Challenges of Creating and Working with Cross-Year-Family-Individual Files – Example from PSID data set

Petia Petrova
Boston College
petrova@bc.edu

1.What could go wrong?
2. How to fix the problem?

1. Why longitudinal data sets?

   - to study individuals and families or firms and plants across time

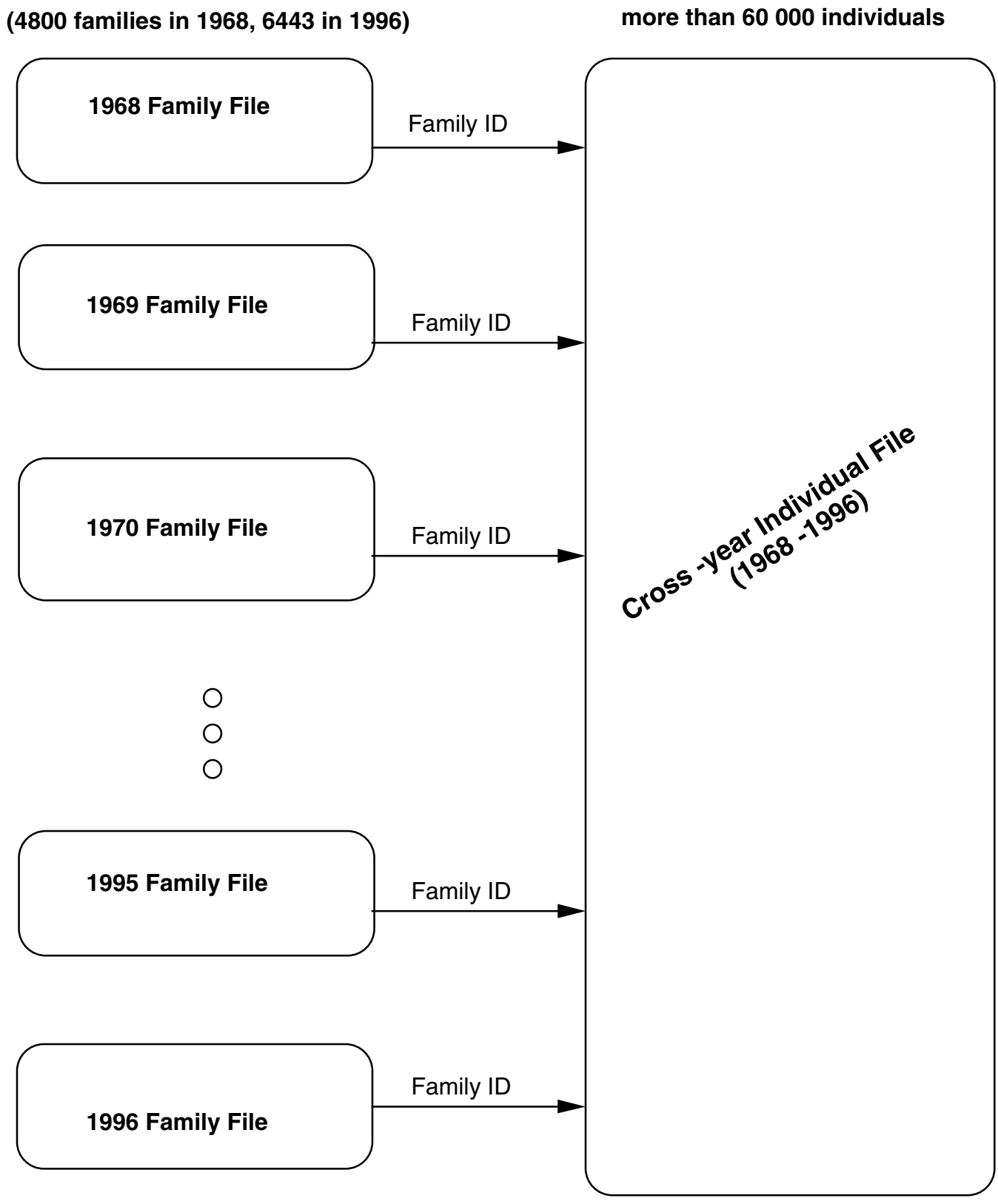
2. Why simply merging, on for example, family and person ID-s leads to wrong records?

