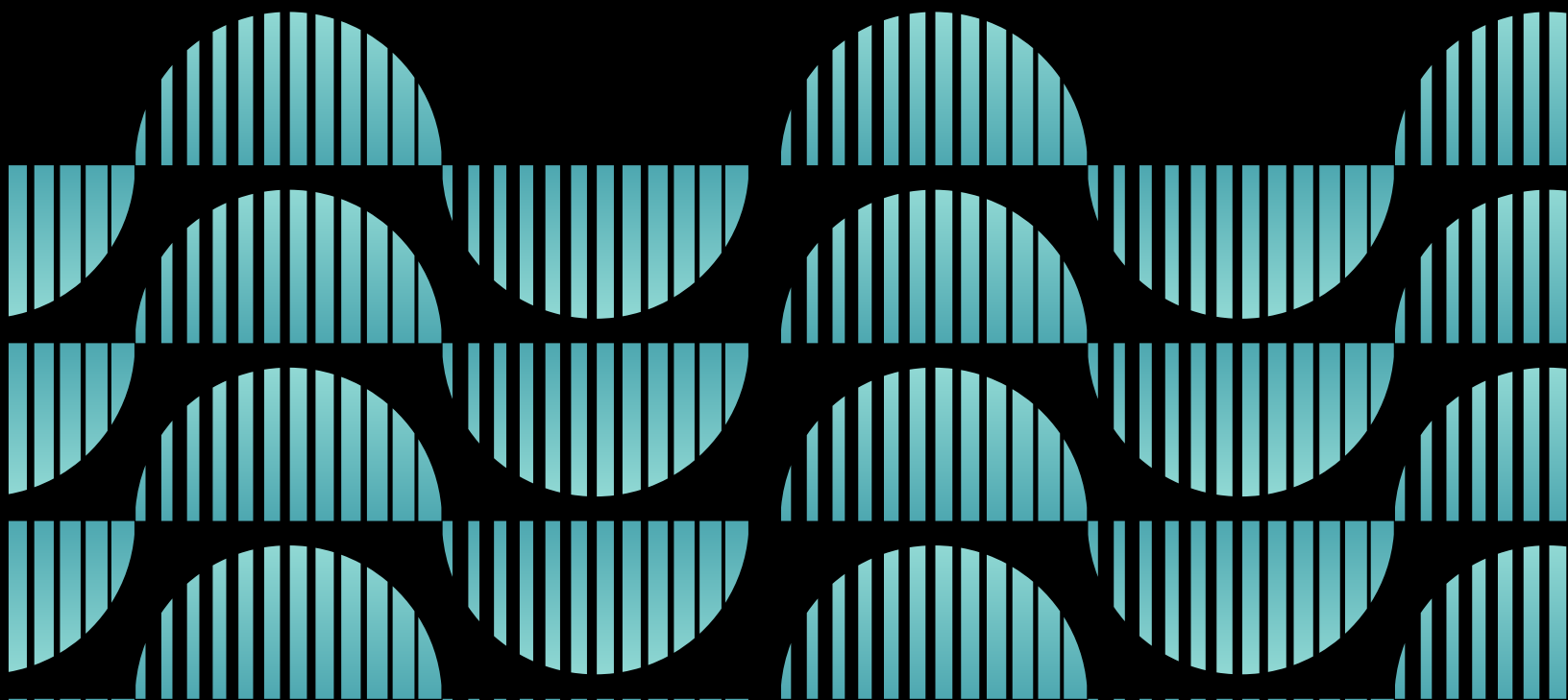
RIPCORD

AI Has a Data Problem, and It's Bigger Than You Think

The growing data deficit undermining enterprise and sovereign AI, and the untapped goldmine of information locked in documents.



THE CURRENT REALITY

Enterprise and Sovereign AI are Starved for Data

Enterprises and governments everywhere are investing billions in artificial intelligence. Yet most overlook what is arguably their most valuable data asset – **the one that can ultimately determine AI success or failure:** decades of data and institutional knowledge trapped within vast archives of unstructured documents.

