

Data Quality: Automated Edit/Imputation and Record Linkage

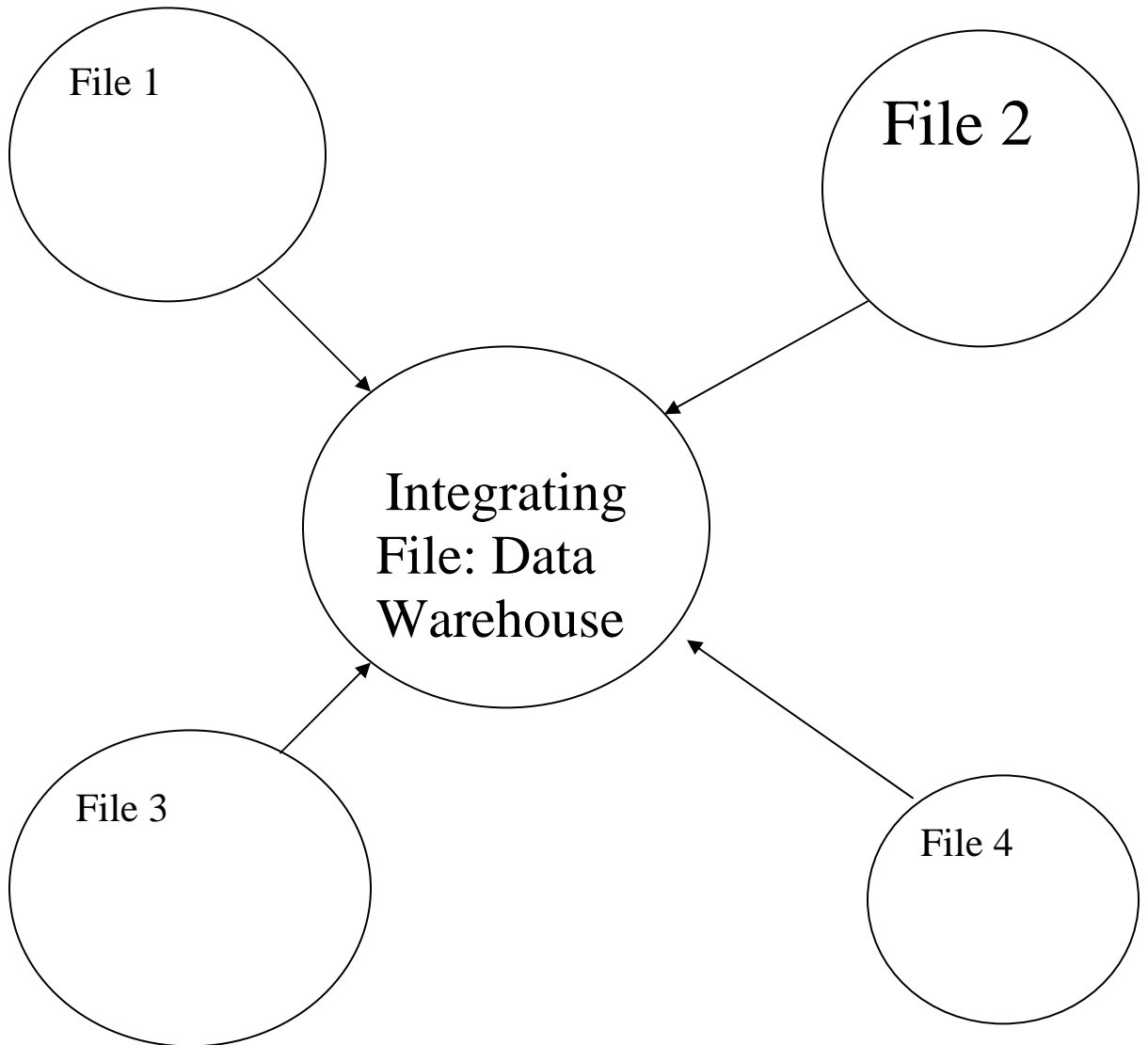
William E. Winkler, william.e.winkler@census.gov

Q2006 Conference, 25 April, 2006

<http://www.census.gov/srd/www/byyear.html>

Outline

1. Introduction
2. Examples of Enhanced Efficiency and Improved Quality
3. Rules of Thumb, Vision, Data Quality Issues
4. Record Linkage and Current Advances
5. Concluding Remarks



Record linkage – link names, addresses, demographic data, ages

Comparison metrics extended to numeric data containing small amounts of error.