   - the data come into a different format (data sets with information about the individuals are separate from those with family or the firm)

   - changes in family's or firm's composition (the problem arises when the questions have been asked not only for the year of interview, but also for the previous year)

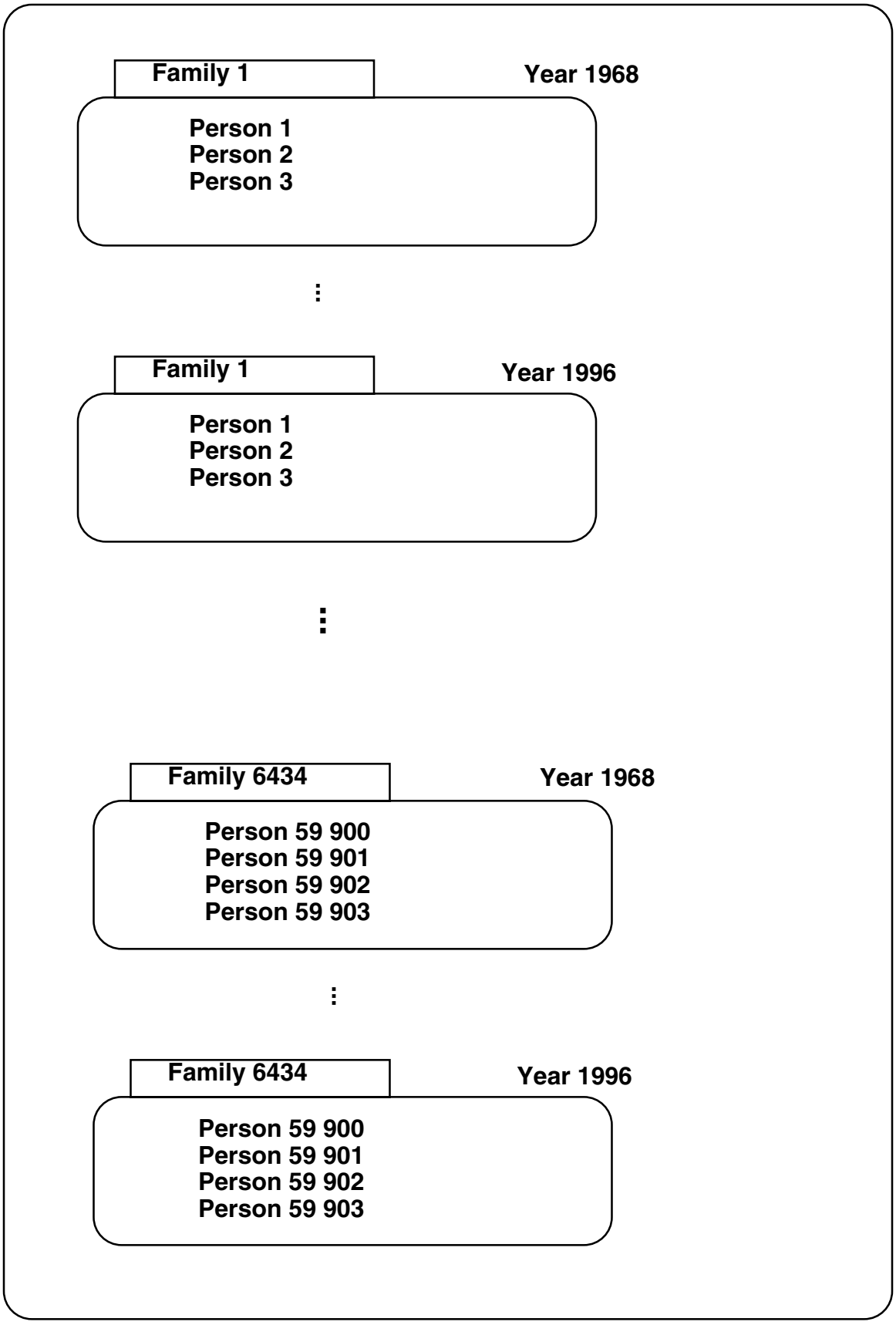
The presentation is very much focused on the Panel Study of Income Dynamics, but the ideas presented could be easily applied to other data sets of the same nature

