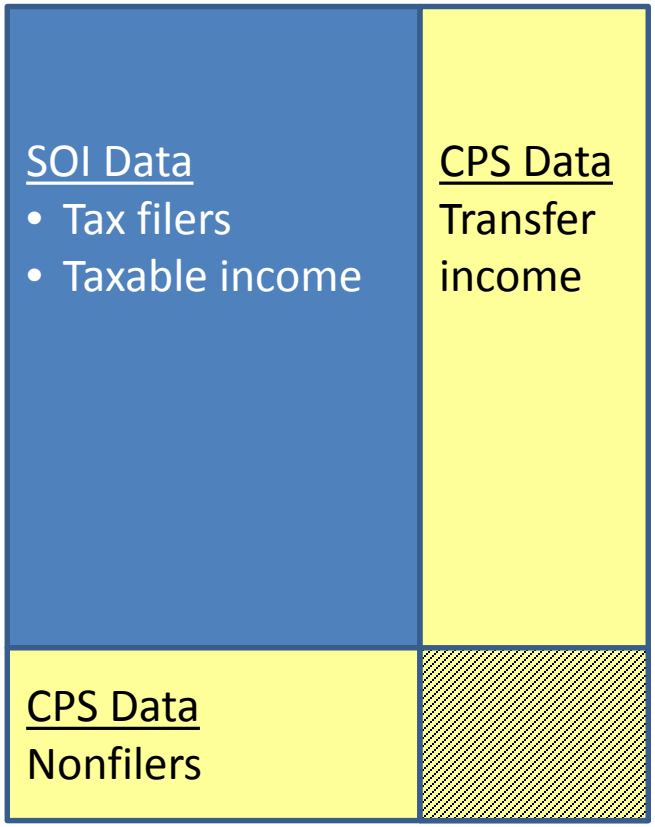
July 21, 2017

# Statistically Matching Administrative Tax Data With Household Survey Data

Presentation at a Workshop Organized by the Washington Center for Equitable Growth

Kevin Perese
Tax Analysis Division

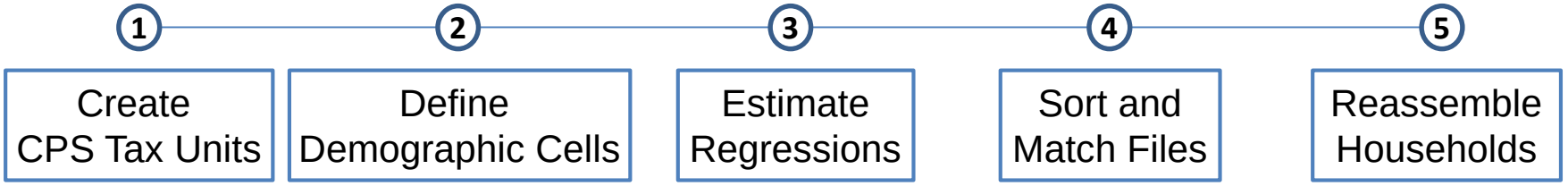
# Why Is It Necessary to Match Tax and Survey Data?



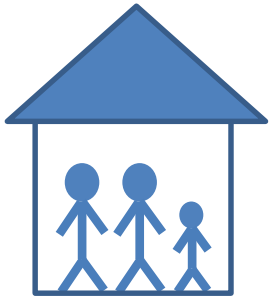

All income sources for nonfilers come from the CPS.

"SOI" is the Internal Revenue Service's Statistics of Income. "CPS" is the Census Bureau's Current Population Survey.

# A Five-Step Process



The unit of analysis in CBO distribution reports is the CPS household.

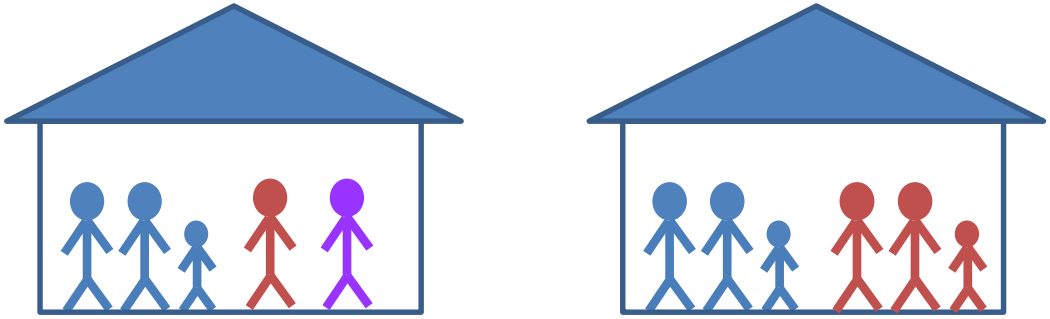


The unit of analysis in CBO distribution reports is the CPS household.

