AI+IM ASEAN FORUM

INDUSTRY PRESENTATION

Illuminating Dark Data:

Using AI to Classify, Govern and Safely Retire Decades of Document Clutter



Hemant Prasad
Founder and CEO, Crest Infosolutions

AI+IM
ASEAN SUMMIT

Illuminating Dark Data

Using AI to Classify, Govern and Safely Retire Decades of Document Clutter



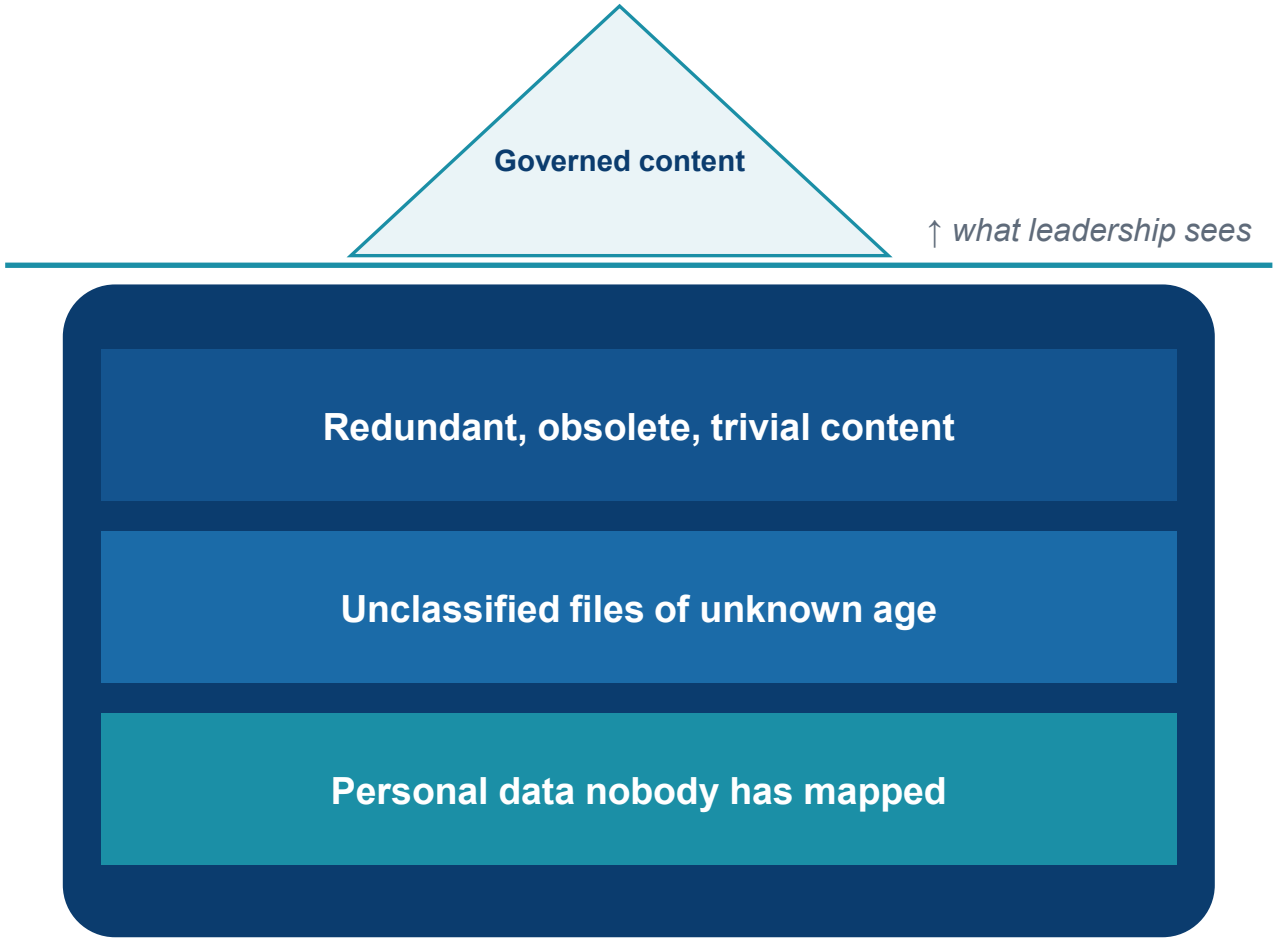
Hemant Prasad

Founder & CEO, Crest Infosolutions Pte Ltd

THE PROBLEM AND THE OPPORTUNITY

Beneath the waterline sits your organisation's untapped memory, and its biggest liability

AI+IM ASEAN SUMMIT



Dark data

Information an organisation holds but cannot describe: what it is, how old it is, whether it is personal, and who owns it.

What leaders believe

"We know what we keep."

What audits and breaches reveal

"We keep far more than we can explain."

