How Best In Class Fuzzy Matching Solutions Work: Combining Established and Proprietary Algorithms
# Table of Content

**The Underlying Challenges with Data**

- Data Matching for Core Business Activities
  - Case Study 1 – State of Nevada Uses SLDS to Make Data-Driven Decisions to Improve Student Achievement
  - But what about unique identifiers, you ask?

**Data Ladder – Making Use of Fuzzy Matching & Proprietary Algorithms to Get 96% Matching Accuracy at Unrivaled Speed**

- How Does Fuzzy Logic Work?
- The Problem with False Negatives in Fuzzy Matching
  - Case Study 2 – Zurich Insurance Uses Fuzzy Matching to Get Accurate Reporting
  - Case Study 3 – St. John Associates Merge Patient Records Saving Hundreds of Man-Hours Annually
- Fuzzy Matching Techniques
- Example of a Real-World Fuzzy Matching Scenario

**DataMatch Enterprise - A Complete Data Matching Framework**

- Data Integration
- Data Profiling
- Data Cleansing
- Data Matching
- On-Premises Solution
- Address Verification
- Final Export

**Conclusion**
How Best In Class Fuzzy Matching Solutions Work: Combining Established and Proprietary Algorithms

Data is no longer just records of customer names, contact number or addresses. It’s a combination of information obtained from multiple sources and it’s mostly flawed, duplicated, inaccurate or incomplete.

Companies need the best-in-class tools to process this data and make sense out of it. Unfortunately, most companies do not have the right tools or processes in place to sort data. Even if they do have a team of data experts, the process of sorting data and ensuring it meets quality standards is a slow, often frustrating one. Raw data is complex beyond measure and requires a combination of algorithms and methodologies to handle its inherently complex nature.

Things become more complicated when organizations have to “match” data from multiple sources to:

- Remove duplicates
- Combine information
- Get a single source of truth
- Get a 360-degree customer view
- Get insights for reporting and analysis

While data matching used to be an easy activity, over the years, the type, format, and complexity of raw data has changed. Today it’s no longer enough to run a match between two fields with similar spellings to weed out duplication, nor is it enough to deploy one kind of algorithm to find matches. Today, the best-in-class data matching tools use a combination of algorithms + proprietary algorithms to weed out even the most difficult data fields to help companies achieve their data matching objectives.

This white paper will explore the challenges of matching, how different types of matching algorithms, how a best-in-class software uses these algorithms to achieve data matching goals.

Let’s dive in.
The Underlying Challenges with Data

Companies large or small collect data for a range of purposes, each purpose unique to the company’s business objectives. That said, there are also universal purposes for data collection, including:

- Understanding & meeting customer expectations
- Providing a modern, digital customer experience
- Making key business operations (marketing, sales, branding etc)
- Generating reports, insights, and studies

Data collection is not limited to private businesses or companies. Government organizations collect data for administering multiple financial programs, administering regulatory requirements or producing economic statistics. For the healthcare sector, data collection could mean producing new, groundbreaking research that could potentially impact funding, treatments and much more. Data collection is the core operation of every business today.

While this data collection is valuable and extremely necessary, it is often rife with duplicative information and errors. For example, a customer may have provided their information to your company twice, each time, using a different variation of their name.

If they filled a web form, they may have given nicknames or abbreviations instead of the complete name. If they are filling information for billing or logistics, they are more likely to give their complete name instead of their nickname. This small, seemingly inconsequential data input pattern has created a duplicate of that record in your database. Now multiply this incident by 1000X times and you have a data duplication crisis on your hands.

Government organizations have it even tougher when it comes to data collection and its issues. It’s not uncommon for different state departments to collect the same information and having to share their multi-unit data for the distribution of federal funds through grant programs. The Bureau of Labor Statistics and The Census Bureau, for example, share data on monthly employment and quarterly wage data for all establishments covered under the one Unemployment Insurance (UI). Based on this information, employment and wage data are classified by industry and by location. The Census Bureau gathers the same information to maintain its Business Register.
The two departments frequently share their data to improve comparability, but with duplicate data, and the lack of a unique identifier, it becomes a challenge for these departments to ensure the accuracy of their data.

It is currently difficult for public and private organizations to access their data and use the information for its intended purpose. In order for data to be useful, business users, researchers, data experts, and key decision-makers must be able to match data within the same data source and across different data sources.

The catch? They must be able to perform complex data matching and still get accurate results, fast, hassle-free.

