WHITE PAPER

An Introduction to

Name Science

and Its Role in Fighting Financial Crimes

Understanding names is key to unlocking the potential of AI-driven AML and anti-fraud automation

QUANTIFIND

ABSTRACT – AN INTRODUCTION TO NAME SCIENCE AND ITS ROLE IN FIGHTING FINANCIAL CRIMES

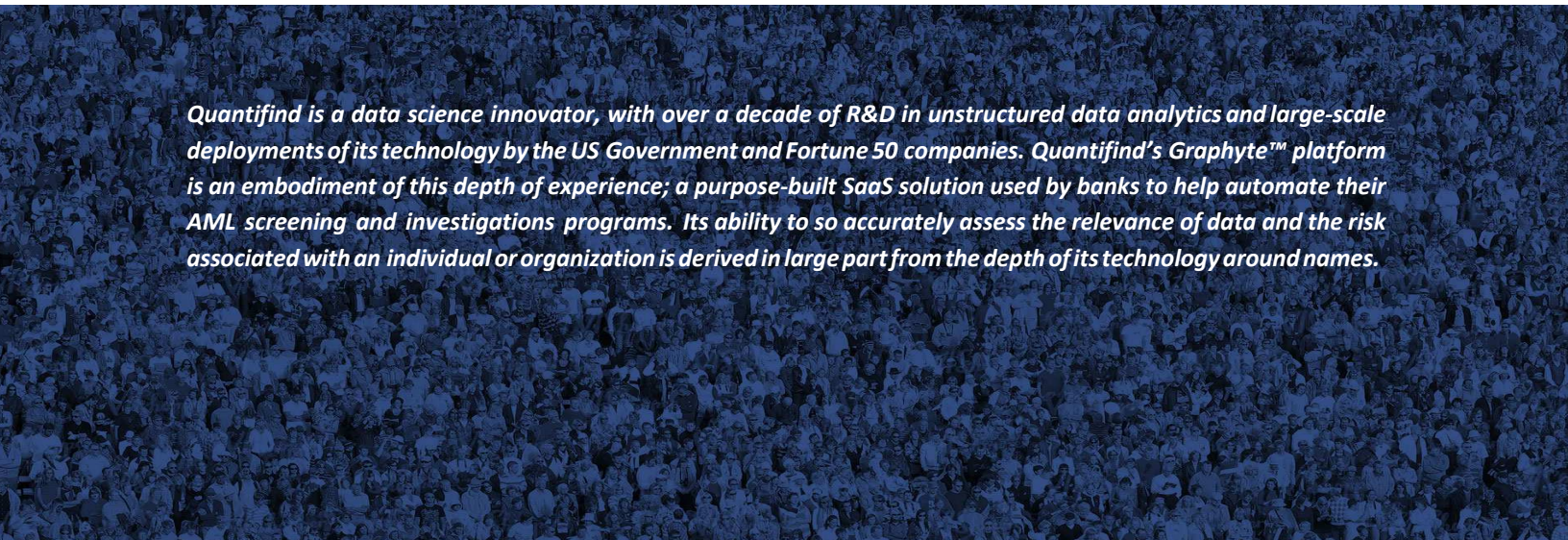
Understanding names is key to unlocking the potential of AI-driven AML and anti-fraud automation

Much of a bank's efforts to prevent global money laundering and fraud can be broadly divided into two categories: monitoring and analyzing transactions, and investigating the people and organizations behind them. Advances in machine learning have made transaction analysis more effective. However, the context contributed by external data is critical to AML and prompts a large proportion of the red flags that lead to successful investigations. Interpreting that context properly is a labor-intensive process that even with large teams can lead to alert backlogs, delayed onboardings, and surveillance gaps; costs are growing by 20% annually. Banks are compelled to make their programs more efficient, and are exploring AI-powered automation not only in their transaction analysis but also in their investigations.

Useful context about people and organizations can be found scattered across vast, disparate stores of public-domain data, but discovering relevant signals relies on first extracting the correct name. Foundational to this process is the emerging field of "*name science*", a subspecialty of data science that is essential to achieving the accuracy needed for true automation. Investigations require that supporting information is attributed to the right entity, and that it is relevant; essential to both is an understanding of names. Thanks in part to name science and underlying advances in machine learning, automation holds the promise to free up time of investigation teams so they may focus their talents on the highest-risk cases.

Banks are beginning to realize the potential, with greater than 40% efficiency gains observed as they automate and reassign headcount. Opportunities are found throughout the KYC/AML workflow and customer lifecycle, from screening, onboarding, due diligence, and surveillance to alerts triage and investigations.

This paper introduces fundamental aspects of name science and how it is applied in practice in support of AI-driven financial crimes automation.



