LESSON 1.10: DATA INCIDENTS: WHAT WENT WRONG IN REAL FIRMS

LESSON 1.10: SUPPLEMENTAL READING

A persistent misconception about data incidents in regulated finance is that they stem from incorrect or fraudulent information, such as bad data entering systems, vendors providing erroneous feeds, or deliberate manipulation of records. While such incidents occur, they represent a minority of regulatory findings. Far more commonly, data incidents involve accurate data that was used inappropriately, reused without review, or interpreted outside its original context.

The data itself may be sound. The systems may function correctly. The calculations may be accurate. Yet governance breaks down because:

- Data collected for one purpose is repurposed informally for another without reassessing permissions, assumptions, or suitability
- **Derived outputs** (scores, rankings, classifications) are treated as objective facts rather than interpretive constructs that embed assumptions
- Internal analyses migrate into client-facing materials without adequate review or validation
- Transformations accumulate over time, with each step adding meaning that is never documented or supervised

In each case, the problem isn't the data's factual accuracy but how meaning is accumulated as the data moves through the organization. Regulators focus on these governance breakdowns because they reveal structural weaknesses that affect not just a single incident but entire classes of data use.

This lesson examines common patterns in real data incidents, explains why issues are often discovered late, and clarifies what regulators expect firms to learn from failures.

Common Patterns Regulators Identify

While data incidents vary in their specifics (different firms, different datasets, different contexts), regulators frequently observe similar structural weaknesses that transcend individual cases. These patterns aren't unique to sophisticated analytics or AI; they appear in routine workflows, manual processes, and simple analyses. Understanding these patterns helps firms recognize vulnerabilities before they result in incidents.

Pattern 1: Informal Reuse Without Reassessment

Data is collected or licensed for a specific, documented purpose, such as regulatory reporting, fraud monitoring, operational efficiency, or client servicing. Over time, teams

discover that the same data could support other uses: marketing segmentation, risk scoring, product recommendations, or performance benchmarking.

Because the data already exists in accessible systems and because extracting it is technically straightforward, teams **repurpose** it informally, without reassessing whether:

- The original permissions allow the new use
- The data is appropriate for the new context
- Assumptions embedded in the original collection remain valid
- Supervision or approval is required

This pattern is insidious because each individual reuse decision may seem reasonable in isolation. The data exists. The team needs insights. The analysis appears harmless. But collectively, these informal reuses create **scope creep**, where data originally constrained to narrow purposes becomes used broadly across the organization without corresponding governance expansion.

Real-world example: A firm collects transaction data to satisfy anti-money laundering (AML) monitoring requirements. Analysts later use the same data to develop client engagement scores for marketing purposes. The data was accurate and lawfully held, but its use for marketing was never reviewed for compliance with privacy obligations, consent frameworks, or purpose limitation principles. When discovered during an exam, the firm could not demonstrate that the reuse had been approved or supervised.

Regulators consider this example a governance failure; they care less about intent and more about controls, documentation, and discipline.

Pattern 2: Documentation Gaps That Obscure Interpretive Decisions

Material decisions about data (how to categorize clients, which thresholds to apply, what transformations to perform, how to handle missing values) are made verbally, in emails, or in undocumented discussions. The resulting outputs are preserved and used, but the rationale for interpretive choices is lost.

When questions arise during exams, audits, or incident investigations, firms can't reconstruct **why** decisions were made. They can show what was done, but not what alternatives were considered, what assumptions were applied, or who validated appropriateness.

This gap becomes especially problematic when data or outputs are reused over time. Teams inherit datasets, scores, or classifications without understanding the interpretive frameworks that created them. They treat these artifacts as objective inputs rather than constructs that embed judgment.

Real-world example: A firm develops an internal client priority ranking based on account size, transaction volume, and engagement metrics. The ranking is used for years to allocate advisor time and resources. When an examiner asks why certain clients were systematically deprioritized, the firm can't explain which factors were weighted, who decided on the weights, or whether the ranking was ever validated for fairness or bias. The original developers have left the firm, and no documentation exists beyond the spreadsheet formula itself.

Pattern 3: Diffuse Accountability Across Teams and Systems

Data flows through multiple teams—operations, analytics, compliance, marketing, risk—with each team assuming someone else is responsible for validating appropriateness, supervising use, or documenting decisions. No single role or function can clearly explain or defend how data was ultimately used.

This diffusion of accountability isn't malicious. It arises naturally in complex organizations where data is shared across functions, each with its own priorities and expertise. Operations focuses on accuracy and availability. Analytics focuses on insight generation. Compliance focuses on regulatory obligations. Marketing focuses on effectiveness.

