Total Error Frameworks for Integrating Probability and Nonprobability Data

Paul P. Biemer
RTI International and
the University of North Carolina – Chapel Hill

Presented on Friday, June 7, 2019 at

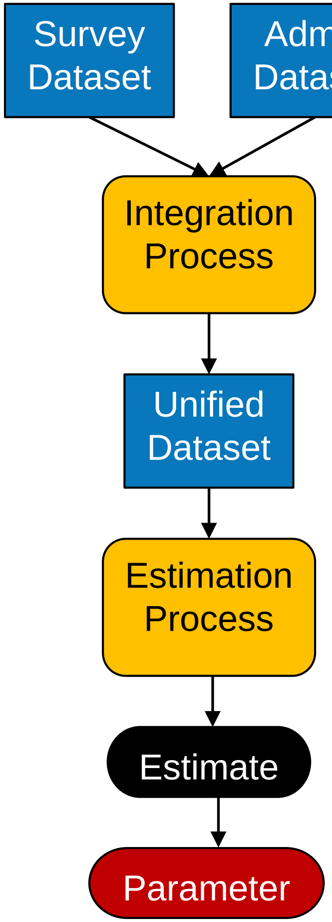
Outline

- Generic data integration process to produce
  - Integrated data sets
  - Hybrid estimates
- An error framework for generic data sets
- An error framework for “hybrid” estimates
- Illustration from 2015 U.S. Residential Energy Consumption Survey

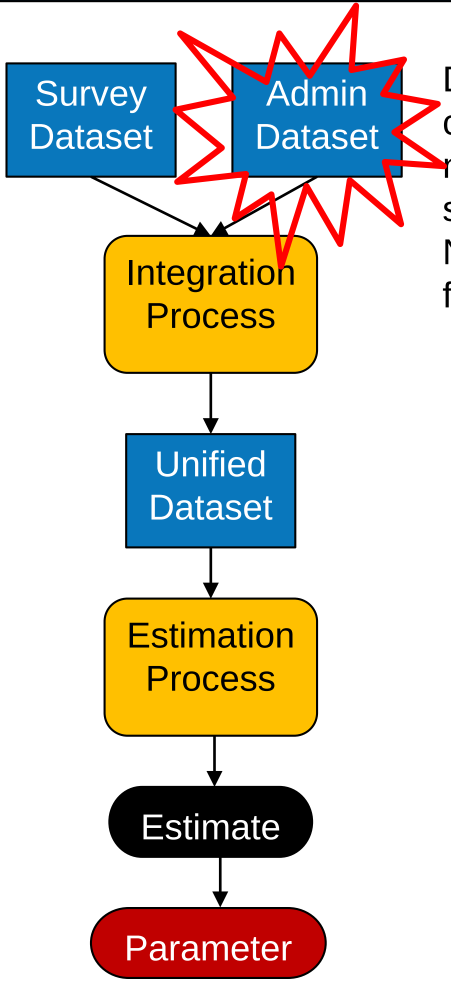
Presentation draws heavily from:

Biemer, P. and Amaya, A. (in press). “Error frameworks for found data,” in Hill, Biemer, Buskirk, Japec, Kirchner, Kolenikov, and Lyberg (eds.) ***Big Data Meets Survey Science: A Collection of Innovative Methods***, John Wiley & Sons, Hoboken, NJ.

The Hybrid Estimation Process

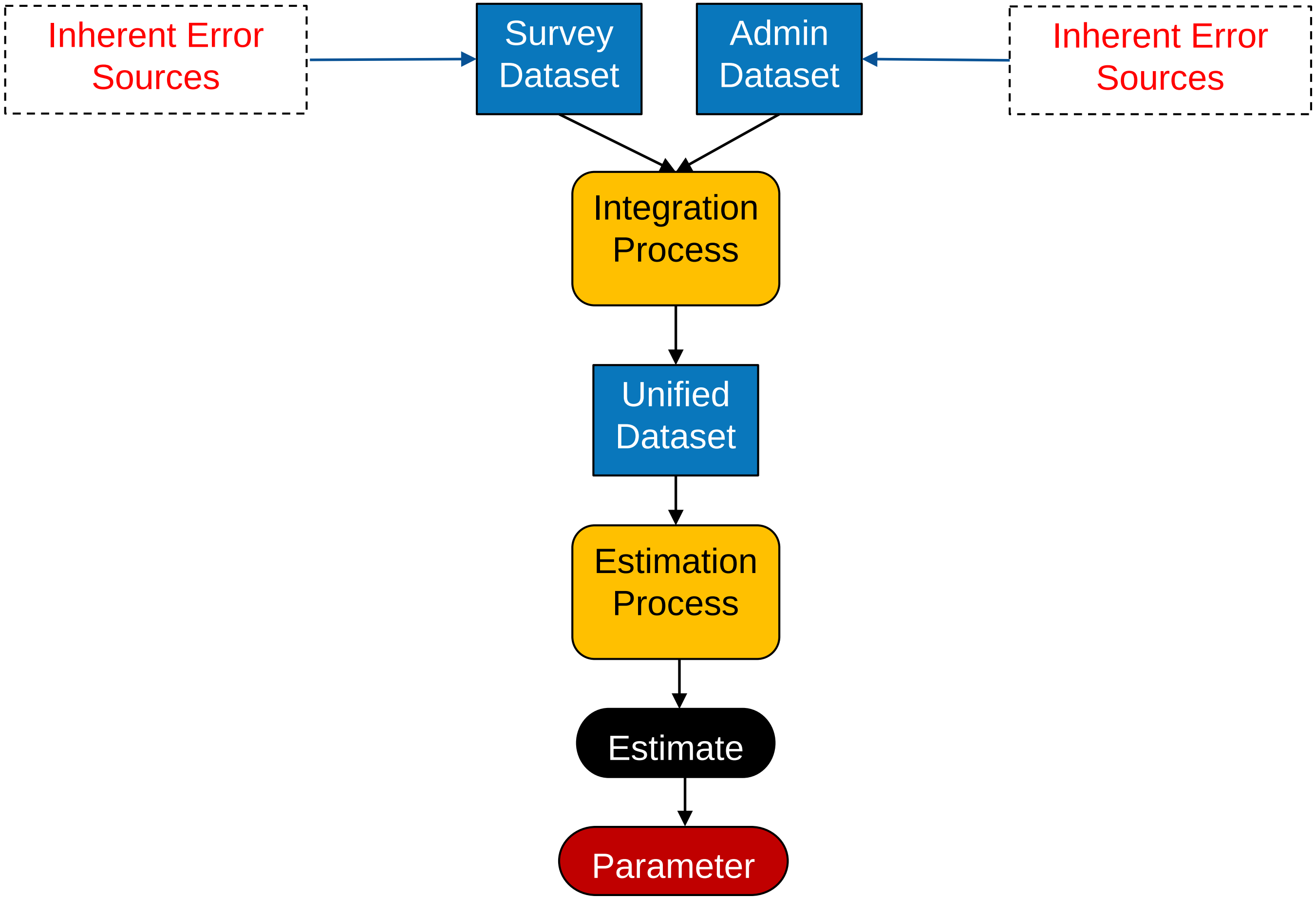


The Hybrid Estimation Process

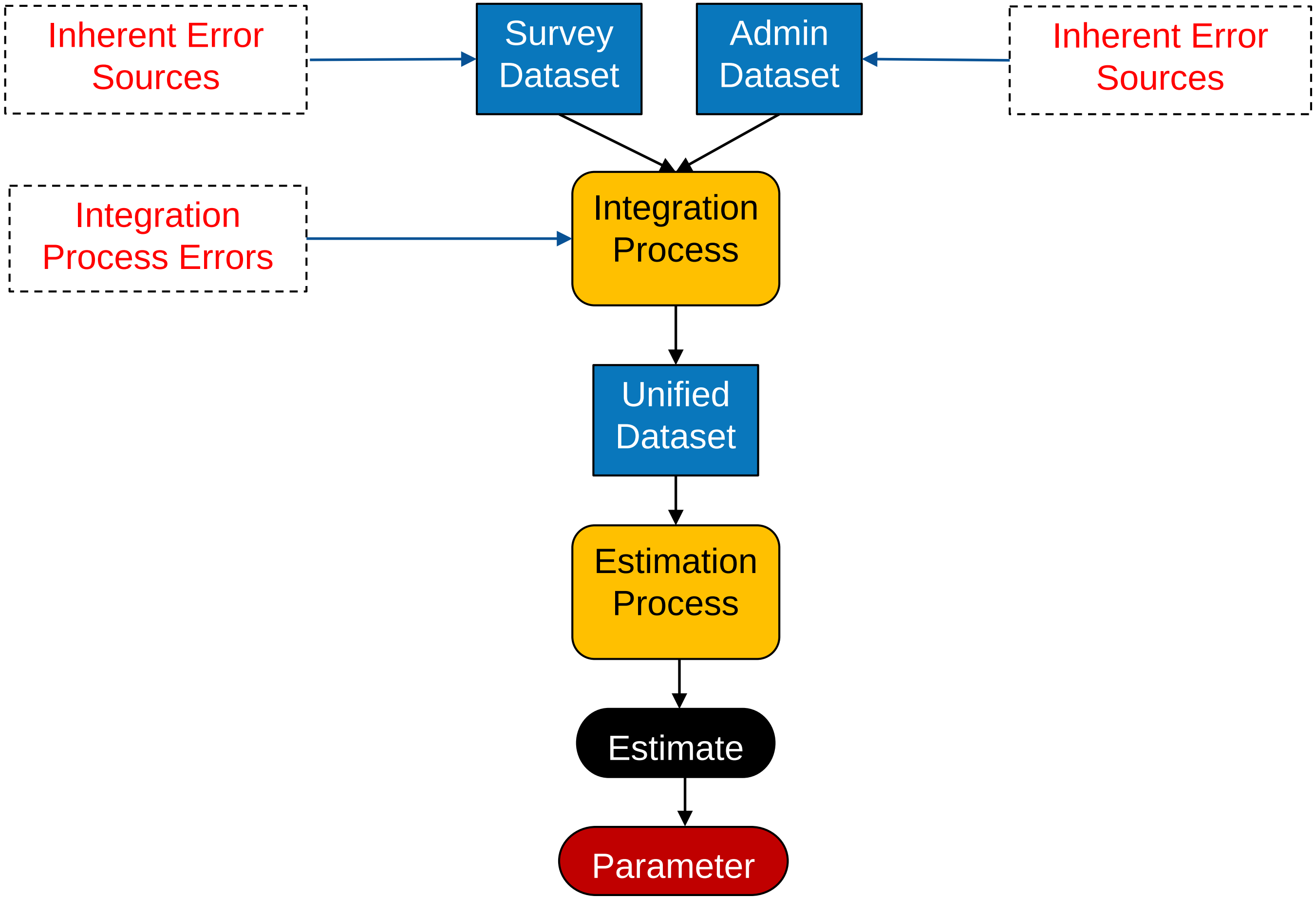


Definition: “data holdings containing information which is not primarily collected for statistical purposes.” United Nations Economic Commission for Europe, 2011