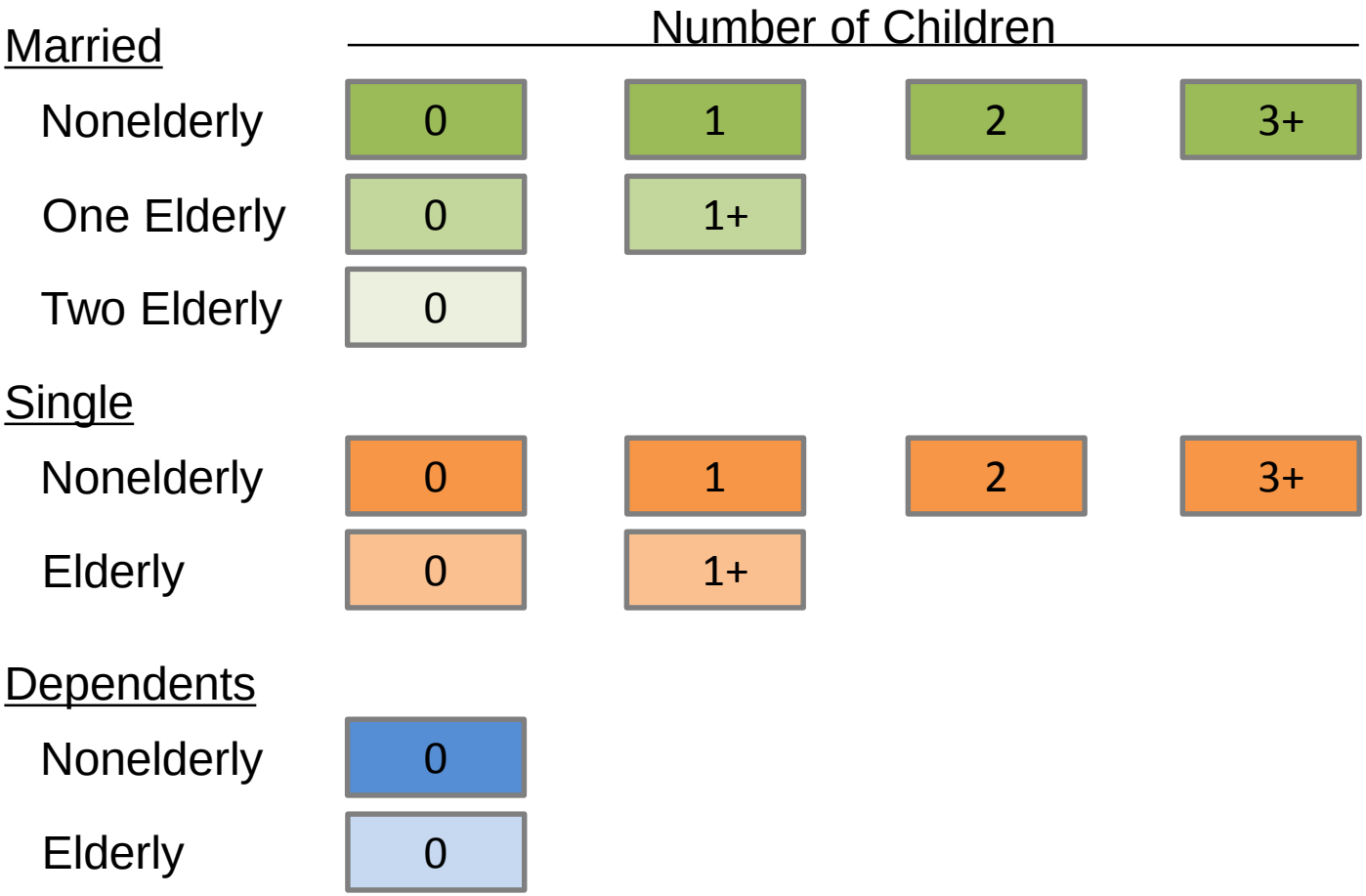
However, there can be multiple tax units in a household.

