Large-scale Data Linkage from Multiple Sources: Methodology and Research Challenges

John M. Abowd
Associate Director for Research and Methodology and Chief Scientist,
U.S. Census Bureau

Based on NBER Summer Institute Methods Lecture
Originally delivered July 27, 2017

Acknowledgments and Disclaimer

- I owe a huge debt to many of my collaborators over the years, in particular Lars Vilhuber and Bill Winkler
- Parts of this talk were supported by the National Science Foundation, the Sloan Foundation, and the Census Bureau (before and after my appointment started)
- The opinions expressed in this talk are my own and not necessarily those of the Census Bureau or other research sponsors

Outline

- Motivation
- Classical Fellegi-Sunter record linkage
- Types of classical record linkages
- Record linkage errors
- Fellegi-Sunter extension for multiple files
- Bayesian methods and virtual populations
- Classical analysis of effects of linkage errors on statistical models
- Bayesian extensions for linkage error analysis
- Some food for thought from the Census Longitudinal Infrastructure Project (CLIP) data
- Critical take-aways

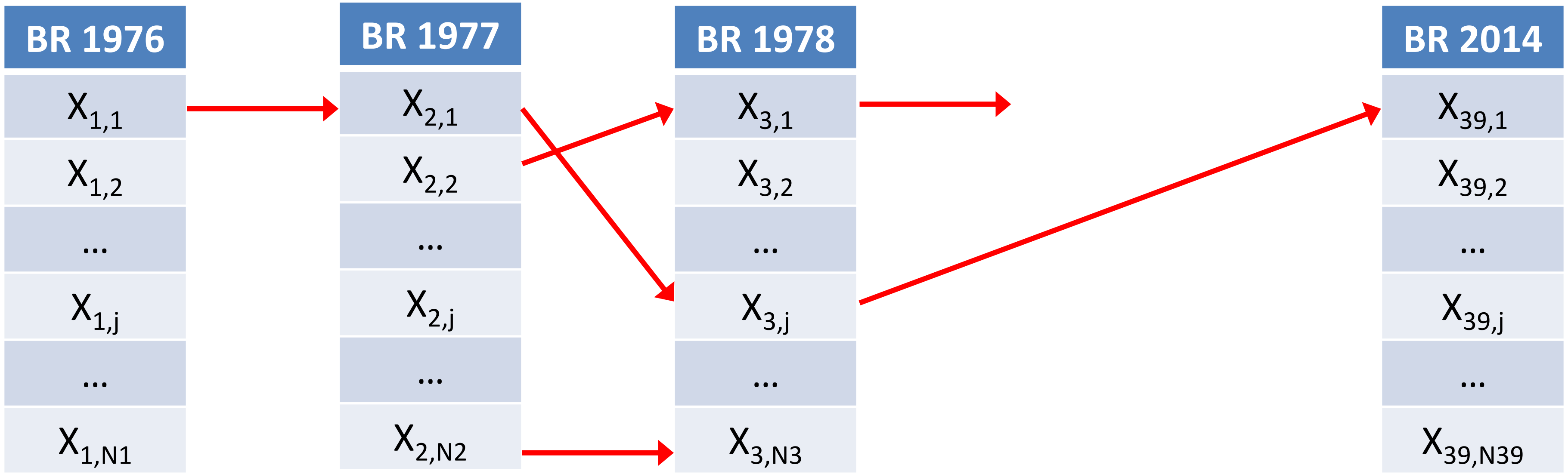
Motivation

- Examples from the Census Bureau and BLS large-scale linkage projects
  - Longitudinal Business Database
  - Longitudinal ES 202 Data
  - Longitudinal Employer-Household Dynamics Infrastructure Files
  - Census Longitudinal Infrastructure Project

Types of Record Linkage

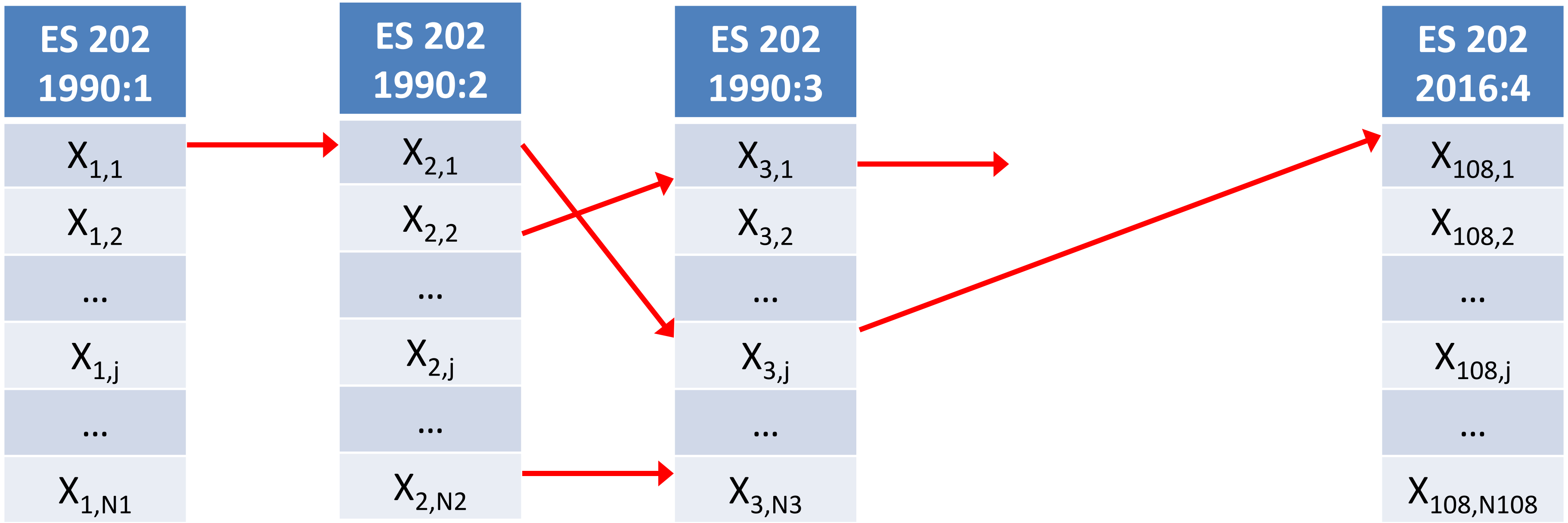
- Deterministic (also called exact)
  - Edited, unique identifiers available on all files
  - Examples: Social Security Numbers (after validation), Employer Identification Numbers (after validation)
- Model-based
  - Comparison variables available on all files (list may be incomplete)
  - Examples: Fellegi-Sunter probabilistic record linkage, distance-based record linkage, posterior predictive models

Longitudinal Business Database

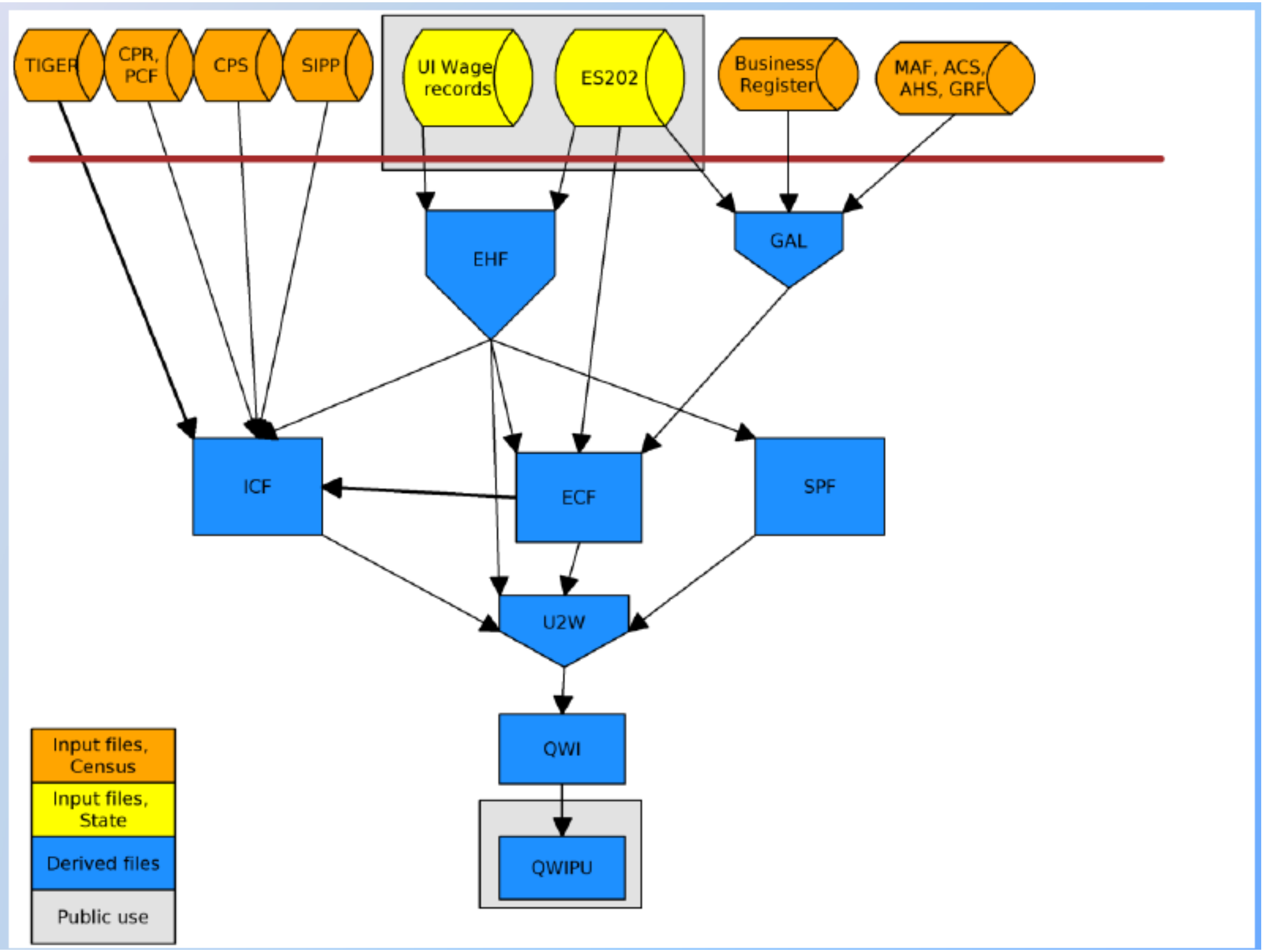


Adapted from: Jarmin and Miranda (2002)

Longitudinal ES 202 from the Bureau of Labor Statistics



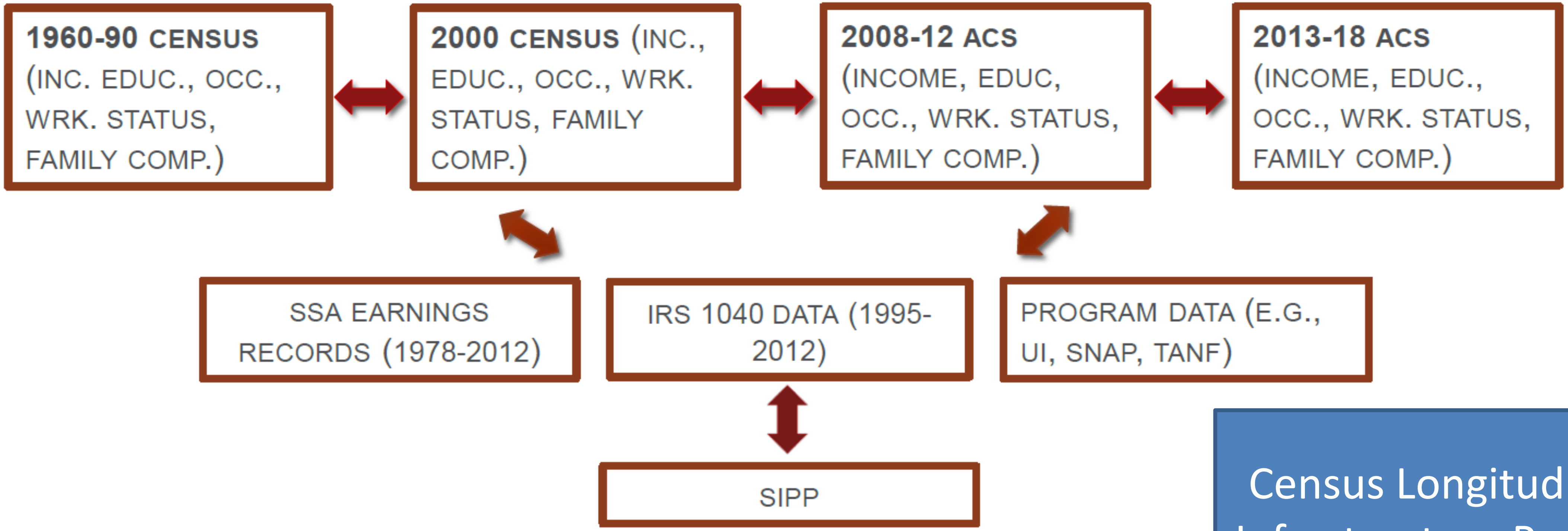
Adapted from: Robertson et al. 1997



The Longitudinal Employer-Household Dynamics Infrastructure File System

Source: Abowd and Vilhuber (2016)

STEP #4: SLIPPING IN THE SURVEY



Census Longitudinal Infrastructure Project

SURVEYS WITH IDENTIFIERS CAN BE SLIPPED IN (E.G., SIPP)

THE SURVEY NOW AS A LEAN AND MEAN VALUE-ADDED INSTRUMENT

STANFORD CENTER ON POVERTY AND INEQUALITY

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Source: Alexander et al. (2016)

Classical Fellegi-Sunter Record Linkage

- Based on Fellegi-Sunter (1969)
- Widely implemented in national statistical agencies
- Used for
  - Deduplication (unduplication, for English majors)
  - Frame management
  - Coverage estimation
- Many refinements, well summarized in Herzog, Scheuren and Winkler (2007)
- Excellent computer science review in Christen and Goiser (2007)

