



Discovering Sensitive Production Patterns in Your AI and Test Environments

Why Data Discovery Is the First Control in Test Data Management and AI Governance



enov8.com | enquiries@enov8.com

Executive Summary

Every enterprise holds sensitive data in places it did not intend. Production databases are copied into test environments, extracts are shared with vendors, log files capture customer details, and AI teams now assemble training sets, evaluation corpora and retrieval stores from whatever data is closest to hand. Each of these movements carries personal and commercially sensitive information into environments that were never designed to protect it.

These are what this paper calls **sensitive production patterns**: a tax file number buried in a free-text notes field, a credit card number nested inside a JSON payload, customer records copied into an AI evaluation set, or personal information embedded in an application log. None of them sit in a column labelled as sensitive, and all of them travel wherever production data is copied.

Most organisations respond by investing in masking, synthetic data and access controls. These are the right controls, but they share a single dependency: they only protect the data you know about. A masking rule cannot be written for a column nobody has identified. A retrieval pipeline cannot exclude records that were never classified. An auditor cannot be shown evidence of protection for data whose location is unknown.

This paper argues that **data discovery is the first control**, not a preliminary task. It examines why sensitive production patterns are harder to find than most teams assume, how the problem differs between test environments and AI environments, and what a modern, continuous discovery model looks like. It closes with a practical adoption path that moves organisations from one-off scans to discovery as an embedded, evidenced part of software and AI delivery.

***You cannot mask, govern or evidence what you have not found.
Discovery is where data protection begins.***

Production Data Has Escaped Production

Production is usually the most controlled environment in the enterprise. It has hardened access, monitoring, change control and a clear owner. The problem is that production data rarely stays there. Over years of delivery activity, it spreads outward into a far less controlled estate.

The Non-Production Estate Is Larger Than Production

A typical enterprise application is supported by several non-production environments: development, system test, integration, performance, pre-production and training. Each is refreshed periodically, often from a production copy. Multiply that across hundreds of applications and the non-production footprint can hold many times more copies of sensitive records than production itself, with weaker controls, broader access and less monitoring.

AI Has Multiplied the Exposure

AI adoption has added a new class of consumer. Data science teams extract datasets for model training and fine-tuning. Engineering teams load documents and records into vector databases to support retrieval-augmented generation. Prompts, completions and agent activity are logged for debugging and evaluation. Each of these creates a new copy of production data, frequently outside the governance processes built for traditional test environments.

The Visibility Gap

The combined effect is a visibility gap: organisations are accountable for data they can no longer locate with confidence. Industry research consistently points to the same pattern.

60%

of organisations have experienced data breaches in non-production environments

84%

allow compliance exceptions in non-production despite having masking mandates

86%

plan to invest in AI data privacy solutions in the next one to two years

The gap between intent and practice is not caused by a lack of policy. It is caused by a lack of knowledge. Teams grant exceptions because they cannot be certain which fields need treatment, and they defer masking because they cannot see the full scope of the work.

Why Discovery Comes First

Data protection programmes are often designed from the control outwards: select a masking tool, define rules, apply them to known systems. This sequence quietly assumes that the inventory of sensitive data is complete and current. In practice, it rarely is.

Discovery Defines the Scope of Masking

Masking is only as complete as the discovery that precedes it. When sensitive fields are identified manually, through interviews, data dictionaries or tribal knowledge, the result reflects what people remember rather than what the data contains. Undocumented columns, repurposed fields and free-text notes are routinely missed, and every miss becomes an unmasked field in every downstream copy.

Discovery Produces the Evidence Regulators Expect

Privacy and AI regulation increasingly asks organisations to demonstrate control rather than assert it. Frameworks such as GDPR, the Australian Privacy Act and emerging AI legislation all expect organisations to know what personal data they hold, where it resides and how it is protected. A current, system-generated discovery inventory is the foundation of that evidence. Without it, compliance reporting relies on attestations that are difficult to defend.

