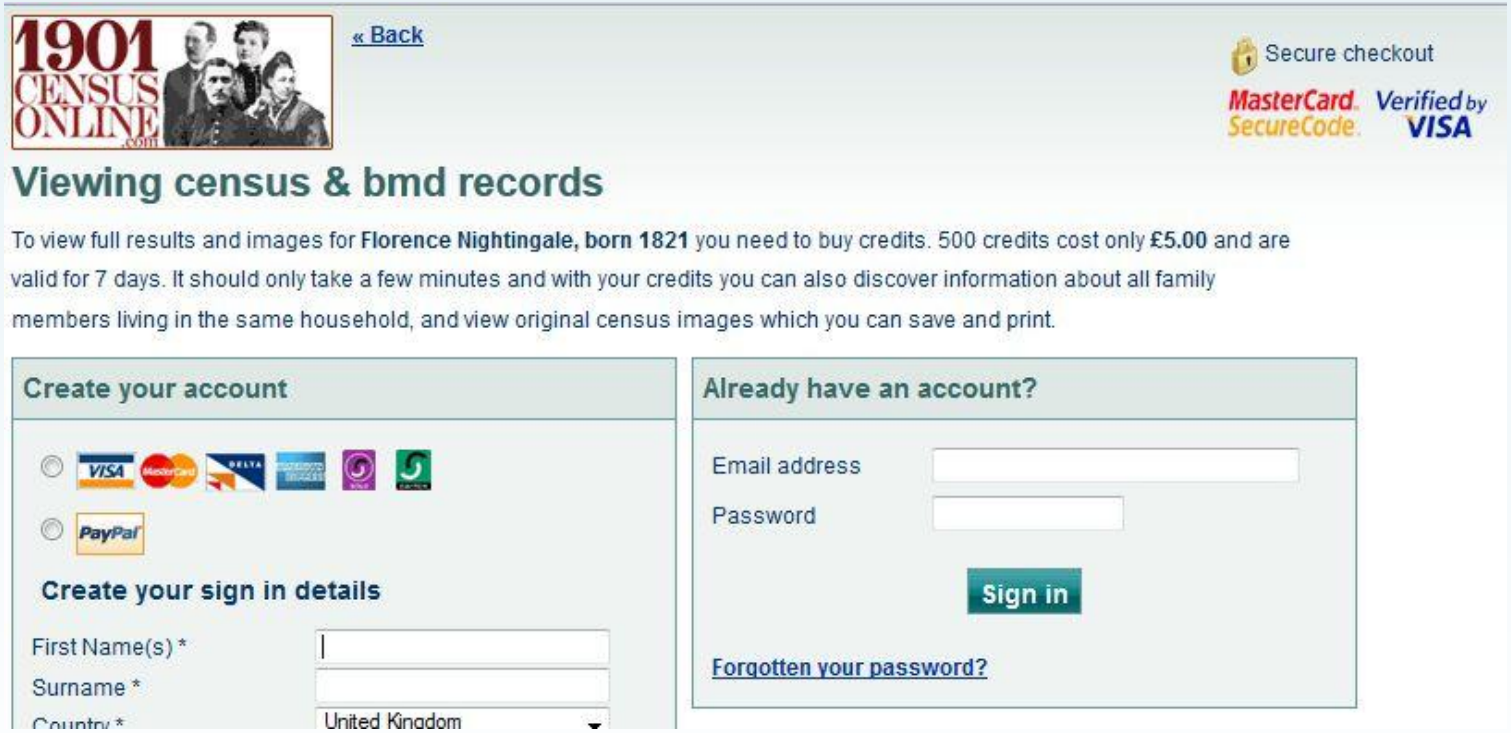
Surname	First name(s)	Mother/Spouse/Age	District	Vol	Page	
Deaths Sep 1927 (>99%)						
Tregenza	Minnie	39	Redruth	5c	194	Info

