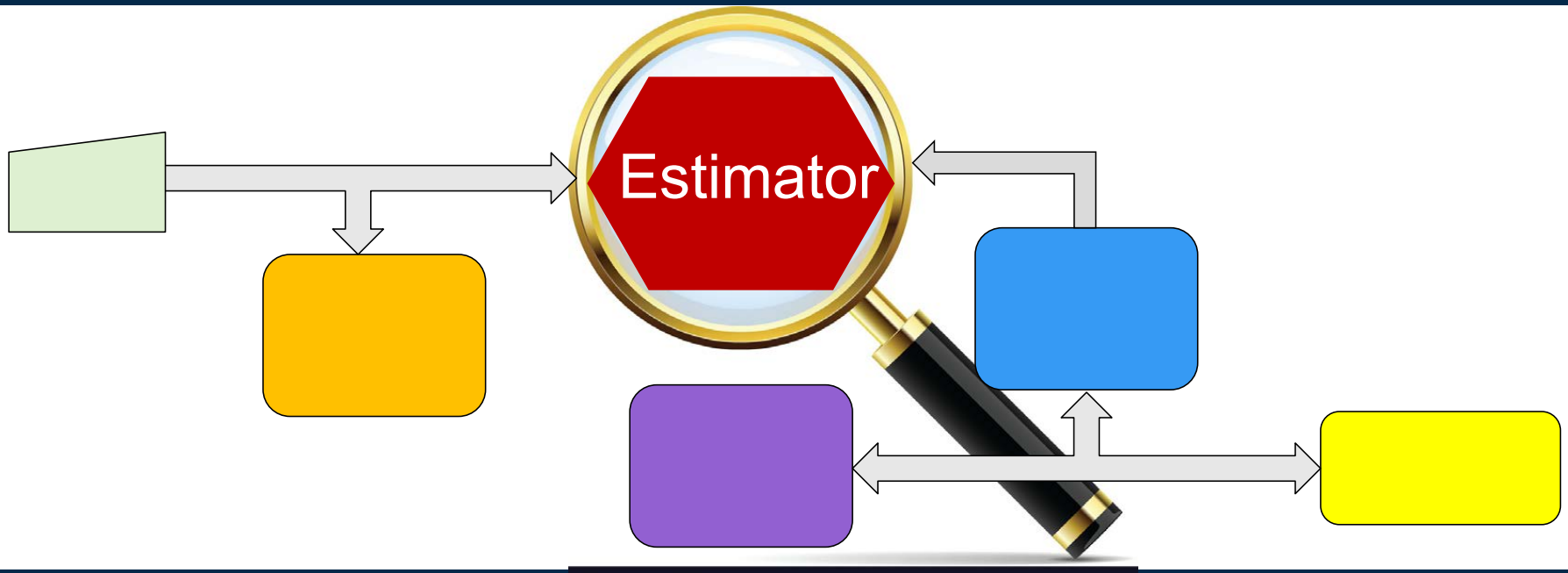
Some Tools for Assessing and Improving the Accuracy of Hybrid Estimators

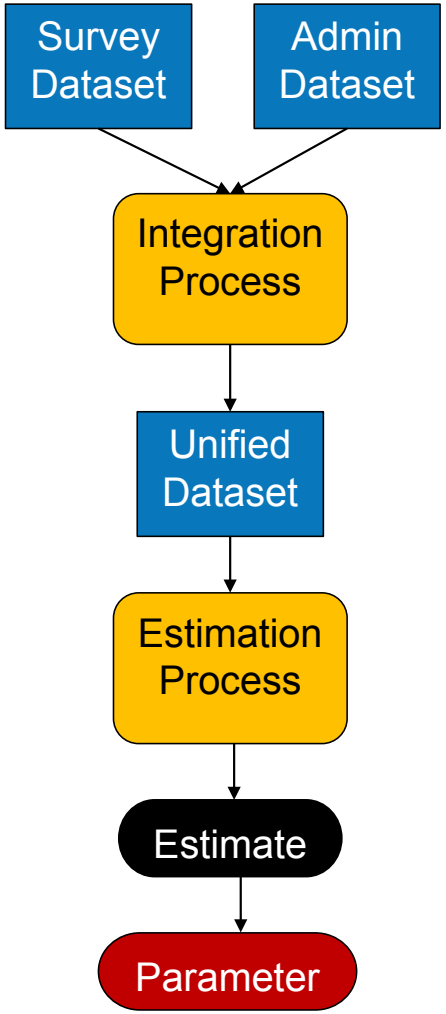
Paul P. Biemer[1,2],
Ashley Amaya[1]
[1] RTI International; [2]University of North Carolina



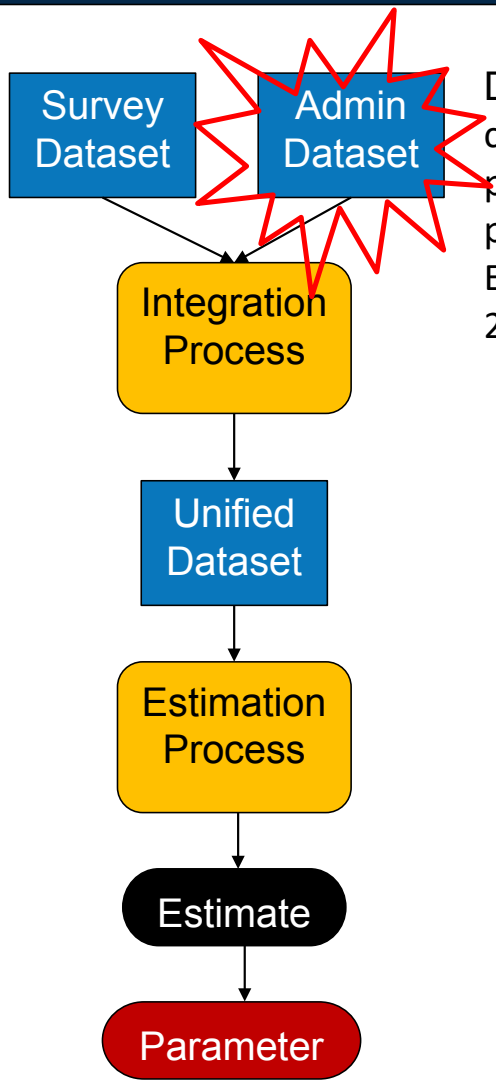
Outline

- Hybrid estimators
- An total error framework for datasets
- An total error framework for hybrid estimators
- Types of error risks
- Error risk profiles
- Illustration of the concepts