Discovery Underpins Trust in AI

AI systems inherit the characteristics of their data. If sensitive attributes enter a training set, a model can memorise and reproduce them. If personal records enter a retrieval store, an assistant can surface them to users who should never see them. Discovery is the control that lets AI teams make an informed decision about what data is allowed into a pipeline before it is ingested, rather than discovering the problem in an output.

Every downstream control inherits the blind spots of discovery. Get discovery wrong and every copy, model and report carries the error forward.

What Sensitive Production Patterns Really Look Like

Sensitive data is rarely confined to neatly labelled columns called Name or Email. Real enterprise estates contain sensitive information in forms that simple keyword searches and column-name checks will not reliably find. Effective discovery must recognise patterns, not just labels.

Pattern Type	Typical Examples	Where It Hides	Why It Is Missed
Direct identifiers	Names, email addresses, phone numbers, tax file numbers, passport and licence numbers	Customer, employee and account tables	Columns are often abbreviated, reused or named generically
Financial identifiers	Card numbers, bank account and BSB numbers, IBANs	Payments, billing and reconciliation systems	Values stored in text fields or split across columns
Quasi-identifiers	Date of birth, postcode, gender, occupation	Demographic and profile data	Harmless alone, identifying in combination
Unstructured content	Case notes, emails, chat transcripts, documents	Comment fields, CLOBs, file stores, ticketing tools	Sensitive values embedded in free text
Semi-structured data	JSON and XML payloads, API messages	Integration tables, queues, event logs	Nested attributes invisible to column scans
Derived and relational	Keys and references that link back to people	Foreign keys, history and audit tables	Identity propagates through relationships

Patterns, Not Just Names

Reliable discovery combines several techniques. Metadata analysis examines schema, column names and data types. Content profiling samples actual values and tests them against known formats. Validation logic, such as check-digit algorithms for card numbers and national identifiers, separates genuine matches from coincidental ones. Reference dictionaries identify names, locations and other domain values. Together, they find what no single method would.

Context Determines Sensitivity

The same value can be sensitive or harmless depending on context. A date is benign in a transaction table but identifying when it is a date of birth attached to a postcode. Mature discovery therefore classifies data by context and combination, assigning a sensitivity rating that reflects real re-identification risk rather than simply flagging every date or number it encounters.

Discovery in Test Environments

Test environments are where the discovery problem is oldest and most entrenched. They are also where the consequences of incomplete discovery are most frequently realised, because they combine production-like data with broad access for internal teams, contractors and offshore partners.

Legacy Estates and Schema Sprawl

Long-lived systems accumulate thousands of tables, many of them undocumented. Columns are added to meet urgent requirements, repurposed years later and rarely retired. Data dictionaries, where they exist, fall out of date quickly. Manual discovery across this sprawl is slow, expensive and incomplete, which is why many masking programmes stall after covering only the most obvious systems.

Release Drift

Discovery is not a one-time exercise because the schema does not stand still. Every release can introduce a new table, a new column or a new integration that carries sensitive values. If discovery is not repeated as part of the delivery lifecycle, masking coverage decays release by release, and the next environment refresh quietly reintroduces unprotected data.

Shadow Copies and Unmanaged Extracts

Alongside formally managed environments sit ad hoc copies: developer databases, spreadsheet extracts, flat files shared with vendors and snapshots taken for defect investigation. These rarely appear in any inventory. Discovery that extends to file stores and unmanaged databases, combined with a clear inventory of environments, is the only practical way to bring them back under control.

Proving Masking Worked

Discovery has a second role after masking: validation. Rescanning a masked environment confirms whether sensitive patterns remain. This closes the loop between policy and outcome, turning masking from an assumed control into a verified one, and provides the evidence needed when auditors ask how the organisation knows its test data is safe.

Enov8 Test Data Management: Profile, discover and classify sensitive data across your estate, then mask, validate and provision compliant test data at enterprise scale. enov8.com

Discovery in AI Environments

AI environments introduce the same discovery challenge in new and less familiar forms. The data is often less structured, the pipelines are newer and the teams building them may sit outside traditional data governance. Discovery must follow data into every stage of the AI lifecycle.