This is a fundamental challenge. AI without high-context data is like a skilled employee with no institutional knowledge or training – capable in theory, but operating blind in practice. The most valuable AI outcomes depend not just on powerful general models (ChatGPT, Claude, Gemini, etc.), but on high-context data that's usually unique and proprietary to a government agency or an enterprise. And in most cases, this high-context data is trapped in unstructured documents, much of it still on paper. The numbers tell the story:

THE REALITY

~90%

of enterprise data is unstructured

THE GAP

<1%

of that unstructured data is used in AI today

THE SCALE

250 T

paper documents contain most of that data

In other words, the data most critical to AI success is the least accessible to AI systems.

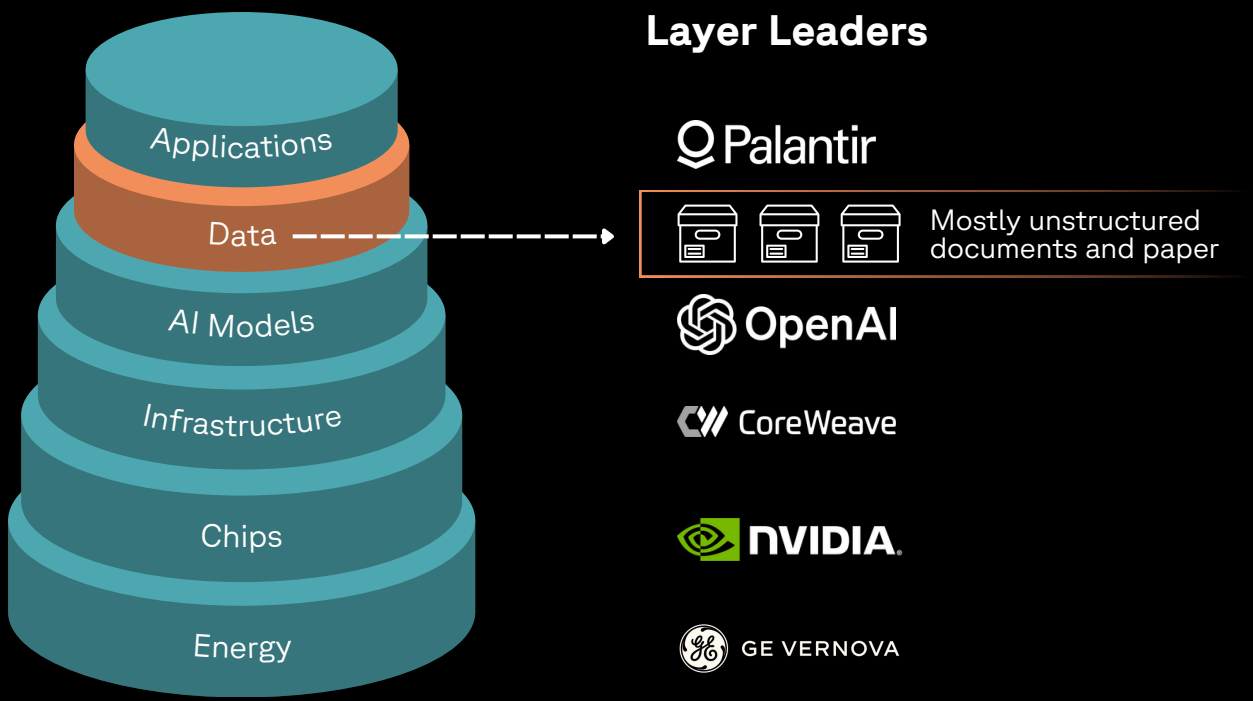
The Six-Layer AI Cake: High-Context Data is an Essential Layer

Building effective AI requires investment across an entire stack of interdependent layers. The market has responded, with trillions flowing into five of those six essential layers, as outlined in Nvidia's "five-layer cake": Energy, chips, infrastructure, AI models, and applications.

