RTI INTERNATIONAL | turning knowledge into practice

# Total ~~Survey~~ Error: Adapting the Paradigm for Big Data

Paul Biemer
RTI International
University of North Carolina

# Acknowledgements


- Phil Cooley, RTI
- Alan Blatecky, RTI

# Why is a ‘total error’ framework needed?

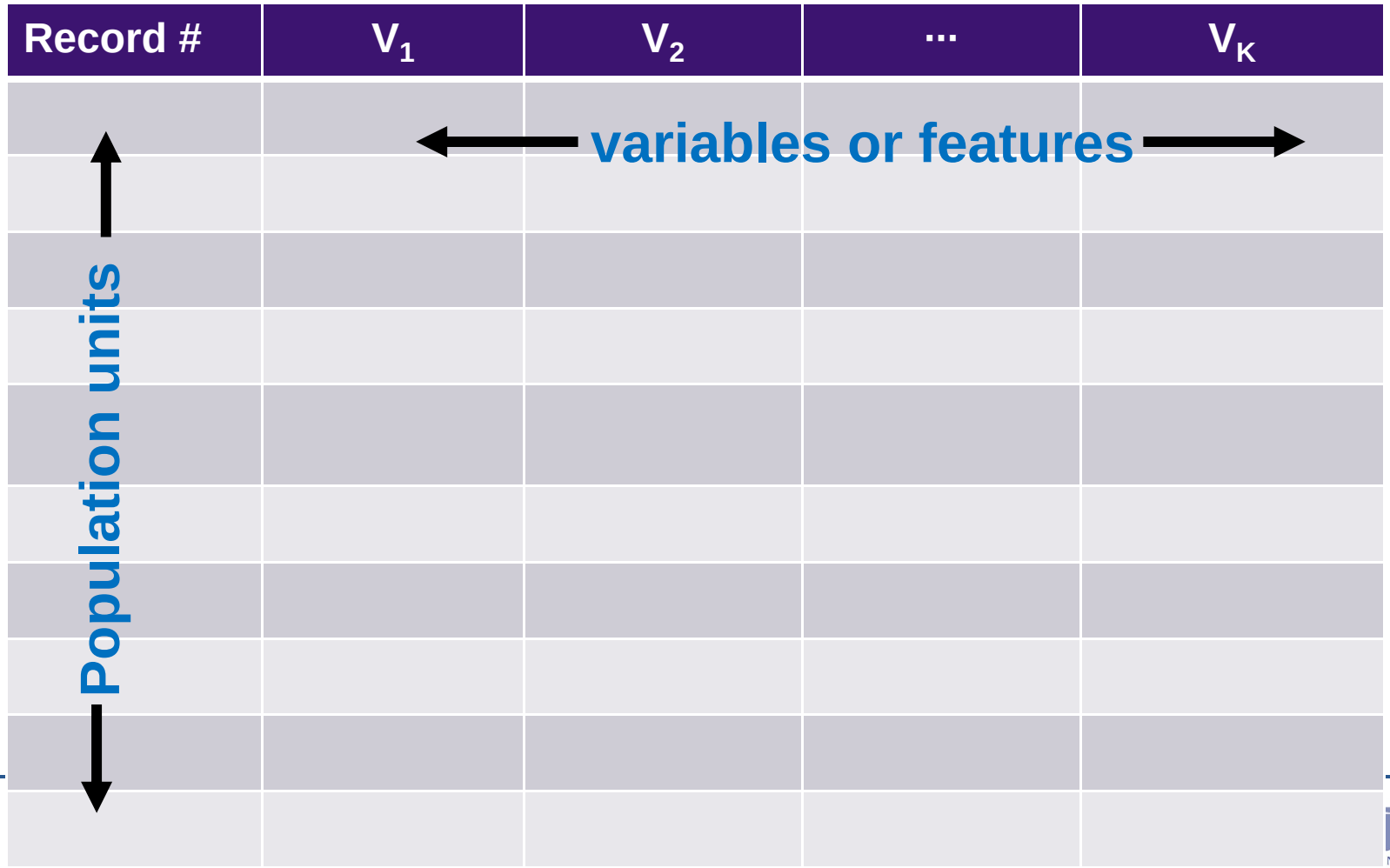
- Large and important errors are inevitable for Big Data
  - Big Data are inherent ‘noisy’
- The total error and its sources are not well-known or studied for Big Data
  - ‘N → all’ does not imply ‘error → 0’
- Such errors lead to erroneous inferences, predictions, conclusions, and decisions
- Awareness of the errors is the first step to addressing their causes and reducing their effects

# What is a total error framework?

- Identifies all major sources of error contributing to data and/or estimator inaccuracy
- Describes the nature of the error sources and how the errors could affect inference
- Maps the errors onto components of uncertainty (for e.g., bias and variance)
- Provides insights regarding how error components affect estimation and inference
- Suggests methods for reducing the effects of errors on inference

