Working with UI Wage Data: Challenges & Triumphs

Jessica Shedd
The University of Texas System
Office of Strategic Initiatives

AIR Forum 2015

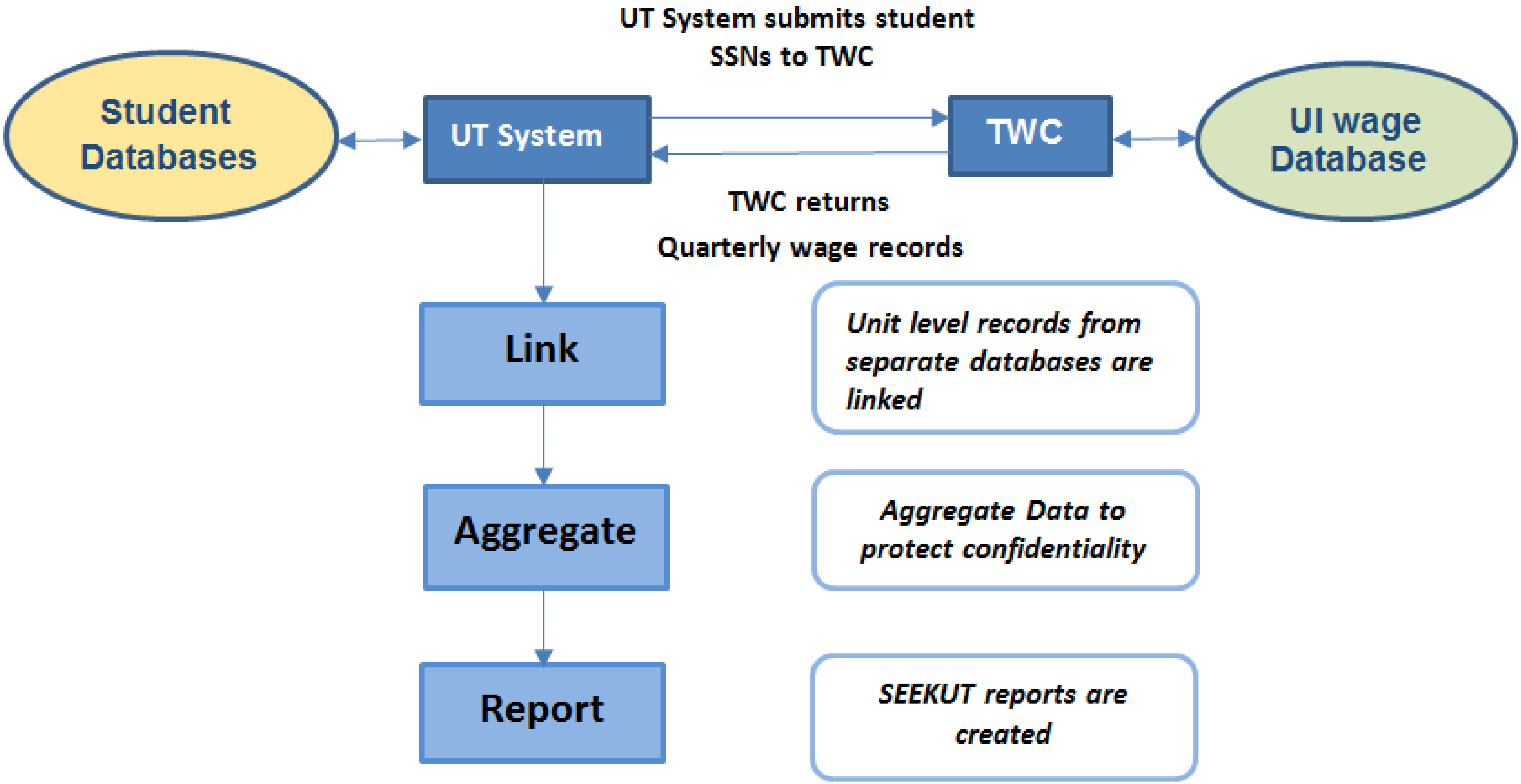
National Picture

- Increased focus on post-collegiate outcomes
  - Gainful Employment
  - Obama admin. proposed college ratings
  - Demand for demonstrating value-add & ROI
  - Current proposed legislation
  - Accreditors

Data Sharing Agreement

- Contract with TX Workforce Commission
  - Individual level data
  - Cost associated with processing time
  - Agreement for 5 years

Overview of the Process



UT System submits student SSNs to TWC

Student Databases

UT System

TWC

UI wage Database

TWC returns

Quarterly wage records

Link

Unit level records from separate databases are linked

Aggregate

Aggregate Data to protect confidentiality

Report

SEEKUT reports are created

Data Preparation for Submission to TWC

- Pull SSNs from THECB degree and enrollment data
  - 1.1 million student SSNs, students enrolled or graduated between 1999 and 2012

