Wenn Ermittlungen auf 100 Millionen Dateien treffen

Lessons Learned beim Aufbau skalierbarer KI-gestützter eDiscovery-Systeme

Teil 1 – Problem & Konzept

Business-Realität, Datenlandschaft, Suchmechanismen

Teil 2 – Technische Umsetzung

Architektur, Pipeline, Skalierung, Governance

DIC Konferenz 2026

Prof. Dr. René Brunner & Dr. Ezra Tampubolon

TEIL 1 – PROBLEM & KONZEPT

Realität moderner Ermittlungen

Ermittlungen sind heute keine Aktenberge mehr – sie sind Datenfluten. Die Quellen sind fragmentiert, die Volumina wachsen ungebremst, und die relevanten Informationen verstecken sich in einem Rauschen aus Millionen von Dateien.

Viele Quellen gleichzeitig

E-Mail-Server, Cloud-Dienste, Mobile Devices, interne Systeme – alles gleichzeitig, selten koordiniert gesichert. gesichert.

Verteilte Kommunikation

Relevante Aussagen finden sich in Teams-Chats, WhatsApp-Verläufen, Slack, Signal – nicht mehr nur in der E-E-Mail.

Massiv wachsende Datenmengen

Moderne Unternehmen produzieren täglich Millionen Dokumente. Ein einziger Custodian kann mehrere TB Daten hinterlassen.

Ermittlungen sind heute datengetrieben – nicht mehr dokumentengetrieben.

TEIL 1 – PROBLEM & KONZEPT

Datenlandschaft: Heterogen und unstrukturiert

Das Spektrum der zu verarbeitenden Daten geht weit über klassische E-Mail-Archive hinaus. Jedes Format bringt eigene technische Anforderungen und Fallstricke mit sich.

E-Mails & Chats

PST, EML, MBOX, Teams-Exports, Slack-Archive – unterschiedliche Formate, unterschiedliche Metadatenstrukturen.

Mobile Backups

iOS- und Android-Backups enthalten Nachrichten, Kontakte, App-Daten – teils verschlüsselt, teils fragmentiert.

Cloud-Systeme

SharePoint, OneDrive, Google Drive, Dropbox – Versionshistorien, Zugriffsrechte und API-Limits erschweren die Extraktion.

Audio & Video

Konferenzaufzeichnungen, Sprachnachrichten, Videokonferenzen – ohne Transkription und Klassifikation nicht auswertbar.

TEIL 1 – PROBLEM & KONZEPT

Das eigentliche Problem

> Nicht was ein System findet zählt – sondern was es **nicht** findet, ohne dass man es merkt.

Systeme verarbeiten nicht alle Daten

Viele eDiscovery-Plattformen verarbeiten verarbeiten nur ausgewählte Formate vollständig. Unbekannte oder verschachtelte Container werden stillschweigend übersprungen oder als Fehler protokolliert – ohne Eskalation.

Keine Transparenz über fehlende Daten

Das kritische Problem ist nicht der Fehler selbst, sondern die fehlende Sichtbarkeit. Wenn ein System 95 % der Daten verarbeitet und 5 % lautlos verwirft, enthält enthält der Bericht trotzdem alle Ergebnisse – ohne Hinweis auf die Lücke.

⊗ **Konsequenz:** Unvollständige Daten erzeugen falsche Schlüsse. In Ermittlungen kann das entscheidend sein.

TEIL 1 – PROBLEM & KONZEPT

Ungewolltes Sampling

Wenn Systeme an ihre Grenzen stoßen, passiert etwas Gefährliches: Sie reduzieren – ohne den Nutzer zu informieren. Das Ergebnis ist kein Fehler, sondern eine stille Verfälschung der Datenbasis.

Große Datenmengen werden reduziert

Plattformen mit Größenlimits kürzen den Datenbestand – z. B. nur die neuesten 500.000 E-Mails werden indiziert, der Rest bleibt unsichtbar.

