Balancing Respondent Confidentiality and Data User Needs

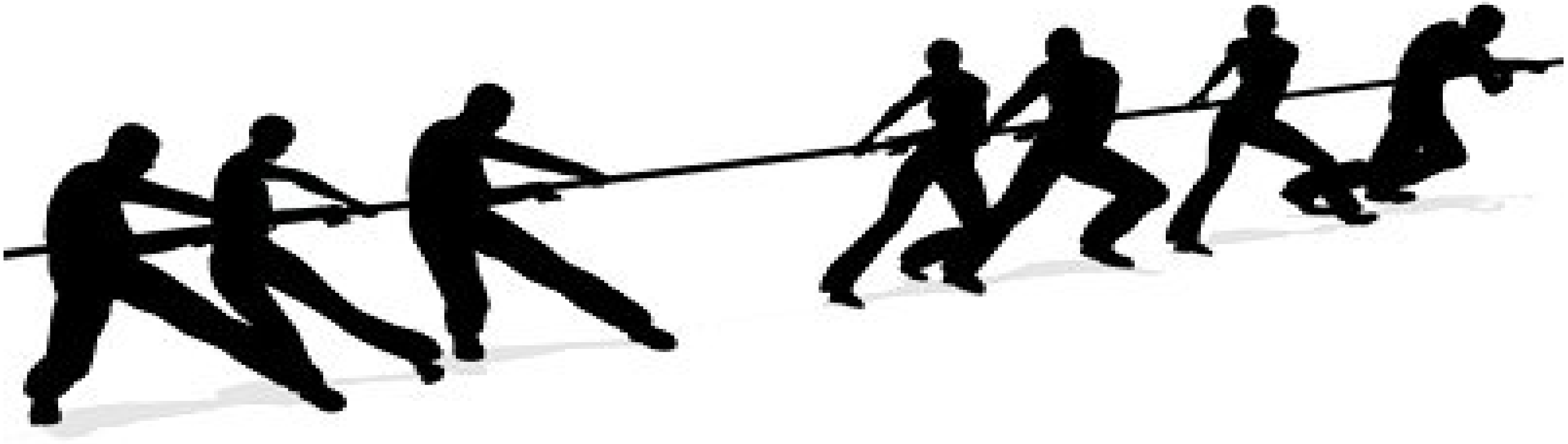
Aaron E. Cobet

Consumer Expenditure Surveys Microdata Users Workshop

July 20, 2017

What is the Issue?

- Conflicting goals
  - Maximize data access
  - Protect respondents identity



Why is Confidentiality Important?

- Ensure trust of respondents for their future cooperation
- Ethical responsibility to protect respondent confidentiality
- It’s the law

What is Title 13?

- U.S. Code: Title 13 allows the government to take a census and provides directives for its administration and enforcement.
- People who took the oath of office who wrongfully disclose information protected under Title 13 are subject to a fine of up to $250,000 or up to 5 years in prison.
- Census and CE staff need Title 13 clearance.

Title 13 Training

- CE staff gain access to internal data *after* completing 2 steps:
  1. Pass a background check by Census
  2. Take the Title 13 training

- CE staff are required to annually retake Title 13 training and pass a knowledge check to maintain Special Sworn Status

Who Determines Disclosure Threats?

- Disclosure Review Board (DRB) by the Census Bureau

United States Census Bureau

How Could Microdata Reveal Respondents' Identity?

- Small PSUs
- High income
- Extreme expenditures

How to Protect Respondents' Confidentiality?

- BLS and Census Bureau conceal information that *could* reveal respondents identity.

How to Protect Respondents' Confidentiality?