Fellegi-Sunter Record Linkage

$$A\colon\ N_A \times (K + K'_A)$$

$$B\colon\ N_B \times (K + K'_B)$$

$$A \otimes B\colon N_A N_B \times (K + K'_A) + (K + K'_B)$$

$$a_i \in A, b_j \in B, ab_r \in A \otimes B$$

Matches: $M \subset A \otimes B$

Non-matches: $U \subseteq A \otimes B - M$

Adapted from: Christen and Goiser (2007) and Herzog, Scheuren and Winkler (2007)

Fellegi-Sunter Record Linkage II

Comparator functions: $\gamma_{ij}^{(k)} \equiv \mathbf{1}^{(k)}(a_{ik} \approx b_{jk}), k = 1, \ldots K; \;\; \gamma_{ij} \in \Gamma$

$$\Gamma: 2^K \times K$$

$$R \equiv \frac{Pr[\gamma_r | ab_r \in M]}{Pr[\gamma_r | ab_r \in U]}$$

$$R^* \equiv \frac{Pr[\gamma_r^1 | ab_r \in M] \ldots Pr[\gamma_r^K | ab_r \in M]}{Pr[\gamma_r^1 | ab_r \in U] \ldots Pr[\gamma_r^K | ab_r \in U]}$$

$$w_r = log(R^*)$$

Adapted from: Christen and Goiser (2007) and Herzog, Scheuren and Winkler (2007)

Fellegi-Sunter Record Linkage III

Classifier Match: $\widetilde{M} \equiv \{w_r \geq T\}$

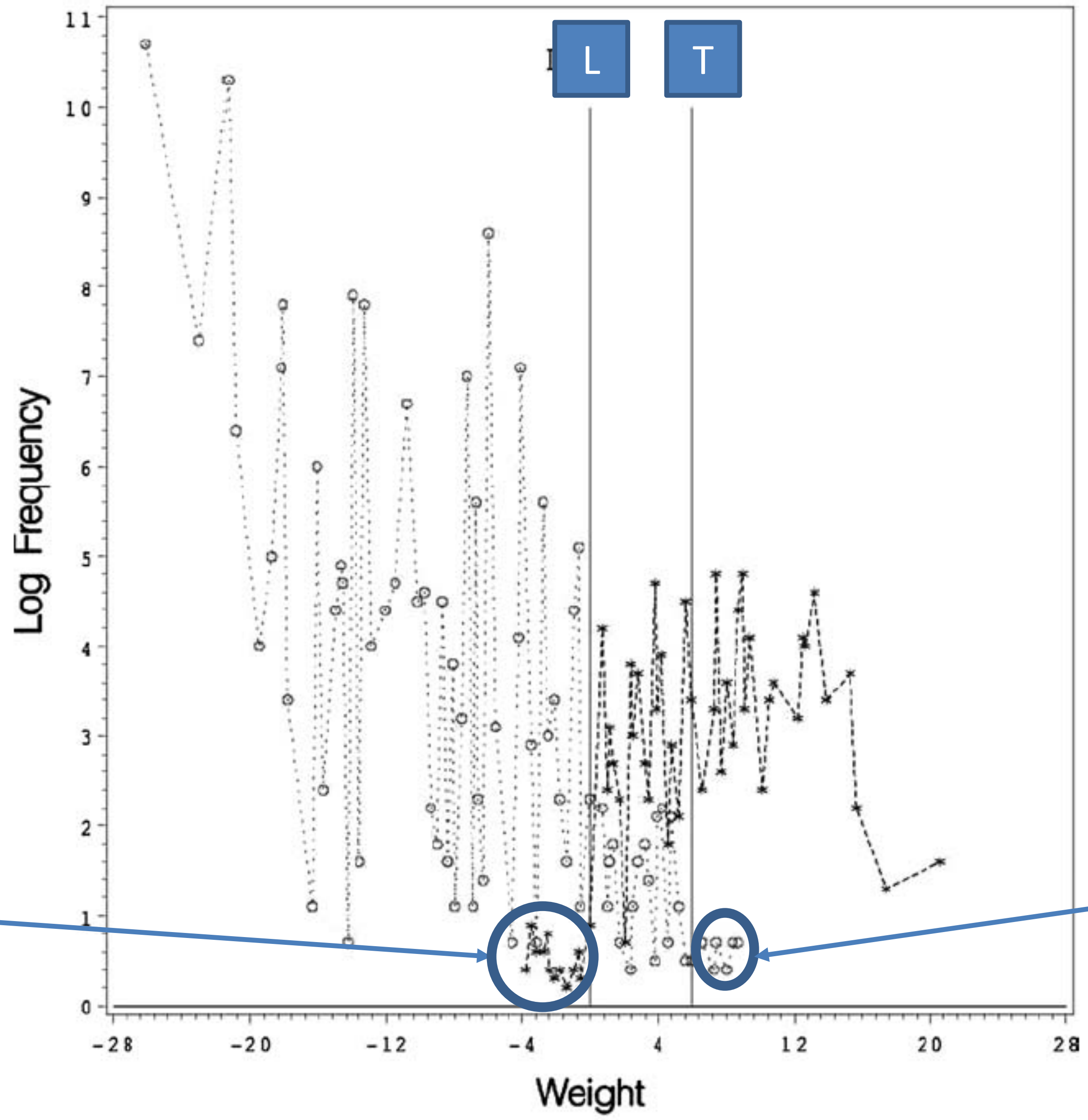
Classifier Non-match: $\widetilde{U} \equiv \{w_r \leq L\}$

Not classified (Clerical resolution): $L < w_r < T$

False Match Rate: $\mu \equiv Pr\left[ab_r \in \widetilde{M} | ab_r \in U\right]$

False Non-match Rate: $\lambda \equiv Pr\left[ab_r \in \widetilde{U} | ab_r \in M\right]$

Adapted from: Christen and Goiser (2007) and Herzog, Scheuren and Winkler (2007)



+ True matches
o = True Non-matches

+ = False Non-matches

o = False Matches

Source: Herzog, Scheuren and Winkler (2007)

Bayesian Record Linkage

$$Pr[ab_r \in M, \gamma_r] = Pr[ab_r \in M|\gamma_r]Pr[\gamma_r] = Pr[\gamma_r|ab_r \in M]Pr[ab_r \in M]$$

$$Pr[ab_r \in M|\gamma_r] = \frac{Pr[\gamma_r|ab_r \in M]Pr[ab_r \in M]}{Pr[\gamma_r]} = 1 - Pr[ab_r \in U|\gamma_r]$$

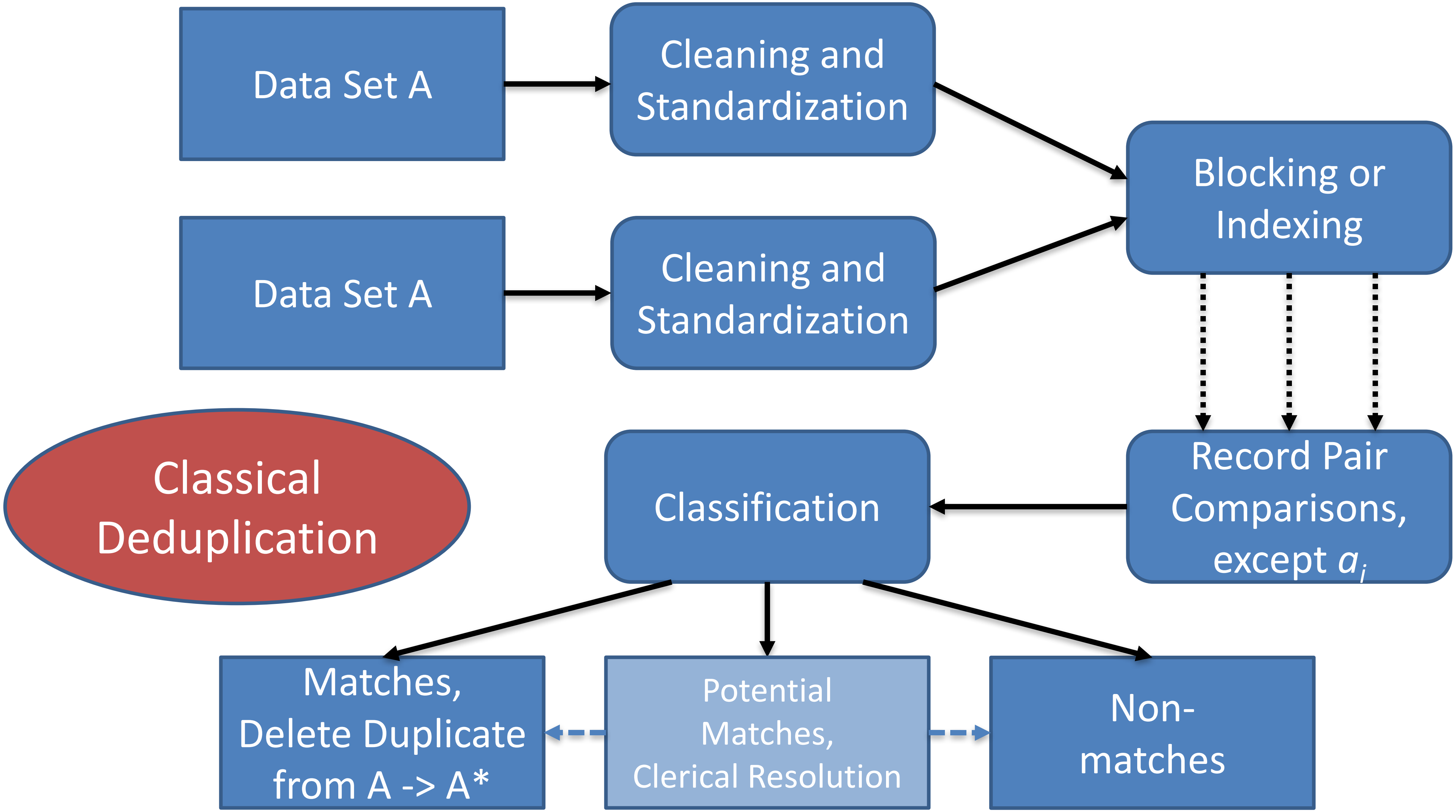
$$Pr[ab_r \in U|\gamma_r] = \frac{Pr[\gamma_r|ab_r \in U]Pr[ab_r \in U]}{Pr[\gamma_r]}$$

Classifier Match: $\widetilde{M} \equiv \left\{\widehat{Pr}[ab_r \in M|\gamma_r] \geq \widehat{Pr}[ab_r \in U|\gamma_r]\right\}$

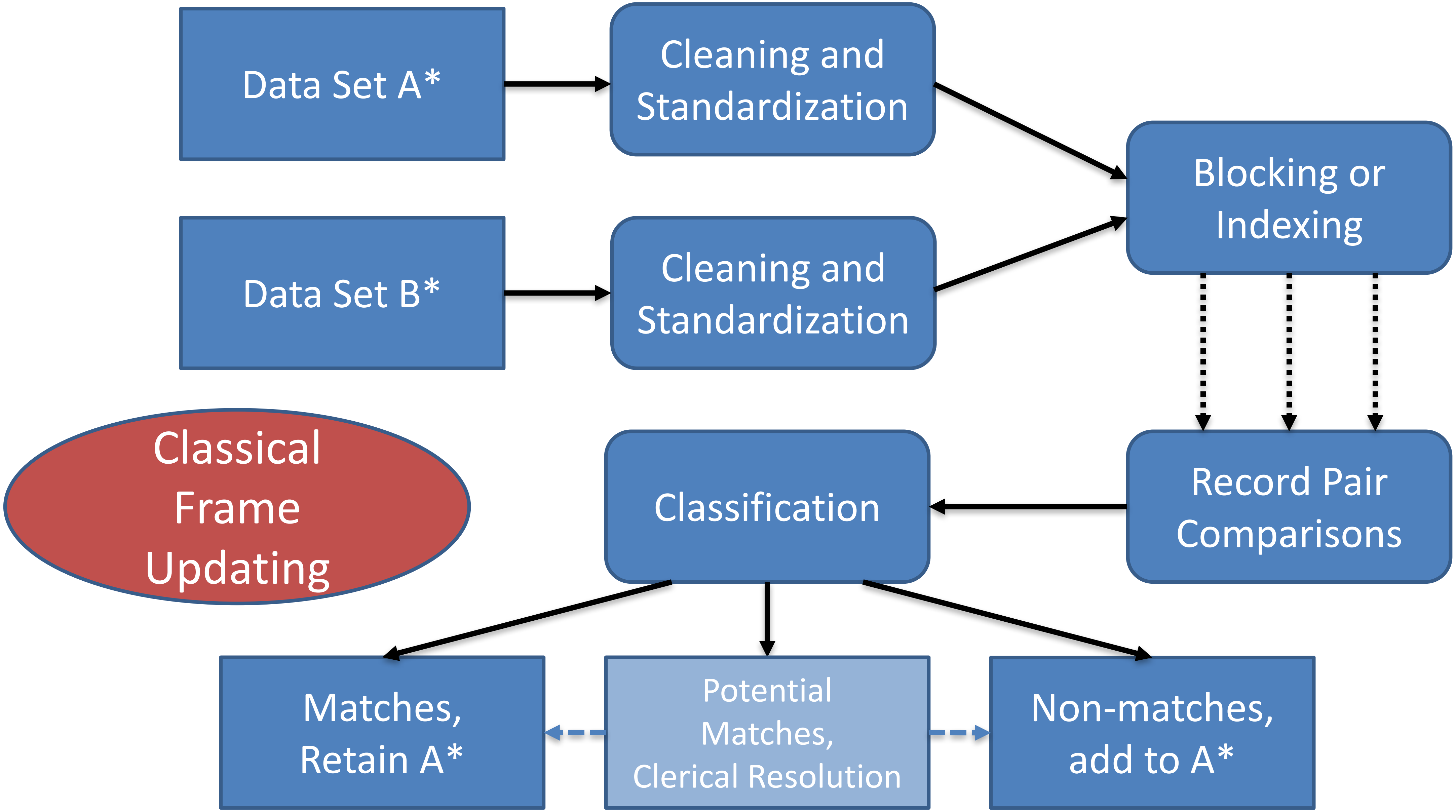
Classifier Non-match: $\widetilde{U} \equiv \left\{\widehat{Pr}[ab_r \in M|\gamma_r] < \widehat{Pr}[ab_r \in U|\gamma_r]\right\}$

