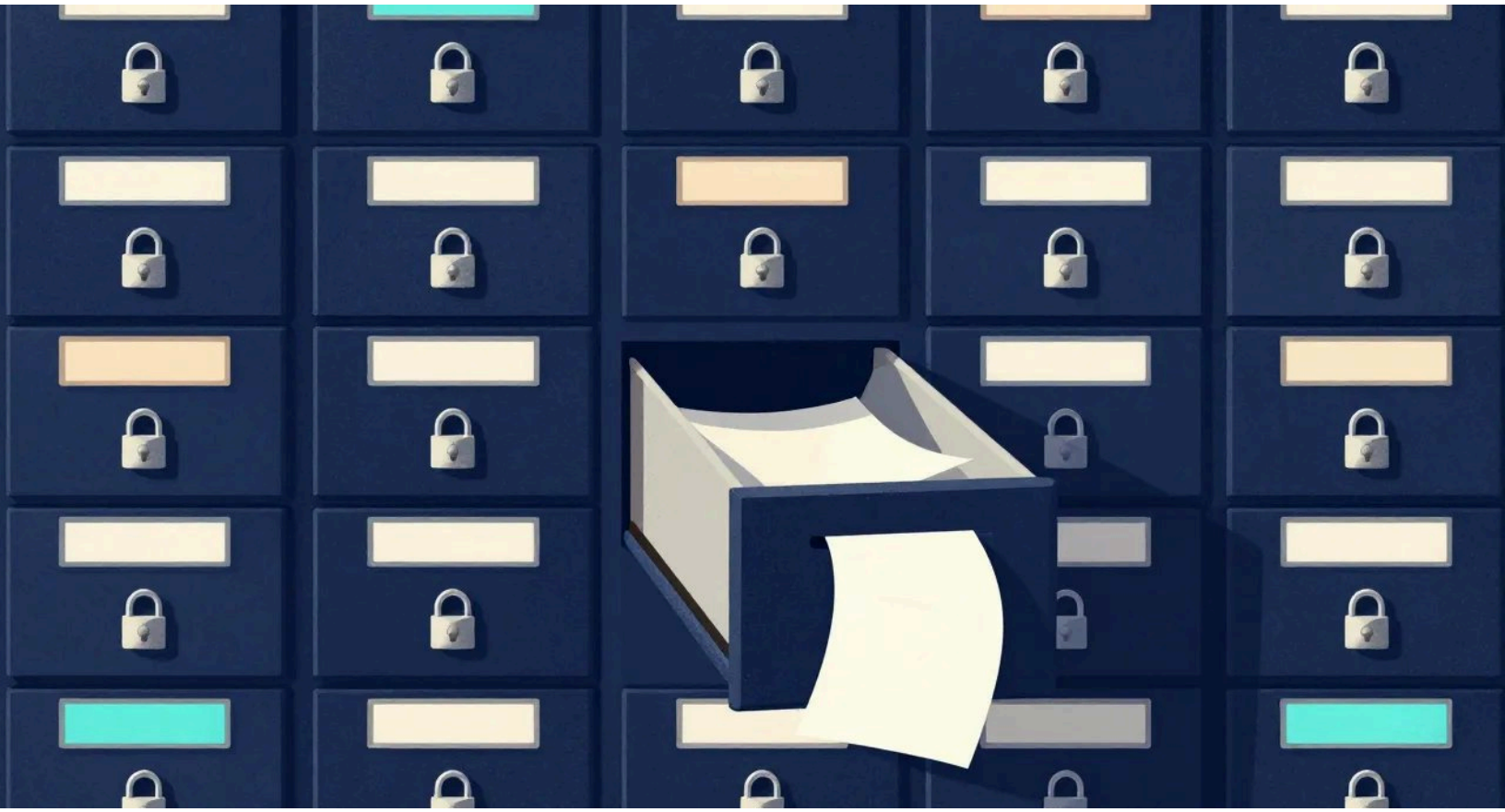


EXECUTIVE PAPER · DATA

Data Classification Is a Breach-Cost Decision

The size of your next breach bill is being set now, by data nobody has labeled and records nobody deleted. CEOs, CFOs, general counsel and CISOs should treat classification as a decision about that bill.