The Structure of PSID Files

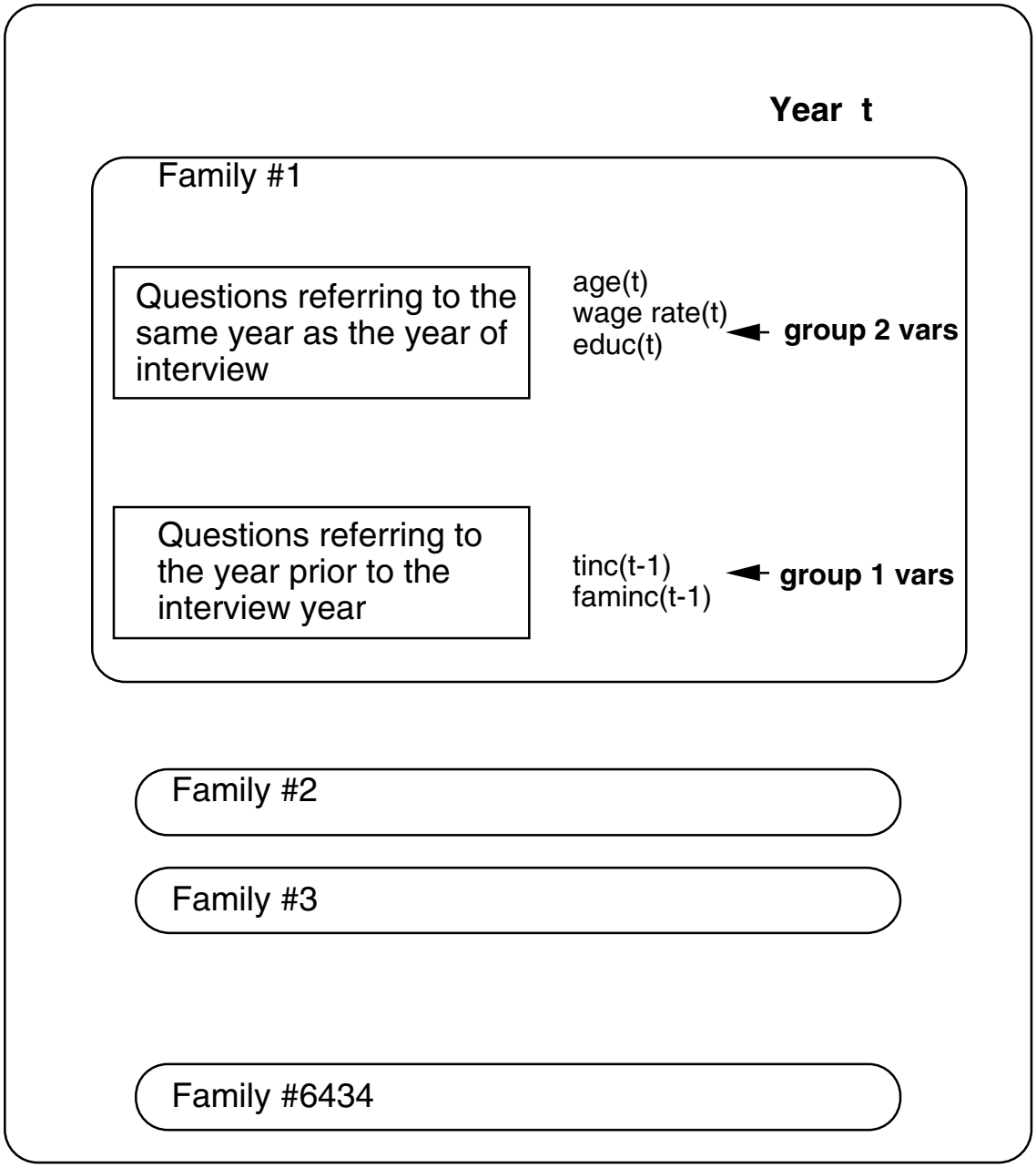


unique identifier - Personal ID
(based on 1968 Family ID number and Person Number in 1968)