THE PROBLEM

Decades of institutional memory, accumulated without a map

AI+IM ASEAN SUMMIT

1990s–2000s

File servers & scanners

Network shares; paper scanned to PDF and TIFF

2000s

Email & ECM

Attachments, departmental archives

2010s

SharePoint & M&A

Inherited repositories from acquisitions

2015+

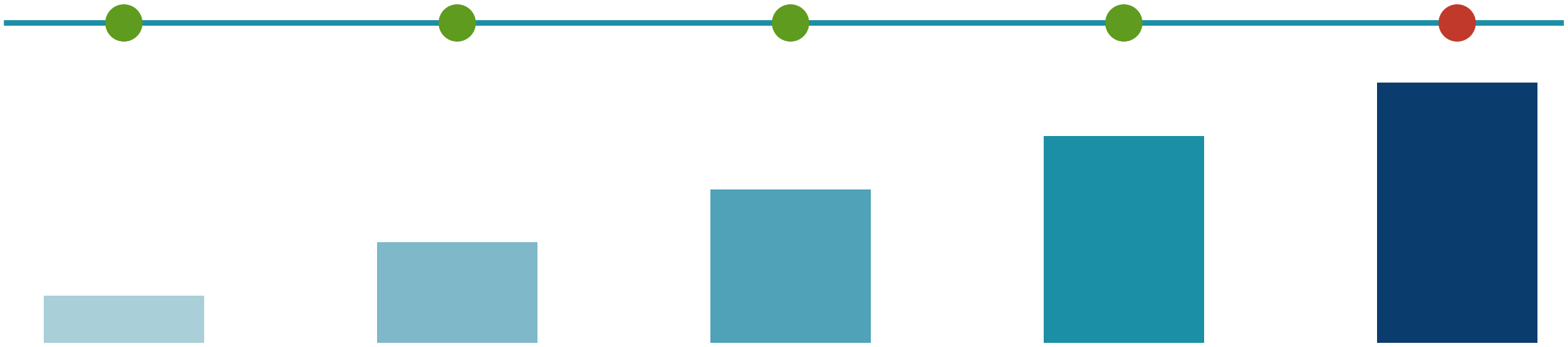
Cloud collaboration

OneDrive, Drive, Box, Dropbox, shadow IT

Today

Hybrid sprawl

Teams, S3, Blob, copies feeding AI tools



Illustrative: unmanaged content accumulates with each generation; deletion rarely gets an owner.

THE PROBLEM

Nobody can use it, and nobody dares touch it

AI+IM ASEAN SUMMIT

Volume

Millions of files. Nobody can read them all.

Opacity

Scans, images, odd formats. Nobody knows what is inside.

Ownership

Owners left, systems merged. Nobody will sign.

Evidence

Even if you delete, can you prove who decided, and why?

THE RETENTION ASYMMETRY

Keeping looks free and safe.
Deleting looks risky and personal.

So nothing is deleted, until a breach, an audit or a regulator forces the question.

AI+IM
ASEAN SUMMIT

Scan to submit your Questions

Scan the Slido QR code to submit your questions.

Response will be shared for as many questions as possible during the closing Q&A.



THE APPROACH

Why the usual answers do not work at estate scale

AI+IM ASEAN SUMMIT

✗

Manual review

Does not scale. Reviewers see a sliver; decisions are inconsistent.

✗

Keyword and regex scanning

Blind to scanned images. Matches patterns, not meaning. Noisy.

✗

LLM on every file

Accurate, but cost scales with the estate and everything goes to a model.

✓

Tiered AI pipeline

OCR, then ML, then LLM only where needed. Explainable, cost-controlled, and a human makes the final call.

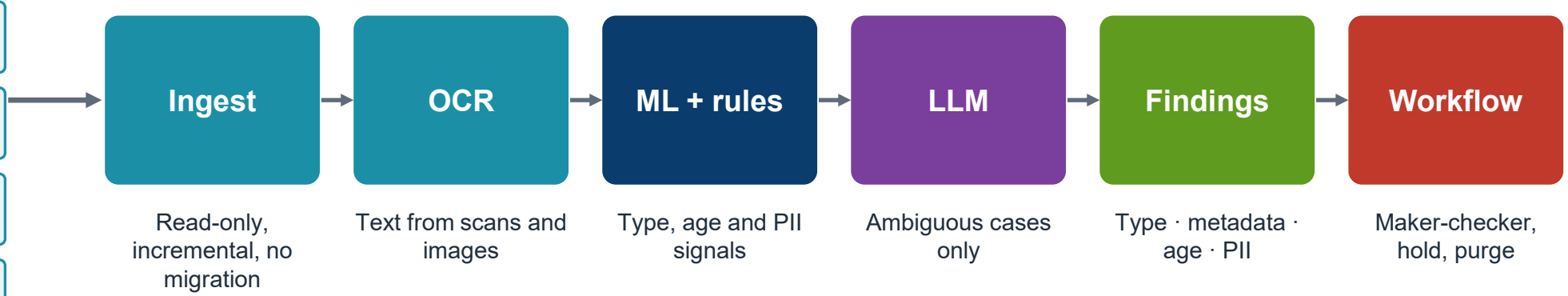
The goal is not to read everything with the smartest model. It is to read everything with the cheapest sufficient one.