**Data Matching for Core Business Activities**

Data matching is not a new concept. For decades now, companies have been using various algorithms and processes to deal with duplication. But in the world of big data, data matching has taken on a whole new role and is beyond the simple function of matching to find for duplicates. Today, data matching is used to fulfill fundamental, complex activities like:

**Finding Data and Deduplicating Data Within a Single Data Set:** At the most basic level, companies need to derive insight and remove duplicated data from within a single data set. For example, an insurance company may want to check for duplicates within their transactional data set or derive some key insight from the same data.

The problem is, basic data fields like the Name, Phone or Address fields are the most prone to human error. A customer’s name may be written as Jamie Smith or J.Smith, not to mention the same customer may have had their data recorded multiple times in the database especially if they have moved frequently, changed their numbers or go by multiple name variations. At this time, a powerful fuzzy data matching tool is required to delete entries based on probability and entries that are reasonably close. Deterministic matching (matching between exact entities) is not effective.

We’ll talk more about fuzzy matching in a while.
Aggregate Data to Get a Single Source of Truth: This is an objective that is currently driving most businesses to search for a data matching solution. More than 80% of companies want to make use of data matching to combine data from multiple sources and get a complete or a single view of truth. To achieve this, multiple data sets are matched, redundancies are removed, and fields are merged to get a consolidated view.

The critical challenge here is the duplication of an entity’s data across these multiple channels. For example, a user’s name may be entered as J.D Smith in a marketing CRM as taken from their LinkedIn account. Now, if the company wants to know the number of purchases J.D Smith has made for a certain product, they will have to check the billing department’s database.

Just for this simple insight, two data sets, siloed away will be combined to get a complete view. There is no guarantee though, that in billing, J.D Smith will be written in the same way. In fact, the company finds a contact called Jonathan D. Smith in billing data. This kind of data variation is what makes it so complex to perform data matching and obtain a single source of truth.

Linking & Matching for Statistical Purposes: Government organizations often need linking and matching for statistical purposes, wherein data sources are integrated to augment one source with fields not initially contained with the source, such as an entity’s educational background, employment details or other relevant data fields. Another aspect of link matching here is the need to merge different data sets and sources to get a complete picture of an entity in the industry.
Case Study 1 – State of Nevada Uses SLDS to Make Data-Driven Decisions to Improve Student Achievement

With over 55 million students in grades K-12 this year, administrators need to make sense of what really impacts student achievement?

With information gathered from statewide P-20 longitudinal data systems (SLDS), educators and administrators will soon have these answers.

The introduction of Common Core State Standards (CSSS) has brought educational leaders and policymakers together with the need to implement an integrated and comprehensive system that will drive student performance and achievement. Big data is driving this effort forward.

To understand the value of this data, record linkage tools are fundamental in evaluating programs for schools. By increasing the matching accuracy of the data available, it becomes much more useful, and decision making improves.

With a record linkage program in place, the Department of Education, Nevada conducted a sample, evaluating the number of students in one year who attended post-secondary education in a specific city. With the old existing program, the sample found that 22 percent of the 5,344 students in that city had gone on to higher education. After using our record linkage software, that number went up to nearly 41 percent, nearly double the first figure! The chart below compares this sample to a city of comparable numbers (5,025 students) with no new program in place.

Imagine the effects that improved match accuracy can have when comparing schools in two cities. Just a small improvement can make the difference in everything from property values to policymaking in those areas.
But what about unique identifiers, you ask?

Databases have unique identifiers that are unlikely to change over time, such as the phone number field which is unique for all. That said, unique identifiers are not without flaws. A phone number may consist of text characters or punctuations that make it difficult for deterministic matching.

If some phone numbers have a country code in the +1 format and the others start with city code or with 001, then the matching would not be possible.

When it comes to government organizations, they have an even greater challenge – they are not at the luxury of using unique identifiers. When multi-unit data is shared, unique identifiers are not shared, making the task of matching even more challenging.

Another critical challenge is cultural nuances when it comes to names and other relevant data.

An Arabic name like Abdul Rashid may also be written like Abdul Rasheed – both of which are completely different names. A deterministic match will not work here because these algorithms take into account only the English alphabet and recognize English names. A probabilistic match may only work if other details with this name are similar. In this scenario, you may need a phonetic or sound-based matching algorithm.