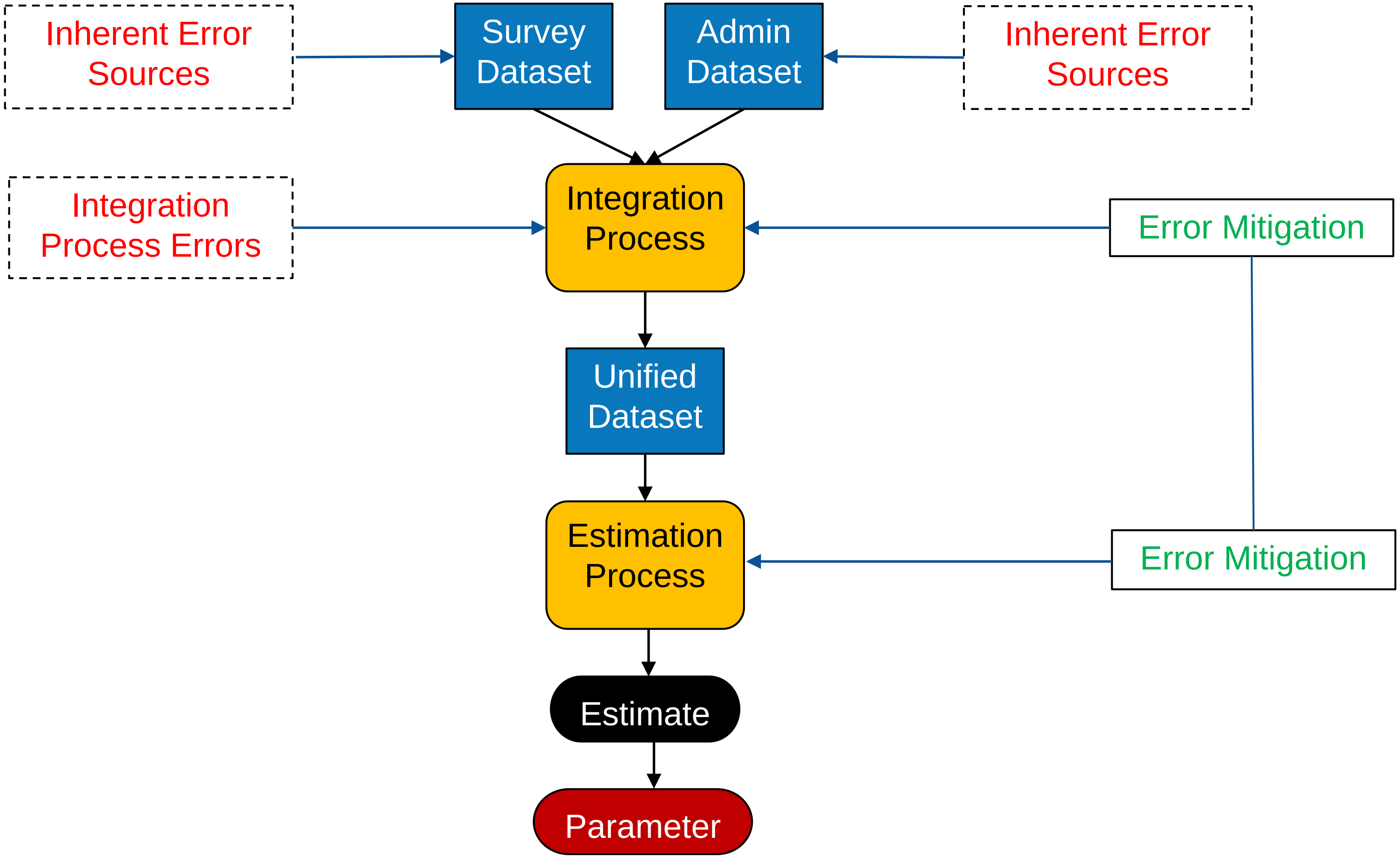
The Hybrid Estimation Process



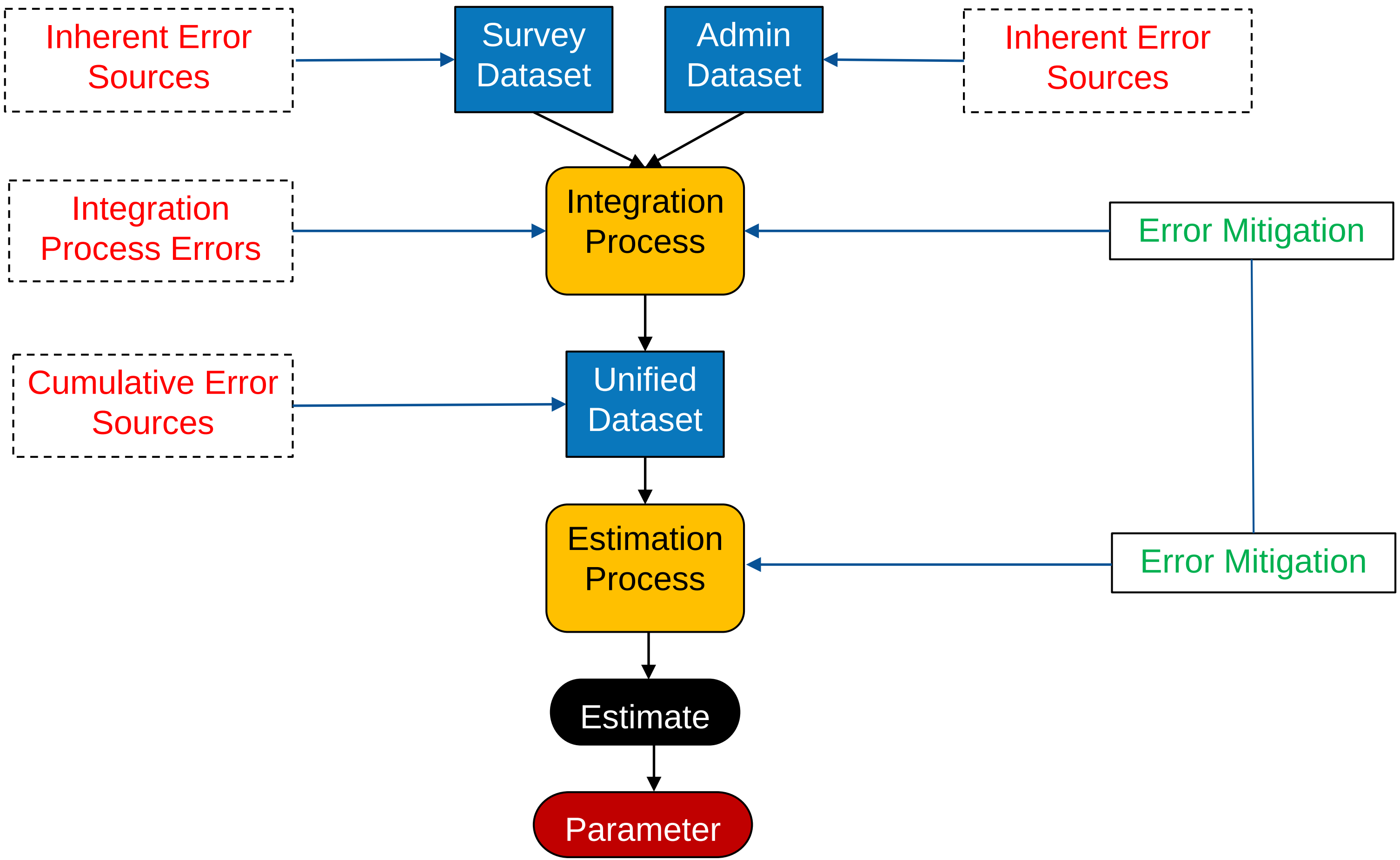
The Hybrid Estimation Process



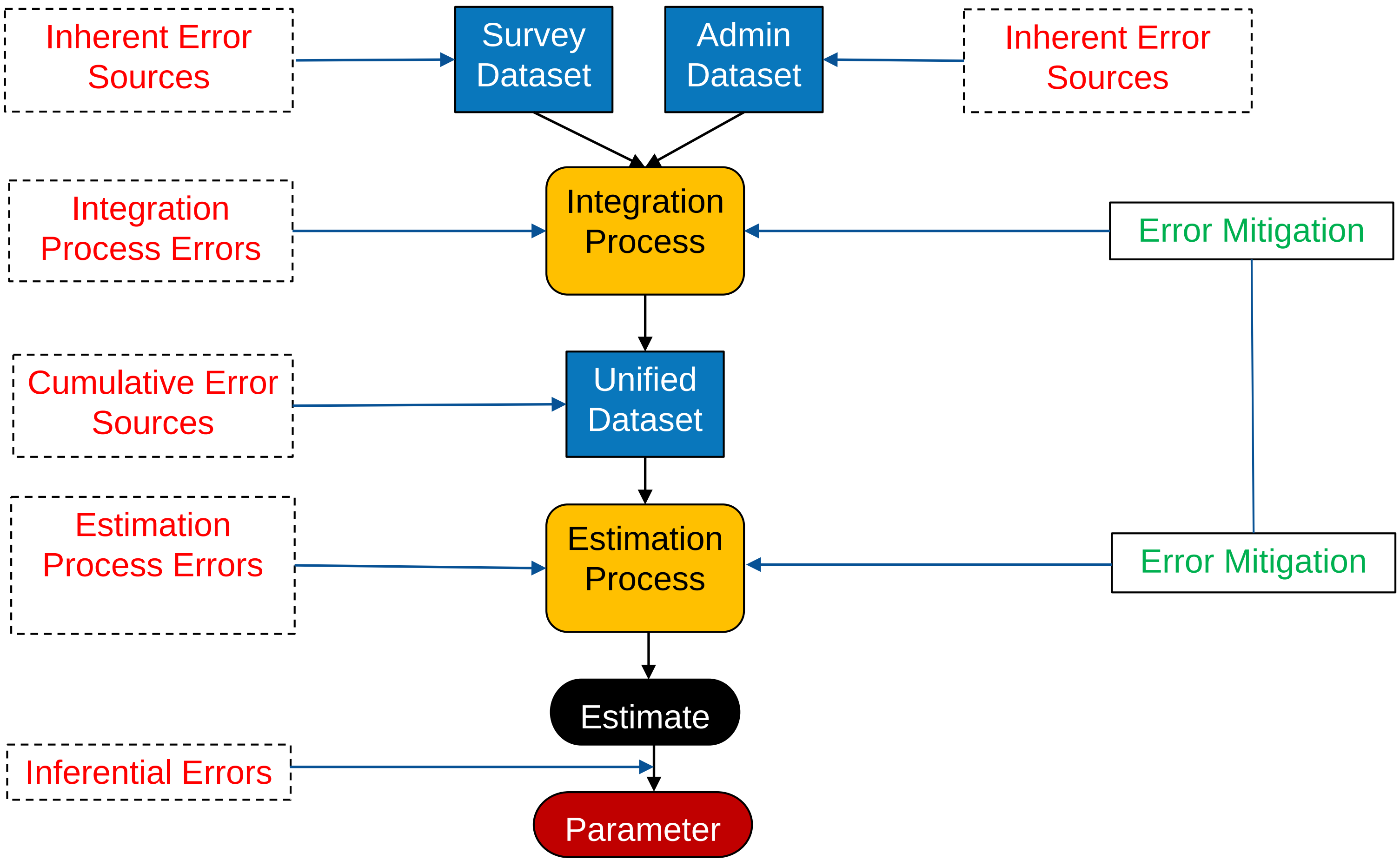
The Hybrid Estimation Process



The Hybrid Estimation Process



The Hybrid Estimation Process



A Total Error Framework for a Generic Dataset

Typical File Structure

