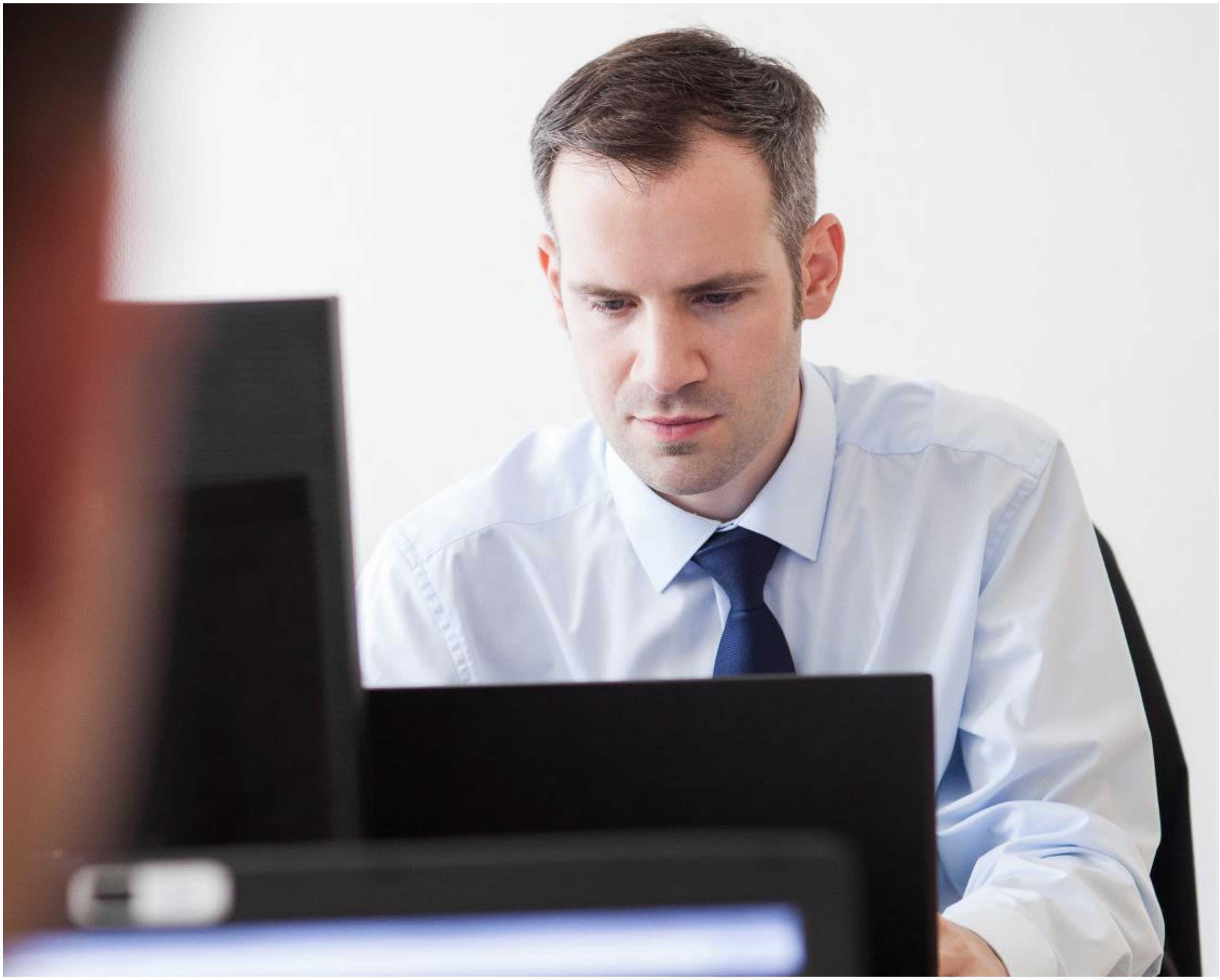
idetect

Name Screening False Positives and Negatives: Can Machine Learning and Artificial Intelligence Really Help?



Contents

Introduction

One of the most important tasks that any compliance officer should complete is the risk assessment of both customers and their transactions. For these tasks there are a number of different activities that they do, such as "data validation," "name screening," "data analytics" and "profiling." The activities can be done manually, but it takes time as they are lengthy processes. For the modern financial services provider (e.g., bank, money service and remittance company, exchange house, insurance company, investment company, financial broker) the speed of executing these activities is crucial. They need fast and reliable systems, true positive alerts only and an automated systematic alert management procedure to complete their risk assessment and decide on further compliance actions.

	Alert Generated	No Alert Generated
Real Risk	True Positive	False Negative
No Real Risk	False Positive	True Negative

Name screening is essential for every customer during the on-boarding process and is frequently repeated to ensure comprehensive Know-Your-Customer (KYC) information. This also applies to every money transfer transaction so as to understand the beneficiary of the funds.

Sanction controls can only be adequately applied using automated name checking powered by appropriate algorithms, with the result being justified alerts for compliance investigators to validate and act on. If the algorithms and/or data quality are poor, false positives will increase along with irrelevant alerts. Consequently, the compliance investigation procedure will be delayed and not much attention will be given to the real risks.

The alerts generated by a well-defined name screening system are categorised into true and false positives. The true positives are the alerts that are meaningful and identify a real risk. False positive alerts are not justified, but represent a false warning as there is no genuine risk.

Therefore, what is important for the compliance investigator is to have as many true positive alerts as possible and as few false positive alerts as possible. To achieve this optimisation in a name screening system, captured data must be of sufficient quality when compared to matched data, and assumptions & correlations must be embedded into the name screening systems and algorithms. There are plenty of legacy algorithms that work fine with different languages, name structures, or matching percentages, but there is no single algorithm that offers a "one-size-fits-all" solution.

Like a human brain, the new generation of name screening solutions use artificial intelligence and machine learning to improve the decision making process so as to reduce the false positives and increase the true positives. The latter will make the work of the compliance officer easier—and the risk assessment activities simpler—to increase efficiency and effectiveness in the implementation of appropriate anti-money laundering and counter terrorism financing controls.