Types of Classical Record Linkages

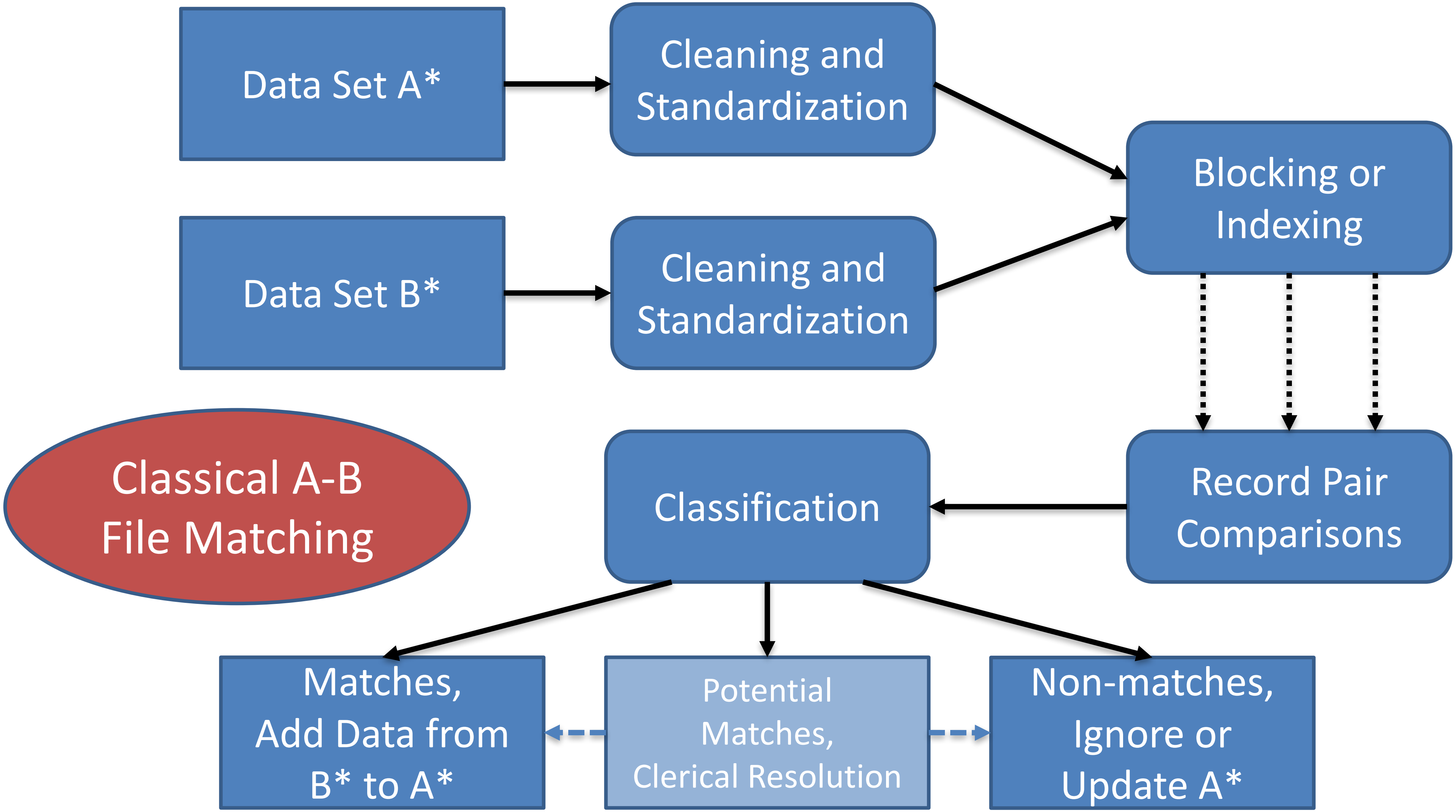
- Deduplication
- Frame updating
- A-B file matching
- Pairwise multiple file matching
- Problems with pairwise multiple file matching



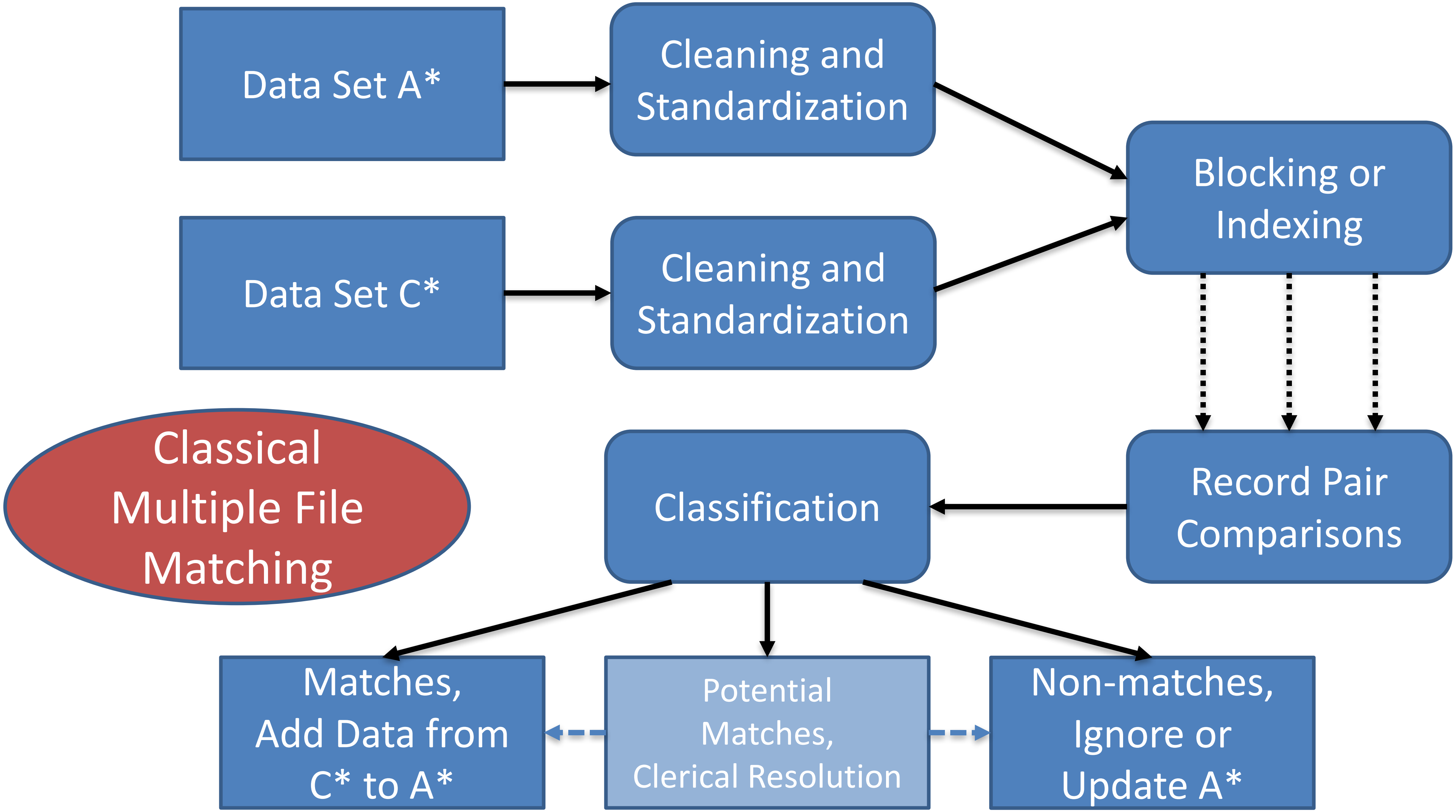
Adapted from: Christen and Goiser (2007)



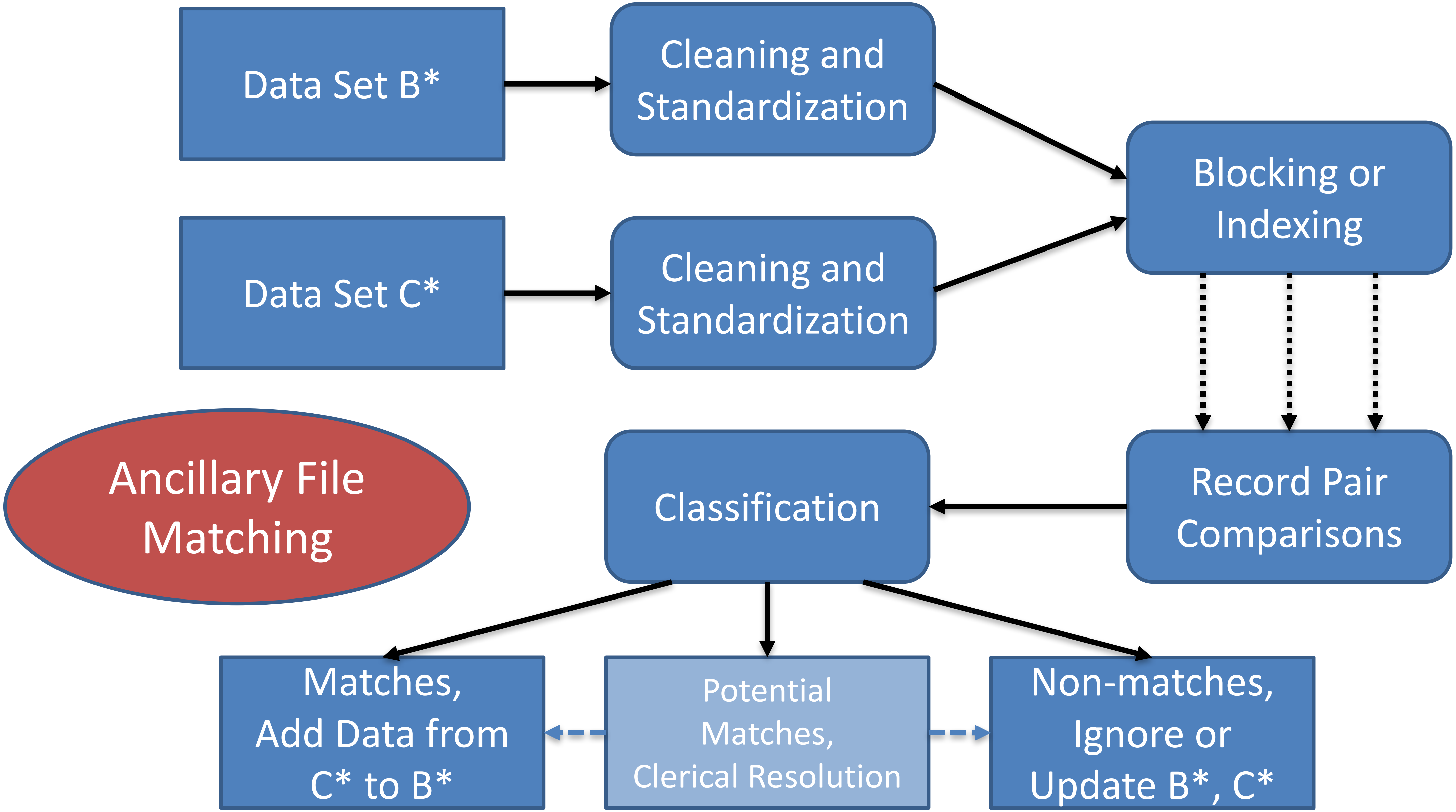
Based on: Christen and Goiser (2007)



Adapted from: Christen and Goiser (2007)



Adapted from: Christen and Goiser (2007)



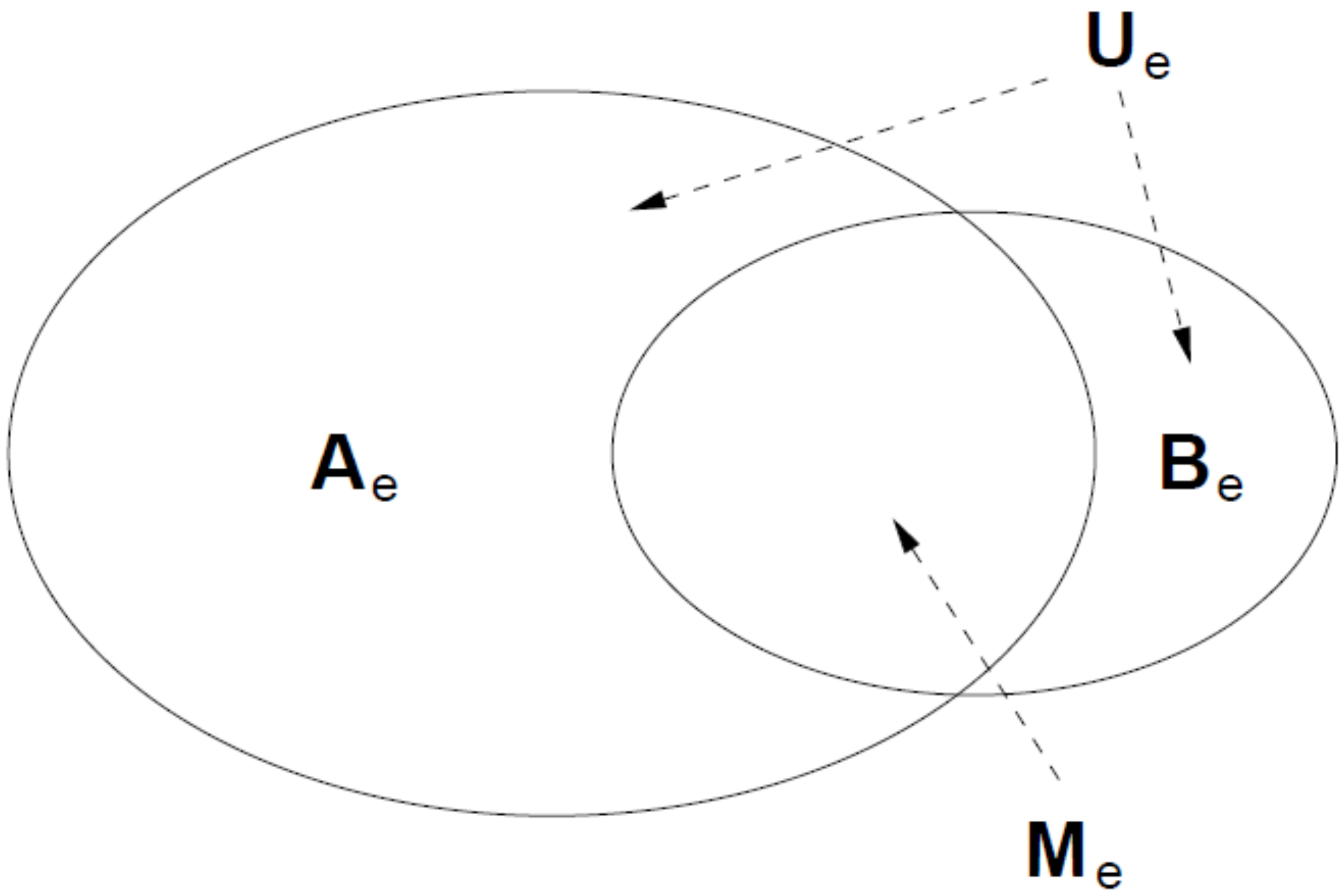
Adapted from: Christen and Goiser (2007)