Quantifind is a data science innovator, with over a decade of R&D in unstructured data analytics and large-scale deployments of its technology by the US Government and Fortune 50 companies. Quantifind's Graphyte™ platform is an embodiment of this depth of experience; a purpose-built SaaS solution used by banks to help automate their AML screening and investigations programs. Its ability to so accurately assess the relevance of data and the risk associated with an individual or organization is derived in large part from the depth of its technology around names.

INTRODUCTION – THE DATA INSPIRES THE SCIENCE

Billions of public-domain data records can be leveraged to automate risk assessment

External data sources including sanctions lists, PEP lists, regulatory proceedings, enforcement actions, and non-standard entity lists (e.g., the Panama Papers)—as well as online news, company data, and legal entity registrations—can all be leveraged for a better understanding of the customer's identity and associations. Unfortunately, the most enlightening data sources are often also the noisiest, and the technical requirements to extract meaningful information can be complicated. Ultimately, the data is only useful if a customer's identity can be accurately matched to the entity extracted from the articles. This process of "entity resolution" must be performed such that relevant information about a given entity is correctly attributed and information about other entities is not.

There is also the matter of assessing relevance, which in financial crimes investigations determines risk. Identifying the entities behind transactions and knowing their full history and relationships are at the core of assessing the likelihood that they pose a risk. A key pillar to accurately incorporating identity into the automation equation is *name science*.

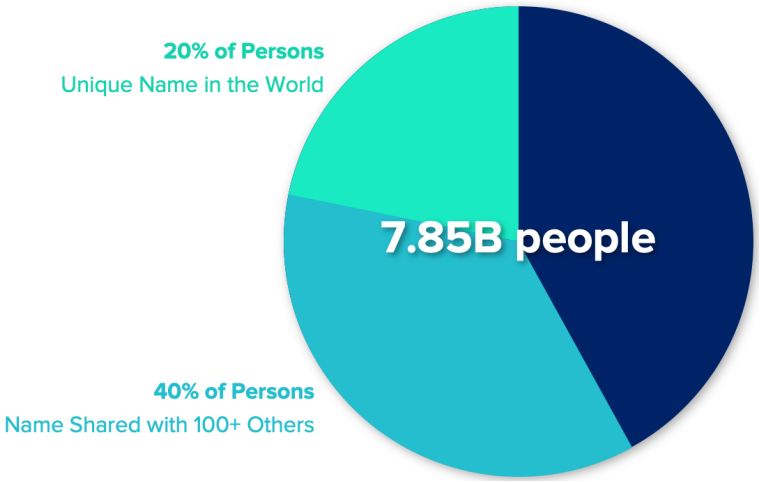


Figure 1 – Modeling name frequency is a critical aspect of name science.

Names are different from language. There are relatively consistent spelling and grammar rules that help define a language, and machine learning is really good at learning rules. Names follow rules too, but they are far more fragmented, temporally and geographically. Name structures, conventions, and standards are driven by culture and society; they differ from language to language and vary greatly around the world. Understanding names as data and accurately modeling how they behave is essential to leveraging global unstructured data sources to perform accurate entity resolution and to incorporate the context of identity into risk assessment automation. This requires a nuanced and purpose-built approach to capture the many idiosyncrasies of names.

NAME SCIENCE IN PRACTICE – MODELS AND ALGORITHMS

The accuracy needed for automation demands a mastery of names

The text in an unstructured document can only be summarized into structured data forms after the subjects of the document are identified with confidence. Names must be discriminated to avoid mistaken confusion with a different entity with a similar name. The subject must be determined to be either a person or an organization; and a name must be classified as a surname, given name or a middle name.

Fuzzy matching algorithms are used to determine when misspellings and missing middle initials are combined. Geographical proximities can be equated within a certain radius, and algorithms can be trained to be culturally aware. The probability of a particular name transformation is dependent on where the subject is from. Training sets that can judge accuracy are readily available, and macro-variables such as localized name rarity drastically improve the precision of matches. Finally, the context of the name within the broader document needs to be ascertained to ensure that the subject is correctly classified; (e.g. as the criminal, the victim, or the article author). This context can also be expanded based on the related entities and link analysis of the subject.