THE APPROACH

An AI pipeline built to scan the whole estate, not a sample of it

PDF · Office · JPEG · PNG · TIFF

- File shares
- SharePoint
- OneDrive
- Google Drive
- Box
- Dropbox
- Amazon S3
- Azure Blob



Cross-cutting: audit log · role and source-level access · contentless-sync option · self-hosted inference option

Same pipeline, any source: new repositories plug in without changing the governance model.

THE APPROACH

Cost engineering: LLM tokens only where uncertainty remains



Low cost

Low cost

Higher cost

Highest cost

Principles

Cheapest sufficient method first

Confidence scored separately for OCR quality and classification

LLM limited to approved types and fields; no free-form categories

Illustrative proportions; the real split depends on the estate, document mix and scan quality.

THE APPROACH

From an anonymous scan to a governed record



Classifier
OCR → ML → LLM
approved taxonomy only

Governed record

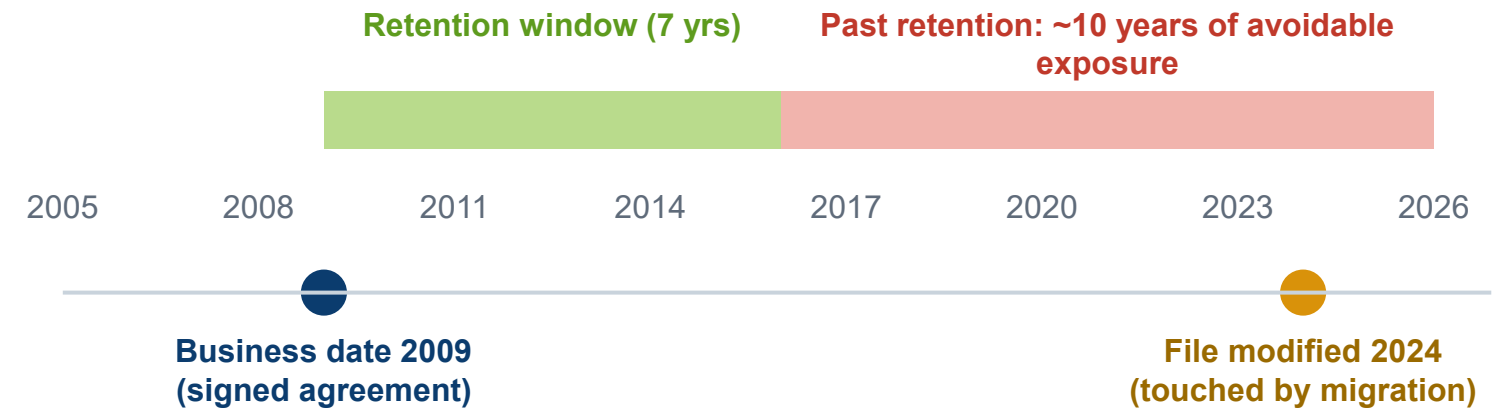
Type	Loan agreement
Fields	Counterparty · effective date · account ref
Business date	12 Jun 2009
PII	Yes: national ID, address
Retention	Per policy for this type

High confidence → accepted automatically

Low confidence → routed to a human review queue

THE APPROACH

Age by what the document says, not by when a migration touched it



LIFECYCLE LABELS

Active | Review due | Past retention | Legal hold

Signals combined

Dates in content and stamps · headers · type-specific rules · references · metadata · ML evaluation