It’s a challenge for data experts to create or use multiple combinations of algorithms to match millions of data rows from dozens of sources. Not only such an undertaking complex in nature, but it is also highly time-consuming – not to mention, the accuracy of such manual matching endeavors often misses the mark.

With the level of complex data, we’re dealing with, a tool is required – one that uses multiple combinations of algorithms and their own proprietary algorithm to analyze and match data without sacrificing speed and accuracy.
Data Ladder - Making Use of Fuzzy Matching & Proprietary Algorithms to Get 96% Matching Accuracy at Unrivaled Speed

A powerful solution, DataMatch Enterprise, has been used by 4,500 clients worldwide. From HP to Deloitte, DHL to Arlington, we’ve worked with industry-leading retailers, manufacturers, insurers, banks and many others with best-in-class data matching. Moreover, we’ve also catered to the U.S. Department of Transportation and the USDA to provide data matching that meets grant funding requirements.

Data Ladder makes use of fuzzy matching and its own proprietary algorithms to handle complex data matching tasks. Our clients have used our data matching solution to:

- Match data within data sources
- Match data across multiple data sets
- Consolidate data to get a single source of truth
- Perform data deduplication and cleansing
- Implement data governance
- Identify and fix data quality issues
- Automate data cleansing

Let’s understand fuzzy matching and how it is used to identify records that deterministic matching would miss.

How Does Fuzzy Logic Work?

Unlike deterministic matching that simply flags records as a ‘match’ or ‘non-match’ fuzzy matching identifies the likelihood that two records are a true match based on whether they agree or disagree on the various identifiers.

Consider the strings “Kent” and “10th”. While there is clearly no match here, a fuzzy matching algorithm will still rate these two strings nearly 50% similar, based on character count and phonetic match.
Our system identifies:
- Acronyms
- Name reversal
- Name variations
- Phonetic spellings
- Deliberate misspellings
- Inadvertent misspellings
- Abbreviations e.g. ‘Ltd’ instead of ‘Limited’
- Insertion/removal of punctuation, spaces, special characters
- Different spelling of names e.g. ‘Elisabeth’ or ‘Elizabeth’, ‘Jon’ instead of ‘John’
- Shortened names e.g. ‘Elizabeth’ matches with ‘Betty’, ‘Beth’, ‘Elisa’, ‘Elsa’, ‘Beth’ etc.

And many other variations.

The Problem with False Negatives in Fuzzy Matching

False negatives refer to matches that are missed altogether by the system: not just a low match score, but an absence of match score. This leads to a serious risk for the business as false negatives are never reviewed because no one knows they exist. Factors that commonly lead to false negatives include:

- Lack of relevant data
- Significant errors in data entry
- System limitations
- Match criterion is too narrow
- Inappropriate level of fuzzy matching

The most effective method to minimize both false positives and negatives is to profile and clean the data sources separately before you conduct matching.

For this very reason, DataMatch Enterprise comes with data profiling and data cleaning abilities that profile your data and allows you to match, clean, dedupe or standardize data “before” matching.

Let’s explain this with a use case.
Case Study 2 – Zurich Insurance Uses Fuzzy Matching to Get Accurate Reporting

Zurich Insurance Group Ltd, commonly known as Zurich is Switzerland’s largest insurer. As of 2017, the group is the world’s 91st largest public company according to Forbes’ Global 2000s list, and in 2011, it ranked 94th in Interbrand’s top 100 brands.

It is organized into three core business segments: General Insurance, Global Life and Farmers. Zurich employs almost 54,000 people serving customers in more than 170 countries and territories around the globe. The company is listed on the SIX Swiss Exchange. As of 2012, it had shareholders’ equity of $34.494 billion.

In the insurance industry, having payee names aggregate and match are critical for the functioning of various payment processes. Zurich Insurance’s current system does not have a hard edit function where payee names can be pre-populated so those managing and entering information in the database can just key in any type of information.

If any query was run against the main data warehouse, a long list of duplicate information would pop up. The result? Vendor names were not being aggregated appropriately, causing massive headaches and operational inefficiency.

Using Data Match Enterprise, the company was able to use fuzzy matching to reconcile the payee names. After using the data cleansing portion of the software to remove all special characters and spaces, a unique identifier was used for each record in question after which, all of payee names could be replaced with a solid vendor name.

From a strategic perspective, fuzzy matching comes into play when you’re conducting record linkage or entity resolution.

