LUCID TRUTH TECHNOLOGIES®

Dealing with the Deluge of Digital Discovery

Leveraging AI to Tackle Modern Legal Challenges

Kenneth G. Hartman
February 6, 2025

Kenneth G. Hartman

CISSP, GSE, GASF, GCFA, GCFE, GCIA, GWAPT, GCPM

Owner and Principal Consultant - Lucid Truth Technologies

Certified SANS Instructor

Mission: *To make the truth clear*

Motto: *I talk to cool people about cool things in cool parts of the world!*

Ken Hartman brings a wealth of expertise in digital forensics and cybersecurity, offering invaluable insights into managing the complexities of modern legal discovery through AI and advanced technologies.

LucidTruthTechnologies.com

The more we solve our problems, the more we widen the definition of "problem" so that our number of problems remains constant.

Don't expect a life without problems.

Progress doesn't mean reducing your quantity of struggles but increasing their quality.

The goal of life is to trade bad problems for better ones.

bit.ly/blue-dot-effect

Objectives

- Discuss best practices for handling legal discovery, including inventory management and version control.
- Highlight AI-driven approaches for identifying and resolving document discrepancies.
- Demonstrate how AI Tools like the Discovery Processing Web App can transform raw data into actionable insights.
- Understand the technical challenges and limitations of such tools.

The Importance of Discovery Management

- Legal discovery often involves thousands of pages of documents, files, and digital records.
- **Challenges:**
  - Unorganized files
  - Missing documents
  - Duplicates and multiple versions
  - Poor-quality scans affecting OCR (Optical Character Recognition)
- **Solution:** Implement inventory tracking and version control systems

Current eDiscovery Software is designed to support discovery production

Properly Handling Criminal Discovery
https://bit.ly/discovery-handling

Creating a Discovery Inventory

Step 1: Log Everything Upon Receipt

- Create an inventory list of all files received
- Categorize by type: reports, emails, photos, transcripts, etc.
- Log the date the discovery transmittal was received
- Record the document metadata (timestamps and authorship, etc.)

Step 2: Verify Completeness

- Compare documents against discovery lists.
- Request missing files immediately.
- Watch for referenced but missing evidence.
- Send a confirmation listing exactly what was received.

Creating a Discovery Inventory

Step 3: Identify Similar Documents

- Different transmittal batches may have the same or similar documents
- Track who sent the document and when it was received as a “Version”
- Compare the wording of similar document versions

Why It Matters

- Changes in police reports, affidavits, and transcripts **can alter legal arguments**.
- **Example:** A single word removed in a later version may impact the case outcome.

Why can’t these steps be automated?

Case Number: 2025-012
Date: 2025-02-03
Reporting Officer: Officer A. Gomez
Location: 789 Oak Avenue, Riverside

Summary:
At approximately 21:30 on February 3, 2025, a break-in was reported at 789 Oak Avenue. The primary suspect is described as a 29-year-old male wearing a black leather jacket and dark jeans. A nearby resident, Ms. Karen Fields, observed a second male individual wearing a white tank top and acting suspiciously near the scene shortly before the incident, noting that the person appeared to be carrying several small items. Evidence collected includes surveillance footage and a set of fingerprints lifted from a shattered window. The investigation is continuing.

Best Practices

- Store **original versions** separately from working copies.
- Use **hash values** to verify integrity (cryptographic fingerprinting).
- Compare versions using **redlining tools** to highlight differences.
- Ensure **chain of custody** documentation is intact
- Use forensic techniques to **preserve metadata** (timestamps, EXIF data, document history)

BIG IDEA: Why not treat discovery documents with the same care as a forensic image of a computer hard drive?

The Perils of Improper Evidence Collection
https://bit.ly/evidence-collection

AI-Powered Discovery Management

Leveraging AI for Inventory & Document Control 💡

- AI-driven tools categorize discovery files automatically.
- LLM models can summarize key documents and highlight changes
- AI-generated timelines track events referenced in documents

Case Study: Discovery Analysis Web App

- Indexes and summarizes Case Documents
- Generates a detailed timeline of all events
- Clusters and describes entities (people, places, evidence, etc.)