Challenges with Multiple Files

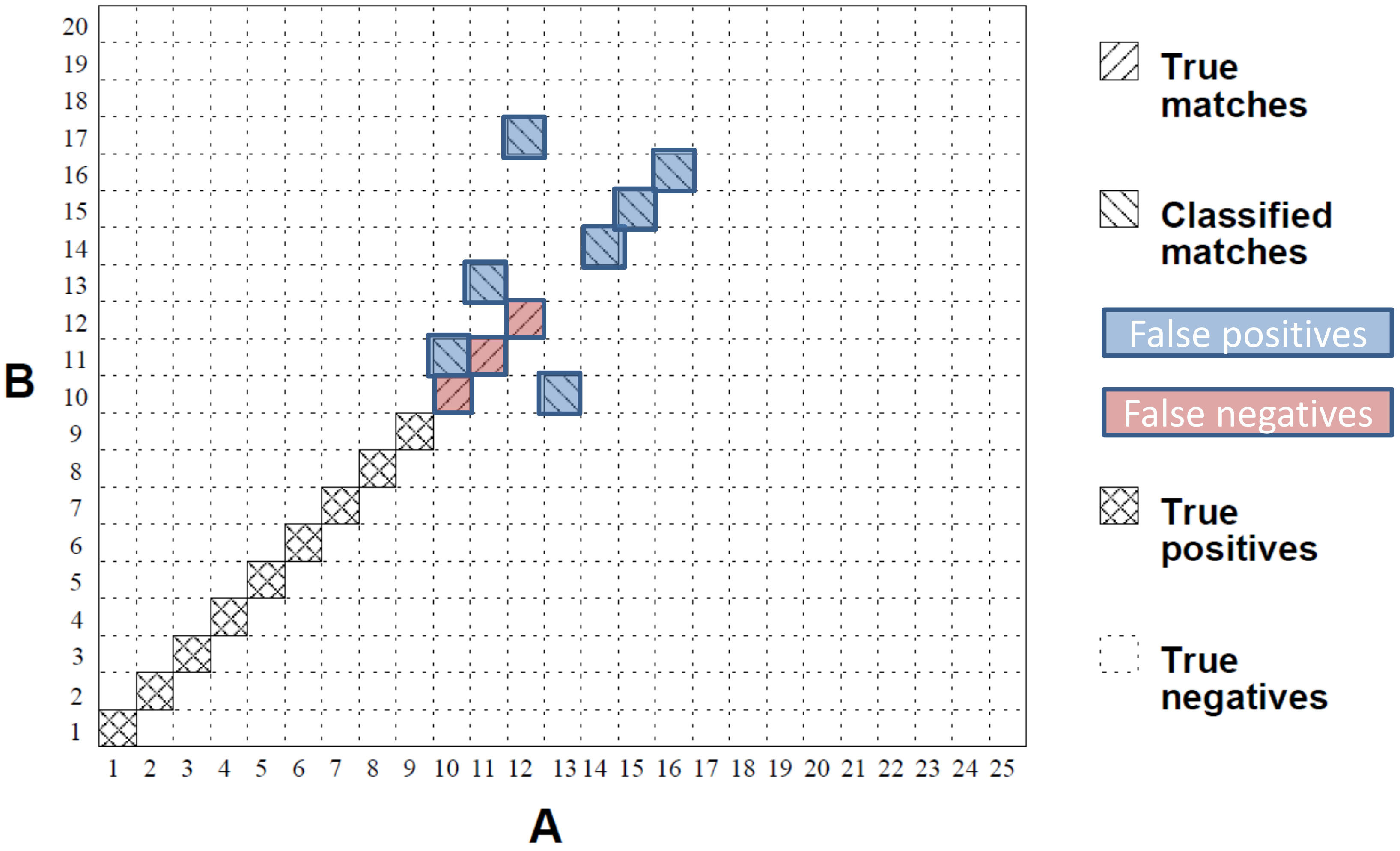
- How should we treat the situation where the multi-file linkage relation is non-transitive?
  - Record a_1 links to b_1.
  - Record b_1 links to c_1.
  - But a_1 does not link to c_1.
- Happens frequently in both business and household data
- Bayesian methods discussed below can handle either case
- Important to specify the outcome set correctly

Record Linkage Errors

- Entity space
- Comparison space
- Suggested error rate measures
- Example



Source: Christen and Goiser (2007)



Source: Christen and Goiser (2007)

Table 1. Confusion matrix of record pair classification		
Actual	Classification	
	Match ($\widetilde{M}$)	Non-match ($\widetilde{U}$)
Match (M)	True match	False non-match
	True positive (TP)	False negative (FN)
Non-match (U)	False match	True non-match
	False positive (FP)	True negative (TN)
Note: defined on the comparison space (all pairs).		

Adapted from: Christen and Goiser (2007)

Measure	Definition	Comment
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Dominated by TN
Precision	TP/(TP+FP)	True matches/Classified matches
Recall (Sensitivity, TPR)	TP/(TP+FN)	True positive rate
Precision-Recall Breakeven	Precision = TPR	
F-measure	2(Prec x Rec)/(Prec + Rec)	Compromise between precision and recall
Specificity (TNR)	TN/(TN+FP)	True negative rate, dominated by TN
False Positive Rate (FPR)	FP/(TN+FP)	= 1 – TNR, also dominated by TN
False Discovery Rate (FDR)	FP/(TP+FP)	= 1 – Precision (preferred to FPR)
ROC Curve	FPR (x-axis) v. TPR (y-axis)	Too optimistic

Adapted from: Christen and Goiser (2007)

Table 2. Quality results for the given example		
Measure	Entity space	Comparison space
Accuracy (TP+TN)/(TP+TN+FP+FN)	94.340%	99.999994%
Precision TP/(TP+FP)	72.222%	72.222%
Recall (True positive rate) TP/(TP+FN)	92.857%	92.857%
F-measure 2(Prec x Rec)/(Prec + Rec)	81.250%	81.250%
Specificity TN/(TN+FP)	94.565%	99.999995%
False positive rate FP/(TN+FP)	5.435%	0.000005%
False discovery rate FP/(TP+FP)	27.778%	27.778%

Source: Christen and Goiser (2007)

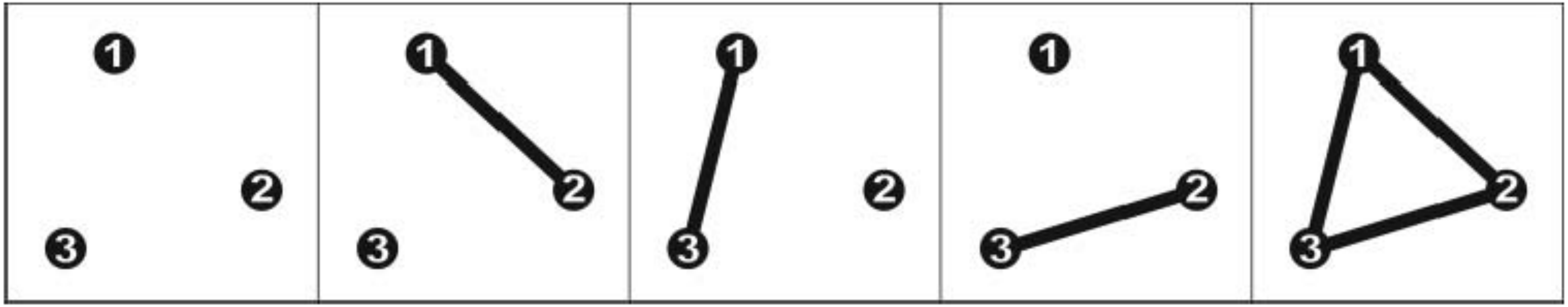
Fellegi-Sunter Extension for Multiple Files

- Based on Sadinle and Fienberg (2013)
- Principled way to extend Fellegi-Sunter to multiple files

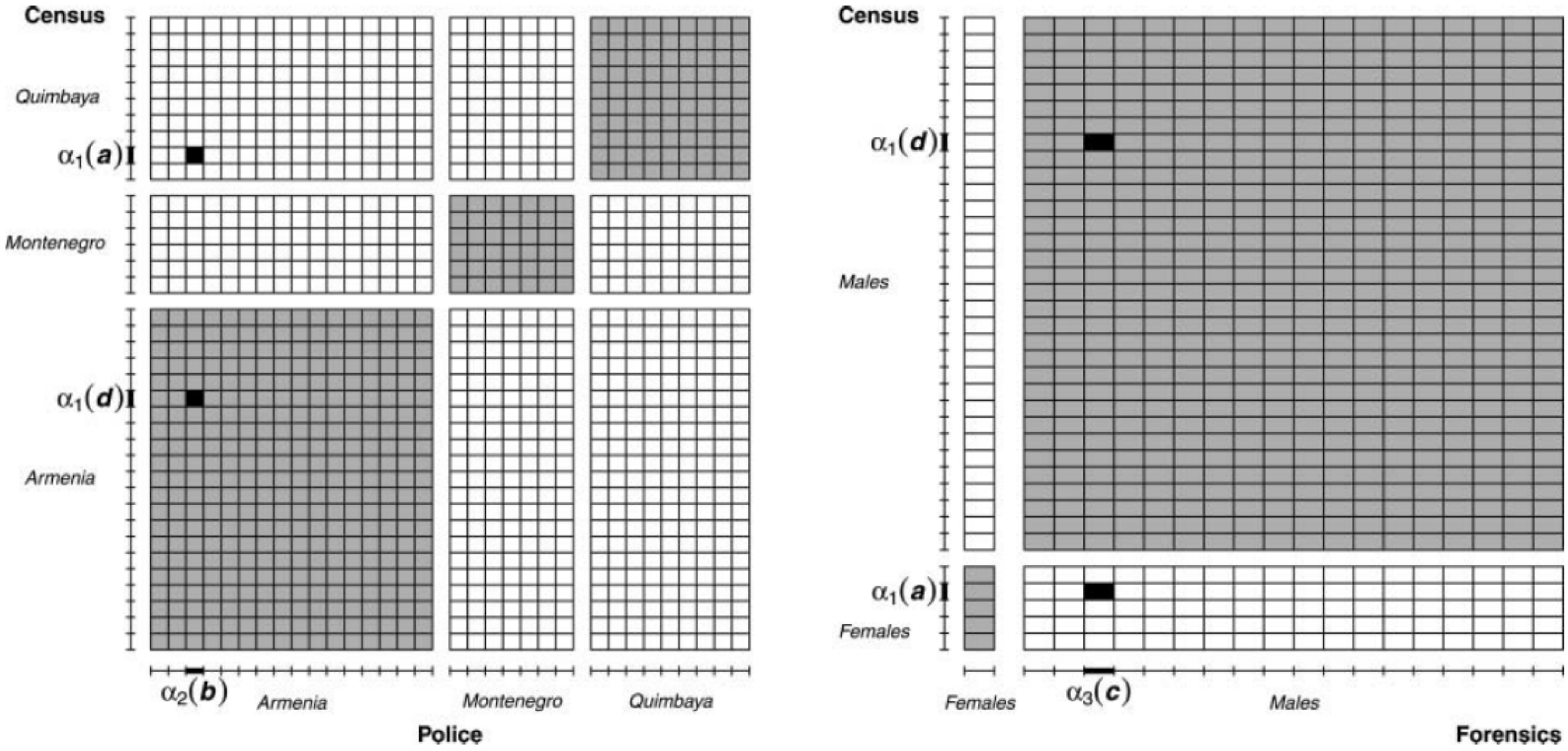
Table 1. Each matching pattern of a record triplet can be associated with a partition of the set $\{1, 2, 3\}$

Notation	$\mathbb{P}_3$	$(\alpha_1(a_1), \alpha_2(a_2), \alpha_3(a_3))$
1/2/3	$\{\{1\}, \{2\}, \{3\}\}$	$a_1 \neq a_2 \neq a_3 \neq a_1$
12/3	$\{\{1, 2\}, \{3\}\}$	$a_1 = a_2; a_3 \neq a_1, a_2$
13/2	$\{\{1, 3\}, \{2\}\}$	$a_1 = a_3; a_2 \neq a_1, a_3$
1/23	$\{\{1\}, \{2, 3\}\}$	$a_2 = a_3; a_1 \neq a_2, a_3$
123	$\{\{1, 2, 3\}\}$	$a_1 = a_2 = a_3$

Source: Sadinle and Fienberg (2013)



Source: Sadinle and Fienberg (2013)



Source: Sadinle and Fienberg (2013)

Implementation of Sadinle and Fienberg

- Works very much like Fellegi-Sunter
- Classifier chooses the predicted match pattern for each K-tuple of records (one from each file) using K agreement indices and controlling the error rate versus unclassified for each one
- Won’t dwell on these methods, instead pass directly to the Bayesian case

Bayesian Methods and Virtual Populations

- Key insight is that the population consists of *J* virtual entities, with *J* unknown
- Specify and estimate the linkage structure, which specifies a posterior probability for each record being assigned to any of the *J* virtual entities
- Allows for errors in measurement of all classifying variables
- Implemented via Markov Chain Monte Carlo
- Full posterior distribution can be used for error assessment

Bayesian Multiple File Linkage

Files: $A_i, i = 1, \ldots, K$

Data in file i: x_{ij} $1 \times M, i = 1, \ldots, K; j = 1, \ldots, N_i; \ell$=1,…, M