Here are some ways that fuzzy matching is used to improve the bottom-line:

- Realize a single customer view
- Work with clean data you can trust
- Prepare data for business intelligence
- Enhance the accuracy of your data for operational efficiency
- Enrich data for deeper insights
- Ensure better compliance
- Refine customer segmentation
- Improve fraud prevention
Case Study 3 – St. John Associates Merge Patient Records Saving Hundreds of Man-Hours Annually

St. Johns Associates provides placement and recruiting services in Cardiology, Emergency Medicine, Gastroenterology, Neurological Surgery, Neurology, Orthopedic Surgery, PM&R/Pain Medicine, Psychiatry, Pulmonary/Critical Care, Surgical Specialties, and Urology.

With a growing database of recruitment candidates, St. John Associates was in need of a way to dedupe, clean and match records. After a number of years of performing this task manually, the company decided it was time to deploy a tool to reduce the time spent on cleaning records.

After having looked at several competitive solutions, St. John Associates settled on DataMatch Enterprise™ as their tool of choice. A key factor in the decision was the ease of use. Competitor products that they tested were found to be too complex or too cumbersome for an end-user to use in an efficient way.

Using DataMatch Enterprise, the organization was able to automate data cleaning and matching, saving manual labor hours, ensuring accuracy and achieving their data linkage goals – which in the case of healthcare could be a matter of life and death.

Fuzzy Matching Techniques

Now you know what fuzzy matching is and the many different ways you can use it to grow your business. Question is, how do you go about implementing fuzzy matching processes in your organization?

Here’s a list of the various fuzzy matching techniques that we use today:

- Levenshtein Distance (or Edit Distance)
- Damerau-Levenshtein Distance
- Jaro-Winkler Distance
- Keyboard Distance
- Kullback-Leibler Distance
- Jaccard Index
- Metaphone 3
- Name Variant
- Syllable Alignment
- Acronym
Example of a Real-World Fuzzy Matching Scenario

The following example shows how record linkage techniques can be used to detect fraud, waste or abuse of federal government programs. Here, two databases were merged to get information not previously available from a single database.

A database consisting of records on 40,000 airplane pilots licensed by the U.S. Federal Aviation Administration (FAA) and residing in Northern California was matched to a database consisting of individuals receiving disability payments from the Social Security Administration. Forty pilots whose records turned up on both databases were arrested.

A prosecutor in the U.S. Attorney’s Office in Fresno, California stated, according to an AP report: “There was probably criminal wrongdoing.” The pilots were either lying to the FAA or wrongfully receiving benefits. The pilots claimed to be medically fit to fly airplanes. However, they may have been flying with debilitating illnesses that should have kept them grounded, ranging from schizophrenia and bipolar disorder to drug and alcohol addiction and heart conditions.”

At least twelve of these individuals “had commercial or airline transport licenses,” the report stated. The FAA revoked 14 pilots’ licenses. The other pilots were found to be lying about having illnesses in order to collect Social Security payments.

The quality of the linkage of the files was highly dependent on the quality of the names and addresses of the licensed pilots within both of the files being linked. The detection of the fraud was also dependent on the completeness and accuracy of the information in a particular Social Security Administration database.
Based on decades of research and 4,000+ deployments across more than 40 countries, DataMatch Enterprise is a highly visual data cleansing application specifically designed to resolve data quality issues. The platform leverages multiple proprietary and standard algorithms to identify phonetic, fuzzy, miskeyed, abbreviated, and domain-specific variations.

Here’s a sneak peek into the software’s functions and features:

As you can see, DataMatch Enterprise offers a complete data quality and data matching framework. With this solution you get:

**Data Integration**

As businesses continue to use new CRMs and apps, they will want a quality solution that connects directly to the data source.

DataMatch Enterprise (DME) allows for easy integration with 150+ apps. This means businesses can directly clean, edit, match and fix data quality issues right within their data source without having to worry about importing or moving data. You can integrate:

- Files (Excel, Delimited Text File, Excel (97-2003), Fixed Width Text File)
- Databases (Direct integration with SQL Server, Oracle, Teradata, and many more)
- CRM (Integration with Salesforce, MS Dynamics CRM, Sugar CRM)
- Social (Integration with Facebook and Twitter)
- Services (XML and JSON)
- Email apps
- Other (ODBC-supported databases)
Data Profiling

The primary goal of data profiling is to provide an overview of the problems within the data set. For example, incomplete addresses, invalid phone numbers, ZIP codes with letters etc are data quality issues that are highlighted during the profiling process.