Content of the Individual File

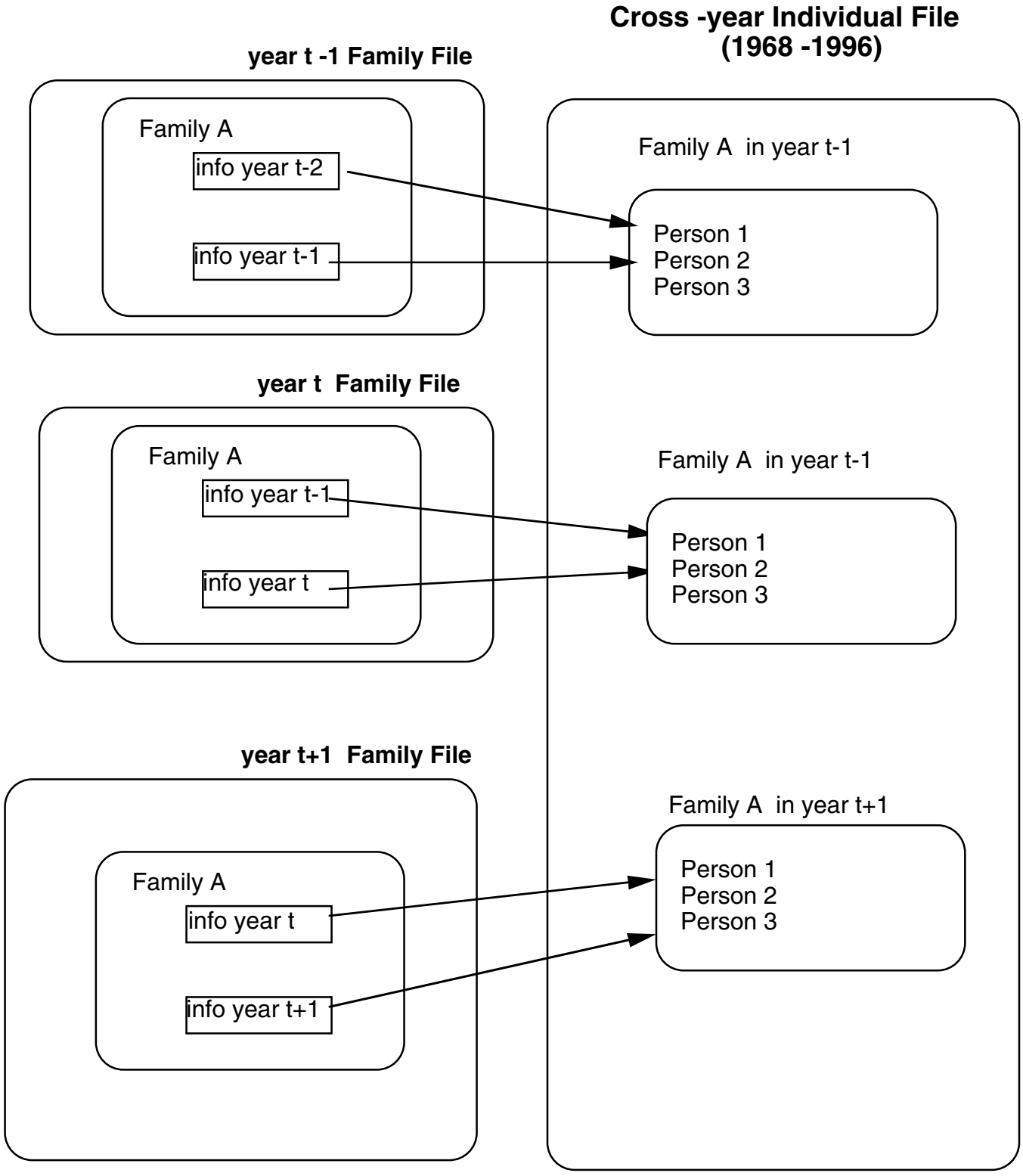


Content of a year t Family File

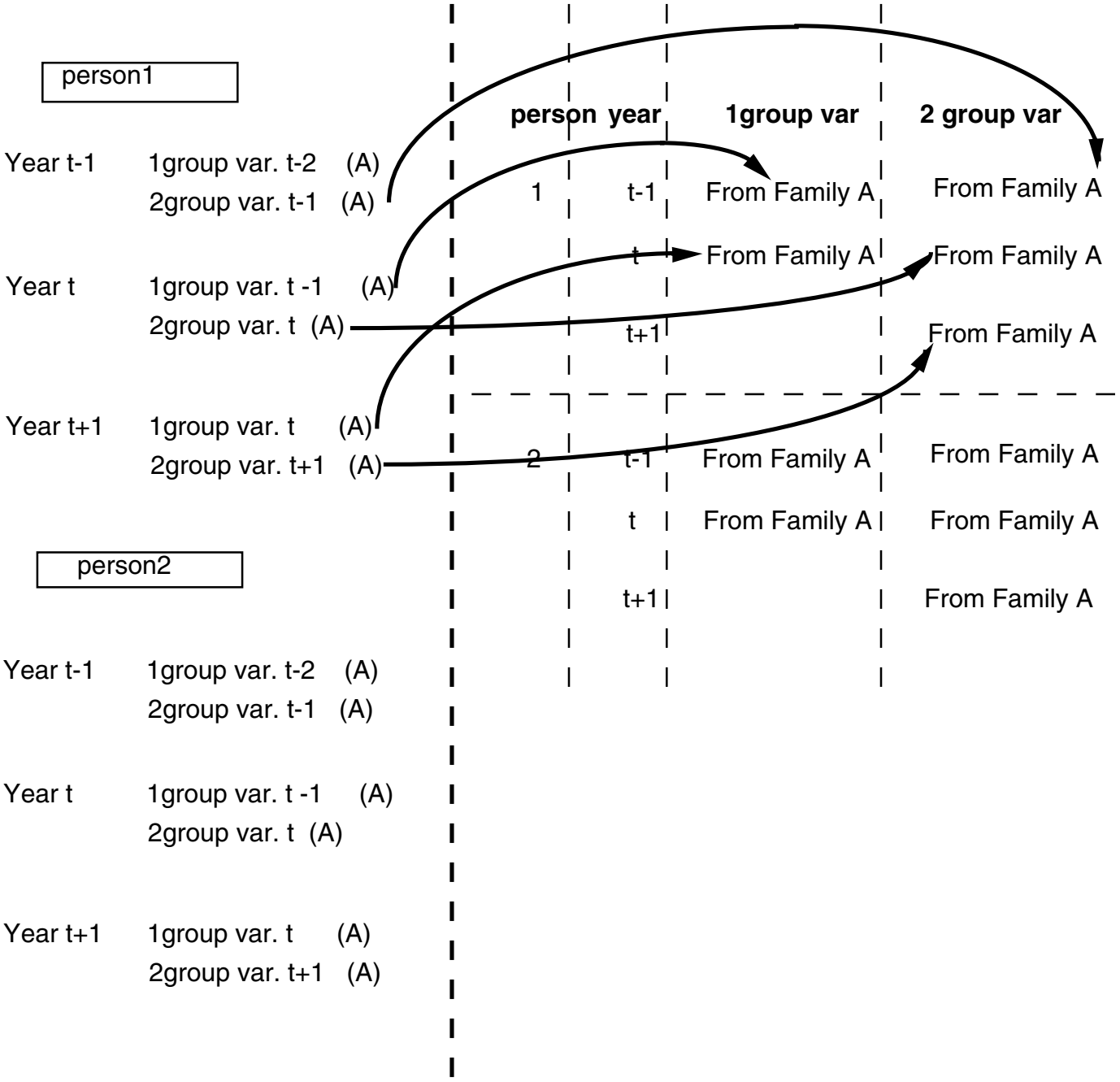


One record per family
(the head answered the questions for himself and for his wife)