THE APPROACH

Find the personal data, then rank it by how long it has been sitting there

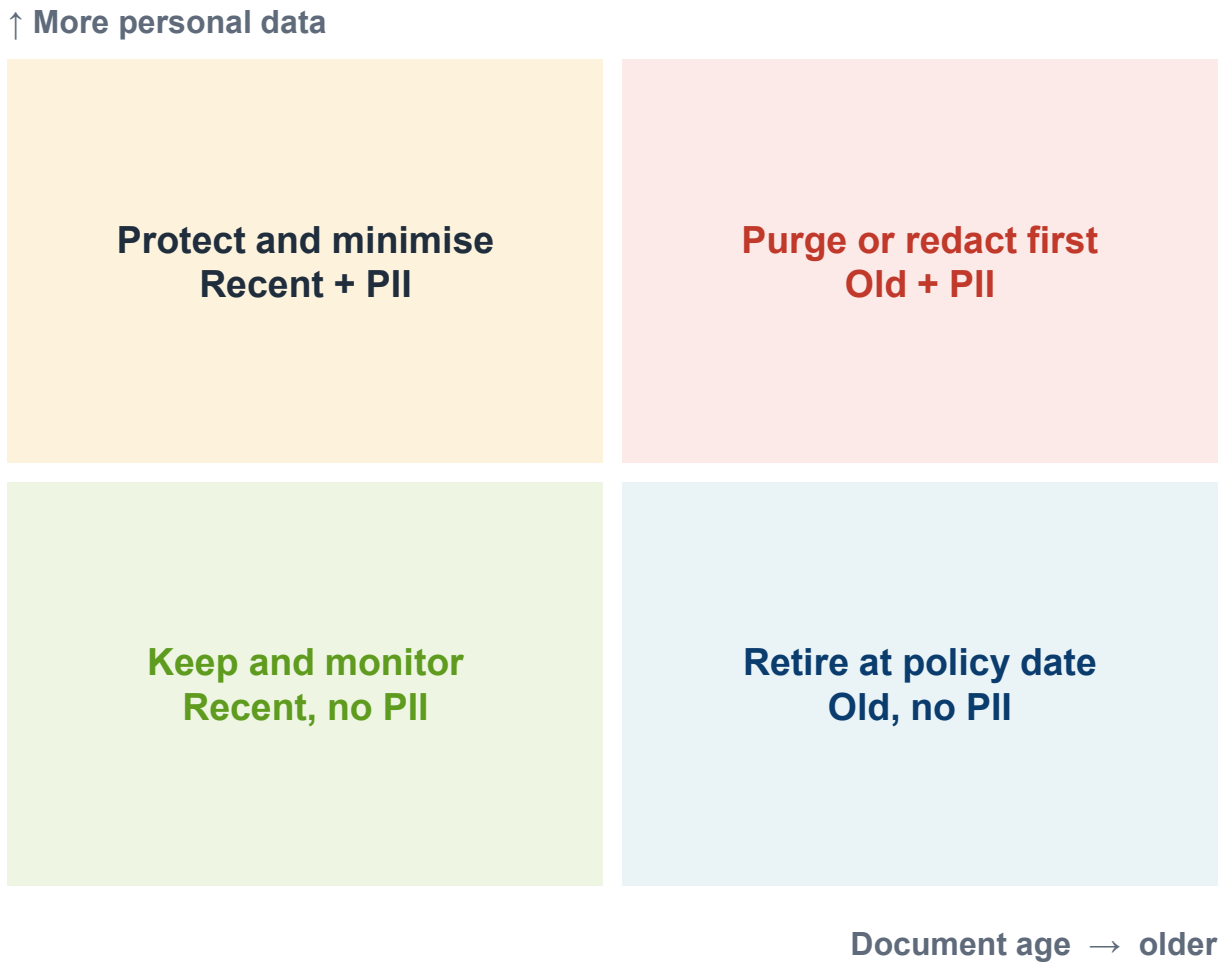
AI+IM ASEAN SUMMIT

DETECTED AND MAPPED TO OBLIGATIONS

- National ID and passport numbers
- Names, addresses and contact details
- Financial and account data
- Health and medical information
- Employee and HR records

GDPR | **PDPA** | **Sector rules**

Flagged by type, library and age: so exposure is queryable, not anecdotal.