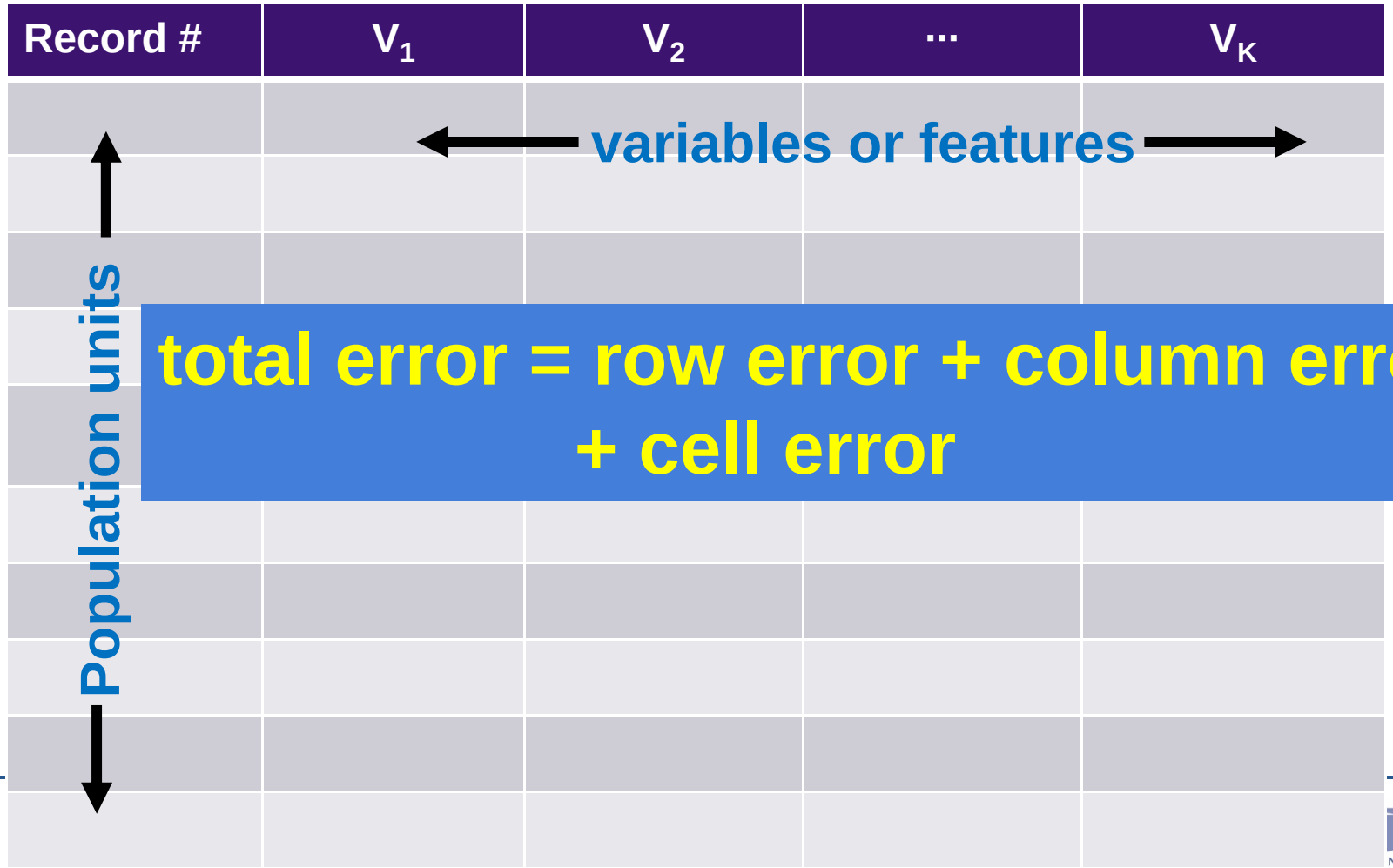
# Total Error Framework for Traditional Data Sets