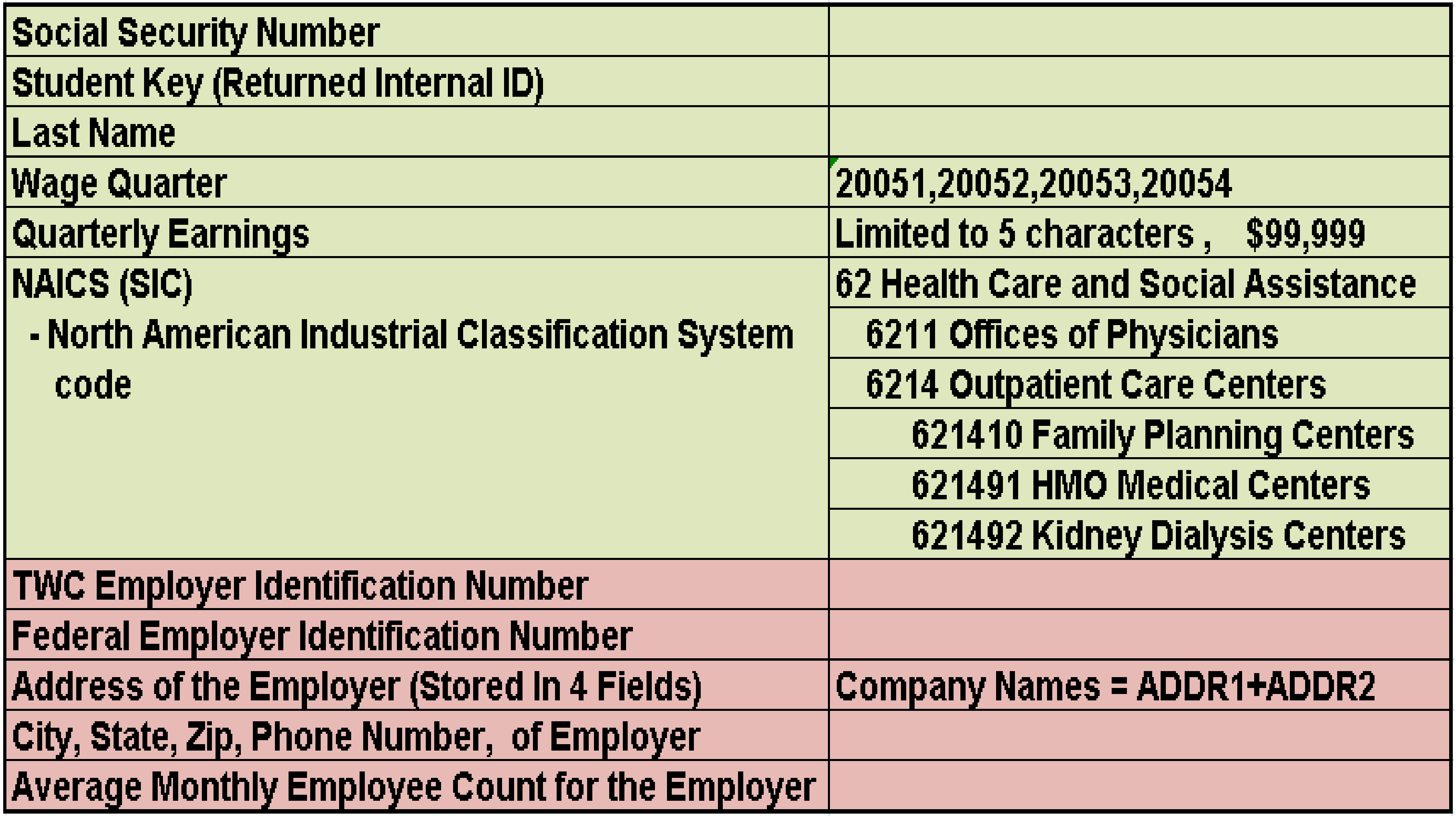
- Send Social Security Number and IDs used to match student and UI wage data
  - Data in TXT format
  - Transfer via secure FTP

- Set up a secure folder to house wage data
- Security training for all users

Data Returned from TWC

- Over 26 million records (Rows)
  - Hierarchical file containing one record per:
    - SSN
    - Employer
    - Quarter
- Less Than 20 Fields (Columns)

UI Wage Data Elements



Social Security Number	
Student Key (Returned Internal ID)	
Last Name	
Wage Quarter	20051,20052,20053,20054
Quarterly Earnings	Limited to 5 characters , $99,999
NAICS (SIC) - North American Industrial Classification System code	62 Health Care and Social Assistance
	6211 Offices of Physicians
	6214 Outpatient Care Centers
	621410 Family Planning Centers
	621491 HMO Medical Centers
	621492 Kidney Dialysis Centers
TWC Employer Identification Number	
Federal Employer Identification Number	
Address of the Employer (Stored In 4 Fields)	Company Names = ADDR1+ADDR2
City, State, Zip, Phone Number, of Employer	
Average Monthly Employee Count for the Employer	

UI Wage Data Limitations

- Includes data for the state of TX only
  - Exception: U.S. gov't employees of U.S. Postal Service, Dept. of Defense military, or U.S. Office of Personnel Management
  - Does not include self-employed

- Does not include number of hours worked
- Does not include full-time or part-time status

- Does not include occupation information
  - Cannot determine if employed in field of study, includes industry codes of employers only
- Does not include location where person is working

Resulting Data

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2002	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2003	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2004	✓	✓	✓	✓	✓	✓	✓	✓	✓	
2005	✓	✓	✓	✓	✓	✓	✓	✓		
2006	✓	✓	✓	✓	✓	✓	✓			
2007	✓	✓	✓	✓	✓	✓				
2008	✓	✓	✓	✓	✓					
2009	✓	✓	✓	✓						
2010	✓	✓	✓							
2011	✓	✓								
2012	✓									

Match Rates for Degree Recipients

- One-Year and 10 Years after Graduation
  - Bachelor’s Degree = 78% and 65%
  - Master’s Degree = 66% and 51%
  - Doctoral Degree = 43% and 32%
  - Professional = 63% and 54%

Predicting Missing Wage

- Determined probabilities of degree recipients’ wage data not being found one year after graduation

- Logistic Regression
  - Binary dependent variable: wage found or not
  - Included 10 yrs of UT (all system institutions) graduates by degree level

Predicting Missing Wage

	Probabilities		
	Bachelor's (22%)	**Master's (34%)**	**Doctor's (57%)**
RACE:			
African American	0.196	0.222	0.483
Asian American	0.261	0.284	0.573
Hispanic	0.176	0.188	0.445
International	**0.477**	**0.581**	**0.689**
Other	0.242	0.306	0.518
White	0.211	0.236	0.486
RESIDENCY:			
Foreign	0.312	0.361	0.545
Out of State	**0.460**	**0.572**	**0.711**
Texas	0.199	0.231	0.474

Predicting Missing Wage

	Probabilities
	Professional (37%)
RESIDENCY:	
Foreign	0.502
Out of State	**0.618**
Texas	0.338
GROUPED MAJORS:	
Health	0.294
Legal Professions	**0.571**

Creating an Analytic File on Degree Recipients

Data Cleaning

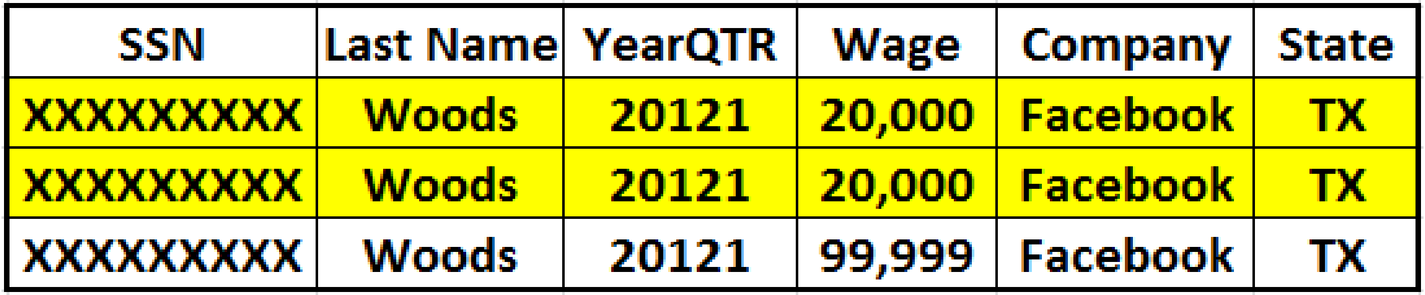
File Restructuring

Wage Data Cleaning Process

- What did we discover?
  - One SSN tied to multiple names
- Why does it matter?
  - Inflate total annual earnings
- Some extreme wages