Data distortion indicator: $z_{ij\ell} = \begin{cases} 1, \text{if } x_{ij\ell} \text{ is distorted} \\ 0, \text{otherwise} \end{cases}$

Size of latent population: $J = 1, \ldots, \sum N_i$

Linkage structure: $\lambda_{ij} = 1, \ldots, J$

Latent data: y_j $1 \times M, j = 1, \ldots, J$

Source: Steorts, Hall and Fienberg (2016)

Bayesian Multiple File Linkage II

Posterior predictive distribution: $Pr[\Lambda, Y, Z|X]$

$$\widehat{Pr}[\lambda_{ij} = \lambda_{i'j'}|X] = \frac{1}{S}\sum_{h=1}^{S} \mathbf{1}\left(\lambda_{ij}^{(h)} = \lambda_{i'j'}^{(h)}\right)$$

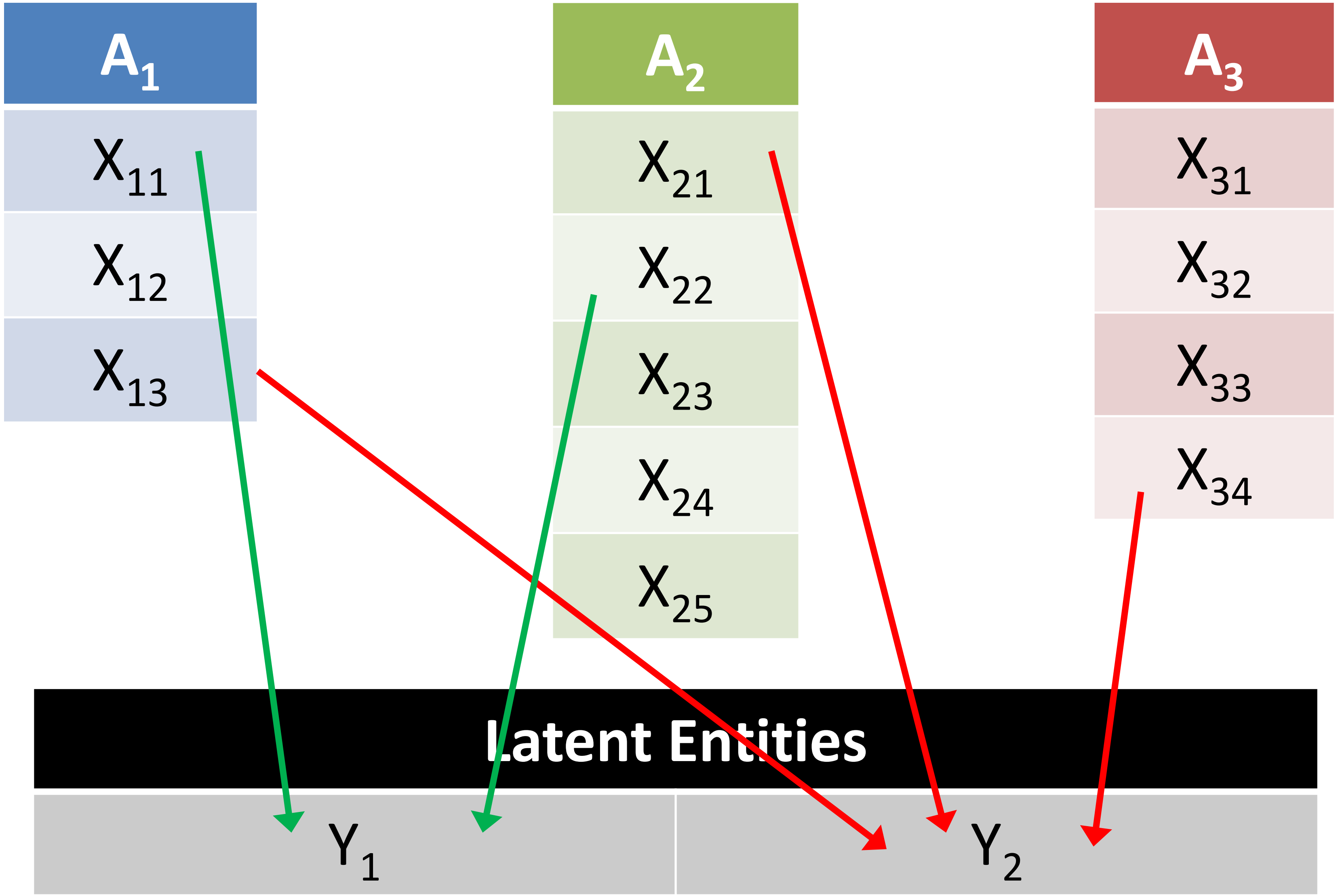
Sets of Records: $\mathcal{A} \equiv \{(i,j)|i \in \{1,..,K\}, j \in \{1,..,N_i\}\}$

Maximal Matching Set (MMS): $\Omega(\mathcal{A},\Lambda) = \sum_{j'}\left(\prod_{(i,j)\in\mathcal{A}} \mathbf{1}(\lambda_{ij} = j') \prod_{(i,j)\notin\mathcal{A}} \mathbf{1}(\lambda_{ij} \neq j')\right)$

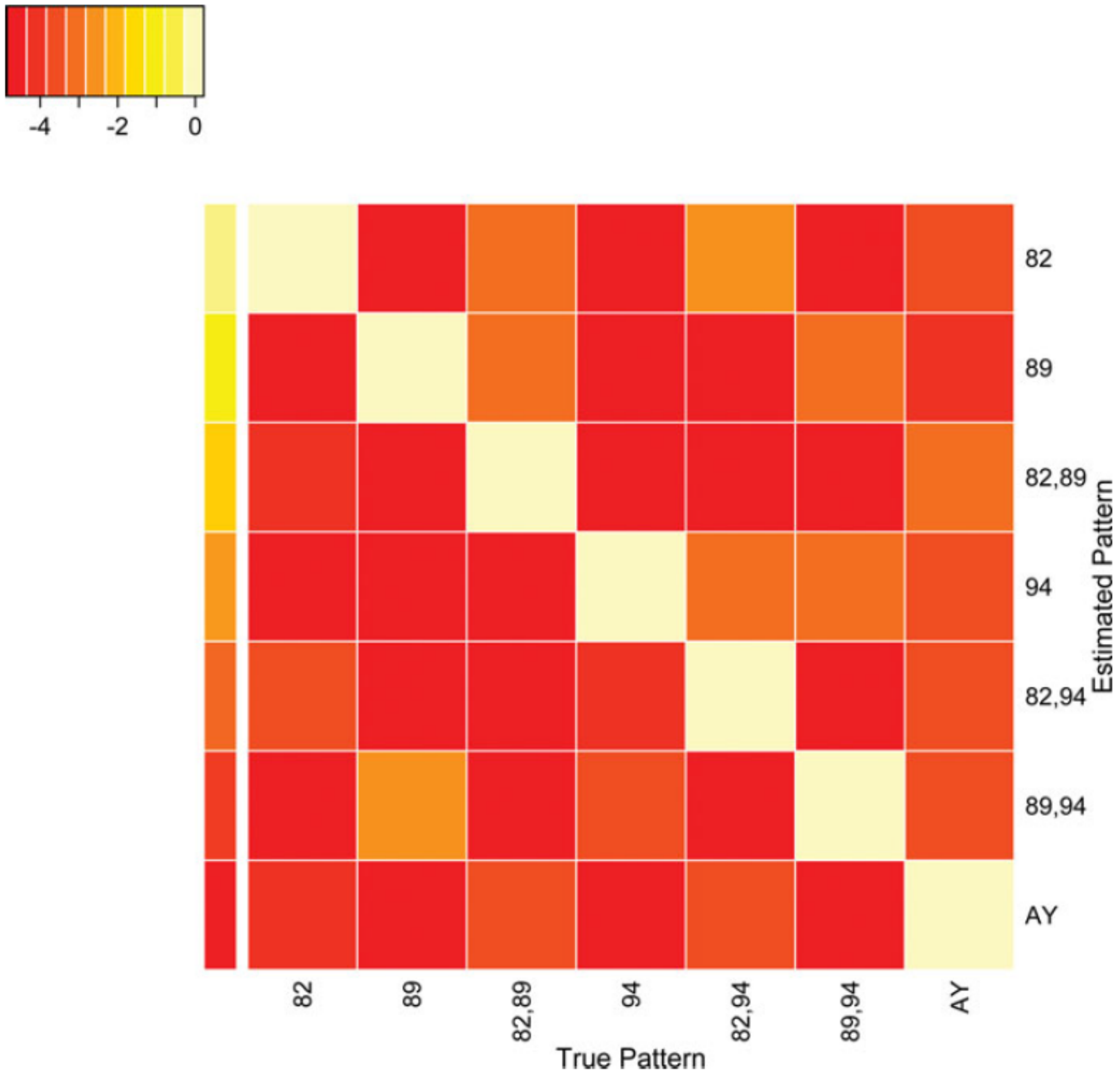
Most Probable MMS (MPMMS): $\mathcal{M}_{i,j} = argmax_{\mathcal{A}:(i,j)\in\mathcal{A}} Pr[\Omega(\mathcal{A},\Lambda) = 1|X]$

Shared MPMMS: $\{(i,j)|\forall(i,j),(i',j'): \mathcal{M}_{i,j} = \mathcal{M}_{i',j'}\}$

Source: Steorts, Hall and Fienberg (2016)



Adapted from: Steorts, Hall and Fienberg (2016, appendix)



Source: Steorts, Hall and Fienberg (2016)

Classical Analysis of the Effects of Linkage Errors on Statistical Models

- Linkage errors due to positive false match rate
- Linkage errors due to positive false non-match rate
- Frame errors due to faulty correspondence between the linked data and the conceptual frame
- Specification errors due to compromises in the implementation of the linkage model

Positive False Match Rate

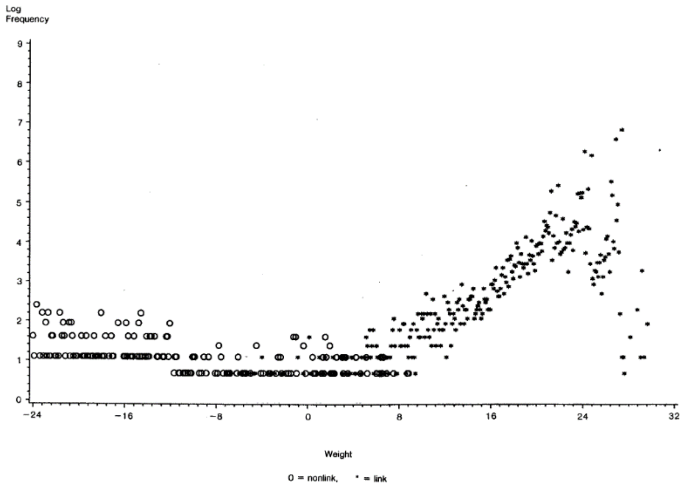


Figure 2. Log of Frequency vs. Weight Good Matching Scenario, Links and Nonlinks

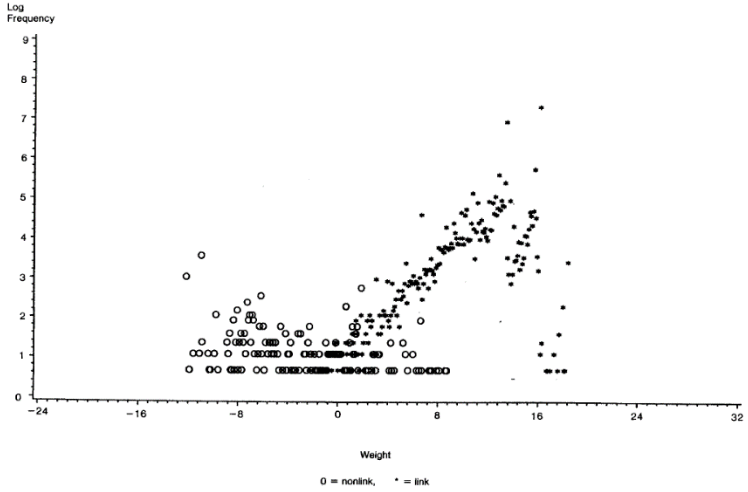


Figure 3. Log of Frequency vs. Weight Mediocre Matching Scenario, Links and Nonlinks

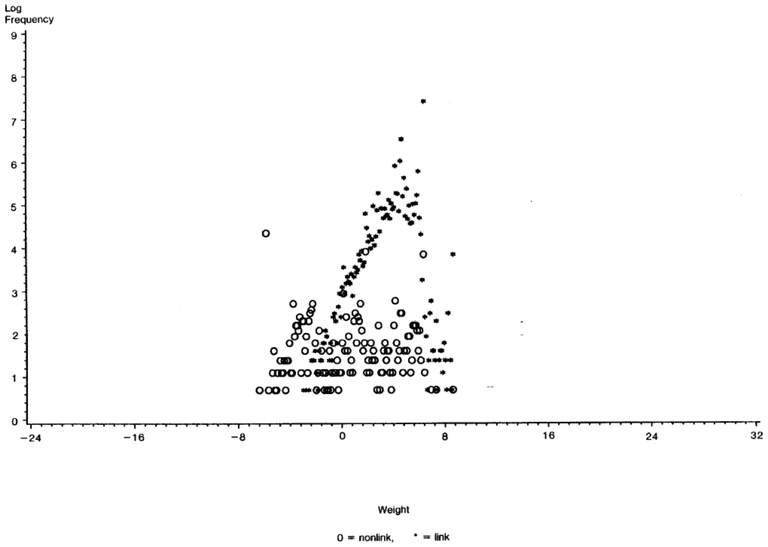


Figure 4. Log Frequency vs. Weight Poor Matching Scenario, Links and Nonlinks

Source: Scheuren and Winkler (1993)

Table 1

Counts of True Links and True Nonlinks and Probabilities of an Erroneous Link in Weight Ranges for Various Matching Cases; Estimated Probabilities via Rubin-Belin Methodology