## Typical File Structure

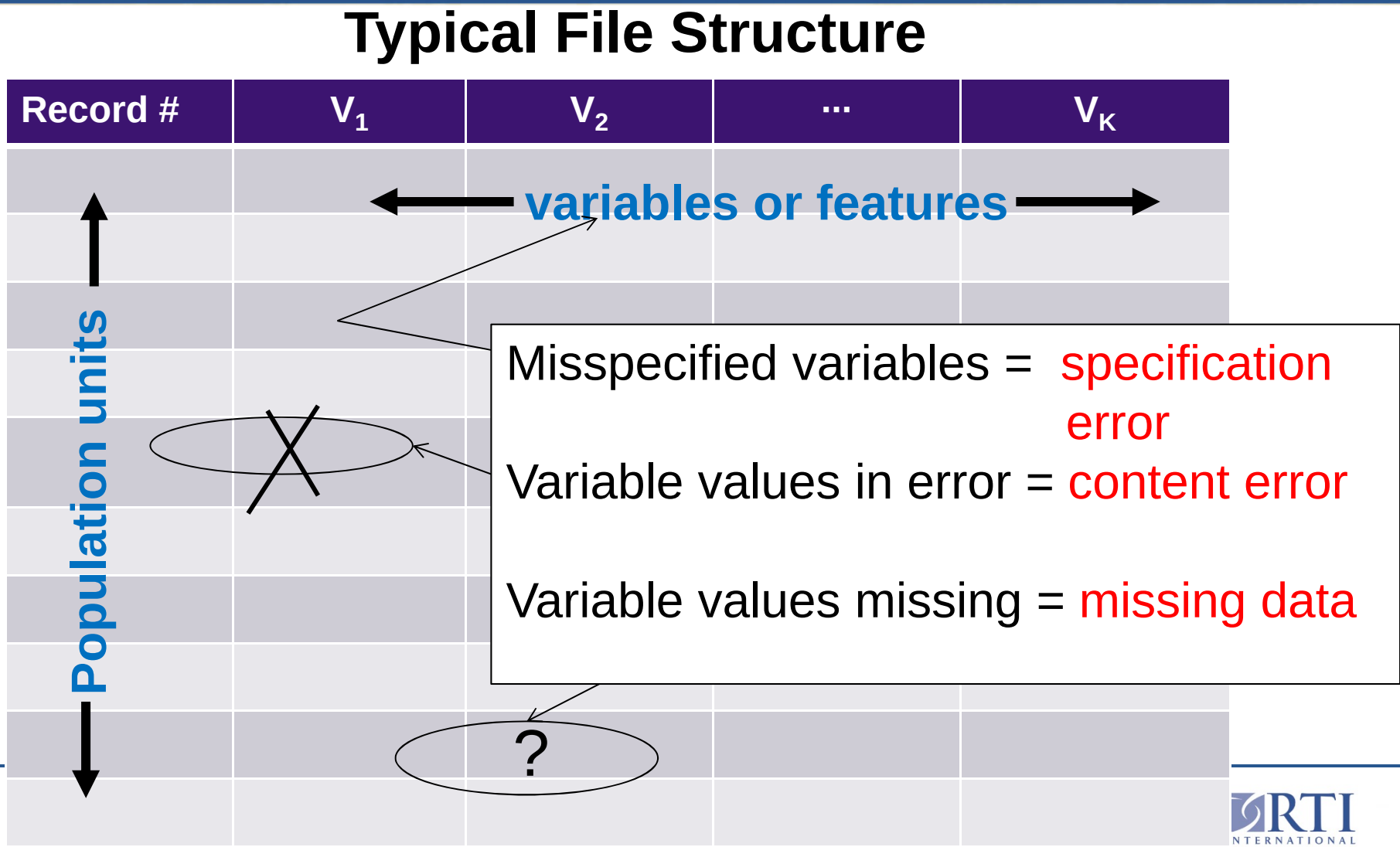


# Total Error Framework for Traditional Data Sets

## Typical File Structure

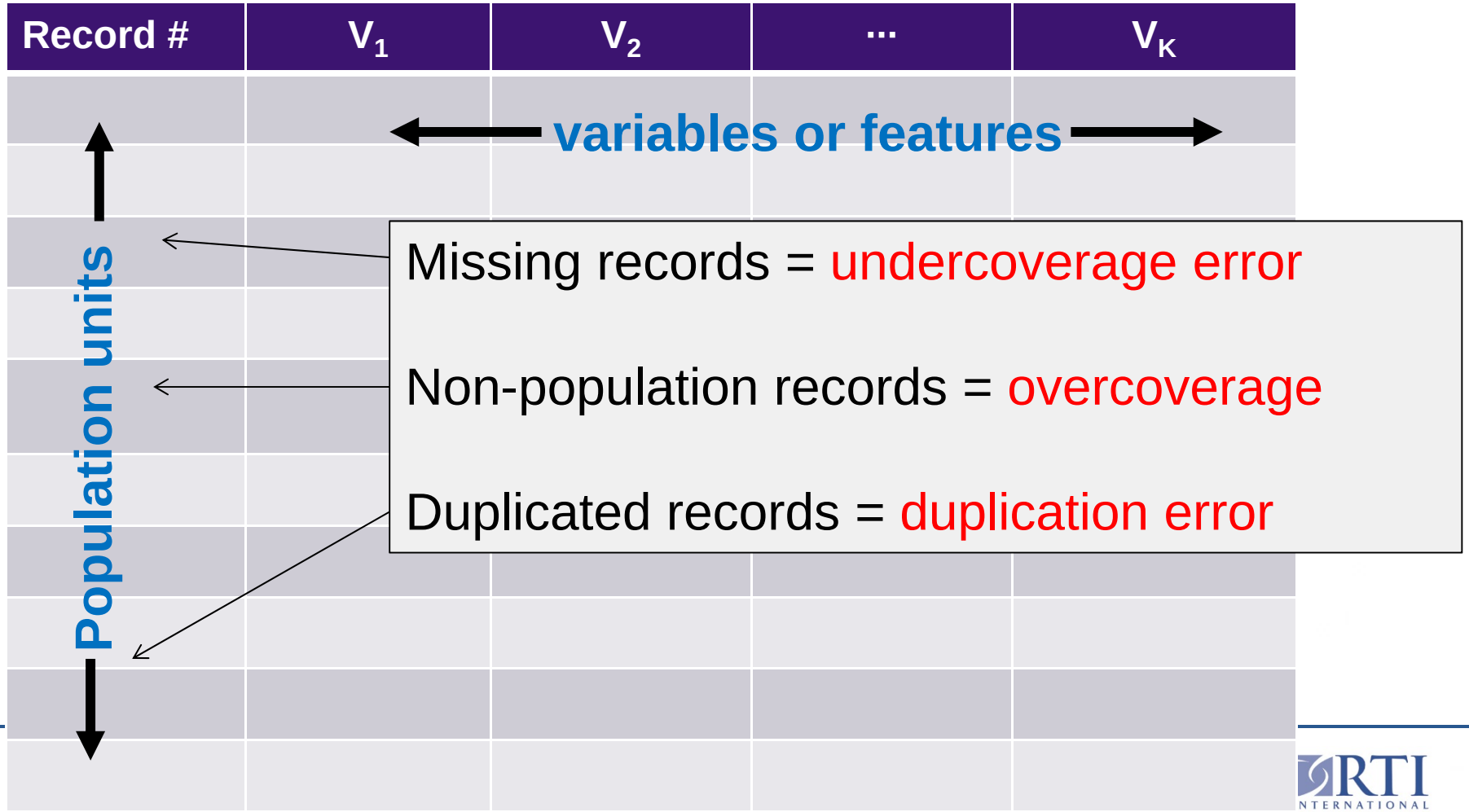


# Possible Column and Cell Errors



# Possible Row Errors

## Typical File Structure



# Shortcomings of the Framework for Big Data

- Big Data files are often not rectangular
  - hierarchically structure or unstructured
- Data may be distributed across many data bases
  - Sometimes federated, but often not
- Data sources may be quite heterogeneous
  - Includes texts, sensors, transactions, and images
- Errors generated by Map/Reduce process may not lend themselves to column-row representations.

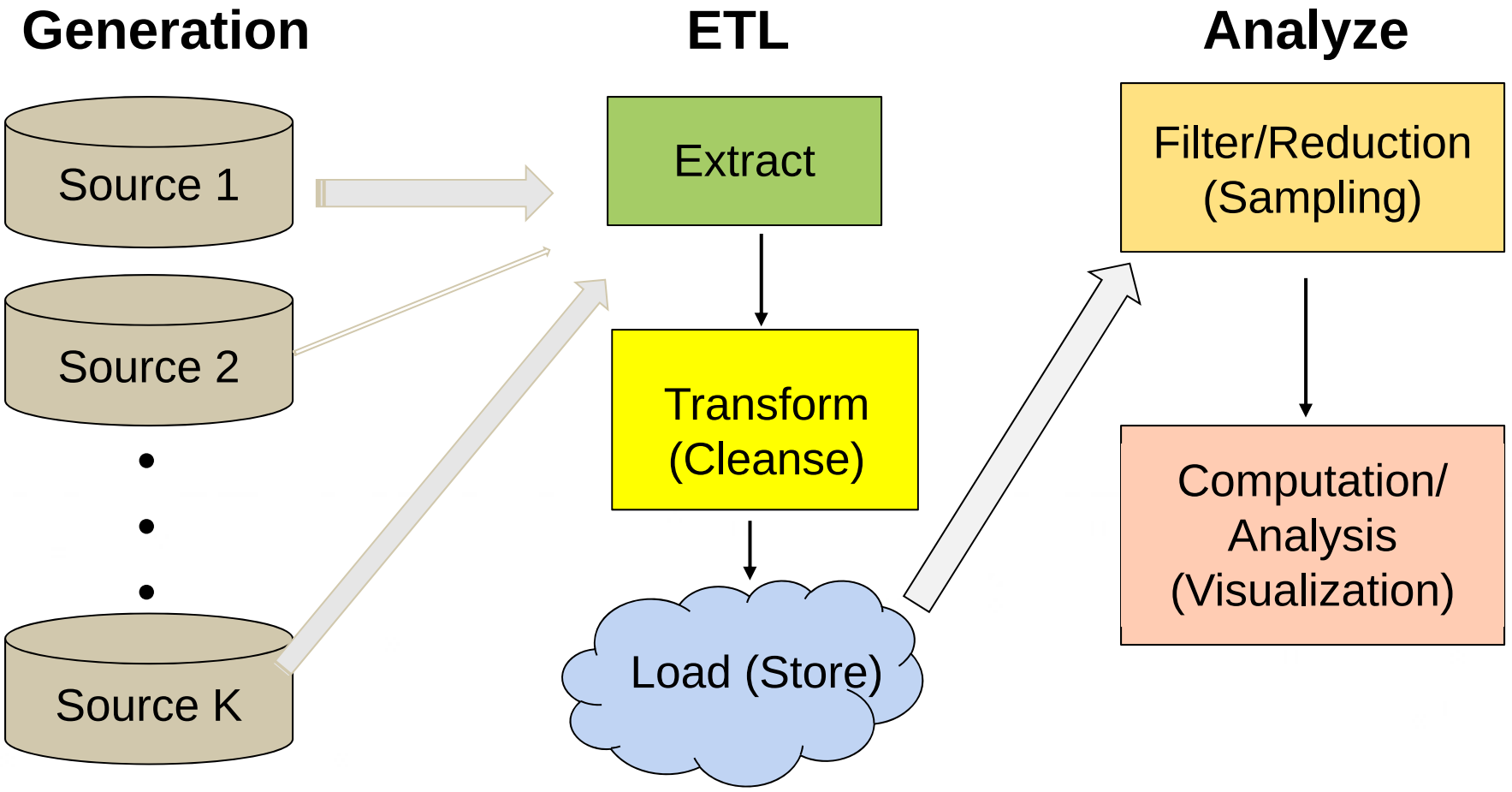
# Big Data Processing Steps that Affect Total Error