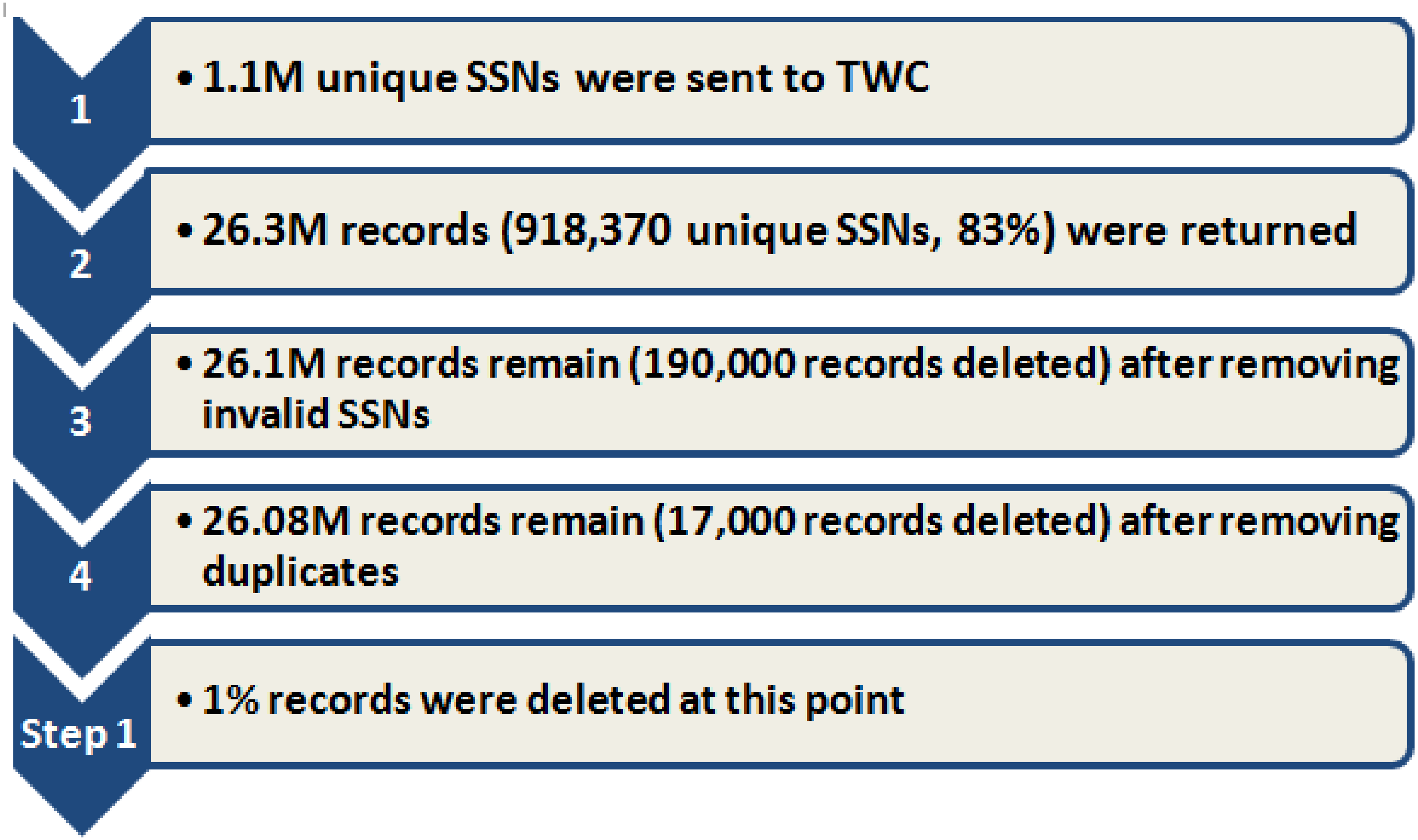
Step1 : Cleaning Based on Invalid SSN Guideline

- Removing SSNs with:
  - All zeros in a digit group: 000-XX-XXXX, XXX-00-XXXX, XXX-XX-0000
  - 666 or 900-999 in the first digit group
  - 123456789,111111111,012345678,001234567
- Removing exact duplicates records:



SSN	Last Name	YearQTR	Wage	Company	State
XXXXXXXXX	Woods	20121	20,000	Facebook	TX
XXXXXXXXX	Woods	20121	20,000	Facebook	TX
XXXXXXXXX	Woods	20121	99,999	Facebook	TX