Record #	V_1	V_2	…	V_K

variables or features

Population units

A Total Error Framework for a Generic Dataset

Typical File Structure

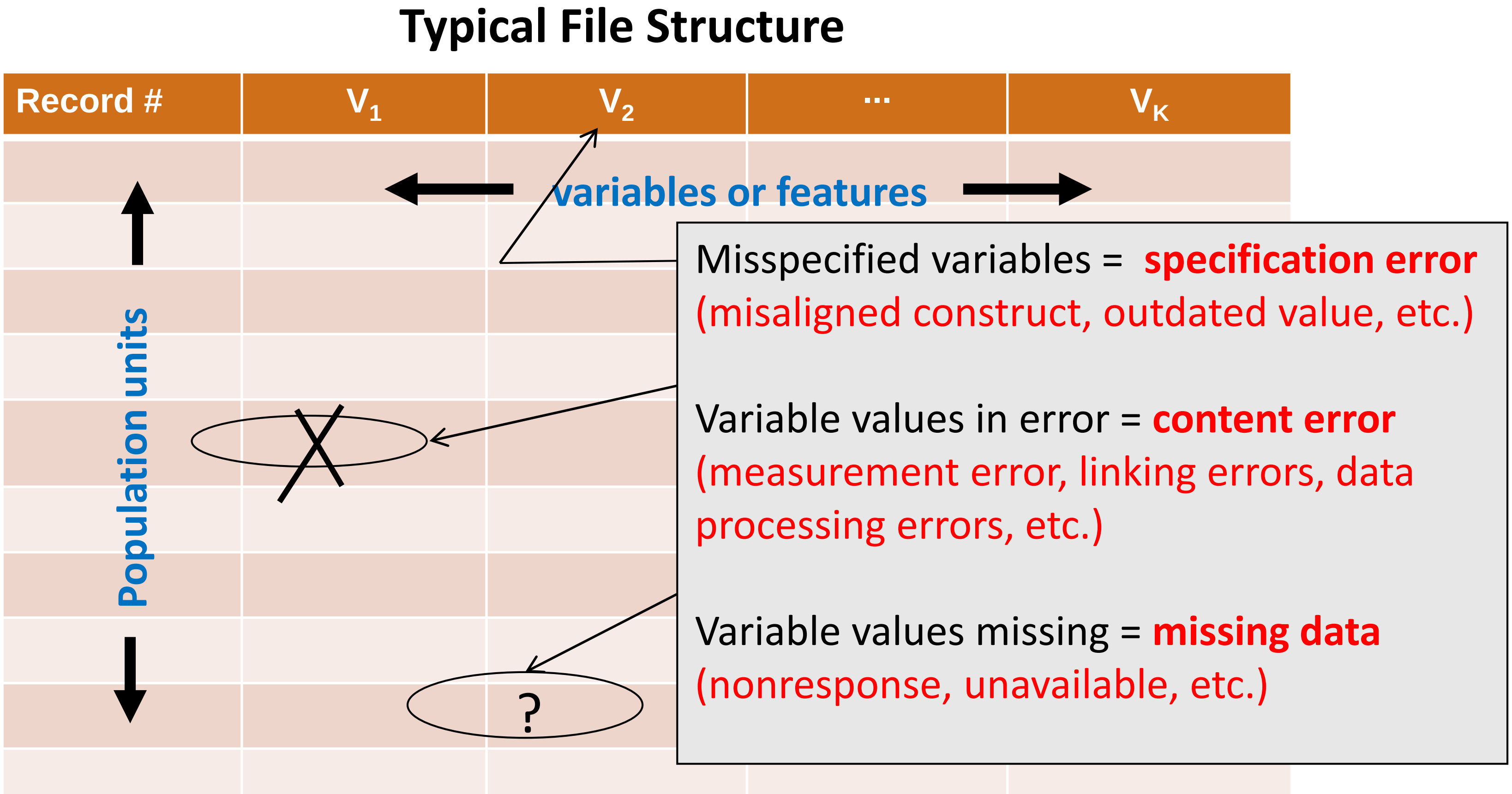
Record #	V_1	V_2	…	V_K

variables or features

Population units

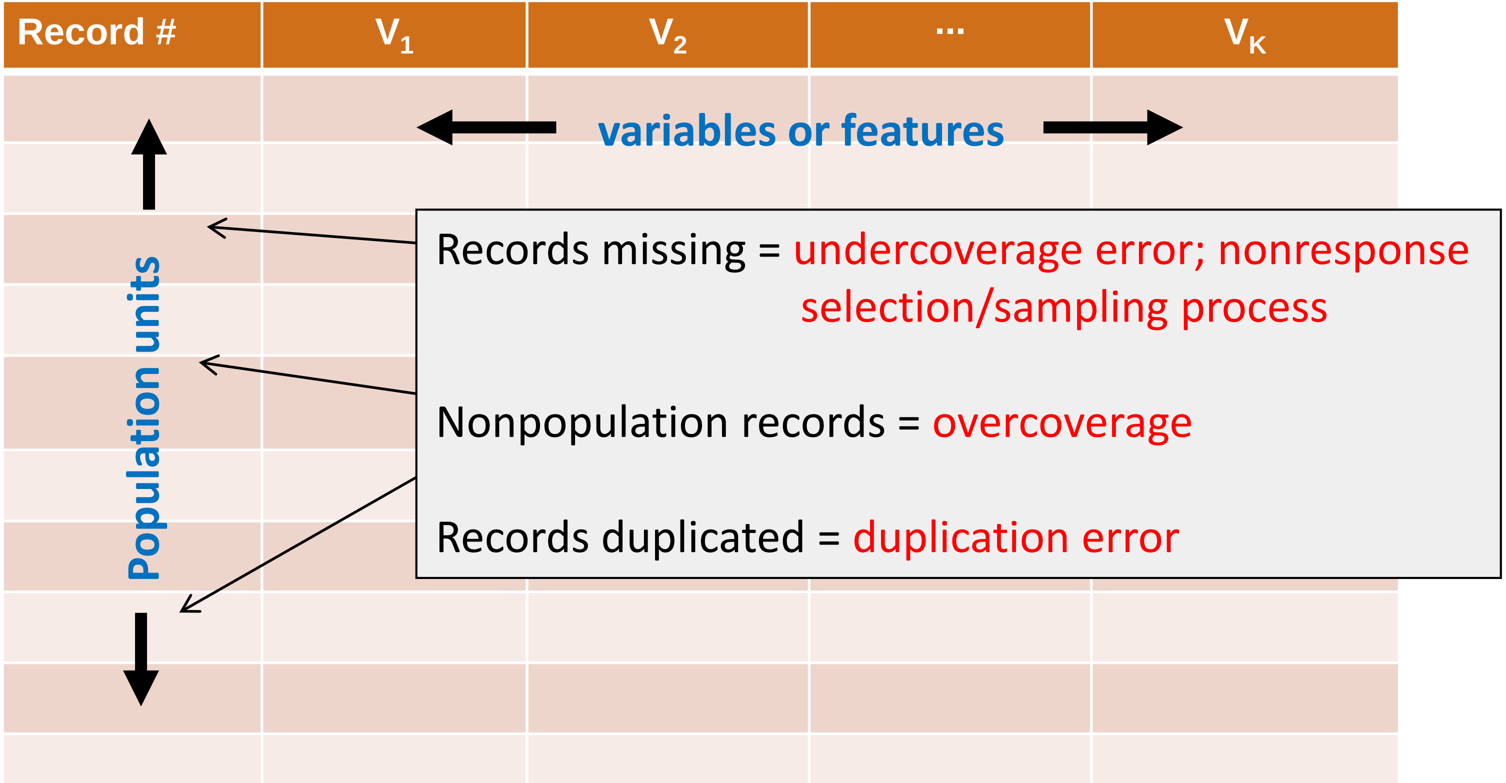
Total error = column error
+ row error
+ cell error