Two stages:

- Census removes *direct* identifiers, i.e. addresses
- BLS suppresses *indirect* identifiers, i.e. high expenses

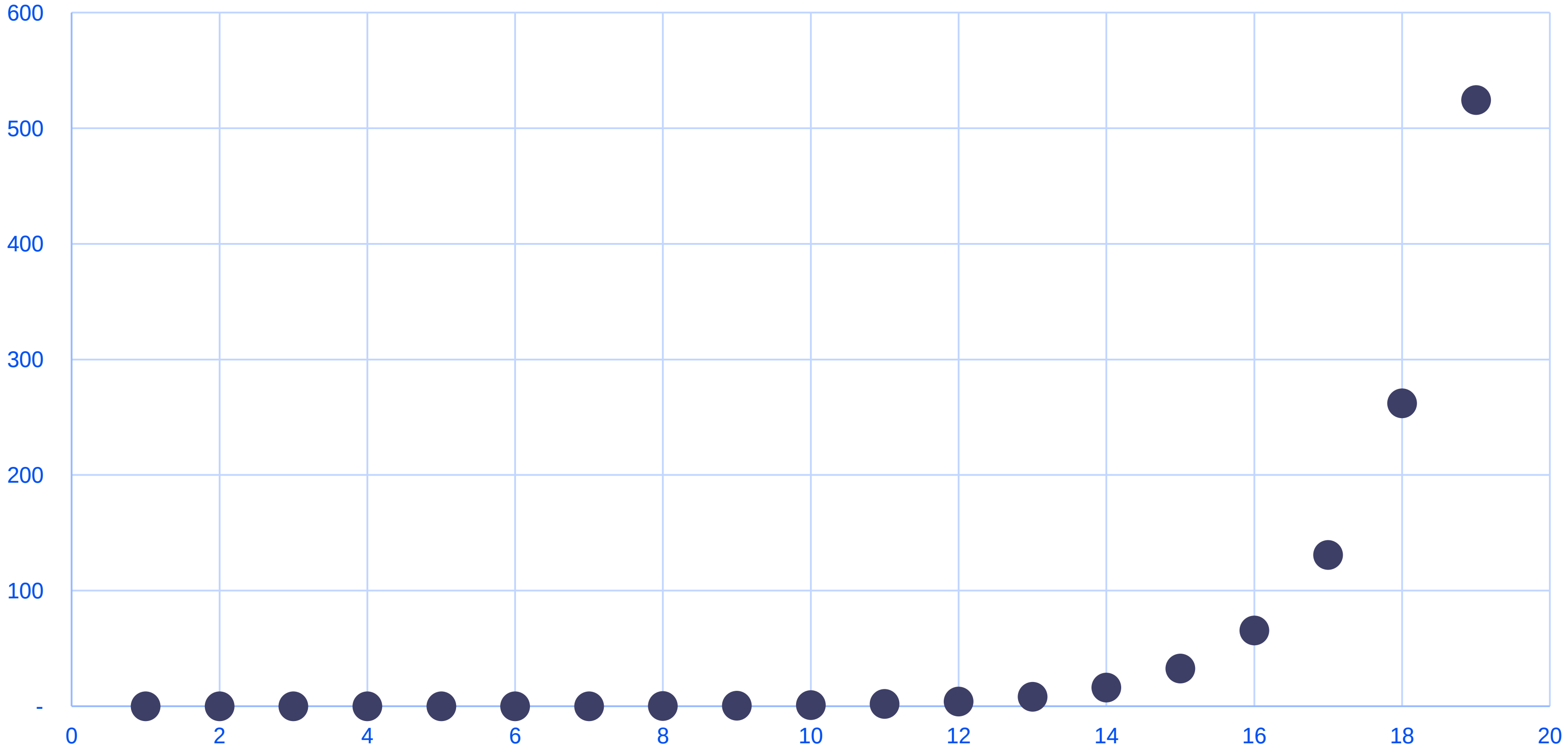
How to Conceal Indirect Identifiers?