Source: freebmd.org.uk

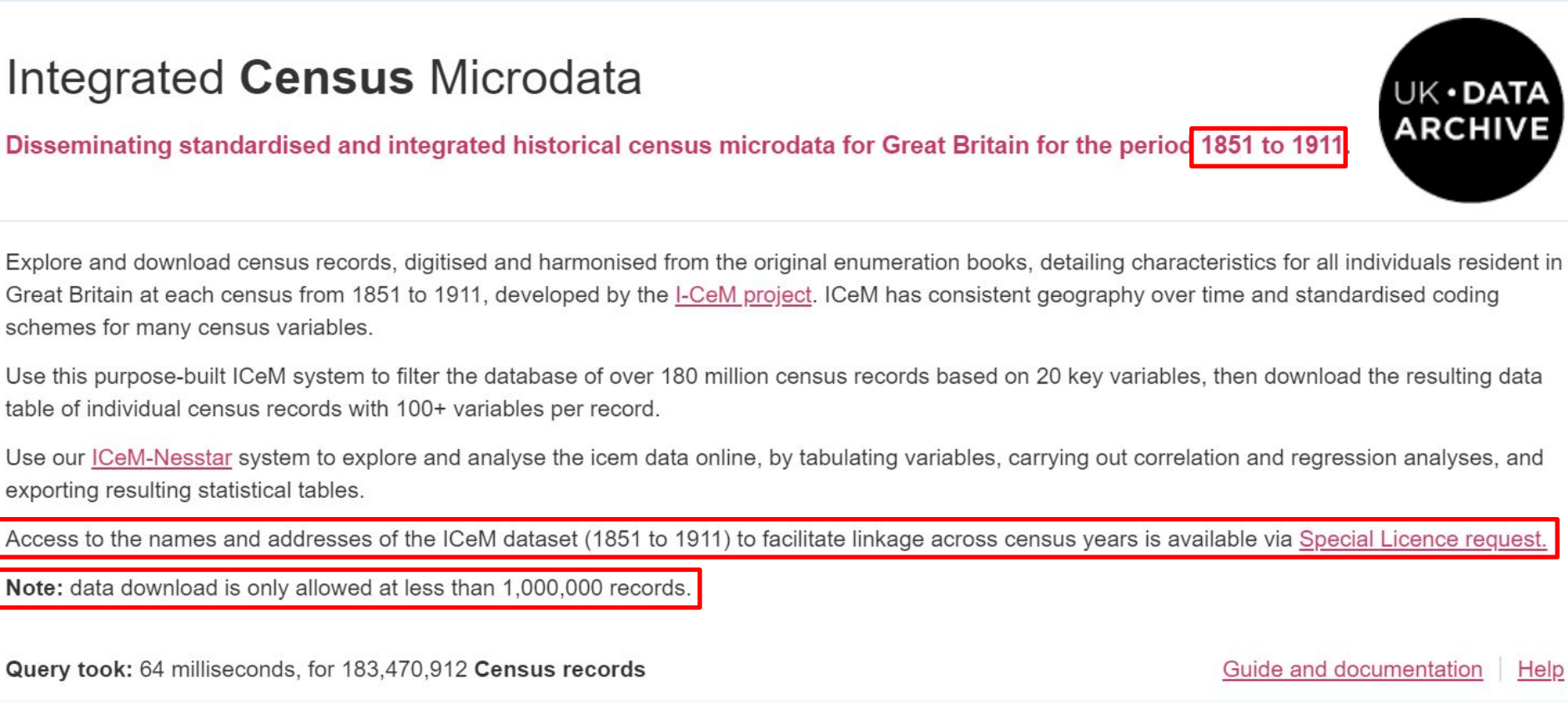
Alternatives

- Commercial sources
- Academic sources

Problems with commercial sources



Academic sources



Integrated **Census** Microdata

Disseminating standardised and integrated historical census microdata for Great Britain for the period 1851 to 1911.

UK • DATA ARCHIVE

Explore and download census records, digitised and harmonised from the original enumeration books, detailing characteristics for all individuals resident in Great Britain at each census from 1851 to 1911, developed by the I-CeM project. ICeM has consistent geography over time and standardised coding schemes for many census variables.

Use this purpose-built ICeM system to filter the database of over 180 million census records based on 20 key variables, then download the resulting data table of individual census records with 100+ variables per record.

Use our ICeM-Nesstar system to explore and analyse the icem data online, by tabulating variables, carrying out correlation and regression analyses, and exporting resulting statistical tables.

Access to the names and addresses of the ICeM dataset (1851 to 1911) to facilitate linkage across census years is available via Special Licence request.

Note: data download is only allowed at less than 1,000,000 records.

Query took: 64 milliseconds, for 183,470,912 **Census records**

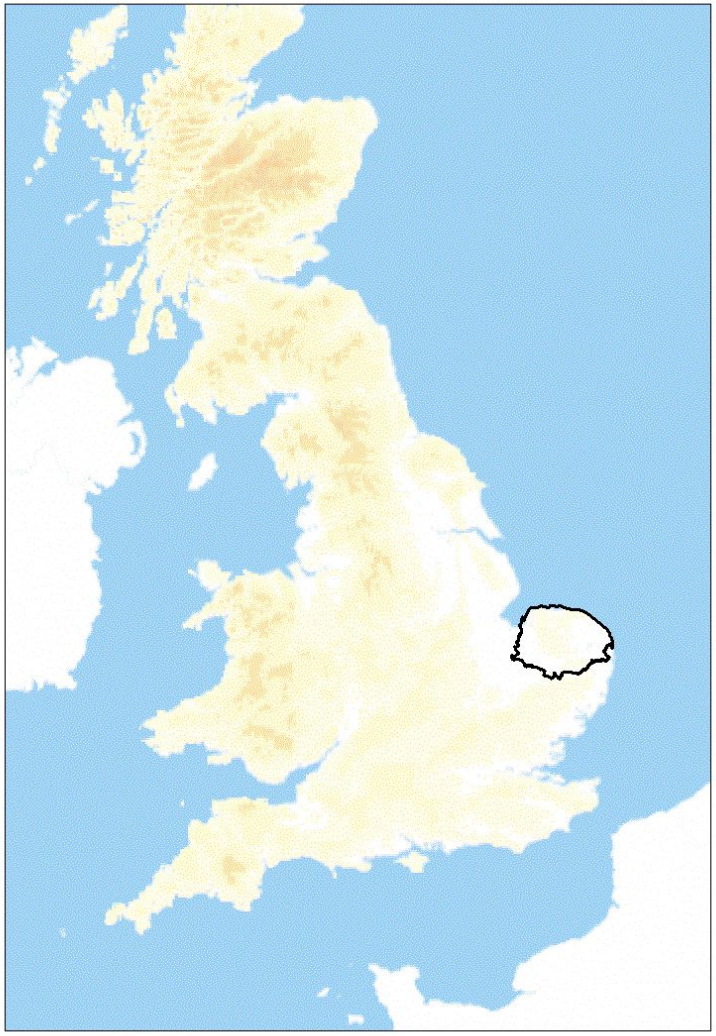
Guide and documentation | Help

- Doesn't include 1841
- Access to names and addresses restricted
- Not designed as a genealogy resource
- Bulk download size constrained