The Hybrid Estimation Process

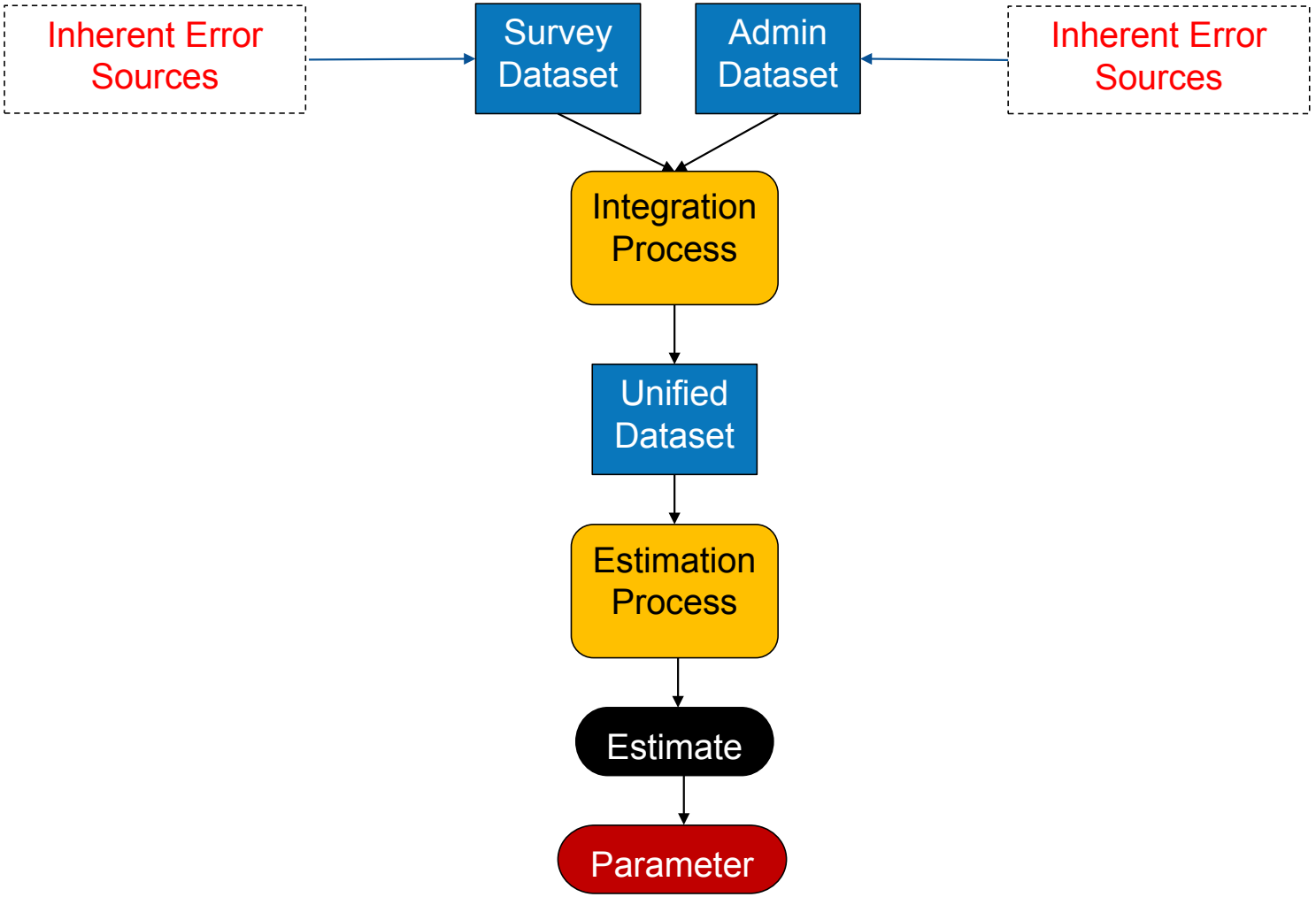


The Hybrid Estimation Process

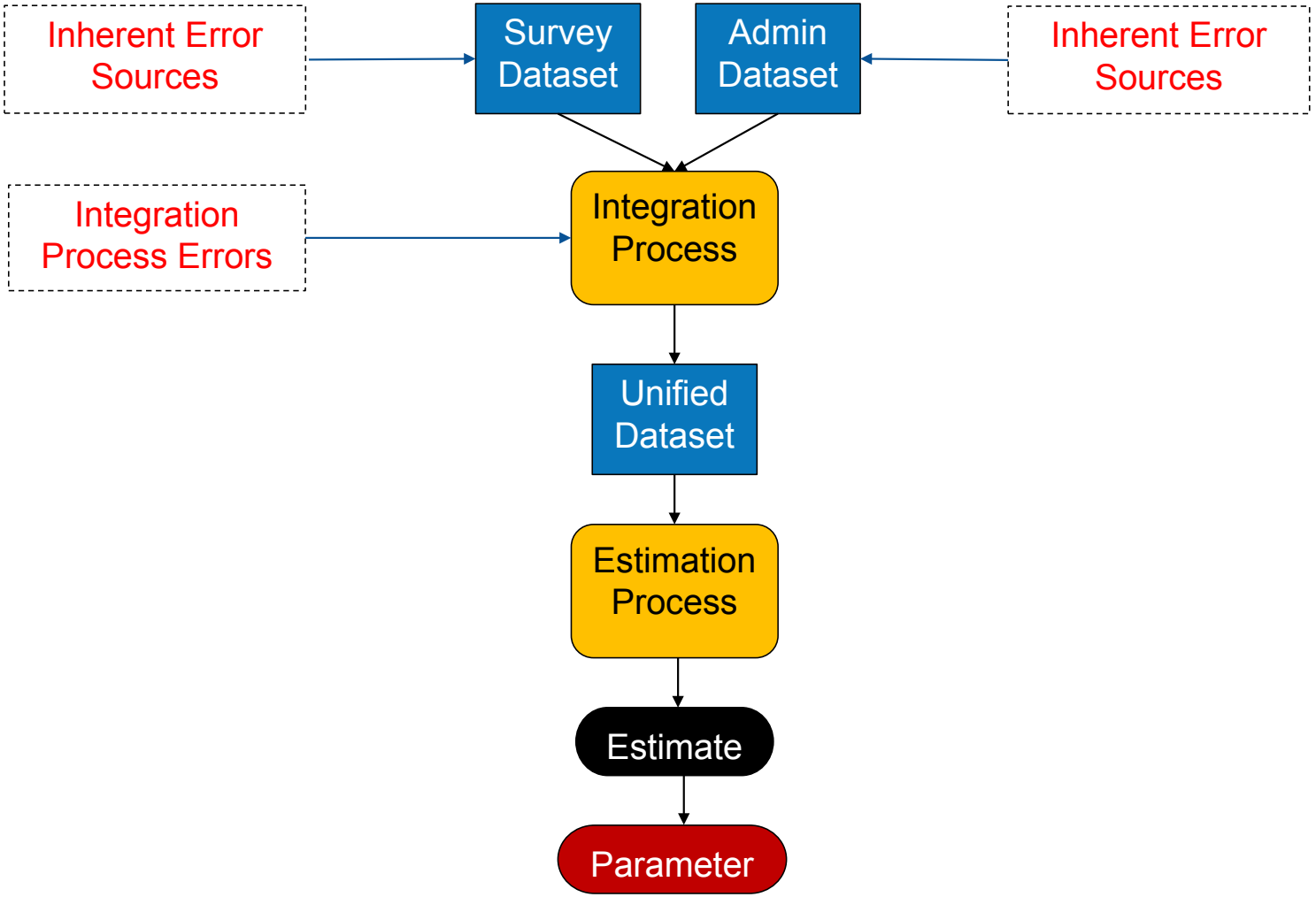


Definition: "data holdings containing information which is not primarily collected for statistical purposes." United Nations Economic Commission for Europe, 2011