Scanned Documents – OCR

- Many discovery documents are poor-quality scans
- Optical Character Recognition (OCR) identifies text in images
  - Adobe Acrobat – Does a good job of OCR
  - tesseract – An opensource command line program
  - pytesseract – A Python module for tesseract
- OCR requires a “readable image” and characters with clean, sharp edges
- AI may be able to clean up OCR text when some characters are incorrectly recognized.

➔ Insist on discovery that is readable!

Case Number: 2025-012

Date: 2025-02-03

Reporting Officer: Officer A. Gomez

Location: 789 Oak Avenue, Riverside

Summary:

At approximately 21:30 on February 3, 2025, a break-in was reported at 789 Oak Avenue. The primary suspect is described as a 29-year-old male wearing a black leather jacket and dark jeans. A nearby resident, Ms. Karen Fields, observed a second male individual wearing a white tank top and acting suspiciously near the scene shortly before the incident, noting that the person appeared to be carrying several small items. Evidence collected includes surveillance footage and a set of fingerprints lifted from a shattered window. The investigation is continuing.

Photocopied Text Messages

⚠️ The Problem

- **Easily Fabricated**: Free online tools allow users to create fake text conversations.
- **Photocopied Screenshots Lack Integrity**: No metadata, timestamps, or verifiable source.
- **Danger in Legal Cases**: Fake texts can lead to wrongful convictions or misleading defenses.

Don't Be Fooled by Fabricated Evidence
Demand the source file from the phone or the records from the cellular carrier.

Photocopied Text Messages
https://bit.ly/photocopied-sms

Email Headers

⚠️ The Risk of Fake Emails

- Printouts of emails are easily **manipulated or forged**.
- The **"From" field** can be **spoofed**, misleading courts and attorneys.
- Without **email headers**, authenticity is **impossible to verify**.

🔍 What Do Email Headers Reveal?

- 📌 **Sender & Recipient:** True email addresses & routing details.
- 📌 **Timestamps:** When the email was sent, received, & processed.
- 📌 **IP Addresses:** Origin & path the email traveled.
- 📌 **Technical Details:** Email client, security checks, and possible tampering.

Don't let your case rely on email printouts—demand the headers!

Why Email Headers are Critical in Court
https://bit.ly/email-evidence

Audio Evidence

Growing Challenge: Body-worn camera footage, interview recordings, and phone call evidence can span **hundreds of hours**.

- **Search Warrants:** Officers executing warrants capture hours of footage.
- **Witness Interviews:** Some cases involve **dozens of witness interviews**, each with critical statements.

Key Issue: How do you efficiently extract, analyze, and track relevant statements?

AI-Powered Transcription

Automated AI Transcription

- Transcribes audio quickly, **reducing manual review time**.
- Converts speech to text with **timestamped utterances**.

Speaker Diarization (Who Said What?)

- Detects different speakers in conversations.
- Labels and organizes **statements from different speakers**.
- Helps investigators focus on **high-value portions** of audio.

⚠️ **Caution:** AI transcription is **not perfect**—overlapping speech can cause errors.

Best practice: Use **transcripts for navigation** but verify with original recordings.

Enhancing Discovery with AI Transcription

- **Batch Processing:** AI can transcribe **hundreds of hours overnight**, allowing investigators to rapidly scan for key evidence.
- **Indexing & Searchability:** Transcripts can be **searched, color-coded, and categorized** for quicker case analysis.
- **Sentiment Analysis:** Witness statements can be analyzed for positive and adverse statements regarding the defendant.
- **Future Feature:** Hyperlinks to specific **audio timestamps** for quick playback.

Who said What? Coping with hundreds of hours of audio evidence
https://bit.ly/audio-evidence

LLM Halucination

What is AI Hallucination?

- **AI hallucination** occurs when an AI system **generates false, misleading, or non-existent information** that appears plausible but is factually incorrect.
- This can happen in **document analysis, transcription, and legal research**, leading to **errors in case interpretation**.

Why Does AI Hallucinate?

- **Pattern Recognition Without Understanding** – AI models **predict** words based on patterns but do not “know” truth.
- **Lack of Source Verification** – AI **fills in gaps** when data is incomplete or ambiguous.
- **Training Data Bias** – If the AI **learns from flawed or incomplete data**, it may produce inaccurate outputs.
- **Overconfidence in Predictions** – AI **does not recognize uncertainty**, so it may present **wrong information as fact**.