Column and Cell Errors

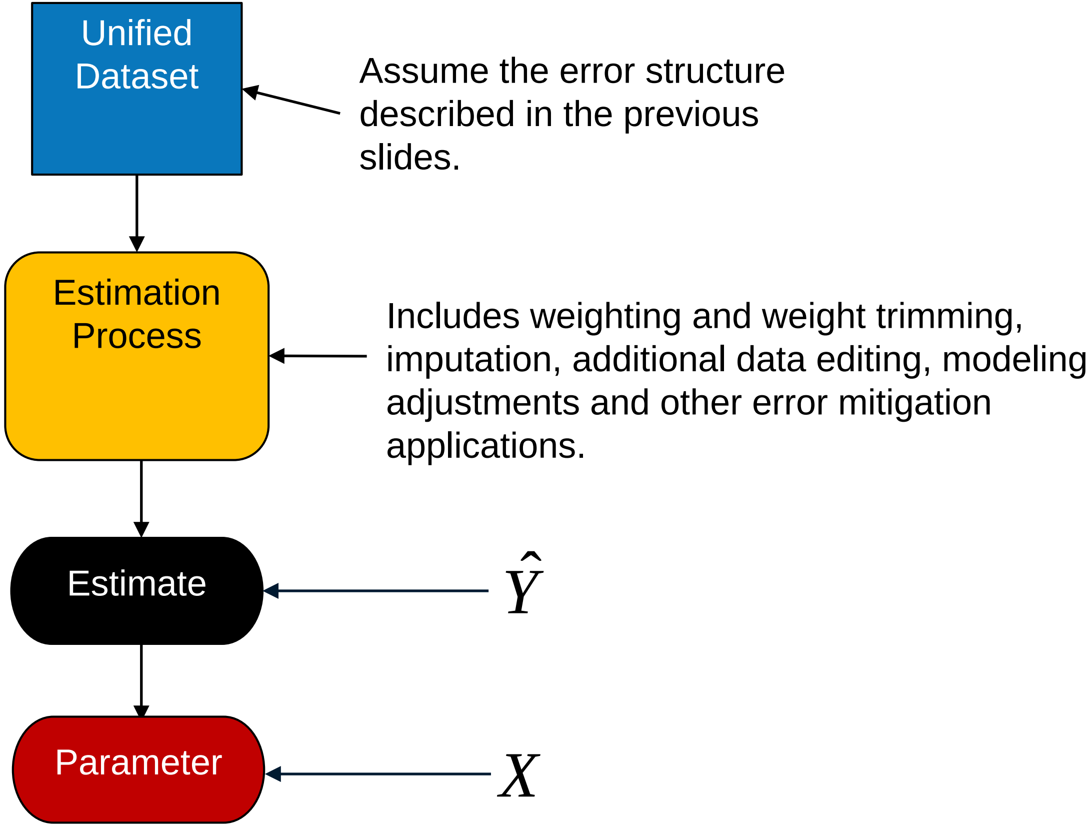


Row errors

Typical File Structure



Errors Associated with the Hybrid Estimation Process



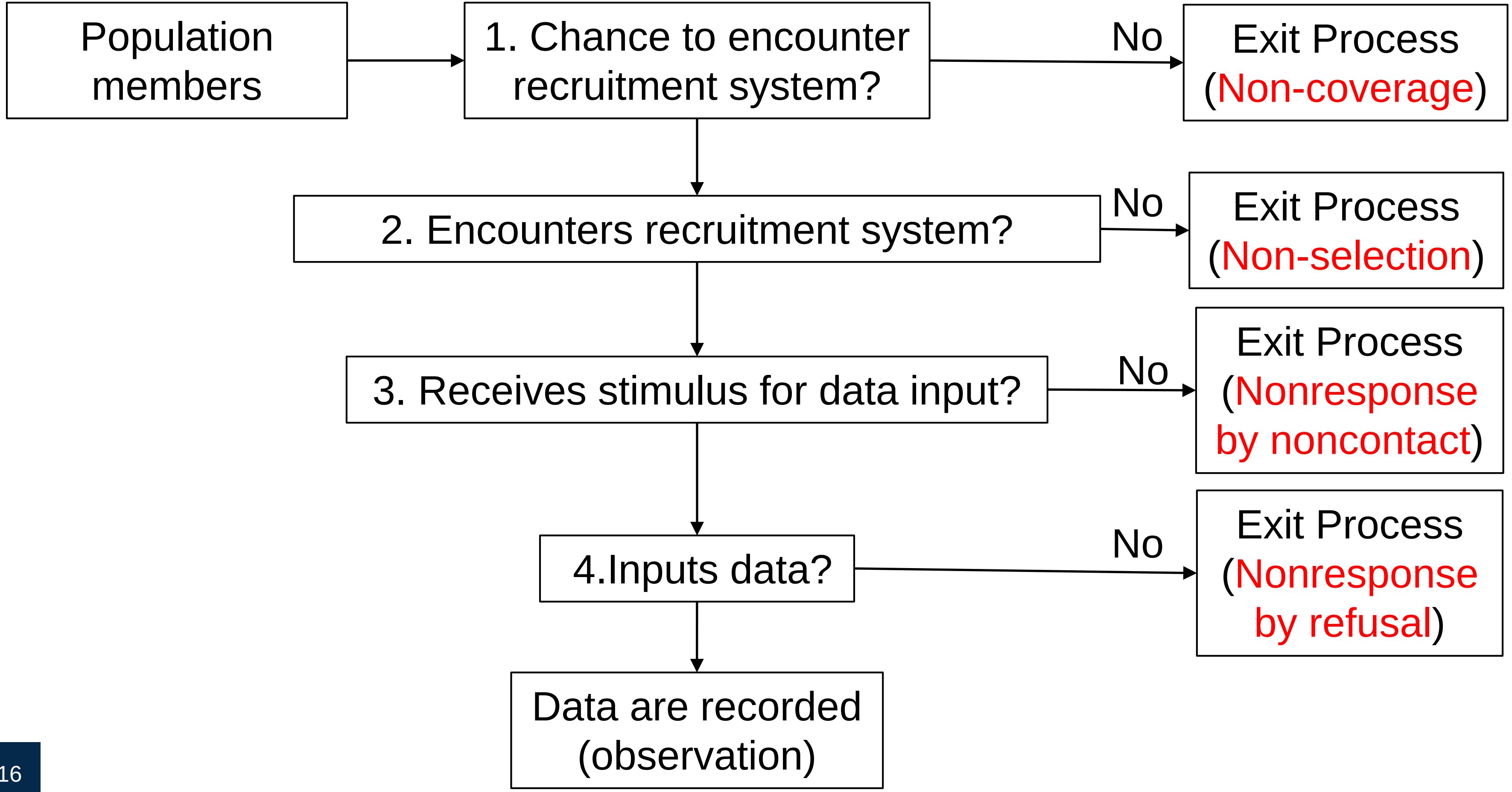
Generalized TE Framework

Total Error = Sample Recruitment Error
+ Data Encoding Error

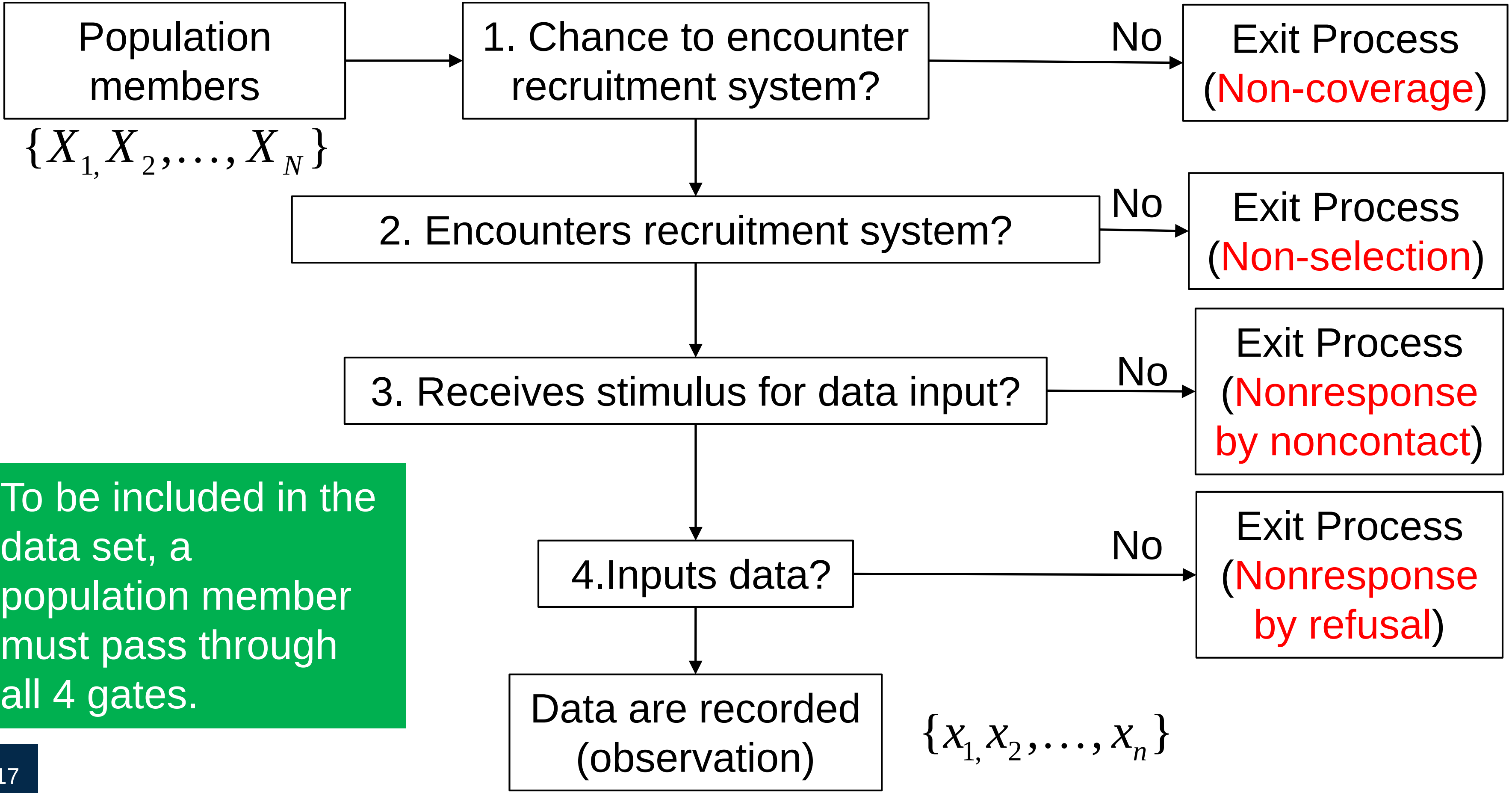
Sample Recruitment Error is a generalization of the concept of representation error

Data Encoding Error is a generalization of the concept of measurement error

Generalized TE Framework – Sample Recruitment Process



Generalized TE Framework – Sample Recruitment Process



To be included in the data set, a population member must pass through all 4 gates.