- **Generate** –data are generated from some source either incidentally or purposively
- **Extract/Transform/Load (ETL)** – brings all data together in a homogeneous computing environment
  - Extract – data are harvested from their sources, parsed, validated, curated and stored
  - Transform – data are translated, coded, recoded, aggregated/disaggregated, and/or edited
  - Load – data are integrated and stored in the data warehouse

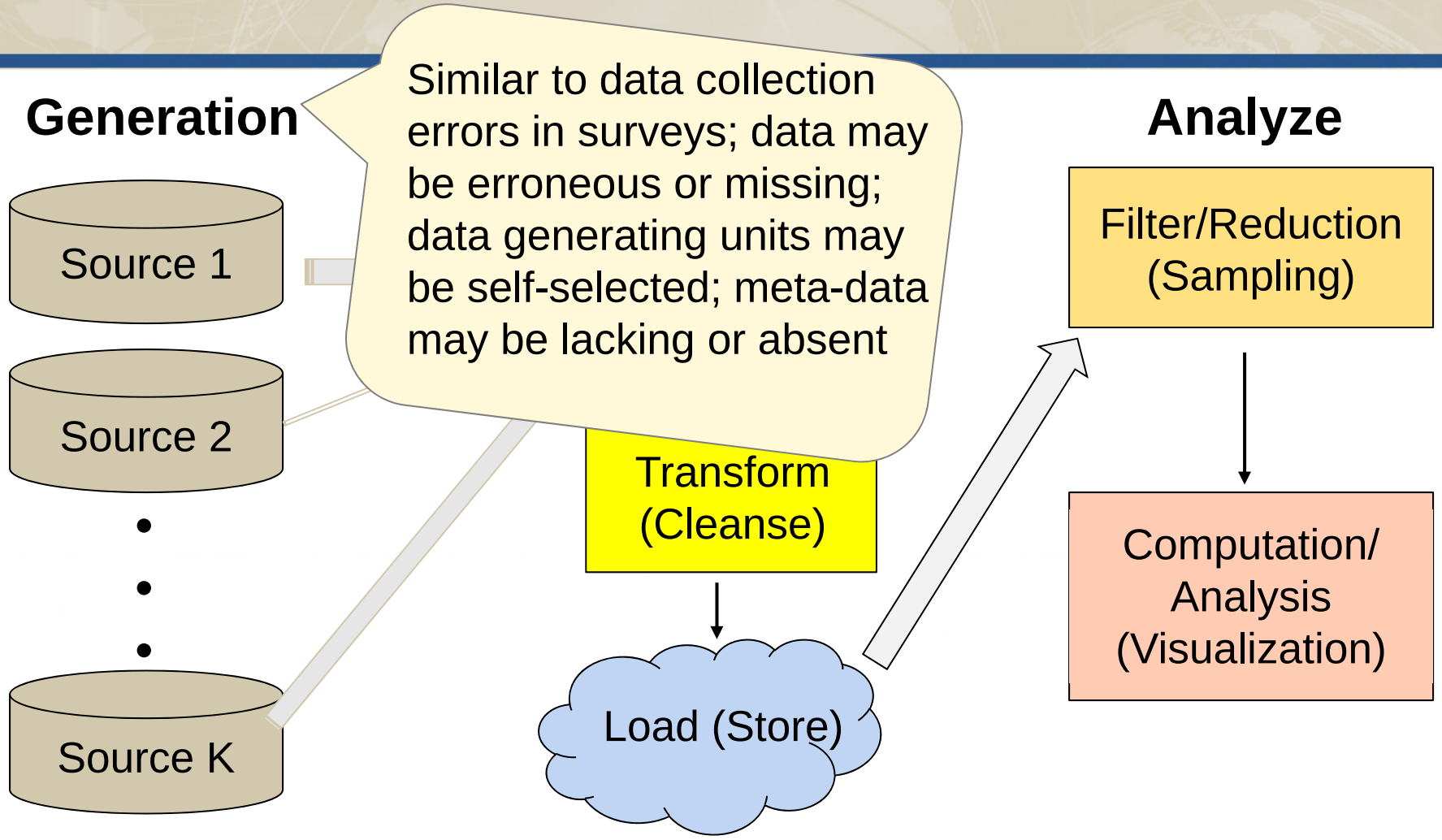
# Big Data Processing Steps that Affect Total Error (continued)

- **Analyze** – Data are converted to information
  - **Filtering (Sampling)/Reduction** –
    - Unwanted features and content are deleted;
    - features may be combined to produced new ones;
    - data elements may be thinned or sampled to be more manageable for the next steps.
  - **Computation/Analysis/Visualization** – data are analyzed and/or presented for interpretation and information extraction.

