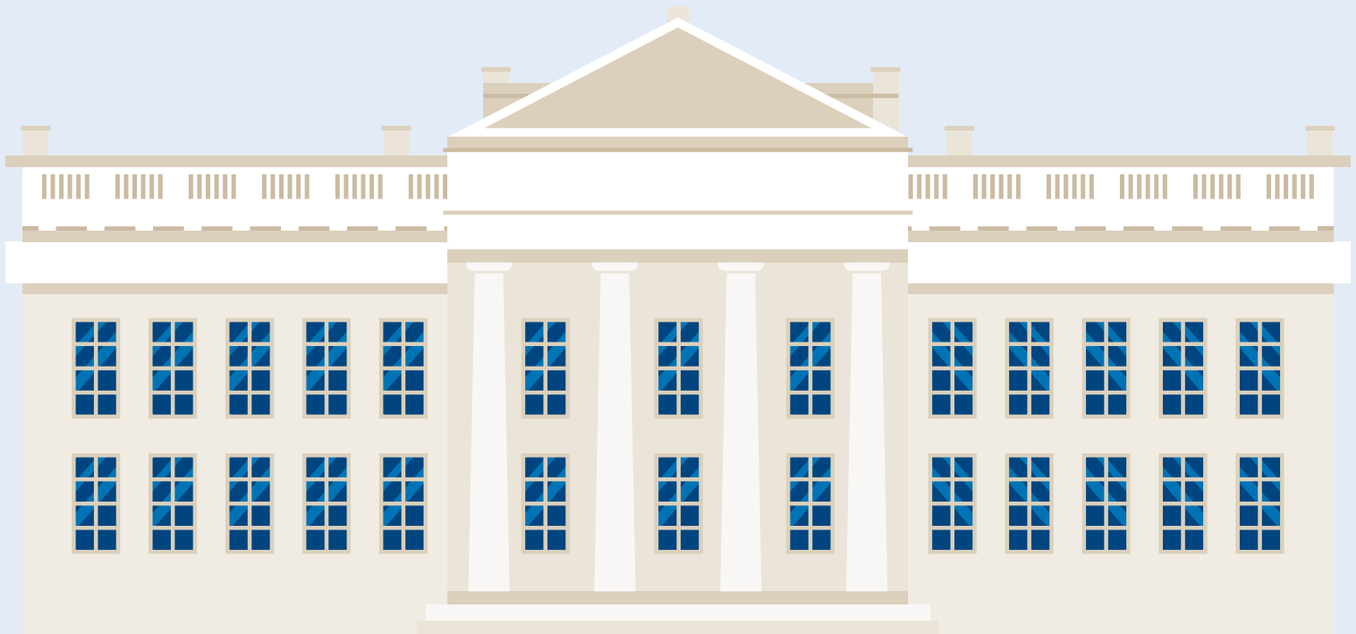
FocalCXM
REMARKABLE EXPERIENCES

Case Study



databricks

DOCUMENT INTEGRITY & AI FORENSICS

PDF Forensics Pipeline to Identify Risk of Fraud in Over 10 Million Documents with Databricks

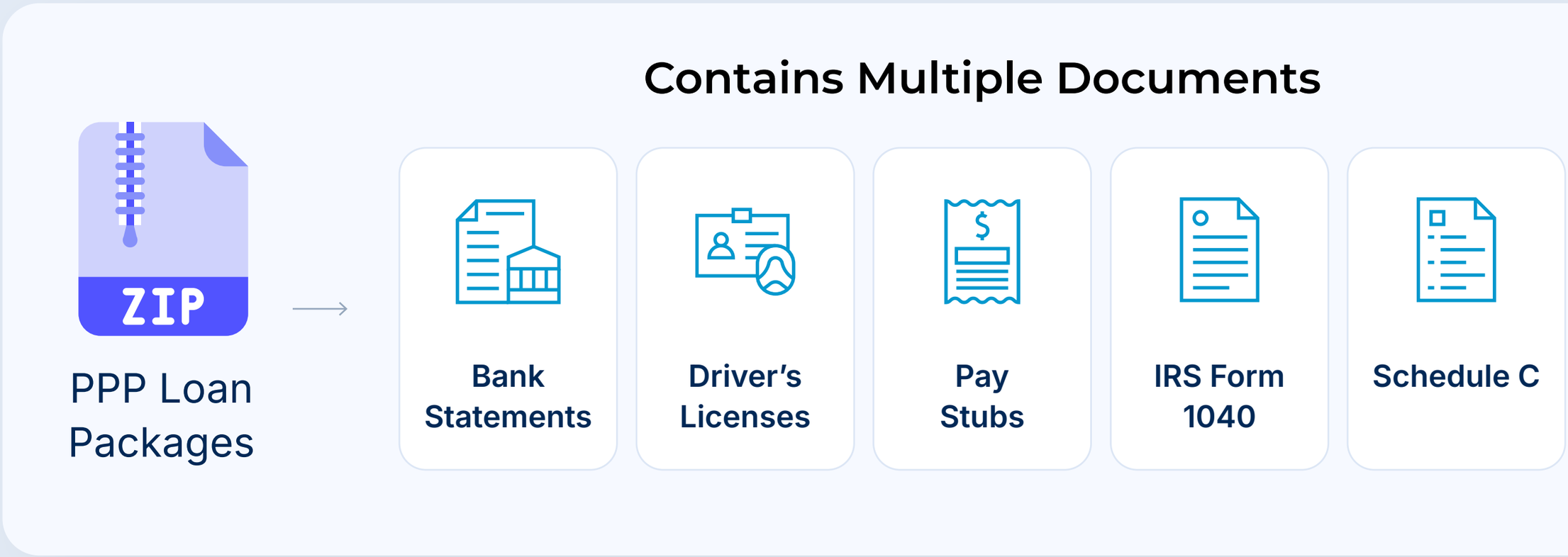
OCR, document forensics, and explainable risk scoring help reviewers identify potentially tampered and AI-generated documents.