Generalized TE Framework – Sample Recruitment Process

Population members

$\{X_{1,} X_2, \ldots, X_N\}$

1. Chance to encounter recruitment system?

No

Exit Process (Non-coverage)

2. E

3. R

R_i = 1 if a population unit makes it through all 4 gates;
R_i = 0 otherwise

4.Inputs data?

No

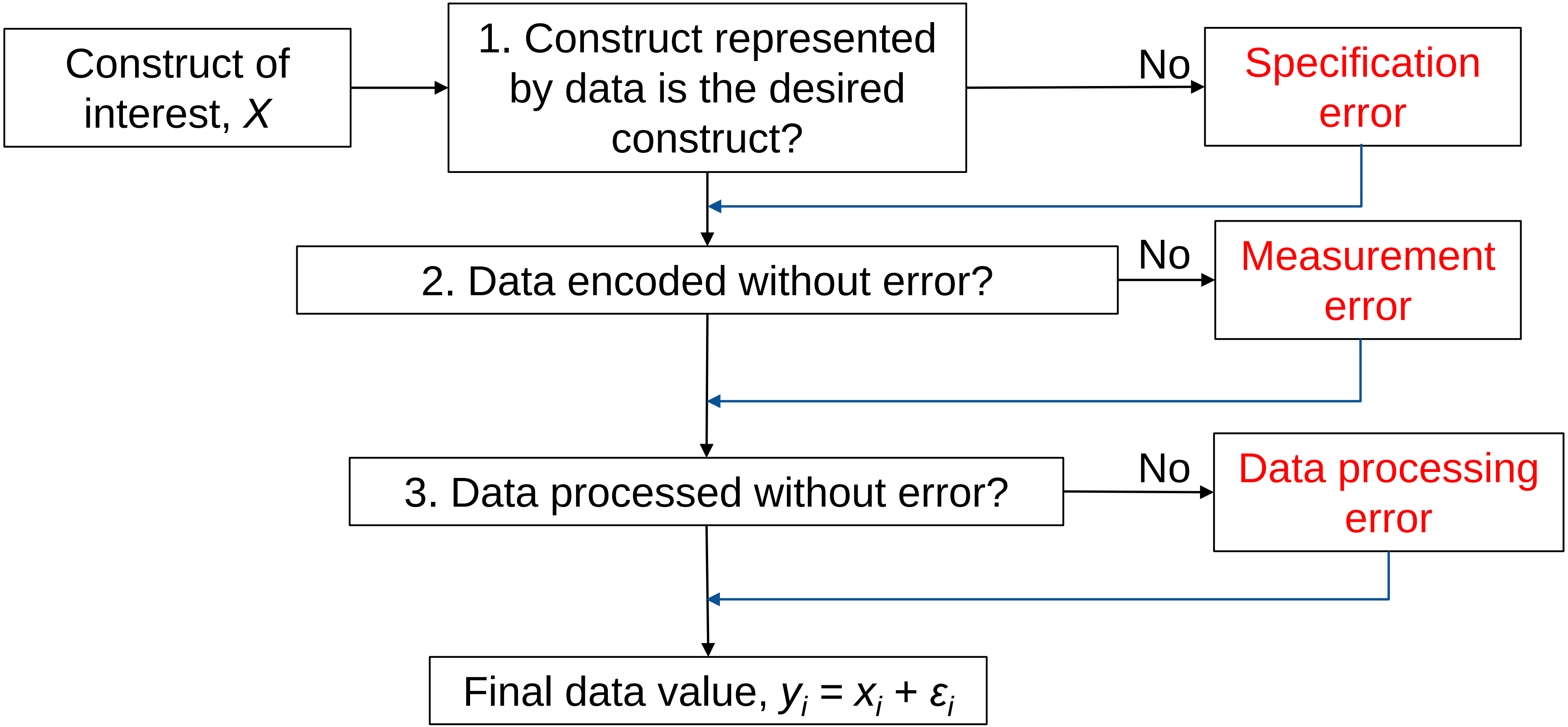
Exit Process (Nonresponse by refusal)

Data are recorded (observation)

$\{x_{1,} x_2, \ldots, x_n\}$

To be included in the data set, a population member must pass through all 4 gates.

Generalized TE Framework – Data Encoding Process



Total Error Identity for the Mean of the Encoded Data

Total Error = Data Enc Error + Samp Recr Error

$$\bar{y}_n - \bar{X}_N = (\bar{y}_n - \bar{x}_n) + (\bar{x}_n - \bar{X}_N)$$

Total Error Identity for the Mean of the Encoded Data

Total Error = Data Enc Error + Samp Recr Error

$$\bar{y}_n - \bar{X}_N = (\bar{y}_n - \bar{x}_n) + (\bar{x}_n - \bar{X}_N)$$

Notation:

$\bar{X}_N$ is the true population mean

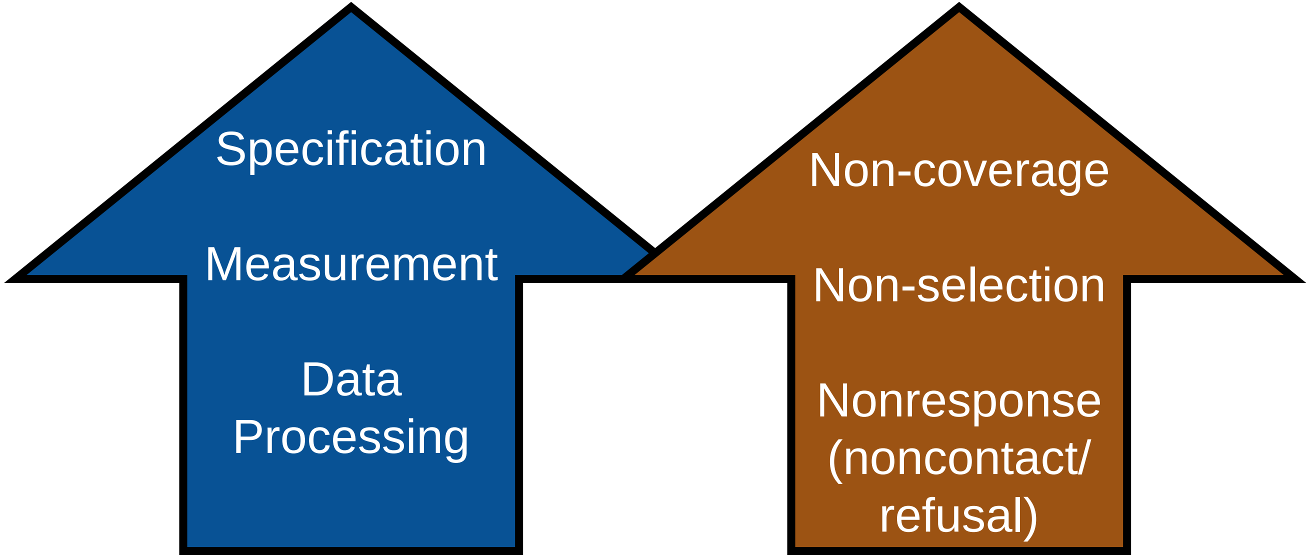
$\bar{y}_n$ is the observed sample mean

$\bar{x}_n$ is the true sample mean

Total Error Identity for the Mean of the Encoded Data

Total Error = Data Enc Error + Samp Recr Error

$$\bar{y}_n - \bar{X}_N = (\bar{y}_n - \bar{x}_n) + (\bar{x}_n - \bar{X}_N)$$



Total Error Identity for the Mean of the Encoded Data

Total Error = Data Enc Error + Samp Recr Error

$$\bar{y}_n - \bar{X}_N = (\bar{y}_n - \bar{x}_n) + (\bar{x}_n - \bar{X}_N)$$

Thus, the total MSE of the sample mean is

$$\begin{aligned}
\mathrm{E}(\bar{y}_n - \bar{X}_N)^2 &= \mathrm{E}(\bar{y}_n - \bar{x}_n)^2 + \mathrm{E}(\bar{x}_n - \bar{X}_N)^2 + 2\mathrm{E}(\bar{y}_n - \bar{x}_n)(\bar{x}_n - \bar{X}_N) \\
&= \left[\mathrm{E}(\bar{y}_n - \bar{x}_n)^2 + \mathrm{E}(\bar{y}_n - \bar{x}_n)(\bar{x}_n - \bar{X}_N)\right] \longleftarrow \text{Data Enc Error} \\
&\quad + \left[\mathrm{E}(\bar{x}_n - \bar{X}_N)^2 + \mathrm{E}(\bar{y}_n - \bar{x}_n)(\bar{x}_n - \bar{X}_N)\right] \longleftarrow \text{Samp Recr Error}
\end{aligned}$$

Data Encoding Error

- x_i is the true characteristic for the *i*th sample unit
- y_i is the encoded value of x_i
- $\varepsilon_i = y_i - x_i$ is the error in the encoded value for the *i*th sample unit
  - Assume $\varepsilon_i \sim i.i.d\,(B_\varepsilon, \sigma_\varepsilon^2)$

$$E_\varepsilon(\bar{y} - \bar{x} \mid R)^2 = B_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n}$$

$R = \{R_i, i=1,\ldots,N\}$

Data capture error bias

Data capture error variance

Sample Recruitment Error Component

- X_i denotes the characteristic measured for the *i*th person in the Recruitment Process
- $\rho_{RX} = \mathrm{Corr}(R_i, X_i \mid R)$, a measure of selection bias

$$\mathrm{E}_R(\bar{x}_n - \bar{X})^2 = \sigma_X^2 \frac{N-n}{n} E_R(\rho_{RX}^2)$$

Population variance

Bias induced by the sample recruitment process

Meng, 2017;
Bethlehem, 1988

Sample Recruitment Error Component

- x_i denotes the true characteristic measured for the *i*th person in the Recruitment Process
- $\rho_{RX} = \mathrm{Corr}(R_i, x_i \mid R)$, a measure of selection bias

$$\mathrm{E}_R(\bar{x}_n - \bar{X})^2 = \sigma_X^2 \frac{N-n}{n} E_R(\rho_{RX}^2)$$