- **Topcoding**: Provide average numerical value that are above a threshold
- **Recoding**: Change metadata but provide numerical value
- **Suppression**: Delete numerical value only or entire record

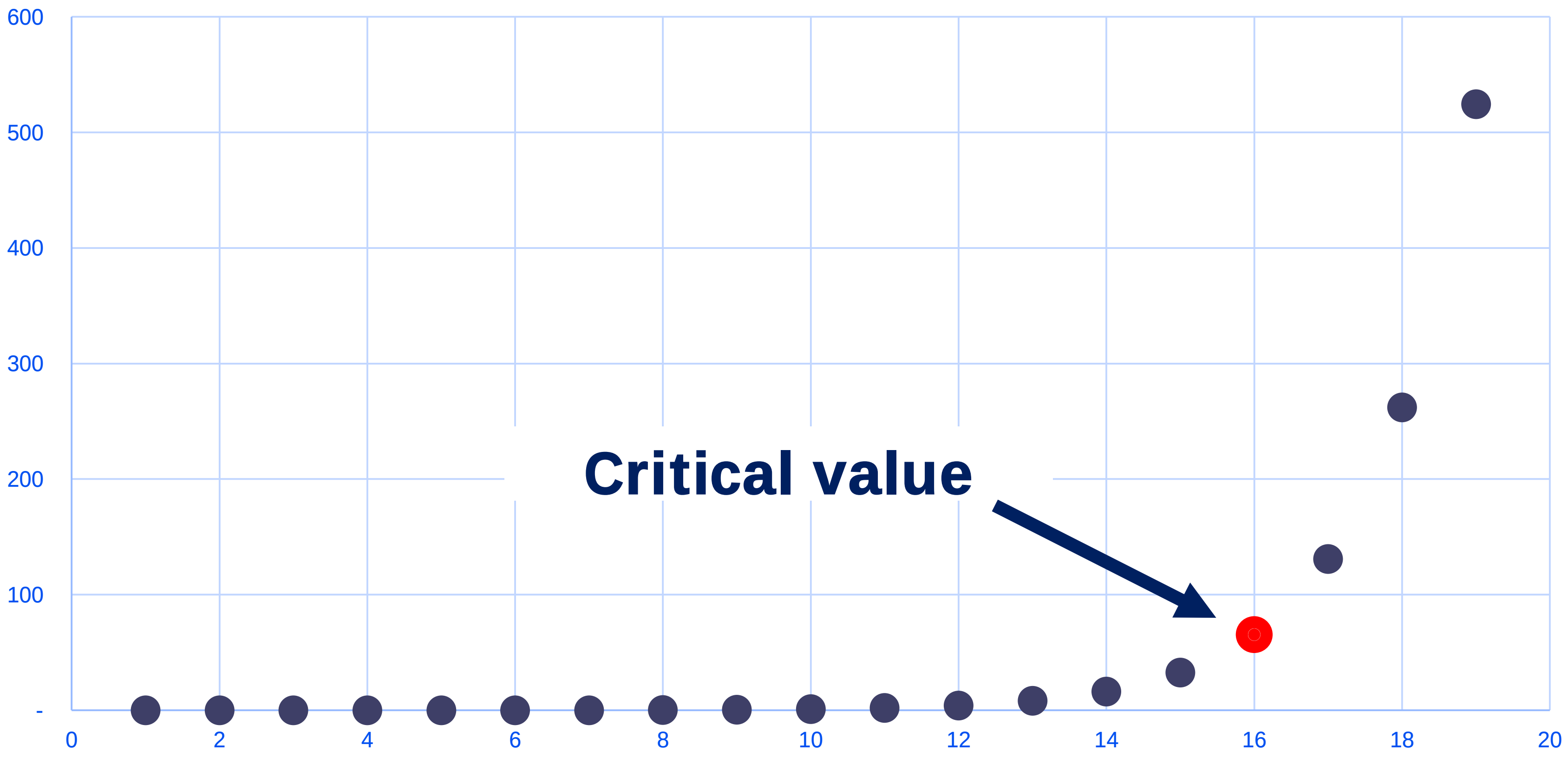
How do we Topcode?