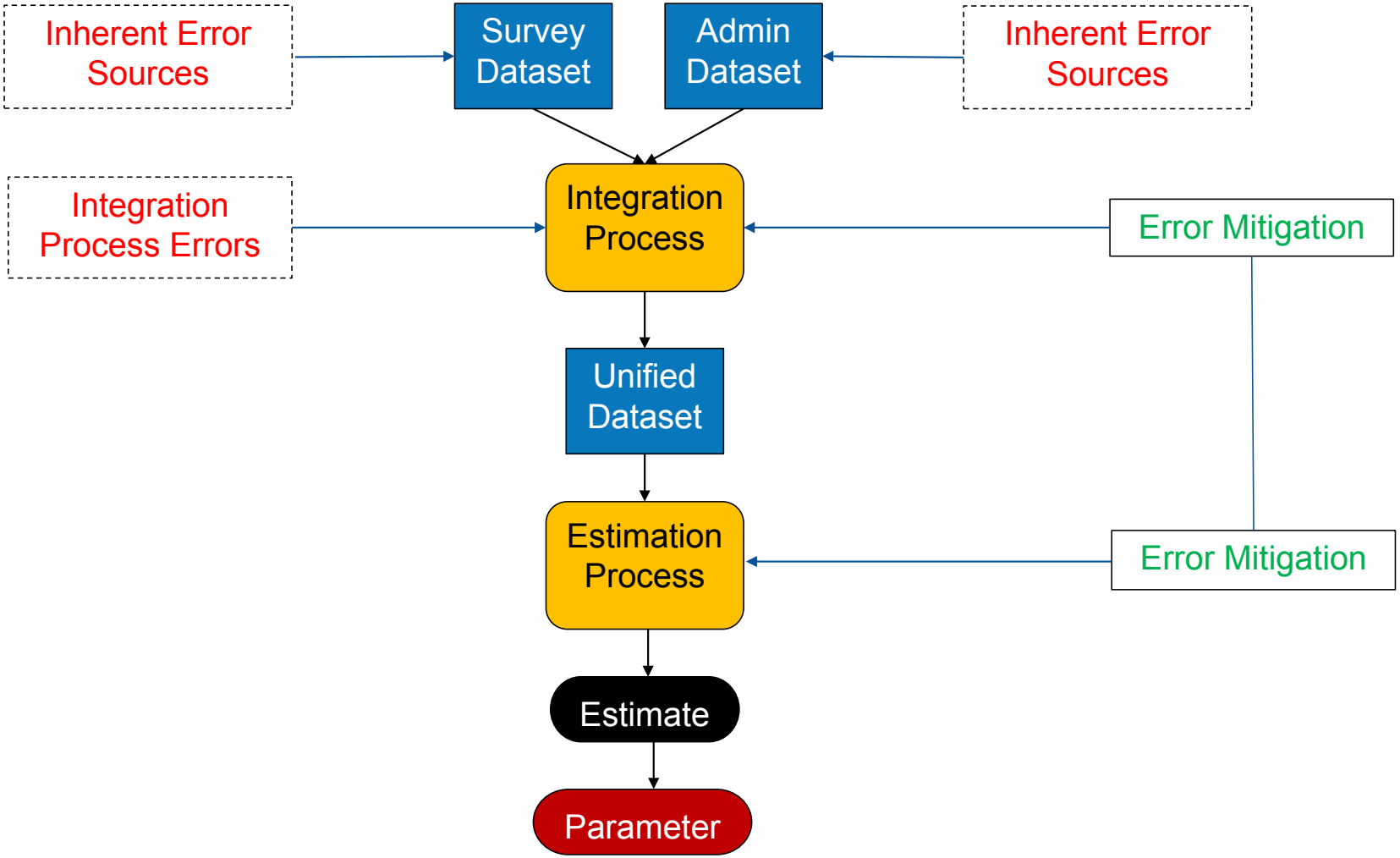
The Hybrid Estimation Process



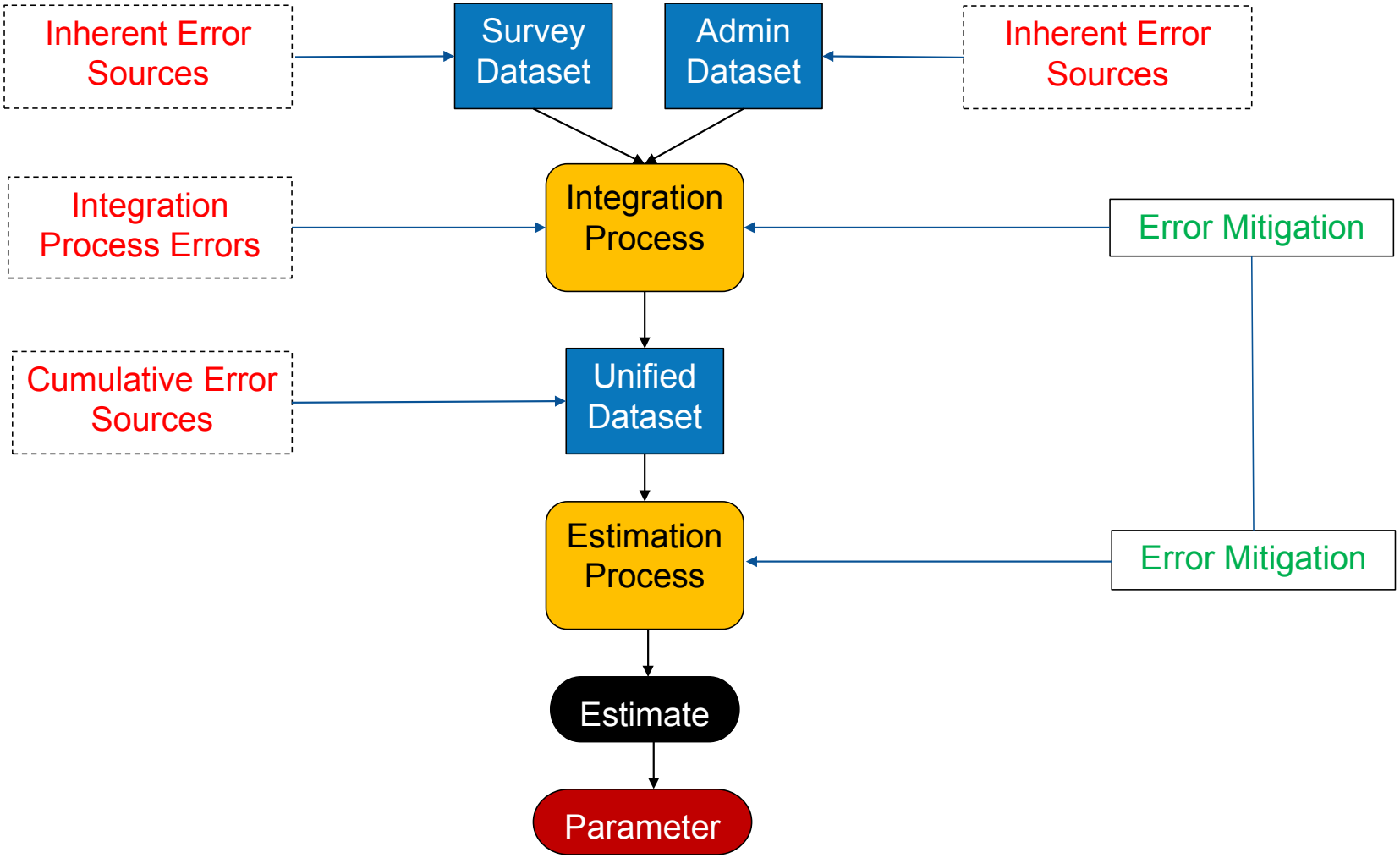
The Hybrid Estimation Process



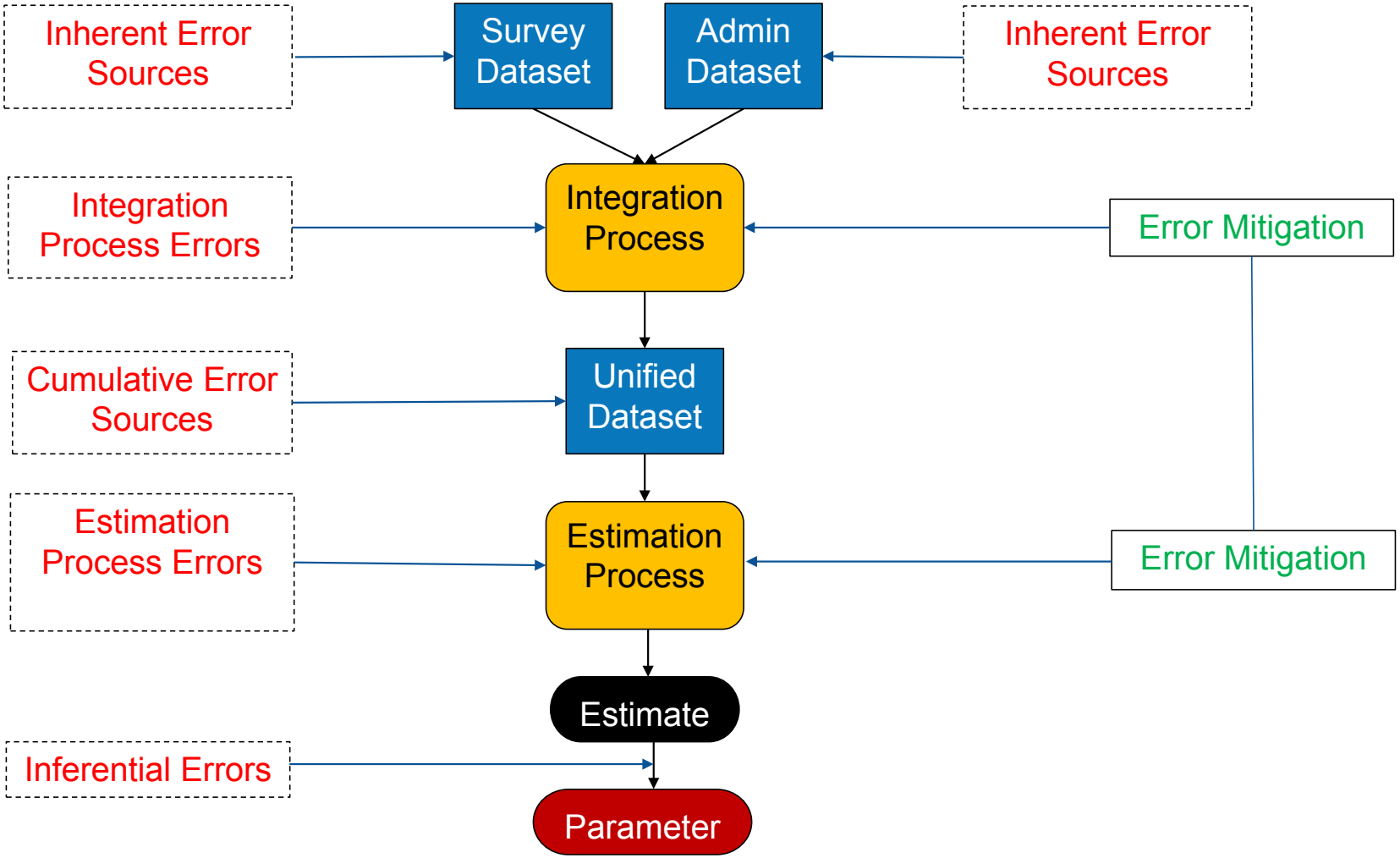
The Hybrid Estimation Process



The Hybrid Estimation Process



The Hybrid Estimation Process

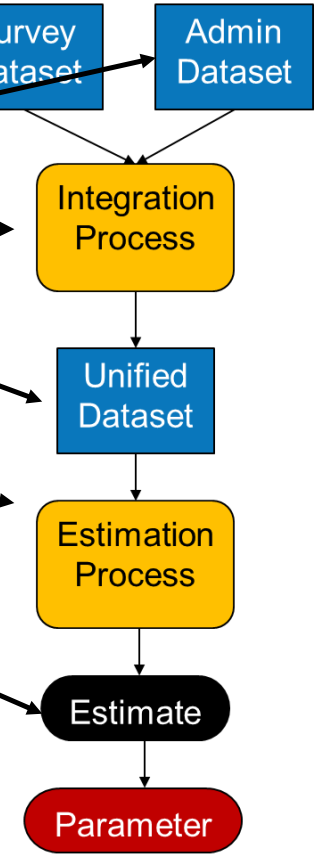


Questions Regarding Hybrid Estimator Accuracy

- What error sources are associated with the unified dataset?
- Which of these pose the greatest *intrinsic* risks to data accuracy?
- Among the hybrid estimators that might be constructed from the unified dataset, which estimator minimizes the *total error risk*?
- What are the major *intrinsic* and *residual* error risks associated with the **hybrid** estimator?
- Which of these error risks could be further mitigated to maximally increase the accuracy of the hybrid estimator?

A Total Error Framework Can be Specified for Each Stage of the Process

- Inherent errors associated with survey dataset
- Inherent errors associated with administrative dataset
- Integration processing errors
- Cumulative errors associated with the unified dataset
- Estimation processing errors
- Final errors associated with the estimate