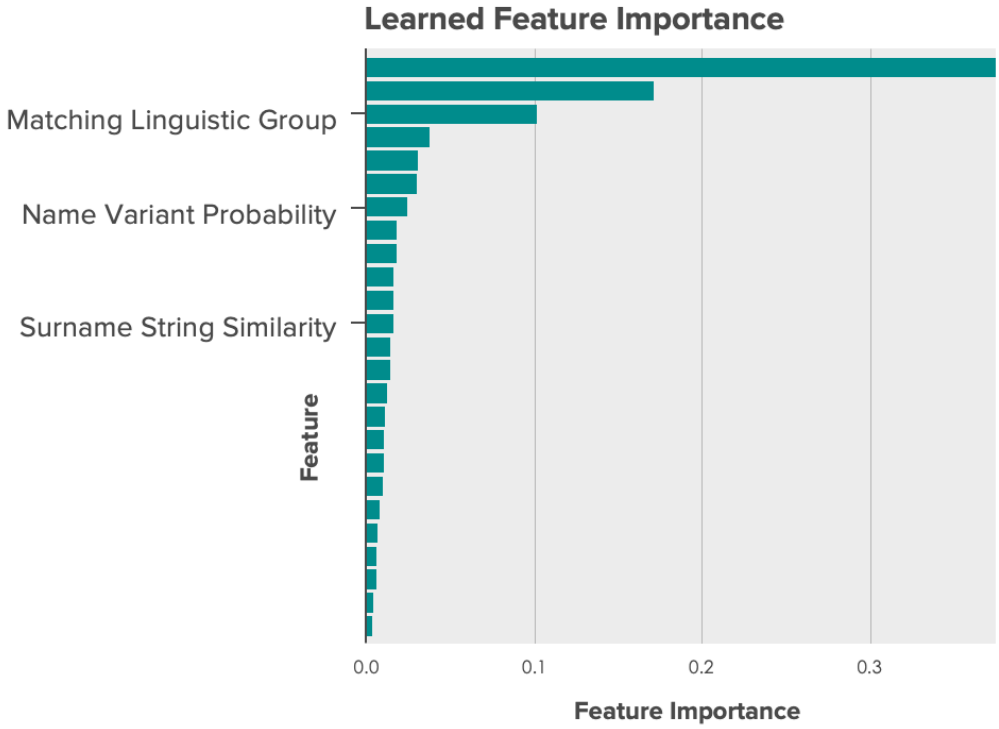


Figure 2 - Name science features stand out in the models.

The accuracy of the algorithm will depend on the metadata available, as well as the commonality of the name, and therefore, in cases where the data is sparse or the name common, the confidence of the match is unavoidably low. Therefore, in all cases, the *confidence* of the classification also needs to be transparent. Financial institutions can thereby quantify their risk exposure with precision, automate when possible, and manually review when necessary. This decision on when to automate requires knowledge about how common the subject's name is, which metadata to trust the most, and what constitutes a reliable fuzzy match.

Behind these capabilities are technology building blocks—entity recognition, fuzzy matching, entity resolution, entity linkage, and contextualization—that are critical for leveraging the vast and growing body of public-domain, unstructured data to make screening and investigations more efficient. Enabling these algorithms to work reliably enough to make automation operationally viable mandates mastery of *name science*.

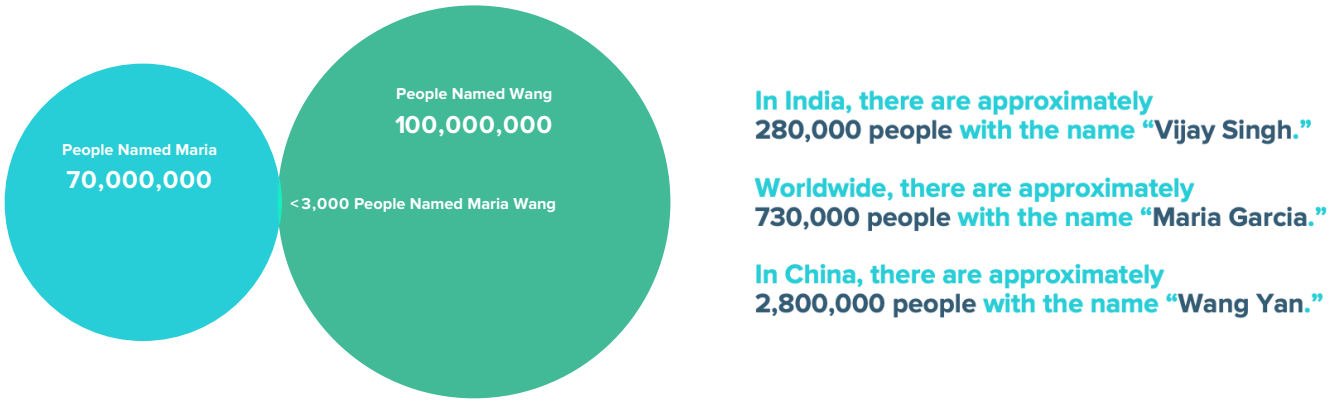


Figure 3 - Many names are extremely common.

Onomastics – the study of names

The study of names is called *onomastics*, and has applications from history and genealogy to law enforcement. Names give identity to people and places across times and cultures in a way that provides predictive signals. In the context of financial crimes, these properties are particularly useful to search and discover valuable information about people and organizations.