- Determine critical value
- Find values exceeding critical value
- Average values exceeding critical value
- Replace values with top-coded values

Topcoding Example

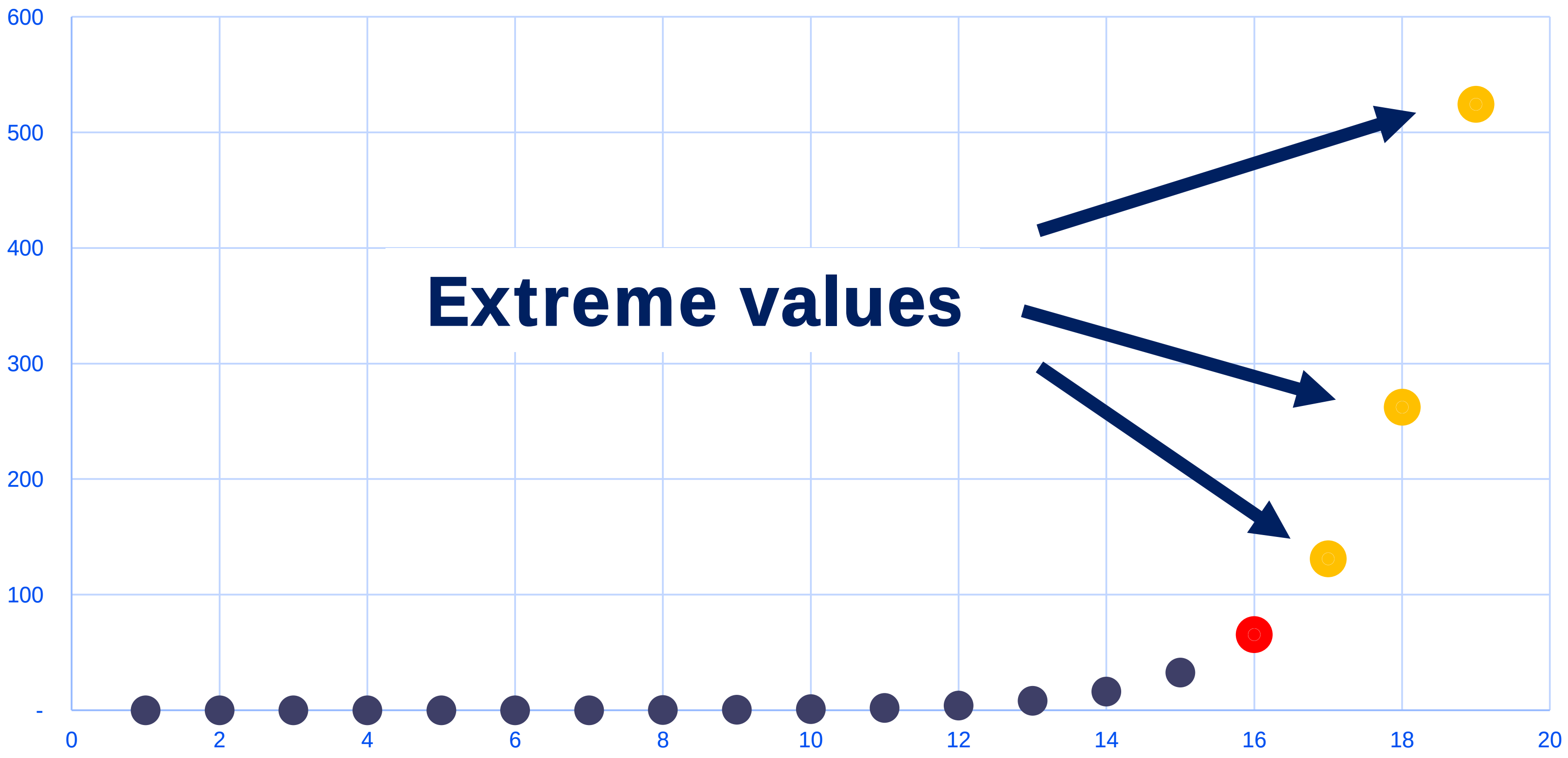


