Sanctions Screening: The Needle Paradox

Finding one terrorist in a million transactions – while 999,999 false alarms drown the signal

Digital Finance

BSc Digital Finance Course

Why Does Every Bank Spend $500M on Compliance and Still Miss Sanctions Violations?

The Needle Paradox

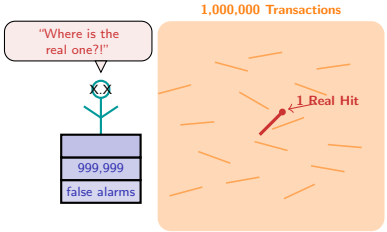
Every international wire transfer passes through a screening engine that checks names, entities, and countries against global sanctions lists. The goal: intercept one terrorist or sanctioned state actor hiding among millions of legitimate transactions.

What the screening system must do:

- Screen billions of transactions per year against 20,000+ listed names
- Flag any match for human review before funds move
- Operate in real time without delaying legitimate payments
- Comply with OFAC, UN, EU, and UK sanctions simultaneously

What actually happens in practice:

- 95–99% of alerts are false positives – innocent people with common or similar names
- Analysts spend most of their day clearing alerts that pose zero risk
- True hits are buried under thousands of noise cases
- Banks spend more on alert clearing than on finding actual violations



The screening signal exists.
The noise is the problem.

Banks screen billions of transactions per year but 95–99% of alerts are false positives – the system is overwhelmed by noise before it can find the signal.

What If Your Wire Transfer Were Frozen for Three Weeks Because Your Name Matches a Sanctioned Person?

Reflection Prompt

You wire CHF 4,200 for rent to a landlord named "Hassan Al-Rahman." The bank's screening engine flags the transfer. Your money is frozen while an analyst reviews the alert. The landlord threatens eviction. **1.** Was the freeze justified? **2.** How long is a "reasonable" review delay? **3.** Who compensates you for the error – and how?

This scenario happens thousands of times a day across global banks. The screening engine cannot distinguish between a sanctioned entity and an innocent person sharing a similar name. So it flags both – and humans must sort them out under time pressure.

The real-world stakes:

- **For individuals:** delayed rent, missed payroll, frozen aid transfers to conflict zones – legal, urgent payments stopped by an algorithm
- **For businesses:** trade finance held, supplier relationships damaged, contracts missed – costs that are invisible to the regulator
- **For correspondents:** banks sometimes refuse entire corridors ("de-risking") rather than screen every transaction individually
- **For the analyst:** clearing 300 alerts a day, knowing one real hit buried in the queue could mean criminal liability if missed

Bring your reaction to class. What is the acceptable false positive rate when the cost of a false negative is enabling terrorism financing?

The false positive is not a technical inconvenience – it is a real cost paid by real people, often the most financially vulnerable, who cannot navigate a compliance dispute.

What Is the Difference Between a Sanctions List, a Watch List, and a PEP Database?

Dimension	Sanctions List	Watch List / Adverse Media	PEP Database
Authority	Government / UN	Commercial / internal	Commercial / regulatory
Status	Legal prohibition	Risk indicator	Enhanced scrutiny
Consequence of match	Transaction blocked	Trigger EDD review	Apply enhanced due diligence
Coverage	Named entities, vessels, regimes	Criminals, suspects, PEPs	Politicians, officials, family
Update frequency	Real-time (OFAC) to weekly (UN)	Continuous	Monthly–quarterly
False positive risk	Very high (name similarity)	Moderate (keyword hits)	Moderate (common names)
Typical list size	10K–25K entries	500K–2M entries	3M–10M entries

Three distinct compliance obligations

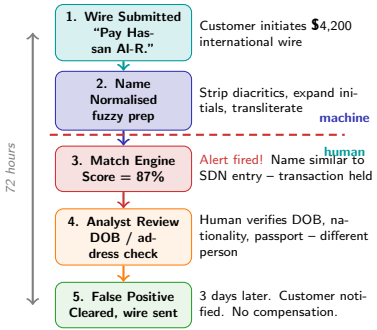
- **Sanctions lists** are legal prohibitions – transacting with a listed party is a crime regardless of intent. OFAC, EU, UN, and UK lists must all be checked simultaneously
- **Watch lists** and adverse media are risk signals, not prohibitions. A match triggers investigation, not automatic blocking
- **PEP databases** cover politically exposed persons – not prohibited, but subject to enhanced due diligence (EDD) because their position creates corruption risk

The compounding problem:

A bank screening against all three layers simultaneously generates overlapping alerts. The same person may appear in all three categories, generating three separate alerts requiring three separate dispositions.

