The Complete Guide to Name Matching

What It Is, How It Works, and Deciding Which Approach to Take

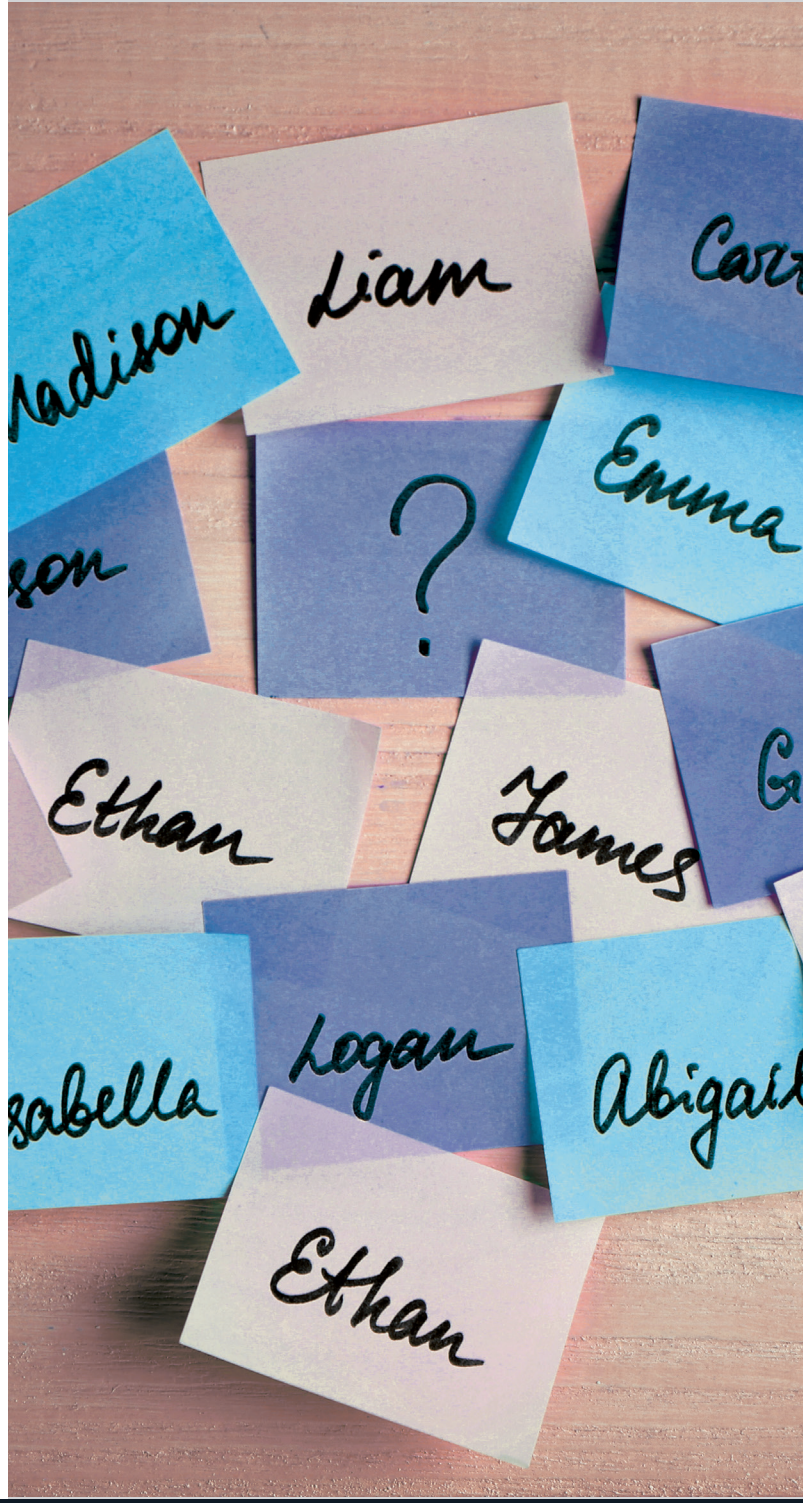
BABELSTREET

Name matching is difficult

Modern commerce, finance, and security require fast, accurate data from multiple sources. But many times the only common link between data sources is the name of a person, place, or organization. And unlike other forms of data, names are highly variable. Spelling, misordered name components, and even nicknames (to name just a few challenges) can result in a failed or false match. The consequences of mismatching a name are no joke, either. Border patrol could fail to detain a watchlisted individual. A hospital could create duplicate patient records. A bank could transfer funds to a sanctioned organization.