FocalCXM
REMARKABLE EXPERIENCES

The Challenge

FocalCXM built a PDF forensics pipeline to process PPP loan packages submitted as ZIP files, with each package containing multiple documents such as **bank statements, driver's licenses, pay stubs, and IRS tax returns including Form 1040 and Schedule C.**

The scale and variety of these documents create challenges for manual review. **Potentially tampered files and AI-generated content** can be difficult to identify through casual inspection, while scanned and flattened documents can create blind spots for automated forensic checks.

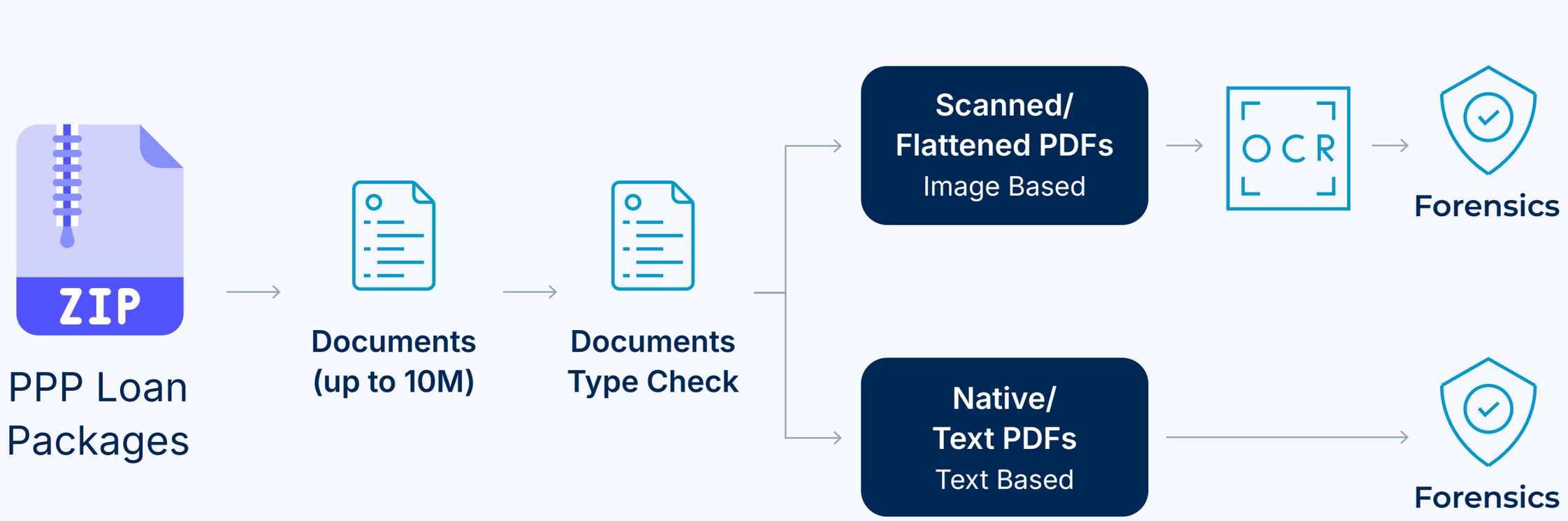


FocalCXM
REMARKABLE EXPERIENCES

The Solution

FocalCXM built a PDF forensics pipeline that scans ZIP packages containing up to 10 million documents, scoring them for tampering and AI-generation risk before they reach a decision-maker.