Queries werden vereinfacht

Komplexe Suchanfragen werden automatisch vereinfacht oder abgebrochen, wenn sie zu viel Ressourcen verbrauchen. Das Ergebnis wirkt vollständig – ist es aber nicht.

Systeme stoßen an technische Limits

Timeout-Grenzen, RAM-Limits, Index-Größenbeschränkungen – jedes dieser Limits führt zu implizitem Sampling.

Konsequenz: Beweise fehlen unbemerkt. Kein Fehlermeldung, keine Warnung – nur ein lückenhafter Bericht.

TEIL 1 – PROBLEM & KONZEPT

Technische Realität: Suche unter Last

Ein konkretes Beispiel zeigt, wie Technik das Ergebnis beeinflusst – unabhängig davon, was der Ermittler beabsichtigt.

Das Beispiel: Wildcard-Suche `*ver*`

Eine scheinbar einfache Suchanfrage mit führenden und nachgestellten Wildcards zwingt das System zu einem vollständigen Index-Scan. Jedes einzelne Dokument muss geprüft werden – kein Index kann genutzt werden.

- Full Index Scan auf 100 Millionen Dokumenten
- Millionen potenzieller Treffer ohne Priorisierung
- System-Timeouts nach wenigen Minuten
- Ergebnis: abgebrochene Suche oder leere Rückgabe

Was das bedeutet

Die Suchanfrage war korrekt. Der Ermittler hat nichts falsch gemacht. Aber das System konnte sie nicht ausführen. Das Ergebnis: Null Treffer – obwohl relevante Dokumente vorhanden sind.

Technische Grenzen der Sucharchitektur bestimmen, was gefunden wird. Das ist kein Edge Case – das ist der Alltag bei großen Datenmengen.

Die Qualität der Ergebnisse hängt direkt von der Sucharchitektur ab – nicht nur vom Suchbegriff.

TEIL 1 – PROBLEM & KONZEPT

Semantische Suche – warum das Ergebnis nicht ausreicht

Ausgangsfrage an das System: „Finde Hinweise auf Bestechung"

Gefundene Dokumente (Beispiele)

- „Beratungshonorar für Projekt X"
- „Sonderzahlung für Partner"
- „Beschleunigungsgebühr"
- „Zahlungsabwicklung über Drittanbieter"
- „Provision wurde genehmigt"

Alle Dokumente sind inhaltlich ähnlich – aber bedeuten etwas völlig anderes.

Unkritisch – z. B. Zahlungsabwicklung

Normal – z. B. genehmigte Provision

Potenziell kritisch – z. B. Beschleunigungsgebühr

Was das in der Praxis bedeutet

- Tausende Treffer mit ähnlicher Bewertung
- Keine klare Unterscheidung zwischen relevant und irrelevant
- Hoher manueller Prüfaufwand

Warum ist das kritisch

Das System kann nicht erklären, warum ein Dokument relevant ist – nur: „weil es ähnlich ist." Das ist nicht nachvollziehbar. Und nicht gerichtsfest.

⚠ Semantische Suche findet ähnliche Inhalte – aber keine belastbaren Beweise.

TEIL 1 – PROBLEM & KONZEPT

Was funktioniert: Hybrid + Kontext

Die Lösung liegt nicht in einer einzigen Technologie, sondern in der gezielten Kombination. Jede Methode kompensiert die Schwächen der anderen.

Keyword = Präzision

Exakte Begriffe, juristische Termini, Aktenzeichen, Namen – Keyword-Suche findet, was eindeutig benannt ist. Reproduzierbar und nachvollziehbar.

Vektor = Kontext

Semantische Einbettungen finden relevante Dokumente auch dann, wenn die exakten Begriffe fehlen – z. B. bei umgangssprachlicher Formulierung oder anderssprachigen Dokumenten.

Metadaten = Filter

Zeitraum, Absender, Empfänger, Dokumenttyp – Metadaten-Filter reduzieren den Suchraum präzise, bevor teure semantische Suchen ausgeführt werden.

