LESSON 1.3: DATA LINEAGE, PROVENANCE, AND WHY REGULATORS CARE

LESSON 1.3: SUPPLEMENTAL READING

When regulatory examiners review how a firm uses data, their questions rarely begin with model outputs, algorithmic logic, or final decisions. Instead, they start upstream: Where did this data come from? How did it move through your organization? What changed along the way? Who approved these transformations? What permissions governed its use?

These questions reflect a fundamental regulatory principle**: outcomes are only as defensible as the process that produced them**. A technically flawless analysis built on data of unclear origin, undocumented transformation, or ambiguous permission isn't compliant; it's a risk waiting to materialize.

Data lineage and provenance provide the narrative structure that allows firms to answer these questions with confidence. They aren't administrative artifacts created to satisfy data engineers or compliance checklists. Rather, they're governance records that trace accountability, demonstrate control, and enable reconstruction of how information became decisions, communications, or actions affecting clients and markets.

Without that narrative, even correct analysis can become a compliance problem. With it, firms can defend their processes under scrutiny, respond to incidents with precision, and build systems that scale responsibly.

Data Lineage: The Map of Movement and Transformation

Data lineage describes the complete path that information takes from its original source through the systems, processes, and transformations that shape its eventual use. It's a record of movement and change—a map showing where data entered the organization, how it was processed or enriched, which systems touched it, and where responsibility rested at each stage.

A robust lineage record answers questions such as:

- What was the original **source** of this data field?
- Which systems or **processes** extracted, copied, or transmitted it?
- What **transformations** were applied—cleaning, aggregation, enrichment, scoring?
- When did these transformations occur, and under what **logic** or business rules?
- Who **approved** or validated each step?
- What **controls** or validations are applied at each stage?
- Where did the data ultimately land, and for what **purpose**?

Lineage isn't a static diagram created once and filed away. It must reflect the actual history of data movement, including ad hoc processes, manual interventions, and iterative changes that occur over time. When lineage documentation diverges from

reality (or when it describes an idealized process rather than what actually happened) it loses its value as a governance tool.

The test of effective lineage is simple: **Can the firm reconstruct exactly how a specific output was produced at a particular point in time**? If the answer is no, the lineage is incomplete, and accountability is compromised.

Provenance: Origin, Authority, and Legitimacy of Use

While lineage tracks movement, **provenance establishes origin and legitimacy**. Provenance answers foundational questions about where data came from, under what authority it was obtained, what permissions govern its use, and for what original purpose it was collected.

These distinctions matter because regulatory permission is often tied, not to technical access, but to **intended use**. Data that is lawfully collected for one purpose—account servicing, fraud prevention, regulatory reporting—may be restricted from use in another context, such as marketing, credit decisioning, or risk modeling. Even when a firm has physical access to data in its systems, it may not have the legal or ethical right to apply that data to every conceivable use case.

Provenance records provide the basis for determining whether a given use remains consistent with:

Contractual constraints: Licensing agreements with data vendors often specify permitted and prohibited uses. Provenance documentation must reflect these limitations.

Regulatory permissions: Client data collected under one regulatory framework (such as account opening requirements) may not be freely repurposed under another (such as consumer credit reporting).

Consent and disclosure: When clients provide information, they do so with certain expectations about how it will be used. Provenance records help ensure that subsequent uses align with those expectations and any associated consents.

Ethical boundaries: Even when data use is technically permitted, provenance documentation forces firms to articulate why a particular application is appropriate—not just legal.

When provenance is unclear or undocumented, firms may inadvertently use data in ways that exceed their rights. While the data itself may appear innocuous(think standard account information, publicly available signals, vendor-provided feeds) the **context** of its collection determines what can be done with it.

Why Lineage and Provenance Must Work Together

Lineage and provenance are complementary, but neither is sufficient on its own.

A firm may have perfect lineage—a detailed record of every system, transformation, and process that touched a dataset—while still lacking a defensible justification for using that data in the first place. If the original collection lacked proper consent, if the licensing agreement prohibited the intended use, or if the data was repurposed beyond its original scope, the lineage becomes a map of compliance failure rather than a defense.

Conversely, a firm may have clear provenance (legitimate acquisition, proper permissions, documented purpose) but lack visibility into how that data was subsequently transformed, combined, or reused. When derived outputs produce unexpected or problematic results, the firm can't trace back to understand what went wrong, where judgment was introduced, or who was accountable.

Regulators expect both dimensions to be addressed. Together, they allow examiners to understand not only what happened to the data, but whether it should have happened at all. They enable firms to demonstrate control over data use, not just data storage.

Regulatory Focus: What Examiners Look For

From a regulatory perspective, lineage and provenance serve several critical functions that go beyond technical documentation:

Audit replay and reconstruction: Regulators must be able to reconstruct how a specific decision, communication, or model output was generated at a given moment in time. This is especially important when investigating complaints, adverse outcomes, or potential violations. If a firm cannot show exactly what data was used, how it was transformed, and who approved its application, the reconstruction becomes impossible.

Accountability and oversight: Lineage and provenance clarify where interpretation occurred and who was responsible for validating it. When data moves through multiple systems and teams, responsibility can become diffused. Effective documentation assigns clear ownership at each stage, ensuring that someone is accountable for every transformation, derivation, or reuse decision.

Risk concentration and control gaps: Gaps in lineage or unclear provenance are rarely isolated incidents. They tend to emerge in environments where data reuse accelerates faster than governance, where informal practices bypass documented processes, or where assumptions go unreviewed. These gaps are early warning signs of broader control weaknesses. Examiners look for patterns: Are lineage gaps concentrated in certain business lines? Do they involve specific types of data? Are they more common when third-party vendors are involved?

Permissions and compliance verification: Provenance documentation allows examiners to verify that data use complies with contractual, regulatory, and ethical

constraints. If a firm claims it has the right to use certain data, provenance records provide the evidence. If those records are missing or inconsistent, examiners may conclude that the firm does not have adequate control over its data practices.

Conclusion

Data lineage and provenance are not compliance formalities. They're the foundation of accountability in data-driven organizations. They provide the narrative that allows firms to explain how information became decisions, to defend those decisions under scrutiny, and to detect and correct errors before they propagate.

Regulators care about lineage and provenance because they understand that control over data is control over outcomes. Firms that can't trace where data came from, how it moved, and whether its use was appropriate have lost control.

Before introducing analytics, automation, or AI, firms must establish the lineage and provenance practices that ensure every use of data can be explained, every transformation can be defended, and every decision can be reconstructed. This is not a barrier to innovation. It is the precondition for building systems that scale safely and responsibly.