Sanctions, watch lists, and PEP databases are legally distinct – but screening systems must check all three simultaneously, multiplying alert volume with each additional layer.

Follow One Wire Transfer Through the Screening Engine – and Count the False Alarms?



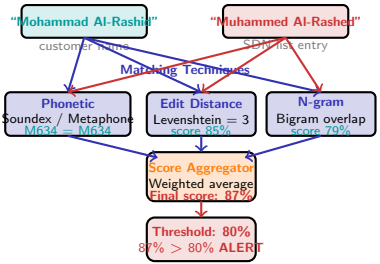
Anatomy of a false positive lifecycle

- **Name normalisation** is the first line of defence: stripping diacritics, expanding initials, and phonetic encoding all widen the matching net – but also generate more false positives
- **Match score 87%** triggers the alert because the bank's threshold is 80%. A score of 79% would have passed automatically – a single percentage point separates a frozen account from a completed payment
- **Analyst review** relies on identifiers the customer may not have provided: date of birth, passport number, nationality. If those fields are empty, resolution takes longer
- **Release with no compensation** is standard practice. Banks are legally protected for good-faith screening delays – the customer bears all costs of the error

The irony: the 87% match score sounded like it was probably right. It was not. And the analyst who cleared it in three days still reviewed 299 other alerts that day.

A false positive that takes 72 hours to clear is not a system error – it is the system working exactly as designed. The design is the problem.

How Does Fuzzy String Matching Find 'Mohammad' When the List Says 'Muhammed'?



Three matching strategies, one score

- **Phonetic matching** (Soundex, Metaphone) converts names to codes based on how they sound. "Mohammad" and "Muhammed" both encode to M634 – identical phonetically
- **Edit distance** (Levenshtein) counts the minimum insertions, deletions, and substitutions needed to transform one string into another. Three edits for an eight-character name gives a high similarity score
- **N-gram overlap** counts shared character sequences. Bigrams like "Mu", "uh", "ha", "mm" appear in both spellings

The threshold is everything:

- Lower the threshold: catch more real hits but drown in alerts
- Raise the threshold: reduce alerts but risk missing real sanctions
- The "right" threshold does not exist – it is a policy choice between operational capacity and legal risk

Fuzzy matching is not a failure of technology – it is a deliberate design choice that trades false positives for coverage, because a missed sanction carries criminal liability.

What Happens When the Screening System Learns to Ignore Alerts – and Misses a Real One?

Alert fatigue: the systemic risk inside compliance
When analysts clear hundreds of false positives daily, a predictable pattern emerges: the human review becomes perfunctory. Analysts begin pattern-matching to close alerts quickly, rather than investigating each one carefully.

The degradation pathway:

- **Day 1:** Analyst reviews each alert thoroughly – 20 minutes per case, 15 alerts cleared per day
- **Month 3:** Analyst learns most alerts are false positives – applies shortcuts, 5 minutes per case, 60 per day
- **Month 6:** Analysts develop "disposition patterns" – similar names cleared without checking all identifiers
- **The miss:** A real sanctioned entity with a common name is cleared via pattern match – the shortcut that worked 999 times fails on the one case that matters

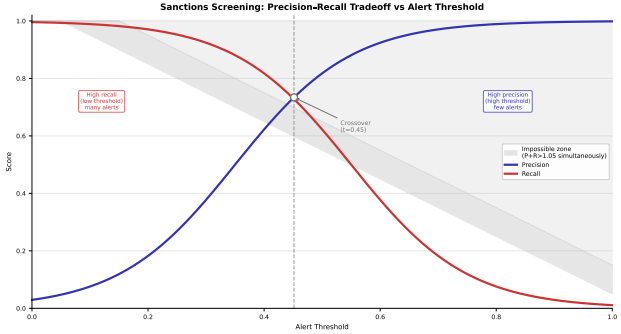
Real-world consequence:

- OFAC fines have reached **$**1B+ for systemic alert mishandling
- The failure is attributed to "lack of controls" – but the root cause is an unmanageable alert volume that made careful review impossible in practice

The system worked too well at generating alerts.
That is why it failed.

Alert fatigue is not a human weakness – it is a predictable, measurable outcome of designing a system that generates more alerts than humans can carefully review.

Where Does Screening Accuracy Break Down – and What Drives False Positive Rates?