Erst die Kombination aller drei Methoden erzeugt verwertbare, priorisierte Ergebnisse.

TEIL 1 – PROBLEM & KONZEPT

Ingest ist der kritische Punkt

Der Ingest-Prozess entscheidet, ob ein Dokument überhaupt in die Analyse gelangt. Fehler hier sind unsichtbar – und meist nicht korrigierbar, wenn die Ergebnisse bereits vorliegen.

Die drei häufigsten Ingest-Probleme

- **Verschachtelte Container:** ZIP-in-ZIP, PST mit eingebetteten MSG, E-Mails mit angehängten Archiven – ohne rekursive Extraktion bleiben Ebenen unverarbeitet.
- **Datenexplosion beim Entpacken:** Ein 1 GB PST kann nach Extraktion 50 GB erzeugen. Unkontrollierter Speicherverbrauch bricht Pipelines ab.
- **Metadatenverlust:** Wer hat wann was gesendet? Ohne sorgfältige Metadaten-Erhaltung beim Entpacken gehen diese Informationen verloren – dauerhaft.

Warum das so gefährlich ist

Ein Ingest-Fehler bei 0,1 % der Daten klingt marginal. Bei 100 Millionen Dokumenten sind das 100.000 Dokumente, die nie analysiert wurden – und niemand weiß es.

Fehler im Ingest lassen sich nachträglich kaum rekonstruieren, weil die Rohdaten oft überschrieben oder nicht mehr zugänglich sind.

Kernaussage: Fehler im Ingest zerstören die Analyse – noch bevor sie beginnt.

TEIL 1 – ABSCHLUSS

Outcome: Vom Datenchaos zum Ermittlungsbericht

Das Ziel ist kein Suchergebnis – sondern ein Bericht, der vor Gericht standhält. Jede Aussage darin muss auf ein konkretes Quelldokument zurückführbar sein.

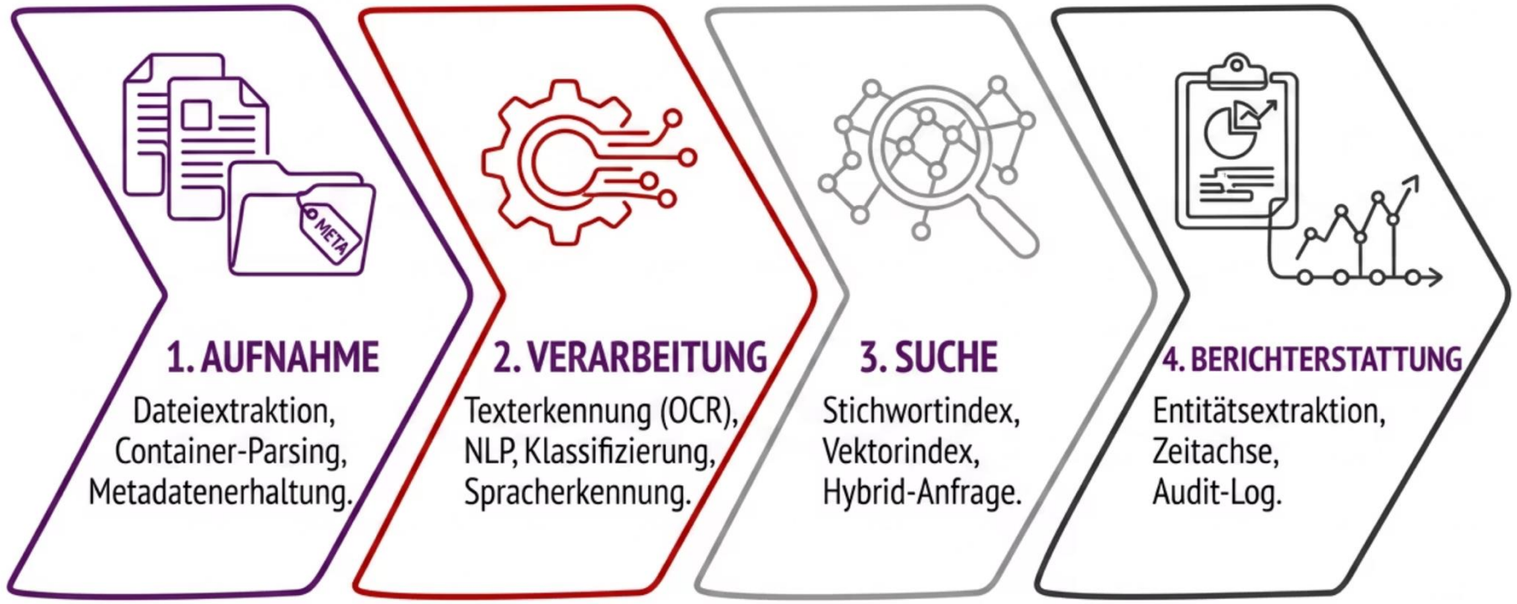
- Ein Bericht über 50 beteiligte Personen und 3.000 relevante Dokumente wird in Minuten generiert – nicht in Wochen.
- Jede Aussage im Bericht verlinkt direkt auf das Quelldokument, inklusive Seitenzahl und Zeitstempel.