You may be thinking, okay, why not just use regular text search? But regular text search doesn't go far enough to address the challenges of name matching or searching. Traditional methods for matching names do exist, but their effectiveness is greatly limited by a trifecta of frustrating side effects. They're weak, slow, and expensive to maintain and run.

Businesses need something better — and that something is modern, AI-driven, intelligent fuzzy name matching.



What is name matching used for?

Numerous critical situations require being able to verify an identity, and that almost always begins with looking up a name. Is the person at the border entry point on a list of suspected terrorists? Is that organization trying to move money to a bad actor? You have to get those questions right — and you only get a few minutes to answer. If you can't do it quickly, you hold up law-abiding citizens. And if you can't do it right, the bad actors get away. Here's how name matching can improve speed, accuracy, and efficiency in a few specific use cases:

Border security

Missing even a single match against a watchlist puts citizens at risk. Boarder security agencies must deliver the most accurate results possible within real-world time constraints to ensure the seamless flow of people and goods through points of entry.

Anti-money laundering

Financial institutions are required by anti-money laundering regulations to avoid doing business with known bad actors. They must be able to check an entity against sanctions lists, verify information against business directory listings, and reduce the time required for manual remediation of false positive results.

Patient matching

Healthcare providers and payers need to quickly find and link patient records, creating a 360-degree view that enables better care and avoids duplicate records, which lower productivity and increase costs.

Investigations

Criminal investigations hinge on unambiguous identification of suspects, making name matching technology crucial for law enforcement agencies.

Getting fuzzy: Why name matching is far from easy

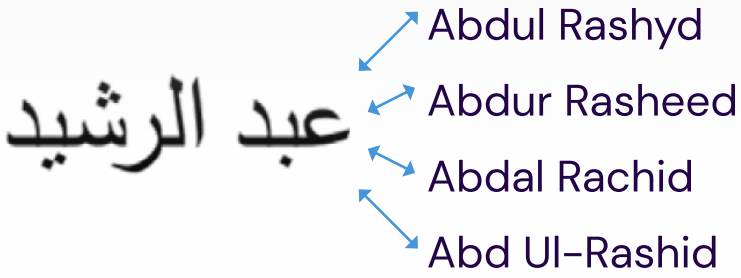
In a structured database, a name is a data point within a record, just like an email address, phone number, or unique ID number. But what happens if you only have a name to look up a record? It happens more often than you think, as privacy regulations may prevent the creation or sharing of ID numbers.

When names are your only unifying data point, you have a problem. (Actually, you have a lot of potential problems!) Unlike Social Security and other ID numbers, names are highly variable. There isn't one correct way to spell a name. And a name isn't unique to a single person or entity. Nicknames, transliteration spellings, multiple spellings of the same name, or the same name written in different languages or scripts are just a few of the ways matching can fail.

We've come to expect our search engines to overcome spelling errors, like when you type Searsha, and the search asks if you meant Saoirse. You might expect the same of name matching, but it's technically much harder behind the scenes.

By the way

This one common Arabic given name can be transliterated more than 1,000 different ways in English. Here are just four possible spellings.



Why name search is unique

While there is an abundance of search tools on the market, name search is a different animal than document search, and requires a fundamentally different approach. Here's why:

1. Typos have an outsized impact on name search. For example, "teh" is 1/250th of a one-page document; the typo in Jhon Smith is 50% of the name.

2. Spell-checking names is impossible. Cindy, Cyndi, Cindi, Cyndy, Syndy, Syndi, Sindi, and Sindy are all correct!

3. Common names (such as John) are as important as unusual names (like Zappa), yet internet search engines demote the importance of frequently occurring words.

A few challenges to name matching

Phonetic similarity
Kailey ↔ Caylee ↔ Kaylie

Transliteration spelling differences
Abdul Rasheed ↔ Abd al-Rashid

Nicknames
William ↔ Will ↔ Bill ↔ Billy

Missing spaces or hyphens
MaryEllen ↔ Mary Ellen ↔ Mary-Ellen

Titles and honorifics
Dr. ↔ Mr. ↔ Ph.D.