THE UPSIDE

Every document gets its best future: Activate it, Retain it, or Retire it

AI+IM ASEAN SUMMIT

Illuminated estate classified · aged · PII-mapped

RETIRE

Past retention, no business value: purge with approval

HOLD

Legal or business reason: keep, with a recorded reason

ACTIVATE

Useful and governed: curate into the AI data lake

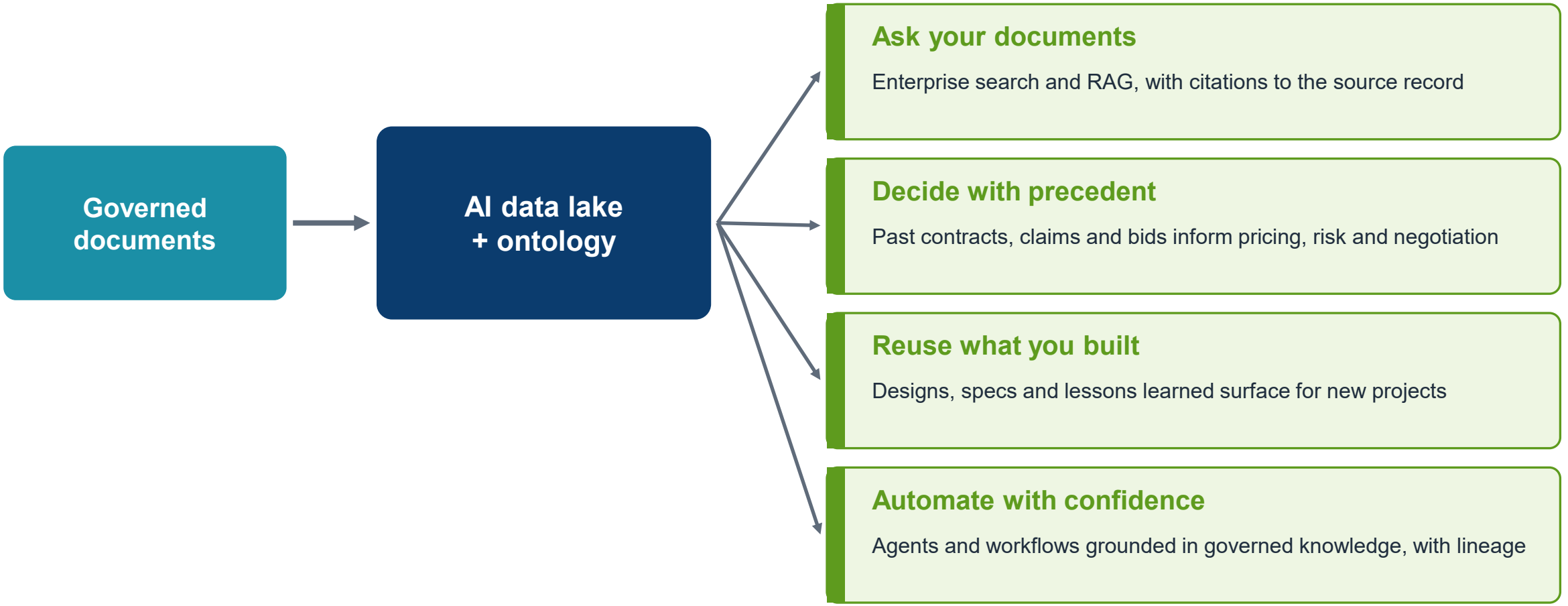
Entry criteria for the AI data lake

- ✓ Classified against approved types
- ✓ PII redacted or cleared
- ✓ Business date and provenance known
- ✓ Owner and retention label attached
- ✓ Lineage recorded in the audit log

AI is only as trustworthy as the documents it learns from. Governance is the entry ticket.

ACTIVATE

Activated knowledge powers search, insight and better decisions



ACTIVATE

Ontology in action: a 2009 policy meets a 2026 flood claim

AI+IM ASEAN SUMMIT

1 DECADES OF DOCUMENTS

Policy PDF · 2009
scanned · clause 14: water damage

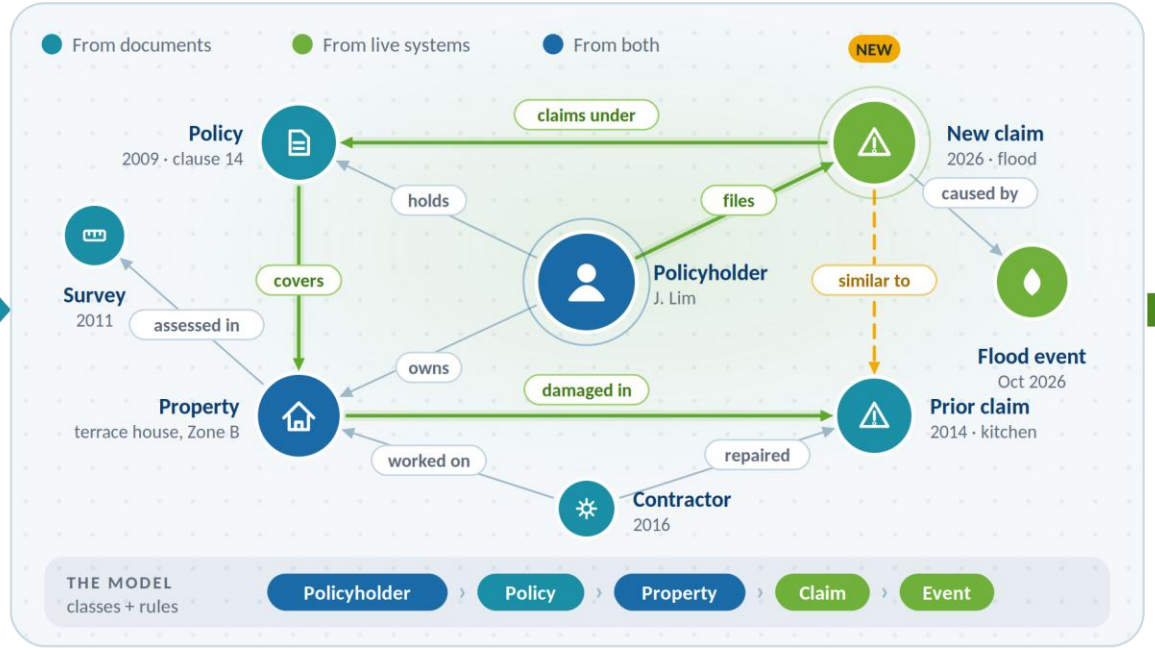
Survey report · 2011
TIFF · flood zone, elevation

Claim file · 2014
JPEG photos · kitchen damage

Repair contract · 2016
Office · contractor, scope of work

OCR · ML · LLM

2 ONTOLOGY: SHARED MEANING



joined by ID

3 LIVE BUSINESS DATA

CRM · customers
policyholder, tenure, channel

Claims system
new flood claim #2026-118

ERP · payments
invoices, payees, reserves

IoT and weather feeds
river level, rainfall alerts

WHAT THE BUSINESS CAN NOW ASK

1. **Is this flood covered?**
   Clause 14 of the 2009 policy, read from the scan, applies to this property.