The unit of analysis in CBO distribution reports is the CPS household.

However, there can be multiple tax units in a household.



An algorithm is used to create tax units based on CPS relationship, age, and income variables.


Number of Children
Married
Nonelderly
0
1
2
3+
One Elderly
0
1+
Two Elderly
0
Single
Nonelderly
0
1
2
3+
Elderly
0
1+
Dependents
Nonelderly
0
Elderly
0

First, using SOI data, define total income.

Total income = Wages
+ Interest and dividends
+ Business income
+ Rental income
+ Unemployment insurance
+ Pension income
+ Capital gains
+ Social Security benefits
+ Other income

Then, in each year and each demographic cell, estimate the following regression (using SOI data):

Total income = $\beta_0$ * Wages
+ $\beta_1$ * Interest and dividends
+ $\beta_2$ * Business income
+ $\beta_3$ * Rental income
+ $\beta_4$ * Unemployment insurance
+ $\beta_5$ * Pension income
+ $\alpha$ * Intercept
+ *Error Term*

Variables that are in both the SOI and the CPS

Capital gains
Social Security benefits
Other income

Finally, calculate predicted total income in the CPS and the SOI, using the estimated regression coefficients:

$$\widehat{\text{Total income}} = \hat{\beta}_0 \text{ * Wages} + \hat{\beta}_1 \text{ * Interest and dividends} + \hat{\beta}_2 \text{ * Business income} + \hat{\beta}_3 \text{ * Rental income} + \hat{\beta}_4 \text{ * Unemployment insurance} + \hat{\beta}_5 \text{ * Pension income} + \hat{\alpha} \text{ * Intercept}$$

**Demographic Cell$_i$**

| CPS File | SOI File |
|---|---|
| Record ID | Record ID |
| 1 | A |
| 2 | B |
| 3 | C |
| 4 | D |
| | E |

Within each demographic cell, each file is sorted from highest to lowest predicted total income.

**Demographic Cell$_i$**

CPS File

| Record ID | Sample Weight |
|---|---|
| 1 | 5 |
| 2 | 3 |
| 3 | 5 |
| 4 | 3 |

SOI File

| Record ID | Sample Weight |
|---|---|
| A | 1 |
| B | 1 |
| C | 1 |
| D | 3 |
| E | 3 |

**Demographic Cell$_i$**

| CPS File | | SOI File | | Merged File | |
|---|---|---|---|---|---|
| Record ID | Sample Weight | Record ID | Sample Weight | Record ID | Sample Weight |
| | | A | 1 | 1A | 1 |
| 1 | 5 | B | 1 | | |
| | | C | 1 | | |
| | | D | 3 | | |
| 2 | 3 | | | | |
| | | E | 3 | | |
| 3 | 5 | | | | |
| 4 | 3 | | | | |

**Demographic Cell$_i$**

| CPS File | | SOI File | | Merged File | |
|---|---|---|---|---|---|
| Record ID | Sample Weight | Record ID | Sample Weight | Record ID | Sample Weight |
| | | A | 1 | 1A | 1 |
| | | B | 1 | 1B | 1 |
| 1 | 5 | C | 1 | | |
| | | D | 3 | | |
| 2 | 3 | | | | |
| | | E | 3 | | |
| 3 | 5 | | | | |
| 4 | 3 | | | | |

**Demographic Cell$_i$**

| CPS File | | SOI File | | Merged File | |
|---|---|---|---|---|---|
| Record ID | Sample Weight | Record ID | Sample Weight | Record ID | Sample Weight |
| | | A | 1 | 1A | 1 |
| | | B | 1 | 1B | 1 |
| 1 | 5 | C | 1 | 1C | 1 |
| | | D | 3 | | |
| 2 | 3 | E | 3 | | |
| 3 | 5 | | | | |
| 4 | 3 | | | | |

**Demographic Cell$_i$**

| CPS File | | SOI File | | Merged File | |
|---|---|---|---|---|---|
| Record ID | Sample Weight | Record ID | Sample Weight | Record ID | Sample Weight |
| | | A | 1 | 1A | 1 |
| | | B | 1 | 1B | 1 |
| 1 | 5 | C | 1 | 1C | 1 |
| | | D | 3 | 1D | 2 |
| 2 | 3 | | | | |
| | | E | 3 | | |
| 3 | 5 | | | | |
| 4 | 3 | | | | |

Pick up the remaining weight on the first CPS record, and split the weight on the fourth SOI record.