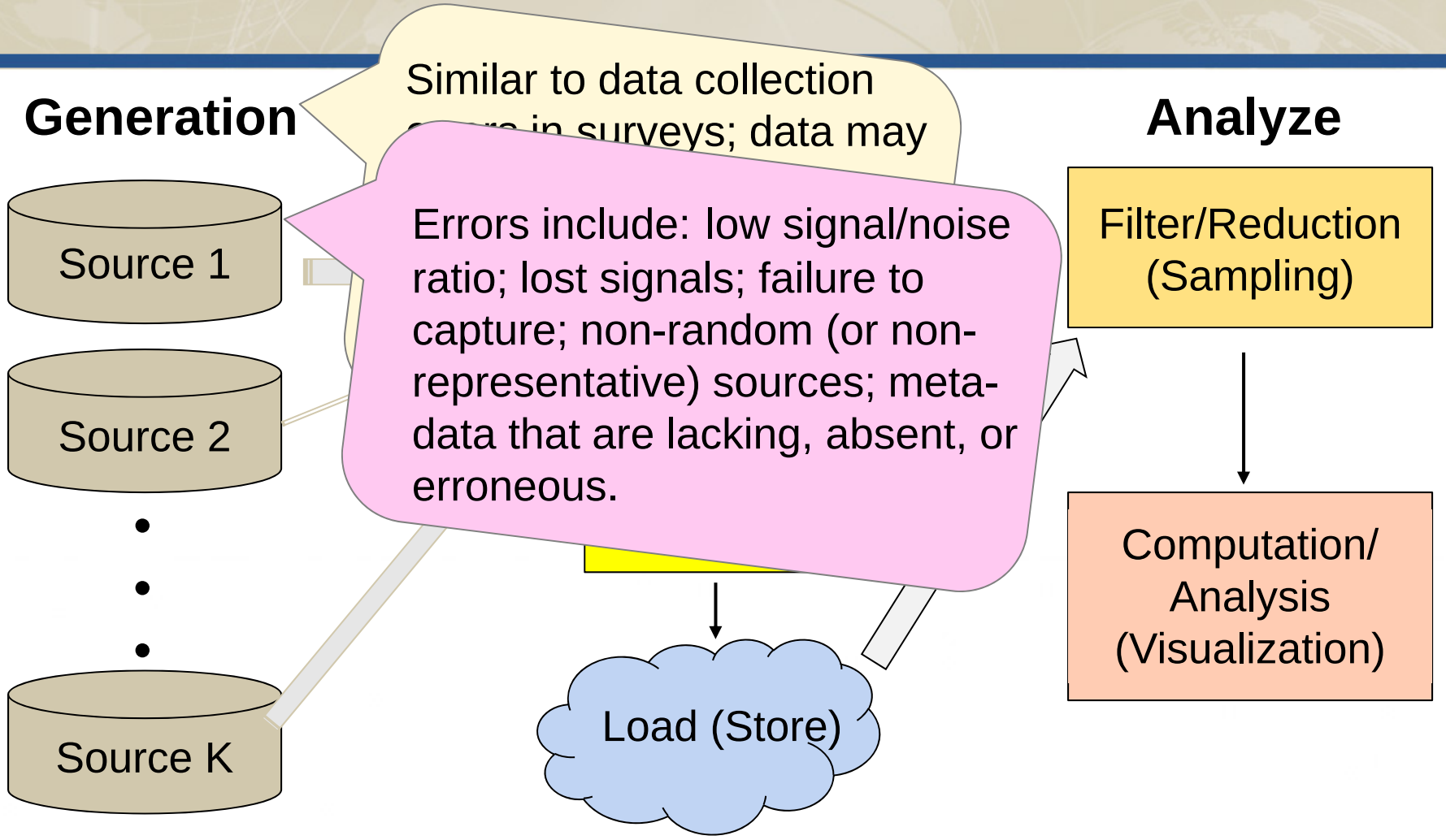
# Big Data Process Map



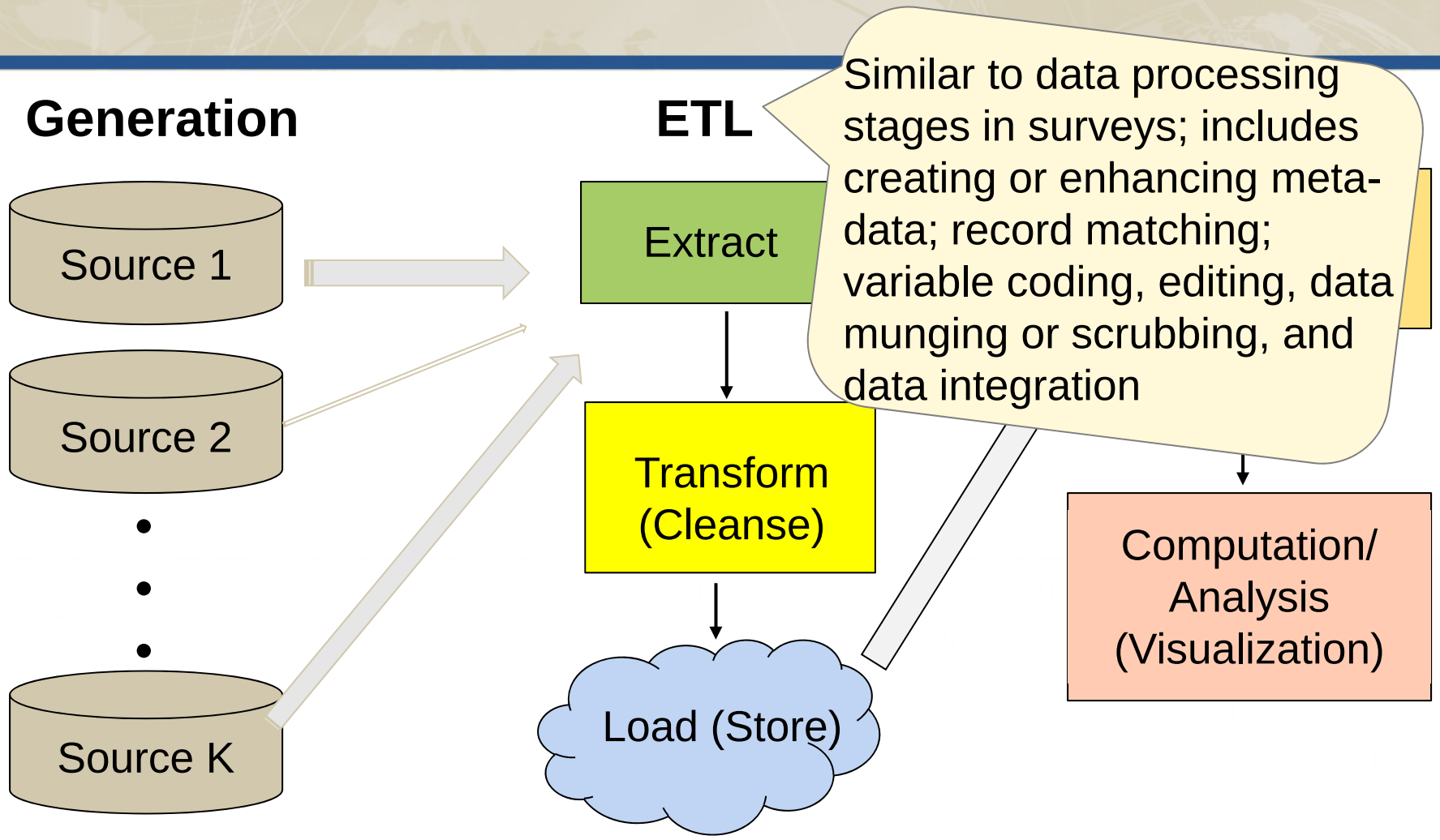
# Big Data Process Map



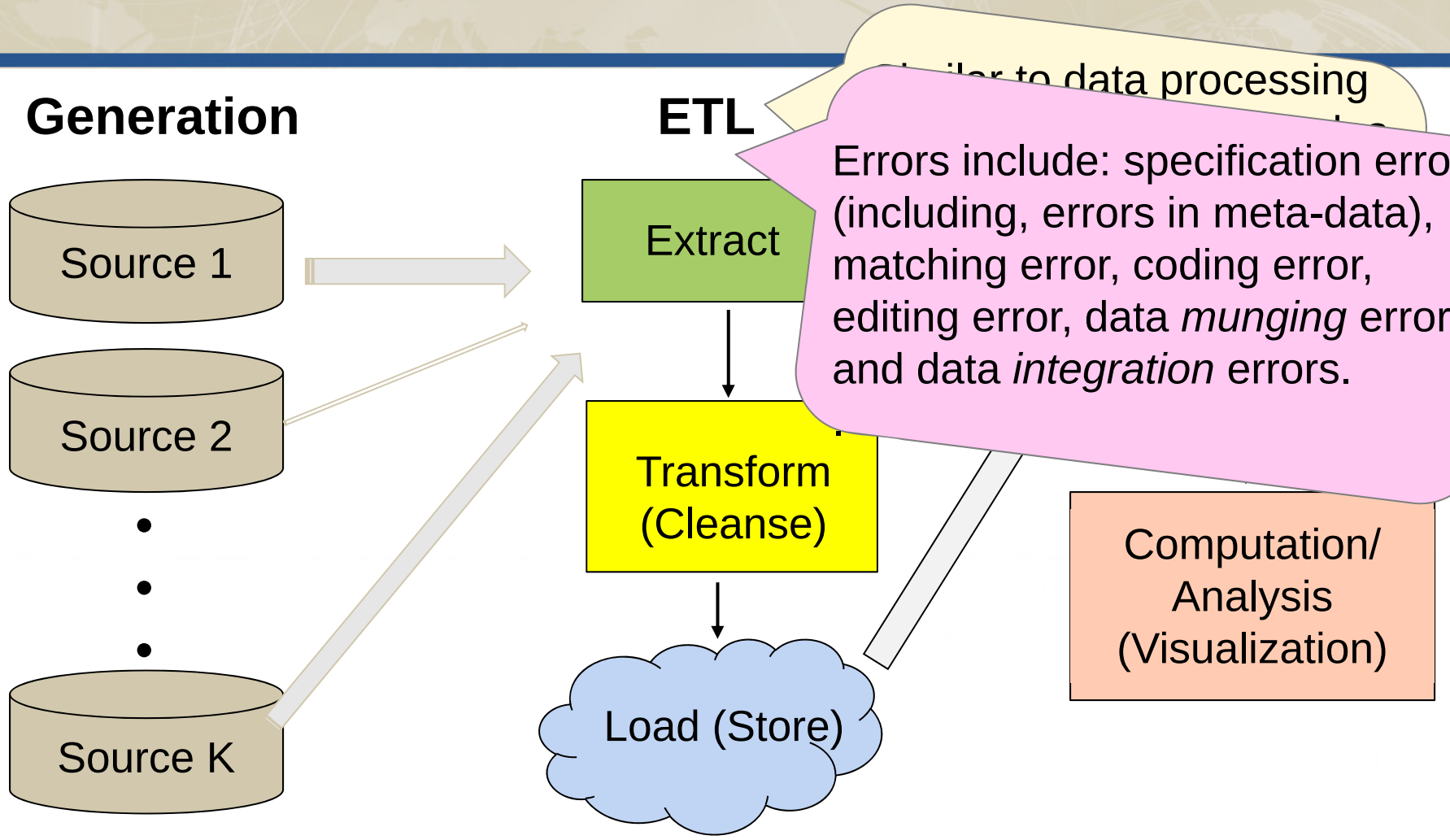
# Big Data Process Map



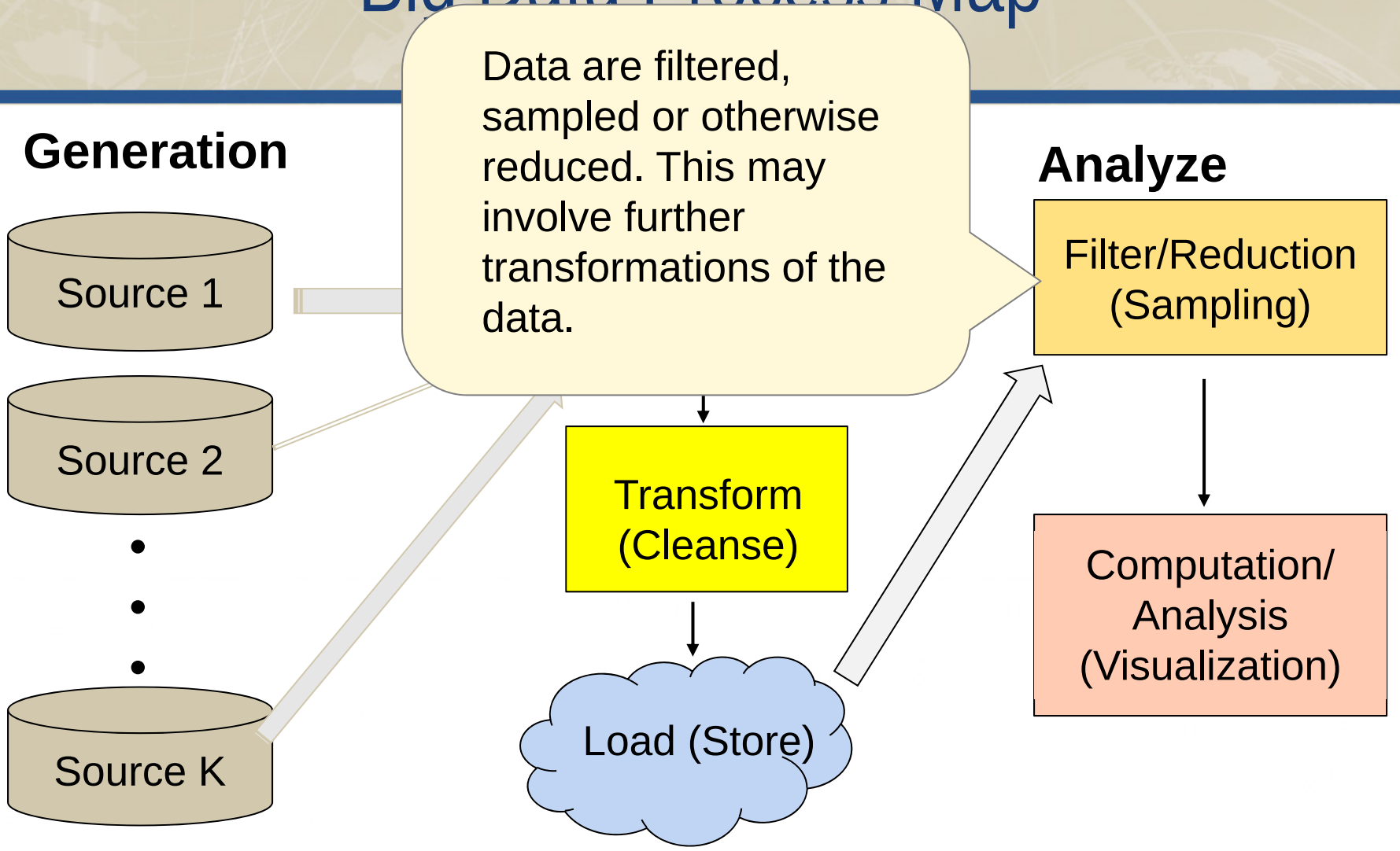
# Big Data Process Map



# Big Data Process Map



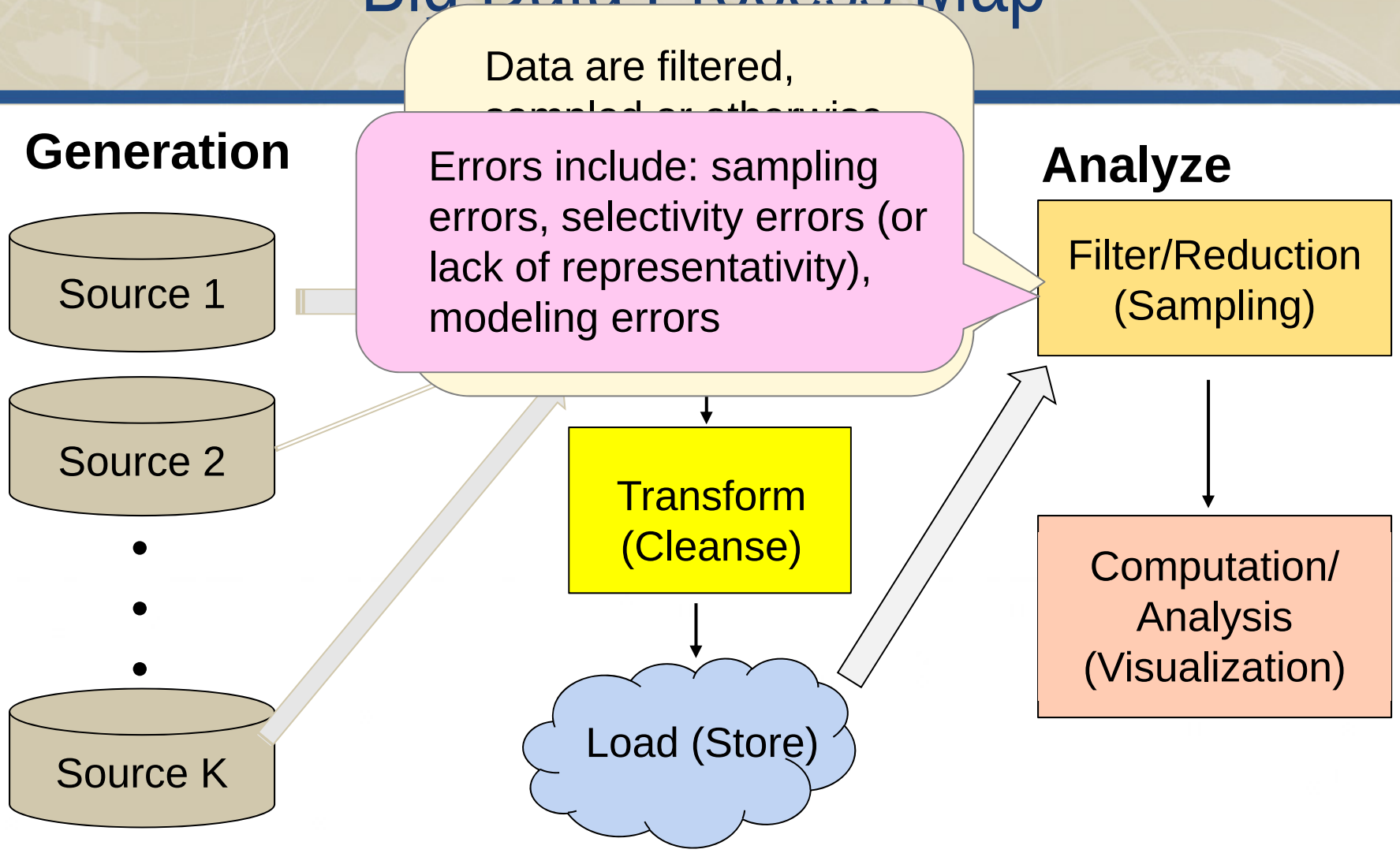
# Big Data Process Map






RTI INTERNATIONAL