Weight	False match rates: Good, True Link	Good, True NL	Good, Prob True	Good, Prob Est	Mediocre, True Link	Mediocre, True NL	Mediocre, Prob True	Mediocre, Prob Est	Poor, True Link	Poor, True NL	Poor, Prob True	Poor, Prob Est
15 +	9,176	0	.00	.00	2,621	0	.00	.00	0	1	.00	.00
14	111	0	.00	.00	418	0	.00	.00	0	1	.00	.00
13	91	0	.00	.01	1,877	0	.00	.00	0	1	.00	.00
12	69	0	.00	.02	1,202	0	.00	.00	0	1	.00	.00
11	59	0	.00	.03	832	0	.00	.00	0	1	.00	.00
10	69	0	.00	.05	785	0	.00	.00	0	1	.00	.00
9	42	0	.00	.08	610	0	.00	.00	0	1	.00	.00
8	36	2	.05	.13	439	3	.00	.00	65	1	.02	.00
7	30	1	.03	.20	250	4	.00	.01	39	1	.03	.00
6	14	7	.33	.29	265	9	.03	.03	1,859	57	.03	.03
5	28	4	.12	.40	167	8	.05	.06	1,638	56	.03	.03
4	6	3	.33	.51	89	6	.06	.11	2,664	62	.02	.05
3	12	7	.37	.61	84	5	.06	.20	1,334	31	.02	.11
2	8	6	.43	.70	38	7	.16	.31	947	30	.03	.19
1	7	13	.65	.78	33	34	.51	.46	516	114	.18	.25
0	7	4	.36	.83	13	19	.59	.61	258	65	.20	.28
− 1	3	5	.62	.89	7	20	.74	.74	93	23	.20	.31
− 2	0	11	.99	.91	3	11	.79	.84	38	23	.38	.41
− 3	4	6	.60	.94	4	19	.83	.89	15	69	.82	.60
− 4	4	3	.43	.95	0	15	.99	.94	1	70	.99	.70
− 5	4	4	.50	.97	0	15	.99	.96	0	25	.99	.68
− 6	0	5	.99	.98	0	27	.99	.98	0	85	.99	.67
− 7	1	6	.86	.98	0	40	.99	.99			.99	.99
− 8	0	8	.99	.99	0	41		.99			.99	.99
− 9	0	4	.99	.99	0	4		.99			.99	.99
− 10 −	0	22			0	22		.99			.99	.99

Notes: In the first column, weight 10 means weight range from 10 to 11. Weight ranges 15 and above and weight ranges −9 and below are added together. Weights are log ratios that are based on estimated agreement probabilities. NL is nonlinks and **Prob** is probability.

Source: Scheuren and Winkler (1993)

Table 2

Summary of Adjustment Results for Illustrative Simulations

Basis of adjustments	Matching scenarios		
	Good	Mediocre	Poor
True probabilities	Adjustment was not helpful because it was not needed	Good results like those in Section 4.1	Good results like those in Section 4.1
Estimated probabilities	Same as above	Same as above	Poor results because Rubin-Belin could not estimate the probabilities

Source: Scheuren and Winkler (1993)

Table 4. Percent Coverage of 95% Confidence Intervals With and Without Bootstrap Adjustment of Standard Errors

	Coverage before bootstrap	*Coverage after bootstrap*
Simulation Case 1		
Naive	34	34
Robust	50	50
Scheuren–Winkler	59	60
Lahiri–Larsen	83	88
Simulation Case 2		
Naive	4	4
Robust	8	8
Scheuren–Winkler	40	41
Lahiri–Larsen	85	89

Source: Lahiri and Larsen (2005)

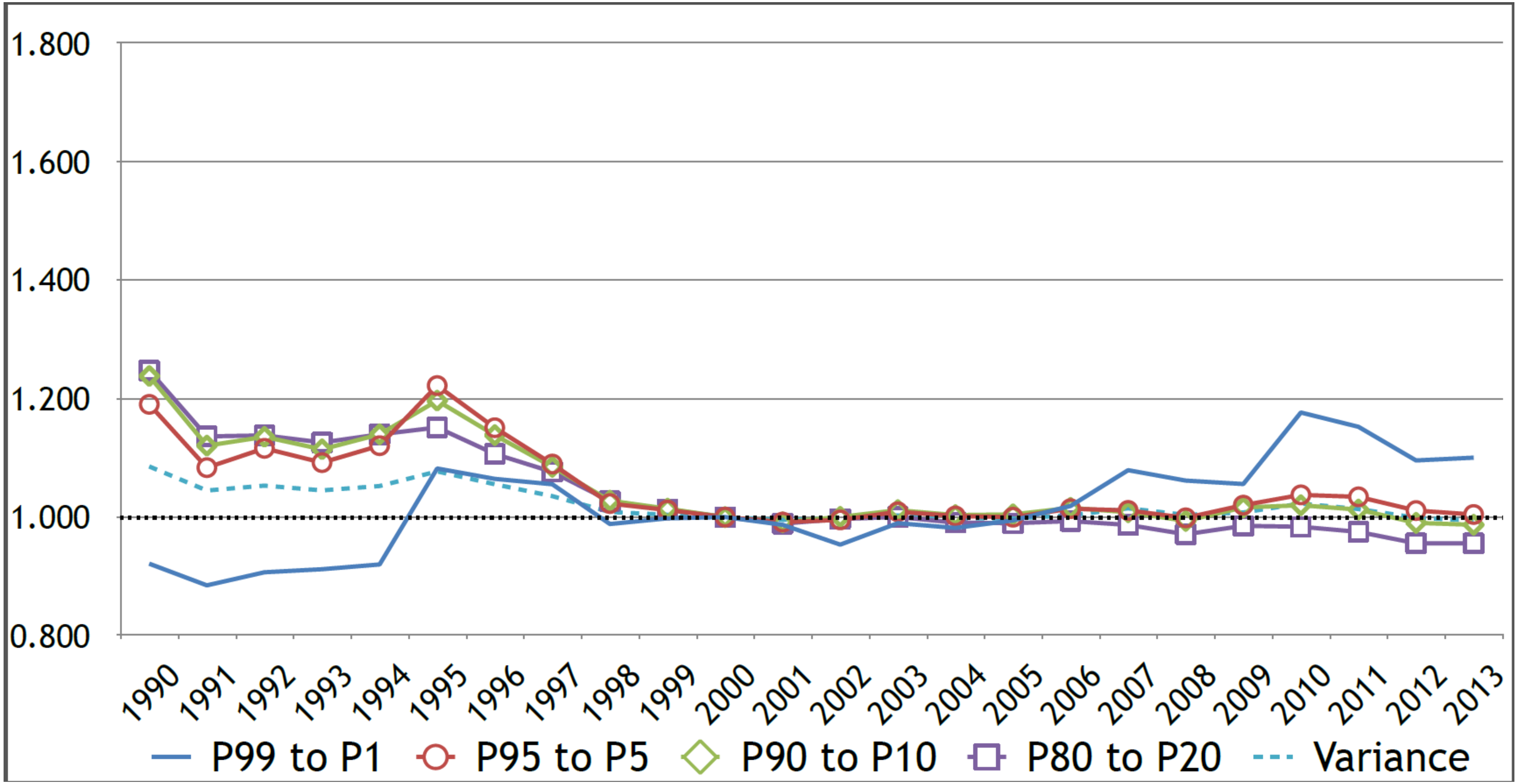
Table 10. False Match Rate and False Non-Match Rate in 2002

Level	Firm Size Group	Observation	Unique SEIN	False Match Rate Lower Bound	False Match Rate Upper Bound
Establishment-level	0 - 19	318	300	0.00 (0.006)	0.133 (0.017)
	20 - 499	713	300	0.0067 (0.006)	0.0867 (0.017)
	500 +	7127	300	0.01 (0.006)	0.0867 (0.017)
Employer-level	0 - 19	300	300	0.0133 (0.006)	0.153 (0.023)
	20 - 499	300	300	0.0033 (0.006)	0.0766 (0.017)
	500 +	301	300	0.0133 (0.006)	0.0800 (0.017)
Level	**Firm Size Group**	**Observation**	**Unique SEIN**	**False Non-Match Rate Lower Bound**	**False Non-Match Rate Upper Bound**
Establishment-level	0 - 19	300	300	0.44 (0.029)	0.667 (0.029)
	20 - 499	410	300	0.603 (0.029)	0.733 (0.029)
	500 +	2006	300	0.733 (0.029)	0.813 (0.029)
Employer-level	0 - 19	300	300	0.457 (0.029)	0.697 (0.029)
	20 - 499	478	300	0.67 (0.029)	0.82 (0.029)
	500 +	2813	300	0.767 (0.029)	0.86 (0.029)

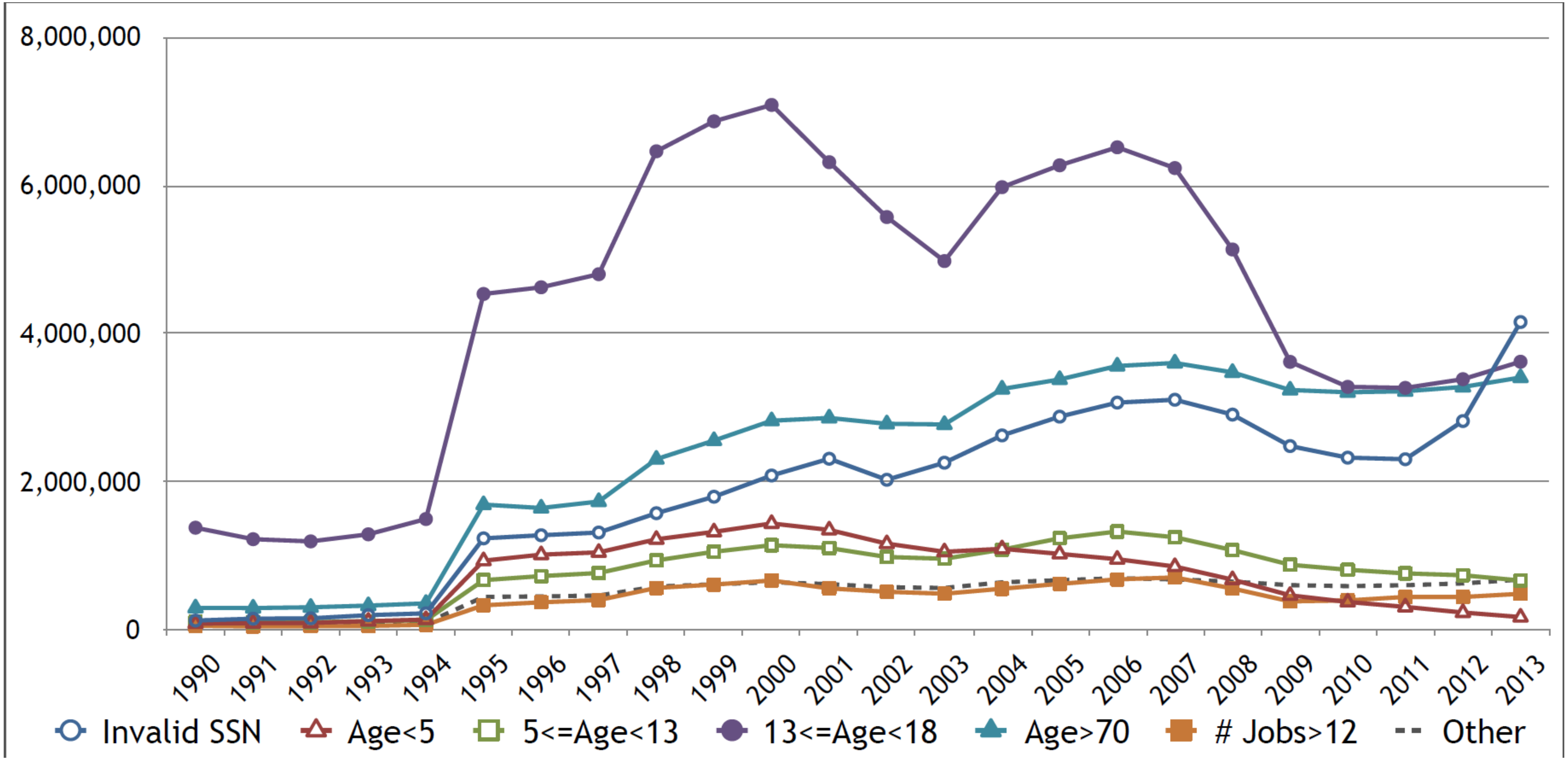
Source: Zheng (2014)

Frame Errors

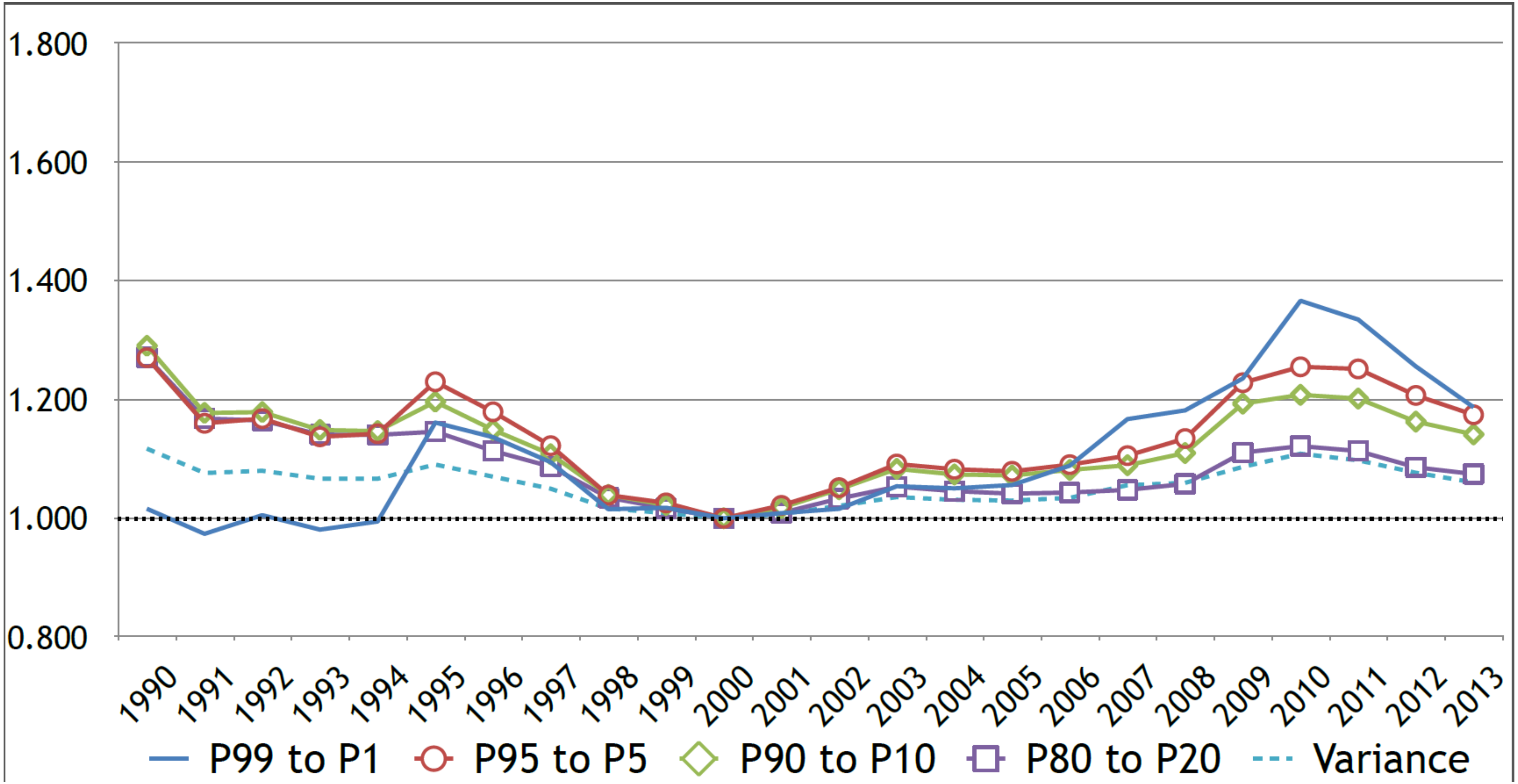
- Example from Abowd, McKinney and Zhao (2018)



