AI Can’t Fix Bad Data: Why Semantic Technologies Are Key to R&D Acceleration

Contributed Commentary by Verena Mertes and Wolf Lichtenstein, LabVantage Solutions

December 30, 2024 | What role will AI play in the future of drug development, especially as novel therapeutic approaches such as precision medicine drive up the complexity and cost of research? Some industry headlines hype AI as a magic bullet, capable of cutting through the headwinds of slow development cycles, shifting regulations, and high rates of failure. For companies racing to reach the market with safe and compliant new drug products, this vision of AI is compelling—and it’s driving heavy investment.

Here’s the thing, though. When it comes to accelerating research, running an AI algorithm is the relatively easy part. Collecting, cleaning, and managing the data feeding that algorithm—that’s the heavy lift. Fail to get that right, and it’s “garbage in, garbage out.” But feeding AI with meaningful data is difficult, especially when that data is unstructured, unorganized, and scattered across different locations.

Knowledge Graphs, Ontologies, and the Pathway to Faster Innovation

For your AI solution to play a reliable role in your research pathway, you need to complement it with semantic technologies like ontologies and knowledge graphs. Working together in a platform-style ecosystem, these technologies will establish a strong upstream data framework to support meaningful, high-quality results from your downstream AI algorithms.

What makes ontologies and knowledge graphs so powerful? They are foundational data management strategies designed to organize and contextualize diverse datasets, like those commonly used in life science research. Ontologies standardize and define the relationships between concepts such as “genes” or “disease pathways,” while knowledge graphs leverage those ontologies to reveal the relationships and connections between disparate data points.

Together, these strategies are critical enablers of the FAIR principles, ensuring that data is findable, accessible, interoperable, and reusable. In the specific context of life science research, they offer four benefits that work hand-in-hand with AI-driven R&D.

1. **Contextualization**

   A common scenario: Your research team has access to a massive data lake, but they’re struggling to use it in a meaningful way.

   That’s likely because the data you’ve captured lacks context—a crucial element, especially in applications like drug discovery where the biological, clinical, and regulatory context can significantly impact outcomes.

   Knowledge graphs and ontologies are the solution. These strategies identify the relationships between disparate data points inside the data lake, and they frame those relationships within a relevant context. In this way, the data lake becomes a life science data lakehouse, where information is structured, organized, and available for meaningful use. From that framework, you will gain a clear and contextualized understanding of the calculations and predictions emerging from your AI model. When researching complex concepts like gene-protein interactions, disease pathways, or regulatory requirements, this level of contextualization is critical.

2. **Interoperability**

   A common scenario: The data you need is distributed across many unconnected systems, making it difficult to use in a meaningful way.

   To find novel associations or predict accurate outcomes, AI models may rely on data from a variety of sources, including R&D datasets, clinical trial results, public scientific literature, and other siloed locations. Collecting this data is already difficult; translating it into an integrated and unified format capable of driving meaningful research is even more challenging.

   This challenge is not well-suited to AI models alone, which struggle to connect pools of standalone data. That’s where semantic interoperability plays a key role. Enabled by knowledge graphs, semantic interoperability is all about building meaningful interconnections between otherwise isolated data points. An AI solution can leverage this semantically integrated data to extract knowledge and rapidly generate research insights, such as the relationship between a drug response and a particular genetic marker.

3. **Explainability**

   A common scenario: Your AI model generates a game-changing research insight, but there's a problem: you can't explain the biological cause-and-effect behind that insight, which makes it impossible to trust from a scientific and a regulatory perspective.

   The solution to this "black box problem" is a data framework that supports transparent and explainable AI outcomes. Semantic technologies are the building blocks of this framework. When an AI model identifies a potential drug target, for example, a knowledge graph will visualize the relationships and hierarchies that contributed to that identification process, helping researchers and regulators map the route from data to discovery.

   The semantic integration that supports robust explainability can also empower your AI solution to ask 'smarter' questions. For example, rather than simply predicting the efficacy of a drug, your AI solution could leverage the biological relationships encoded in a knowledge graph to explore causal pathways or suggest alternative compounds.

4. **Reusability**

   A common scenario: Your team is using valuable resources to regenerate existing datasets that are hard to find and even harder to repurpose.

   Redundant research efforts can drive up lab costs, particularly when it comes to recreating comprehensive, high-quality, and ethically sensitive experiments. A structured knowledge graph can solve this problem by recovering relevant data that was previously locked inside manually impenetrable datasets. By surfacing that data and making it available for meaningful AI analysis, knowledge graphs can help significantly reduce repetitive experiments and their associated costs while accelerating novel research outcomes.

The bottom line: For high-quality AI insights, establish a high-quality data management strategy. How can we position life science researchers to better address the critical needs of patients around the world—and to do so with greater speed and precision?

The emergence of AI provides a partial solution, but it requires ontologies, knowledge graphs and other semantic technologies to operate at its full potential. With these data management strategies working upstream to organize and connect disparate datasets, and a robust AI solution working downstream to generate novel research insights from that data, you have a complete solution—and a gateway to faster, more streamlined R&D research.

***Wolf Lichtenstein**, President at LabVantage Europe, is an expert in digital transformation, AI, and business growth. Based in Berlin, he has spent decades leading and advising performance-driven organizations such*

as McKinsey & Company, C3.ai, and Webtrekk GmbH. Wolf holds a master's in psychology from Universität Augsburg and has attended the World Economic Forum five times. He can be reached at wlichtenstein@labvantage.com.

***Verena Mertes, PhD** in Technical Biochemistry, serves as Director of Project Management at Labvantage Biomax. With over a decade of experience in life sciences digitalization, she has spearheaded numerous projects in semantic data integration. She also leads BioTech360, driving innovation in life sciences R&D with Labvantage Biomax's semantic integration platform. She can be reached at verena.mertes@labvantage-biomax.com.*