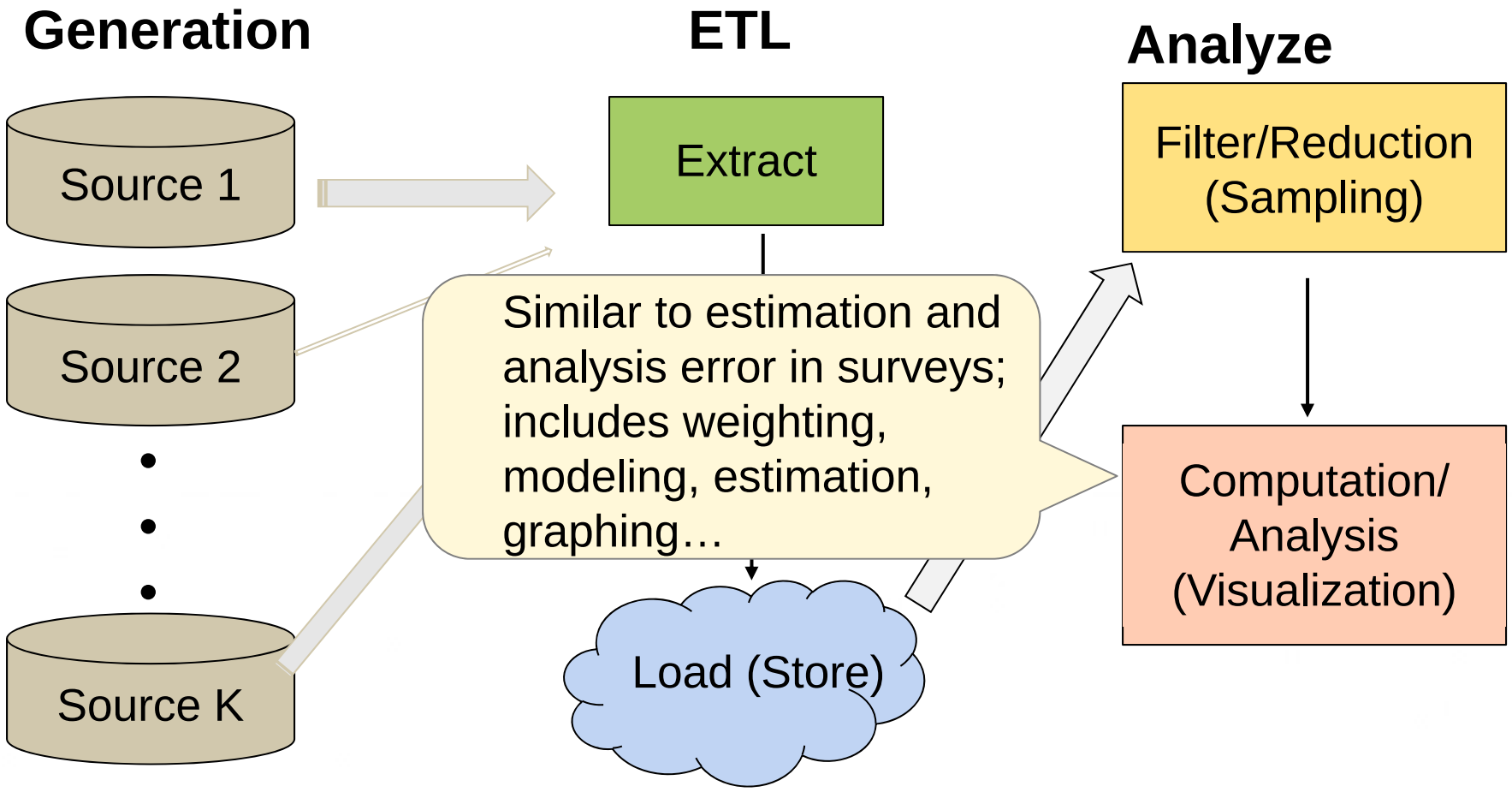
# Big Data Process Map






RTI INTERNATIONAL

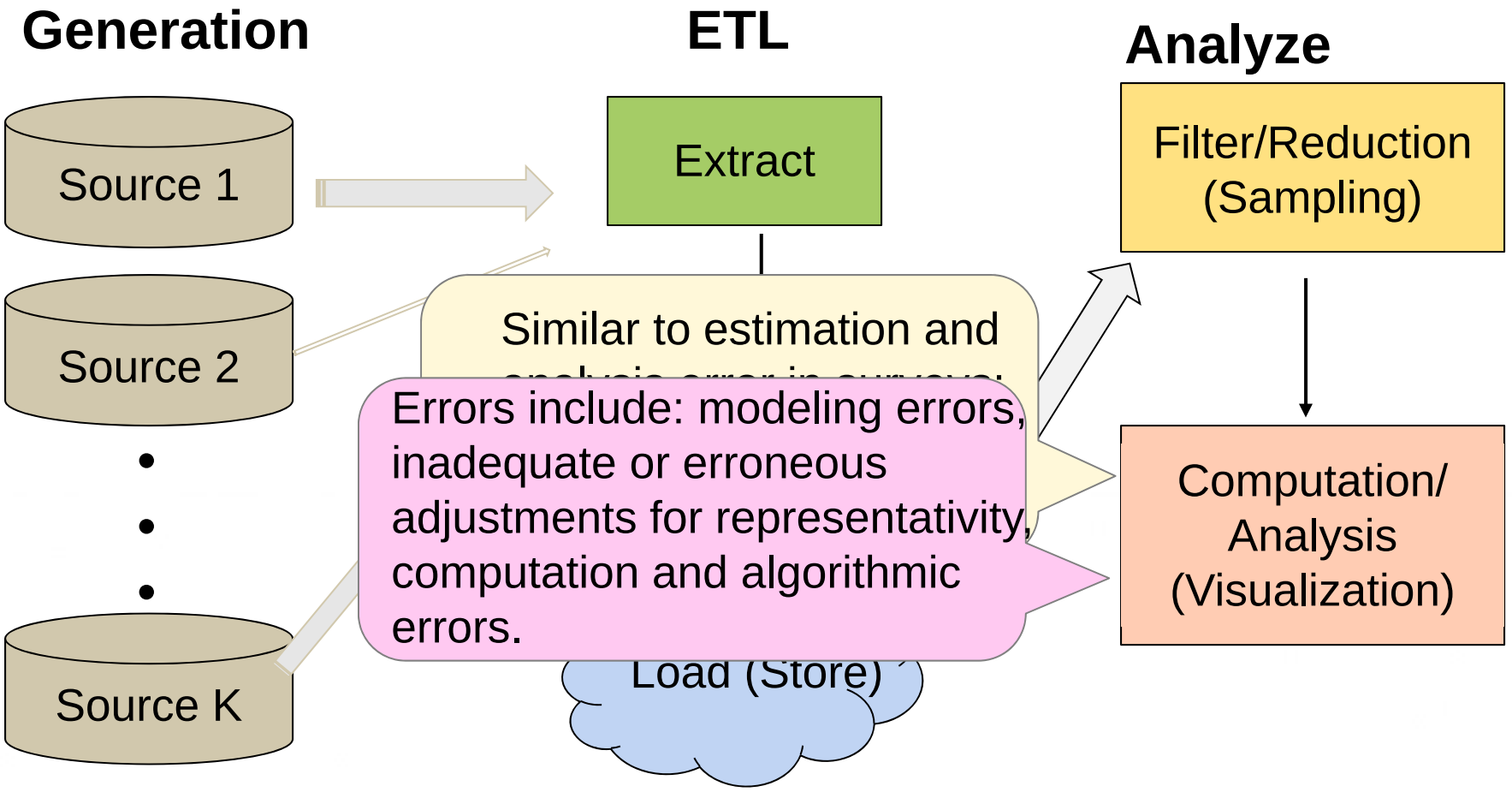
# Big Data Process Map






RTI INTERNATIONAL

# Big Data Process Map






RTI INTERNATIONAL

# Other Big Data Analysis Errors

Fan, Han, and Liu (2014) show that high dimensionality leads to three analysis issues :

a. noise accumulation – inability to identify correlates
b. spurious correlations - false discoveries
c. incidental endogeneity – Cov(error, covariates)

- These issues are a concern even if the data could be regarded as error-free.
- Data errors can considerably exacerbate these problems.
- Current research is aimed at demonstrating this.

# Summary

- Big data can be extremely complex and subject to selectivity bias, missingness and content errors
- Errors that apply to surveys can also apply to Big Data, including sampling
- Traditional approaches for describing errors in data bases may be too simplistic
- Distributed and unstructured data bases processed by Map/Reduce approaches create new opportunities for errors that may vary across applications
- A taxonomy with standardized definitions for these errors is needed