MERGING FAMILY DATA TO INDIVIDUALS

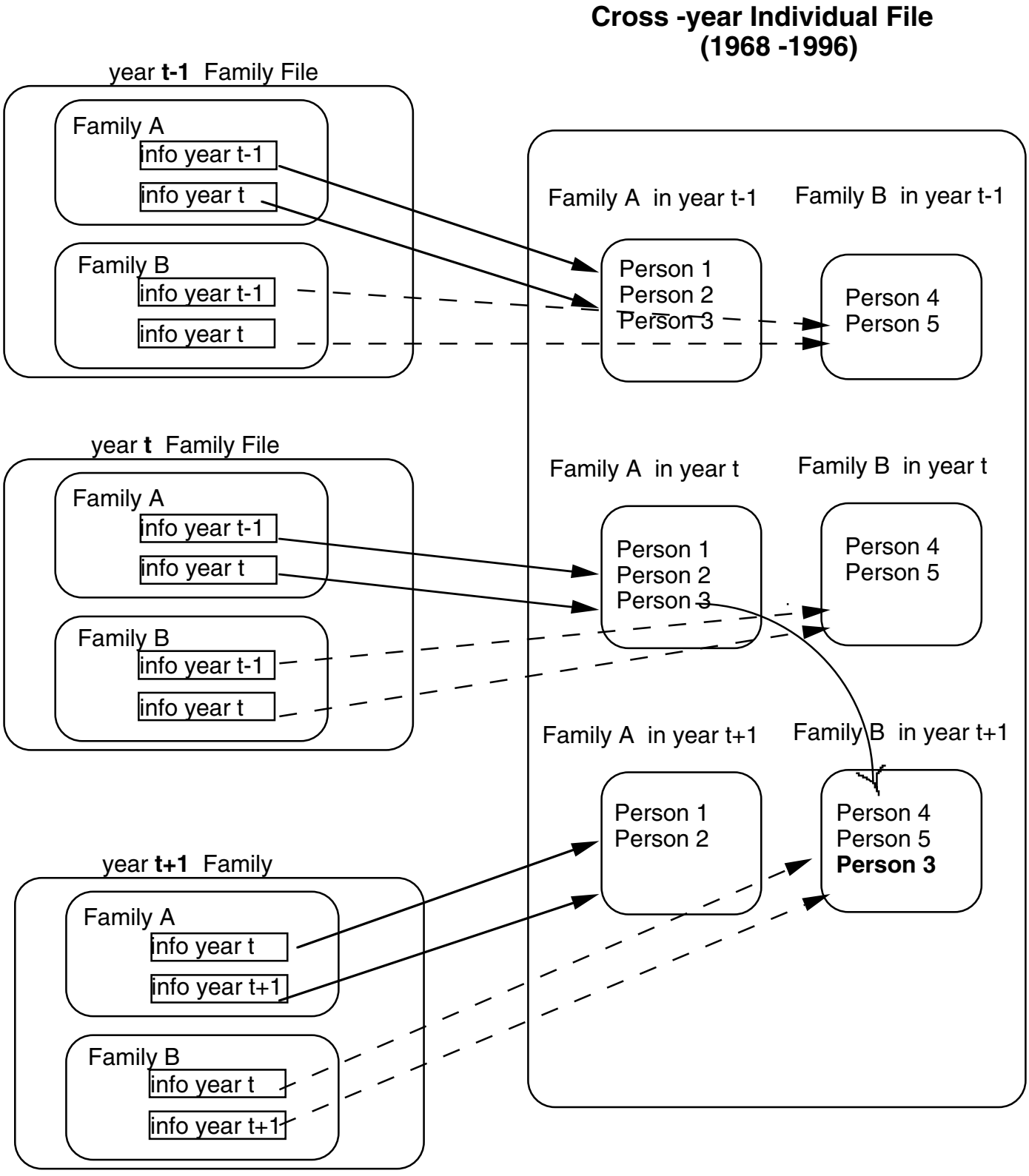


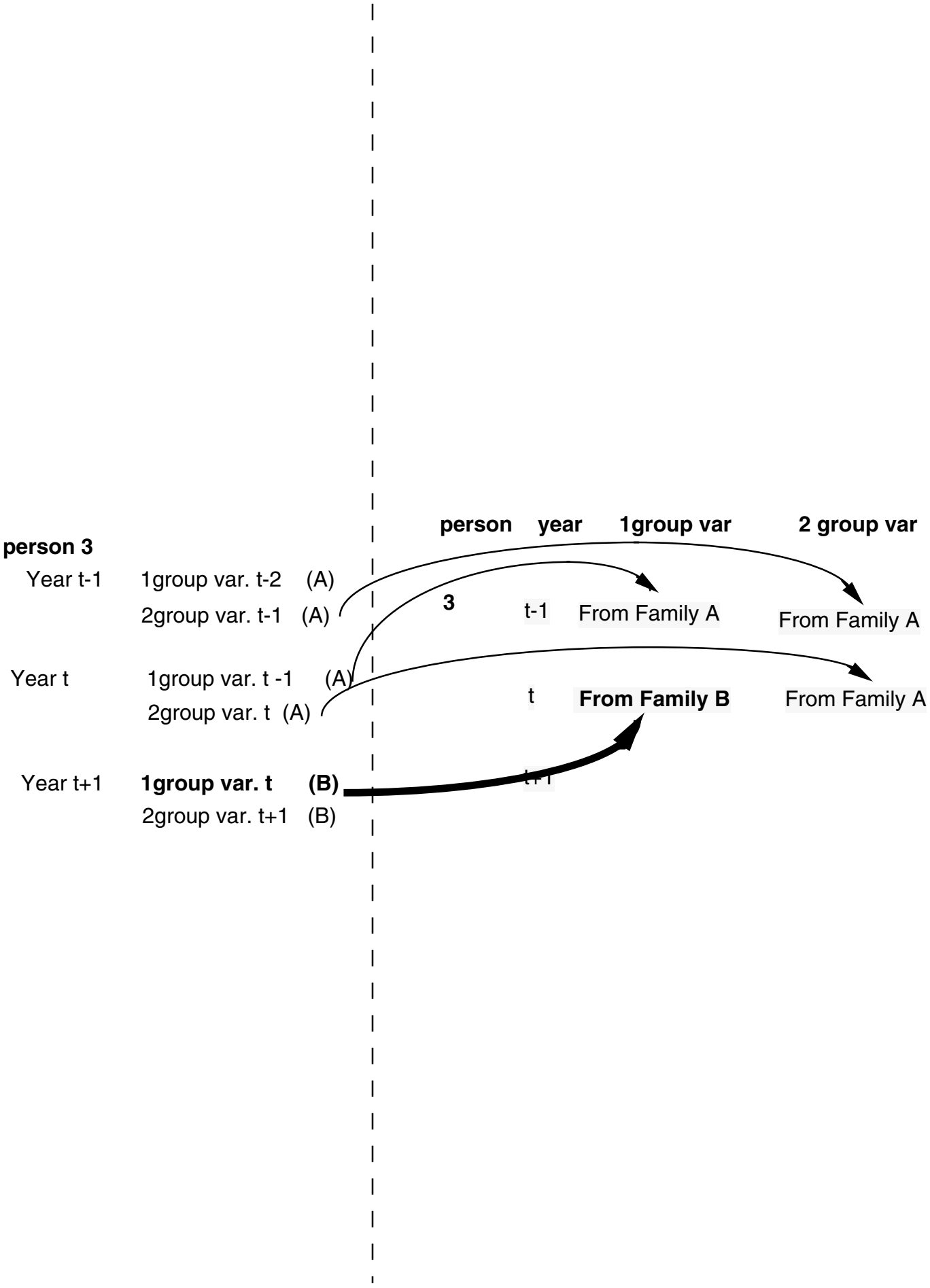
This works perfectly fine under the assumption that the Family Composition does not change!!


person1
person year
1group var
2 group var
Year t-1
1group var. t-2 (A)
2group var. t-1 (A)
1
t-1
From Family A
From Family A
t
From Family A
From Family A
Year t
1group var. t -1 (A)
2group var. t (A)
t+1
From Family A
Year t+1
1group var. t (A)
2group var. t+1 (A)
2
t-1
From Family A
From Family A
t
From Family A
From Family A
person2
t+1
From Family A
Year t-1
1group var. t-2 (A)
2group var. t-1 (A)
Year t
1group var. t -1 (A)
2group var. t (A)
Year t+1
1group var. t (A)
2group var. t+1 (A)