Artificial intelligence and machine learning facilitate a deeper assessment of similarities between pairs of data objects. Traditional matching algorithms, which rely on pre-defined similarity measures, typically focus on a single aspect of name similarity (e.g., pure syntactical similarity). In contrast, artificial intelligence and machine learning simultaneously examine the similarity from a multitude of angles and effectively combine hundreds of attributes between two instances to produce a much more sophisticated estimate of the true overall matching similarity between two instances.

The idetect® solution is one of the first financial crime investigation platforms that uses artificial intelligence and machine learning techniques in its real-time name screening. For each name pairing to be screened it collectively exploits a high number of attributes, where each attribute examines a different aspect of similarity.

Due to this combination, idetect® has successfully reduced the rate of false positives, effectively enabling compliance investigators to concentrate on likely suspicions of money laundering, terrorism financing or violation of sanctions implemented by the critical sanctions regimes. Thus, idetect® offers high efficiency to the financial services provider to comply with the strict regulatory requirements as well as the controls implemented by correspondents.

Increased Regulatory Pressure for Real-Time Name Screening and Risks of Non-compliance

Based on the recommendations[1] of the Financial Action Task Force (FATF), as they relate to anti-money laundering and counter terrorism financing controls, there are specific requirements for identifying persons and entities related to money laundering and/or terrorism financing. These recommendations also cover risk assessment and implementation of controls on executed transaction types in order to determine if they are violating sanctions imposed by United Nations Security Council or other countries.

The United Nations Security Council has, from time to time, implemented specific sanctions on persons and entities to support peaceful transitions, deter non-constitutional changes, constrain terrorism, protect human rights and promote non-proliferation[2]. These persons and entities are included in different sanctions lists that must be used by financial services providers while on-boarding or maintaining any business relation with their customers. Any violation of these sanctions is a threat to worldwide peace and security and constitutes a crime that can be severely punishable, both on a personal basis (the person or persons that violated the sanctions) or on a corporate basis (the financial services provider that allowed the violation). Severe financial penalties have been imposed by regulators to banks and non-banking financial institutions for failing to implement adequate controls and thus allowed their customers to violate sanctions and execute transactions to/from sanctioned entities and countries. The United Nations Security Council is constantly updating the data of their sanctions lists and publishes these to be used by all financial services providers.

A number of sanction regimes (i.e., the financial sanctions which relate to a specific country or terrorist group[3]) have been implemented by different countries—mostly members of the G20[4] —like the United States of America, the European Union and some of its major member states (e.g., United Kingdom). These sanctions include countries, persons and entities in sanction lists to be used by the financial services providers worldwide while maintaining a transactional relationship with the persons and entities or to/from the defined countries. If any violation of the sanction regimes is identified, then there are severe financial actions taken against the offenders, including freezing and confiscating of financial assets or even imposing financial penalties to the persons and or financial services providers involved.

Regulators all over the world issue strict laws, regulations and directives to their licensed financial services providers to take appropriate measures to identify—and take action against—persons, entities and their transactions that violate financial sanctions regimes. These

[1] FATF (2012) - International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation, updated October 2016, FATF, Paris, France – www.fatf-gafi.org/recommendations.html

[2] https://www.un.org/sc/suborg/en/sanctions/information

[3] https://www.gov.uk/government/collections/financial-sanctions-regime-specific-consolidated-lists-and-releases

[4] https://www.g20.org/en

regulatory measures impose a need for financial services providers to apply name screening to identify if any customer, money remitter, beneficiary or any executed transaction is related to sanctions — especially those related to the United Nations Security Council Resolutions.

Definitely, the controls imposed to identify relations with sanctioned countries or persons or entities cannot be done without real-time name screening activities, most of which are automated. The manual validation or checking of sanctions lists can be time consuming, difficult and often provides misleading data, frequently resulting in significant decision making processes for the compliance investigation. As expected, these are not welcomed from the compliance teams.

There are a number of risks associated with non-compliance in terms of the proper identification of persons, entities or transactions related to sanction regimes. This includes the confiscation of financial assets in possession by the regulators, the freezing of assets in possessions of banks or other financial institutions and the confiscation of funds transmitted from cross-border payments. Apart from those risks, there is always the risk of reputational damage — the imposition of financial penalties can be severe and there is also the possibility of revocation of an organisation's license to execute financial services.

Name Screening – Methods and Algorithms

The phrase "name screening" is synonymous with the compliance investigation activity of **identifying the risky data within a transaction set of data**; this requires a number of different name screening methodologies to be applied, and/or matching algorithms, so as to identify the risky data and act accordingly.

Methods of Name Screening

- **Manual recognition of risky data:** The compliance investigator goes through the whole transaction data manually and identifies similarities, or equal data, based on tables or lists — also maintained manually. This is a very lengthy process and cannot be done for high volumes of transactions, messages or data;

- **Manual use of search engines for risky data:** There are compliance investigators that use different search engines to key-in names and other information which will be searched over a database and provide specific matching results. This is a time consuming process, especially with high volumes, but is necessary for ad-hoc or special searches executed during the compliance investigation process. It requires the maintenance of a very analytical and complete database and the search engine algorithms must be specific;

- **Use of batch name matching software:** There is software on the market that can be used to scan transactions and/or customers and—based on the algorithm used—they can identify the risky data within. This is made asynchronously (i.e., done on a specific time, typically at night) and has no effect on the transaction workflow; after the alert is created, it is sent to the compliance investigator for further manual investigation, validation and decision. This method is useful for scanning a customer database, but is not favorable for risk mitigation of money transfers and remittances. The money transfers and remittances are transmitted and then the compliance investigator will find any risky data to identify possible matches with sanction persons, entities, countries or other black-listed persons/entities; and,

- **Use of real-time name matching software:** This type of software is configured in a similar way as the batch name screening software, but the difference is that it uses algorithms to scan data and generate an alert immediately, in real-time; the compliance investigator can/must investigate the alert immediately and decide before the transaction is completed or stop the transaction for further due diligence measures when required. This method is best for money transfers, remittances and for customer on-boarding. It can be configured to provide maximum efficiency for anti-money laundering, counter terrorism financing, and protection from sanctions to the company, as the compliance officer controls the data and transaction workflow.