icem.data-archive.ac.uk/

FreeCen: Norfolk sample

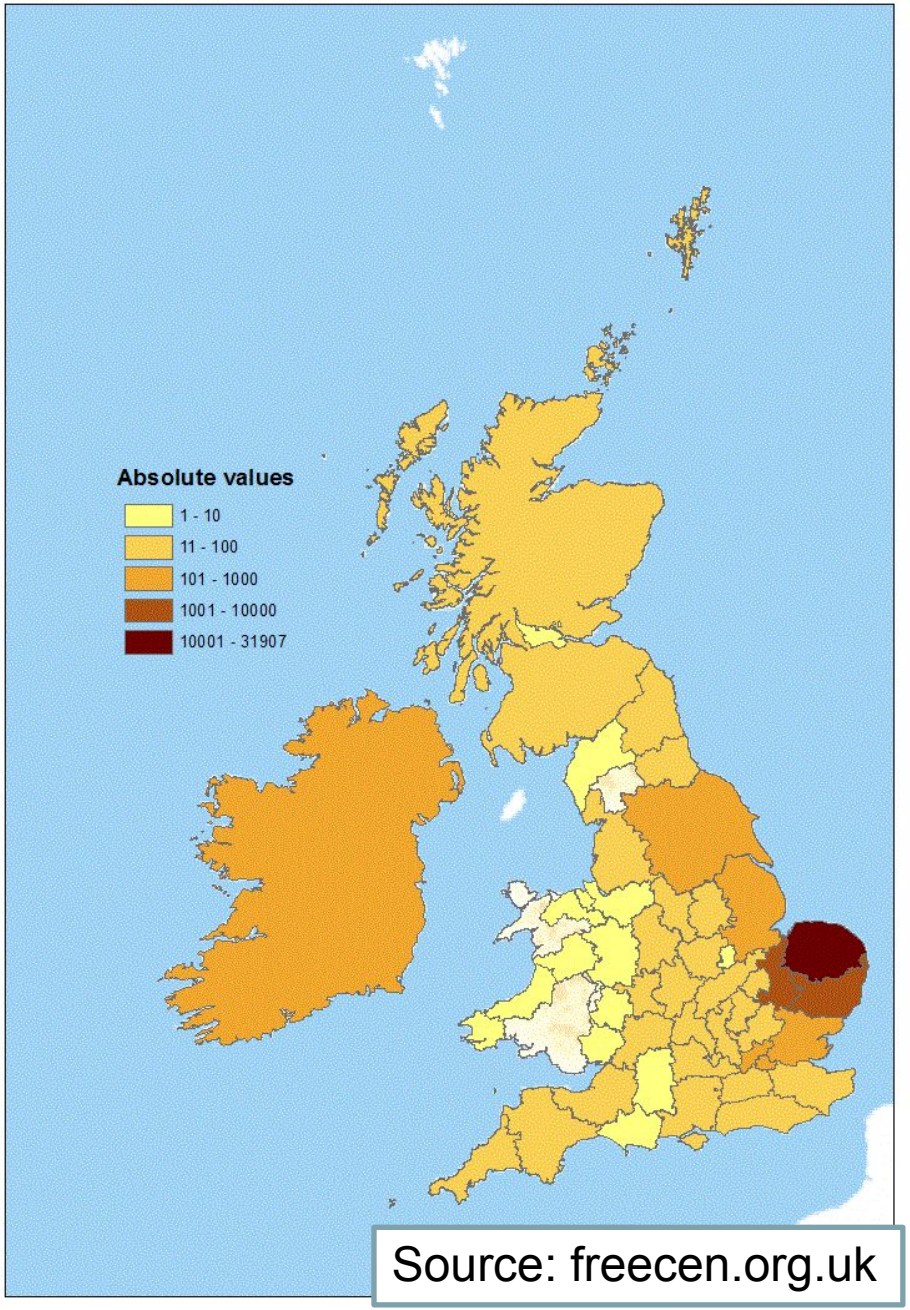
- Sample of raw data from 1861
  - c. 40,000 records in sample
  - 423,000 records in county (100% transcription)



Source: freecen.org.uk

County-of-birth of sample members

- Generally, county coded as Registration County
  - ‘Yorkshire’ as single county
  - General references to ‘England’, ‘Wales’, ‘Scotland’



Source: freecen.org.uk

Constructing new tables

- It is possible to produce tables that were never published at the time
  - Multi-generational households
    - Link generation coding to ‘relation to head of household’
    - Identify earliest and latest generations per household
    - Aggregate over other constraints