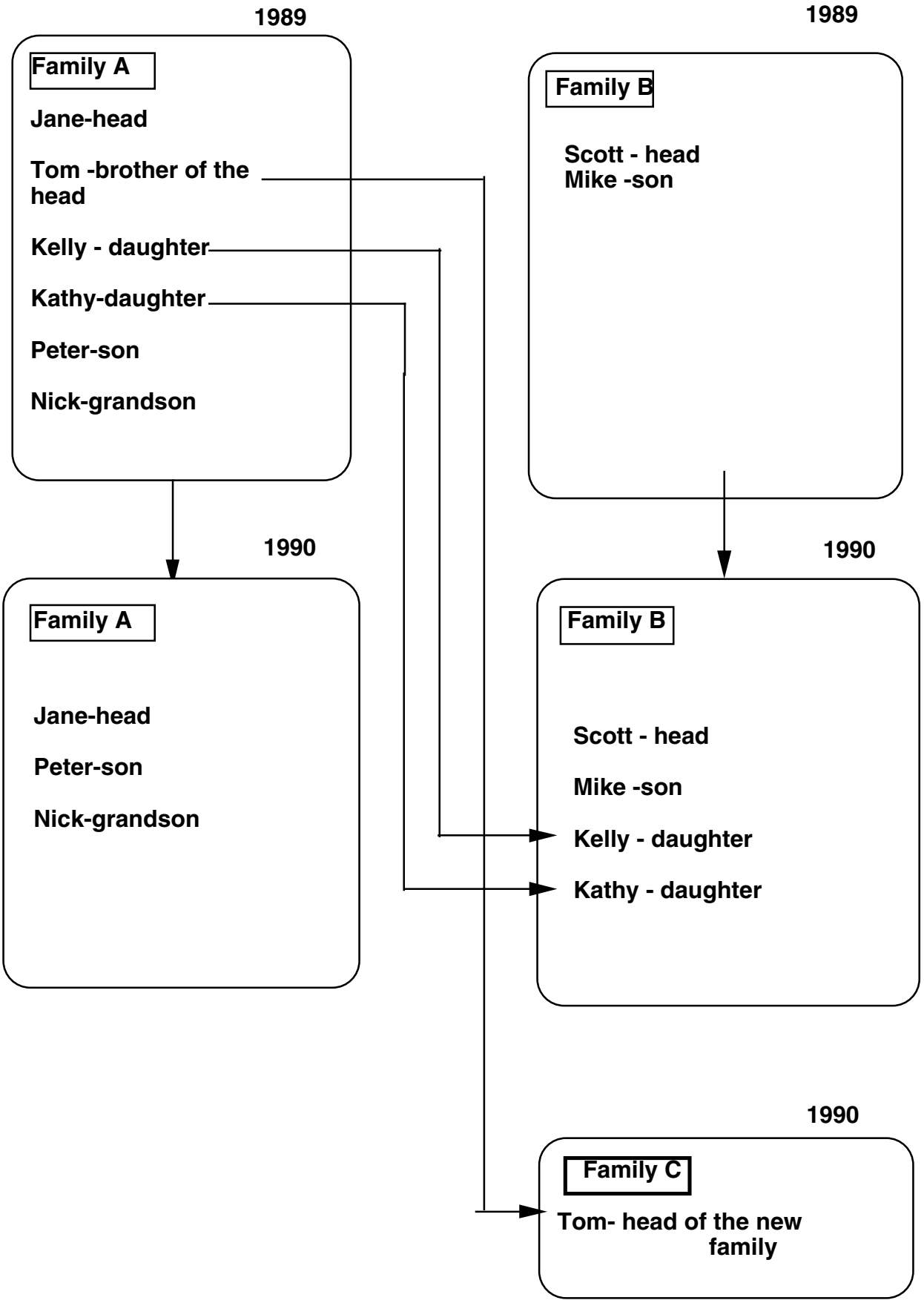
MERGING FAMILY DATA TO INDIVIDUALS

PROBLEMS WHEN FAMILY COMPOSITION CHANGES OVER TIME

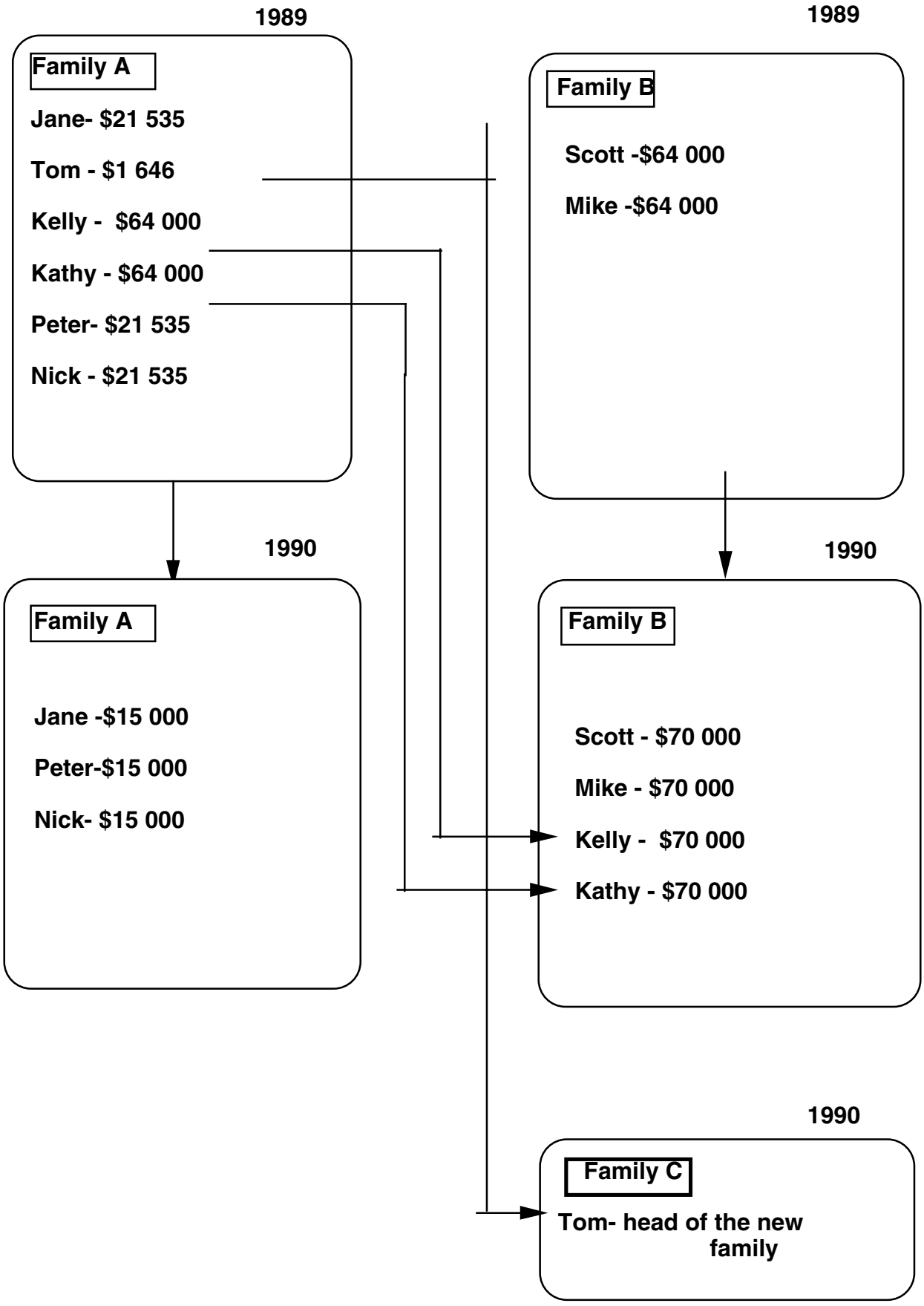



person year 1group var 2 group var
person 3
Year t-1
1group var. t-2 (A)
2group var. t-1 (A)
3
t-1
From Family A
From Family A
Year t
1group var. t -1 (A)
2group var. t (A)
t
From Family B
From Family A
Year t+1
1group var. t (B)
2group var. t+1 (B)
t+1

REAL EXAMPLE FROM PSID



REAL EXAMPLE FROM PSID - Family Income



Extract three files, containing information for all the years of interest:

- Individual File, containing individual level variables
- Family File1, which has info for the same year as the interview year
- Family File2, which has info for the previous of the interview year

(make the files in long-long form)

Merging individual data set with the family data set, which has info for the same year as the interview year

(Individual file+ Family File1)

```
gen pid=id68*1000+pernum68
gen count=_n
reshape long id hwage rearn educ age , i(count) j(year)
```

pid	year	hwage	educ	age
1	1968	$17998	12	27
1	1969	$26989	12	28
.				
.				
.				
2	1968	$20000	16	30
2	1969	$35000	16	31

```
use Individual File.dta
sort pid year

merge pid year using Family File1.dta
```

Merging individual data set with the family data set, which has info for the previous of the interview year

(Individual File + Family File1) + Family File2

(in Family File 2, the family information was attached to every person, but sometimes in a wrong way , we will correct for that)

sort pid year

gen idspl=idsp[_n+1] /* Main Family ID for Splitoff*/
gen movel=move[_n+1] /* Moving Indicator*/

gen a=faminc if idspl==0 & movel==0

Since the value is the same for all persons non-splitoffs in the family, assign the max (or min) to all persons in the family (i.e. these with the same mid in t)

sort mid year /*(mid is the Family ID)*/

egen b=max(a), by (id year)

gen c_faminc=b
replace c_ faminc =. if id==0 /*If a family is non-response in t but not in t+1*/
(we need this to be consistent with the other missing values in t)

merge pid year using (Individual File +Family File1).dta