But the utility of names in properly attributing information to the right individual relies on our ability to address two conditions: 1) a single person being referred to by multiple variations (polysemy), and 2) many people sharing the same name (synonymy). Much of name science is motivated by the need to address these two conditions. In the crime-fighting context, it's also important to recognize that some names represent an intentional attempt to obfuscate a true identity, such as with aliases, synthetic identities, and shell companies.

Data Source Type	Bias(es)	Gap(s)	Other Limitations
Genealogy and Ancestry	Skews wealthy and white	Developing nations	Focus on high-frequency surnames
Official census		Non-US Undocumented immigrants	Aggregated by name parts
Social media samples	Skews wealthy and young	Developing nations	Informal use of names Unparsed names
Company data: officers	Skews male and old	Developing nations	Small sample size

Figure 4 - Models are needed because data sources are flawed.

Fundamental name science functions – recognition, parsing, and linking

Naming conventions vary drastically across different cultures and linguistic backgrounds. In some cultures, individuals possess multiple given names, surnames, and religious names. In others, a given name and surname is most prevalent. Name science algorithms apply cultural awareness to help not only sort out these factors but leverage them to derive better insights and information.

Machine learning-based *named-entity recognition (NER)* models automatically identify a name within an article, and then classify the name by the type of entity it identifies. *Global name parsing* is the process of applying algorithms as well as cultural awareness to the task of breaking a full name into its subcomponents; e.g. for a person those might be salutation, given, middle, surname, and suffix. *Named-entity linking (NEL)* describes the task of matching a named entity of interest to a document by leveraging associated metadata such as birth year, location or employer.

Full Name: Maria Gabriella de Gonzalez Fuentes	
Issue	**Example**
Swapped Tokens	Gabriella Maria de Gonzalez Fuentes
Added Tokens	Gabriella de la Gonzalez Fuentes
Missing Tokens	Gabriella Fuentes
Alternate Spellings	Gonzales
Dropped Letters	Gabriela

Figure 5 - Name parsing challenges are important to solve, as they affect downstream processes.

Understanding the linguistic group behind a name is crucial to these processes. A data-driven Bayesian approach can be used to predict the linguistic group for a name and a decision tree model is used to parse the name accordingly. Accuracy is particularly important because the results are used downstream to perform higher-order analysis and processing of names. Errors at this early stage in the process can become magnified. Many ambiguities have the potential to reduce accuracy, such as:

Middle name vs. last name - Different cultures use middle names as either part of a family name or part of a personal name. Some married names are hyphenated, some are not, and the structure of names varies widely by geography.

Organization vs. person - There are many examples of organizations that contain common human names, and so it's necessary to be able to reliably distinguish between, for example, S. C. Johnson the person and S.C. Johnson the company.

Name ordering - In different cultures, first, middle, and last names are placed in different orders. Algorithms can use empirical data to quantify the likelihood that the first, middle, and last names of a subject in a particular country are identified correctly. Data with first and middle initials must also be parsed. String distance algorithms are helpful but notoriously imprecise.

Global name variants – different spellings of the same name around the world

There are well over a hundred writing systems in use globally, and yet the majority of publicly accessible data is based upon Latin, or Roman characters. The transliteration of the words and sounds from other systems—and romanization of names into those of the Latin alphabet—is not a linear one. Thus, it is common to have multiple spellings of Romanized names, with each influenced by the translator and target audience.

A commonly cited example is the name "Mohammed", which has more than 300 different spellings[1] in Latin characters. Sources have their own biases that have an impact on how names are spelled, and this bias can contribute to inaccuracies in which data is attributed to a particular entity. In addition, the same name will take on a modified spelling under influence of the local language, dialects, and pronunciations. One example is the anglicization of names such as Kathleen, derived from the Irish "Caitlín".

Metaphone and Soundex are phonetic algorithms for indexing words and names, respectively, by sound, and "string distance" algorithms such as Jaro-Winkler and Levenshtein are a common method to calculate the difference between two strings. They can be used in concert to quantify the similarity of how two names sound towards recognizing differently spelled names. This approach achieves excellent match rates but also significant false positives. For example, a search for "Oleg Deripaska" returns the true match, but using string distance also scores high for "Francois Olenga", despite having obvious and definitive mismatch qualities. Using name science, we can more reliably determine the probability that a searched name and matched alias belong to the same person, by:

1) segmenting names into cultural groups,
2) parsing names into tokens of different types,
3) using culturally aware query expansion for searching potential variants of a name, and
4) scoring name matches in such a way that determines the probability of a match.