Topcoding Example



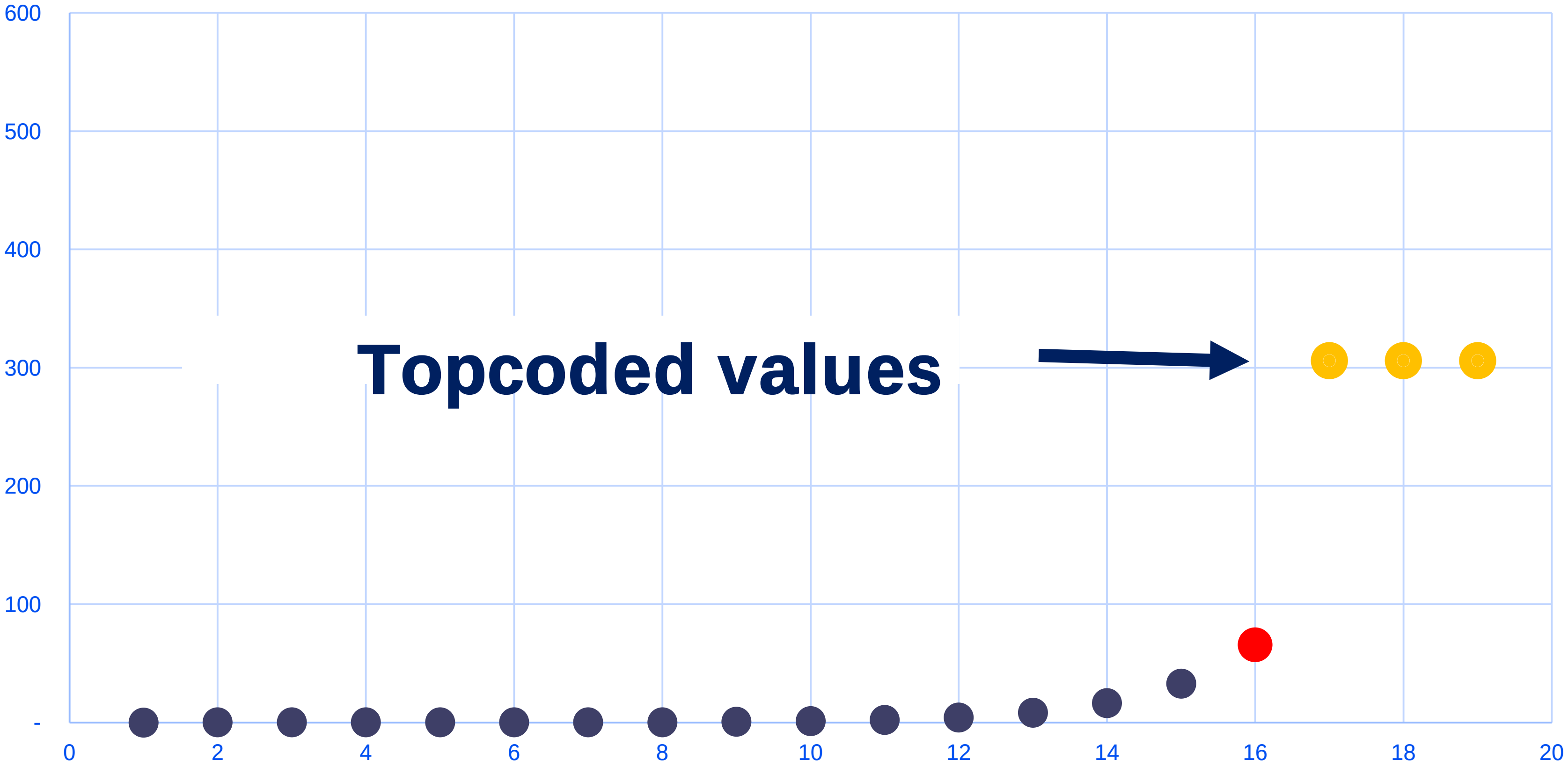
Critical value

Topcoding Example



Extreme values

Topcoding Example



Topcoded values

How to Determine Critical Values?