Truncated name components
Blankenship ↔ Blankensh

Gender
Jon Smith ↔ John Smith (but not Joan Smith)

Missing name components
Phillip Charles Carr ↔ Phillip Carr

Out-of-order name components
Diaz, Carlos Alfonzo ↔ Carlos Alfonzo Diaz

Initials
J. E. Smith ↔ James Earl Smith

Name split inconsistently across database fields
Rip · Van Winkle ↔ Rip Van · Winkle

Same name in multiple languages
Mao Zedong ↔ Мао Цзэдун ↔ 毛泽东 ↔ 毛澤東

Semantically similar names
PennyLuck Pharmaceuticals, Inc. ↔ PennyLuck Drugs, Co.

Semantically similar names across languages
San'in Telegraph and Telephone Corporation ↔ 山陰電信電話株式会社

Organizational aliases
Boston Brewing Company ↔ BeantownBeer

Precision, recall and accuracy

The key metrics to use when evaluating a search solution or the performance of a system are precision and recall. Together, these define accuracy. Already confused? Here's the breakdown.

Precision measures the percentage of "matches found" that are correct. High precision means fewer false positives (incorrect matches).

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall measures the percentage of all possible correct matches that are actually found. High recall means fewer false negatives (missed matches).

$$\text{Recall} = \frac{tp}{tp + fn}$$

Key

tp = true positives (correct matches found)

fp = false positives (non-matches labeled as "matches")

fn = false negatives (missed matches)

The pros and cons of common name matching methods

As a business user, you're interested in results. You probably prefer leaving the technological details to the pros. But you still need to understand a few key terms and solutions when talking about name matching — fuzzy or otherwise — with your technical team. You have to know what kind of accuracy your business needs (see "Precision, recall, and accuracy"). Once you've gained an understanding of some popular approaches to name matching, and their pros and cons, you'll see that a new, more sophisticated approach is necessary.