Only scanned and flattened documents go through OCR extraction, converting image-based PDFs into searchable, analyzable text. **Native,** text-based documents proceed directly to the forensic analysis without unnecessary OCR processing.



FocalCXM
REMARKABLE EXPERIENCES

Extensible Across Document Types

The pipeline can be extended beyond PPP loan packages to process multiple document types, including:

Legal Documents

Real Estate Contracts

Lease Agreements

Vendor Contracts

Deeds & Titles

Insurance Documents

FocalCXM
REMARKABLE EXPERIENCES

PDF forensics pipeline checks metadata-level signals, including:

- **Standard, XMP & XREF metadata**
  editing history, application traces, and authorship clues
- **EOF markers**
  structural signs of post-creation edits
- **Creator/Producer Mismatches**
  inconsistencies between declared and actual authoring tools
- **Document ID Discrepancies**
  signals that may suggest reassembly or splicing
- **Font metadata**
  irregularities that may indicate font substitution or injection
- **Layered vs. flattened content**
  differences in document structure and composition
- **Revision timing**
  edit patterns that may be inconsistent with the claimed document history

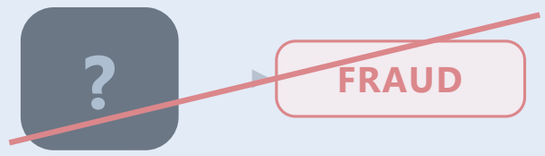
FocalCXM
REMARKABLE EXPERIENCES

The Design Principle That Matters

The pipeline reports indicators for a human reviewer, not automated fraud verdicts.

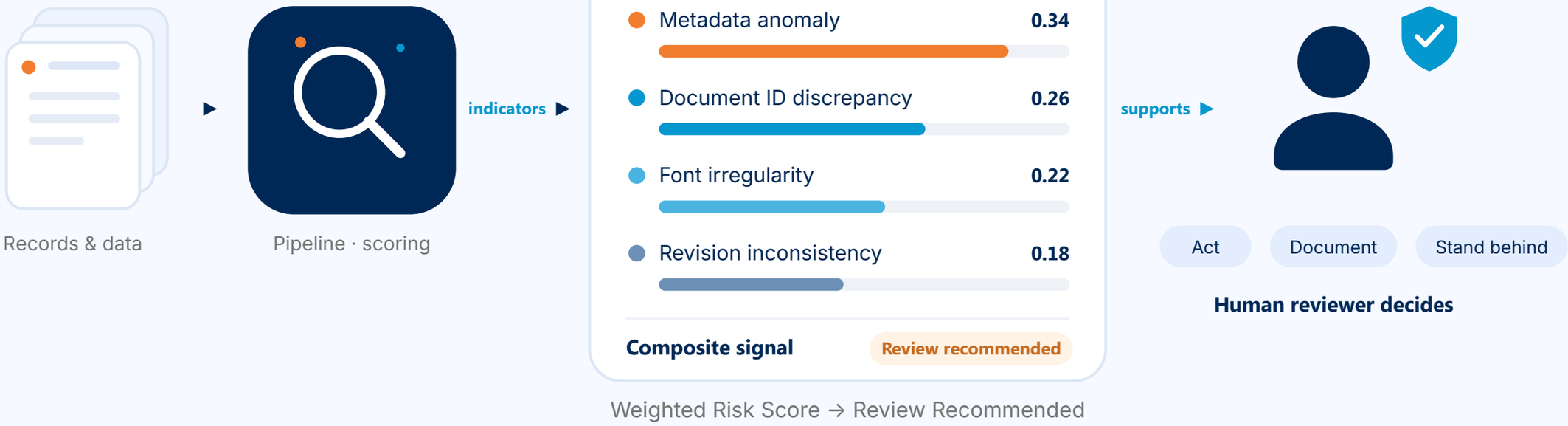
In compliance driven programs such as federal contracting, loan and grant review, and audit programs, this distinction provides a defensible, explainable signal for reviewers to act on rather than a black-box accusation.

SIGNAL, NOT VERDICT



Not this — a black-box verdict.

Hard to defend, hard to appeal, easy to challenge.



Weighted Risk Score → Review Recommended

FocalCXM

REMARKABLE EXPERIENCES

Trust every document.
Verify every decision.
Scan to see how.

contact@focalcxm.com

www.focalcxm.com