UI Wage Data Cleaning process - Step 1 Summary



- 1.1M unique SSNs were sent to TWC

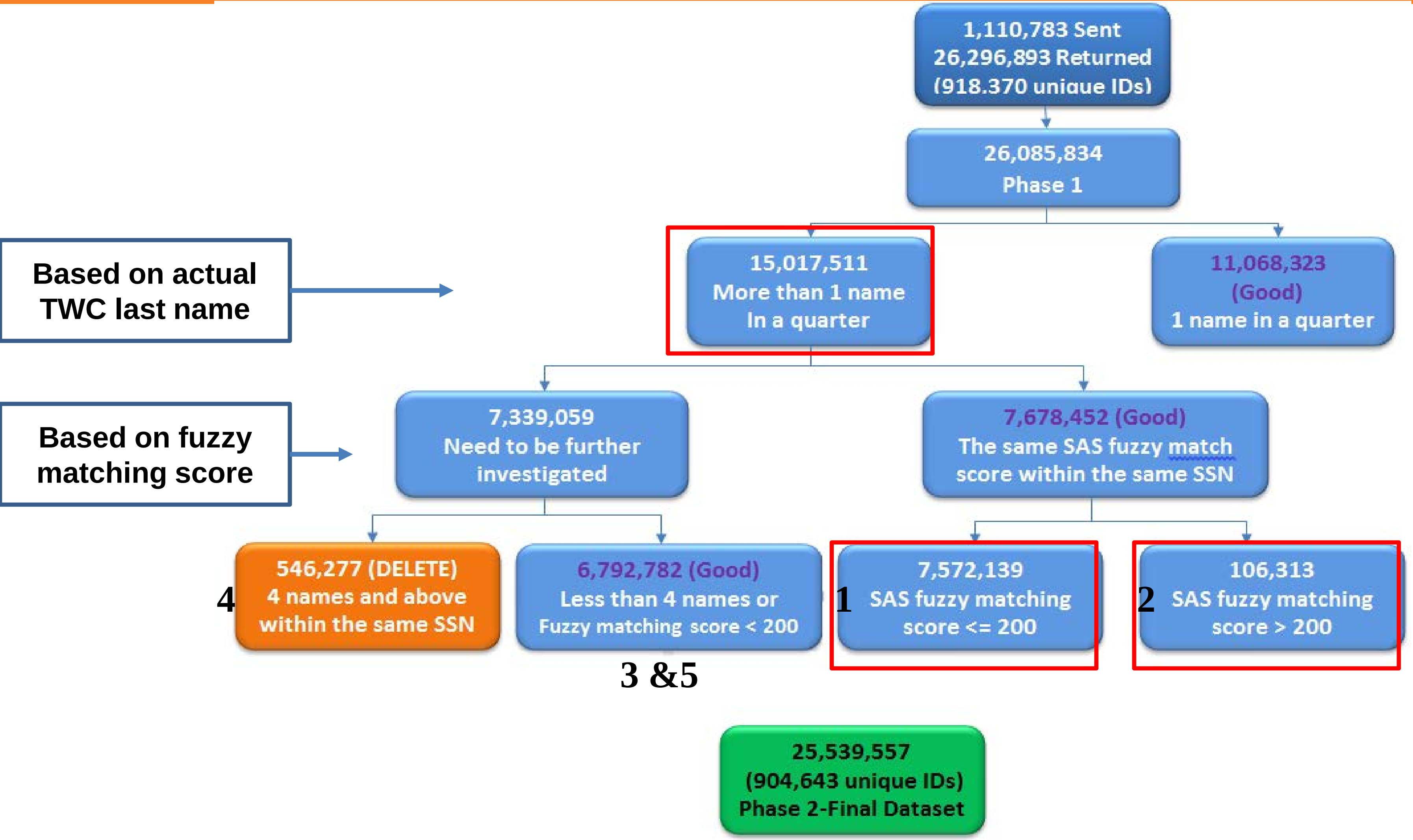
- 26.3M records (918,370 unique SSNs, 83%) were returned

- 26.1M records remain (190,000 records deleted) after removing invalid SSNs

- 26.08M records remain (17,000 records deleted) after removing duplicates

- 1% records were deleted at this point

UI Wage Data Cleaning process – Step 2 Summary



Based on actual TWC last name

Based on fuzzy matching score

1,110,783 Sent
26,296,893 Returned
(918,370 unique IDs)

26,085,834
Phase 1

15,017,511
More than 1 name
In a quarter

11,068,323
(Good)
1 name in a quarter

7,339,059
Need to be further investigated

7,678,452 (Good)
The same SAS fuzzy match score within the same SSN

4 546,277 (DELETE)
4 names and above within the same SSN

6,792,782 (Good)
Less than 4 names or Fuzzy matching score < 200

3 &5

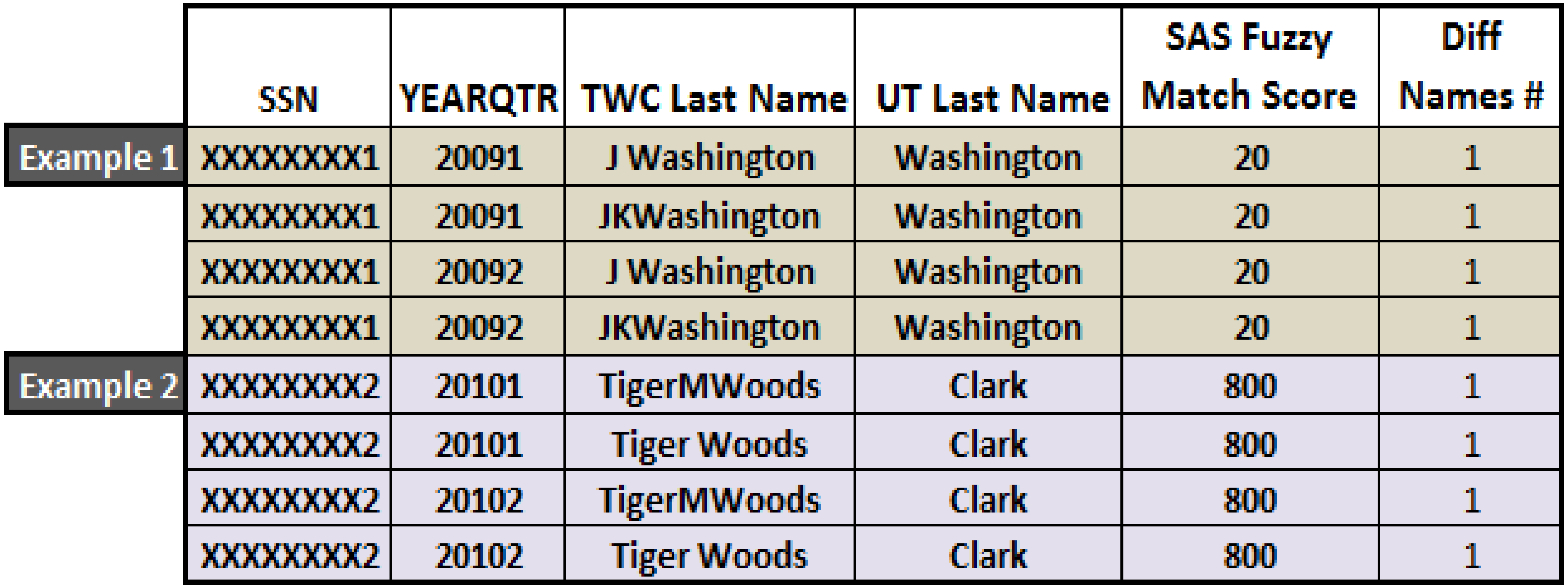
1 7,572,139
SAS fuzzy matching score <= 200

2 106,313
SAS fuzzy matching score > 200

25,539,557
(904,643 unique IDs)
Phase 2-Final Dataset

Step 2: Cleaning Based on UT Last Name (example 1 & 2)