How to Mitigate AI Hallucination

✔ **Always Link to Source Documents** – Ensure AI-generated insights include references to **original evidence**.

✔ **Use Retrieval-Augmented Generation (RAG)** – AI should **retrieve data from verified sources**, rather than generating new information.

✔ **Human Review is Essential** – Attorneys and investigators must **validate AI outputs** before relying on them.

✔ **Set AI Confidence Thresholds** – AI should **flag uncertain responses**, rather than fabricating answers.

📌 **Key Takeaway:** AI is a powerful tool for legal discovery, but its outputs must always be verified against source documents to prevent misinformation.

What Are Deepfakes?

- **Deepfake technology** uses AI and machine learning to create **highly realistic yet fake** images, videos, and audio recordings.

- Can manipulate **facial expressions, voice, and body language** to create **false evidence**.

- Increasingly accessible through **open-source tools**, making deception easier than everdeepfakes.

Legal Challenges Posed by Deepfakes

⚖️ **Fabricated Evidence** – Fake videos or audio can **falsely incriminate or exonerate** suspects.

⚖️ **Authentication Issues** – Courts struggle to **differentiate between real and AI-generated content**.

⚖️ **Liar's Dividend** – Defendants may **dismiss real evidence as fake** to create doubtdeepfakes.

⚖️ **Erosion of Trust** – Affects **jury perception**, **law enforcement credibility**, and **case integrity**.

Defending Against Deepfakes

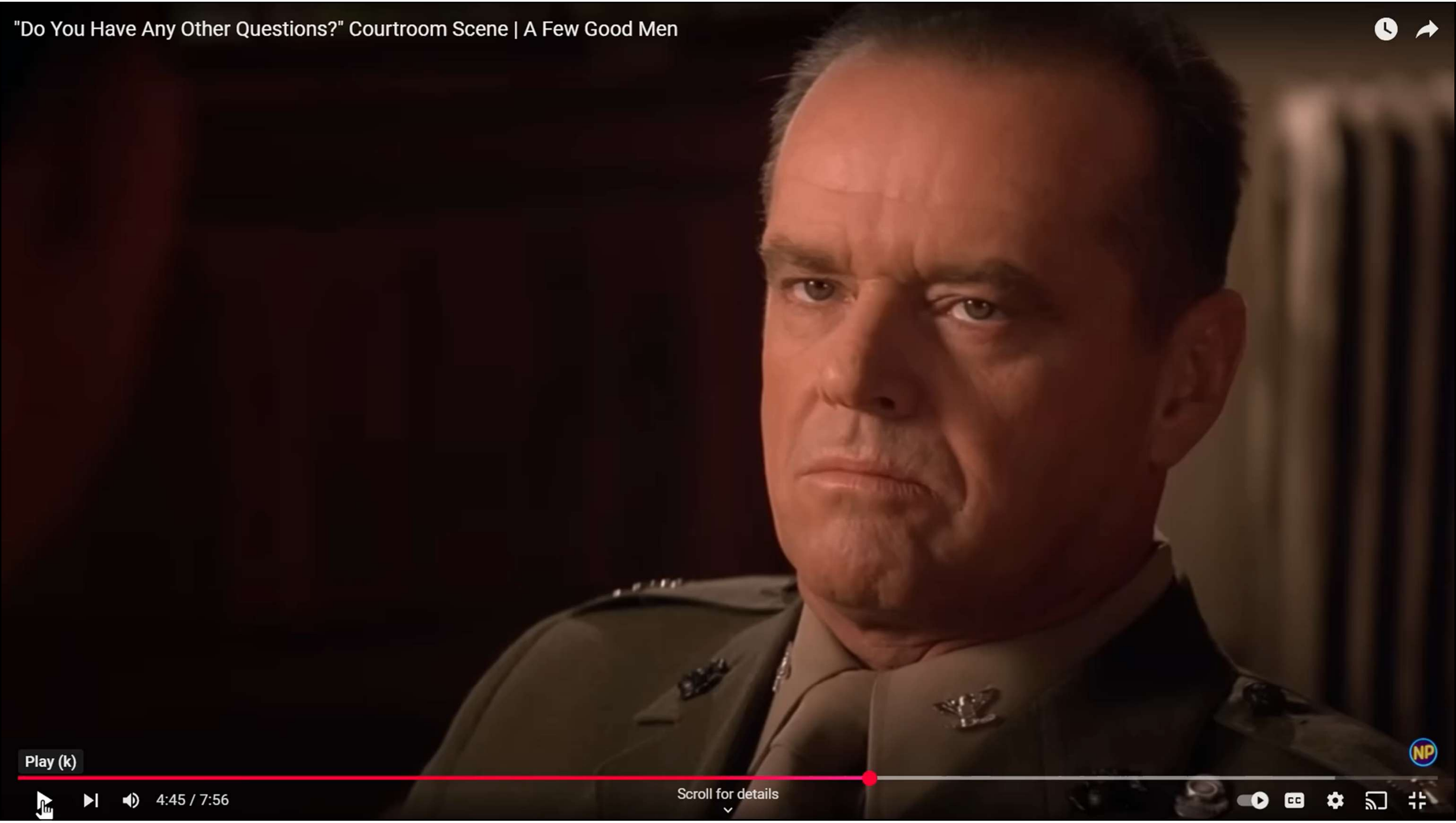
✔ **AI Detection Tools** – Use deepfake detection algorithms to analyze **inconsistencies** in media.