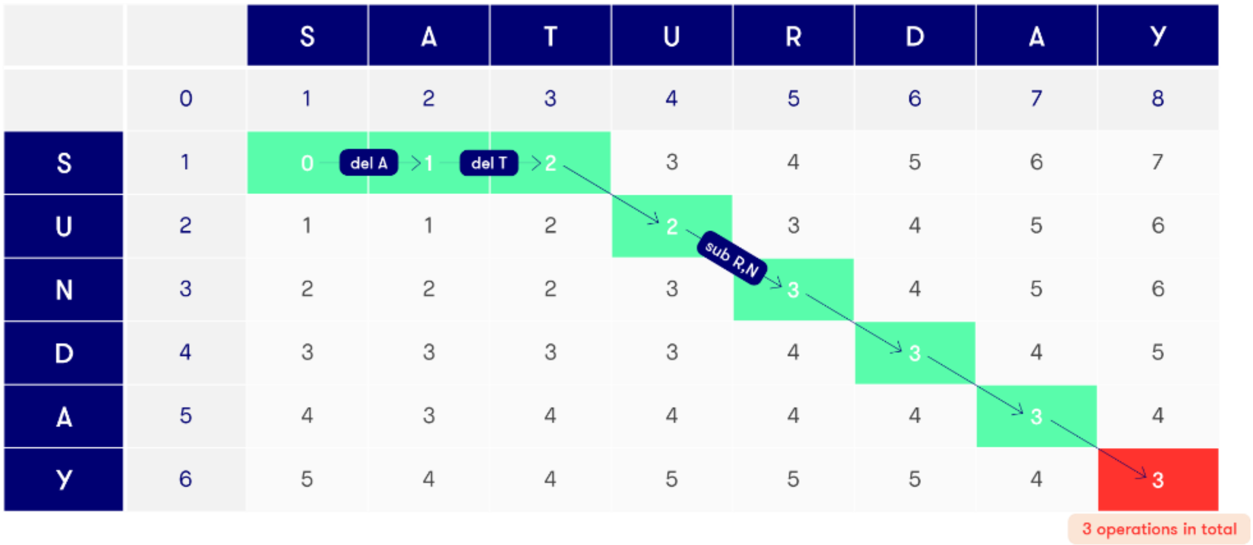
Types of Matching Algorithms and Techniques

There are a number of algorithms that are used to identify risky data, and the selection of each is based on the risk appetite of the company and or the compliance investigation procedure. An algorithm is "*an unambiguous specification of how to solve a class of problems*[5]"; algorithms in computers can perform calculations, data processing and automated reasoning tasks. Therefore there are many different algorithms used in name screening and some commonly used by compliance investigators and software. There were many researchers that used different algorithms to evaluate the results of name matching on names; a comprehensive list set of algorithms used by Gabriel Recchia and Max Lawrence[6] to prove different similarities between languages and names, includes the following:

1. **Edit Distance Measures:**

   Edit distance measures quantify the difference between strings in terms of a sometimes-weighted sum of the number of insertions, deletions, substitutions and/or transpositions required to yield the first string from the second.

   The standard Levenshtein[7] algorithm is the most common edit distance measure, but there are many modifications that can be done in the algorithm, like the Damerau-Levenshtein distance, that additionally counts a transposition between adjacent characters as an edit operation[8].



		S	A	T	U	R	D	A	Y
	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
U	2	1	1	2	2	3	4	5	6
N	3	2	2	2	3	3	4	5	6
D	4	3	3	3	3	4	3	4	5
A	5	4	3	4	4	4	4	3	4
Y	6	5	4	4	5	5	5	4	3

Example of the Levenshtein distance between "Saturday" and "Sunday"

[5] https://en.wikipedia.org/wiki/Algorithm

[6] ACM SIGSPATIAL COMP'13, November 5, 2013. Orlando, FL, USA Copyright (c) 2013 ACM ISBN 978-1-4503-2535-6/13/11.

[7] Levenshtein, V. I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10, 707-710.

[8] Damerau, F. J. 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM, 7, 3, 171-176.

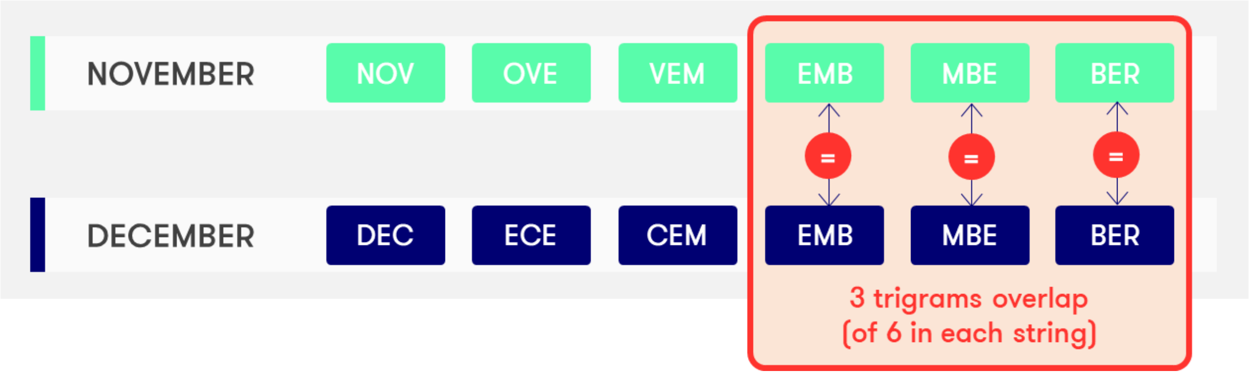
Another measure of edit distance is the Jaro algorithm that defines matching characters when they are the same and their indices are no farther than a value. A common variant, generally referred to as Jaro-Winkler, considers the fact that spelling errors are less likely to occur at the beginning of the names than elsewhere, therefore assigning a higher weight to initial characters.

There are other parameterisable measures that use particular rules of edit operations, all of which are designed to fit a specific purpose. For instance, the list below highlights the key characteristics of some popular edit distance measures:

- The Levenshtein distance is calculated using deletion, insertion and substitution;
- The Damerau–Levenshtein distance uses insertion, deletion, substitution and the transposition of two adjacent characters;
- The Longest Common Substring (LCS) distance uses only insertion and deletion, not substitution; and,
- The Hamming distance uses only substitution when both strings have the same length.