Multi-generational households

Number of generations in household	Observed cases
1	2,381
2	5,114
3	740
4	17

Source: freecen.org.uk

Cornwall in FreeCEN

- Cornwall is the only county that is 100% coded for the complete FreeCEN run

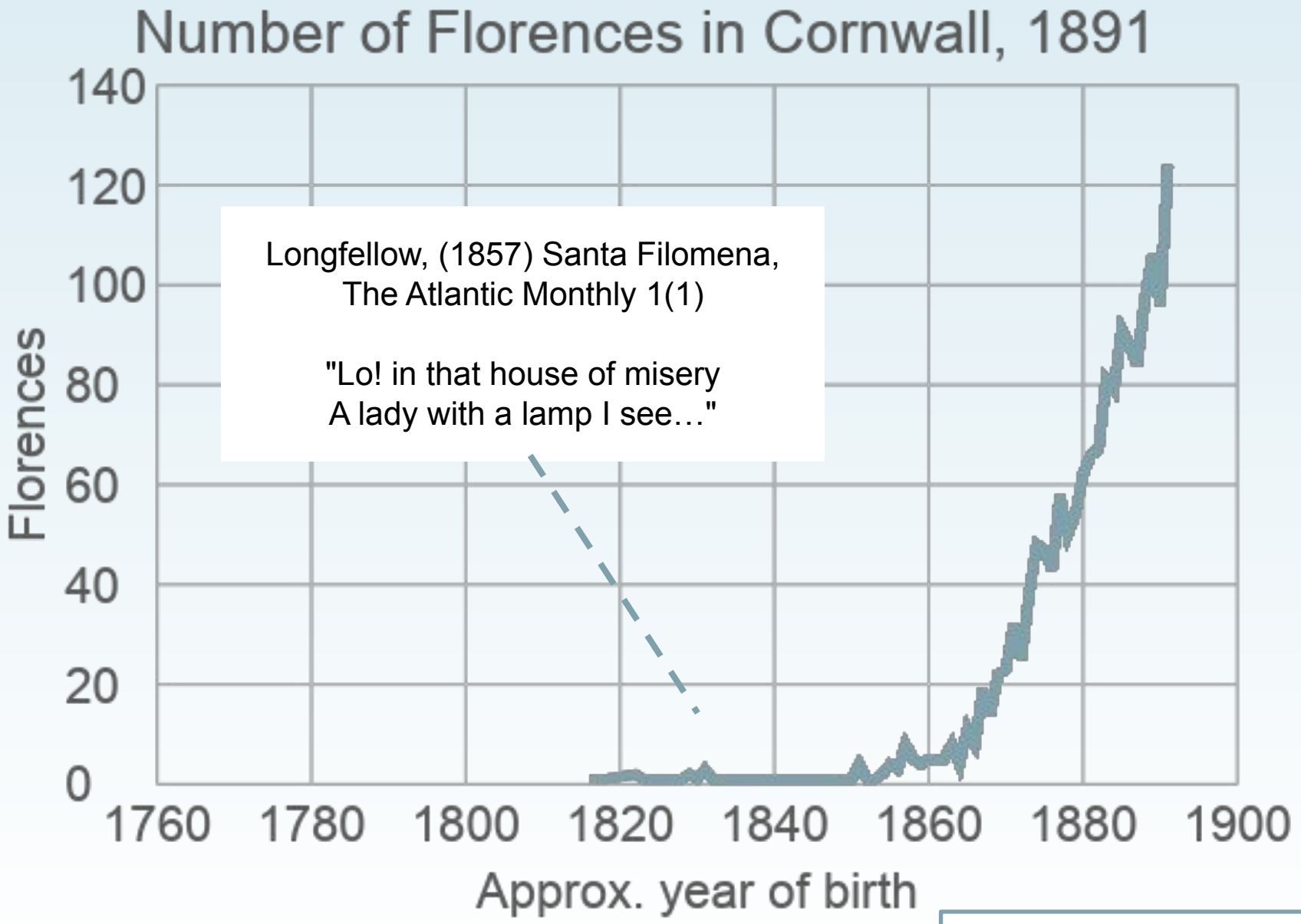
Census	Records	%complete
1841	340,901	100
1851	354,744	100
1861	362,111	100
1871	358,043	100
1881	326,187	100
1891	318,637	100

Source: freecen.org.uk

Female names, 1841

Female names, 1891

Popularity of the name ‘Florence’



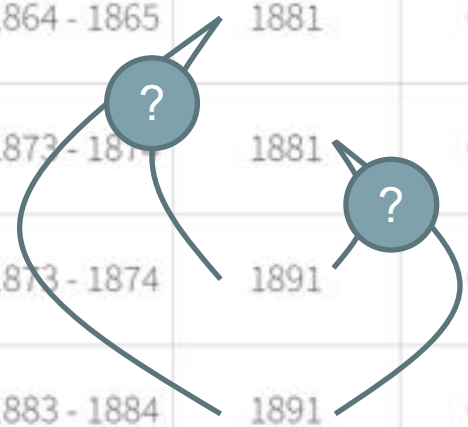
Source: freecen.org.uk

Constructing longitudinal data

- Longitudinal data sets are constructed from modern censuses
  - ONS Longitudinal Study; Scottish Longitudinal Study; Northern Ireland Longitudinal Study
  - These rely on name and date of birth for linkage
  - Date of birth was first asked in the 1971 Census
- Can we attempt to build longitudinal data from FreeCen data?
  - Can we find the same person in two or more censuses

Matching people by name

Detail	Individual	Birth County	Birth Year	Census Year	Census County	Census Place
Detail	Minnie B TREGENZA	Cornwall	1864 - 1865	1881	Cornwall	Uny Lelant
Detail	Minnie TREGENZA	Cornwall	1873 - 1874	1881	Cornwall	Penzance
Detail	Minnie TREGENZA	Cornwall	1873 - 1874	1891	Cornwall	Penzance
Detail	Minnie TREGENZA	Cornwall	1883 - 1884	1891	Cornwall	Liskeard