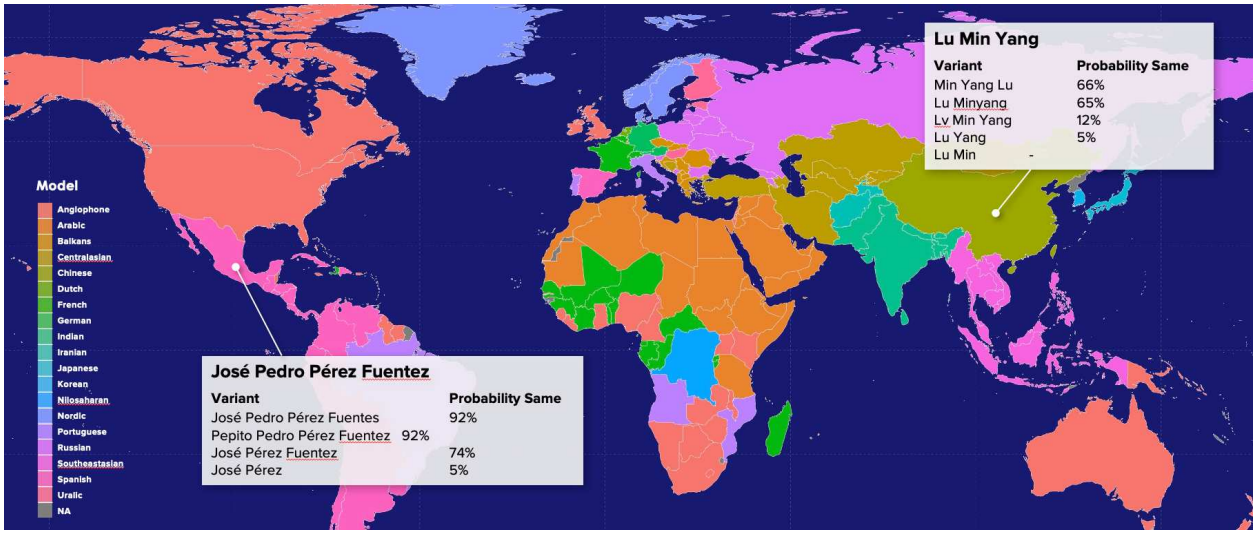


Figure 6 - Name variants are different around the world.

[1] Christopher Westphal, Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies, 1st ed. CRC Press, Inc. Boca Raton, FL, USA ©2008.

Global name rarity – strong clues that inform match/no-match decisions

Using context and a deep set of training data, a name's geographical distribution can be used to better separate good matches from bad ones. Machine learning can be used to determine global name rarity, which—in combination with fuzzy matching and contextual models—brings cutting-edge technology to the AML workflow.

Consider the name "Mike Hamilton." In the United States alone, there are more than 3,000 people with that name. With all the possible variations and without additional information such as age, how can institutions quickly, reliably, and accurately find the right Mike among all the noise? To make the problem more interesting, what if the Mike in question were not living in the United States?

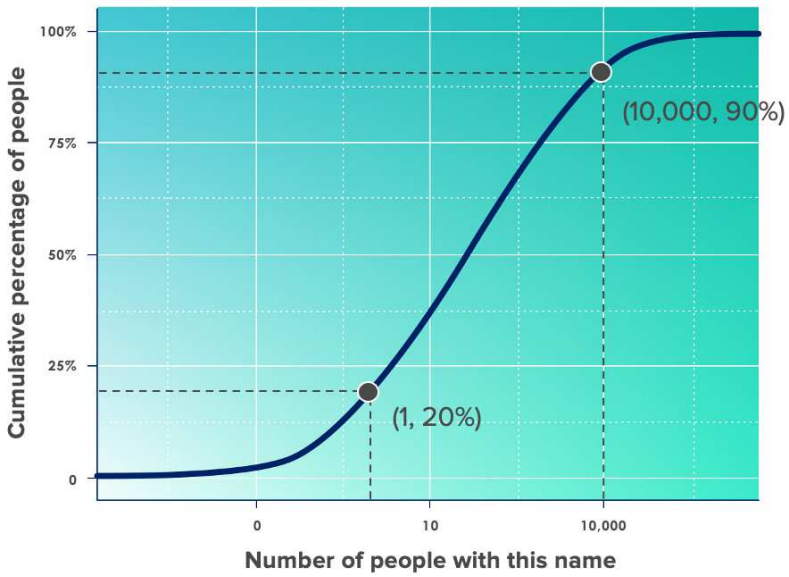


Figure 7 - Name frequency can be described with a mathematical model.

An article about Mike Hamilton residing in the United States, with no accompanying metadata, is unlikely to be referring to the same Michael Hamilton alerted in a transaction based on its commonness. However, if a Mike Hamilton were identified in UAE, in both an alert and a news article, the likelihood of them referring to the same individual increases significantly.