2. **N-gram Measures:**

The n-gram measures count the number of substrings with length "n" (n-grams) that are common in the two strings that are compared. The similarity is obtained either by the division of the count by the number of n-grams (overlap coefficient), or divided by the number in the longer string (Jaccard index) or the average number in both strings (Dice coefficient[9]). Common measures are unigrams (substrings with length 1), bigrams (substrings with length 2), trigrams (substrings with length 3) or skip-grams with gap length 0, 1, 2, etc.

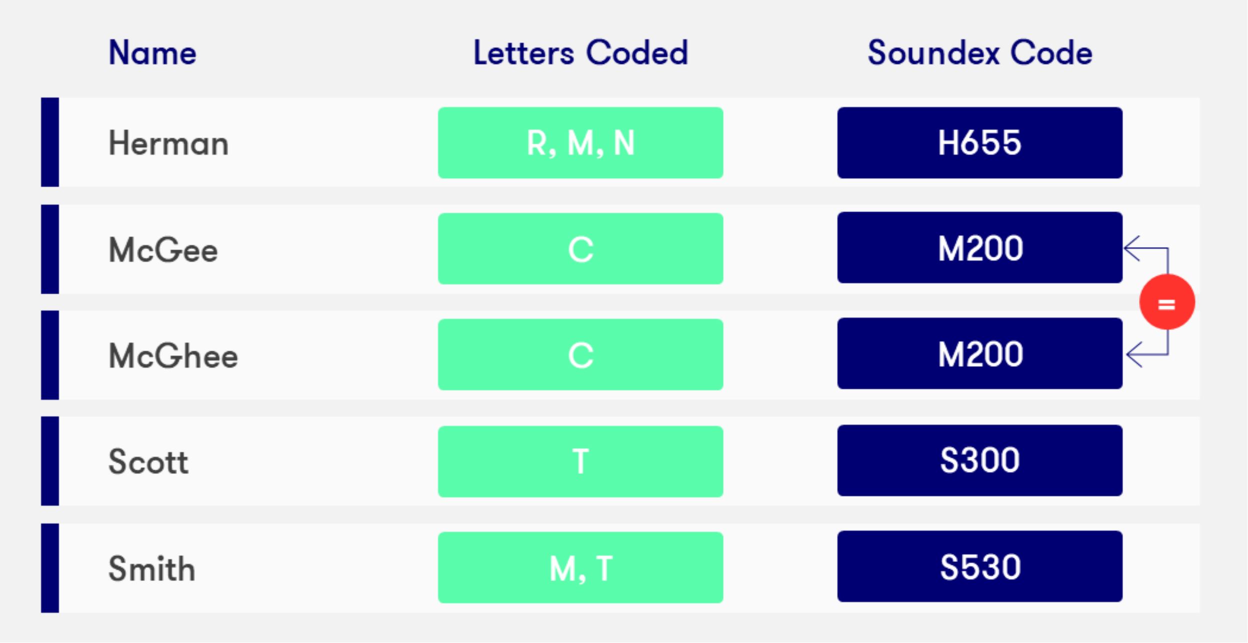


Example with trigrams

[9] Lennon, M., Peirce, D. S., Tarry, B. D., and Willett, P. 1981. An evaluation of some conflation algorithms for information retrieval. Journal of Information Science, 3, 4, 177-183.

3. **Phonetic Measures:**

   There are a number of phonetic algorithms that take into consideration the phonetics and sounds within a name; Soundex is commonly used to match names with difficult conversions to a Latin or English alphabet, therefore the algorithm uses the pronunciation of the word to match (e.g., Arabic or Greek names).



Name	Letters Coded	Soundex Code
Herman	R, M, N	H655
McGee	C	M200
McGhee	C	M200
Scott	T	S300
Smith	M, T	S530

Soundex coding examples

Other Types of Data Matching Algorithms

While name screening algorithms apply for person or entity names, they are equally applicable to toponyms; therefore, the use of multiple variables or objects within the different data strings used during the name screening process is important in order to produce more accurate results.

In name screening there is a possibility to use different algorithms related to "dates" only; some date-specific algorithms include:

- **Same Year, Month, and Day:** This is for checking for an exact date match;
- **Same Year and Month:** This is for checking if the year and month match;
- **Same Year:** This is for checking for a year match; and,
- **Lustrum Approximate:** This is for checking if the year within the date is within a five year period.

Effect of Data Quality on Name Matching Results

In order for name screening algorithms to operate and provide the required results, the computer software using them will require two separate sets of data: a) the lists, tables or database wherein the static name data is found (e.g., public UNSC Resolution Sanctions Lists, OFAC SDN Lists, EU Sanctions Lists, etc.) or b) paid-for databases containing names of persons and entities that have been published by official resources to be of compliance interest (e.g., Thomson Reuters' World Check[10], Dow Jones[11], Lexis Nexis[12]).

Data Quality of Lists

Different public/paid lists or databases of persons and entities related to sanction regimes, financial crime convictions, adverse media publicity, law enforcement investigations include different types of data and information. This data may include names (first name, middle name, last name), dates (date of birth, publication dates, ID issue dates), country of nationality, identification document numbers and passport numbers, known addresses of residency, etc. The quality of this data, and the consistency of storage and representation, is critical when used by different matching algorithms.

In case of inconsistency of the database structure, the results from the matching engine may not produce what is expected and may result in false alerts. For example, the name structure may be different for separate records, such as the first name, middle name, and last name being misplaced between data fields. In this respect, the matching algorithm has to be "wide" enough to capture the majority of this misplaced data and match them properly. The result may not be what the compliance investigator is seeking.

Data Quality on Input of Primary Data

A financial services provider uses its core systems to record the data captured for its customers and their transactions; this means that the way the users capture the data, logged in the different fields, is critical for the appropriate use of the name screening process. For example, required static customer data like "name," "address," "identification" and "nationality" may not be logged in the appropriate fields, logged in the correct sequence, not completed at all or mistakenly logged.

The matching algorithms may be unable to produce correct or adequate alerts; therefore, the process of name screening will be incorrect or produce inappropriate results. This is one of the biggest risks that the company will face even if it has implemented the appropriate computer software.