Source: freecen.org.uk

Matching people by name

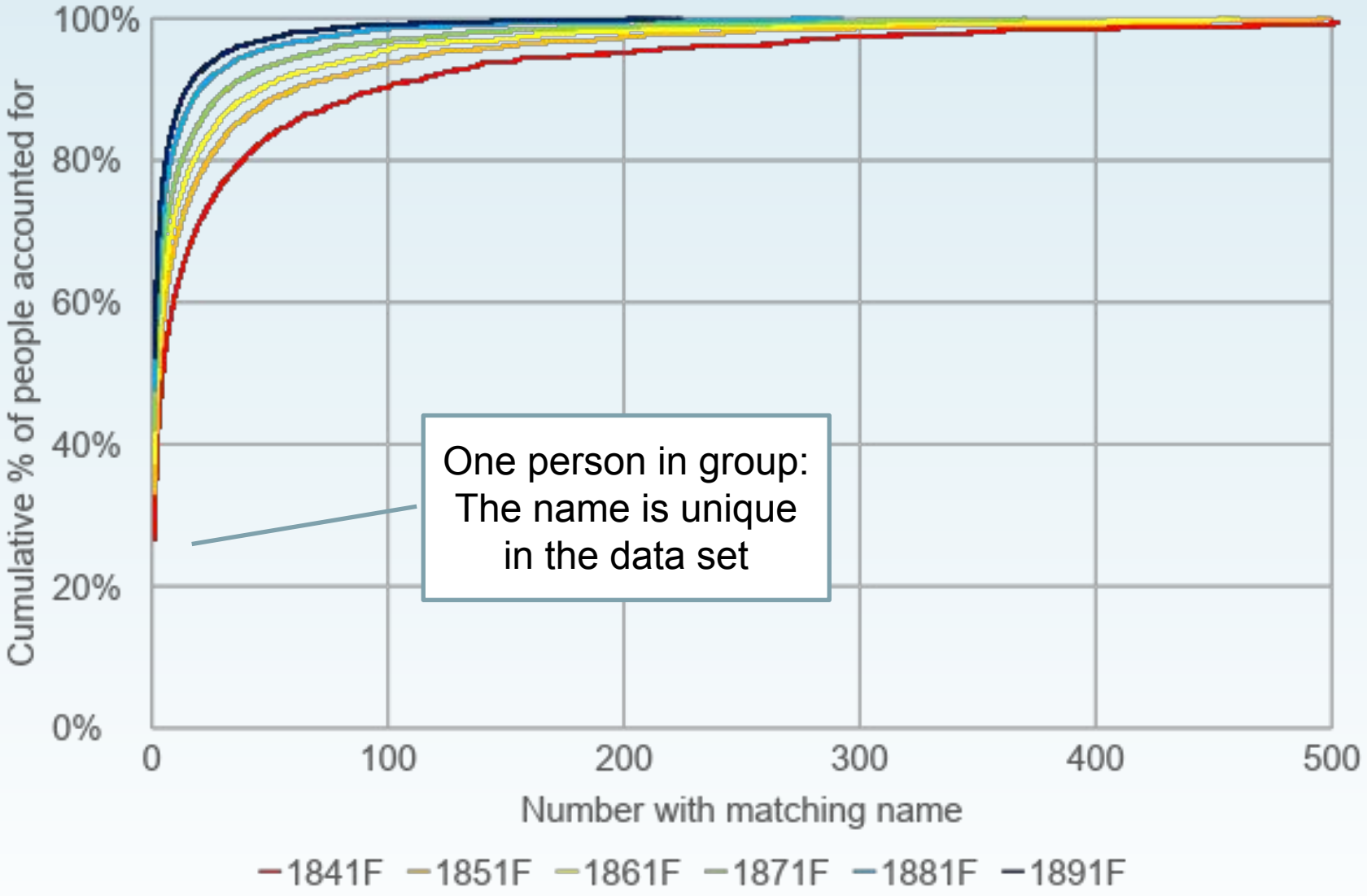
Detail	Individual	Birth County	Birth Year	Census Year	Census County	Census Place
Detail	Minnie B TREGENZA	Cornwall	1864 - 1865	1881	Cornwall	Uny Lelant
Detail	Minnie TREGENZA	Cornwall	1873 - 1874	1881	Cornwall	Penzance
Detail	Minnie TREGENZA	Cornwall	1873 - 1874	1891	Cornwall	Penzance
Detail	Minnie TREGENZA	Cornwall	1883 - 1884	1891	Cornwall	Liskeard