Error mitigation can occur at various stages of each the process

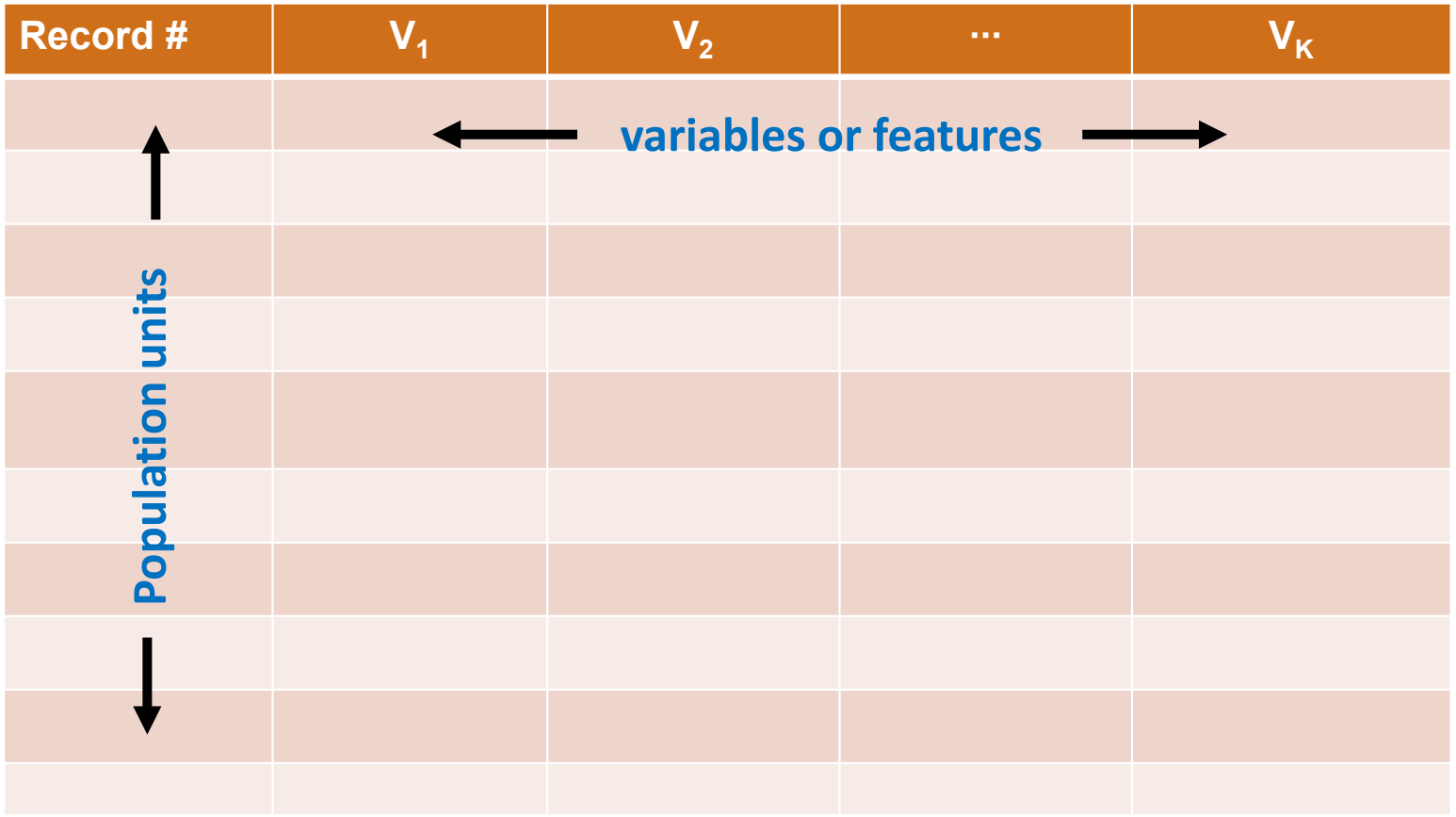
(see, for example, 3 stage framework of Reid, et al, 2017)

In many cases it suffices to simply describe the errors in the final output

- Total error model for registers, frames and other datasets
- Total error model for survey point estimates
- Total error model for hybrid estimates
- Total error models for compilations such as the GDP and various price indexes

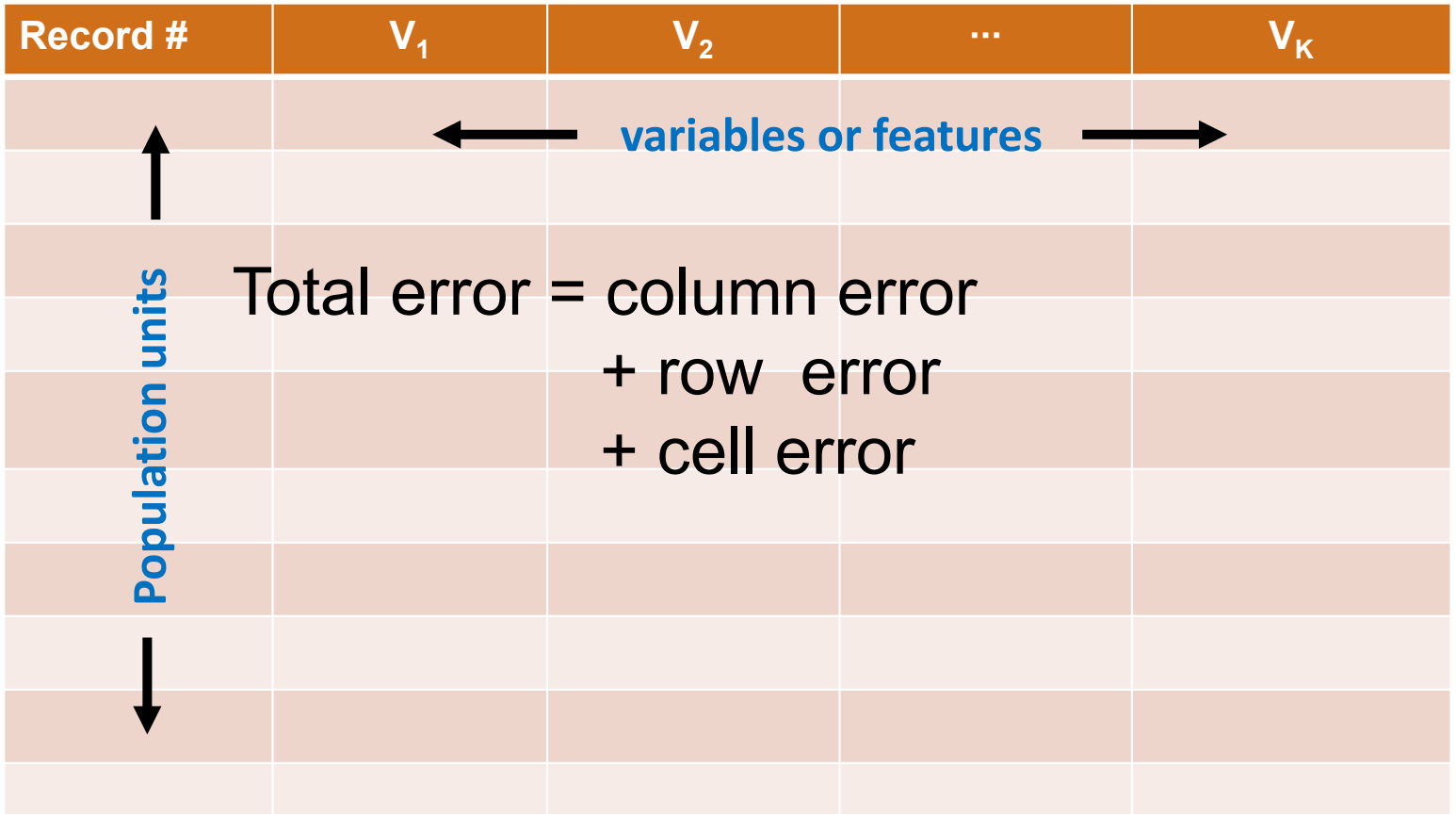
A Total Error Framework for a Generic Dataset