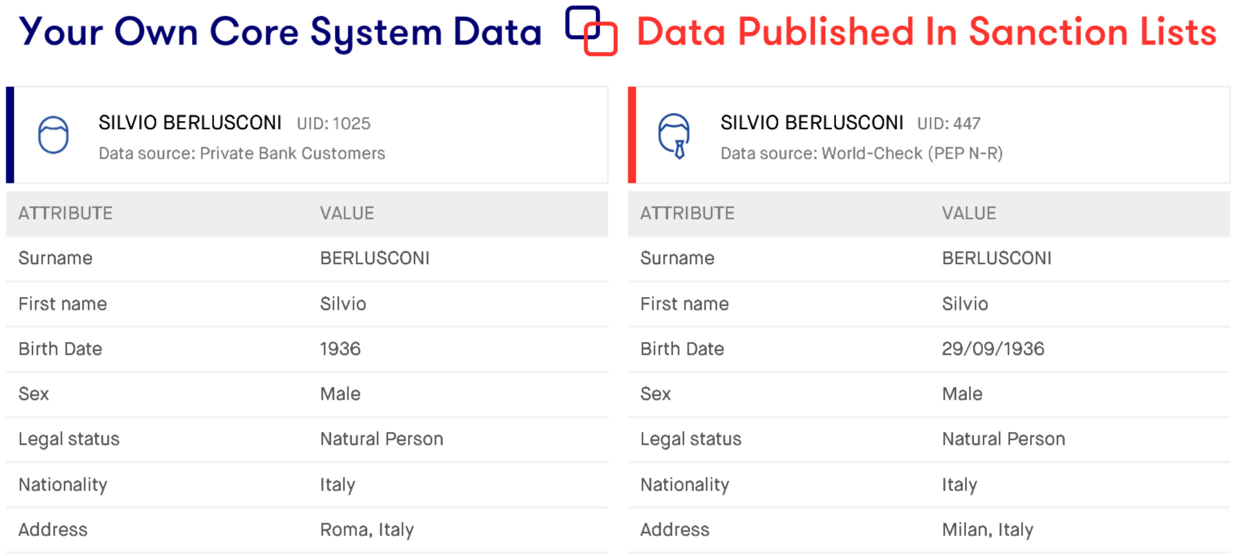
[10] http://risksolutions.thomsonreuters.com/world-check-global?utm_source=google&utm_medium=cpc&utm_campaign=413706060905

[11] https://www.dowjones.com/products/risk-compliance/

[12] https://www.lexisnexis.com/en-us/products/lexis-diligence/sanction-peps-and-watch-list-verification.page

Matching Results, Alerts and their Benchmarking

Basically, the matching engine used during the name screening process compares two sets of data:



Your Own Core System Data | Data Published In Sanction Lists

SILVIO BERLUSCONI UID: 1025
Data source: Private Bank Customers

ATTRIBUTE	VALUE
Surname	BERLUSCONI
First name	Silvio
Birth Date	1936
Sex	Male
Legal status	Natural Person
Nationality	Italy
Address	Roma, Italy

SILVIO BERLUSCONI UID: 447
Data source: World-Check (PEP N-R)

ATTRIBUTE	VALUE
Surname	BERLUSCONI
First name	Silvio
Birth Date	29/09/1936
Sex	Male
Legal status	Natural Person
Nationality	Italy
Address	Milan, Italy

During the name screening process, the application of the matching engine must deliver a result based on a set of different data elements. These element are configurable within the name screening process, and can include name, nationality, country or address of residency, identification documents, etc.

The result of this name screening process is categorised into four different cases:

	Screening Result	
	Match ALERT	No Match NO ALERT
Customer is on Sanction List	True Positive	False Negative
Customer is not on Sanction List	False Positive	True Negative

When the name screening process reports a match between a customer and an entity on a sanction list, an alert is generated. If this customer corresponds to one of the entities on the sanction list, this case is called a **true positive** (i.e., the alert is justified). However, a name screening can also be prone to errors. Thus it can happen that a match is reported by the name screening, but the flagged customer does not correspond to any entity on the sanction list. This case is called a **false positive**, because the name screening incorrectly classified a customer as being on the sanction list.

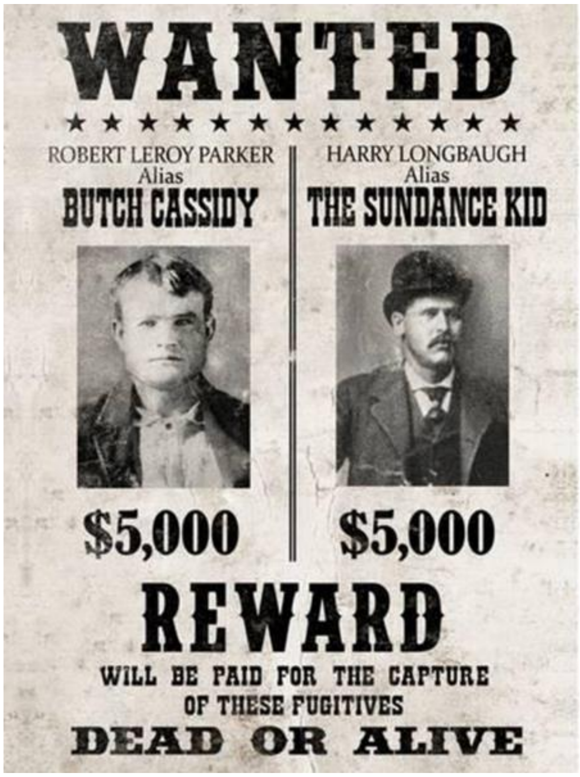
When the name screening process reports no match between a customer and an entity on a sanction list, no alert is generated. If this customer does not correspond to any of the entities on the sanction list, this case is called a **true negative** (i.e., the alert is justified). However, a name screening can also make a second kind of error. Thus it can happen that no match is reported by the name screening, but the screened customer does in fact correspond to an entity on the sanction list. This case is called a **false negative,** because the name screening incorrectly classified a customer as not being on the sanction list and thus missed detecting an eventual high risk.

Due to either data quality issues, inappropriate matching algorithms or matching configuration for the name screening process, both kinds of errors—false positives and false negatives—can be severe to the extent that compliance investigators lack effective risk management and efficiency in their investigation process.

Therefore, the basic requirement of a compliance investigator is to have a more effective screening system that produces meaningful and reliable alerts, with the result being a more effective risk management function. During the real-time name screening procedure, the time factor plays a major role — spending time on false positive alerts does not help the effectiveness of the compliance function and is thus counter-productive.