Source: freecen.org.uk

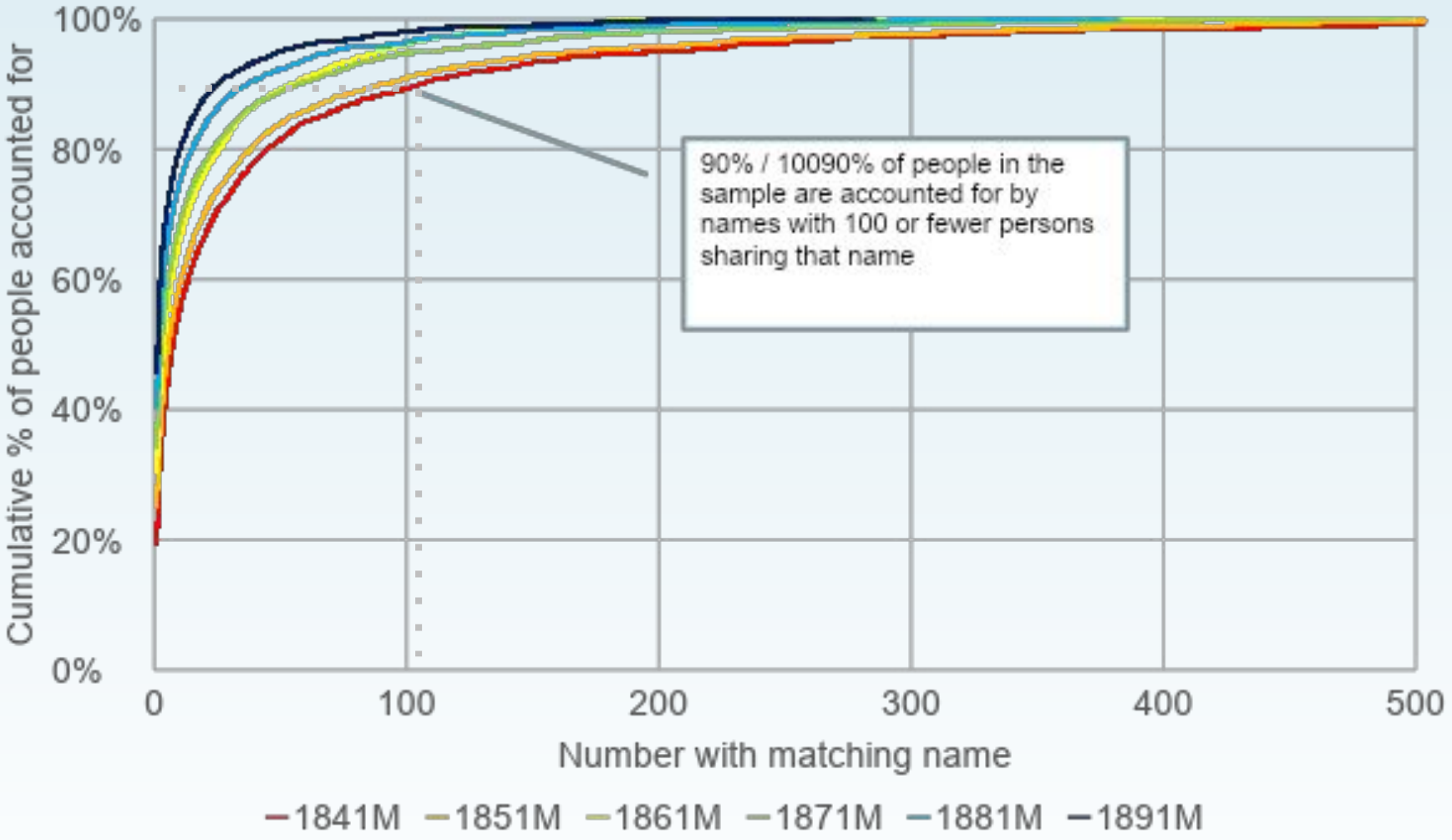
Searching for uniques

- How do we isolate different people?
  - We can look for people with unique names
  - We can look for people for whom the combination of name and something else is unique

How unique are names?



