



Identifying Duplicate Customers

Jim Harris
Blogger-in-Chief
www.ocdqblog.com



Jim Harris

Blogger-in-Chief
www.ocdqblog.com

Email

jim.harris@ocdqblog.com

Twitter

twitter.com/ocdqblog

LinkedIn

[linkedin.com/in/jimharris](https://www.linkedin.com/in/jimharris)





Introduction

This will be a vendor-neutral presentation:

- Focusing on the common data quality challenge of identifying duplicate customers
- Discussing why a robust methodology, not technology alone, should form the solution
- Using an interactive review of data to demonstrate the highly subjective nature of this problem



A Common Problem

- A common and challenging data quality problem is the identification of duplicate records, especially:
 - Redundant representations of the same customer
 - Within and across systems throughout the enterprise
- A solution to this specific problem is a primary reason to invest in data quality software and services



With Many Viable Solutions

- Many data quality vendors to choose from and all of them offer viable data matching solutions
- Many solutions are driven by impressive technology using advanced mathematical techniques:
 - Deterministic Record Linkage
 - Probabilistic Record Linkage
 - Fellegi-Sunter algorithm
 - Bayesian statistics
 - Bipartite Graphs
 - or my personal favorite...



The Redundant Data Capacitor

Makes identifying
duplicates possible
using only:

- 1.21 gigawatts
- DeLorean DMC-12
- 88 miles per hour





A Business Problem

- What is sometimes overlooked is that although technology provides the solution...
- ...what is being solved is a *Business Problem*
- Technology can carry with it a dangerous conceit:
 - The laboratory and engineering department are similar to the boardroom and accounting department
 - The mathematician and computer scientist think like the business analyst and data steward



Business Rules

- What truly determines that a duplicate customer has been identified is NOT what can be justified by:
 - Scientific techniques
 - Mathematical models
- Duplicate customers ARE identified by:
 - Your business rules definitions



Technology *and* Methodology

- Symbiosis of technology and methodology is necessary when approaching this common data quality problem
- Effective methodology will help maximize the time and effort as well as the subsequent return...
- ...on whatever technology you invest in



A Highly Subjective Problem

- Data characteristics and associated quality challenges are unique from company to company
- Data quality can be defined differently by different functional areas within the same company
- Business rules can be different from project to project within the same company
- Decision makers on the same project can have widely varying perspectives



Defining “Duplicate Customer”

- Important to define the term “duplicate customer”
- Without methodology, it can be frustratingly difficult
- Common challenges include:
 - “A duplicate customer is a duplicate customer”
 - “Data denial”
 - Conservative concerns



“A duplicate is a duplicate”

- A common definition states that a duplicate customer occurs when the *exact same information* is repeated
- Exact same information usually defined as all required attributes are populated with the same value
- Definition can be expressed as passive-aggressive doubt that such a problem could be very prevalent (i.e. “data denial”)



Self-Defense *not* “Data Denial”

- “Data Denial” is not necessarily a matter of blissful ignorance or a fear to confront the truth
- More often, it is a natural self-defense mechanism
- Neither data owners nor process owners want to be blamed for causing or failing to fix the quality problem
- This human factor must be considered because it is the people involved that truly make projects successful



A Common Concern

- Aggressively identifying duplicates can negatively impact business decisions after consolidation
 - Either physical removal or logical linkage of duplicates
- This common concern will often influence a cautious approach to duplicate identification
- Motivated by the fact that there is generally far greater concern about “false positives” than “false negatives”

The Two Headed Monster

False Positives

- Identified duplicates that do NOT represent the same customer

False Negatives

- Redundant representations of the same customer that are NOT identified

- Primary business problem is the reduction of false negatives
 - Reducing false negatives risks creating false positives
- False negatives and false positives can never be eliminated



Not A Theoretical Problem

- The duplication of customer information is:
 - NOT a theoretical problem
 - IS a real business problem negatively impacting the quality of decision-critical enterprise information
- Data-driven problems require data-driven solutions