**Demographic Cell$_i$**

| CPS File | | SOI File | | Merged File | |
|---|---|---|---|---|---|
| Record ID | Sample Weight | Record ID | Sample Weight | Record ID | Sample Weight |
| | | A | 1 | 1A | 1 |
| | | B | 1 | 1B | 1 |
| 1 | 5 | C | 1 | 1C | 1 |
| | | D | 3 | 1D | 2 |
| | | | | 2D | 1 |
| 2 | 3 | E | 3 | | |
| 3 | 5 | | | | |
| 4 | 3 | | | | |

Pick up the remaining weight on the fourth SOI record, and split the weight on the second CPS record.

**Demographic Cell$_i$**

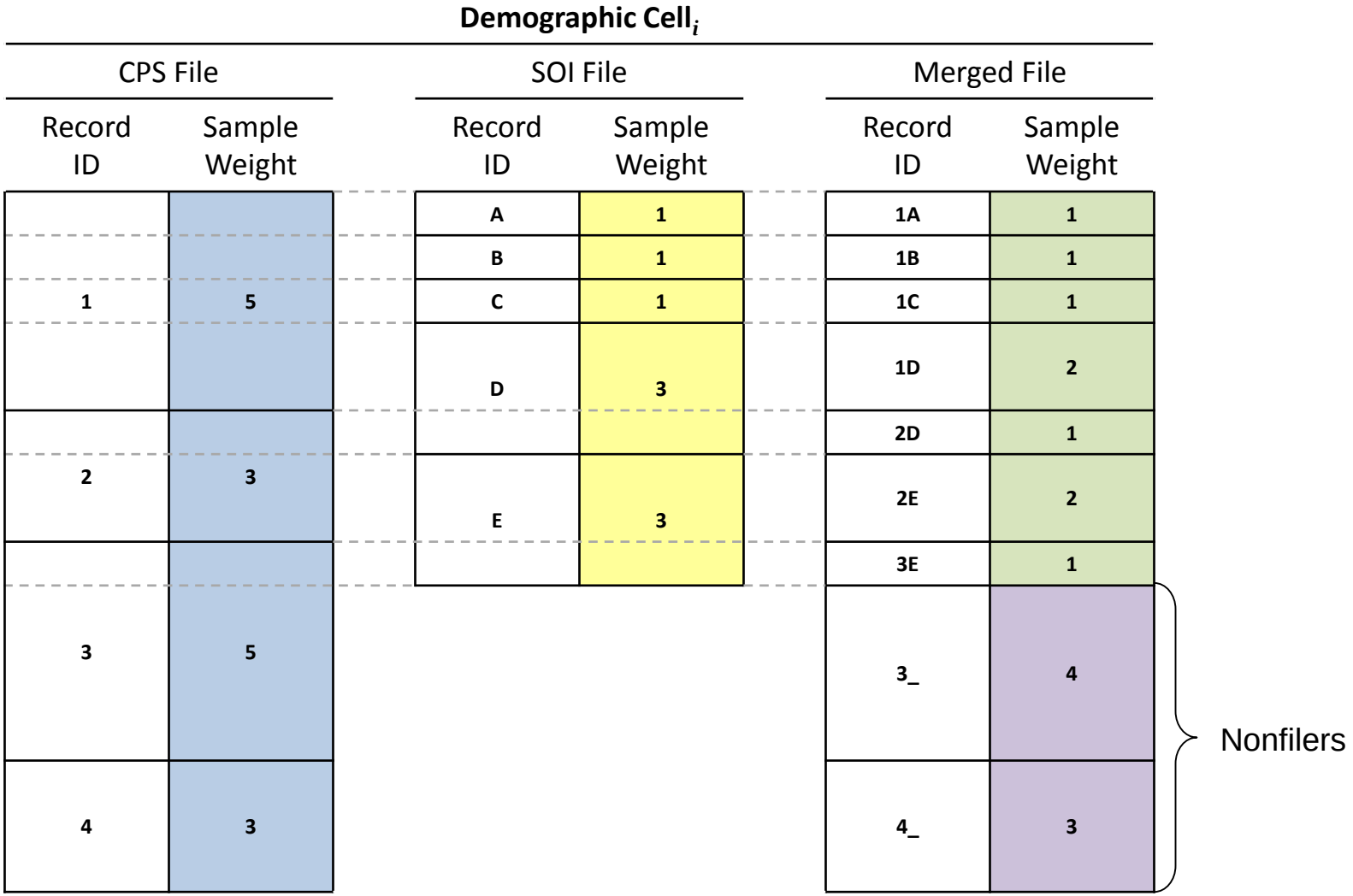
| CPS File | | SOI File | | Merged File | |
|---|---|---|---|---|---|
| Record ID | Sample Weight | Record ID | Sample Weight | Record ID | Sample Weight |
| | | A | 1 | 1A | 1 |
| | | B | 1 | 1B | 1 |
| 1 | 5 | C | 1 | 1C | 1 |
| | | D | 3 | 1D | 2 |
| | | | | 2D | 1 |
| 2 | 3 | E | 3 | 2E | 2 |
| 3 | 5 | | | | |
| 4 | 3 | | | | |