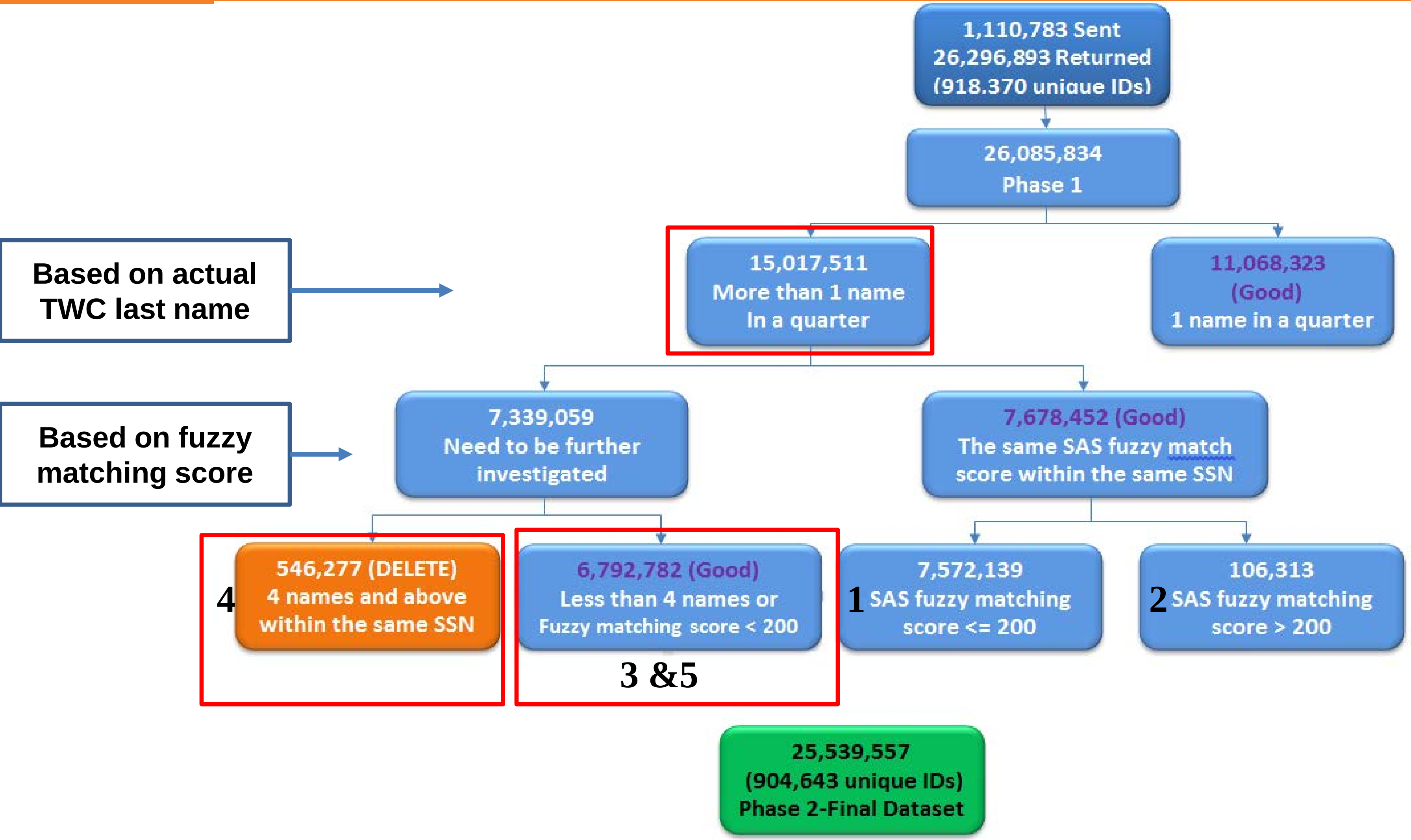
- Only one UT last name at graduation
- TWC names in a quarter- one or potentially more names
- Create SAS fuzzy matching score to identify different names
  - <=200 = UT name and TWC name similar
  - > 200 = UT name and TWC name very different



	SSN	YEARQTR	TWC Last Name	UT Last Name	SAS Fuzzy Match Score	Diff Names #
Example 1	XXXXXXXX1	20091	J Washington	Washington	20	1
	XXXXXXXX1	20091	JKWashington	Washington	20	1
	XXXXXXXX1	20092	J Washington	Washington	20	1
	XXXXXXXX1	20092	JKWashington	Washington	20	1
Example 2	XXXXXXXX2	20101	TigerMWoods	Clark	800	1
	XXXXXXXX2	20101	Tiger Woods	Clark	800	1
	XXXXXXXX2	20102	TigerMWoods	Clark	800	1
	XXXXXXXX2	20102	Tiger Woods	Clark	800	1

- **Result:** Keep SSNs in Example 1 and 2 in the wage data

UI Wage Data Cleaning process – Step 2 Summary



1,110,783 Sent
26,296,893 Returned
(918.370 unique IDs)