By determining name rarity by location, resources can be prioritized for the sources with the highest confidence. Determining this name rarity is straightforward in the United States, as census data is easily available, and the sample sizes are statistically significant. However, this calculation is much harder when offering a global solution[2].

TELLING THE STORY – NAME SCIENCE HELPS ENSURE THE RIGHT NARRATIVE

Entity resolution – attributing information to people and organizations

Entity resolution is a culmination of the investment in name science. It's the part in the process where all the data from the various sources about an entity is identified as having a high probability of relevancy and associated with that particular entity. The result is a complete narrative that most accurately correlates with their actual risk profile. We may find hundreds of articles that discuss an individual named Mike Hamilton, but they might also be articles sharing facts about a hundred different Mike Hamiltons. Name science helps us get it right.

[2] Learn more about "Mike Hamilton" and global name rarity modeling at https://www.quantifind.com/blog/global-name-rarity-model/

Contextualization – assigning risk

Name science algorithms are designed to find the right person with the correct age and location, but they must also look beyond accuracy and recognize whether the information of a particular article is relevant and useful for a financial crimes investigation.

Contextualization ensures that not only is the right person identified, but that the information is relevant and useful to the investigation. A news article with a subject's name is often returned in a search because there are high-risk terms found in the article, but unfortunately, they are potentially being used metaphorically to describe an event. Lawyers can be scored as high-risk in articles in which they are mentioned based on negative keyword matches describing their client. Celebrities and PEPs can have hundreds of hits, but only a small percentage may be relevant to a financial crimes investigation.

For example, an alert generated by anomalous wiring behavior from a prominent individual in Kansas generated 84 hits from negative news sources. Many of these sources happen to include a negative term. However, even if all 84 hits were confidently about the correct person, only the three hits that indicate risks traditionally associated with a SAR filing are most relevant.

Relationship detection - bad actors rarely act alone

Big wins in AML have occurred when bad actors are uncovered as working together. Thus, while an investigation might begin with a single name, it often is expanded with social graph analysis, co-occurrence in news articles, and linking to internal records.

However, the mere existence of a tangential connection between individuals, while accurate, may not necessarily be useful or relevant. An entity graph can be expanded infinitely based on tangential co-occurrences, which would drastically hurt efficiency.

In order for a node on a related entity graph to be useful, the relationship with a suspect must be scored based on the modeled risk features of the related entity as well as the details of that relationship. This modeling allows us to understand which nodes are relevant in discovering new potential suspects and which are coincidental or not usefully related.

REALIZING AUTOMATION – INTRODUCING GRAPHYTE

Quantifind's Graphyte platform puts accurate, relevant information and powerful tools at the fingertips of investigators

Financial institutions use Graphyte to maximize the efficiency of their screening, due diligence, and investigations by driving automation from end-to-end throughout the KYC/AML workflow. Graphyte is differentiated by its accuracy in assessing risk on an individual or organization, achieved through best-in-class name science, entity resolution, and dynamic risk typologies.

Quantifind has brought these technologies together through over a decade of R&D and experience with large-scale deployments by US Government agencies and Fortune 50 companies. Today, Graphyte is purpose-built to put it all at the fingertips of AML investigation teams in the form of easy-to-integrate APIs and highly intuitive, richly featured investigation applications.

Name science in action – an example investigation

A financial transaction originating in Tanzania and containing the name *Yang Feng Lan* is flagged as potentially suspicious, but no other information is known. An investigator uses GraphyteSearch to perform a basic query, and with only these two pieces of information, Quantifind algorithms are able to extract critical information towards getting to the bottom of this case:

1. **Named-Entity Recognition and Global Name Parsing** - NER algorithms detect multiple mentions of Mrs. Yang under this same alias across numerous news websites and watchlists. This name is first identified as a Chinese name, and global name parsing yields "Yang" as the surname and "Feng Lan" as the given name. In contrast to Western linguistic groups, "Feng" would be recognized as the middle name. Selecting the correct linguistic group and parsing the name with cultural awareness correctly ensures a higher-quality set of name variants, a more accurate name variant score, and ultimately a better entity linking result.

2. **Global Name Variants** - *Feng Lan* (usually written as one token *Fenglan*) is a common forename, so the variant *Fenlan Yang* is a likely representation of this name. In most African languages, this type of variant would be unlikely, but since the name is Chinese it's quite common.