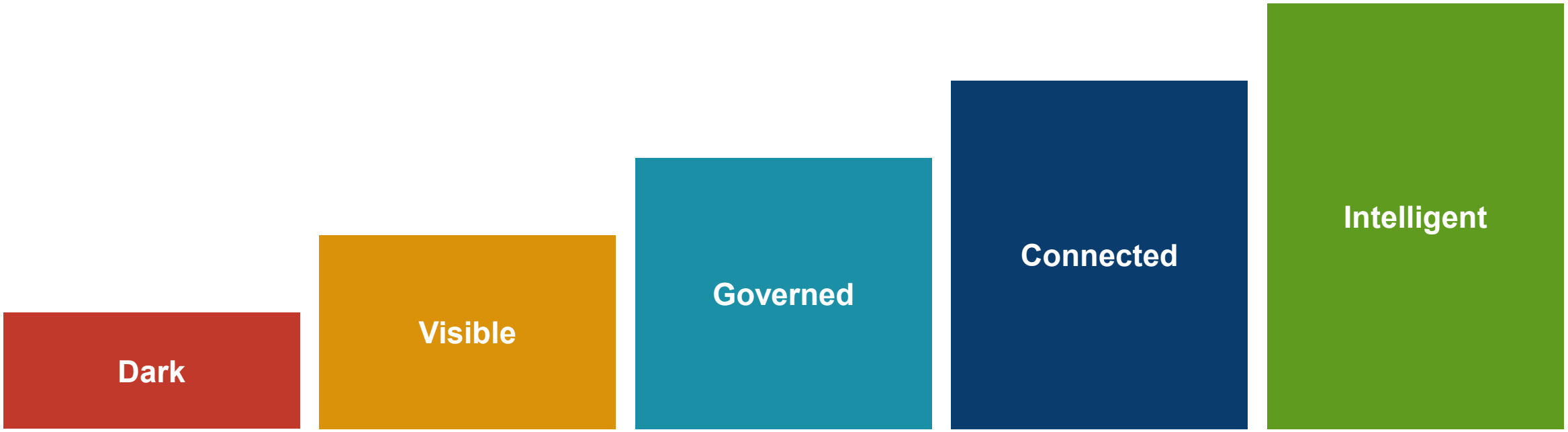
2. **Was this damage claimed before?**
   A 2014 kitchen claim matches: flagged for review before payout.

3. **Who else is exposed?**
   Other Zone B properties carrying the same clause, across the portfolio.

THE UPSIDE

AI maturity starts with knowing what you have

AI+IM ASEAN SUMMIT



ACROSS INDUSTRIES

Same pattern, different paperwork, different prizes

AI+IM ASEAN SUMMIT

Banking & Finance

PILES UP

KYC scans, loan files, closed-account records

UNLOCKS

Credit precedent, customer history, faster checks

CARRIES RISK

Identity-grade PII on former customers

Insurance

PILES UP

Claims files, medical reports, adjuster notes

UNLOCKS

Fraud patterns, claims precedent, underwriting insight

CARRIES RISK

Health data, a sensitive category

Telco

PILES UP

SIM registrations, contracts, billing disputes

UNLOCKS

Contract intelligence, dispute and churn insight

CARRIES RISK

ID documents of former subscribers

Public Sector

PILES UP

Citizen applications, case files, scanned archives

UNLOCKS

Policy precedent, case search, service insight

CARRIES RISK

Citizen PII plus records-law disposal rules

Retail

PILES UP

Loyalty sign-ups, returns, supplier contracts

UNLOCKS

Supplier-terms intelligence, customer insight

CARRIES RISK

High-volume consumer PII, consent drift

Engineering

PILES UP

Drawings, specs, tenders, project files

UNLOCKS

Design reuse, lessons learned, bid intelligence

CARRIES RISK

IP exposure plus personnel PII

Typical patterns across large organisations; retention periods and legal bases vary by jurisdiction and policy.

THE COST OF THE REST

Keeping what has no value is a recurring cost with no return

AI+IM ASEAN SUMMIT

- Primary storage
- Backup and replication
- eDiscovery and legal review
- Search time and lost productivity
- Audit and compliance effort
- A larger surface to defend

Every layer recurs, every year.

Nobody can say which files matter, so everything is paid for as if it did.

REAL RISK

When old data meets a bad day: three cases worth remembering

AI+IM ASEAN SUMMIT

BANKING & FINANCE

Latitude Financial · 2023

~14M

customers, past customers and applicants affected (first estimate: 328,000)

- 6.1M records dated back to at least 2005; ~94% supplied before 2013
- 7.9M driver-licence numbers

Lesson: Closed accounts still carried live identity risk.