Example: For SRS sampling and no nonresponse, $\mathrm{E}_R(\rho_{RX}^2) = \frac{1}{N-1}$

$$\frac{N-n}{n}\sigma_X^2 \mathrm{E}_R(\rho_{RX}^2) = \left(1 - \frac{n}{N}\right)\frac{S_X^2}{n}$$

Total Mean Squared Error of the Mean of a Generic Data Set

Total Error = Data Enc Error + Samp Recr Error

$$\bar{y}_n - \bar{X}_N = (\bar{y}_n - \bar{x}_n) + (\bar{x}_n - \bar{X}_N)$$

$$\underbrace{\mathrm{E}(\bar{y}_n - \bar{X}_N)^2}_{\mathrm{MSE}(\bar{y}_n)} = \underbrace{\mathrm{E}(\bar{y}_n - \bar{x}_n)^2}_{B_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n}} + \underbrace{\mathrm{E}(\bar{x}_n - \bar{X}_N)^2}_{\frac{N-n}{n}\sigma_X^2 \mathrm{E}_R(\rho_{RX}^2)} + \underbrace{2\mathrm{E}(\bar{y}_n - \bar{x}_n)(\bar{x}_n - \bar{X}_N)}_{B_\varepsilon \sqrt{\frac{N-n}{n}}\sigma_X \mathrm{E}_R(\rho_{RX})}$$

Bias from
- Specification error
- Measurement error
- Data processing error

Variance from
- Measurement error
- Data processing error

Variance & Bias from
- Noncoverage
- Nonselection
- Nonresponse

Encoding error /recruitment error interaction

Interpretation of ρ_{RX}

- Not much is known about ρ_{RX} for nonprobability samples.
- However, ρ_{RX} has been studied extensively for surveys (through the estimation of nonresponse bias).
- ρ_{RX} will be smaller for nonprobability samples when gates 1, 2 and 3 are entered for all members of the population. →
- Ability to adjust for sample recruitment bias is better for surveys because
  - We have more control over who enters gates 1-3 and thus more control over ρ_{RX}
  - We often know a lot about sample recruitment failures and how to adjust for them through weighting and imputation.

Alternative Form of the MSE

$$\mathrm{RelMSE}(\bar{y}_n) = RB_{\varepsilon}^2 + \frac{CV_X^2}{n}\left[\frac{1-\tau_y}{\tau_y} + (N-n)\mathrm{E}_R(\rho_{RX}^2)\right] + 2CV_X\,RB_{\varepsilon}\sqrt{\frac{N-n}{n}}\mathrm{E}_R(\rho_{RX})$$

Relative MSE is Often More Convenient to Work With

$$\text{RelMSE}(\bar{y}_n) = RB_\varepsilon^2 + \frac{CV_X^2}{n}\left[\frac{1-\tau_y}{\tau_y} + (N-n)\text{E}_R(\rho_{RX}^2)\right] + 2CV_X RB_\varepsilon \sqrt{\frac{N-n}{n}}\text{E}_R(\rho_{RX})$$

$$\frac{MSE}{\bar{X}^2} \qquad RB_\varepsilon = \frac{B_\varepsilon}{\bar{X}} \qquad CV_X = \frac{\sigma_X}{\bar{X}} \qquad \tau_y = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\varepsilon^2} \text{ (reliability ratio)}$$

Alternative Form of the MSE

$$\mathrm{RelMSE}(\bar{y}_n) = RB_\varepsilon^2 + \frac{CV_X^2}{n}\left[\frac{1-\tau_y}{\tau_y} + (N-n)\mathrm{E}_R(\rho_{RX}^2)\right] + 2CV_X\,RB_\varepsilon\sqrt{\frac{N-n}{n}}\mathrm{E}_R(\rho_{RX})$$

Data Encoding Error

Alternative Form of the MSE

$$\mathrm{RelMSE}(\bar{y}_n) = RB_{\varepsilon}^2 + \frac{CV_X^2}{n}\left[\frac{1-\tau_y}{\tau_y} + (N-n)\mathrm{E}_R(\rho_{RX}^2)\right] + 2CV_X RB_{\varepsilon}\sqrt{\frac{N-n}{n}}\mathrm{E}_R(\rho_{RX})$$

Sample Recruitment Error

Illustration of Some Uses of these Results

Which is more accurate?

1. An estimate of the population average based upon an administrative data set with almost 100,000,000 records and over 80% coverage? or
2. A national survey estimate based upon probability sample of 6000 respondents with a 55% response rate?

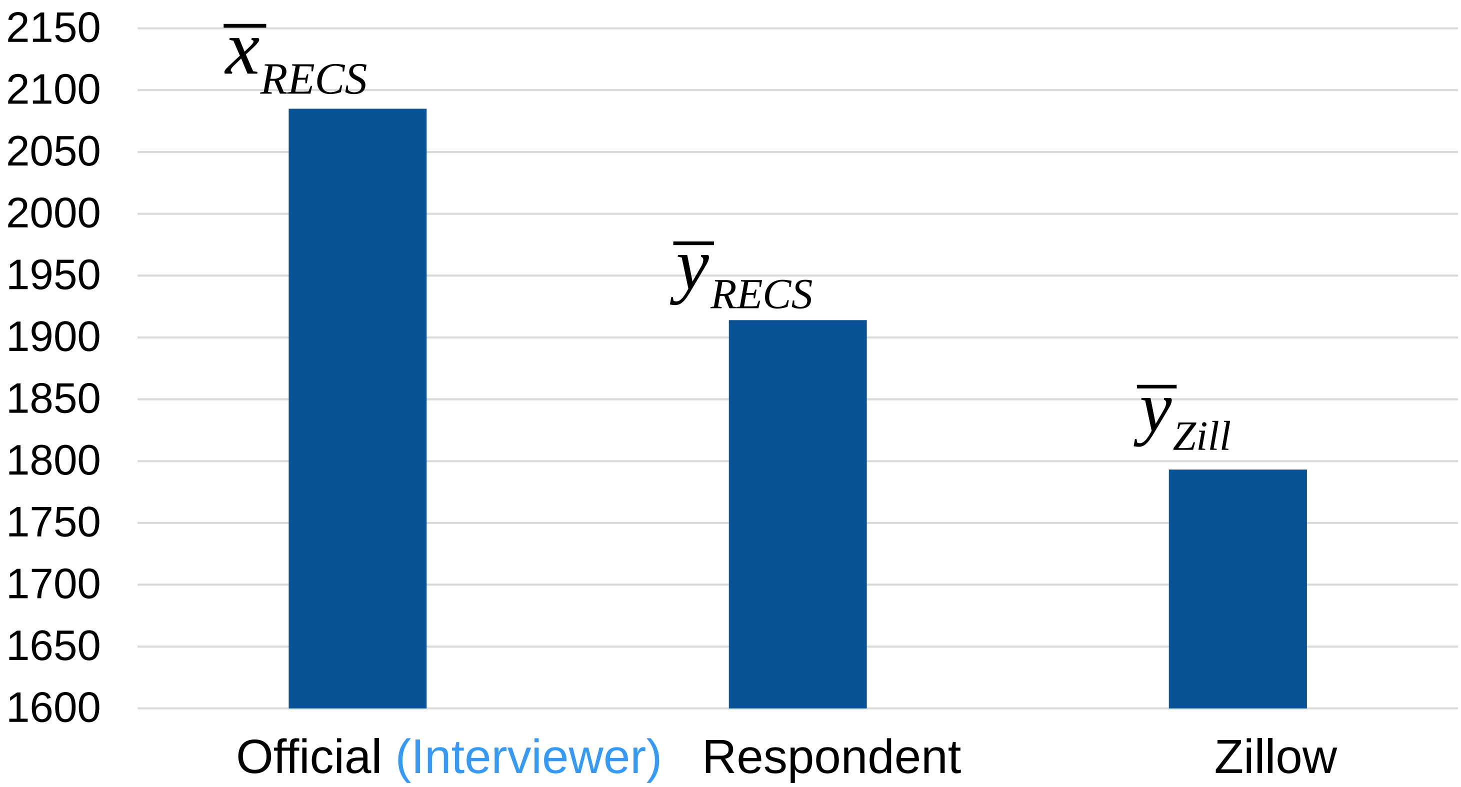
We try to answer this question for the US Residential Energy Consumption Survey (RECS)

- **Survey Data: 2015 RECS**
  - Mode changed from face to face to web/mail
  - Respondent reports of housing unit square footage not reliable
  - Substituting administrative data could be more accurate
  - $n \approx$ 6,000 completed cases
  - response rate ≈ 55%
- **Administrative data: Zillow real estate data base**
  - Coverage ≈ 82%
  - $n \approx$ 100,000,000 records
  - Other data bases were also considered (Acxiom and CoreLogic)