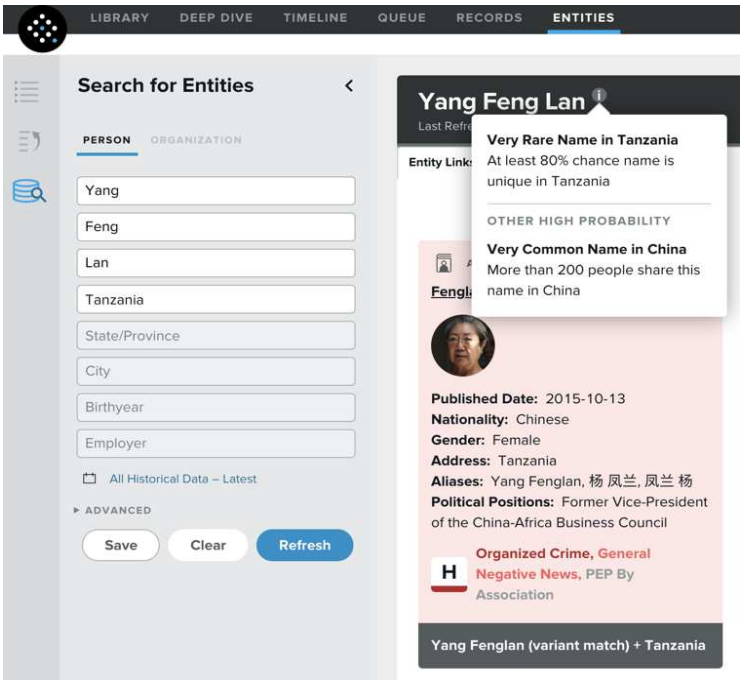


Figure 8 - Name science works in the background to assure that the most relevant and accurate results rise to the surface upon a simple query entered into GraphyteSearch.

3. **Global Name Rarity** - This name is very rare in Tanzania, so it's unlikely there is anyone else with the same name there. Thus, if we see a name match in Tanzania, it's highly likely to be the right person. This name is much more common in China; *Yang* is a common Chinese surname.

4. **Contextualization** - Many of the news reports contain information and keywords give a strong indication that Mrs. Yang has been involved in criminal activity, and therefore indicate "high risk". The first search result shows that Mrs. Yang is a politically exposed person (PEP) involved in organized crime and wildlife trafficking.

5. **Relationship detection** - Algorithms are used to detect other entities mentioned in the same articles, assess their risk and the risk by association imposed upon Mrs. Yang.

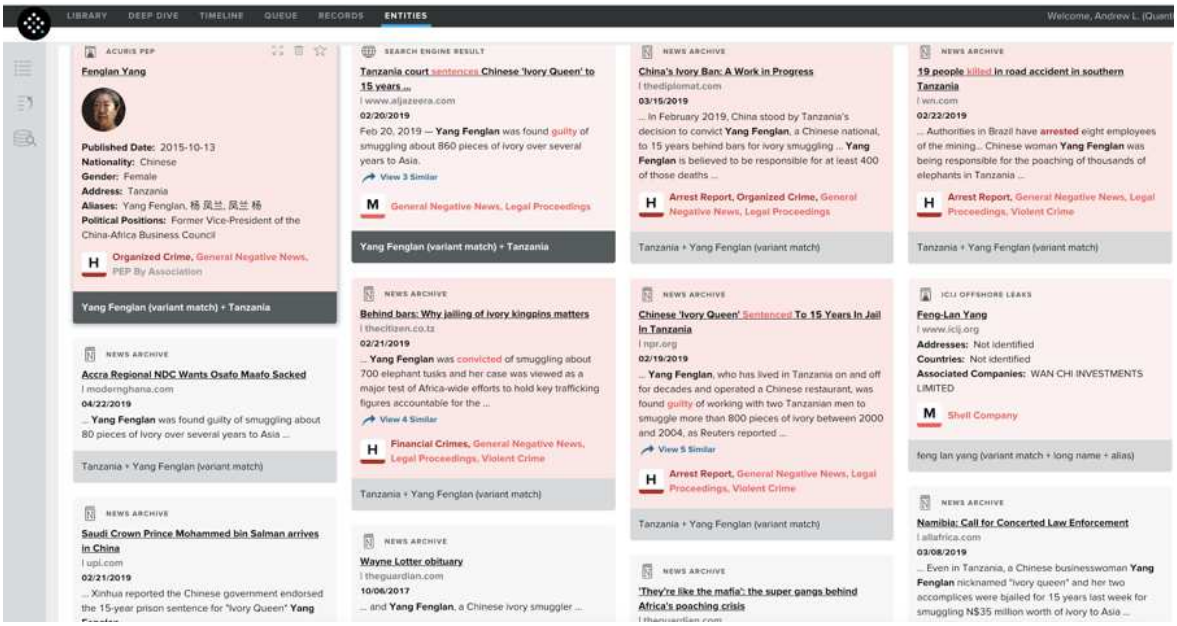


Figure 9 - Several other relevant records show that it's just the tip of the iceberg.