Preliminary Data Analysis

- Business rules are best illustrated by *data examples*:
 - Real data that *exemplifies* the problem
 - Not *data metaphors* that meaningfully demonstrate the problem but are nonetheless *fictional*
- Before the requirements gathering phase:
 - Perform preliminary analysis on representative samples from the project's actual data sources
 - This preparation will enable a far more productive discussion of the business rules



Data Metaphors

- We will look at fictional examples that illustrate:
 - The importance of real data analysis
 - The highly subjective nature of this problem
- For simplicity, three customer attributes will be used:
 1. Customer Name – only personal names
 2. Postal Address – only United States address formats
 3. Tax ID – for better fictional values, dates related to the customer name have been used

False Negatives

False negatives can be caused when conservative concerns motivate a cautious approach to duplicate identification.

Leading many projects to adopt an exact match strategy...

Exact Matching

Key	Customer Name	Postal Address	Tax ID
111	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
112	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
121	Tek Jansen	513 West 54th Street, New York, NY 10019	10262005
122	Tek Jansen	513 West 54th Street, New York, NY 10019	10262005
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
141	Joseph Heller	22 Catch Circle, Washington, DC 20004	19230501
142	Joseph Heller	22 Catch Circle, Washington, DC 20004	19230501
151	Franz Joseph Haydn	94 Surprise Symphony Street, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Street, Austria, CO 80467	17321809

An Important Choice

Take the **Blue Pill**

- Implement an exact match strategy and some duplicate customers will be identified

Take the **Red Pill**

- Stay in Wonderland and be shown just how deep the rabbit hole goes...



Back to the Future...

Key	Customer Name	Postal Address	Tax ID
111	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
112	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
113	Marty Calvin McFly	Lone Pine Mall, Hill Valley, CA 94941	11121955
114	McFly, Marty	Lone Pine Mall, Hill Valley, CA 94941	
115	McFly	Hill Valley, California	

Abbreviation Aggravation

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	
134	T.S. Eliot	Wasteland, Texas	26091888
211	J.D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
212	Jerome D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	
213	J. David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
214	Jerome Salinger	Holden Caulfield Highway, Agerstown, PA 19102	
215	David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951

Who is T. Kundera?

Key	Customer Name	Postal Address	Tax ID
221	Tomas Kundera	245 Daniel Day-Lewis Drive, Kitch, NY 10022	19681985
222	Thomas Kundera	245 Daniel Day Louis Drive, Kitch, NY 10022	
223	T. Kundera	Daniel Day Lewis Drive, Kitch, NY 10022	19681985
231	Tereza Kundera	245 Daniel Day-Lewis Drive, Kitch, NY 10022	20061988
232	Teresa Kundera	245 Daniel Day Louis Drive, Kitch, NY 10022	
233	T. Kundera	Daniel Day Lewis Drive, Kitch, NY 10022	20061988

Key	Customer Name	Postal Address	Tax ID
223	T. Kundera	Daniel Day Lewis Drive, Kitch, NY 10022	
233	T. Kundera	Daniel Day Lewis Drive, Kitch, NY 10022	



Name and address variations

Key	Customer Name	Postal Address	Tax ID
151	Franz Joseph Haydn	94 Surprise Symphony Street, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Street, Austria, CO 80467	17321809
153	Joseph Haydn	45 Farewell Symphony Street, Austria, CO 80467	
154	Papa Haydn	Austria, CO 80467	17321809
241	Emmett Lathrop Brown	1640 John F. Kennedy Drive, Hill Valley, CA 94942	11051955
242	Brown Emit Latrop	1640 Riverside Drive, Hill Valley, CA 94942	
243	Doc Brown	Hill Valley, California	11501955



Marriage, good for people, bad for data

Key	Customer Name	Postal Address	Tax ID
251	Lorraine Baines	55 Enchantment Sea Park, Hill Valley, CA 94941	19382015
252	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	
253	Lorraine McFly	Hill Valley, California	19382015
261	Elinor Frost	1 Road Not Taken, Derry, NH 03038	18961938
262	Eleanor Rost	1 Road Not Taken, Derry, NH 03038	
263	Mrs. Robert Frost	122 Rockingham Road, Derry, NH 03038	18961938



To Dupe or Not To Dupe...