Typical File Structure



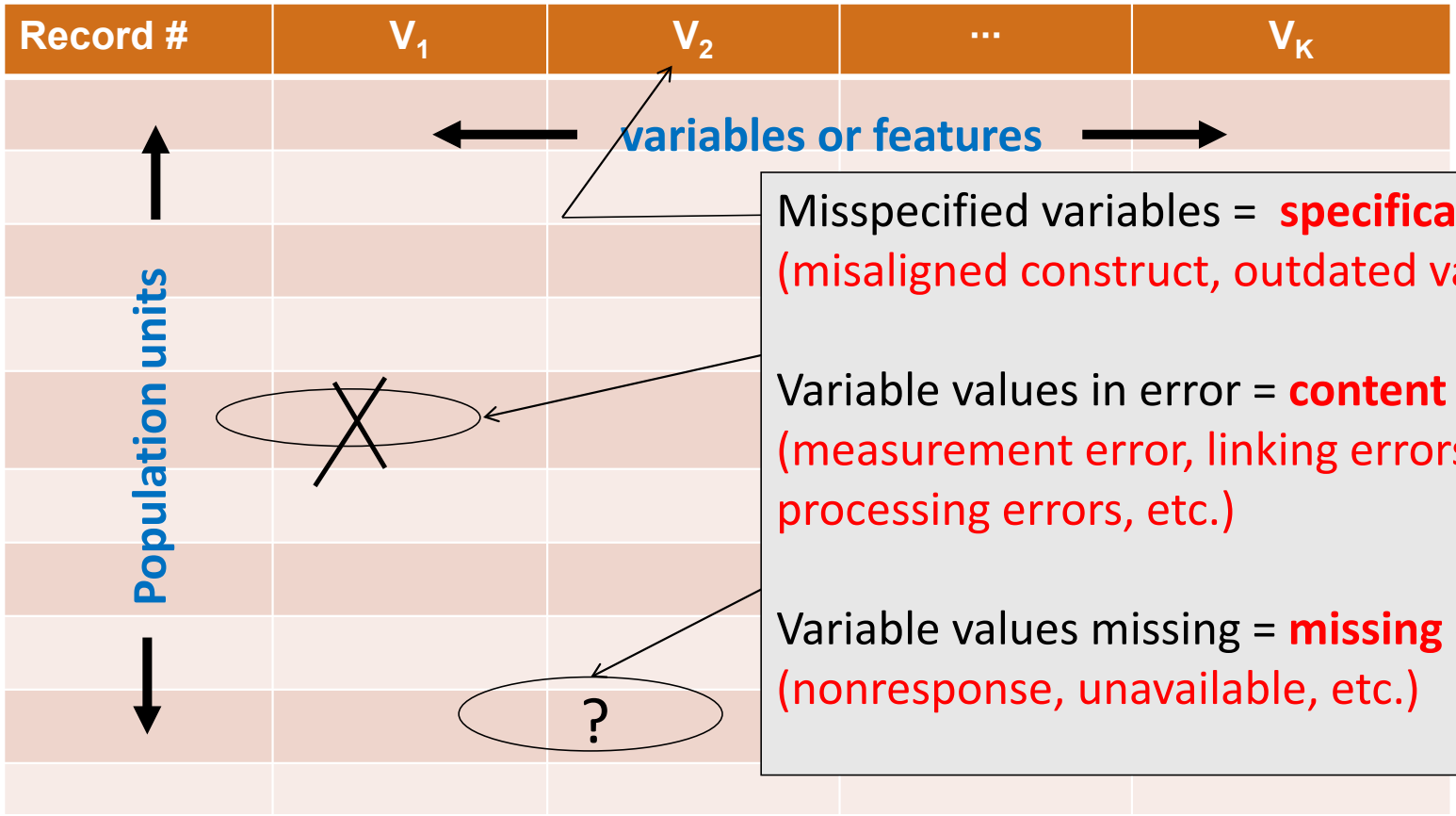
A Total Error Framework for a Generic Dataset

Typical File Structure



Column and Cell Errors

Typical File Structure

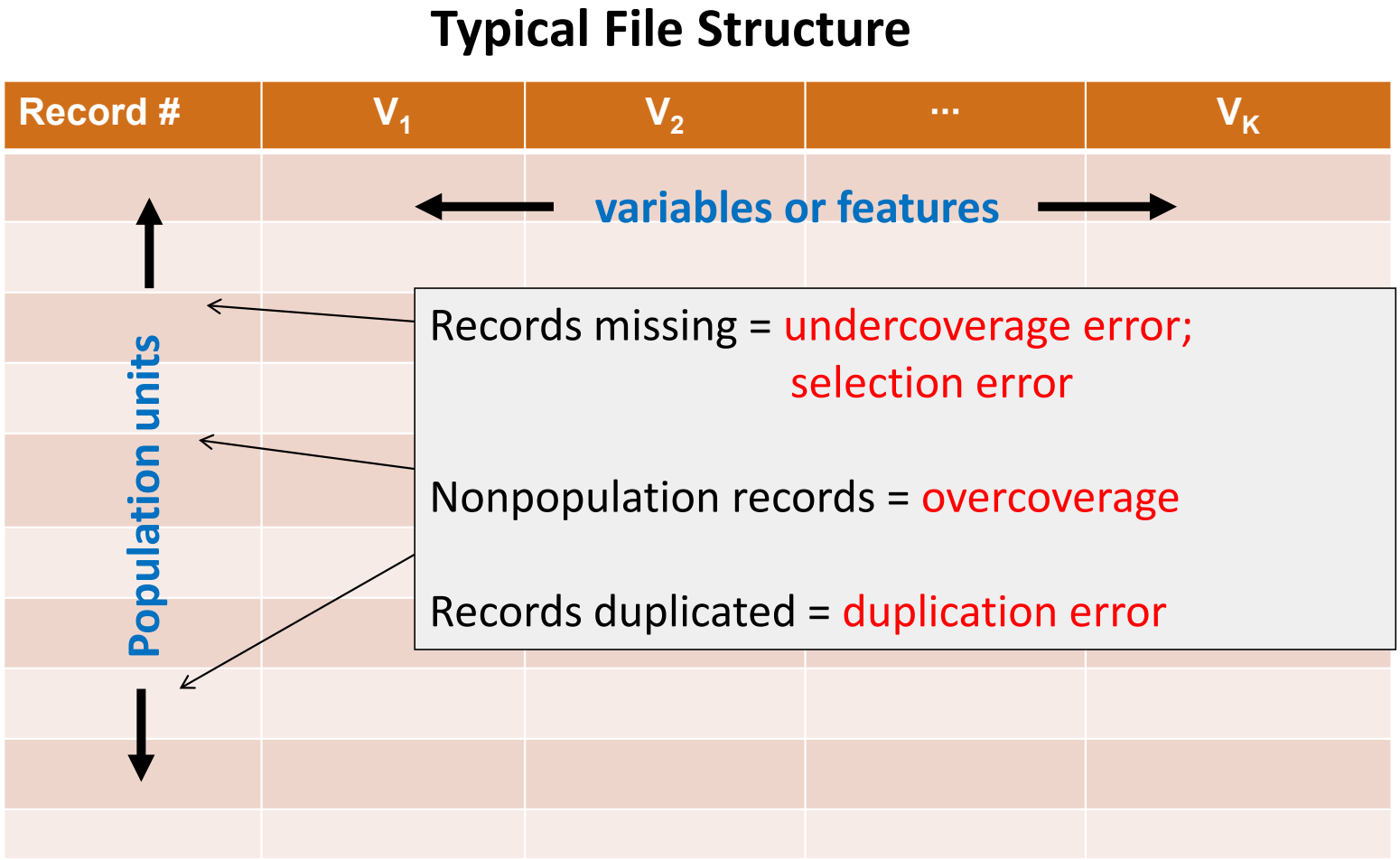


Misspecified variables = **specification error**
(misaligned construct, outdated value, etc.)