Übergang zu Teil 2: Die nächsten Folien zeigen, wie diese Pipeline konkret aufgebaut ist – von der Ingest-Stufe bis zum fertigen Bericht.

TEIL 2 – TECHNISCHE UMSETZUNG

Architekturüberblick: Pipeline statt Tool

Ein einzelnes Tool löst das Problem nicht. Was benötigt wird, ist ein End-to-End-System aus entkoppelten Komponenten, die unabhängig skalieren und austauschbar sind.



Die Entkopplung der Komponenten ist entscheidend: Fällt eine Stufe aus oder muss ausgetauscht werden, bleibt der Rest des Systems betriebsfähig. Jede Stufe schreibt in eine definierte Schnittstelle – kein proprietäres Gesamtsystem.

TEIL 2 – TECHNISCHE UMSETZUNG

Ingest Pipeline (konkret)

Die Ingest-Stufe ist technisch die anspruchsvollste. Hier entscheidet sich, ob alle Daten vollständig und korrekt in die Verarbeitungskette gelangen.

Technische Anforderungen

01

Rekursive Extraktion

Container werden in beliebiger Tiefe entpackt. ZIP in PST in MSG mit angehängtem ZIP – jede Ebene wird vollständig traversiert.

02

Prozess-basierte Verarbeitung

Jeder Dateiformat-Typ wird in einem isolierten Prozess verarbeitet. Absturz oder Fehler in einem Prozess betreffen keine anderen Dateien.

03

Fehlerisolierung

Fehlgeschlagene Extraktion wird explizit protokolliert, nicht stillschweigend verworfen. verworfen. Jede nicht verarbeitete Datei erscheint im Audit-Log.

Plugin-System für neue Formate

Neue Dateiformate werden über ein Plugin-Interface eingebunden, ohne dass die Kern-Pipeline angepasst werden muss. Ein Format-Plugin implementiert eine definierte Schnittstelle: Eingang ist ein Byte-Stream, Ausgang sind strukturierte Metadaten und Rohtext.

Neue Chat-Plattformen, proprietäre Unternehmens-Archive oder zukünftige Formate Formate können so innerhalb von Tagen integriert werden.

TEIL 2 – TECHNISCHE UMSETZUNG

Datenverarbeitung & Anreicherung

Nach dem Ingest sind Dokumente als Rohtext vorhanden. Die Verarbeitungsstufe verwandelt diesen Rohtext in eine strukturierte, durchsuchbare Datenbasis.

OCR

Gescannte PDFs, Fotos von Dokumenten, eingescannte Faxe – OCR macht diese Inhalte durchsuchbar. Qualität und Sprache werden automatisch erkannt.

Entity Extraction

Automatische Erkennung von Personen, Organisationen, Orten, Geldbeträgen, Datumsangaben. Basis für Graphaufbau und Timeline-Analyse.