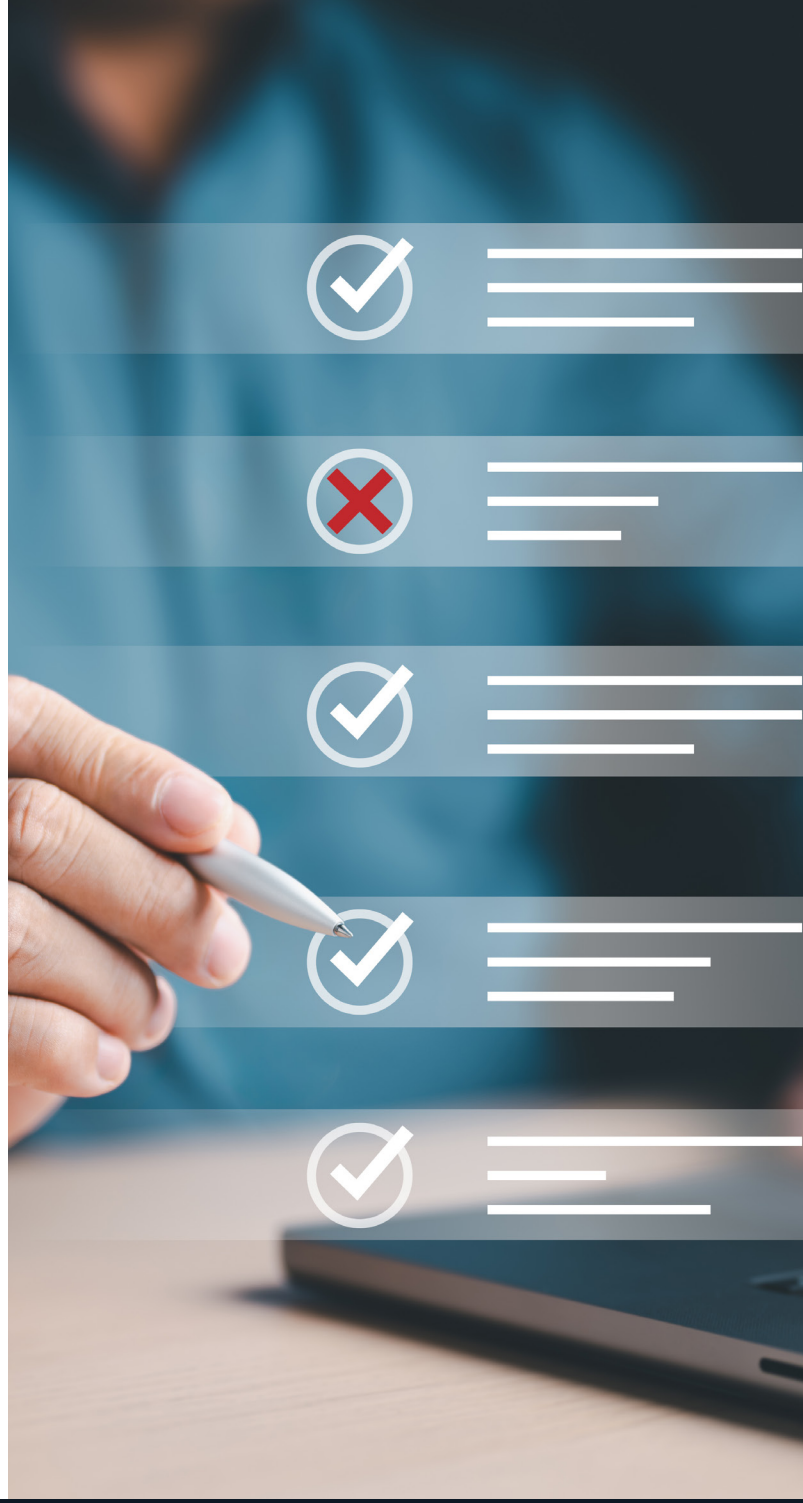
Simple and unsophisticated

Edit distance

Edit distance looks at how many character changes it takes to get from one name to another, but it lacks the linguistic smarts to understand that Jack to Jcak is more likely a match than Jack to Mack.

Pros: Easy to implement; fast

Cons: Limited to Latin-based languages (e.g., English, French, Spanish); all swaps are weighted evenly, missing linguistic nuances



Smarter, but limited

Rules-based methods

Nearly 30 years ago, the go-to name matching technology was to create huge static lists of variations for every name on a watchlist. These rules-based systems are complicated, top-heavy, and difficult (therefore costly) to maintain. There might be hundreds of variations for an average name with three components — an Arabic name might easily have five. Multiply that times a million names on a list, and you have a good sense of how this is bad for real-time processes. And you still might not make the match.

Rules-based methods are restricted by human knowledge. They can only capture situations that people encounter or imagine. As a result, they can be as frustrating as a game of whack-a-mole: a new variant pops up, leading to a new rule, which may affect how other rules work, adding complexity atop complexity.

There are two broad categories of rules-based methods.

The list method attempts to list all possible spelling variations of each name component and looks for matching names from these lists of name variations.

Pros: Easy to edit or add new rules

Cons: Difficult to coordinate potentially conflicting rules, computationally intensive, requires expensive hardware to run against long lists quickly; can't handle unlisted names, missing/added spaces between components, or name components in incorrect fields

The common key method addresses some of the limitations of lists by reducing names to a key or code based on their English pronunciation, such that similar sounding names share the same key.

Pros: Fast execution; high recall

Cons: Mostly limited to Latin-based languages; transliterating non-Latin names reduces precision

Here's an example

Your matching strategy finds three matches to your target name from a list of 200 names. Of those matches, two are correct, and one is incorrect. There were an additional three correct matches in the sample that your strategy did not uncover.

tp = 2
true positives (correct matches found)

fp = 1
false positives (non-matches labeled as "matches")

fn = 3
false negatives (missed matches)

The precision of your results is .67 = 2/(2+1).

The recall of your results is .40 = 2/(2+3).

Sad but true: As precision increases, recall decreases and vice versa. If every name was picked as a match, you'd have perfect recall, but dismal precision. If you picked only the top match and it was correct, precision would be perfect, but recall would be low.

So what is accuracy? Accuracy (known as an "F-score") is a calculation that combines precision and recall. What most businesses need to understand is whether they need higher precision or higher recall.

What kind of accuracy do you need? It depends!

Higher precision searches are those that hone in on the likeliest matches in order to reduce time-wasting false hits. They're designed for finding patient records or bank compliance screenings.

Higher recall searches are best for high-stakes situations, such as border security, where a miss could mean a potential terrorist attack.

Smartest, but slowest

AI-powered statistical models

A statistical approach takes hundreds, if not thousands, of matching name pairs and trains a model to recognize what two “similar names” look like, so that the model can calculate the probability that two names are a match and assign a similarity score.

Some statistical models can compare the semantic similarity of common words (i.e., how close in meaning they are), such as “drug” and “pharmaceutical,” to spot a possible match in “PennyLuck Drugs” and “PennyLuck Pharmaceuticals.” Semantic similarity is especially powerful across languages.