AUTHOR

Chris Watkins

SERVICE

Data

FOR

CEO, CFO, General Counsel, CISO · 12 min read

AT A GLANCE

The problem. A breach is priced by what the attacker could reach. Customer records the business stopped needing years ago, copies in analytics buckets and test databases, and exports sitting in shared folders all count. IBM's 2024 study found shadow data in about a third of breaches, and those breaches cost 16% more and took longer to contain.

Why it persists. Keeping data is free on the day you keep it. Deleting it takes a decision from someone who owns the risk, and in most companies nobody holds that job. Classification projects stall as labeling exercises that never connect to retention or access.

The recommendation. Put every data store into one of four sensitivity tiers. Use automated scanning to find personal and regulated data where it should not be. Then make the tier decide how long the data lives and who can read it, with deletion that runs on a schedule. Report the count of sensitive records held, and the share past their retention date, to the executive team every quarter.

The Decision in Front of You

Every company holds personal data it no longer needs. The decision is whether to keep paying to hold it, in storage and in breach exposure, or to fund the work that finds it, labels it and deletes it on a schedule.

The stakes are set by arithmetic regulators already use. Notification duties attach to each affected person. Fines under the EU General Data Protection Regulation (GDPR) for breaching its core principles reach €20 million or 4% of worldwide annual turnover, whichever is higher [7]. In IBM's 2025 study of 600 breaches, 32% of organizations paid a regulatory fine, and nearly half of those fines exceeded $100,000 [3].

Classification sounds like an IT hygiene task. In practice it decides three numbers in the incident you have not had yet: how many people you must notify, which laws apply to them, and how long it takes to work out what was taken.

What's Happening

Breach costs are high, and unmanaged data raises them

IBM's 2026 Cost of a Data Breach study covered 602 breaches across 17 industries between March 2025 and February 2026. The global average cost reached $4.99 million and the US average $11.5 million [4]. Breaches involving data stored across multiple locations took the longest to identify and contain, at 256 days [4].

IBM's 2024 study measured shadow data directly, meaning data held outside the systems the security team manages. It appeared in 35% of breaches. Those breaches cost 16% more, took 26% longer to identify and 20% longer to contain [1]. The study covered 604 organizations [1].

The type of data matters too. Coverage of IBM's 2025 study puts the average cost of a compromised customer personal data record at $160, and reports that more than half of all breaches involved customer personal data [2]. That coverage comes from a company that sells data discovery software, so treat it as a reported figure and check it against the full report before you budget on it.

The rules already assume you know what you hold

The US federal government has required sensitivity categorization since 2004. FIPS 199 rates each type of information as low, moderate or high impact for confidentiality, integrity and availability, and a system inherits the highest rating of any information it holds [5]. NIST SP 800-60 maps information types to those levels. NIST's 2024 working draft of SP 800-60 Revision 2 proposes adding privacy considerations to categorization and folding in the agency's guidance on protecting personal data [6]. FIPS 199 binds federal agencies. Its method of rating the information first and letting the system inherit the rating works in any company.

GDPR makes the retention side explicit. Article 5 requires personal data to be limited to what is necessary for its purpose, and kept in identifiable form for no longer than necessary [8]. The UK regulator's guidance expects organizations to set retention periods, justify them and review what they hold [8]. Both principles sit in the top fine tier [7].

In the United States, every state has a breach notification law; South Dakota and Alabama became the 49th and 50th in 2018 [9]. The laws differ on deadlines, covered data and who must be told [9]. A customer database with residents of all 50 states triggers all 50 laws, each with its own definition of personal information. Public companies also face the SEC's 2023 rule, which requires a Form 8-K within four business days of deciding that a cybersecurity incident is material, describing its nature, scope and likely impact [10]. Describing scope in four business days is hard when nobody knows which systems held which data.

Recent cases priced unknown and over-retained data

National Public Data, 2024. This background-check data broker was breached and a 277.1 GB file was posted for sale, described as holding about 2.9 billion records with names, addresses and Social Security numbers [11]. The company first said 1.3 million people were affected. Its October 2024 bankruptcy filing said hundreds of millions were potentially affected, that it was likely liable under state laws to notify and pay for credit monitoring for them, and that its insurer had declined coverage [11]. The business reported about $1.15 million in revenue in its last financial year and estimated assets of $25,000 to $75,000. It faced more than a dozen class actions [11].