Training and Fine-Tuning Datasets

Datasets assembled for model training are frequently drawn from production systems, data lakes and document repositories. Once a model has been trained on sensitive data, removing that influence is difficult and costly. Discovery before ingestion allows teams to exclude, mask or synthesise sensitive attributes while the decision is still cheap to make.

Retrieval Stores and Vector Databases

Retrieval-augmented generation depends on loading enterprise content into a searchable store. Documents, tickets, emails and knowledge articles often contain personal and confidential information. Once converted into embeddings, that content is much harder to inspect. Scanning source content before it is chunked and embedded is significantly more practical than attempting to audit a vector store after the fact.

Prompts, Logs and Agent Activity

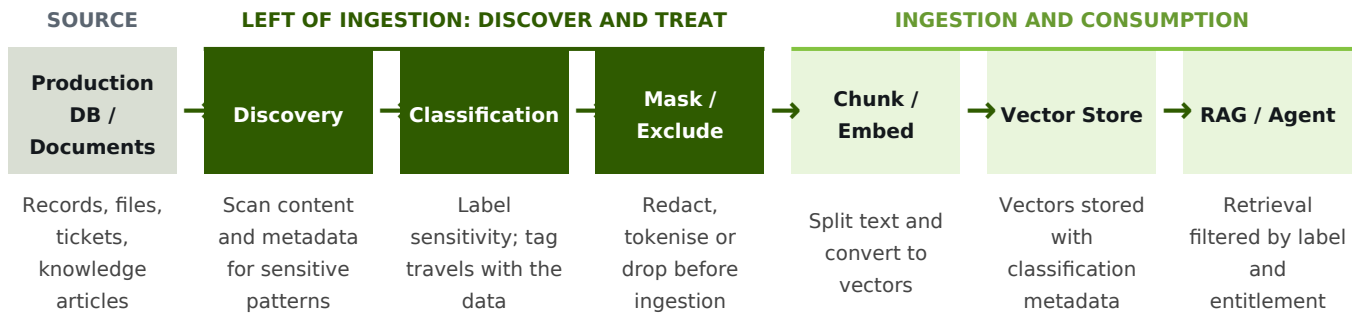
AI systems generate data as well as consume it. Prompts submitted by users, model responses, tool calls made by agents and evaluation traces are routinely logged. These logs can contain the very information the organisation worked hard to keep out of training data. They need to be treated as a data source in their own right and brought within the scope of discovery.

Evaluation and Test Datasets for AI

AI systems require testing like any other software, and their test datasets must be representative of production behaviour. This is where AI and test environments converge. The same discipline that governs test data, discover first, then protect, then provision, applies directly to AI evaluation data.

Example Flow: Left of Ingestion in a RAG Pipeline

The flow below shows where discovery sits in a typical retrieval-augmented generation pipeline. The critical design choice is that discovery, classification and treatment all happen before content is chunked and embedded. Everything to the right of that line inherits the protection, or the exposure, established on the left.



🔄 **Feedback loop:** prompts, responses and agent tool calls are logged, then rescanned by discovery as a new data source

Why treatment must happen before embedding. Chunking splits a document into many overlapping fragments, so a single sensitive value can be replicated across several chunks. Embedding then converts each fragment into a numeric vector that cannot be inspected with conventional scanning tools. Research into embedding inversion has shown that meaningful portions of source text can be reconstructed from vectors, so an embedding should be treated as a derivative of the original data, not an anonymised version of it. Masking or excluding sensitive values at source is far more reliable than attempting to find them once they are dispersed and encoded.

Why classification should travel with the data. When discovery assigns a sensitivity label at the document or record level, that label can be carried into the vector store as chunk metadata. Retrieval can then filter results by both classification and the requesting user's entitlements, so an assistant or agent only surfaces content the user is permitted to see. Without that metadata, the vector store becomes a flat index in which every chunk is equally retrievable, regardless of how sensitive its source was.