- **Percentiles:** If sample matches population
  - Expenditures: 99.5 %
  - Income: 97.0 %
- **Outside sources:** If sample differs from population

How to Conceal Indirect Identifiers?

- **Top-coding**: Provide average of expenditures above a threshold
- **Re-coding**: Change metadata but provide numerical data
- **Suppression**: Delete numerical data or entire record

How do we Recode?

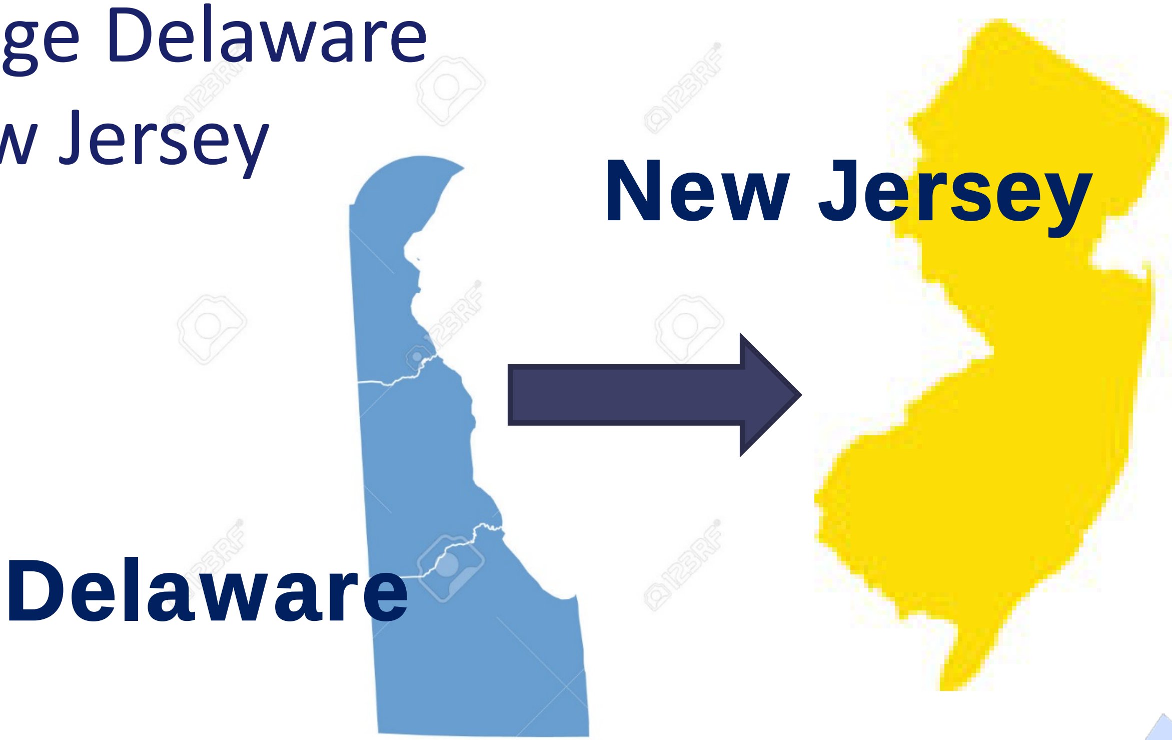
- Find metadata that meet criteria
- Determine method:
  - Generalize
  - Change
- Replace original metadata with recoded metadata

Re-coding: Generalize Information

- Broaden production year of cars
  - From Toyota Corolla 1999
  - To Toyota Corolla 1990s

Re-coding: Change information

- Change states to comparable states
  - Change Delaware to New Jersey



How to Conceal Indirect Identifiers?

- **Top-coding**: Provide average of expenditures above a threshold
- **Re-coding**: Change metadata but provide numerical data
- **Suppression**: Delete numerical data or entire record

Suppression

- Erase numerical data and leave metadata
  - Blank out numerical values of infrequent purchases
  - Example: Boat purchase

Suppression

- Complete eradication of numerical and metadata
  - Erase entire record
  - Example: Airplane purchase

Reverse Engineering

What’s X?

$$5 = 3 + X$$

How to Prevent Reverse Engineering?