Source: Abowd, McKinney and Zhao (2018)



Source: Abowd, McKinney and Zhao (2018)



Source: Abowd, McKinney and Zhao (2018)

Linkage Errors in the Business Employment Dynamics Series

- Had to create an algorithm to link establishments across quarters when the UI account number changed.
- Tested various linkages based upon different blocking variables:
  - Each linkage led to different amounts of job creation and job destruction being placed in the "expansion and contraction" category versus the "openings and closings" category
  - Important because it affected the answer to the policy question of how much job creation was attributable to entrepreneurs, to continuing small firms, or to continuing large firms.
  - BLS reported the results of the various linkages in a technical paper
- Key statistics depend upon "behind-the-scenes" decisions made by statisticians in constructing the data
- Likely that these decisions and their implications are not clearly communicated to the policymakers
- Assessed in Robertson et al. (1997)
- Abowd and Vilhuber (2005) and Benedetto et al. (2007) make similar assessments for statistics from the LEHD infrastructure

Adapted from: Robertson et al. 1997

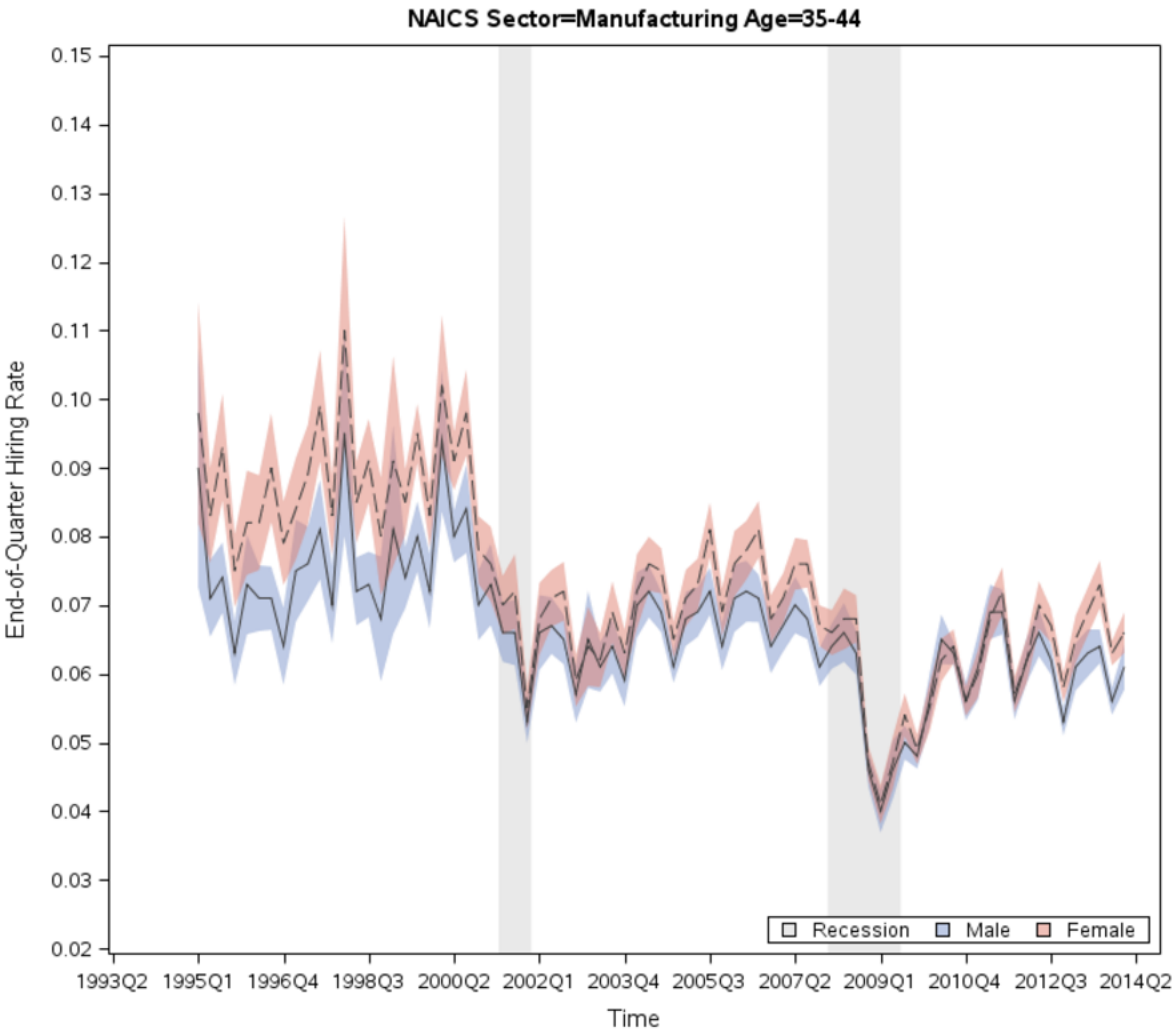
Specification Errors

- Large differences in validation rates by person and housing unit characteristics in the 2009 ACS
- The characteristics of persons the ACS who can be linked to external data sources vary considerably from the full set of ACS persons
- Should consider adjusting survey weights accordingly when conducting analysis
- Changes tested in the PVS process for the 2010 ACS validation attenuate the bias by characteristics

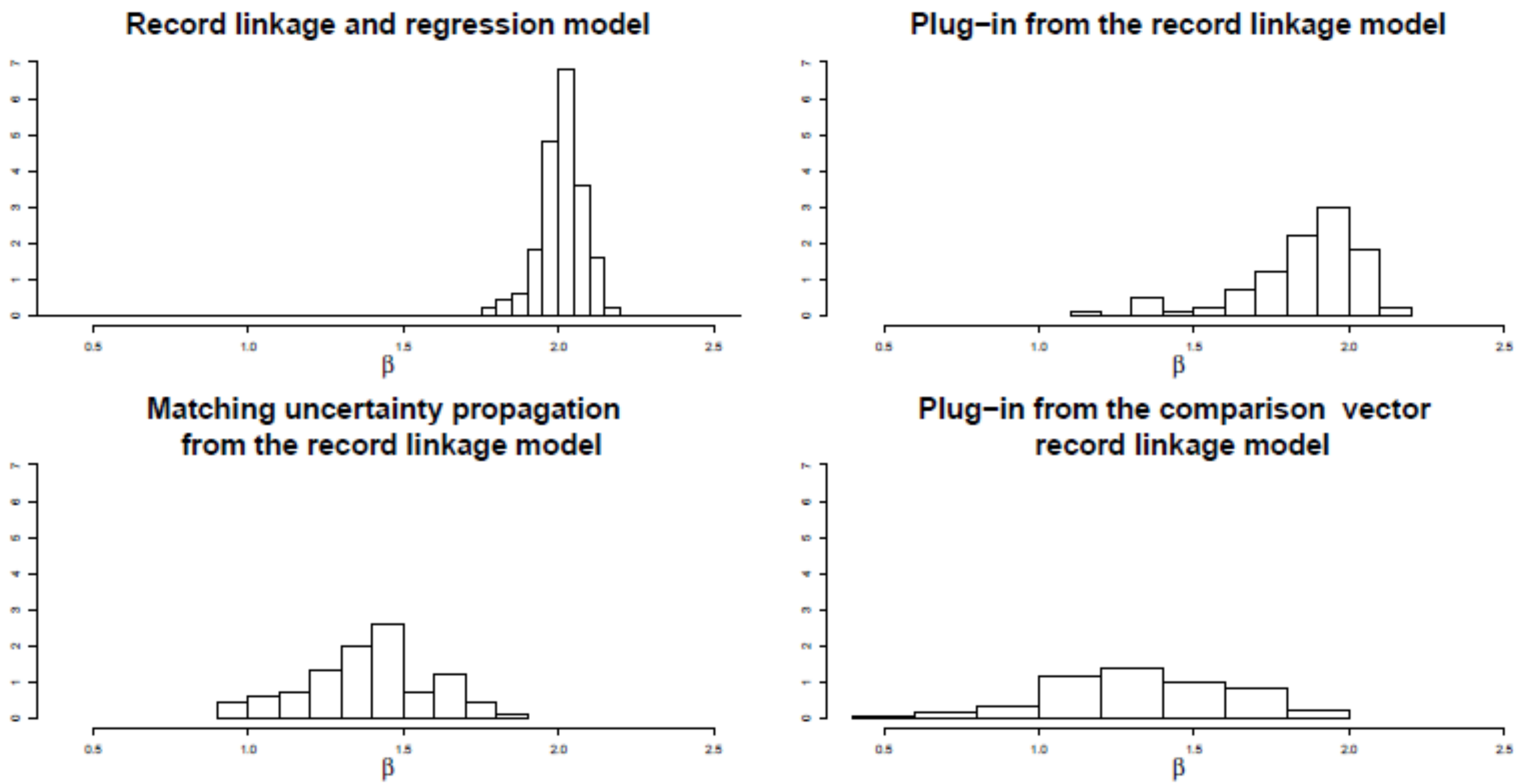
Source: Bond et al. (2014)

Bayesian Extensions for Linkage Error Analysis

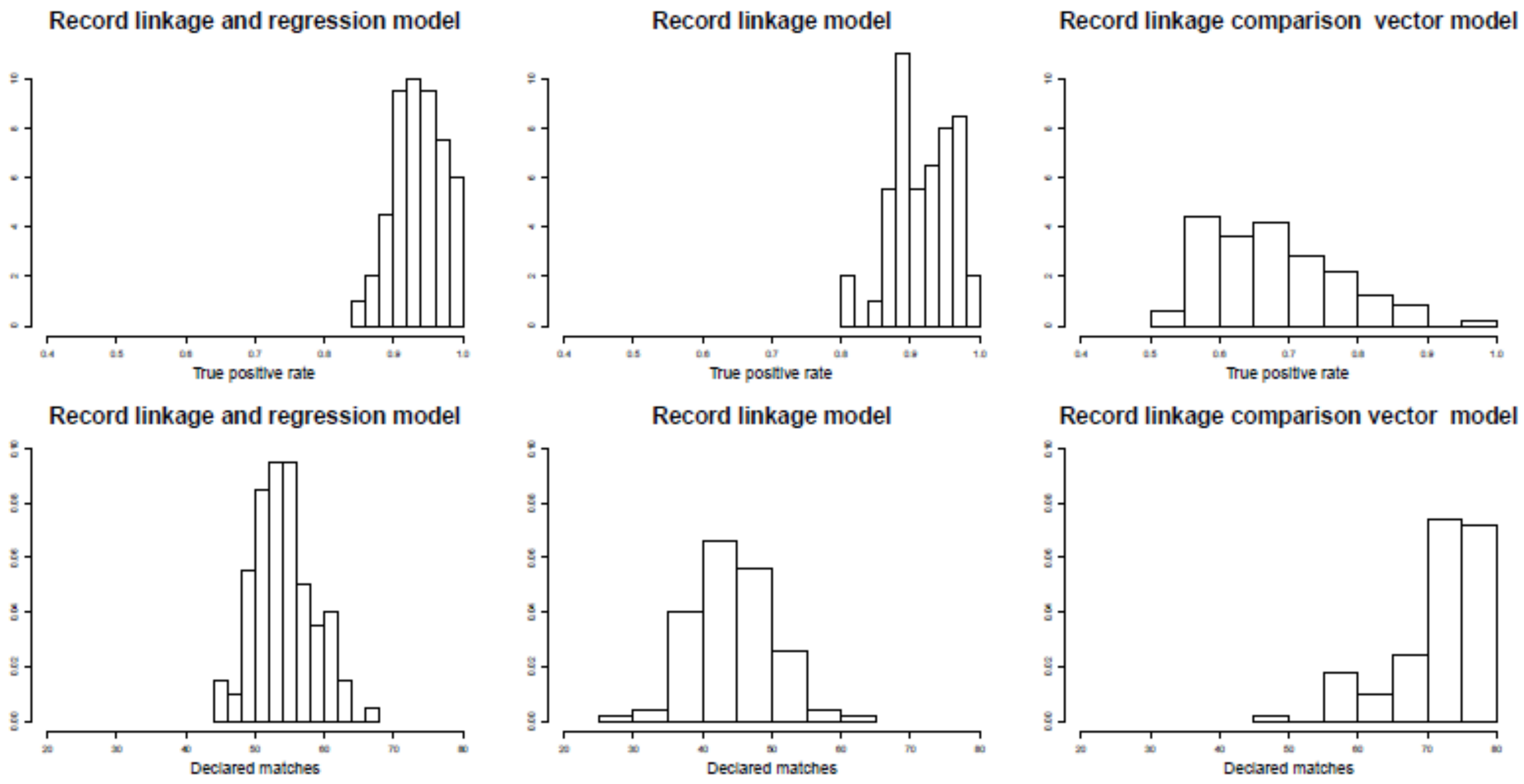
- The full posterior distribution is available (good news)
- Compromises in constructing the posterior distribution (similar to the conditional independence assumption in Fellegi-Sunter) can result in specification errors