26,085,834
Phase 1

Based on actual TWC last name

15,017,511
More than 1 name
In a quarter

11,068,323
(Good)
1 name in a quarter

Based on fuzzy matching score

7,339,059
Need to be further investigated

7,678,452 (Good)
The same SAS fuzzy match score within the same SSN

4 546,277 (DELETE)
4 names and above within the same SSN

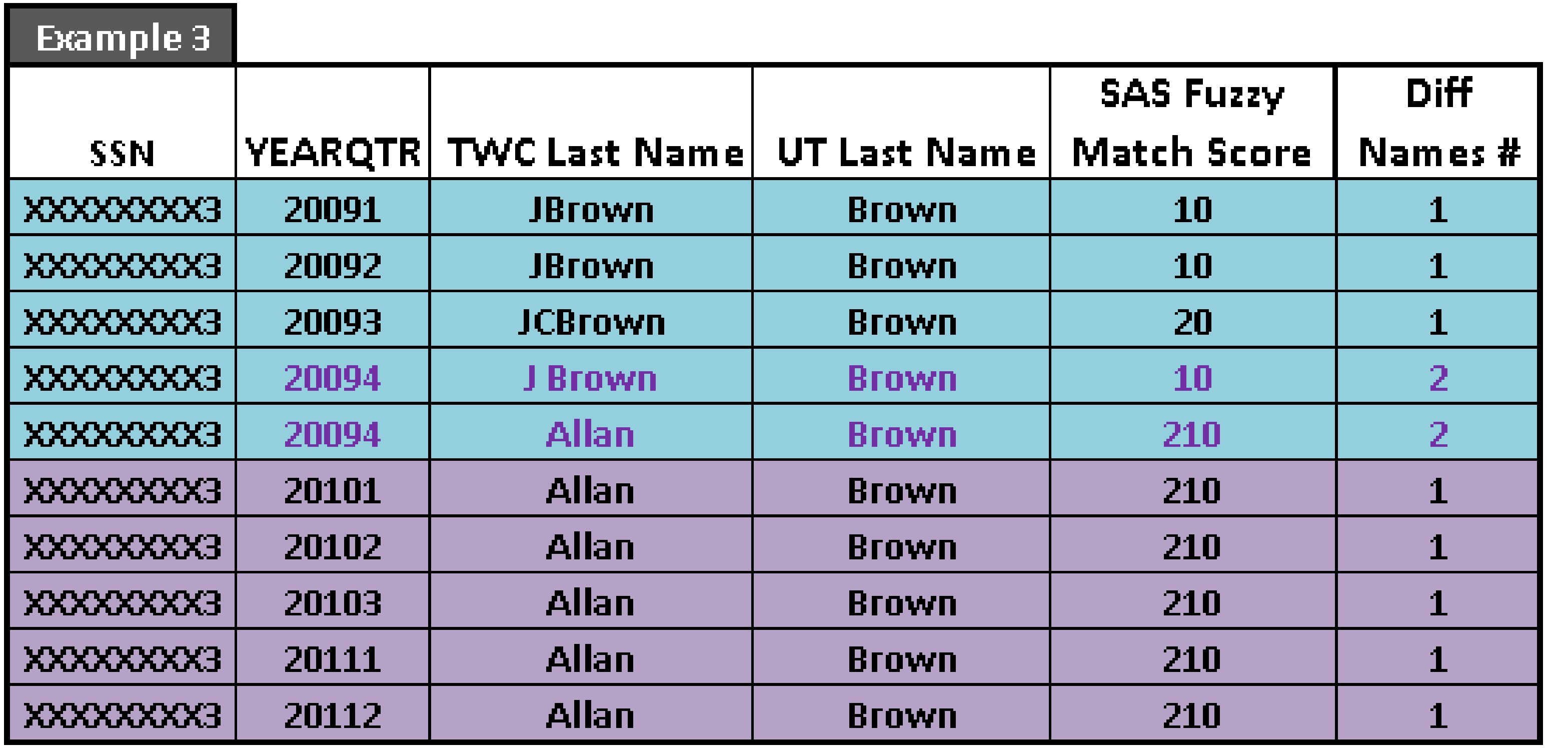
6,792,782 (Good)
Less than 4 names or Fuzzy matching score < 200
3 &5

7,572,139
1 SAS fuzzy matching score <= 200

106,313
2 SAS fuzzy matching score > 200

25,539,557
(904,643 unique IDs)
Phase 2-Final Dataset

Example 3 – Kept SSNs



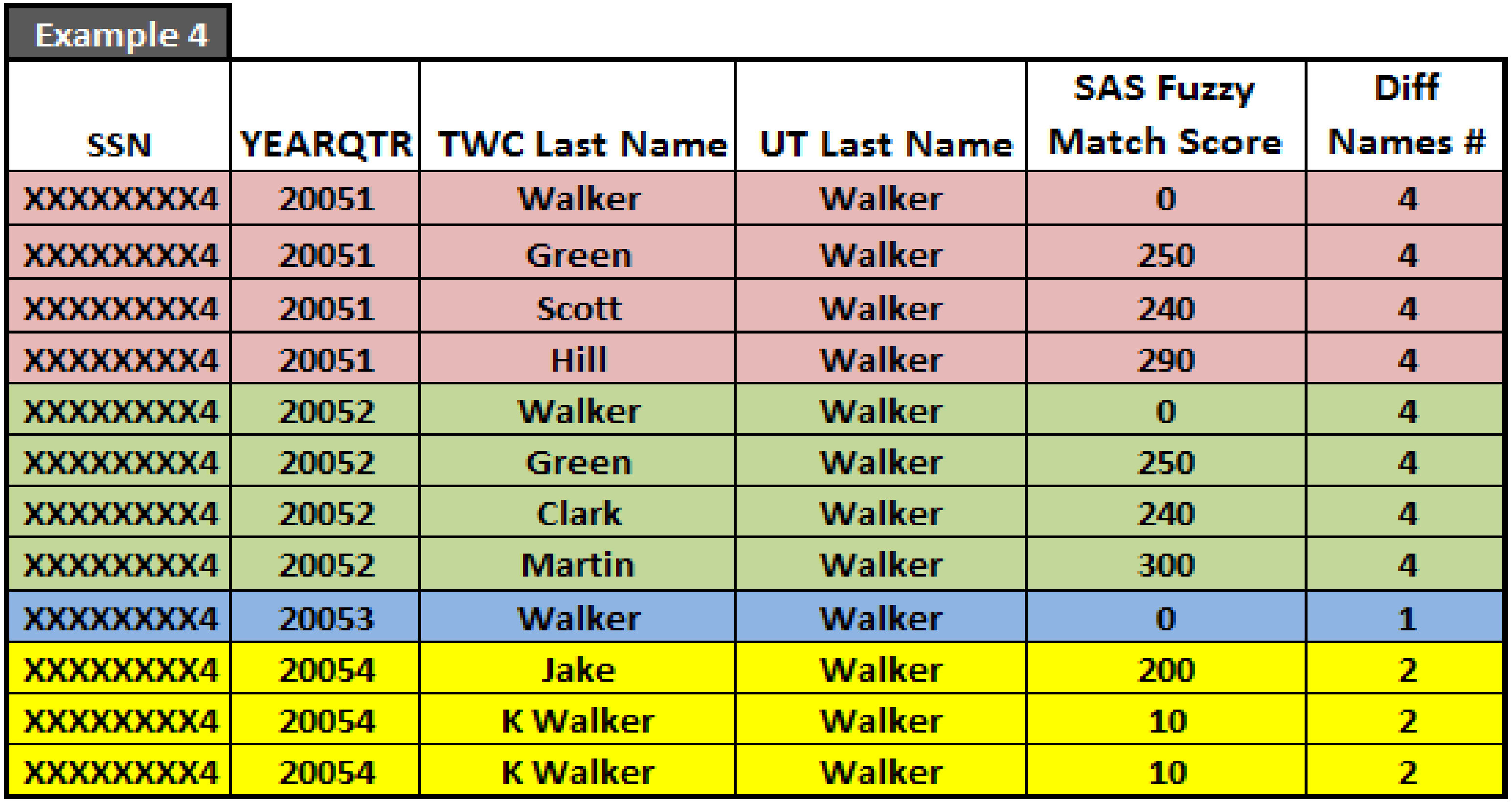
Example 3

SSN	YEARQTR	TWC Last Name	UT Last Name	SAS Fuzzy Match Score	Diff Names #
XXXXXXXX3	20091	JBrown	Brown	10	1
XXXXXXXX3	20092	JBrown	Brown	10	1
XXXXXXXX3	20093	JCBrown	Brown	20	1
XXXXXXXX3	20094	J Brown	Brown	10	2
XXXXXXXX3	20094	Allan	Brown	210	2
XXXXXXXX3	20101	Allan	Brown	210	1
XXXXXXXX3	20102	Allan	Brown	210	1
XXXXXXXX3	20103	Allan	Brown	210	1
XXXXXXXX3	20111	Allan	Brown	210	1
XXXXXXXX3	20112	Allan	Brown	210	1