Why the loop must close. AI systems generate new data with every interaction. Prompts may contain customer details typed by users, responses may echo retrieved content, and agent tool calls may pull records from connected systems. Treating these logs as a data source and returning them to discovery ensures that exposure created downstream is identified and treated with the same discipline as the original source content.

This is the practical meaning of Enov8's Left of Ingestion Control Model: control is applied where data is still visible, structured and attributable, rather than after it has been transformed into a form that is difficult to audit.

The safest point to control AI data is before it enters the pipeline. Once data is embedded or learned, it is far harder to find and remove.

Two Environments, One Discovery Model

Test environments and AI environments have historically been owned by different teams, governed by different processes and supported by different tools. From a data risk perspective, however, they share the same root problem: production data flowing into places with weaker controls. Treating them separately doubles the effort and leaves gaps between the two.

Dimension	Test Environments	AI Environments
Primary data sources	Production database copies, extracts and refreshes	Data lakes, document stores, production extracts and logs
Dominant data shape	Structured, relational	Unstructured and semi-structured, plus structured features
Key risk	Unmasked sensitive fields exposed to broad internal and partner access	Sensitive data learned by models or surfaced through retrieval and agents
When discovery must happen	Before and after each refresh, and with every schema change	Before ingestion into training, retrieval or evaluation pipelines
Downstream control	Masking, subsetting, synthetic data and virtualised provisioning	Exclusion, masking, synthetic data and access-controlled pipelines
Evidence required	Masking coverage and validation reports	Dataset lineage, content classification and exclusion records

A unified discovery model applies one classification taxonomy, one set of detection rules and one inventory of sensitive data across both. Test and AI teams then draw on the same understanding of where sensitive data lives and how it must be treated. This consistency is what allows organisations to answer a regulator, an auditor or a board with a single, coherent account of data risk.

Principles of Modern Data Discovery

Discovery practices that were adequate for periodic compliance reviews are no longer sufficient. The pace of delivery and the scale of AI adoption call for a different operating model, built on five principles.

- **Automated, not manual.** Discovery should be driven by scanning and profiling, not interviews and spreadsheets. Automation provides coverage that manual effort cannot reach and removes dependence on individual knowledge.
- **Continuous, not one-off.** Discovery should run whenever data or schema changes: on each release, before each environment refresh and before each ingestion into an AI pipeline. A point-in-time inventory is out of date almost as soon as it is produced.
- **Pattern-aware and contextual.** Detection should combine metadata, content sampling, format validation and reference data, and should classify risk by context and combination rather than by isolated values.
- **Estate-wide.** Discovery should span relational databases, file stores, semi-structured payloads and the data sources feeding AI pipelines, including environments that sit outside formal inventories.
- **Actionable and evidenced.** Discovery results should flow directly into masking rules, provisioning policies and AI ingestion decisions, and should produce reports that stand as audit evidence without additional manual effort.

This approach reflects Enov8's Left of Ingestion Control Model: identify and treat sensitive data before it is ingested into a test environment or AI pipeline, rather than attempting to contain it after it has spread.

Operational Discovery Is Not Data Cataloguing

A reasonable question at this point is whether an existing data catalogue already solves the problem. Enterprise catalogue and governance platforms such as Microsoft Purview, Collibra and Informatica provide valuable estate-wide visibility: metadata harvesting, lineage, business glossaries, ownership and, in many cases, sensitivity classification. Organisations that have invested in these platforms should keep them.

The distinction is one of purpose. A catalogue answers the question of what data the organisation holds and what it means. Operational discovery answers a narrower and more urgent question: **is this data safe to move into this environment or pipeline, right now?** That question is asked by delivery teams, at delivery speed, every time a refresh, release or AI ingestion takes place.