And so on…

**Demographic Cell$_i$**

| CPS File | | SOI File | | Merged File | |
|---|---|---|---|---|---|
| Record ID | Sample Weight | Record ID | Sample Weight | Record ID | Sample Weight |
| | | A | 1 | 1A | 1 |
| | | B | 1 | 1B | 1 |
| 1 | 5 | C | 1 | 1C | 1 |
| | | D | 3 | 1D | 2 |
| | | | | 2D | 1 |
| 2 | 3 | E | 3 | 2E | 2 |
| | | | | 3E | 1 |
| 3 | 5 | | | | |
| 4 | 3 | | | | |

…until all SOI records (portions of SOI sample weights) have been exhausted.


Demographic Cell$_i$
CPS File
Record ID
Sample Weight
1
5
2
3
3
5
4
3
SOI File
Record ID
Sample Weight
A
1
B
1
C
1
D
3
E
3
Merged File
Record ID
Sample Weight
1A
1
1B
1
1C
1
1D
2
2D
1
2E
2
3E
1
3_
4
4_
3
Nonfilers

**Demographic Cell$_i$**

CPS File

| Record ID |
| --- |
| 1 |
| 2 |

A household with two tax units

**Demographic Cell$_i$**

| CPS File | SOI File | Merged File |
|---|---|---|
| Record ID | Record ID | Record ID |
| 1 | A | 1A |
| 2 | B | 1B |
| | C | 1C |
| | D | 1D |
| | E | 2D |
| | | 2E |

A household with two tax units

**Demographic Cell$_i$**

| CPS File | SOI File | Merged File | Household File |
|---|---|---|---|
| Record ID | Record ID | Record ID | Record ID |
| 1 | A | 1A | 1A-2D |
| 2 | B | 1B | 1A-2E |
| | C | 1C | |
| | D | 1D | 1B-2D |
| | E | | 1B-2E |
| | | 2D | |
| | | 2E | 1C-2D |
| | | | 1C-2E |
| | | | 1D-2D |
| | | | 1D-2E |

A household with two tax units

The Household file has every combination of CPS-SOI matches in the Merged file, with each household record getting a scaled weight so that the sum of weights is the same as the original CPS household weight.