Used in microdata confidentiality experiments.

Record linkage can yield significant cost savings.

Resources for US Decennial Census Matching

	<u>clerical</u>	<u>computerized</u>	
		<u>1988</u>	<u>1990</u>
# clerks	3000	600	200
# month	6	1.5	1.5
false match rate	5%	0.5%	0.2%
computer match proportion	0%	70%	75%

Updating and unduplicating a survey frame

Identify duplicates in 6 million records
from 12 lists

	1987	1992
duplicates	6.6%	12.8%
potential duplicates	28.9%	19.7%
final file duplication	~10%	~2%
clerical resources	75 clerks for 3 months	6500 person hours

Edit/Imputation (missing data, contradictory data)

1. Kovar & Winkler (1996) – Install and run generalized GEIS and SPEER systems in less than one day.
2. Winkler (1997) – Install and run DISCRETE in less than 4 hours each on two discrete, demographic surveys. Complete edit/imputation.
3. Garcia & Thompson (2000) – Large survey complicated edit patterns. 12 analysts for six months. Changed 3 times as many fields as FH system that took 24 hours to run.
4. Herzog, Scheuren, & Winkler (2006) – One analyst/programmer installed and ran generalized system in less than 3 weeks. Replaced 6 programmers (6-12 months of effort) and 12 analysts (6 months of effort).

Common Issues: Generalized software and skilled individuals

Record Linkage – Generalized systems based on Fellegi-Sunter model (JASA 1969) that are portable across matching situations and computers.

Edit/Imputation – Generalized systems based on Fellegi-Holt model (JASA 1976) that are extended to generalized imputation via methods of Little and Rubin (1987, 2002) and Winkler (2003).

Vision: Analyst with suitable software can do most edit/imputation and record linkage to ‘clean up’ a database or set of databases

Data Quality Issues (Metrics often missing)

Within a single file

- Duplicates

- Errors and missing data in fields

- Lack of coverage of population

Across files

- Lack of common identifiers

- Errors and missing data in fields

- Inconsistency of values in fields

- Duplicates

Suitability for analyses

First rule of thumb: List problems (both duplicates and lack of coverage) can cause greater difficulties with survey estimates than all other sources of error combined.

Second rule of thumb: If there are no list problems, then it is possible that problems with statistical data editing and imputation can cause greater difficulties with survey estimates than all other sources of error combined.

Elementary Examples

Name	Address	Age
John A Smith	16 Main Street	16
J H Smith	16 Main St	17
Javier Martinez	49 E Applecross Road	33
Haveir Marteenez	49 Aplecross Raod	36
Gillian Jones	645 Reading Aev	24
Jilliam Brown	123 Norcross Blvd	43

Names

Starting with a free-form name, how do we get at the components that need to be compared?

Table **Examples of Name Parsing**

Standardized								
<hr/>								
1.	DR	John	J	Smith	MD			
2.		Smith	DRY	FRM				
3.		Smith & Son	ENTP					
<hr/>								
Parsed								
<hr/>								
	PRE	FIRST	MID	LAST	POST1	POST2	BUS1	BUS2
<hr/>								
1.	DR	John	J	Smith	MD			
2.				Smith			DRY	FRM
3.				Smith		Son	ENTP	
<hr/>								

Extensions of Fellegi-Sunter Model

Missing identifiers (i.e., fields)

Typographical errors – String Comparators

Frequency (Smith vs Zabransky)

Numeric Data

Estimation of Error Rates

Key Advances (unsupervised learning)

Automatic estimation of optimal parameters without training data (Winkler 1988, 1989, 1993; Ravikumar & Cohen *UAI* 2004, Shen, Li, & Doan *AAAI* 2005).