Different Questions, Different Moments

Dimension	Enterprise Data Catalogue	Operational Discovery
Primary question	What data do we have, what does it mean and who owns it?	Is this dataset safe to move into this environment or pipeline now?
Primary users	Data governance, data stewards, analytics and the privacy office	Test, DevOps, platform engineering, and AI and data science delivery teams
Scope	Registered production and analytical data assets	Production sources plus the non-production and AI copies derived from them
Trigger	Scheduled scanning of governed assets	Delivery events: refreshes, releases, schema changes and AI ingestion
Output	Inventory, glossary, lineage and policy documentation	Masking rules, exclusion decisions, validation results and environment-level exposure
What it proves	Sensitive data has been identified and classified	Sensitive data was treated before it moved, and the treatment was verified

Discovery That Acts at the Point of Movement

Operational discovery is embedded in software delivery rather than sitting alongside it. Before a **test environment refresh**, it confirms what sensitive data the source contains, and after the refresh it validates that nothing unprotected arrived. Its findings drive **masking** directly, generating and updating rules rather than informing a separate manual exercise. When a **release** introduces new tables or fields, it rescans as part of release readiness. It maps sensitive data to the **environments**

that hold it, including who has access and when those environments are next refreshed. And before **ingestion**, it classifies and treats content before it is chunked, embedded or used for training.

This is where Enov8 focuses. Because Enov8 already governs environments, test data and releases, discovery results are connected to the operational events that create exposure, rather than stored as an inventory that delivery teams must interpret for themselves.

Complementary, Not Competing

The two approaches work best together. Where a catalogue already holds sensitivity classifications, those provide a useful starting point. Operational discovery then extends coverage to the non-production and AI copies that catalogues are not typically designed to track, verifies that protection has actually been applied, and produces evidence of control at the point of use. The catalogue describes the data estate. Operational discovery governs how that data moves through delivery.

A catalogue tells you what sensitive data you have. Operational discovery makes sure it is treated every time it moves.

From Discovery to Control

Discovery creates value when it is connected to the controls that follow it. The operating change organisations need is a closed loop in which every stage builds on the knowledge produced by the one before.

Stage	What Happens	Outcome
1. Discover	Scan databases, files and AI data sources for sensitive patterns	A current inventory of where sensitive data lives
2. Classify	Rate sensitivity by type, context and combination	Consistent risk ratings across test and AI estates
3. Protect	Apply masking, subsetting, synthetic generation or exclusion	Sensitive values removed before data moves
4. Validate	Rescan treated data to confirm protection is complete	Verified, not assumed, compliance
5. Provision	Deliver compliant data to test and AI teams through governed channels	Faster access without new exposure
6. Evidence	Record discovery, treatment and validation outcomes	Audit-ready reporting as a by-product of delivery

Connecting Data to Environments

Knowing where sensitive data lives is only half the picture. Organisations also need to know which environments hold that data, who uses them and when they are refreshed. Linking the discovery inventory to an environment inventory turns a list of sensitive fields into a map of real exposure, and allows teams to prioritise the environments where risk is highest.

Connecting Data to Releases

Because schema change is a primary source of discovery drift, discovery belongs in the release process. When a release introduces new tables or fields, a discovery scan and updated masking rules can become part of release readiness. This prevents new sensitive data from reaching non-production environments unnoticed.

The Value of Getting Discovery Right

Discovery is sometimes treated as a cost of compliance. In practice, it is one of the highest-leverage investments a data protection programme can make, because every downstream control becomes faster, cheaper and more reliable when it is built on accurate knowledge. The benefits below should be read as indicative planning assumptions, with actual outcomes depending on estate size, maturity and scope.

Value Area	What Changes
Reduced exposure	Fewer unprotected copies of sensitive data in test and AI environments, and a smaller surface for breach and misuse
Faster masking programmes	Masking effort is directed at confirmed sensitive fields rather than spent on manual analysis or blanket treatment
Fewer compliance exceptions	Teams no longer need to request exceptions because the scope of required protection is clear
Audit readiness	Discovery, treatment and validation records are available on demand rather than assembled under pressure
Safer AI adoption	AI teams can move quickly with confidence about what data is entering their pipelines
Better quality test data	Understanding data structure and relationships improves subsetting and synthetic generation, producing more realistic test data

Importantly, these benefits compound. A single discovery capability serves test data management, AI data governance, privacy compliance and audit. The same investment supports all four, which is why discovery is best treated as shared infrastructure rather than a project owned by any one team.