Key	Customer Name	Postal Address	Tax ID
271	William Shakespeare	Globe Theatre, Stratford-upon-Avon, NH 03576	26041564
272	Christopher Marlowe	Globe Theatre, Stratford-upon-Avon, NH 03576	26041564
281	Samuel Langhorne Clemens	11 Hucklebery Finn River, Tom Sawyer, MS 38967	30111835
282	Mark Twain	11 Hucklebery Finn River, Tom Sawyer, MS 38967	30111835

False Positives

Business rule adjustments for preventing false negatives can result in linking related (not duplicated) customers.

False positives may reveal meaningful data relationships that are useful in other enterprise information initiatives.

The very model of a false positive

Key	Customer Name	Postal Address	Tax ID
311	Gilbert A. Sullivan	571 H.M.S. Pinafore Plaza, Mikado, MI 48738	18711896
312	W.S. Gilbert	363 Pirates of Penzance, Patience, PA 19114	18711896
313	Arthur Sullivan	246 Princess Ida Island, Ruddigore, RI 02841	18711896
321	Abbott and Costello	First Baseman Boulevard, Cooperstown, NY 13326	
322	Bud Abbott	First Baseman Boulevard, Cooperstown, NY 13326	
323	Lou Costello	First Baseman Boulevard, Cooperstown, NY 13326	
331	David Starsky	93 Striped Tomato Ford, Bay City, CA 90731	19751979
332	Kenneth Hutchinson	93 Striped Tomato Ford, Bay City, CA 90731	19751979
333	Huggy Bear Brown	93 Striped Tomato Ford, Bay City, CA 90731	19751979

Family Data – Blessing or Curse?

Key	Customer Name	Postal Address	Tax ID
411	Elizabeth Barrett Browning	43 Portuguese Sonnets, Counting Ways, IL 62650	18061861
412	Robert Browning	43 Portuguese Sonnets, Counting Ways, IL 62650	18061861
413	Robert Barrett Browning	43 Portuguese Sonnets, Counting Ways, IL 62650	18061861
421	Felix Mendelssohn	46 East Elijah Expressway, Oratorio, OR 97289	18090203
422	Abraham Mendelssohn	46 East Elijah Expressway, Oratorio, OR 97289	18090203
423	Moses Mendelssohn	46 East Elijah Expressway, Oratorio, OR 97289	18090203
431	Horatio Alger Jr.	One American Dream Avenue, Revere, MA 02151	
432	Horatio Alger Sr.	One American Dream Avenue, Revere, MA 02151	
433	Horatio Alger	One American Dream Avenue, Revere, MA 02151	
441	Patrick Thames	1831 London Bridge, Lake Havasu City, AZ 86403	
442	Patricia Thames	1831 London Bridge, Lake Havasu City, AZ 86403	
443	Pat Thames	1831 London Bridge, Lake Havasu City, AZ 86403	

Sharing Data – Family Style

Key	Customer Name	Postal Address	Tax ID
511	Peter and Lois Griffin	31 Spooner Street, Quahog, RI 02903	19990131
512	Lois and Stewie Griffin	31 Spooner Street, Quahog, RI 02903	20020214
513	Stewie and Brian Griffin	31 Spooner Street, Quahog, RI 02903	20050501

Key	Customer Name	Postal Address	Tax ID
511-a	Peter Griffin	31 Spooner Street, Quahog, RI 02903	19990131
511-b	Lois Griffin	31 Spooner Street, Quahog, RI 02903	19990131
512-a	Lois Griffin	31 Spooner Street, Quahog, RI 02903	20020214
512-b	Stewie Griffin	31 Spooner Street, Quahog, RI 02903	20020214
513-a	Stewie Griffin	31 Spooner Street, Quahog, RI 02903	20050501
513-b	Brian Griffin	31 Spooner Street, Quahog, RI 02903	20050501