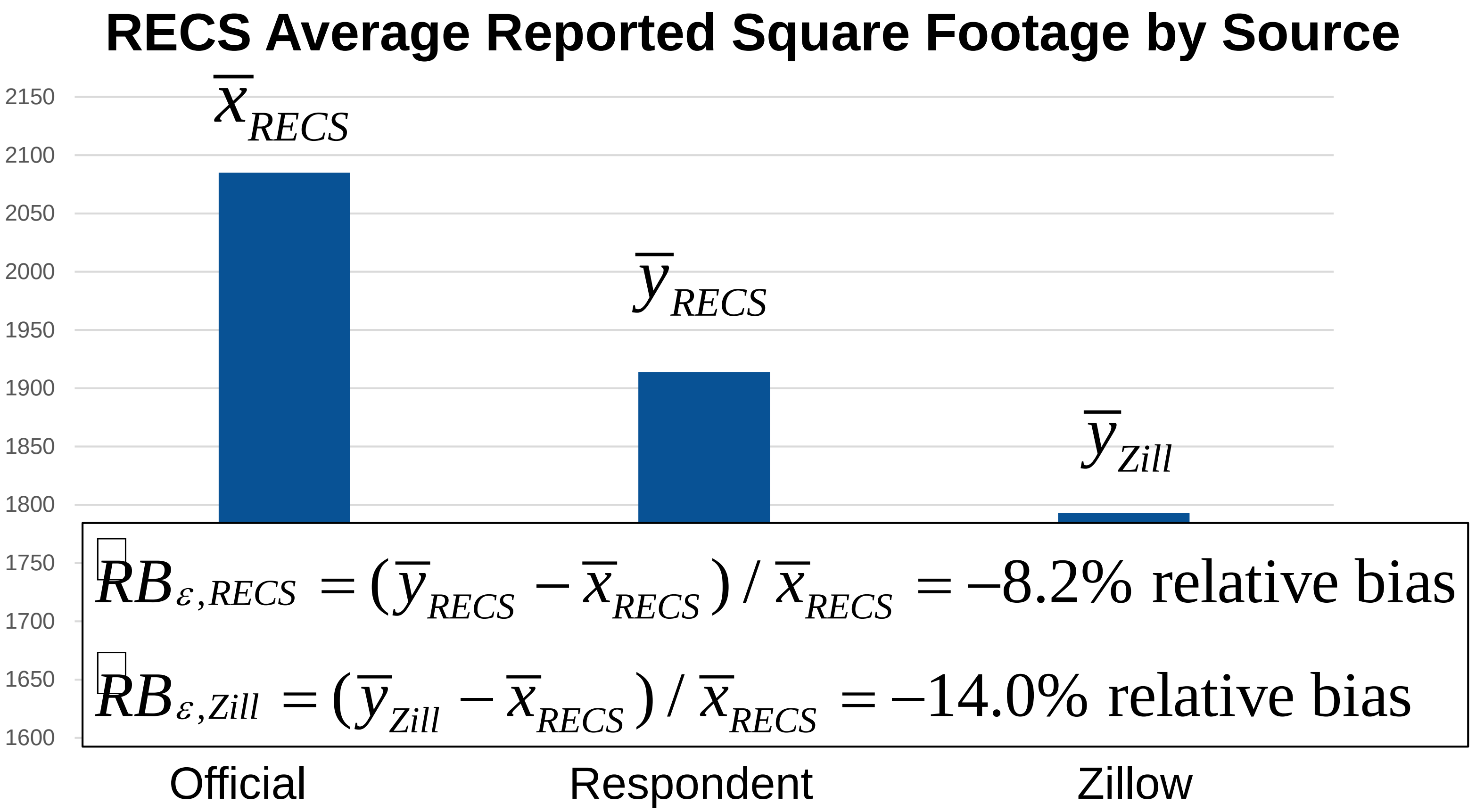
We try to answer this question for the US Residential Energy Consumption Survey (RECS)

- HU square footage is primarily used for micro-econometric modeling
- We will consider estimation of the U.S. average HU square footage to demonstrate the MSE analysis
- Similar analysis could be considered for other parameters of interest (i.e., regression coefficients)
- However, the current formulation would not be appropriate.
- Our analysis will use results from Amaya (2017)

Evidence of Nonsampling Error from the RECS



Evidence of Nonsampling Error from the RECS



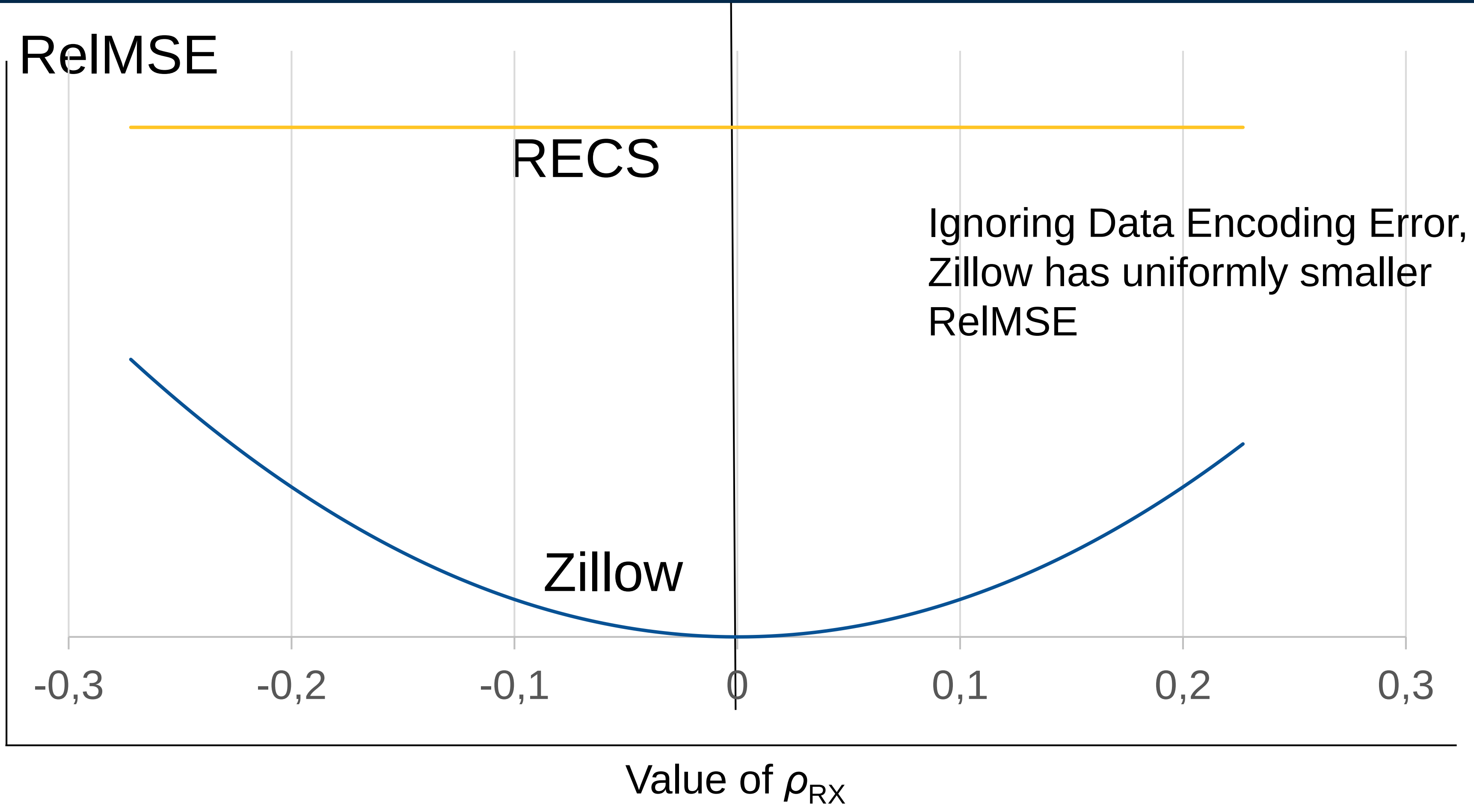
$$RB_{\varepsilon,RECS} = (\bar{y}_{RECS} - \bar{x}_{RECS}) / \bar{x}_{RECS} = -8.2\% \text{ relative bias}$$

$$RB_{\varepsilon,Zill} = (\bar{y}_{Zill} - \bar{x}_{RECS}) / \bar{x}_{RECS} = -14.0\% \text{ relative bias}$$

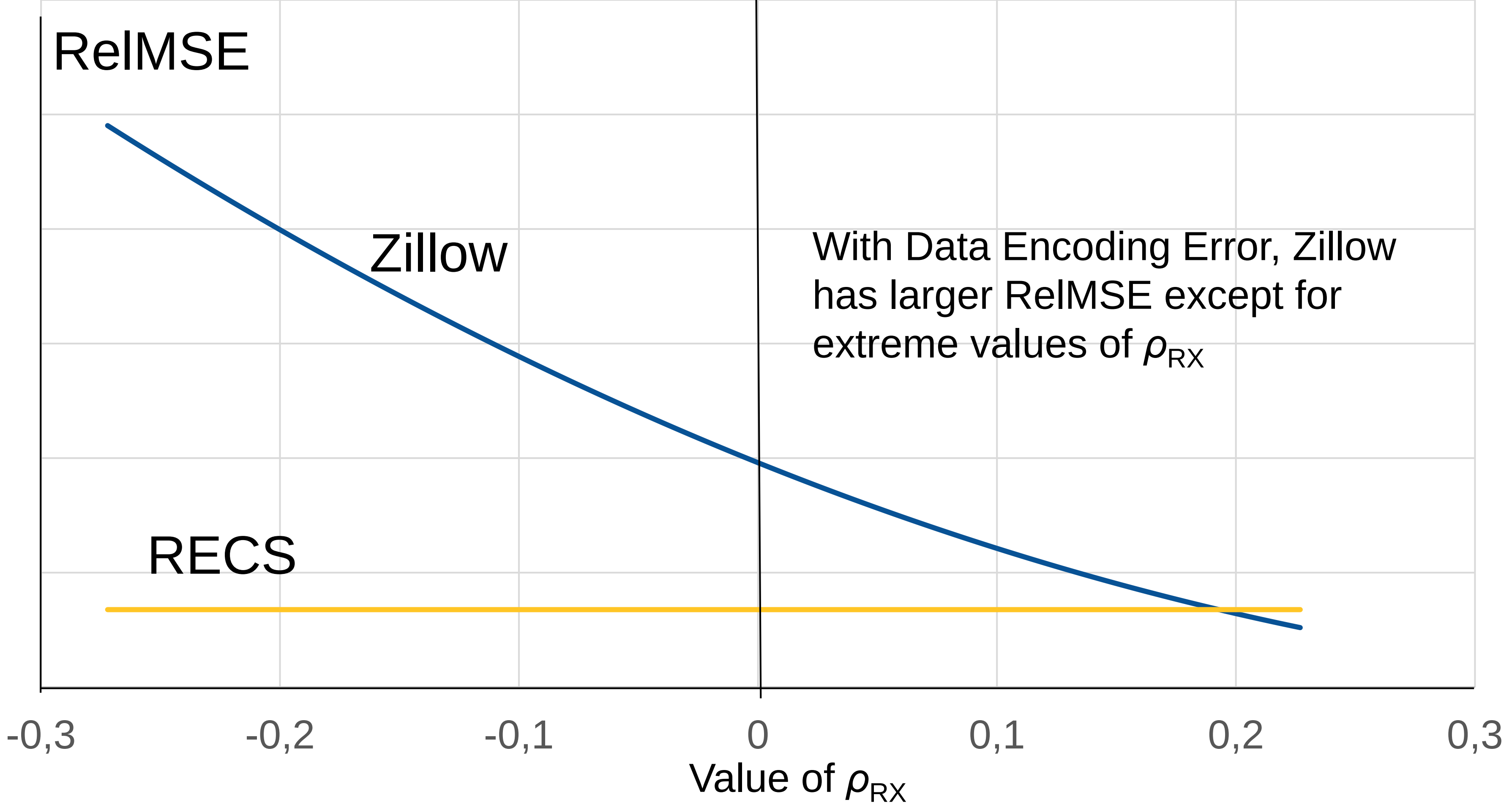
Input Parameters for Computing MSE

MSE Component	RECS	Zillow
Relative Bias	-0.082	-0.14
Pop'n CV	0.64	0.64
Reliability	0.59	0.66
ρ_{RX}	-0.000295	[-0.27,0.22]
N	118,208,250	118,208,250
n	6,000	96,930,765
Response rate	55.4%	82%
Coverage rate	≈ 99%	
Selection rate	0.009%	