Source: Adapted from Abowd and Vilhuber (2011)



Source: Tancredi and Liseo (2015)



Source: Tancredi and Liseo (2015)

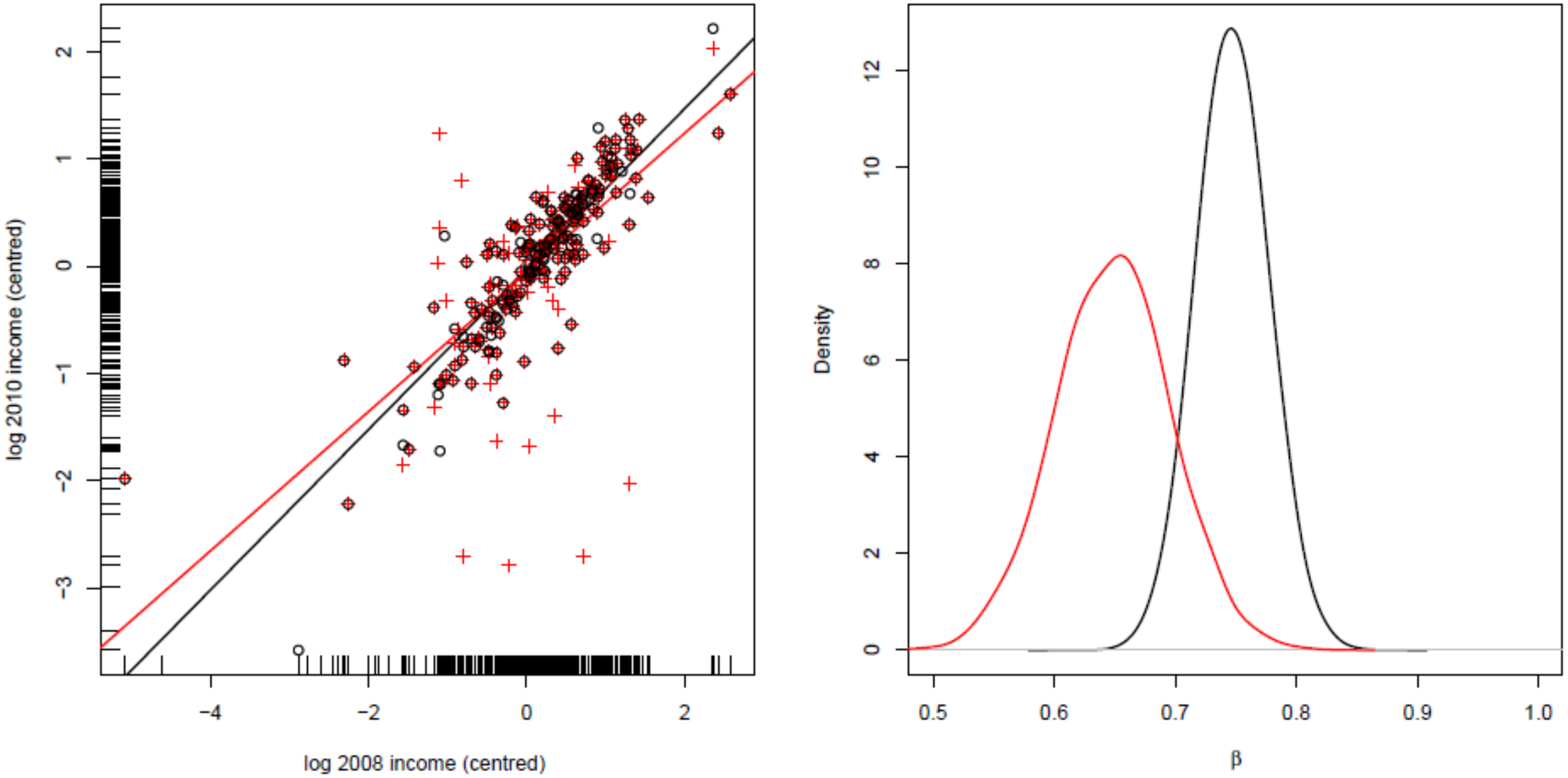


Figure 3 – SHIW data (Friuli block, $n_1 = 434, n_2 = 355$). Regression analysis with the 2010 individual income as the response variable and the 2008 individual income as a covariate. *Left panel*: $\circ$ =true matches, $+$ =declared matches after a perturbation procedure. *Right panel*: posterior distributions for the regression coefficientes with the true matches (black line) and the declared matches.

Source: Tancredi and Liseo (2015)

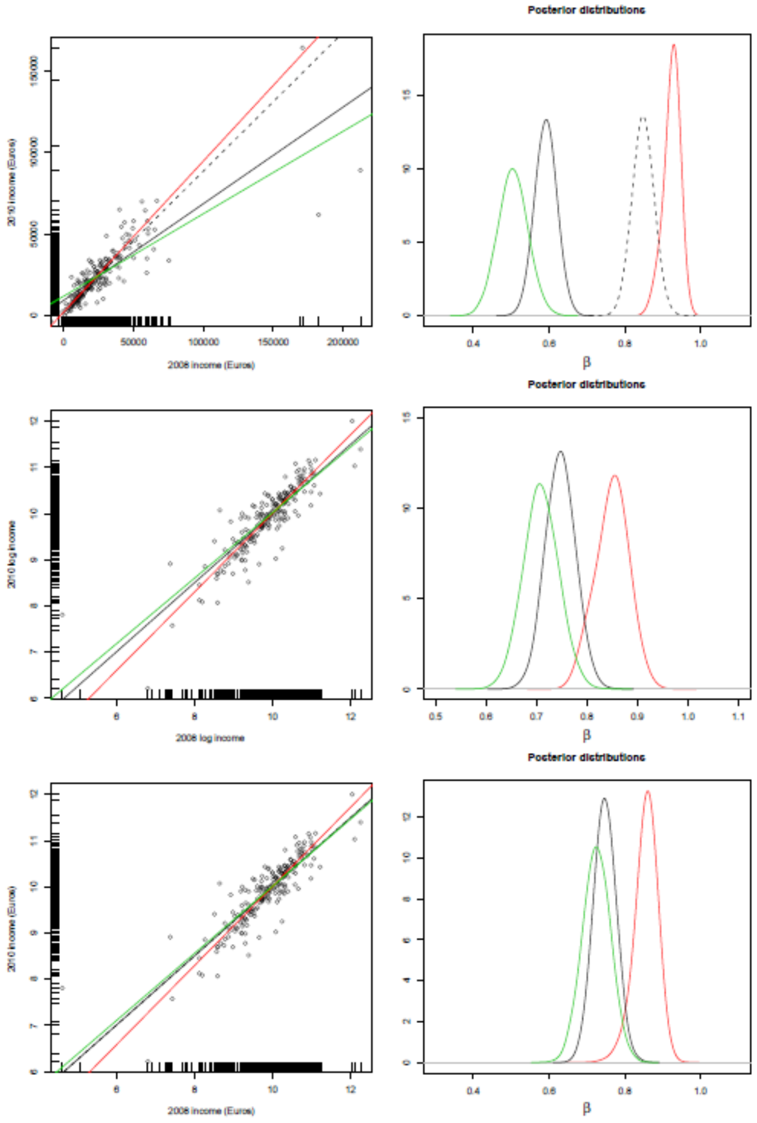


Figure 4 – Results for the SHIW data (Friuli block). Black line: true regression line using the 203 true matches. Black dashed line: true regression line without 2 very influential observations. Red line: Bayesian estimate with the "regression and record linkage". Green line: Bayesian estimate with the "record linkage only" model and posterior regression on the matched pairs. First row: six key variables, non transformed data. Second row: six key variables, log transformed response and covariate. Third row: nine key variables, log transformed response and covariate

Source: Tancredi and Liseo (2015)

Some Food for Thought from the CLIP Data

- Successful matches are very high
- False match rates are very low when using the full set of linking variables and passes
- False match rates can be troubling when only a subset of the linking variables are available
- Commercial data does not link as successfully as administrative records
- There are no estimates of the false non-match rates

Exhibit 1: Match Percentages for Census Bureau PVS Projects

Incoming Data	Matched in Verification	Matched in GeoSearch	Matched in NameSearch	Validated All Incoming
Survey Records				
ACS 2001	N/A	86.30	58.12	93.49
ACS 2002	N/A	86.27	57.57	93.12
ACS 2003	N/A	87.05	54.15	92.39
ACS 2004	N/A	88.16	53.63	92.60
ACS 2005	N/A	89.93	44.77	92.90
ACS 2006	N/A	87.87	47.53	92.03
ACS 2007	N/A	89.06	41.76	91.65
ACS 2008	N/A	88.08	46.07	91.71
ACS 2009	N/A	84.02	52.23	90.82
SIPP 2001*	93.74	69.57	33.19	93.06†
CPS 2001*	94.07	82.20	32.28	76.53
Census Records				
Census 2010	N/A	83.04	57.57	91.14
Federal Administrative Records (2009)				
HUD Public and Indian Housing Information Center File	99.27	42.05	43.53	99.54
IRS Individual Master File and Returns Transaction File (1040)	96.61	7.97	0.30	96.73
IRS Information Returns (1099)	97.28	50.61	0.46	98.66
CMS Active Medicare Enrollment Database	99.92	17.42	30.60	99.89
Indian Health Services Patient Registration File	97.17	29.41	67.23	97.43
Selective Service System Registration File	98.72	46.03	60.01	98.82
HUD Tenant Rental Assistance Certification System File	96.98	55.82	70.19	99.43

ACS yearly results were obtained from "ACS PVS Results All Years for Groves Briefing.xls"
CPS and SIPP results were obtained from "PVS Final Evaluation Report 10242006.doc"
Census 2010 Decennial Response File (DRF) results were obtained from "2010 Char Imp Results by State Table.rtf"
Federal Administrative Records results were obtained from "StARS 2009 PVS Results.doc"
*Results shown are for PVS reruns that occurred after improvements to the system where implemented during the 2004 timeframe.
† The refusals for SIPP 2001 were removed before the file was sent for PVS. Had they been in the file—as they were for the CPS 2001 file—the percent validated of all incoming records would have been much lower.

Source: Mulrow et al. (2011)

Table 1. Observed Error – 2011 MEDB

	2011 MEDB			
	Number of Observations	Search PIK Matches Verified PIK	Search PIK Doesn't Match Verified PIK	% Observed False Matches
Total Verified	**53,058,202**			
GeoSearch				
Passes 1-4	52,186,950	52,184,681	2,269	0.004%
Passes 5-6	157	140	17	10.828%
Pass 9	219,874	219,575	299	0.136%
TOTAL	**52,406,981**		**2,585**	**0.005%**
Zip3 Spatial Adjacency				
Pass 1	11,737	11,735	2	0.017%
Pass 2	5,159,187	5,098,480	60,707	1.177%
TOTAL	**5,170,924**		**60,709**	**1.174%**
NameSearch				
Passes 1-4	49,374,794	49,245,314	129,480	0.262%
DOBSearch				
Passes 1-4	50,327,034	50,237,940	89,094	0.177%

Source: 2011 MEDB.

Source: Layne, Wagner and Rothhaas (2014)

Table 3. Observed Error – 2010 Commercial

	2010 Commercial			
	Records	Search PIK Matches Verified PIK	Search PIK Doesn't Match Verified PIK	% Observed False Matches
Total Verified	**Vendor 1: Total Verified = 210,587,934**			
GeoSearch				
Passes 1-4	161,984,045	161,747,967	236,078	0.146%
Passes 5-6	12,351,322	12,276,474	74,848	0.606%
Pass 9	87,033	75,620	11,413	13.113%
TOTAL	**174,422,400**		**322,339**	**0.185%**
Zip3 Spatial Adjacency				
Pass 1	7,412	7,402	10	0.135%
Pass 2	212,834	209,756	3,078	1.446%
TOTAL	**220,246**		**3,088**	**1.402%**
NameSearch				
Passes 1-4	98,862,854	94,733,395	4,129,459	4.177%
	Vendor 2: Total Verified = 179,860,081			
DOBSearch				
Passes 1-4	65,313,236	60,999,837	4,313,399	6.604%

Source: 2010 Vendor 1 and Vendor 2.

Source: Layne, Wagner and Rothhaas (2014)

Table 5. Modeled and Observed False Match Rates at Cut-Off Weights

	2011 MEDB		2011 IHS		2010 Commercial	
	Probability of a False Match Rate at Cutoff	% Observed Error	Probability of a False Match Rate at Cutoff	% Observed Error	Probability of a False Match Rate at Cutoff	% Observed Error
GeoSearch[1]						
Passes 1-4; Cut-off=14.64	2.220%	0.004%	0.001%	0.046%	1.700%	0.146%
Zip3 Spatial Adjacency[1]						
Pass 2; Cut-off=32.13	3.550%	1.177%	0.261%	0.443%	NA[2]	1.446%
NameSearch						
Passes 1-4; Cut-off=32.14	2.230%	0.262%	0.272%	0.714%	2.550%	4.177%
DOBSearch						
Passes 1-4; Cut-off=32.83	4.050%	0.177%	0.106%	0.392%	1.670%	6.604%

Note:

[1] Passes 5-6 and 9 are omitted because of small sample sizes. Zip3 pass 1 omitted because of a low number of false matches.

[2] The model for this data, module, and pass did not converge and we are examining why.

Source: 2011 MEDB, 2011 IHS, 2010 commercial Vendor 1 and Vendor 2.

Source: Layne, Wagner and Rothhaas (2014)

Critical Take-aways

- Consider sensitivity analyses when using linked data
  - Estimates of the false match rates
  - Use these to assess predictive models like regressions
  - Address representativeness of the analysis sample after linking
  - Perform the analysis with alternative linking strategies
- Begin to experiment with full-scale virtual population models
  - Important for linking business data
  - Likely important for linking decennial censuses

References (in order of appearance)

- To be completed

Thank you

john.maron.abowd@census.gov