DataMatch Enterprise (DME) lets you:

- Identify which of your lists have irregularities or oddities
- Gives a confidence score to each of your fields
- Statistics on your dataset’s size, median, mean, average, standard deviation, etc
- Tailor-made expressions that let you sort data over 19 different fields
- Create custom expressions using a Pattern Builder option
- View all your profile history changes
- Visual data exploration to help you see your audience clusters
- Identify nicknames from a pre-defined dictionary of nicknames

Data Cleansing

Data fields with inaccurate spellings, case settings, and other structural problems need to be cleansed by fixing the inconsistencies. Ideally, this needs to be done right within the data source so that errors are nipped in the bud. DME has a full-fledged data cleansing and standardization system that allows users to:

- Change cases and ensure that each field follows a defined standardization
- Find and replace different characters
- Choose whether you want to preserve abbreviations
- Set your own exceptions
- Choose delimiter that separates your fields such as tabs, comma, semicolons etc
- Remove spaces, numbers, letters, non-printable characters, zeros and zeroes with letters

DME also has a unique WordSmith tool that allows you to standardize names and addresses. This is especially useful if you want to ensure that a certain word is automatically replaced or removed in a list.
Data Matching

DataMatch Enterprise’s data matching allows you to:

- Define how you want to match data sources using four options – all, none, between and within. All looks for matches between each data source as well as matches within each data source. If you don’t want to do all, you can just choose between or within to look for duplicates between data sources or within data sources.
- Choose the fields you want to map by the fields of your data list.
- Run multiple matching sessions simultaneously.
- Get matches based on exact, phonetic, numeric and fuzzy matching techniques.
- View match results and create a Master Record that you can use as a surviving record.
- Save and export your results for later review.
- Use a merge and survivorship option to overwrite data based on your criteria.

On-Premises Solution:

DataMatch Enterprise is primarily an on-premises solution. It can be deployed on a company’s own server and can be used behind your firewall. This ensures safety, security and complete control over your data assets.

Address Verification:

DataMatch Enterprise is a CASS certified address verification and validation solution that lets users verify address fields against government databases like the USPS and Canada Post among others. This feature is a "must-have" if you need to clean a mailing list. There are 54 fields according to which you can match, verify and validate your address. From checking 5 digits zip codes to identifying street names to longitude and latitude fields, you have a wide range of address validation field to use.

Survivorship and Data Enrichment

Once you have performed the cleansing, matching, and fixing of your records, you might want to save that master record and keep it as the central record. DME allows for data survivorship and data enrichment where you can decide what record you will use to overwrite, determine master record or merge data.
The merging of data allows you to create a complete master record. This is pretty handy if you have customer info dispersed over several platforms or channels. With this option, you can merge specific fields of those other data sources and create a complete picture of your customers.

**Final Export**

What do you eventually do with the duplicates? With the final export option, you can export all of the records from your data sources and create a column for the match group ID in order to flag the matches and duplicates. The final export option lets you create a master record for each group of duplicates. Additionally, you can also choose to export duplicates only or use a cross-reference to export the results that have similar values in a cross-column format.

**Conclusion**

Data is the golden goose of businesses today - both private and public organizations need to collect, process and make sense of this data if they want to grow. Data matching, therefore, is no longer just about removing duplicates. It plays a central role in all data management and data processing objectives. The complexity of data requires institutions to invest in solutions that can tackle underlying challenges and help save time, yet ensures accuracy and data reliability.
Data Ladder is a data quality software company dedicated to helping business users get the most out of their data through data matching, profiling, deduplication, and enrichment tools. Whether it’s matching millions of records through our fuzzy matching algorithms, or transforming complex product data through semantic technology, Data Ladder’s data quality tools provide a superior level of service unmatched in the industry.

DataMatch Enterprise, was proven to find approximately 5-12% more matches than leading software companies IBM and SAS in 15 different studies.

- Unparalleled Matching Accuracy and Speed For Enterprise Level Data.
- Cleansing beating IBM and SAS.
- Proprietary Matching Algorithms with a high level of matching accuracy at blazing fast speeds on Desktop/Laptop Hardware.
- Big Data Capability with data sets up to 100 Million Records.
- Deduplication and Merge Purge within and across any number of files.
- Suppression of existing customers or Do Not Contact from marketing lists.
- Advanced record linking technology to create data warehouses.
- Quick Data Profile tool finds and fixes Data Quality issues within the first 5 minutes of setup to improve match quality.

Free Download