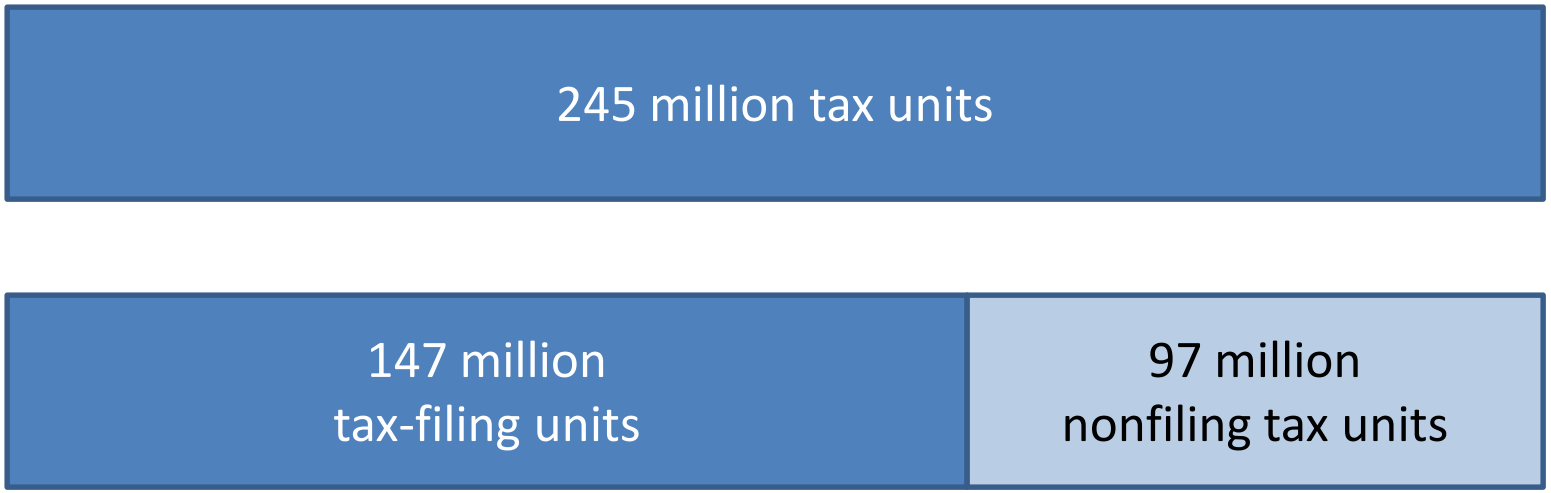
# A Taxonomy of Tax Units

In 2013, there were:

245 million tax units

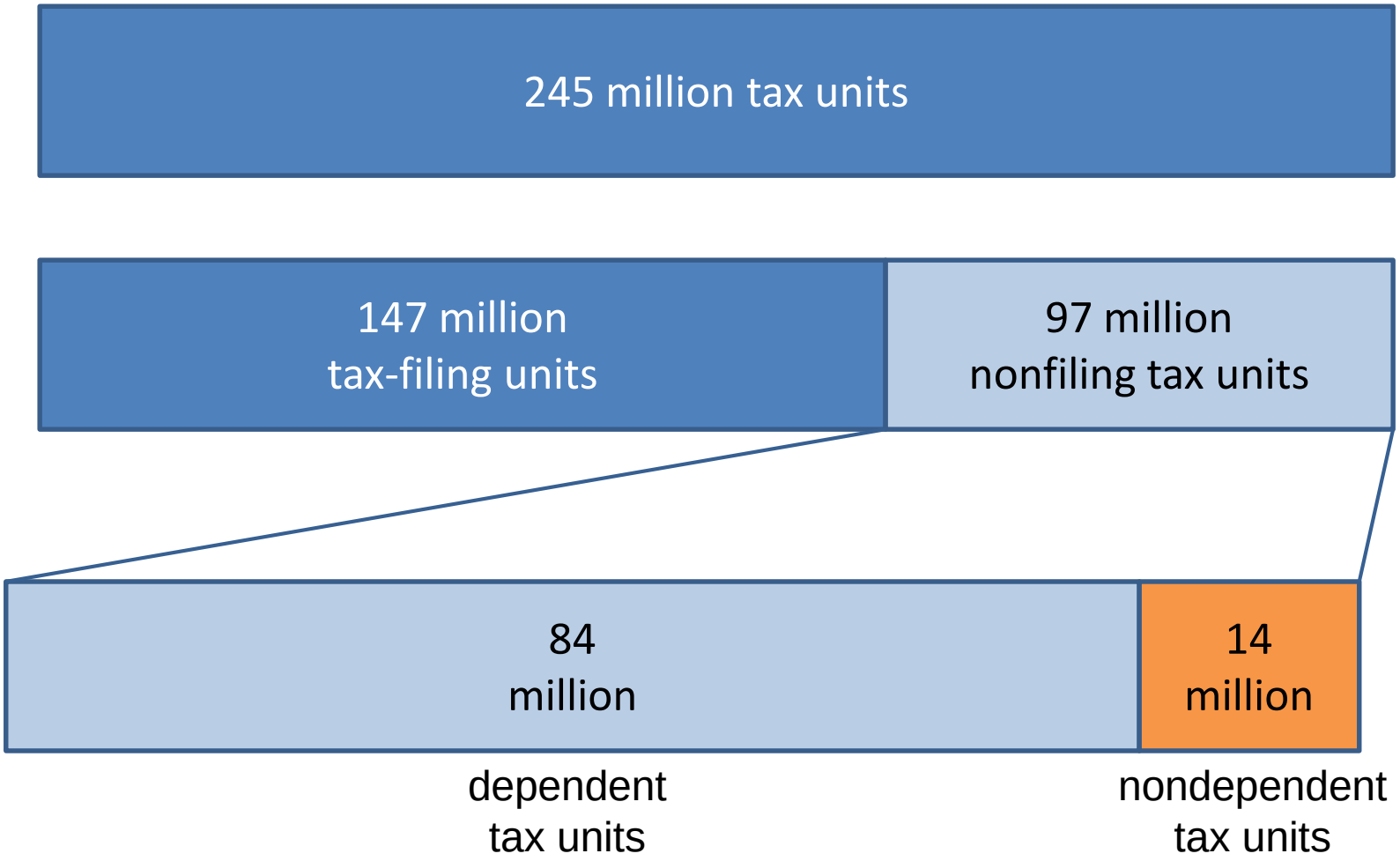
# A Taxonomy of Tax Units

In 2013, there were:



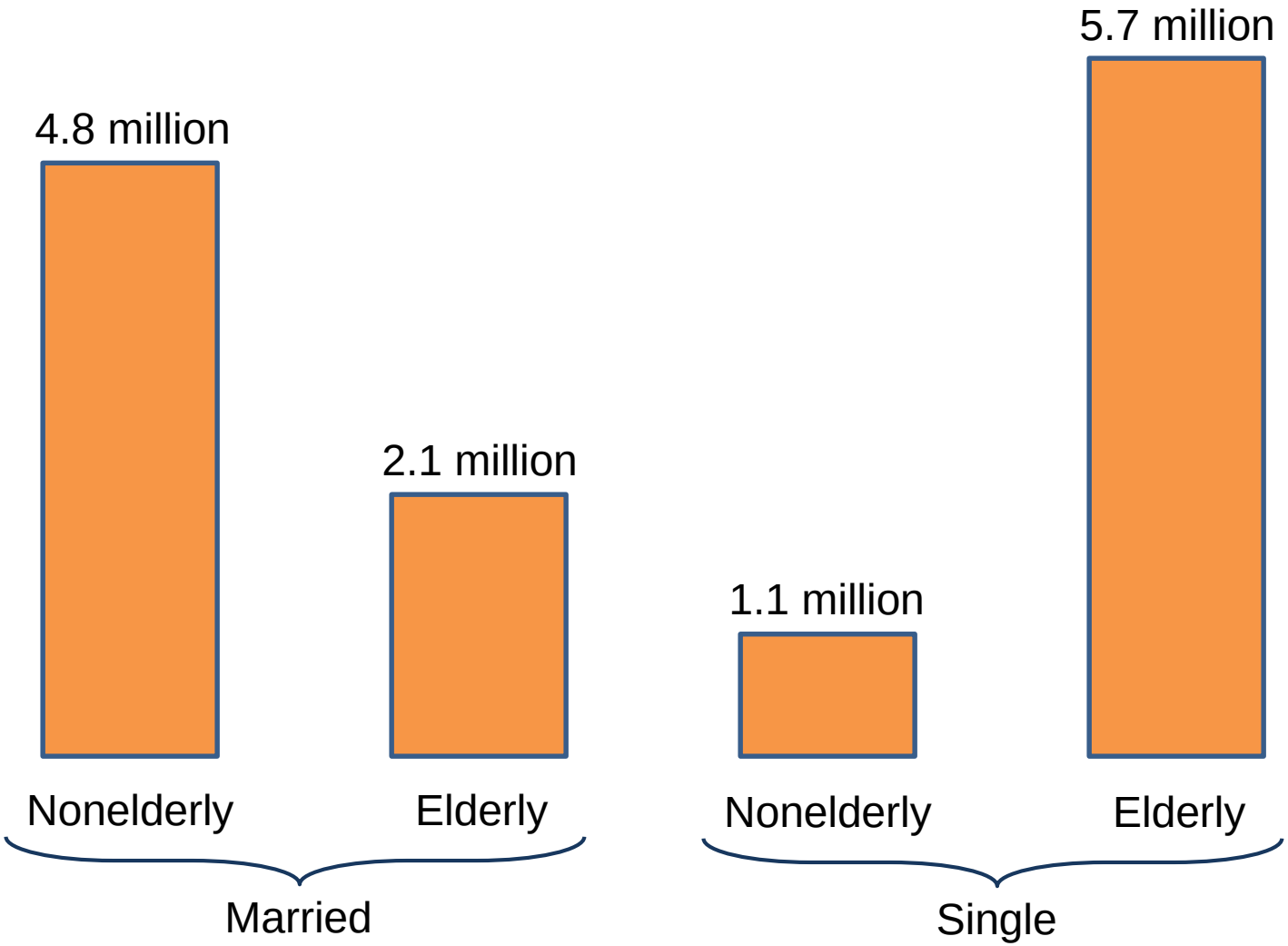
# A Taxonomy of Tax Units

In 2013, there were:



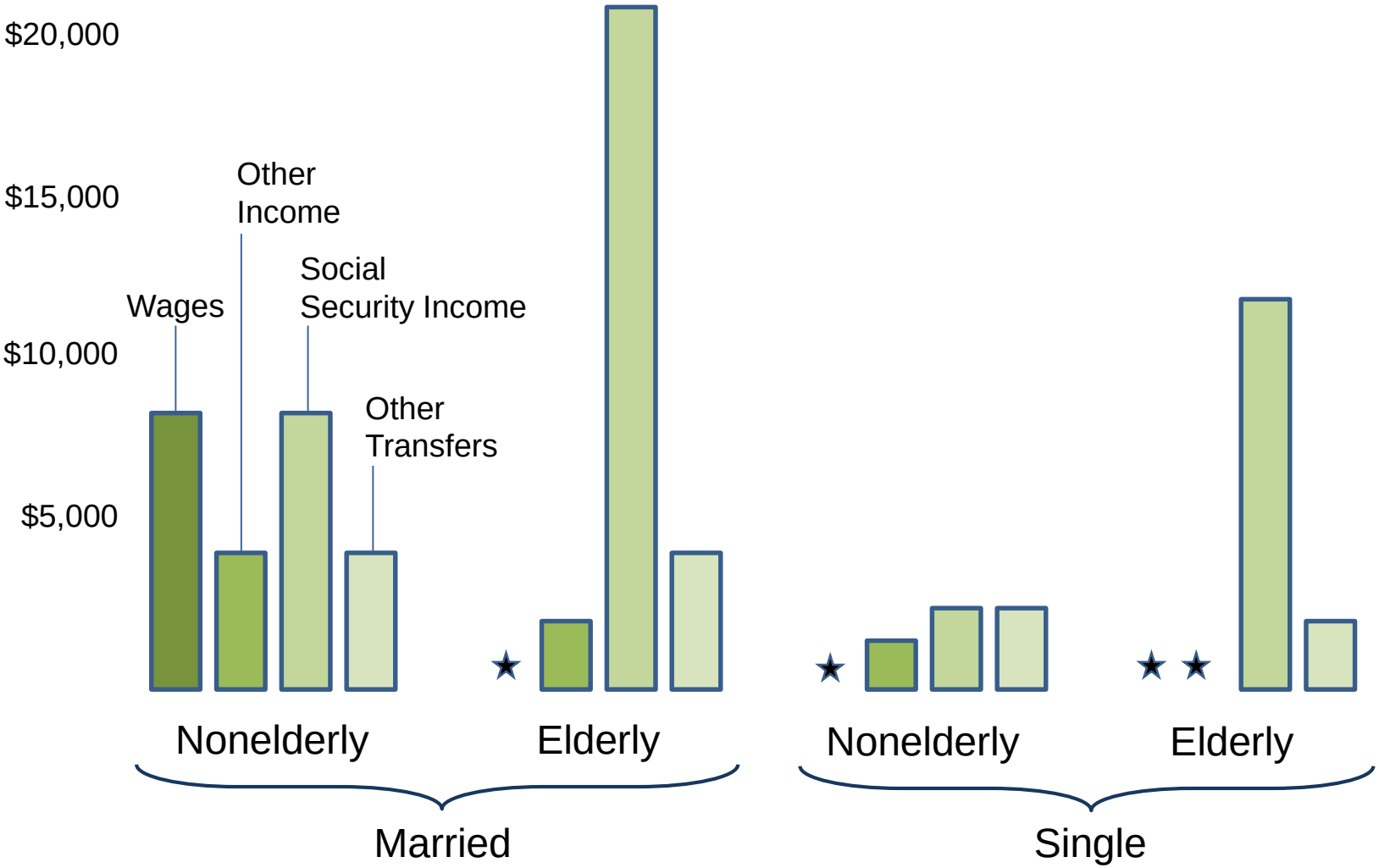
# Some Results

**Number of Nondependent, Nonfiling Tax Units**



# Some Results

**Average Income of Nondependent, Nonfiling Tax Units**




★ = less than $500.