The History of Name Matching

The history of financial crime goes way back in time. It has transformed through the years and evolved with the use of modern technology but, ultimately, it has the same scope: *to make use of the money received through illegal activities or channels*. The Law Enforcement Agencies of all the countries in the world identify, arrest, prosecute and publish data about criminals, direct and indirect money launderers, terrorists and other types of condemned persons and entities. Moreover, illicit funds that are derived from drugs, guns and arms, human trafficking, hacking, cyber-crime, etc. have been logged by the Law Enforcement Agencies throughout the world, and persons, entities, organisations and cartels have been identified and listed.



The origin of name screening?

All these lists were—and are still—used to identify any transactional activities and funds belonging to these persons. The different sanction regimes are imposing controls over the banks and non-banking financial service providers, and are requesting the screening of the data within the financial transactions to identify possible matches with the listed entities. Initially through manual processes, but eventually the name screening procedure has been automated to a great extent.

Back in the early 2000s, the systems used the name screening automated procedure ("**batch**" process) to periodically identify listed persons and entities from the financial service provider's databases. The alerts generated were investigated by the compliance officers, and any true positive match was marked in the database as "high risk" so as to be monitored closely throughout the relationship. This procedure required a significant number of manual resources to comply with the legal and regulatory requirements.

Later, due to the imposition of controls deriving from FATF Recommendations[13] (the so-called "40 + 9 FATF Recommendations") the financial service providers – especially banks – applied near real-time automated procedures to identify any connection of cross border payments with any listed person or entity. They used the so called "SWIFT scanning" due to the fact that the majority of the cross-border payments they executed were via SWIFT[14] messages. The term "**near real-time**" is used because the transaction was completed within the core transactional system — the SWIFT message was created, but was stopped for validation before being sent to the receiver. If there was any match on any list, the message was blocked and the transaction reversed or dealt with by the compliance officer according to the risk and/or internal workflow procedure.

Due to heavy legal and regulatory pressure on financial service providers coupled with the imposition of heavy fines on large banks and other financial service providers for violation of sanction regimes and involvement in money laundering, the last 5+ years have seen financial service providers investing in "**real-time**" name screening on almost all types of transactions (not just those related to SWIFT transfers). The transaction is screened against lists during and before the completion of the transactional workflow within the core transactional system. This provides an additional control measure against the use of the financial service provider for money laundering, terrorism financing or any other type of financial crime. The heavy de-risking measures taken by correspondent banks throughout the world also make real-time name screening a basic requirement — funds are frozen more easily and correspondent accounts and relationships are closed if there are no proven real-time screening controls before executing a transaction and sending the cross-border payment.

[13] FATF (2012) - International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation, updated October 2016, FATF, Paris, France – www.fatf-gafi.org/recommendations.html

[14] https://www.swift.com

Efficiency of the Name Screening Process

Within any financial services provider, there is always a conflict between operations and compliance in terms of the implementation of controls and customer service efficiency. This is because transactions are often processed faster than the controls implemented. What is the efficiency of the name screening process within a financial services provider and how complete should the control function be?

Implementation of Adequate Automation

The selection of computer software to implement an appropriate automated, real-time name screening procedure is paramount for every financial services provider. According to regulatory and correspondent bank requirements, financial services providers must implement automated sanctions identification, financial crime investigation and adverse publicity notification procedures in order to ensure that they are not used for any type of money laundering and terrorism financing activity. The configuration of the algorithms used during this name screening automated procedure depend heavily on the capability of the system being used.

The second variable in the implementation of the adequate automation procedure are the lists used and how they are used. In addition to sanction regimes and designated person & entity lists published by the UN Security Council and G20 countries, there are a number of paid list providers on the market. Any computer software, therefore, must be able to efficiently use both public and paid lists.

Minimisation of False Positives

As mentioned earlier, detected elements from the different algorithms used within the automated name screening procedure can result in true positive or false positive alerts. True positive alerts are used by compliance investigators to identify money laundering or terrorism financing risks. False positive alerts require additional time for the compliance investigator to verify & validate elements and decide if there is a risk of money laundering or terrorism financing.

Too many false positive alerts increase the possibility of further errors during the compliance investigation procedure and result in the reduced efficiency of the compliance investigation process; the effect is the depletion of important resources along with added costs. Therefore, the ultimate task of the compliance officer is to reduce false positive rates, increase the effectiveness of its name screening process, increase the efficiency of the compliance investigation procedure and reduce the compliance risk & cost.

However, the classification of false and true positives require knowledge and experience. In some cases this can be subjective by nature — some people would consider virtually any reported match as a true positive while others would consider it as a false positive. For example, would you consider **"S.A.S E.U."** matching with **"SERVICIO AERO DE SANTANDER E.U."** as a true positive or a false positive? Is **"SOMEX S.A."** matching with **"SOMEX L.L.C."** a true positive or a false positive?

Ultimately, the responsibility of proper classification lies with the business expert or the compliance officer — they must liaise with solution providers or internal IT to define precisely what is considered as a true positive or a false positive. A definition of the matching policy is necessary in order to define an acceptable level of matching based on available data. The compliance officer must understand the different languages that the data originates from, and then—with the assistance of data scientists—execute realistic tests to determine optimum classification levels based on real-life statistics.

Decreasing the threshold on traditional algorithms, like the Levenshtein distance or Soundex, is not the solution. Doing so results in a significant increase in the amount of alerts, with the false positive rate increasing exponentially while the false negative rate decreases only linearly.

In his article "False Positive is not a pain area for Financial Services[15]," Dr. S. Ambiga observes that a compliance officer may investigate thousands of alerts, but only a few are filed with the Financial Intelligent Units as suspicious — this means that their false positive rate could be quite high compared to the true positive rate. He concludes that "*wrong way false positive is reducing the matching score and risk score,*" and he continues that the "*right way of reducing false positive is to implement better algorithms, fuzzy logic and analytics.*"

[15] https://www.ponsun-amlacademy.com/false-positive-is-not-a-pain-area-for-financial-institutions/

Artificial Intelligence and Machine Learning

Machine learning is an application of artificial intelligence that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed[16]. In a simplistic way, machine learning is the automated methodology used in order to compute results in such a way that will be repeated without actual human intervention and adjust actions accordingly.