Under the hood – patented real-time search techniques

Graphyte's real-time search performance is made possible by patented in-memory data management techniques optimized for efficient storage, search, and retrieval across disparate unstructured data sources. At its core is Graphyte's "Real Time Cluster" (RTC), a purpose-built distributed query and storage engine that implements name science for speed and scale. Its primary purpose is to efficiently store terabytes of unstructured data in-memory in packed binary form, and to provide optimized querying capabilities over this data.

Normalize the name	*3000 names/second*
Parse full name into name parts	*1000 names/second*
Predict linguistic group	
Predict expected name variants	
Predict name rarity	*500 names/second*
Full-text search (20mm+ news articles)	*200 names/second*

Figure 10 – Benchmark performance: implementation in production requires highly efficient search and processing.

Traditional query approaches involve the use of regular expressions, but they are difficult to scale when applied to billions of documents at runtime. RTC is an engine that allows for rapid, interactive exploration of data as well as data modeling without pre-computation.

For financial crimes applications, Graphyte executes a full-text search across tens of millions of documents, filters results through various NLP models, and returns a summary in milliseconds. This breakthrough capability developed over the course of over ten years is a big part of what differentiates Graphyte from other solutions.

THE IMPACT – FOCUS ON WHAT MATTERS

Name science provides the foundation for a technology-driven, risk-based approach to AML and anti-fraud

The models, algorithms, and architecture behind Graphyte—combined with powerful APIs, case management integrations, and feature-rich applications for investigators—makes it a game changer in the AML automation marketplace. Graphyte has been demonstrated by Tier 1 financial institutions to reduce false negatives and positives by orders of magnitude compared to other automation solutions. This degree of accuracy means that institutions can automate their investigation processes with confidence that they are doing so without encumbering their investigators with false positives and without false negatives that pose a compliance risk.

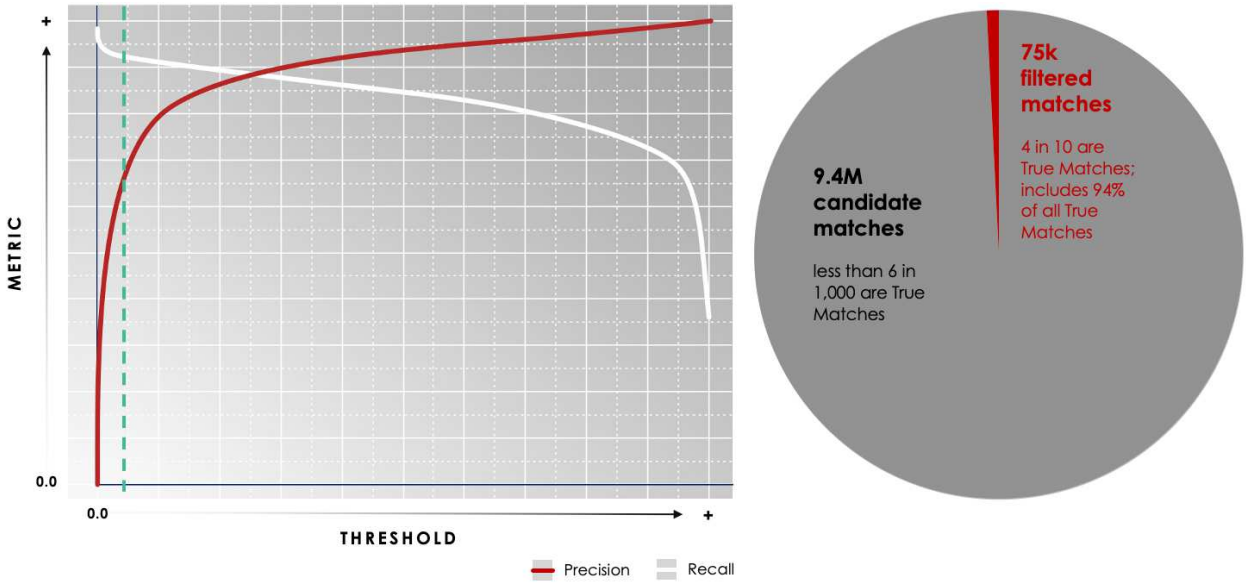


*Figure 11 - A benefit of applying name science is a **10-100x** reduction in the volume of screening hits needing review.*

The ultimate result is that these institutions can reduce the complexity and inconsistency of their investigation efforts. This helps to control the growth of their investigation teams, which can be thousands of people in high burnout roles. These resources can also be diverted to high-risk transactions and applicants, enabling a path from a rules-based to a risk-based approach to AML compliance.

ABOUT QUANTI**FIND**

Quantifind was founded in 2009 upon pioneering work building machine learning technology to discover meaningful patterns across large, disparate, unstructured datasets. Quantifind is headquartered in Menlo Park, California, with teams in Boston, New York, and Washington.

Learn more about Quantifind and request a demo of Graphyte at www.quantifind.com.

March 2021