How Enov8 Helps

Enov8 provides an integrated platform for governing data, environments and releases across the software and AI delivery lifecycle. Discovery sits at the start of that model and connects directly to the controls that follow.

Test Data Management

Enov8 Test Data Management profiles data sources to discover and classify sensitive information, then applies masking, synthetic data generation and validation to produce compliant datasets. Validation scans confirm that treatment has been applied, and audit reports record what was found, how it was treated and when.

Environment Management

Enov8 Environment Management maintains an inventory of environments, their components and their usage. Combined with discovery results, this shows which environments contain sensitive data, who has access and when refreshes occur, giving teams a clear view of where risk is concentrated.

Database Virtualisation

Enov8 VirtualizeMe (vME) delivers lightweight, virtual copies of compliant datasets in minutes. When virtual datasets are created from a masked and validated source, protection becomes a property of every copy provisioned, reducing both storage costs and the number of physical copies holding sensitive data.

Release Governance

Enov8 Release Management brings discovery into release readiness, so that schema changes can trigger rescans and updated masking rules before new data reaches downstream environments.

Enov8 Platform: Discover sensitive data, protect it at source and provision compliant datasets to test and AI teams, with governance across data, environments and releases. enov8.com

A Practical Adoption Path

Discovery maturity is built progressively. Enov8's Stabilise, Standardise, Optimise approach provides a practical path that delivers early risk reduction while building towards continuous, embedded discovery.

STAGE 1 Stabilise	STAGE 2 Standardise	STAGE 3 Optimise
→ Identify the highest-risk systems and AI data sources → Run baseline discovery scans → Establish an initial sensitive data inventory → Map sensitive data to existing test environments → Prioritise masking of confirmed high-risk fields	→ Adopt one classification taxonomy for test and AI data → Bind masking and exclusion rules to discovery results → Rescan after masking to validate coverage → Embed discovery in environment refresh processes → Produce standard audit reports	→ Trigger discovery automatically on schema change → Scan data before AI ingestion by default → Provision compliant data through self-service → Surface compliance posture in dashboards → Generate evidence continuously

The first stage is deliberately narrow. Organisations that attempt to discover everything at once often stall under the weight of the task. Starting with the systems and pipelines that carry the most sensitive data produces early, visible risk reduction and builds the credibility needed to extend coverage across the wider estate.

Final Thoughts

Data protection does not begin with masking. It begins with knowing what you have.

Enterprises are asking their test and AI environments to do more than ever, and both depend on data that looks and behaves like production. That dependency is legitimate, but it carries risk that grows with every copy, every pipeline and every new use case. The organisations that manage this risk well will not be those with the most controls. They will be those whose controls are built on an accurate, current understanding of where sensitive data lives.

Data discovery provides that understanding. Made automated, continuous and connected to masking, provisioning, environments and releases, it turns data protection from a periodic effort into an operational capability. It gives test teams the confidence to use realistic data, gives AI teams the confidence to build on enterprise information, and gives compliance leaders the evidence they need to demonstrate control.

Discovery is the first control. It is also the one that determines whether every other control works.

Find Out Where Sensitive Production Data Is Hiding in Your Test and AI Estate

Most organisations underestimate how far production data has spread, and how much of it sits in fields, files and pipelines nobody has classified. A Data Discovery engagement with Enov8 scans your priority systems and AI data sources, identifies sensitive production patterns, and maps them to the environments where they create real exposure. You finish with a clear, evidenced view of where to act first.

Talk to Enov8 about a Data Discovery: enov8.com | enquiries@enov8.com