日本電信電話株式会社 would be phonetically transliterated as Nippon Denshin Denwa Kabushikigaisha, but its official English name is Nippon Telegraph and Telephone Corporation. The semantic match would be telegraph to 電信 (denshin); telephone to 電話 (denwa); and corporation to 株式会社 (kabushikigaisha).

Pros: Matches across languages and scripts; offers greater precision

Cons: Slower performance; high barrier to entry, as it requires training data and customized algorithms

The winner: AI-powered hybrid two-pass

All of the previous methods excel at solving a specific problem, but only one. Edit distance is a blunt instrument for the complexity of name matching. Lists are customizable, but hard to maintain and slow to execute. Common keys are fast to execute, but offer limited extensibility and accuracy, lacking the smarts of the slower statistical approach.

But a method can't be good at just one thing when it comes to successful name matching. It must be able to address all of the name matching challenges listed above to succeed. The solution is a hybrid strategy that uses the strength of one approach to overcome the weakness of another. This is known as the AI-powered hybrid two-pass method.

Artificial intelligence can dynamically and simultaneously consider all the key ways that names vary, not just the ones found in a list. It can weigh computational methods (hence the "hybrid") to refine a search and return the likeliest match. The dynamic quality of AI enables the identification of name variations in real time, instead of iterating over static lists.

The two-pass approach is a strategy for maximizing both recall and precision. It works like this:

1. **The first pass** uses the common key method, which quickly produces a high-recall match set. In other words, it gets rid of obvious non-matches.
2. **The second pass** uses myriad computational methods combined with a statistical model (the AI). These take longer, so starting with a list that's already winnowed down saves a lot of time. This pass is extremely high precision because the system is making smart decisions to assign a match score to each pair of names it considers.

The two-pass approach can handle a huge variety of variations: nicknames, misspellings, misplaced field data — even different languages and alphabets (scripts). This greatly improves accuracy compared with the common key method alone. Instead of being locked into a coarse comparison of derived keys (for better or worse), the second pass of the hybrid approach takes a fresh look at the original names in their original scripts before scoring their similarity.

The hybrid method also avoids the weaknesses of the list approach by not relying on pre-generation of name variations. Instead, it dynamically considers (via AI) the linguistic variations of names in each language. This linguistic knowledge of name variations also gives the hybrid approach an edge over the edit distance method, which lacks linguistic knowledge and cannot directly compare names in different scripts. The result is a fast, accurate name matching solution.

The Winning Combination

Hybrid

Uses multiple methods to maximize accuracy and speed across a broad range of name variations."province/state," and "country."