Machine learning has been instrumental in solving important business problems, such as detecting e-mail spam, providing focused product recommendations, determining accurate medical diagnoses, etc. The adoption of machine learning has been accelerated with increased processing power, availability of big data and advancements in statistical modeling. In short, machine learning converts data intensive and confusing information into a simple format that suggests actions to decision makers.

In the name screening process, machine learning enables investigators to attain a deeper assessment of the similarity between pairs of data objects. Traditional matching algorithms, which rely on pre-defined similarity measures, typically focus on a single aspect of similarity between two names (e.g., pure syntactical similarity). In contrast, machine learning simultaneously examines similarities from a multitude of angles and effectively combines hundreds of attributes between two instances to produce a more sophisticated estimate of the true overall matching similarity between two instances. For example, machine learning can assess a pair of names taking into consideration all of the following attributes:

- Smart syntactical matching;
- Smart phonetic matching;
- Ethnicity awareness; and,
- Smart semantic matching.

In comparison with traditional matching algorithms, the use of machine learning results in drastically higher quality and performance in the matching process, meaning **fewer errors** due to:

- Reduced false positives, and
- Reduced false negatives.

The fight against money laundering, terror financing and sanctioned individuals is based on lists of millions of names of individuals and companies that financial companies must not deal with. Yet even if these extensive, ever-changing lists were accurately compiled, it would be challenging to check them against client databases in near real time — the names in client databases often feature inaccuracies (e.g., typos) and are often in non-standardised formats.

In addition, there are no globally accepted rules for the transcription of words written using non-Latin characters (e.g., Chinese, Cyrillic, Arabic, Taiwanese, etc). For example, someone with British, Chinese and Egyptian heritage might have their name written "Mohammed Lee-Smith," "Li-Smyth, Muhamet," "Mohammed Leesmith" and so

[16] https://www.expertsystem.com/machine-learning-definition/

on. Searching for company names can be harder still as there is even less standardisation.

Until now, system designers have typically employed logic that scores the syntactical match between names and uses thresholds to define which names are considered as matching those in the system. The major disadvantage of these traditional processes is the high occurrence of false positives.

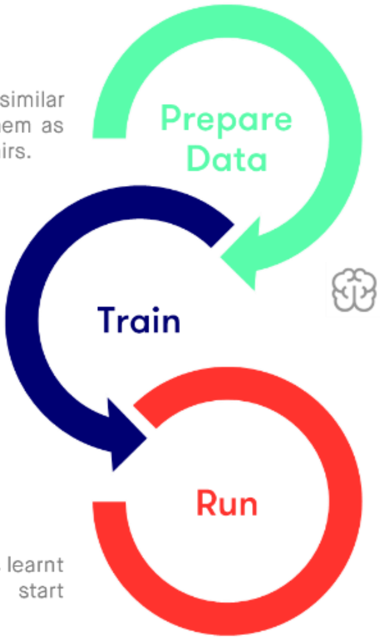
Hence, the only long-term efficient solution is machine learning.

How False Positives are Reduced in idetect

What makes idetect® different is how its machine learning capabilities can scan large data sets and quickly retrieve multiple aspects of name pairs to deal with the trickiest matching occurrences. idetect®'s machine learning engine is able to automatically evaluate the probability of a match using a pre-trained model that combines hundreds of name-matching features. The parameters of the model can self-tune after being trained with labelled samples of challenging name pairs (i.e., few of them being marked as matching, few others being marked as non-matching).

1. Prepare Data

Prepare a sample of tricky pairs of similar names in XLS/CSV format. Label them as either matching or non-matching pairs.



2. Train

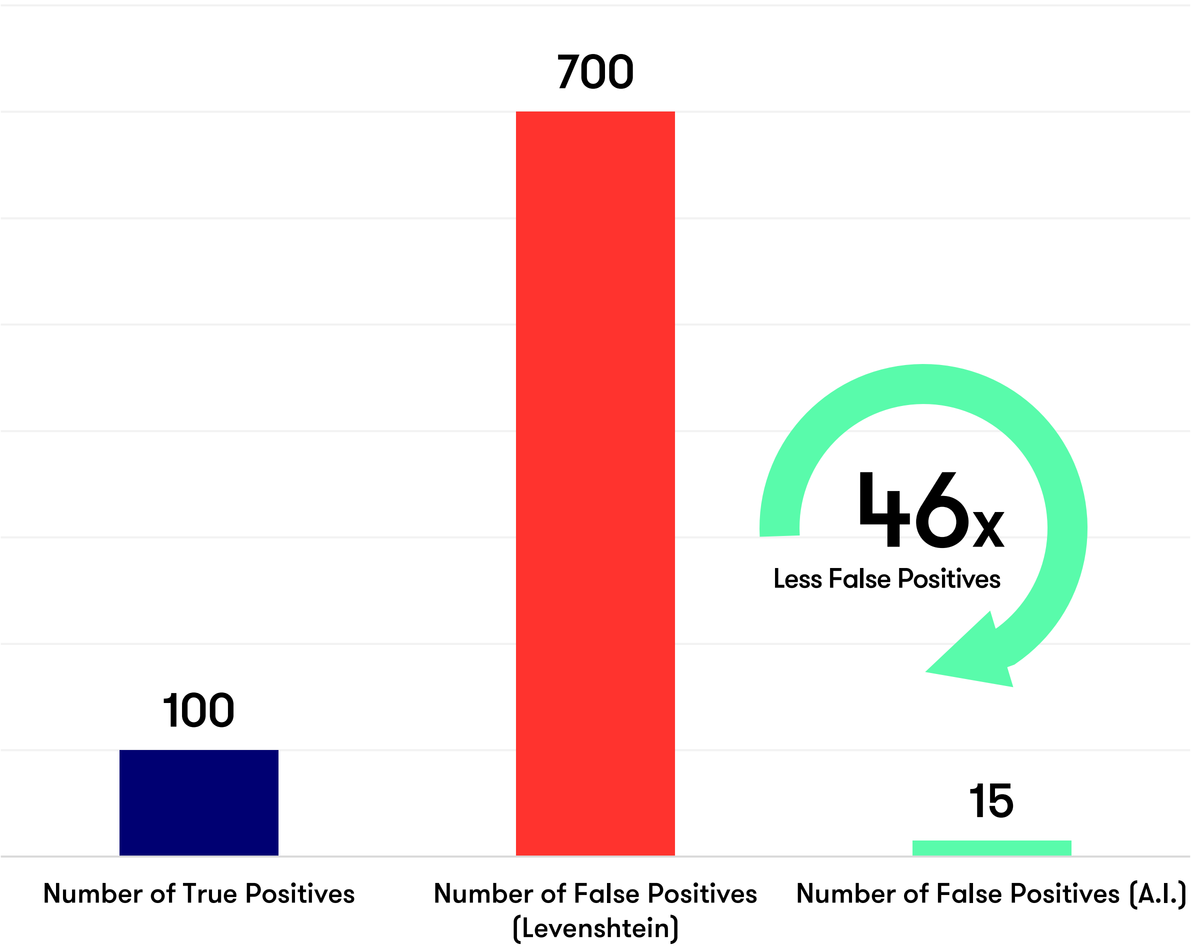
Upload your labelled sample and train the build-in model (hundreds of name matching features will be compared for each name).