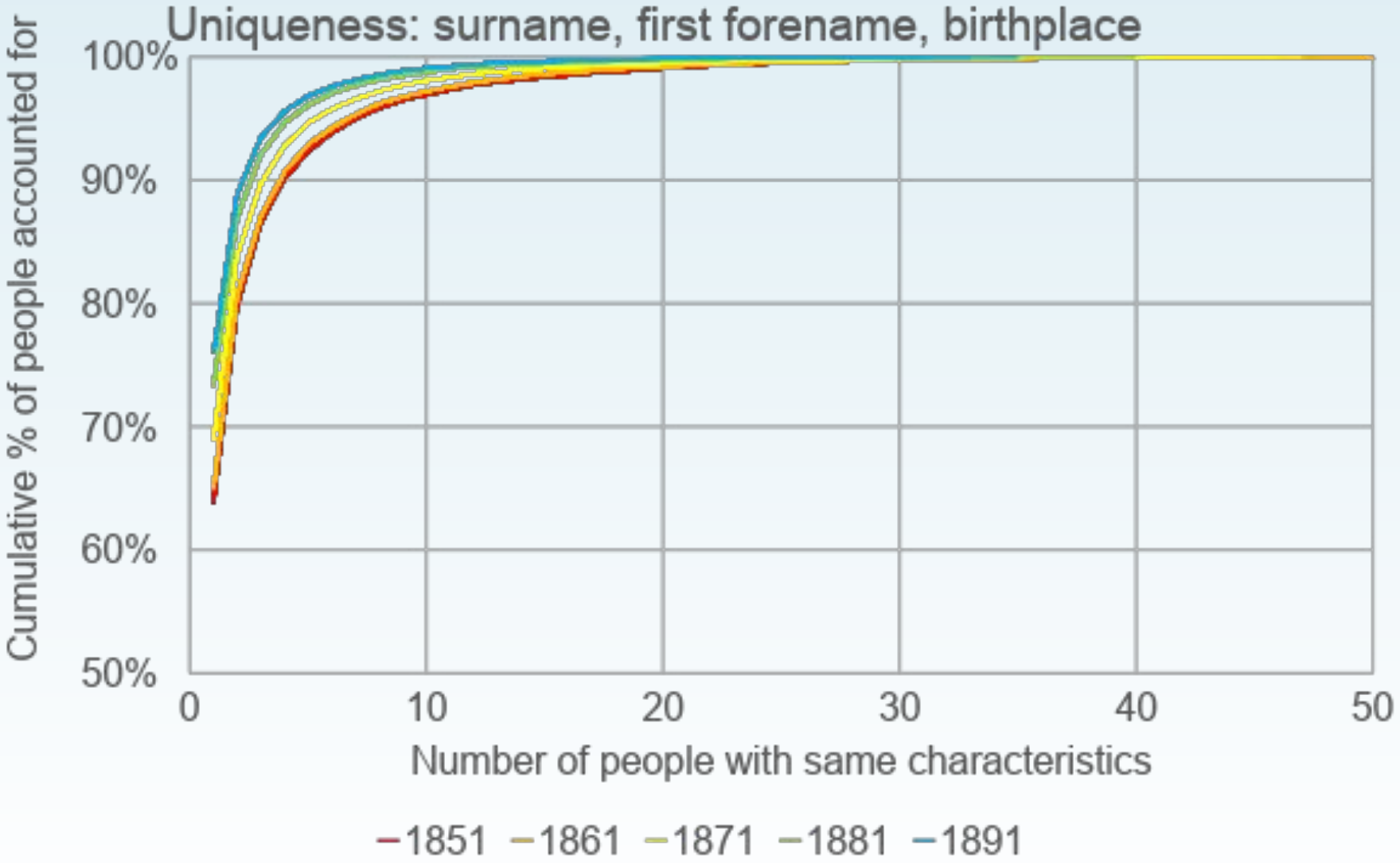
Source: freecen.org.uk



Source: freecen.org.uk

Increasing the chance of people being unique

- We can increase our chances of finding ‘unique’ people by looking at more characteristics than just name
  - Age
  - Place of birth
- We might increase the chance of having a match by ignoring initials and second names etc



Source: freecen.org.uk

How many people are unique?

Characteristics used	1851	1891
Name (male)	25%	45%
Name (female)	33%	52%
Name, birth place (persons)	64%	76%
Name, age (persons)	67%	84%
Name, age, birth place (persons)	98%	99%

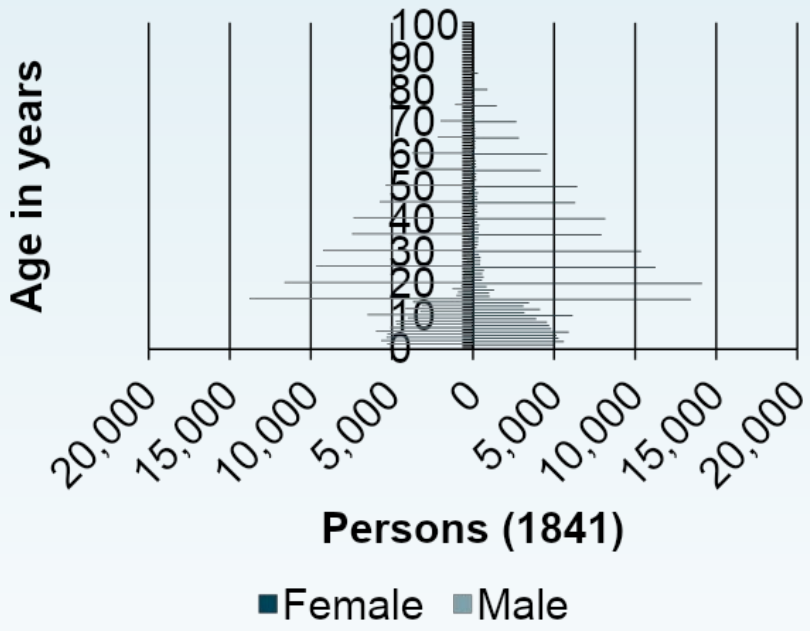
Source: freecen.org.uk

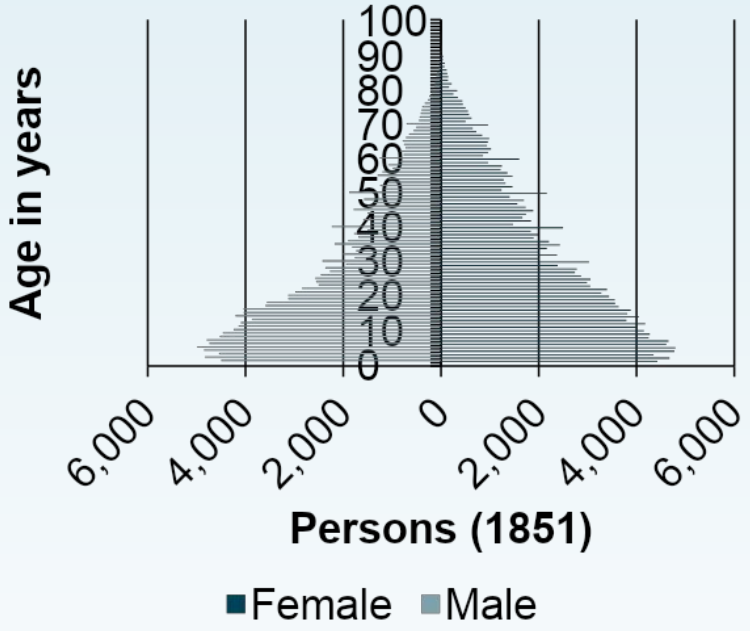
- A very high proportion of people are unique on
  - Name
  - Age (years)
  - Birth place
- This bodes well for potential to match across years
- As we move from Cornwall to UK, uniqueness would drop

But…

- Just because people are unique *within* a census, they may not match *between* censuses
  - Error in completing census
  - Variations in spelling or presentation of name
  - Variation in spelling of birth place
  - Errors introduced by transcribers