Result:

- Though 2 names are identified for the same person…
- Keep SSN in Example 3 in the wage data as 1 person

Example 4 – Removed SSNs



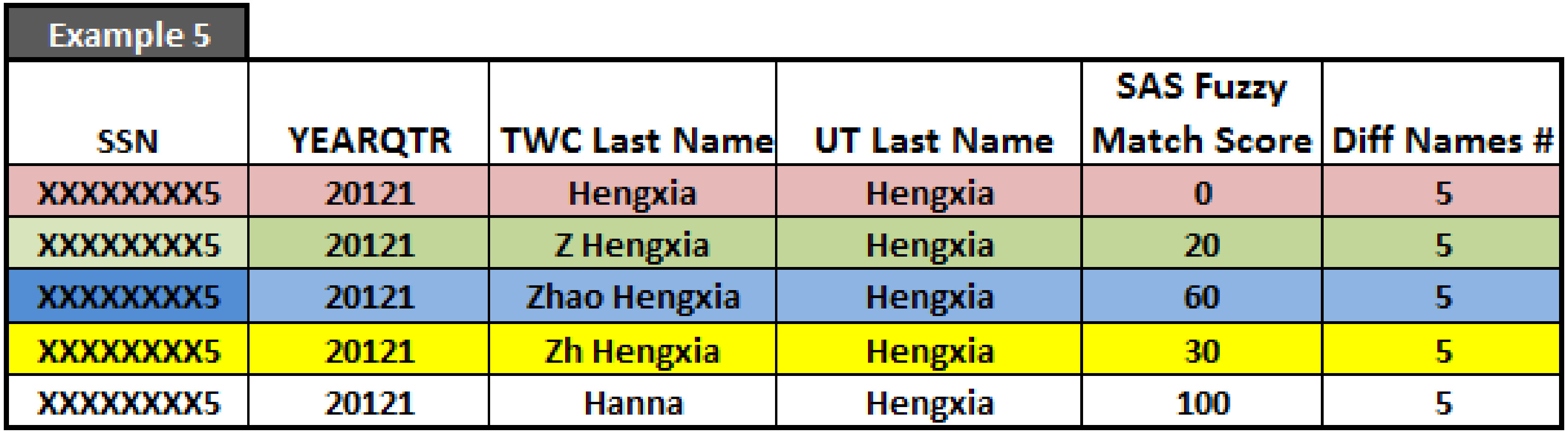
Example 4

SSN	YEARQTR	TWC Last Name	UT Last Name	SAS Fuzzy Match Score	Diff Names #
XXXXXXXX4	20051	Walker	Walker	0	4
XXXXXXXX4	20051	Green	Walker	250	4
XXXXXXXX4	20051	Scott	Walker	240	4
XXXXXXXX4	20051	Hill	Walker	290	4
XXXXXXXX4	20052	Walker	Walker	0	4
XXXXXXXX4	20052	Green	Walker	250	4
XXXXXXXX4	20052	Clark	Walker	240	4
XXXXXXXX4	20052	Martin	Walker	300	4
XXXXXXXX4	20053	Walker	Walker	0	1
XXXXXXXX4	20054	Jake	Walker	200	2
XXXXXXXX4	20054	K Walker	Walker	10	2
XXXXXXXX4	20054	K Walker	Walker	10	2

Results:

- 4 names are identified (and at least one name > 200)
- REMOVE SSN in Example 4

Example 5 – Kept SSNs



Example 5

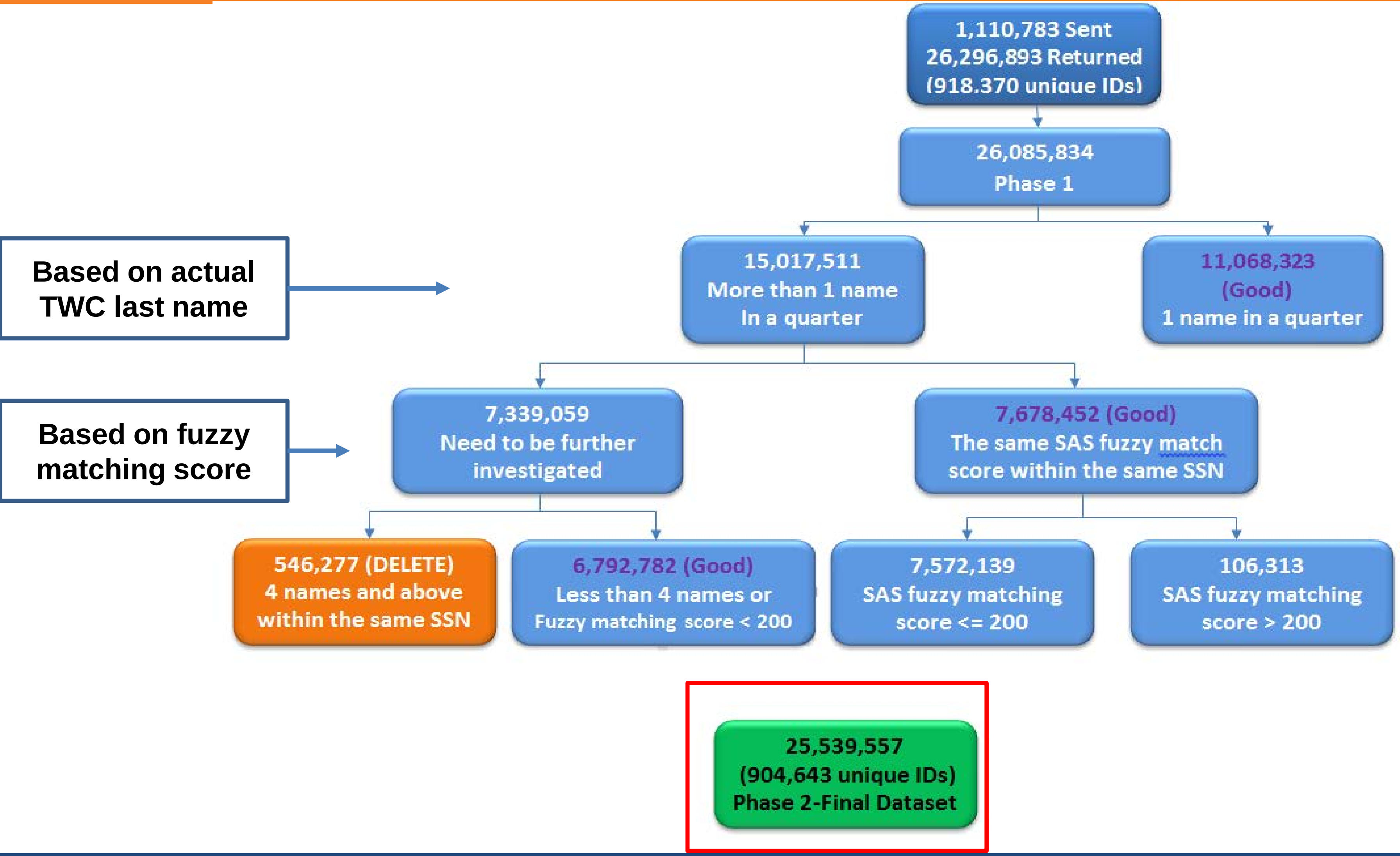
SSN	YEARQTR	TWC Last Name	UT Last Name	SAS Fuzzy Match Score	Diff Names #
XXXXXXXX5	20121	Hengxia	Hengxia	0	5
XXXXXXXX5	20121	Z Hengxia	Hengxia	20	5
XXXXXXXX5	20121	Zhao Hengxia	Hengxia	60	5
XXXXXXXX5	20121	Zh Hengxia	Hengxia	30	5
XXXXXXXX5	20121	Hanna	Hengxia	100	5