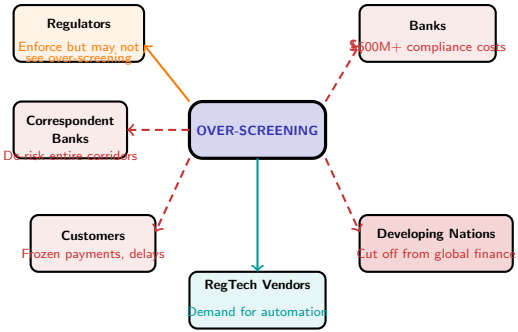


Reading the precision–recall tradeoff

- **Precision** (purple) rises as the alert threshold increases: at a high threshold, the alerts that do fire are more likely to be real hits – but many real hits are silently missed
- **Recall** (red) falls as the threshold increases: a high threshold means fewer total alerts, but the system misses more sanctioned parties who scored just below the cutoff
- **The crossover point** is where precision equals recall. Most banks operate to the left of this point – they prefer high recall (catch everything) even at the cost of massive false positives
- **The grey band** marks the "impossible zone" where both precision and recall cannot simultaneously be high. The fundamental tradeoff cannot be engineered away
- **The policy implication:** choosing a threshold is not a technical decision – it is a regulatory and ethical one about acceptable risk on both sides of the tradeoff

Illustrative estimates. The precision–recall tradeoff is not a flaw in the algorithm – it is a mathematical inevitability that forces compliance teams to choose between two kinds of error.

Who Pays the Cost of Over-Screening – Banks, Customers, or Entire Countries?



Over-screening costs fall on:

- **Banks** spend $500M–$1B per year on compliance infrastructure – analyst salaries, screening software, fines for false negatives when they do occur
- **Customers** bear the direct cost of delays, frozen accounts, and denied services with no legal right to compensation
- **Developing nations** are most exposed: correspondent banks withdraw from high-risk corridors entirely ("de-risking"), leaving entire populations cut off from dollar-clearing

Who benefits:

- **RegTech vendors** see surging demand for AI-driven false-positive reduction tools

Ambiguous:

- **Regulators** enforce compliance rules that create the over-screening incentive but rarely penalise over-screening itself

The hidden cost of over-screening is not paid by the banks – it is paid by the people and countries least able to absorb it, often through loss of access to basic financial services.

The Screening Optimization Framework: How Do You Tune for Risk Without De-Banking the Innocent?

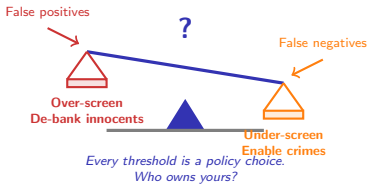
When evaluating any sanctions screening program – as a compliance officer, regulator, or system designer – apply these four questions:

1. What is your current false positive rate?
If it is above 95%, the alert queue is unmanageable by design. A system generating 10,000 alerts to find 10 real hits is not screening – it is generating liability cover for the bank while creating harm for customers.

2. What identifiers resolve alerts fastest?
Date of birth and passport number resolve 80% of false positives instantly. If your onboarding process does not capture these, you are guaranteeing long resolution times before screening even starts.

3. Who bears the cost of each error type?
A false positive costs the customer time and money. A false negative costs the bank a regulatory fine and potential criminal liability. Are these costs calibrated, or does the bank externalise all false positive costs onto customers?

4. Is your threshold a policy choice or an accident?
Most alert thresholds were set years ago and never reviewed. Does someone in the organisation own this number, and is it reviewed against current false positive rates?



Sanctions screening is not a technical system with a policy wrapper – it is a policy system built on technical infrastructure. The threshold is a governance decision, not an engineering one.

Your Challenge: Redesign an Alert Triage Workflow to Cut False Positives by 50%

Mini-Challenge (15 minutes)

A regional bank screens 200,000 transactions per day. Its current system generates 8,000 alerts daily. Analysts clear 7,900 as false positives. The remaining 100 are escalated. Of those, 2 are confirmed sanctions violations per month. The bank wants to cut alerts by 50% without missing any real violations. What do you recommend?

Apply the four-question framework:

1. **What is the false positive rate?** Calculate it. Is 7,900/8,000 = 98.75% normal, high, or catastrophic? What industry benchmark would you compare against?
2. **What identifiers could resolve alerts faster?** If date of birth were collected at onboarding, how many of the 7,900 could be auto-cleared without analyst review? What data would you add to the customer onboarding form?
3. **Who bears the cost of each error type?** If the bank raises its threshold by 5 percentage points and misses one real sanction per year, what is the expected OFAC fine? If it over-screens and delays 500 legitimate payments per day, what is the customer harm?
4. **Is the threshold a policy choice?** Who in your bank should own the alert threshold number – the compliance officer, the CRO, the board? What governance process should surround changing it?

Discuss: What would you recommend first – raise the threshold, improve data quality, or add a machine learning pre-filter?

The four-question framework applies to any screening system – payment screening, trade finance, correspondent banking, or crypto. The numbers differ; the paradox is the same.