This framing reflects a fundamental shift: AI is no longer just software. As it scales and permeates every aspect of our lives, it has necessarily become a vast ecosystem, an essential utility that can operate autonomously and produce intelligence in real time. While five of these layers are advancing rapidly, one layer seems to be lagging: Data. Specifically, specialized high-context data to fuel enterprise and sovereign AI.

Data is the Connective Tissue Between AI Models and Applications

Data – the critical layer that sits between AI models and the applications that run on them – remains underdeveloped. This is an essential layer for specialized AI, including enterprise AI and sovereign AI, which depend on proprietary, high-context data that's unique and exclusively available to those organizations or nations.



The reason this data layer has not yet been built or made available is not due to a lack of awareness. Stakeholders understand the value of this data. The challenge is that activating it – particularly from unstructured documents – has historically been too complex and too costly to do, and virtually impossible to achieve at scale.

And yet, high-context data is the layer that ultimately determines whether AI performs effectively in your specific environment for your specific needs. A powerful model running on generic information produces generic results. The same model, fueled by vast amounts of proprietary data, historical decisions, and institutional knowledge, delivers something fundamentally different: reliable, high-confidence outputs tailored to your domain.

In practical terms, the AI stack is incomplete without the high-context data layer, where data represents proprietary information that's fully accessible and operationalized. The "five-layer cake" is actually six layers. Bon appétit!

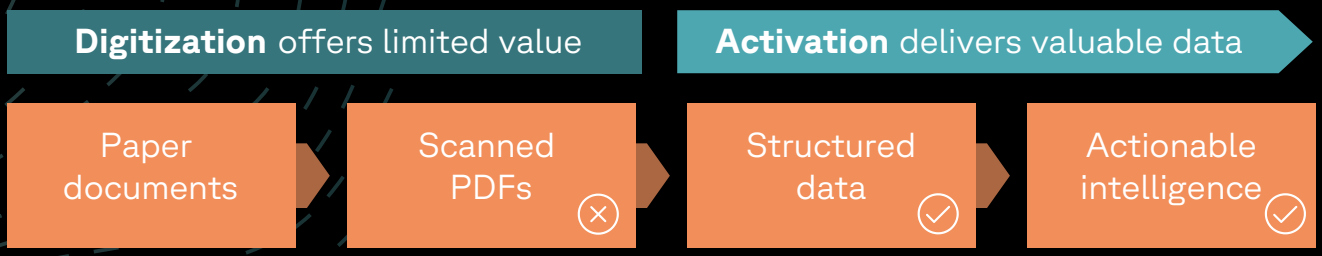
Energy → Chips → Infrastructure → AI Models → **Data** → Applications

While the market continues to invest heavily across the broader AI stack, the data layer remains a critical opportunity. This is especially true in enterprise and sovereign AI use cases, where high-context data is not optional but essential. **The data most critical to AI success is often the least accessible. It already exists today, but has yet to be fully activated.**

THE CORE PROBLEM

AI Redefines Document Activation

For decades, organizations treated document digitization as a storage decision. In most cases, the choice did not favor digitization. The cost and complexity of digitizing documents at scale far outweighed the return on investment, resulting in negative ROI. That reality may have held then, but it doesn't work now. Today, data is a primary bottleneck to progress in the most consequential technology evolution of our time: AI.



A scanned PDF is operationally equivalent to a paper document. You cannot easily query fields, automate workflows, analyze trends, train models, or integrate systems. The result: digitization without productivity gain, weak ROI, and projects that stall. To capture the real data value of documents, the process must include accurate indexing and conversion of the underlying content into structured data and the inference of intelligence from that structured data (classification, indexing, association, ontology). It's what we call activation.