Blackbaud, FTC order, 2024. Blackbaud provides software to nonprofits, schools and health organizations. In a 2020 ransomware attack, an intruder went undetected for about three months and took unencrypted personal data, including Social Security and bank account numbers, affecting millions of consumers [12]. The FTC alleged that Blackbaud held data far longer than its purpose required, including data about former customers [12]. The 2024 order requires it to delete personal data it no longer needs, publish and follow a retention schedule stating why it keeps each category and when it will delete it, and delete backup files holding customer data it is not retaining to serve customers [12] [13].

23andMe, ICO fine, 2025. Attackers used reused passwords from other breaches to log in to 23andMe accounts between April and September 2023. The UK Information Commissioner's Office fined the company £2.31 million in June 2025 over the exposure of 155,592 UK residents' data, which included health reports, race, ethnicity and family trees [14]. The ICO cited weak authentication, missing verification before raw genetic data could be accessed or downloaded, and slow response [14]. The same login weakness on a shopping site would have exposed order history. Here it reached health and ancestry data, the tier that carries the heaviest regulatory weight.

At the largest scale, Change Healthcare's 2024 ransomware attack ended with notices covering about 192.7 million individuals, and the company said it could not fully deduplicate the count [15].

Why the Usual Approach Fails

Classification programs usually start with a policy document that defines four or five labels. Staff are asked to mark documents. A year later, labels appear on some email and some files, and none of them changes what happens to the data.

The common mistake is treating the label as the goal. A label has value only when it changes behavior: who can read the store, whether it can be copied into a test environment, and when it is deleted. Without that binding, labeling is clerical work.

The second mistake is relying on people to find sensitive data. Engineers copy production tables into analytics platforms. Support staff export customer lists to spreadsheets. Integrations land files in cloud buckets nobody owns. Manual inventories describe the systems people remember, and IBM found shadow data in about a third of breaches [1].

The third mistake is leaving retention to legal and storage to engineering. Legal writes a schedule. Engineering runs systems with no delete function. Nobody checks whether the schedule runs. Blackbaud's order required deleting backups as well as live data [13], and backups are often where schedules go unenforced.

The Framework: Find, Tier, Bind, Shrink

The model has four steps, run continuously.

Find. Scan every data store, including cloud storage, databases, file shares, SaaS exports and backups, for personal and regulated data. Automated tools do this at scale. Amazon Macie, for example, samples representative objects across all storage buckets daily, detects data such as card numbers, passport numbers and credentials with built-in identifiers, accepts custom identifiers for employee or account numbers, and scores each bucket's sensitivity [16]. Microsoft's auto-labeling uses predefined sensitive information types and trainable classifiers to apply labels to documents and email [17]. Expect false positives and tune them. Broad, imperfect coverage of every store is worth more than a precise inventory of a few.

Tier. Give every store one of four tiers. Like FIPS 199, a store takes the highest tier of anything in it [5].

Bind. Attach controls to the tier through configuration that enforces them. The tier sets who can read the store, whether data can leave it, how it is encrypted and how long records live.

Shrink. Delete on schedule, including backups and copies. Mask or replace personal data in test and analytics environments. Measure the count of sensitive records held, and the share past their retention date, and report both.

Exhibit 1. Four sensitivity tiers: the tier sets retention and access, and a store takes its highest tier

Tier	Typical contents	How it is found	Default retention	Access and handling	Breach consequence
Regulated	Social Security, bank and card numbers, health, genetic, biometric data	Automated scanning with built-in identifiers [16] [17]	Shortest period the governing law or contract allows; deletion verified	Named individuals only; no copies to test or analytics; encrypted; access logged and reviewed	Notification in every affected state [9]; top-tier GDPR fines [7]; possible 8-K [10]
Personal	Names, contact details, account and order history	Scanning plus custom identifiers for internal IDs [16]	Business need stated per category, then delete or anonymize [8] [12]	Role-based; masked outside production	Notification depends on state definitions and data combinations [9]
Confidential	Contracts, pricing, source code, financial plans	Owner declaration plus classifiers [17]	Set by legal and business owner	Need-to-know groups	Contract, competitive and disclosure exposure
Public	Published material, marketing	Owner declaration	No limit	Open	Minimal

Source: author's framework, adapted from FIPS 199 high-water-mark categorization [5], GDPR storage limitation [8] and the FTC Blackbaud order [12][13].

