LESSON 1.2: DATA TYPES & RISK

LESSON 1.2: SUPPLEMENTAL READING

Financial firms generate, acquire, and transform vast amounts of information daily. Yet not all of that information carries the same regulatory weight or introduces the same level of risk. Two datasets may appear functionally similar—both feeding into dashboards, models, or operational systems—but they can trigger vastly different supervisory expectations, documentation requirements, and accountability obligations.

These distinctions don't arise from analytical complexity or technological sophistication. They arise from what the data represents, how it was created, and how far it has traveled from its original source. Understanding these differences isn't an abstract exercise in taxonomy. It's a prerequisite for defensible governance, especially as firms introduce automation, advanced analytics, and AI-driven decision systems.

Common Data: Familiar, But Not Risk-Free

Common data refers to information that is widely used, well understood, and historically accepted within regulated financial activities. It includes market prices, security identifiers, issuer fundamentals, economic indicators, credit bureau reports, and standardized client account information. This data is typically sourced from established providers—vendors with regulatory oversight, transparent methodologies, and industry recognition—or from internal systems designed specifically for compliance and regulatory reporting.

Because common data is familiar, firms often treat it as inherently low-risk. This assumption is only partially accurate. Common data still carries obligations around accuracy, timeliness, and appropriate use. A market price sourced from a reputable vendor doesn't become exempt from validation simply because it's widely available. If that price is stale, incorrect, or applied to an unsuitable context, the firm remains accountable.

What distinguishes common data from other categories isn't the absence of risk, but the legibility of that risk. Regulatory expectations are well established. Documentation standards are clear. Supervisory frameworks have matured over decades. When common data is disclosed in client-facing materials or regulatory filings, its provenance is generally understood and accepted.

The governance lesson here is straightforward: common data doesn't require less oversight—it requires clearer oversight. Firms know what controls are expected, what disclosures are standard, and what documentation satisfies examiners. Discipline, not ambiguity, is often the challenge.

Alternative Data: Context, Permission, and Embedded Assumptions

Think of **alternative data** as information that falls outside traditional financial datasets but is used to generate insights, signals, or classifications. This includes transactional metadata, geolocation patterns, web browsing activity, satellite imagery, sentiment indicators derived from social media or news sources, and aggregated behavioral signals.

The defining feature of alternative data isn't novelty. Many alternative datasets have existed for years. What distinguishes them is contextual ambiguity, which is the lack of established frameworks for understanding what they represent, how they should be governed, and what constraints apply to their use.

Alternative data introduces questions that regulators prioritize:

- Why does this data exist? Was it created for the purpose the firm is now using it for, or has it been repurposed? If repurposed, what assumptions underlie that new use?
- How was it collected? What consents, permissions, or licensing agreements govern its acquisition? If the data was collected without the explicit knowledge of the individuals it represents, what privacy or fairness concerns arise?
- What interpretive choices are embedded? Even aggregated or anonymized alternative data reflects decisions about what to include, how to categorize, and what patterns to emphasize. Those decisions can introduce bias that is difficult to detect or explain.
- What happens when context changes? Alternative data that was appropriate for one use case may become unsuitable when applied to another. For instance, foot traffic data used to assess retail performance may not be appropriate for evaluating creditworthiness, even if statistical correlations exist.

The regulatory challenge with alternative data is that its risks are often less obvious than those associated with common data. There are fewer established norms, less regulatory guidance, and more room for misinterpretation. This makes alternative data harder to supervise—and therefore more likely to draw scrutiny when questions arise.

Importantly, the moment alternative data is used to inform decisions, segmentation, prioritization, or client recommendations, it becomes a governance issue. This is true even if the data never appears in client-facing communications. The interpretive framework that determines how alternative data is applied must be documented, defensible, and subject to oversight.

Derived Data: Where Judgment Becomes Encoded

Derived data is created when existing data—whether common, alternative, or other **derived data**—is transformed, combined, summarized, scored, labeled, or inferred. Examples include client rankings, risk scores, propensity models, segmentation schemes, summary statistics, and machine learning model outputs.

Derived data doesn't exist independently. It's the product of assumptions, logic, and objectives embedded in its creation. Every step in the derivation process—choosing which inputs to include, defining how they are weighted, selecting what thresholds apply—reflects interpretive judgment. That judgment may be encoded in algorithms, business rules, or statistical models, but it remains judgment nonetheless.