Prevent users to deduce protected information from available data

1. Find protected values
2. Protect them in all locations
3. Protect related values

Reverse Engineering

- Scenarios
  - Within file
  - Across files

Reverse Engineering: Within File

- Income = Wages + taxes
- 1000 = 800 + 200
- 1000 = 750 + 200
- 950 = 750 + 200

- Critical value: 700
- Topcode value: 750

Wages **exceeds** the critical value

Reverse Engineering: Within File

- Income = Wages + taxes
- 1000 = 800 + 200
- 1000 = 750 + 200
- 950 = 750 + 200

- Critical value: 700
- Topcode value: 750

Wages **match** the critical value

Reverse Engineering: Within File

- Income = Wages + taxes
- 1000 = 800 + 200
- 1000 = 750 + 200
- 950 = 750 + 200

- Critical value: 700
- Topcode value: 750

Wages and taxes **match** the income

Reverse Engineering:
Across Files

- **Income:** Topcoded income in FMLI
  => Topcode associated UCCs in ITBI

- **Expenditure:** Topcoded expenditures in EXPN and FMLI
  => Topcode associated UCCs in MTBI

How Do We Document?

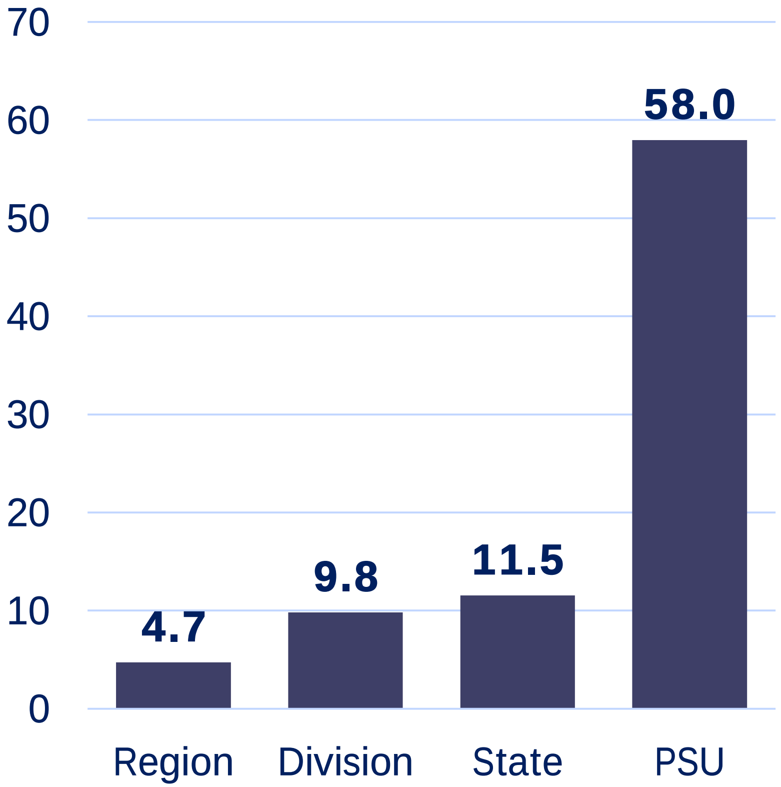
- Flag values
  - **T**: Topcoded value
  - **D**: Valid value

Impact of topcoding

- CE topcodes few observations
- Most affected data slices:
  - Geographic data non-self representing cities
  - Income for high earners.

Impact of Suppression of Geographic variables, Percent

- Almost 60 % of PSUs suppressed
- Below 15 % of states, divisions, and regions suppressed



Source: FMLI and FMLD files for 2015.

Additional Information

- Protection of respondent confidentiality provides additional information on protecting the confidentiality of respondents.

Thank You!

Aaron Cobet
Senior Economist
Consumer Expenditure Surveys
(202)-691-5018
Cobet.Aaron@bls.gov