What over-retention costs in one breach

The following is an illustrative model for a hypothetical online retailer. Substitute your own figures.

Exhibit 2. Illustrative notification cost of keeping former-customer records

Assumption or result	Keep everything	Retention schedule enforced
Active customer records (purchase in last 3 years)	1,500,000	1,500,000
Former customer records still held	4,500,000	0
People to notify after a full database breach	6,000,000	1,500,000
Direct notification and monitoring cost at $5 per person	$30.0 million	$7.5 million
Difference		$22.5 million
Share of exposure from records with no current use	75%	0%

Source: author's illustrative estimate. Assumes a breach of the full customer database, $5 per notified person for mailing, call center and credit monitoring offers, and no other costs. Legal fees, regulatory penalties, litigation and lost business are excluded and would scale in the same direction. IBM's per-record averages [2] are far higher and do not scale linearly to large breaches, so they are not used here.

The model is deliberately narrow. It counts only the bill that follows directly from the number of people in the file, which is the number retention controls.

Costs, Risks, and Tradeoffs

Exhibit 3. Tradeoffs of classification bound to retention and access

Area	Cost or risk	Mitigation
Tooling	Discovery tools bill by data scanned; full scans of large estates are expensive	Sample, as Macie does [16]; scan new and changed data first; exclude logs
False positives	Scanners flag test data and public numbers as sensitive, and staff learn to ignore alerts	Allow lists for known exceptions [16]; owner review of each new high-tier finding
Lost analytics value	Deleting old customer records ends some long-range analysis	Keep aggregates and properly anonymized datasets; delete the identifiers
Legal holds	Deleting data under litigation hold is spoliation	Legal hold overrides the schedule, tracked by system and case
Engineering effort	Many systems have no reliable delete function, and backups are hardest	Fix the top five stores by record count first; set backup expiry to match the tier
False confidence	Labels with no enforcement look like control on an audit slide	Report records past retention date, measured from the systems themselves
When this is wrong	A small business with one SaaS platform and no regulated data gains little from scanning infrastructure	Use the vendor's built-in retention settings and an annual manual review

Source: author's analysis.

Two costs deserve candor. Deleting data is irreversible, and the first deletion run will alarm someone who wanted that data. And retention periods set by law sometimes require keeping records longer than the business wants; the schedule must follow the stricter rule.

How to Act

Exhibit 4. A one-year roadmap

Window	Actions	Owner	Milestone
First 30 days	List every data store with customer or employee personal data, including SaaS, backups and analytics. Name a business owner for each. Approve the four tiers.	CISO with General Counsel	Store inventory with an owner for every entry; tier definitions signed
First 30 days	Count personal records held and the share belonging to people with no relationship in the last three years.	Data or engineering lead	First count reported to CEO and CFO
First 90 days	Run automated discovery across cloud storage, databases and file shares. Tier every store by its highest finding.	CISO	Every store tiered; shadow copies of regulated data listed with owners
First 90 days	Write the retention schedule by data category with a stated purpose and deletion period. Set legal hold handling.	General Counsel	Schedule approved and published internally
First 90 days	Remove regulated data from test and analytics environments, or mask it.	Engineering leads	Zero unmasked regulated records outside production
First 365 days	Automate deletion in the five largest stores, including backup expiry. Restrict regulated-tier access to named individuals with logged review.	CTO with CISO	Records past retention under 5% in those stores; quarterly access review completed
First 365 days	Report sensitive record count, records past retention and new shadow findings quarterly. Rehearse the scope statement for an 8-K or state notice from the inventory.	CFO with CISO	Quarterly report delivered; breach scope estimated within 48 hours in a tabletop

Source: author's recommendation.

Questions to Ask Your Team

1. How many people's personal records do we hold today, and how many of them have had no relationship with us in three years?
2. Which systems hold Social Security, bank, health or biometric data, and how do we know the list is complete?
3. Where are the copies: test databases, analytics platforms, exports, backups?

4. When did we last delete customer data because a retention period ended, and from which systems, including backups?
5. If we were breached tomorrow, how long would it take to say which states' residents were affected?
6. Who owns the decision to delete, and who can stop it?
7. What does our scanner flag that nobody has reviewed?