AI-powered

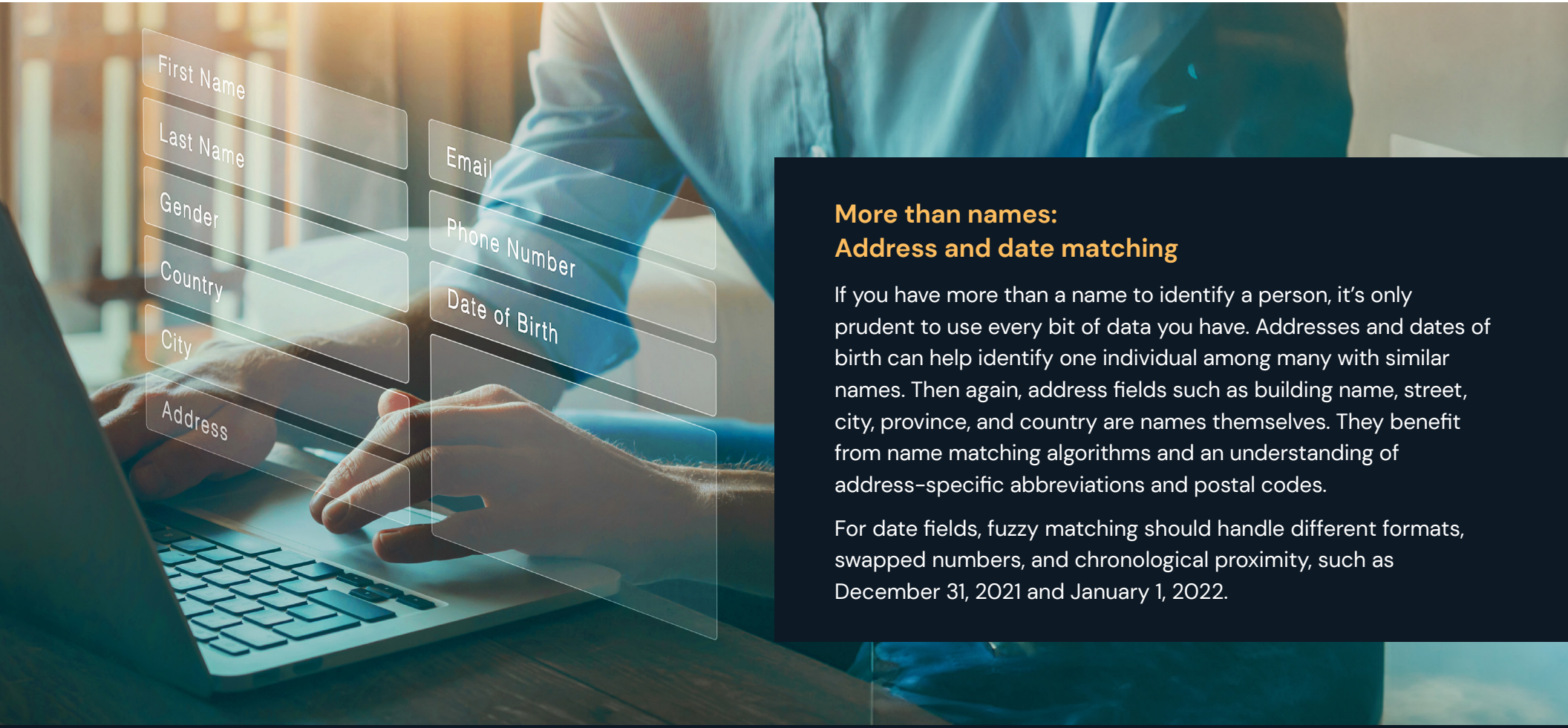
Adds precision in evaluating matches using machine learning.

Two-pass

Pass 1 quickly eliminates obvious non-matches. Pass 2 takes a closer look to rank matches from highest to lowest confidence with high precision.

Congratulations! You know that misspellings, aliases, nicknames, initials, and different languages are but a few of the things that get in the way of making the right match. You also know how important name matching is for keeping people and businesses safe. So why are you using a slow, inaccurate solution?

If you really want to solve the problem, you can begin with a smart, powerful name matching solution that integrates into your current systems without disruptive and risky rip-and-replace implementation. A solution that takes the best of traditional name matching technology, throws in some serious intelligence, and gives you a match score you can trust.



More than names: Address and date matching

If you have more than a name to identify a person, it's only prudent to use every bit of data you have. Addresses and dates of birth can help identify one individual among many with similar names. Then again, address fields such as building name, street, city, province, and country are names themselves. They benefit from name matching algorithms and an understanding of address-specific abbreviations and postal codes.

For date fields, fuzzy matching should handle different formats, swapped numbers, and chronological proximity, such as December 31, 2021 and January 1, 2022.

Babel Street Analytics for identity intelligence

You need Babel Street Analytics. Its hybrid AI-powered name matching technology is the go-to choice for mission-critical applications that need to verify or match identities because it is accurate, fast, and easily integrated into any system. Babel Street Analytics' plugins for Elasticsearch and Apache Solr handle the complexity of name matching and only deliver match-score ranked results.

Babel Street Analytics uses a two-pass approach to take advantage of the speed of the common key method and the precision of machine learning to perform fuzzy name matching, decreasing false positives and false negatives. It lets you tune parameters for each use case by adjusting the precision/recall ratio, accommodating name data idiosyncrasies, and weighting data fields to account for their reliability.

Customers report up to a 90% reduction in false positives and an increase in true positives when using the global name matching capabilities of Babel Street Analytics.

Every day, Babel Street customers conduct over 500 million watchlist checks worldwide.

Babel Street is the trusted technology partner for the world's most advanced identity intelligence and risk operations. The Babel Street Insights platform delivers advanced AI and data analytics solutions to close the Risk-Confidence Gap.

Babel Street provides unmatched, analysis-ready data regardless of language, proactive risk identification, 360-degree insights, high-speed automation, and seamless integration into existing systems. We empower government and commercial organizations to transform high-stakes identity and risk operations into a strategic advantage.

Learn more at babelstreet.com

All names, companies, and incidents portrayed in this document are fictitious. No identification with actual persons (living or deceased), places, companies, and products are intended or should be inferred.

BABELSTREET