TELCO

Optus · 2022

9.5M+

current and former customers' records accessed

- 2.1M customers had ID document numbers compromised
- Data sat on a domain dormant since 2017 (per ACMA filing)

Lesson: Regulators now ask why former customers' data was still held.

REGULATORY ACTION

Deutsche Wohnen · 2019

€14.5M

GDPR fine; no breach was proven

- Archive could not delete data no longer needed
- Included salary slips, tax and bank statements

Lesson: Hoarding is itself the violation.

Sources: Latitude Financial statements (Mar 2023) via The Record and Security Affairs; ACMA/OAIC court filings re Optus (2024); Berlin DPA decision (2019) via Legal500 and Mondaq. Figures as publicly reported.

REAL RISK

From forgotten file to front page

AI+IM ASEAN SUMMIT



Project ends, customer leaves, system retired

No review, no retention clock, no steward

Legacy store, stale domain, over-shared folder

Everything ever kept is now exposed

Notification, remediation, regulators, class actions, churn

Latitude

Records from accounts closed up to ~18 years earlier were stolen

Optus

Dormant-since-2017 domain held customer data until after the attack

Deutsche Wohnen

Fined because the archive could not delete

Deleting what you never needed is the cheapest breach control you will ever buy.

COMPLIANCE

Privacy law now has a maximum retention clock, not just a minimum

AI+IM ASEAN SUMMIT

EU GDPR

Storage limitation and data minimisation (Art. 5); erasure (Art. 17); privacy by design (Art. 25)

Singapore PDPA

Retention Limitation Obligation: stop retaining personal data once no longer needed for legal or business purposes

Malaysia PDPA

Retention Principle: do not keep personal data longer than necessary; recent amendments add breach notification

Regional peers

Thailand, Indonesia and others follow similar purpose-limitation and deletion duties

Minimum retention (records, tax, sector rules) — **Maximum retention (privacy law)**

Defensible retention window

Too early: you destroyed a record you had to keep

Policy must live here, and be provable

Too late: you kept personal data you had no basis to hold

Simplified summary of GDPR Art. 5(1)(c),(e), 17, 25; Singapore PDPA Retention Limitation Obligation; Malaysia PDPA Retention Principle. Not legal advice.

GOVERNANCE

Maker, checker, executor: no one person can delete alone

AI+IM ASEAN SUMMIT

Data governance manager
MAKER

- Reviews AI findings
- Selects documents
- Proposes purge or hold, with the explanation

→

Business owner
CHECKER

- Approves a purge or places a hold
- Records a business reason
- Cannot approve own request

→

Executor
EXECUTES

- Typed confirmation
- Purges or redacts at source
- Verifies the result

Legal hold overrides retention entirely: held documents can never be selected for purge

Every request, approval, rejection, hold and execution lands on an append-only audit log

GOVERNANCE

Evidence, not assertions

AI+IM ASEAN SUMMIT

09:12 — Purge request submitted
1,240 documents · AI explanation attached

11:40 — Approved by business owner
Reason recorded · not the requester

11:41 — Legal-hold check
12 documents excluded automatically

14:05 — Executed and verified
Deletion confirmed at source

14:06 — Evidence exported
CSV for audit or regulator review

Questions you can now answer

What was deleted?
Why was it deleted?
Who approved it?
Under which policy and document type?
Was a hold in place?
Can you prove it?

Illustrative governance log entries.

THE PLAYBOOK

A practical path for a large organisation

AI+IM ASEAN SUMMIT

1 Frame	2 Discover	3 Decide	4 Activate & retire	5 Sustain
DO	DO	DO	DO	DO
Retention schedule, taxonomy, hold rules, owners per repository	Connect high-risk repositories read-only; scan and baseline	Waves, riskiest first; route to business owners	Curate useful content into the AI data lake; retire the rest with verification	Scan new content continuously; review policy
EXIT WHEN	EXIT WHEN	EXIT WHEN	EXIT WHEN	EXIT WHEN
Approved policy and RACI	**Heat-map: type × age × PII**	**Approved purge and hold batches**	**AI-ready corpus; storage reclaimed**	**Dark data stops regrowing**

Start with the repository that combines oldest content, most personal data and weakest ownership.