From a regulatory perspective, **derived data is often where risk concentrates**. Even when the underlying inputs are well understood, permitted, and appropriate, the act of derivation introduces:

- **New meaning**: A credit score is not simply a summary of payment history. It represents a judgment about future behavior, risk, and suitability.
- **New potential for error**: Each transformation step—cleaning, aggregation, normalization, scoring—creates opportunities for mistakes, drift, or unintended bias.
- **New documentation obligations**: Derived data must be explainable. Firms must be able to articulate not just what the output shows, but why the derivation process produces that result and whether it remains fit for purpose over time.
- **New supervisory responsibility**: Someone must own the derivation logic. Someone must validate that it behaves as intended. Someone must monitor for changes in performance or outcomes.

The critical insight is that derived data may appear objective or technical while still encoding subjective judgment. For example, a ranking algorithm may seem purely mathematical, but the choice of which variables to include, how to weight them, and what threshold determines "high" versus "low" priority are all interpretive decisions. Those decisions must be traceable, defensible, and subject to review.

The Risk Gradient: Why Governance Obligations Escalate

As data moves from common to alternative to derived forms, three dynamics unfold simultaneously, each amplifying regulatory risk:

First, interpretive distance increases. The further data moves from its original source, the harder it becomes to explain what it truly represents. A stock price is straightforward. A sentiment score derived from news articles aggregated across multiple sources and weighted by proprietary algorithms isn't. If an examiner asks, "What does this number mean?" the firm must be able to answer not just technically, but conceptually.

Second, accountability becomes diffused. With common data, responsibility is often clear: a vendor provides it, a team validates it, a process governs its use. With derived data, accountability can fragment across data engineers, model developers, business owners, compliance reviewers, and third-party service providers. When something goes

wrong, it may not be obvious who was responsible for the interpretive choices that led to the error.

Third, documentation gaps emerge. Transformations that feel routine—cleaning, joining datasets, recalculating scores—are often poorly recorded, even though they materially affect outcomes. A firm may document its model methodology in detail but fail to document the preprocessing steps that shaped the training data. Those undocumented steps become vulnerabilities during exams or audits.

This escalation explains why regulators focus not only on data sources, but on how data is transformed and reused over time. They understand that the most consequential risks arise not from exotic models or novel datasets, but from poorly governed derivation processes applied to familiar information.

An Illustrative Progression: From Common to Derived

Consider a firm that begins with common data: standard client account records and transaction histories used for regulatory reporting and operational dashboards. This data is well understood, sourced from internal systems, and governed by established controls.

The firm then decides to enhance its understanding of client behavior by incorporating alternative data: external signals such as web browsing activity, app usage patterns, or aggregated purchase behaviors from third-party vendors. These signals aren't inherently problematic, but they introduce questions: Why was this data collected? What permissions govern its use? What assumptions underlie its relevance?

Finally, the firm combines these inputs to create derived data: client propensity scores that rank clients by likelihood to engage with certain products, risk classifications that inform communication strategies, or segmentation schemes that determine eligibility for specific offers. These outputs don't simply summarize the underlying data; they interpret it. They **embed judgments** about what patterns matter, which clients are similar, and what behaviors predict future outcomes.

At each step, the same underlying account information remains present. But with each transformation, interpretive distance grows, accountability becomes less obvious, and documentation requirements escalate. What began as straightforward reporting of data has become a derived asset that drives decisions affecting client treatment. The governance framework must reflect that evolution.

Why This Matters Before Introducing Analytics or AI

Analytics platforms and **AI systems accelerate the creation and use of derived data.** They enable **transformations at scale**, often with minimal human review. This amplification is valuable—but only if the underlying governance is sound.

If a firm does not clearly understand which data is common, which is alternative, and which is derived, it cannot:

- Reliably assess risk or prioritize oversight efforts
- Assign accountability for interpretive decisions
- Respond to regulatory questions about data provenance and transformation
- Detect when derived data drifts from its original purpose or introduces unintended bias

Most AI-related incidents and regulatory findings trace back not to exotic models or cutting-edge techniques, but to poorly governed derived data. The model may function correctly, but the data it was trained on (or the way that data was transformed before training) was flawed, biased, or inadequately documented.

This lesson isn't about restricting innovation or discouraging the use of alternative or derived data. It's about recognizing **where meaning changes** and ensuring that governance keeps pace. Data classification is not a compliance formality. It's a map that shows where interpretive risk accumulates and where supervisory attention must focus.

Conclusion

Not all data carries the same regulatory weight. The distinctions between common, alternative, and derived data aren't only academic. They determine what controls are needed, what documentation is required, and where accountability must be assigned.

As firms adopt automation and AI, the volume and complexity of derived data will only increase. Building governance structures that recognize these distinctions now and before introducing advanced analytics creates a stable foundation for responsible innovation. It ensures that when regulators ask where data came from, how it was transformed, and who approved its use, the firm has clear, defensible answers.

Understanding data types and their associated risks is not a barrier to progress. It is the precondition for building systems that scale safely.