✔ **Metadata Analysis** – Review **timestamps, compression artifacts, and inconsistencies** in file metadata.

✔ **Expert Testimony** – Forensic analysts can **evaluate digital evidence** for signs of manipulation.

✔ **Chain of Custody Verification** – Ensure that **digital evidence has a traceable and untampered history**.

Can AI Detect Deception?


"Do You Have Any Other Questions?" Courtroom Scene | A Few Good Men
Play (k)
4:45 / 7:56
Scroll for details

Discovery Analysis Web App

- Designed to help analyze **large volumes of discovery** efficiently.
- Uses **Artificial Intelligence (AI) and Natural Language Processing (NLP)** to organize and access legal discovery.
- Focused on **organizing, tracking, and analyzing documents**, reducing manual effort.

- A **Minimum Viable Product (MVP)** at this stage
- Even needs a better name ☺

Key Features of the Web App

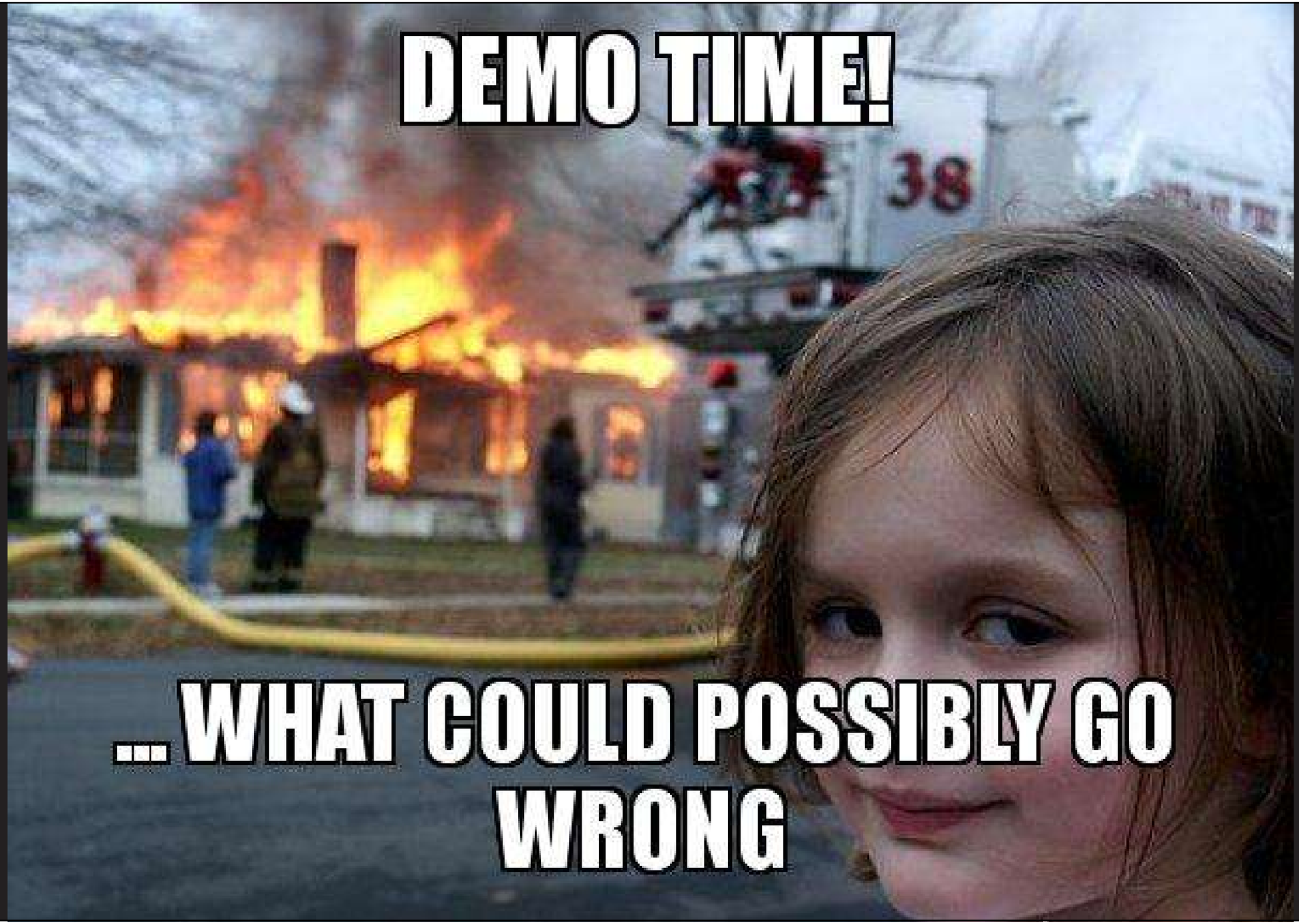
✔ **Case Overview Page** – Provides a **high-level summary** of the case, refreshing user memory.

✔ **Case Documents** – Categorizes documents and allows for **easy access and searchability**.

✔ **Detailed Case Timeline** – Lists **timestamped events**, normalizing **time zones** for consistency.

✔ **Entity Identification** – Groups **people, places, and key elements** across documents.

✔ **AI-Powered Search** – Helps detect **discrepancies, missing evidence, and inconsistencies**.


DEMO TIME!
...WHAT COULD POSSIBLY GO WRONG

MVP Concept & Development Focus

- **Focused on Core Functionalities** – The app is built to address the **most pressing needs first**.