But when no one owns end-to-end accountability for how data moves from collection through transformation to application, supervision becomes fragmented. Each team reviews its own piece of the workflow, but no one validates the entire chain. Assumptions made upstream are invisible downstream. Constraints documented in one system are unknown in another.

Real-world example: A firm acquires alternative data from a vendor under a license that restricts use to internal research. The data is ingested into a central repository managed by the data team, which assumes licensing compliance was validated by procurement. Analysts access the repository and combine the vendor data with other sources to create client propensity scores, assuming that if data is in the repository, it's available for any internal use. Marketing uses the scores to target campaigns, assuming analytics validated appropriateness. When the vendor discovers the use and objects, no single team can explain who was responsible for ensuring the use complied with licensing restrictions.

The three patterns tend to coexist rather than occur in isolation. Informal reuse creates documentation gaps because undocumented decisions accumulate. Documentation gaps enable diffuse accountability because no one has a complete record to own. Diffuse accountability enables further informal reuse because oversight is fragmented. The result

is a data environment where governance is reactive, accountability is unclear, and incidents are discovered only after consequences have occurred.

Why Issues Are Discovered Late

Many data incidents are identified only after outputs reach clients, markets, or regulators. Firms may have controls in place (review processes, approval workflows, compliance checks), but those controls are often concentrated downstream, focused on final communications, published reports, or client-facing products.

When governance emphasizes endpoints while leaving upstream data decisions unsupervised, several dynamics contribute to late discovery:

Earlier decisions escape review: Data selection, transformation, and interpretive choices occur without formal oversight. By the time downstream controls are applied, those earlier decisions are treated as settled, and reviewers focus on presentation rather than substance.

Reuse accelerates invisibly: Data is repurposed across teams and contexts without triggering review checkpoints. Each reuse feels incremental and low-risk, so no single instance prompts scrutiny. The cumulative effect is only visible in hindsight.

Assumptions become embedded: Interpretive frameworks about how categories are defined, what thresholds apply, and what data represents are established informally and then persist. Over time, they're treated as facts rather than judgments, and no one questions whether they remain appropriate.

Documentation gaps prevent early detection: Without records of upstream decisions, firms cannot proactively identify where data may have been used inappropriately. Issues surface only when external parties (clients, regulators, vendors) raise questions.

Regulators view late discovery as evidence that governance was reactive rather than preventive. Controls that catch problems only after impact has occurred are less effective than controls that prevent inappropriate use before it happens. Firms that consistently discover incidents late signal that their governance frameworks aren't aligned with how data actually moves through the organization.

Why This Matters Before Analytics or AI

As was noted in earlier lessons, analytics platforms, machine learning models, and AI-driven systems amplify the consequences of weak governance. When similar incidents occur in automated or scaled environments, their impact is broader, faster, and harder to contain.

A manual process that inappropriately reuses data might affect dozens of decisions. An automated system that inherits the same flaw can execute thousands or millions of decisions before the problem is detected. A documentation gap that made it difficult to explain a single analysis becomes a fundamental barrier to explaining an entire model.

Regulators increasingly expect firms to demonstrate that they have addressed historical data governance weaknesses before deploying advanced tools. Incidents that occurred in manual or simple workflows are viewed not as exceptions but as signals that governance is insufficient to support automation at scale.

Questions regulators ask include:

- Have you identified and corrected the governance gaps that led to past incidents?
- Can you demonstrate that similar issues will not recur in automated systems?
- What controls prevent your AI systems from reusing data inappropriately, embedding undocumented assumptions, or operating without clear accountability?

Firms that can't answer these questions face heightened scrutiny when deploying analytics or AI. Learning from past failures isn't just good practice; it's a **prerequisite for responsible automation**.

Conclusion

Data incidents rarely stem from malicious intent or technical failure. They arise from reasonable decisions made without sufficient governance, documentation, or supervision. Regulators study these incidents to understand where control frameworks failed and to assess whether firms learn and improve systematically.

Common patterns—informal reuse, documentation gaps, diffuse accountability—tend to coexist and reinforce each other, creating environments where governance is reactive, and incidents are discovered late. Firms that treat incidents as isolated errors face repeat findings. Firms that treat them as governance signals and respond with systematic improvements are viewed more favorably.

Incidents in manual workflows are warnings, not exceptions. Learning from past failures is the foundation for deploying advanced AI and automation tools responsibly and at scale.