Klassifikation

Automatische Einordnung: Vertrag, interne Kommunikation, Finanzbeleg, rechtlich relevant. Ermöglicht gezielte Filterung ohne manuelle Sichtung.

Sprachdetektion

Multilinguales Preprocessing: Sprache wird pro Dokument und Abschnitt erkannt. Ermittlungen mit Dokumenten in 10+ Sprachen werden korrekt verarbeitet.

Ergebnis: Aus unstrukturiertem Rohtext wird eine strukturierte, angereicherte Datenbasis – maschinell und human lesbar.

TEIL 2 – TECHNISCHE UMSETZUNG

Metadaten & Graph

Einzelne Dokumente sind isoliert wenig aussagekräftig. Erst wenn Beziehungen zwischen Dokumenten, Personen und Ereignissen sichtbar werden, entsteht Ermittlungsrelevanz.

Kommunikationsbeziehungen

Wer hat mit wem kommuniziert? Wie oft? Über welchen Kanal? Der Graph macht Kommunikationsmuster sichtbar, die in Einzeldokumenten unsichtbar bleiben.

Zeitachsen

Alle Ereignisse werden auf einer gemeinsamen Zeitachse verankert. Auffällige Häufungen, verdächtige Pausen oder koordinierte Aktivitäten werden sofort sichtbar.

Dokumentverknüpfung

Anhänge, Weiterleitungen, Antwort-Ketten, referenzierte Dokumente – der Graph verknüpft zusammengehörige Materialien, auch wenn sie in verschiedenen Quellsystemen liegen.

Der Graph ersetzt die isolierte Dokumentenansicht. Zusammenhänge werden strukturell, nicht nur textuell erfasst.

TEIL 2 – TECHNISCHE UMSETZUNG

Index & Suche: Zwei getrennte Systeme

Keyword-Suche und semantische Suche haben unterschiedliche Anforderungen an Speicher und Abfragelogik. Deshalb werden sie getrennt gehalten – und erst bei der Abfrage kombiniert.

Keyword Index (Elasticsearch)

Invertierter Index auf Volltextbasis. Optimiert für exakte Begriffe, Boole'sche Logik, Wildcard-Suche mit Einschränkungen und Fuzzy-Matching. Jede Abfrage ist deterministisch und reproduzierbar.

- Skaliert auf Milliarden von Tokens
- Sharding und Replikation out-of-the-box
- Vollständige Query-Dokumentation möglich

Vector Index (Embeddings)

Hochdimensionale Vektoren (typisch 768–1536 Dimensionen) werden in einem spezialisierten Vektorspeicher (z. B. Qdrant, Weaviate, pgvector) abgelegt. Approximate Nearest Neighbor (ANN) Suche liefert semantisch ähnliche Dokumente in Millisekunden.