- **Iterative Development** – New features will be added **based on user feedback and real-world testing**.
- **Accuracy** – AI-generated insights are **linked back to source documents** to prevent hallucinations.
- **Scalability** – Future enhancements will improve **document clustering, dossier generation, and interactive case review**.
- **Security** – Currently uses a Single-Tenant VPN Connection

Roadmap & Enhancements

- **Enhanced Document Search**
- **Dossier Generation** – Automated summaries of **key figures in the case**.
- **Interactive Chat with Case Data** – **Retrieval-Augmented Generation (RAG)** for grounded AI-powered insights.
- **Enhanced Document Navigation** – **Hyperlinked references** between documents.
- **Advanced Filtering** – Improved **sorting and categorization** of discovery files.

Conclusion

- Artificial Intelligence is changing the Legal arena. Stay informed.
- The **Discovery Analysis Web App** is an **evolving solution** to streamline legal discovery.
- As an MVP, it will **continue to improve** based on real-world usage.
- Your **feedback is crucial** – help shape the next generation of AI-powered discovery tools.

📩 **Contact Lucid Truth Technologies** if we can help your case in any way: ken@lucid-truth.com or 608-250-0015 cell phone.

Stay in Touch!

Kenneth G. Hartman
231-492-8049 (office)
608-250-0015 (mobile)
Ken@Lucid-Truth.com

LucidTruthTechnologies.com

X: @kennethghartman
Linkedin.com/in/kennethghartman/

Download this presentation

bit.ly/discovery-deluge

Thank You

LUCID TRUTH
TECHNOLOGIES