Family Household

- Keys 411 – 513 were also examples of a non-duplicate data relationship referred to as a *family household*
- Where multiple distinct customers are linked for having the same family name and postal address
- Useful in marketing programs targeting:
 - Family units (e.g. vacation packages, phone plans)
 - Head of household (i.e. purchasing decision-makers)

Seven Smiths on Main Street

Key	Customer Name	Postal Address	Tax ID
611	Agent Smith	1999 Main Street, Toute Ville City, Irgendein Land	
612	Dudley Smith	1987 Main Street, Toute Ville City, Irgendein Land	
613	Jefferson Smith	1939 Main Street, Toute Ville City, Irgendein Land	
614	Hannibal Smith	1983 Main Street, Toute Ville City, Irgendein Land	
615	Winston Smith	1984 Main Street, Toute Ville City, Irgendein Land	
616	Sarah Jane Smith	1973 Main Street, Toute Ville City, Irgendein Land	
617	Doctor Zachary Smith	1965 Main Street, Toute Ville City, Irgendein Land	

Many People, One Address

Key	Customer Name	Postal Address	Tax ID
711	Shawn Spencer	Pineapple Apartments, Santa Barbara, CA 93121	
712	Burton Guster	Pineapple Apartments, Santa Barbara, CA 93121	
721	F. Scott Fitzgerald	Literary Luxury Lofts, Littera, LA 70116	
722	James Joyce	Literary Luxury Lofts, Littera, LA 70116	
723	Jay Gatsby	Literary Luxury Lofts, Littera, LA 70116	
724	Stephen Dedalus	Literary Luxury Lofts, Littera, LA 70116	
731	Leonard Hofstadter, Ph.D.	CALTECH, Pasadena, CA 91125	
732	Sheldon Cooper, Ph.D.	CALTECH, Pasadena, CA 91125	
733	Rajnish Koothrappali, Ph.D.	CALTECH, Pasadena, CA 91125	
741	Jack Carter	Global Dynamics, Eureka, OR 97086	
742	Allison Blake	Global Dynamics, Eureka, OR 97086	



Geographic Household

- Keys 711 – 742 (as well as Keys 321 – 513) were also examples of a non-duplicate data relationship referred to as a *geographic household*
- Where multiple distinct customers are linked for having the same postal address
- Useful in mass mailing programs to eliminate costs of redundant deliveries to the same postal address

Best Practices

- Business Rule Documentation
- Application Development Expectations
- Business-IT Collaboration



Business Rule Documentation

- The goal of a business requirements document (BRD) is to provide clear definitions for both:
 - Business problem statements
 - Associated solution criteria
- Include data examples because they convey business rules far better than either:
 - Concise (but esoteric) statements
 - Detailed (but verbose) pages of attempted explanation



Accentuate the Negative

- Although it may sound counterintuitive, it is simply easier to explain something when you don't like it
- This effect is known as “negativity bias” where bad evokes a stronger reaction than good in the human mind – just compare an insult and a compliment, which one do you remember more often?
- Therefore, focus on documenting the rules that identify what is NOT a duplicate customer



Avoid Technology Bias

- Knowing how your vendor's software works can cause a “framing effect” where rules are defined in terms of software functionality, framing them as a technical problem instead of a business problem
- All data quality vendors have viable data matching solutions driven by impressive technology
- Therefore, focus on stating both the problem and solution criteria in business terms



Development Expectations

- Many data quality initiatives fail because of lofty expectations, unmanaged scope creep, and the unrealistic perspective that problems can be permanently “fixed”
- In order to be successful, application development must always be understood as an iterative process
- ROI will be achieved by targeting well defined objectives that can deliver small incremental returns that will build momentum to larger success over time
- Projects are easy to get started, even easier to end in failure and often lack the decency of at least failing quickly



Focus on the Data

- Every vendor's software ranks match results as a primary method to differentiate the three possible result categories:
 1. Automatic Matches
 2. Automatic Non-Matches
 3. Potential Matches requiring manual review
- Reviewers sometimes focus on the ranking and ignore whether or not records have been properly categorized
- Trending analysis should be performed to measure the effects caused by modifying the matching criteria