The Document Deadlock: Complexity > Value

For 30+ years, the complexity and cost to extract data from documents at scale exceeded the value of the data extracted, and paper archives accumulated faster than digitization by trillions of pages. So organizations digitized only active records and compliance minimums. That amounts to less than 5% of documents. Everything else stayed in boxes, stored away in warehouses, basements, and offices, taking up valuable space and providing zero value beyond retention policies and regulatory compliance.

The barriers to at-scale document activation were structural, not motivational. Traditional digitization required extensive manual work at every stage:

- Manual removal of non-document parts (staples, clips, folders, envelopes, etc.)
- Manual detection and unfolding of folded or non-standard sheets
- Page-by-page scanning with manual feed and removal
- Human reading, metadata entry, and field validation for every document

Traditional "Manual" Digitization Can't Meet AI's Need for At-Scale Data

In this manual world, total document activation costs ranged from $0.25 to $2.50 per page. With an estimated 250 trillion pages globally, activation at manual economics could exceed $100 trillion – structurally and economically impossible, by design.

Beyond costs, the sheer amount of manual work required to activate documents at scale is utterly impractical, if not impossible. Broadly speaking, it would take 5,000 people working an entire year to activate 1 billion pages. For context, the U.S. National Archives and Records Administration (NARA) holds ~14 billion pages of documents, and it would take approximately 70,000 people working a full year to activate those documents at costs exceeding $5 billion. Virtually impossible. To activate all 250 trillion documents globally? That would take 1.25 billion people and trillions of dollars. The math speaks for itself.

Robots and AI Flip the Equation

With robots and AI, the equation flips entirely. All 14 billion documents at NARA can be activated in about a year with robots, AI, and about 2,000 people, at a cost of under $400 million, or less than 10% of manual costs, and, fundamentally, a more practical scale.

Activation of 14 billion NARA documents (estimate)

Beyond costs and labor, document variability broke every past attempt at automation. Traditional optical character recognition (OCR) and intelligent document processing (IDP) systems assumed uniform pages, fixed templates, and clean forms. Reality was the opposite: archives contain handwritten notes, degraded copies, multilingual records, annotations, forms that changed frequently, and documents with a variety of structures or nonstructures. Each new document type required custom rule-writing and engineering. Automation costs scaled linearly with document diversity for a zero-sum outcome.

For regulated industries – like healthcare, banking, and government – the stakes made it even worse. A misread medical dosage, an incorrect loan amount, a missed legal clause; these aren't inconveniences, they're liabilities. So humans stayed in the loop, necessarily.

Manual document activation never crossed the trust threshold, and the deadlock held.

WHY NOW – THE SHIFT

The Fundamental Shift in Data Economics:

From “We Can’t Afford to Activate” to “We Can’t Afford ***Not to*** Activate”

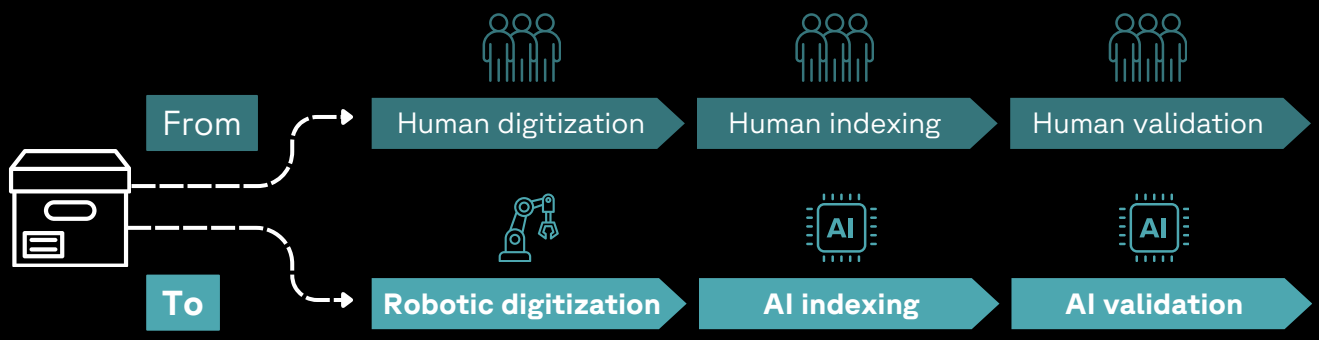
The problem was never a lack of awareness. Every agency leader, CEO, and CIO understands the value of their document archives. The problem was their activation economics: the data value was lower than the high cost of activation. Those economics have now dramatically shifted; data value has skyrocketed, and activation costs have plummeted.