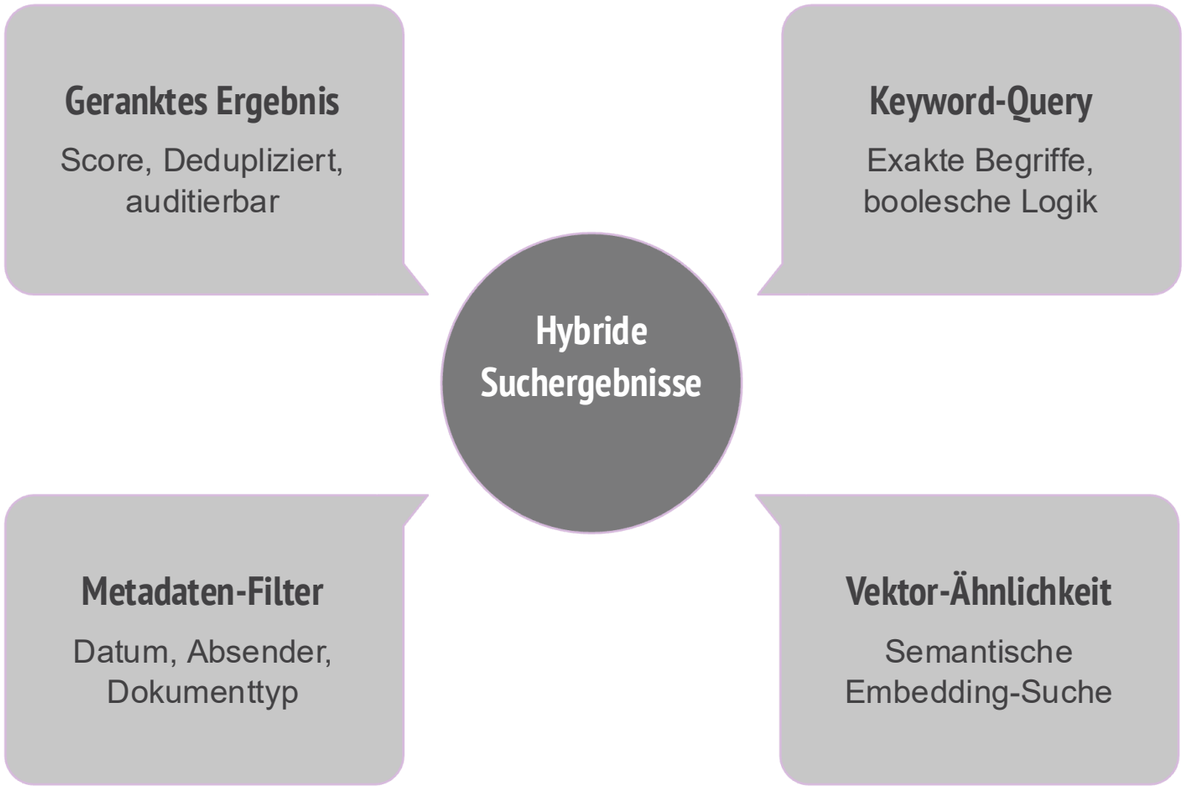
- Getrennte Speicherung vom Keyword-Index
- Embedding-Modell ist austauschbar
- Chunk-basierte Indizierung für lange Dokumente

TEIL 2 – TECHNISCHE UMSETZUNG

Hybride Suche (technisch)

Eine hybride Suchanfrage besteht aus mehreren parallelen Teilabfragen, deren Ergebnisse gewichtet zusammengeführt werden. Das erlaubt Präzision und Kontext gleichzeitig.

Die Score-Fusion (Reciprocal Rank Fusion oder gewichtete Summe) kombiniert die Treffer beider Indizes zu einem gemeinsamen Ranking. Metadaten-Filter werden als Pre- oder Post-Filter angewendet, um den Suchraum einzuschränken, bevor teure Vektorberechnungen stattfinden. Jede Abfrage wird mit allen Parametern im Audit-Log gespeichert.



TEIL 2 – TECHNISCHE UMSETZUNG

Skalierung der Plattform

100 Millionen Dokumente sind keine Ausnahme mehr – sie sind die Realität in großen Unternehmensermittlungen. Die Architektur muss von Anfang an für für diese Größe ausgelegt sein.

100M+

Dokumente

Verarbeitbar in einer einzigen Plattforminstanz durch horizontale Skalierung

<1w

Ingest-Zeit

Für typische Unternehmensermittlungen mit parallelen Worker-Clustern

N+1

Redundanz

Jede Komponente läuft redundant – kein Single Point of Failure im kritischen Pfad

Horizontale Skalierung

Worker-Nodes für Ingest und Verarbeitung werden bei Bedarf hinzugefügt. Kein Umbau der Architektur nötig – nur mehr Ressourcen.

Asynchrone Pipelines

Entkopplung über Message Queues (NiFi, RabbitMQ). Jede Stufe verarbeitet in verarbeitet in eigenem Tempo, Rückstau wird absorbiert statt geblockt.