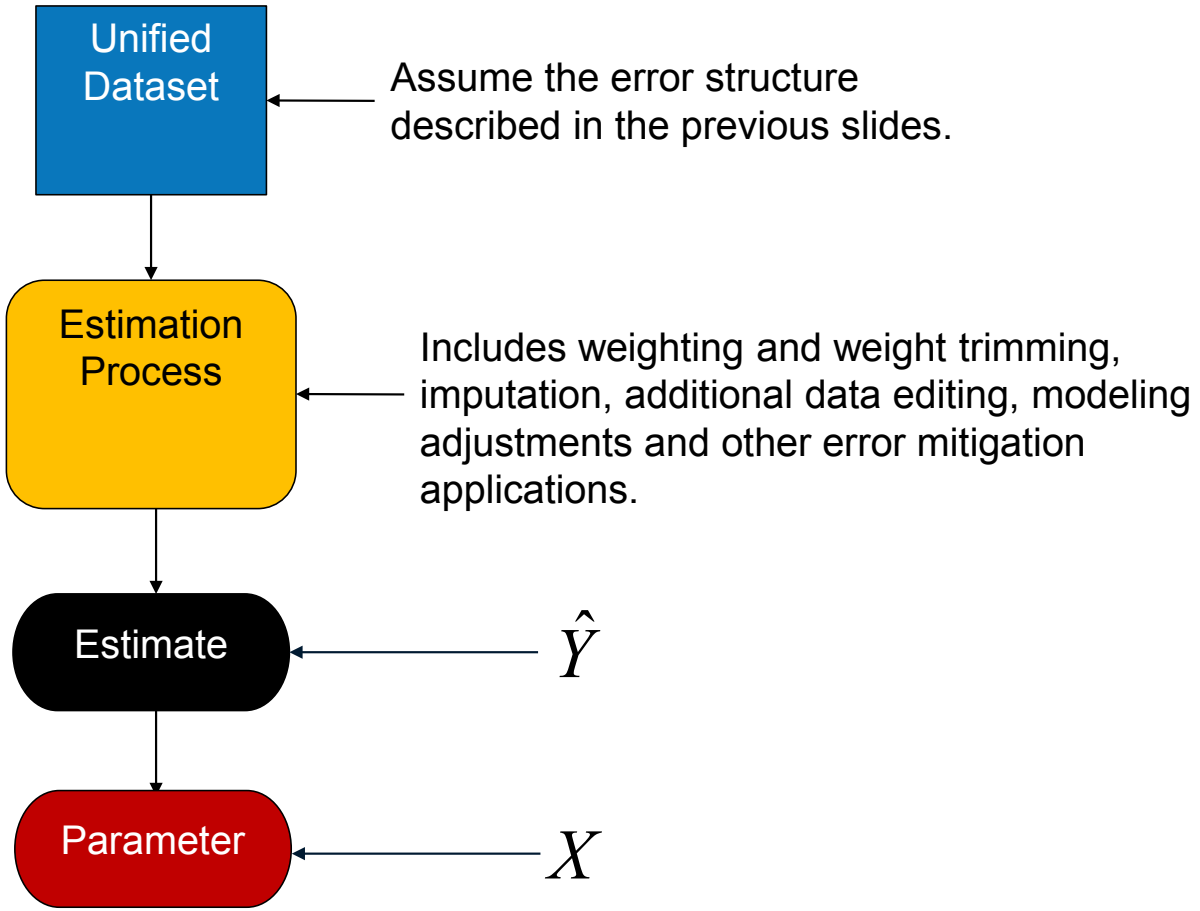
Variable values in error = **content error**
(measurement error, linking errors, data processing errors, etc.)

Variable values missing = **missing data**
(nonresponse, unavailable, etc.)

Row errors


Typical File Structure
Record #
V_1
V_2
…
V_K
variables or features
Population units
Records missing = undercoverage error; selection error
Nonpopulation records = overcoverage
Records duplicated = duplication error

Errors Associated with the Hybrid Estimation Process



Total Error Model for Hybrid Estimators

$$\underbrace{\hat{Y} - X}_{\text{total error}} = \underbrace{(\hat{Y} - Y)}_{(\varepsilon_1 + \cdots + \varepsilon_6)} + \underbrace{(Y - X)}_{\varepsilon_7}$$

$\varepsilon_1 =$ Selection error
$\varepsilon_2 =$ Coverage error (over-, under-, duplication)
$\varepsilon_3 =$ Missing data error
$\varepsilon_4 =$ Content error
$\varepsilon_5 =$ Data processing error
$\varepsilon_6 =$ Model/estimation error
$\varepsilon_7 =$ Specification error

Assessing Error Risk

Types of Error Risks

- Intrinsic risk – risk that an error source poses if no steps are taken to reduce the error; error risk of "doing nothing."
  - Example: The intrinsic risk of nonresponse bias in an linear estimator is

$$B_I = \frac{\mathrm{cov}(y_i, \rho_i)}{\bar{\rho}}$$

- Residual risk – risk of error for a source that remains after mitigation strategies have been applied.
  - Example: After nonresponse weighting adjustments have been applied, the residual risk of bias is

$$B_R \leq B_I$$

Risk Profile Comparing Survey, Administrative and Unified Datasets: Either Intrinsic or Residual Risks

Error Sources	*Survey Dataset*	*Administrative Dataset*	*Unified Dataset*
Specification	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Undercoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Overcoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Duplication	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Selection	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Content	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Missing Data	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)