3. Run

Apply the new similarity parameters learnt by the model's algorithms and start screening your customers.

Test cases conducted in idetect®'s lab with the usage of machine learning proprietary techniques demonstrated improvements in precision as much as a factor of **forty-six** by reducing the number of false positives without increasing regulatory risk.

A.I. vs. Levenshtein Operational Efficiency with Constant Regulatory Risk



Non-Permuted Full Names
Dataset: ~ 3'000'000 Customers
Regulatory Risk: False Negative Rate < 0.5%

Such results are significant because every false positive adds cost due to the need for human intervention and reduces service quality as processing times are slowed. It is worth noting that these results are all the more difficult to achieve given that the risk of missing genuine matches should not be increased. That is precisely where the challenge lies and what simple/traditional matching approaches fail to achieve: producing minimal false positives without increasing the risk of false negatives.

This problem is explained by two well-known factors/measures that enter in conflict in the field of information retrieval:

- **Recall:** Represents the percentage of true positives over the total amount of relevant elements (= TP / (TP + FN)). This percentage is a great indicator of your **regulatory risk** since it is a good way to measure the strength (or weakness) of a system to detect all true positives; and,

- **Precision:** represents the percentage of true positives among the detected elements, including those that are not relevant (= TP / (TP + FP)). This percentage is a great indicator of your **operational efficiency**. A low precision means too much noise generated by the system that pollutes and slows down the work of the compliance investigators.

Improving precision without impacting the recall—and so the regulatory risk—and vice-versa, is scientifically not possible. Precision will always "fight against" recall, hence it can only be limited. Fortunately, machine learning can significantly reduce false positives without increasing the regulatory risk for the compliance officer (i.e., at a constant recall). This breakthrough is achieved by improving the precision by which "fuzzy name" matches and non-matches are identified. idetect®'s machine learning capabilities enables organisations to combine the best features of existing complementary algorithms paired with novel features to then achieve further advances. Another great advantage of machine learning systems is that they do not keep making the same mistakes over and over again: the more you train and run the model, the better the results.

Machine learning is the one and only way to break the vicious relationship between the number of false positives and the number of false negatives, by a factor that all other preceding approaches have consistently failed to get close to. Machine learning might not yet be the solution of the timeless problem of "squaring the circle," but compliance officers have never been that near to do the impossible.

Conclusion

The majority of automated name matching systems employ legacy matching algorithms—using pre-defined similarity matches—to identify matches between the different data sets. These algorithms, though, use pre-defined matching thresholds and therefore have substantially higher false positive rates. The results presented to the compliance investigator require difficult decisions and actions to determine the actual compliance risks. Investigations require a myriad of tests to calculate acceptable false positive and false negative rates, which the compliance officer will then include in the company's risk assessment policy and risk appetite.

The implementation of machine learning in a new generation of name matching systems enables investigators to increase the number of compared features, resulting in a deeper match. idetect®'s proprietary machine learning engine comes out of the box with a pre-trained "cold" model. All the benchmarks performed with this generic model have proven that idetect® delivers outstanding results in comparison with all other name screening solutions in the industry. It puts the long-standing problem of name screening on a totally different paradigm in terms of how to meet the compliance officers' requirements for low false positive rates in a real-time name screening environment.

In addition, further improvements can even be achieved following a deep assessment of your own data. The idetect® machine learning solution offers the advantage of delivering optimum results when the model is warmed up with custom training data sets that take into account all the specifics of your core system's data. idetect®'s analytics experts can assist you in training the model with your own data to maximise the potential of idetect®'s machine learning technology. In this way, compliance officers increase the effectiveness of their investigations and are empowered to concentrate on the efficiency of their overall risk mitigation programmes and processes.

Company Profile

idetect® is a next generation software for Enterprise Fraud, Anti-Money Laundering, Transaction Monitoring, Know-Your Customer (KYC) and Client Onboarding, and Watchlist Monitoring, which *provides the latest and most efficient technological features against financial crime and illicit transactions.*

LOGOS ITS S.A., a company based in Luxembourg and Germany, is the editor and distributor of the technology. Our company delivers high-quality services and solutions for market leaders in the area of finance, industry, and government. With 20 years of experience and an average of 65 highly skilled employees in the 3 last years, the reliability and stability of the team has allowed to collaborate, establish partnerships and official agreements with some of the largest institutions in Europe and the World.

LOGOS ITS has numerous prestigious clients among which Agricultural Bank of China (ABC), China Merchants Bank (CMB), Crédit Agricole-Caisse d'Épargne Investor Services (CACEIS), Banque et Caisse d'épargne de l'État Luxembourgeois (BCCE), Arcelor Mittal, LuLu International Exchange, Bahrain Financing Company, Wafacash, CIHBank, Swisscard and Deutsche Börse Group (Clearstream Services, Regis-TR, Deutsche Börse Security Services, Eurex).

Our company is also investing heavily into research and innovation including a specific partnership framework with the Science University of Luxembourg and the Ministry of Economy. Researches focus on machine-learning and artificial intelligence to combat financial crime.

No legal or accounting advice is provided into this Whitepaper and any discussion of regulatory compliance is purely illustrative.

Contact us at:
info@idetect-soft.eu | +352 26 36 55-1

idetect