Perfection is Impossible

- Always remember that the harsh reality is false negatives and false positives can be reduced, but never eliminated
- For example, imagine having only 100 exceptions to review out of one billion records (i.e. 99.99999% success rate)
- Do not underestimate the difficulty that the human mind has with large numbers or forget about “negativity bias”
- Focusing on exceptions can undermine confidence and prevent acceptance of a successful implementation



Business-IT Collaboration

- Data Quality is neither an IT issue nor a Business issue
Data Quality is everyone's issue
- Neither the Business nor IT alone has all of the necessary knowledge and resources required for success
- Successful projects are driven by an executive management mandate for the Business and IT to forge an ongoing collaboration



Provide Leadership

- An executive sponsor provides oversight and arbitrates any issues of organization politics
- Both the Business and IT must also designate a team leader for the initiative – choose these leaders wisely
- Effective Leaders:
 - Know how to listen well
 - Foster open communication without bias
 - Seek mutual understanding on difficult issues
 - Truly believe it is the people that make projects successful

Duplicate Consolidation

- Techniques for creating a “best of breed” record
- Physical Removal vs. Logical Linkage
- Consolidation vs. Cross Population

Creating a “Best of Breed” Record

- Consolidation evaluates groups of identified duplicate records and creates one “best of breed” representative record for each group
- Typically, consolidation creates the representative (a.k.a. “survivor”) record using one of two techniques:
 1. Record Level Consolidation – choosing one complete record from within the group
 2. Field Level Consolidation – constructing fields from potentially different records from within the group



Consolidation Selection Criteria

- Completeness – selecting the record with the highest number of populated fields or the longest value in a field
- Frequency – selecting most frequently occurring field value or record with most frequently occurring values across fields
- Recency – selecting the record most recently updated or the field value from the record most recently updated
- Source – selecting the record or a field value that originated in a preferred source system

Record Level Consolidation

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	
134	T.S. Eliot	Wasteland, Texas	26091888

Select a record based on the most frequently occurring values across fields

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888

Field Level Consolidation

Key	Customer Name	Postal Address	Tax ID
211	J.D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
212	Jerome D. Salingr	Holden Caulfield Highway, Agerstown, PA 19102	
213	J. David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
214	Jerome Salnger	Holden Caulfield Highway, Agerstown, PA 19102	
215	David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951

Construct a record based on the most complete (longest value) in a field

Key	Customer Name	Postal Address	Tax ID
211	Jerome David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951



“Frankenstein Consolidation”

- Field level consolidation is sometimes referred to as “Frankenstein consolidation”
- Constructing fields from different records within the group can assemble an “unnatural data monster” by creating an invalid combination of field values
- This concern typically makes record level consolidation the far more common consolidation technique



Physical Removal vs. Logical Linkage

- Consolidation of the group of identified duplicate records is implemented using one of two techniques:
 1. Physical Removal – group is replaced with the representative record and duplicates are either deleted from the source or simply excluded from the target
 - Most commonly uses record level consolidation
 2. Logical Linkage – group is updated with an identifier field whose value references the identifier field value of the representative record



Consolidation vs. Cross Population

- An alternative strategy is cross population, where the representative record is used to update the group
- Most commonly uses field level consolidation and is implemented using one of two techniques:
 1. Fill in the blanks – update only unpopulated fields
 2. Create consistent values – update all fields

Fill in the Blanks

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	
134	T.S. Eliot	Wasteland, Texas	26091888

Select a record based on the most frequently occurring values across fields

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888

Update only unpopulated fields

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
134	T.S. Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888

Create Consistent Values

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	
134	T.S. Eliot	Wasteland, Texas	26091888

Select a record based on the most frequently occurring values across fields

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888

Update all fields

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
134	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888



Conclusion

- Perform a preliminary data analysis to provide real examples for productive discussion of business rules
- Anticipate the reality of both false negatives and false positives during duplicate identification
- Understand that projects require a holistic approach involving people, methodology, and technology
- Stay focused on the data, but do not let the pursuit of perfection undermine a successful implementation