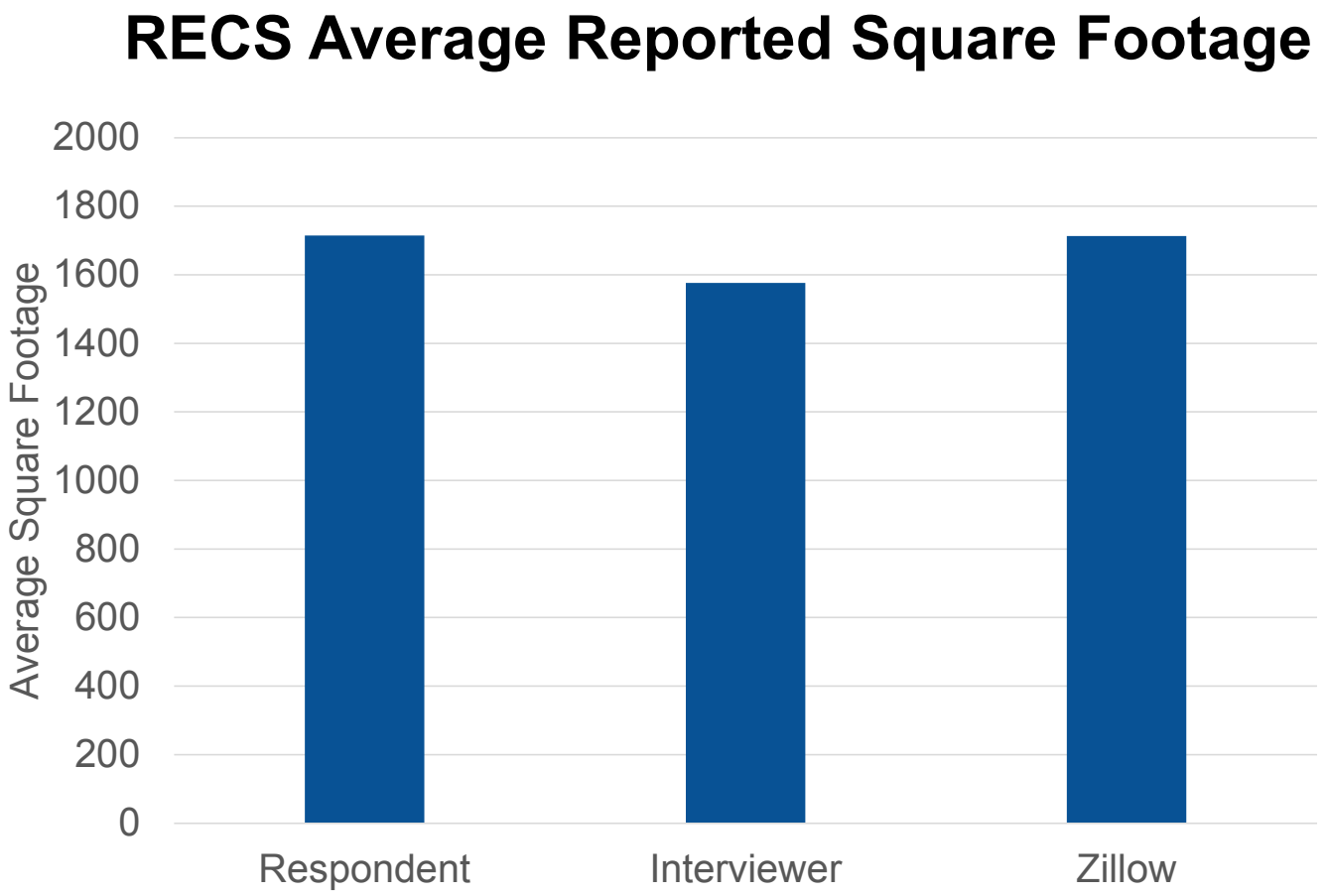
Intrinsic Error Risk Profile Comparing Survey and Hybrid Estimates

Error Sources	*Survey Estimator*	*Hybrid Estimator*
Specification	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Undercoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Overcoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Duplication	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Sampling/Selection	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Measurement/Content	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Data Processing	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Nonresponse/Missing data	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Modeling/estimation	Risk level (1, 2, 3)	Risk level (1, 2, 3)

Case Study: Error Mitigation for Energy Use Survey Square Footage Data using Unified Data

- Data sources
  - Survey data: 2015 Residential Energy Consumption Survey (RECS)
    - n ≈ 2,400 completed cases
  - Big Data (data pulled from various sources)
    - Zillow
    - Acxiom
    - CoreLogic
- Variable of interest: housing unit square footage
- **Goal: Integrate the external data sources with the survey data to improve and/or evaluate the accuracy of survey square footage data**

Evidence of Nonsampling Error from the RECS



More Evidence of Intrinsic Error Risks

	Survey (R)	Zillow
NR/Missing		
Unit NR/missing rate	58.2%	22.0%
Item NR rate	19.2%	
Overcoverage rate	11.8%	~0
Undercoverage rate	~0	15.3%
Reliability	50%	68%

Intrinsic Error Risk Profile for the RECS, Zillow and Unified Datasets

Error Sources	*RECS*	*Zillow*	*RECS U Zillow*
Specification	2	2	2
Coverage: Undercoverage	1	2	1
Coverage: Overcoverage	2	1	1
Coverage: Duplication	2	1	1
Selection	3	1	3
Content	3	3	3
Missing Data	3	2	1
Average	**2.3**	**1.7**	**1.7**

Intrinsic Error Risk Profile for the RECS, Zillow and Unified Datasets

Error Sources	*RECS*	*Zillow*	*RECS U Zillow*
Specification	2	2	2
Coverage: Undercoverage	1	2	1
Coverage: Overcoverage	2	1	1
Coverage: Duplication	2	1	1
Selection	3	1	3
Content	3	3	3
Missing Data	3	2	1
Average	**2.3**	**1.7**	**1.7**

Unified data offers no advantage to Zillow only dataset.

Intrinsic Error Risk Profile RECS and RECS/Zillow Hybrid Estimates

Error Sources	*RECS Estimator*	*RECS/Zillow Hybrid Estimator*
Specification	1	2
Coverage: Undercoverage	2	1
Coverage: Overcoverage	1	1
Coverage: Duplication	2	1
Sampling/Selection	3	1
Measurement/Content	3	2
Data Processing	2	2
Nonresponse/Missing data	3	1
Modeling/estimation	3	3
Average	**2.2**	**1.6**

Initial evaluations suggest a total error reduction with the hybrid estimator even before error mitigation efforts have been fully exploited.

Illustration 2 – Market Research Client

- Currently conducting a large scale survey to evaluate market share for it customers products about 300 markets
- Quarterly estimates tend to be unstable in some markets
- Various administrative sets have been identified that would improve estimator stability, but each brings with it other error risks that have been fully investigated

Question:

Can a hybrid estimator be constructed having greater stability than the current survey estimator without increasing total error?

Intrinsic Risk for the Hybrid Estimator Compared to Estimators Based on the Survey and Administrative Data

Quality Component	Survey	Admin Data	Hybrid Estimator
Specification	**3**	**2**	**3**
Coverage	**2.7**	**2.0**	**2.7**
Undercoverage	3	1	3
Duplication	3	3	3
Within unit	2	2	2
Selection	**2.5**	**3**	**1.5**
Sample size	2	N/A	1
Weight variation	3	3	2
Nonresponse/Missing Data	**1.5**	**2**	**2**
Unit	1	3	2
Item	2	1	2
Measurement	**2**	**3**	**3**
Data processing	**2**	**1**	**2**
Keying/editing	1	1	1
Design weighting	3	N/A	3
Estimation/modeling	**1**	**1**	**1**
Analysis	**1**	**3**	**1**
Overall assessment	***2.0***	***2.1***	***2.0***

Summary

- A total error framework decomposes total error so that key subcomponents can be identified and addressed.
- A unified error risk framework facilitates comparisons across individual and unified data sources.
- An error risk profile can provide insights regarding the quality implications of unified datasets
  - Assesses intrinsic risks by error source
  - Helps determine whether residual risk can be reduced by data unification.

Thank you!

Please direct inquiries to:

Paul Biemer: ppb@rti.org