Automatic estimation of error rates without training data (Belin & Rubin *JASA* 1995, Winkler & Yancey 2006)

Exceptionally general data extraction and standardization (Borkar, Deshmukh, & Sarawagi *SIGMOD* 2001, Agichstein & Ganti *KDD* 2004, Cohen & Sarawagi *KDD* 2004)

BigMatch Software

Yancey & Winkler 2004, 2006; Yancey, Winkler, & Creecy 2006

Match moderate size list having 300 million records against large administrative lists having upwards of 10 billion records

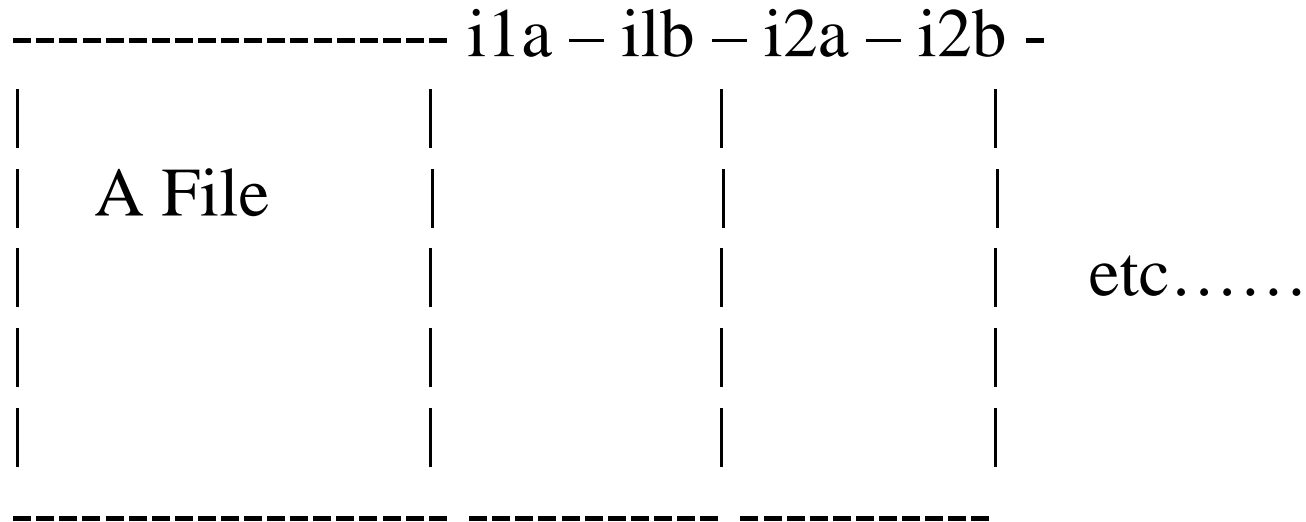
10 sets of blocking criteria

No sorting of files

New indexing, very fast retrieval and comparisons

300 million \times 300 million $\sim 10^{17}$

Small File A and Large File B



IBM PC – 4 gigabytes RAM – 2MHz

FAST: 10 blocking passes, all sorts, all matches

100,000 pairs per second

BigMatch allows exploration of additional sets of blocking criteria.

Very Large Matching Situation – Significant savings in time, disk space, and skilled person intervention
Match 100 million record B-file against 1 billion record A-file with 10 blocking passes. With conventional matching, must do 20 passes on each file.

With conventional: **5+ days CPU** time to sort A-file 10 times. With BigMatch: **0 CPU** time to sort A-file.

Hard-to-find missed matches (*artificial data*)
(date-of-birth missing in file B, address missing in file A)

	Household 1		Household 2	
	First	Last	First	Last
HeadH	Julia	Smoth	Julia	Smith
Child1	Jerome	Jones	Gerone	Smlth
Child2	Shyline	Jones	Shayleene	Smith
Child3	Chrstal	Jcnes	Magret	Smith

Head of Household gave linkage. **No 3-grams in common.**

Concluding Remarks

Identifying and correcting duplicates within and across files is the major challenge. Correcting data via edit rules and filling in missing data are other difficulties.

Powerful methods are available for both record linkage and edit/imputation. General software and algorithms are available and can be adapted.

Research Problem: Developing *effective* metrics that connect specific types of errors with specific effects on analyses.