TEIL 2 – TECHNISCHE UMSETZUNG

AI Report Generator: Mechanik ohne Halluzinationen

Der Report Generator ist keine generative KI im klassischen Sinne – er ist eine strukturierte Aggregations- und Zusammenfassungsmaschine, die die ausschließlich auf verifizierten Daten aus dem System operiert.

Input: Was geht rein

- Strukturierte Entities (Personen, Organisationen, Orte)
- Events mit Zeitstempeln und Quelldokumenten
- Kommunikationsbeziehungen aus dem Graph
- Klassifizierte und priorisierte Dokumente

Pipeline: Was passiert intern

- **Aggregation:** Zusammenführung aller Events pro Person / Zeitraum
- **Bewertung:** Gewichtung nach Relevanz-Score und Häufigkeit
- **Zusammenfassung:** LLM-Prompting mit striktem Grounding auf Quelldaten
- **Verifikation:** Jede Aussage wird mit Quelldokument-Referenz belegt

⚠ **Wichtig:** Das Modell darf keine Annahmen treffen, die nicht durch vorhandene Daten gedeckt sind. Jede Aussage im Bericht enthält eine Quellenangabe – ohne Quelle keine Ausgabe.

TEIL 2 – TECHNISCHE UMSETZUNG

Vom Chat zur automatisierten Analyse

Ein konkretes Beispiel zeigt, wie eine einzige Frage im Chat eine vollständige, automatisierte Ermittlungsanalyse auslöst.

1 Anfrage im Chat

Der Mitarbeiter stellt eine Frage direkt per WhatsApp: „Zeige mir alle relevanten Aktivitäten zur Stadt X."

2 Anfrage geht an Agenten-System

Das System interpretiert die Anfrage und startet automatisch eine strukturierte Analyse.

3 Datenquellen werden durchsucht

Bilder: Erkennung von Orten und Landmarken. E-Mails: Suche nach gleicher Stadt und Kontext. Weitere Dokumente und Chats werden einbezogen.

4 Zusammenführung der Ergebnisse

Treffer werden über Ort, Zeit und beteiligte Personen verknüpft. Relevante Zusammenhänge werden identifiziert.

5 Ergebnis für den Nutzer

Strukturierter Bericht wird erstellt – direkt aus aus dem Chat heraus per E-Mail versendbar, z. B. an Kollegen oder das Ermittlerteam.

Der Nutzer stellt nur eine Frage – das System übernimmt die vollständige Analyse und Aufbereitung.

FAZIT

Was zählt – und was nicht

Nicht jede technische Lösung ist für Ermittlungsszenarien geeignet. Diese Prinzipien haben sich in der Praxis als entscheidend erwiesen.

Vollständigkeit vor Geschwindigkeit

Ein schneller, aber lückenhafter Bericht ist gefährlicher als ein langsamer, vollständiger. Unbemerkt fehlende Daten sind das größte größte Risiko.

Hybrid Search als Standard

Weder reine Keyword-Suche noch reine Vektorsuche reicht aus. Nur die Kombination mit Metadaten-Filtern liefert verwertbare, priorisierte Ergebnisse.

Skalierbare Architektur ist Pflicht

Systeme, die bei 10 Millionen Dokumenten funktionieren, scheitern oft bei 100 Millionen. Skalierbarkeit muss von Anfang an eingeplant sein.

Fokus verschiebt sich: von Suche zu Zusammenhängen

Das Ziel ist nicht mehr das Finden einzelner Dokumente – sondern das automatische Erkennen von Mustern, Verbindungen und Ereignissen.

Kontakt & Demo: Interesse an einer Live-Demo oder weiteren technischen Details? Sprechen Sie uns direkt an oder besuchen Sie unseren unseren Stand.

Danke für Ihre Aufmerksamkeit

Fragen & Diskussion

Wir freuen uns auf Ihre Fragen – und auf das Gespräch.

Live-Demo

Interesse an einer Live-Demo? Sprechen Sie uns direkt an oder besuchen Sie unseren Stand.

Kontakt

Gerne stehen wir für technische Details, Rückfragen und weiterführende Gespräche zur Verfügung.