Closing

A breach costs what the attacker could reach. Every record past its purpose and every unlabeled copy adds people to the notice list, jurisdictions to the response and days to the investigation. Classify every store, let the tier set retention and access, delete on schedule, and report the count of sensitive records to the executive team as a liability that rises or falls by decision.

Sources

[1] Infosecurity Magazine. "Cost of a Data Breach Surges 10% on Shadow Data Challenge." Coverage of IBM Cost of a Data Breach Report 2024. 2024. https://www.infosecurity-magazine.com/news/cost-data-breach-10-shadow-data/.

[2] BigID. "2025 Data Breach Cost." Coverage of IBM Cost of a Data Breach Report 2025; BigID sells data discovery software. 2025. https://bigid.com/blog/2025-data-breach-cost/.

[3] Baker Donelson. "Ten Key Insights from IBM's Cost of a Data Breach Report 2025." 2025. https://www.bakerdonelson.com/ten-key-insights-from-ibms-cost-of-a-data-breach-report-2025.

[4] Baker Donelson. "Ten Takeaways from IBM's 2026 Cost of a Data Breach Report." 2026. https://www.bakerdonelson.com/ten-takeaways-from-ibms-2026-cost-of-a-data-breach-report.

[5] National Institute of Standards and Technology. "Standards for Security Categorization of Federal Information and Information Systems." FIPS PUB 199. 2004. https://nvlpubs.nist.gov/nistpubs/fips/nist.fips.199.pdf.

[6] National Institute of Standards and Technology. "Guide for Mapping Types of Information and Systems to Security Categories." SP 800-60 Rev. 2, Initial Working Draft. 2024. https://csrc.nist.gov/pubs/sp/800/60/r2/iwd.

[7] European Union. Regulation (EU) 2016/679 (General Data Protection Regulation), Article 83, "General conditions for imposing administrative fines." 2016. https://gdpr-text.com/de/read/article-83.

[8] Information Commissioner's Office. "Principle (e): Storage limitation." UK GDPR guidance. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/storage-limitation/.

[9] Debevoise & Plimpton. "Breach Notification Hits the Big 5-0: Now the Law in All U.S. States." 2018. https://www.debevoise.com/insights/publications/2018/04/breach-notification-hits-the-big-50.

[10] WilmerHale. "SEC Adopts Cybersecurity Disclosure Rules." 2023. https://www.wilmerhale.com/en/insights/blogs/keeping-current-disclosure-and-governance-developments/20230728-sec-adopts-cybersecurity-disclosure-rules.

[11] The Register. "National Public Data files for bankruptcy, admits hundreds of millions potentially affected." 2024. https://www.theregister.com/2024/10/09/national_public_data_bankrupt/.

[12] Federal Trade Commission. "FTC Order Will Require Blackbaud to Delete Unnecessary Data, Boost Safeguards to Settle Charges its Lax Security Practices Led to Data Breach." Press release. 2024. https://search.ftc.gov/news-events/news/press-releases/2024/02/ftc-order-will-require-blackbaud-delete-unnecessary-data-boost-safeguards-settle-charges-its-lax.

[13] Federal Trade Commission. "In the Matter of Blackbaud, Inc." Decision and Order. 2024. https://www.ftc.gov/system/files/ftc_gov/pdf/Blackbaud-D&O.pdf.

[14] Information Commissioner's Office. "ICO fines 23andMe £2.31 million for failing to protect UK users' genetic data." Press release. 2025. https://ico-newsroom.prgloo.com/news/ico-fines-23andme-gbp-2-31-million-for-failing-to-protect-uk-users-genetic-data.

[15] CyberInsider. "Change Healthcare Revises Breach Impact to 192.7 Million Individuals." 2025. https://cyberinsider.com/change-healthcare-revises-breach-impact-to-192-7-million-individuals/.

[16] Amazon Web Services. "How automated sensitive data discovery works." Amazon Macie User Guide. https://docs.aws.amazon.com/macie/latest/user/discovery-asdd-how-it-works.html.

[17] Microsoft. "Auto-labeling overview." Microsoft Purview guidance for the Australian Protective Security Policy Framework. https://learn.microsoft.com/en-us/compliance/anz/pspf-auto-labeling-overview.

The views in this paper are the author's own and do not represent any employer. Scenarios are hypothetical. This paper is not legal, financial, or compliance advice.