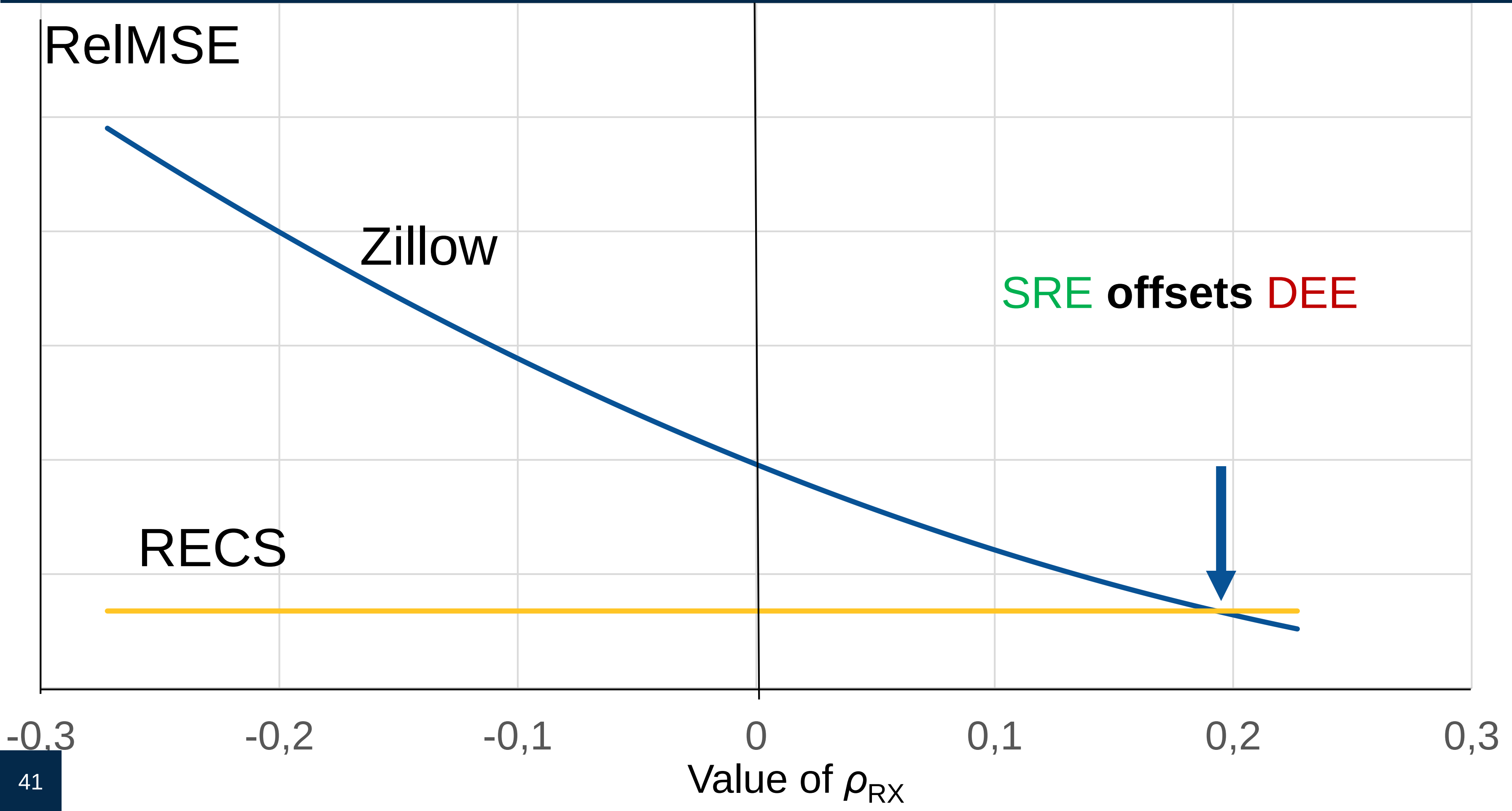
RMSEs as a Function of ρ_{RX} for Zillow and RECS



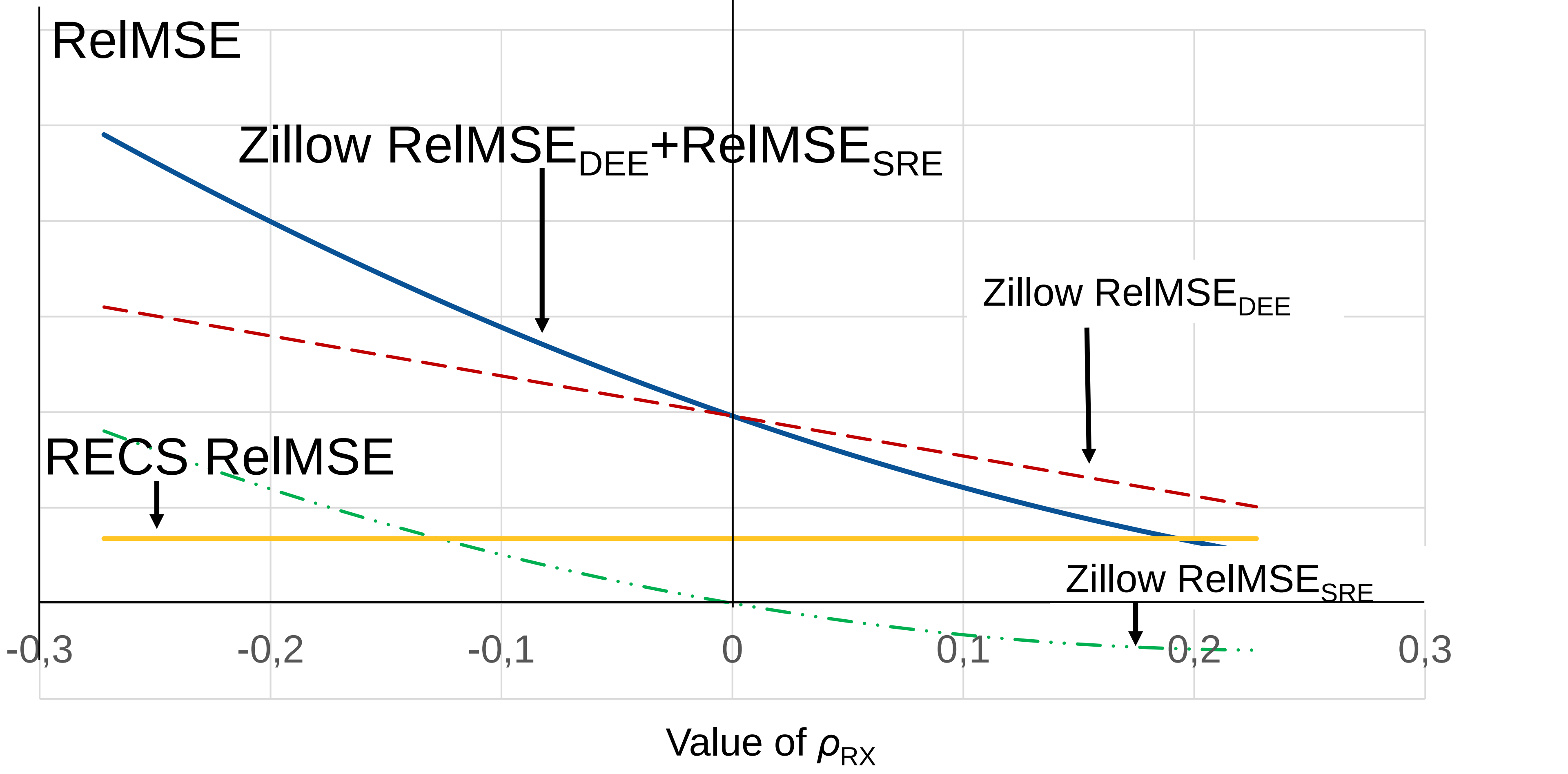
RelMSEs as a Function of ρ_{RX} for Zillow



RelMSEs as a Function of ρ_{RX} for Zillow



RelMSEs as a Function of ρ_{RX} for Zillow



Results Summary

- Whether $MSE_{Zillow} < MSE_{RECS}$ depends on value of ρ_{RX}
- In this case, reducing ρ_{RX} may lead to larger MSE because two biases are offsetting one another
- Ideally, both biases should be minimized because an offsetting biases situation is not sustainable

Potential Zillow Error Mitigation Strategies

Data Encoding Error

- Estimate the bias and adjust for it
- May need ground truth square footage data to model this bias
- Weighting is not an effective strategy for mitigating this error risk

Sample Recruitment Error

- Weight the Zillow data to reduce $|\rho_{RX}|$
- Weights will approximate $[E(R_i)]^{-1}$
- Modeling $E(R_i|X)$ will require understanding how R_i varies by housing unit and other characteristics (X) of the sample recruitment process
- Biemer and Amaya (in press) consider the effects of erroneous weights on the total MSE

A Few Take Aways

As we move towards integrating survey and Big Data, need to consider the **total error**.

- Sample recruitment bias is the least understood component of the total error.
- Data encoding errors (a.k.a measurement errors) are often ignored in Big Data analysis, yet they can have extreme effects on inferences and insights.
- Understanding these components will lead to statistical products of greater quality, utility and efficiency.

Grazie!

ppb@rti.org

Generalized TE Framework – Sample Recruitment Process ←