- Results:
- Though 5 names are identified
- Keep SSN in Example 5 as 1 person
- All SAS fuzzy match scores <=200

Our Rule of Thumb

- SSNs removed when:
  - 4 names and above are identified
  - At least one name has fuzzy matching score > 200

UI Wage Data Cleaning process – Step 2 Summary



Based on actual TWC last name

Based on fuzzy matching score

1,110,783 Sent
26,296,893 Returned
(918.370 unique IDs)

26,085,834
Phase 1

15,017,511
More than 1 name
In a quarter

11,068,323
(Good)
1 name in a quarter

7,339,059
Need to be further
investigated

7,678,452 (Good)
The same SAS fuzzy match
score within the same SSN

546,277 (DELETE)
4 names and above
within the same SSN

6,792,782 (Good)
Less than 4 names or
Fuzzy matching score < 200

7,572,139
SAS fuzzy matching
score <= 200

106,313
SAS fuzzy matching
score > 200

25,539,557
(904,643 unique IDs)
Phase 2-Final Dataset

Analytic Decisions for seekUT tool

- Focus on graduates, working full-time for a full-year, for purposes of creating a tool for students
    - Working all 4 quarters of calendar year
    - Annual earnings >= 35 hours x $7.25 x 52 weeks = $13,195
      [quarterly earning >= $3,298.75]

Handling Multiple Degrees

- In different academic year:
  - Keep all degrees
    - Earned BA in 2005
    - 1-yr post graduation earnings based on 2006
    - Earned PhD in 2010
    - 1-yr post graduation earnings based on 2011

- In the same academic year:
  - Different degree levels →keep the highest degree
  - Same degree levels → keep all degrees

Create Flat Data file

- Flatten the data file to a row per graduate
  - Includes annual wage data for up to 10yrs
  - Employer and Industry where the highest wage was earned in that given year
  - Adjust all wage data to 2013 dollars
- Merge in student information from THECB data (major, financial aid, etc.)
  - Adjust all loan data to 2013 dollars

Used NSC Student Tracker

- For additional context, we sent all records sent to TWC to NSC Student Tracker
    - Able to determine who continued their education after leaving a UT institution

Baccalaureate Recipient Match Rates

Baccalaureate	System		Academic		Health	
	1st	10th	1st	10th	1st	10th
Working in TX, FY & FT*	50%	55%	49%	54%	74%	63%
Working in TX, Other	28%	10%	29%	10%	13%	10%
Working in TX Total	**78%**	**65%**	**78%**	**64%**	**87%**	**73%**
Enrolled Only	7%	2%	7%	2%	2%	2%
Total Found	**85%**	**67%**	**85%**	**66%**	**89%**	**75%**

Master’s Recipient Match Rates

Master’s	System		Academic		Health	
	1st	10th	1st	10th	1st	10th
Working in TX, FY & FT*	53%	44%	52%	43%	59%	49%
Working in TX, Other	13%	7%	13%	7%	12%	7%
Working in TX Total	**66%**	**51%**	**65%**	**50%**	**71%**	**56%**
Enrolled Only	3%	2%	3%	2%	3%	3%
Total Found	**69%**	**53%**	**68%**	**52%**	**74%**	**59%**

Doctoral Recipient Match Rates

Doctoral	System		Academic		Health	
	1st	10th	1st	10th	1st	10th
Working in TX, FY & FT*	29%	27%	29%	25%	35%	34%
Working in TX Other	14%	5%	13%	5%	15%	3%
Working in TX Total	**43%**	**32%**	**42%**	**30%**	**50%**	**37%**
Enrolled Only	2%	1%	1%	1%	3%	3%
Total Found	**45%**	**33%**	**43%**	**31%**	**53%**	**40%**

Professional Recipient Match Rates

Professional	System		Academic		Health	
	1st	10th	1st	10th	1st	10th
Working in TX, FY & FT*	50%	49%	53%	46%	48%	51%
Working in TX, Other	13%	6%	15%	6%	11%	6%
Working in TX Total	**63%**	**54%**	**68%**	**52%**	**59%**	**57%**
Enrolled Only	1%	1%	1%	1%	1%	1%
Total Found	**64%**	**56%**	**69%**	**53%**	**60%**	**58%**

Created the seekUT tool

- Built in SAS Visual Analytics

- Developed for students & families
- Open to the public

Created the seekUT tool

- Contains earnings information & more
  - 1, 5, and 10 yr earnings after graduation
  - Data on debt
  - Time-to-degree
  - Percent of grads continuing their education
  - Job Projections

seekUT www.utsystem.edu/seekUT

Productivity Dashboard: data.utsystem.edu

Explore More Data Visualizations: exploredata.utsystem.edu

OSI website: www.utsystem.edu/offices/strategic-initiatives

OSI Blog: https://utfactsonline.wordpress.com/

Follow us on Twitter: @UTFactsOnline

Questions & Comments

Jessica Shedd

Office of Strategic Initiatives
The University of Texas System

jshedd@utsystem.edu