THE PLAYBOOK

Designed so IT and security can say yes

AI+IM ASEAN SUMMIT

Read-only connectors

Incremental sync; no migration, no rip-and-replace

Contentless option

Keep metadata; process files transiently for OCR and classification

Self-hosted inference

Sensitive document types never leave your boundary

Dual-layer permissions

Workflow role and source-level access must both allow it

Segregation of duties

Enforced by the system, not by a policy document

Verified execution

Redaction and purge are confirmed at source, then logged

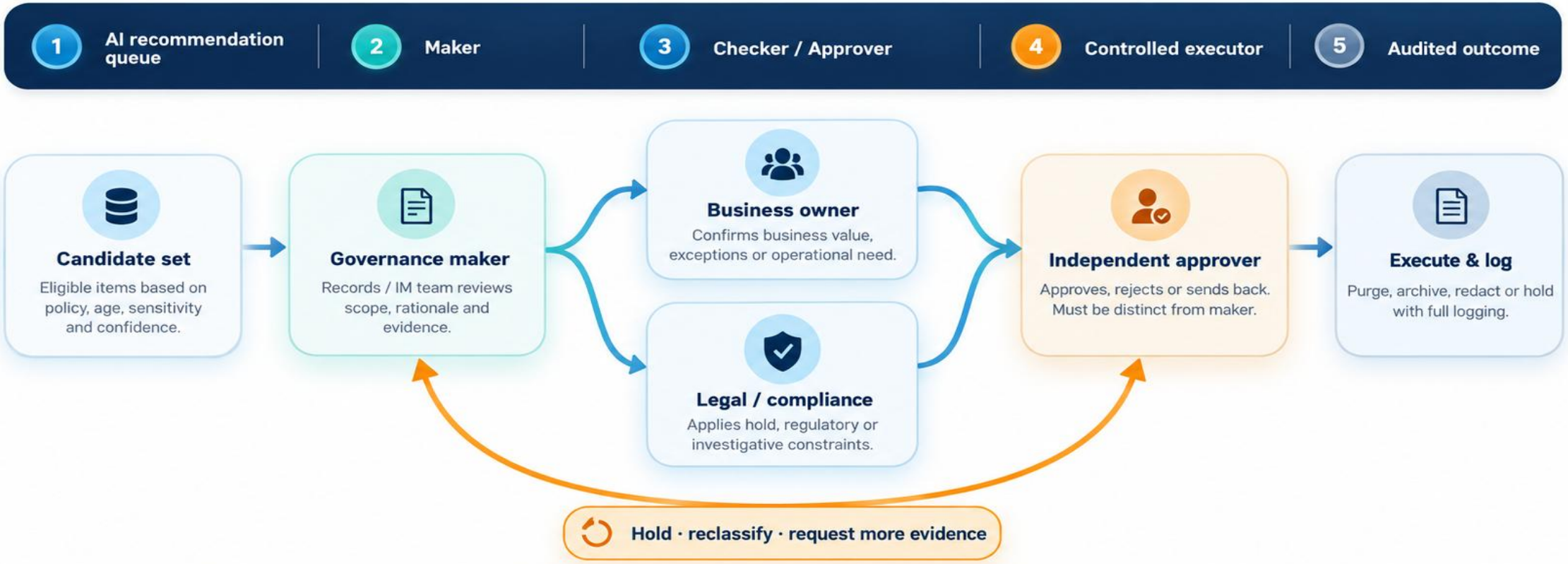
Large organizations need a federated operating model.

Central Governance	Taxonomy · retention schedules · policy conflict resolution
IT / Cyber	Connectors · identity · controls · telemetry · resilience
Business Owners	Purpose, value, business event and disposition authority
Legal / Privacy	Holds, jurisdictional exceptions, PII obligations
Independent Audit	Sampling, segregation and effectiveness evidence

HUMAN CONTROL AND SEGREGATION OF DUTIES

Governance is a decision system, not a purge button.

AI+IM ASEAN SUMMIT



Segregation of duties principle

A defensible workflow separates recommendation, approval and execution — and preserves the decision trail when the answer is not now, not this way or do not dispose.

THE OUTCOME

From clutter to knowledge, from "we think" to "we can prove it"

AI+IM ASEAN SUMMIT

TODAY		WITH AN AI-DRIVEN PIPELINE
Unknown content	→	Classified, searchable estate
Age by file date	→	Age by business date
PII unmapped	→	PII located by type and library
Retention on paper	→	Retention applied with approvals
Evidence after the fact	→	Continuous audit trail

AI-ready knowledge base | **Better decisions from history** | **Lower cost and attack surface** | **Faster audits and holds**

TAKEAWAYS

Find what the business needs. Retire the rest.

AI+IM ASEAN SUMMIT

1. Decades of documents are institutional memory, but most of it is invisible.
2. Tiered AI (OCR, ML, then LLM) makes whole-estate understanding affordable.
3. Activate what matters: feed AI search, analytics and decisions through an ontology.
4. Retire the rest with maker-checker and audit, and cost, security and compliance follow.

YOUR NEXT STEP

Pick one high-risk repository.
Scan it read-only.
See your dark data.

Let the findings make the business case.

Do not edit
How to change the design



Audience Q&A

ⓘ The Slido app must be installed on every computer you're presenting from

slido

AI+IM
ASEAN SUMMIT

The future of information governance is knowing what not to keep.

Understand what you have. Protect what matters.

Retire what no longer serves, with evidence.

Hemant Prasad
Founder & CEO, Crest Infosolutions Pte Ltd
www.crestsolution.com

Thank you

Let's Connect on LinkedIn