Thanks to AI, Data Value has Skyrocketed

Modern AI doesn't just recognize characters, it understands meaning and relationships. Where legacy OCR asked "what letters are these?," foundation models ask "what does this document mean?" – distinguishing an invoice from a contract, identifying signer roles, understanding obligations and dates in context, inferring implicit structure. No templates required. The intelligence value of a digitized document has increased by orders of magnitude.

Robots and AI Eliminate Most Manual Work and Offer Near Infinite Scale

Physical document preparation and scanning – the most labor-intensive bottleneck – can now be handled by physical digitization robots, powered by AI. Meanwhile, manual data extraction, validation, and classification are handled by AI agents. The workflow shifts:



The Economic Inflection Point

A 90%+ reduction in cost and manual work doesn't just make existing digitization projects cheaper – it makes previously impossible projects strategically viable. Backlogs and document archives that were economically irrational to digitize at $1.00/page become strategic investments at $0.05/page. For the first time, activation at scale is no longer a future aspiration. With robotics and AI, it's a present-tense infrastructure imperative.

Robots and AI Improve Over Time

There is one additional property that makes this fundamentally different from any prior automation wave: robotic and AI systems improve with scale. Legacy digitization systems produced more exceptions as volume increased. AI systems do the opposite; more documents mean better statistical understanding of document structure, fewer exceptions, and higher accuracy over time. Digitization flips from a services cost center into a compounding data infrastructure asset.

The trust threshold has also finally been crossed. Modern AI enables confidence scoring, explainability, human-in-the-loop review, and continuous audit trails. Humans move from operators to supervisors. That's the moment adoption unlocks at scale.

THE STRATEGIC OPPORTUNITY

Documents Are the Largest Untapped Data Asset in Existence

Document activation at scale is becoming a competitive, if not a survival imperative. This shift is comparable to what happened with cloud computing vs. on-premise infrastructure or the economics of satellite launches. A technology breakthrough changes the cost and scale curve so dramatically that an entirely new category of activity becomes rational. For documents, the shift is arguably even more impactful:

For AI Outcomes	High-context institutional data grounds AI outputs in real organizational knowledge, reducing hallucination and improving accuracy by as much as 40% compared to generic RAG systems.
For Competitive Differentiation	Foundational AI models are commoditized. Organizations that lead with AI are those with the richest proprietary data – data no competitor can replicate.

Centuries of human knowledge are recorded on paper. **For the first time in history,** we have the technology to unlock that knowledge (robots) and the technology to use it for the betterment of humanity (AI).

Robots and AI are the first technologies capable of converting 100% of paper archives into structured, machine-readable data at scale – turning the largest unindexed dataset ever created directly into usable AI infrastructure. That dataset already belongs to your organization. The question is simply whether you choose to activate it – or let a competitor do it first.

RIPCORD

About Ripcord

Ripcord is disrupting the $62 billion document intelligence market with robots and AI, turning paper and digital documents into powerful data and vital information to automate operations, fuel insights, and advance AI systems.

Founded out of NASA research, Ripcord is based in Silicon Valley and backed by leading investors including Kleiner Perkins, Google Ventures, Icon Ventures, Lux Capital, MUIP, and Apple co-founder Steve Wozniak.

For more information, visit ripcord.com.