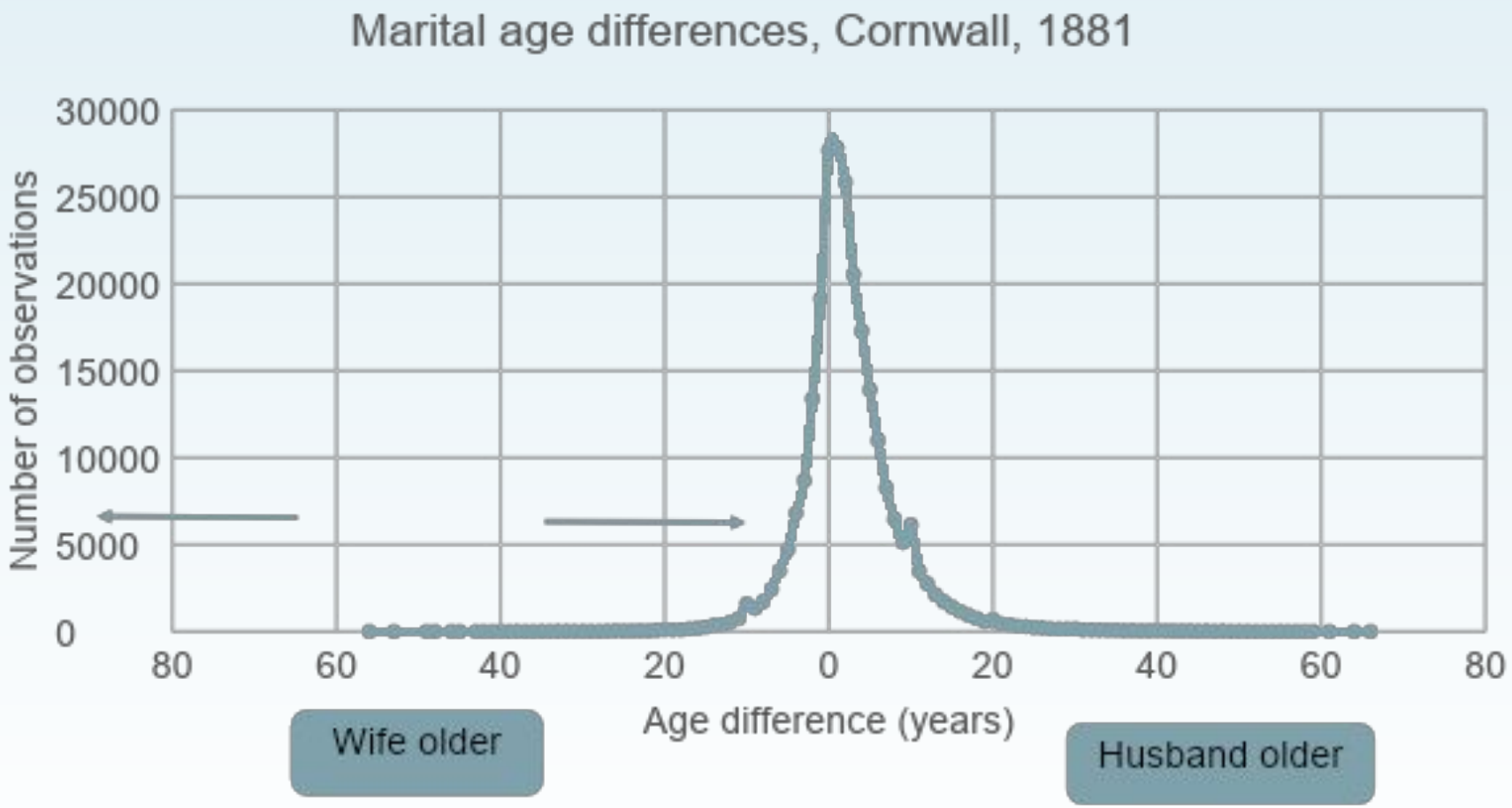
Error in completion: failing to follow instructions





Source: freecen.org.uk

Transcription errors



Source: freecen.org.uk

Transcription errors: extreme marital age differences

Extract from FreeCen data

TREZISE	William	-	Head	M	M	17
TREZISE	Mary	-	Wife	M	F	73
WAITE	Elizabeth	-	Dau	M	F	38
WAITE	William T	-	Grnson	-	M	4
KNIGHT	William	-	Head	M	M	16
KNIGHT	Abigail	-	Wife	M	F	69
KNIGHT	William	-	Son	M	M	36
ALLAN	Josia	-	Head	M	M	16
ALLAN	Mary A	-	Wife	M	F	69
ALLAN	T K	-	Boardr	-	M	2

Source: freecen.org.uk

Assembling matches between censuses

- We can look for matching sets of names
- We can limit our search to people with a ten year age difference
  - Census dates vary, so the age difference could also be 9 or 11
- To be more confident, we might look at pairs of names in a household

Results: 1881 and 1891

Match key	Feasible unique matches
Name, age+10	51,147
Name, age+10, birth place	40,890
Name, spouse name, birth places, ages+10	2,380

Source: freecen.org.uk

Extracts of results: occupation changes

Surname	Forenames	Age	Birth place	Occupation 1881	Occupation 1891
ROWE	Mary Ann	45	Breage	Farmer Wife	Farmers Widow
ROBERTS	Honour	50	Landewednack	Farmers Wife	Farmer's Widow
POLGLASE	Jane	59	Breage	Tin Miner Wife	Tin Miners Widow
TRAYES	Gertrude	35	St Teath	Assistant Cook (Dom)	Dressmaker,Retired
VIVIAN	Ada	22	Budock	Assistant Draper	Draper's Assistant(Em'ee)
MATTHEWS	Albert	13	St Buryan	Assistant Farm Servnt (Indoor)	Hedger(Em'ee)
HEARD	Elijah	29	Tintagel	Assistant Farmer	Farmer(Notem)
SNELL	Hugh	26	Menheniot	Assistant Farmer	Farmer(Em'er)
TREGENZA	Minnie	7	Mousehole	NULL	General Servant

Source: freecen.org.uk

What's next

- Validation of results
- More flexible match keys (variant spellings etc)
- More use of household structure
- Comparison of transcription quality
- Extension beyond Cornwall