



**data  
iku**

EBOOK

# **AI & Advanced Data Analytics in Life Sciences**

---

A Use Case Library



# Accelerating the Transformation of Pharma With Data



Technology shifts tied with digitalization in recent years are driving unparalleled transformation in the healthcare and life sciences industry. More than any other industry, we are seeing leaders in biotech and pharma investing in AI, including generative, to improve strategic business objective outcomes. Moreover, we see these transformations not only refining their key business objectives but redefining new business models, such as investments in digital health. And there is more to it than just AI. There is collective ambition and imperative to adopt streamlined data-driven approaches that include broad data democratization to enable self-service analytics, particularly in established global organizations that often carry burdens of legacy systems, redundancy and historical silos in people, processes and data.



New digital-native players, biotech start-ups, medtech, and technology services companies are also seeking to create an ingrained strategy around data literacy that allows them to optimize their processes with minimal resources, particularly for companies working in specialty therapeutics or with novel biotech advancements such as cell-based gene therapies or mRNA.

In this library, you will find a representative sample of use cases which is bound to expand and grow as novel solutions to common problems in the industry are found. From optimal reach and engagement strategies with healthcare professionals and patients, to speeding discovery and development of new therapeutics, there are dozens of ways that companies can harness AI, data, and analytics to drive value at scale — all while maximizing corporate efficiency and patient care.

In the following pages, we'll run through many of the solutions and applications that have helped life sciences companies get ahead in their data game, from transversal and corporate data strategies, to data visibility, and accessibility, and reuse, to highly specific domain areas across the pharmaceutical value chain including:

- Accelerating drug discovery, development and pipeline diversity
- Optimizing clinical operations to bring new therapies to market
- Implementing digital manufacturing strategies and resilient supply chains
- Improving market access, commercialization, market engagement
- Deepening patient-centric engagement and insights to improve care outcomes
- Medical affairs and regulatory compliance
- Human resources, financial operations, and corporate functions

Fully integrated and centralized data practices that support an upskilled and data-savvy workforce across all of an organization's departments are the future of data science for healthcare. And as large language models (LLMs) become more sophisticated and more deeply embedded in AI models and processes, this future will arrive at an increasingly accelerated pace.

Dataiku as a modern orchestration platform breaks down the silos both within and across myriad departments while balancing transparency with security, providing instead a central, governed hub from which teams can access and collaborate. Organizations trust Dataiku to tackle some of the hardest challenges the industry faces including data access and integration, regulatory oversight, enabling and accelerating scarce data experts while upskilling domain experts and analysts, and the ethical implications of adopting intelligent automation and AI to speed the delivery of life-saving or life-changing therapies to the world.

# Table of Contents



Click any heading to navigate directly to the section.

## ■ Accelerating Discovery and Development

Drug Repurposing Knowledge Graph

Novel Drug Target Identification

Scientific Publication Relevance

Molecular Property Prediction From Chemical Structures

Optimal Medical Device Utility

Scientific Research Imaging Classification

## ● Optimizing Clinical Research and Operations

Predict Clinical Site Risks and Impacts With NLP

Predicting the Need for an IDMC in Clinical Studies

RWE Patient Insights

## ■ Digital Manufacturing and Resilient Supply Chains

Predictive Process Modeling

Realtime Product Defect Detection on the Production Line

Improving Manufacturing Processes With Predictive Maintenance

Improving Data Collation Processes

Inventory Visibility

Self-healing Supply Chains

AI-Driven Drug Stocking and Transportation

## ● **Market Launch and Commercialization**

Optimizing Omnichannel Marketing in Pharma  
Physician Scoring for Adopting New Prescriptions  
AI for Personalized Marketing Segmentation  
Optimizing Content in Email Marketing Campaigns  
Enabling 360° Healthcare Professional Customer-Centricity  
Next Best Action (NBA) Recommendation Engine  
Automated Reports for Sampling Operations

## ■ **Patient Engagement and Medical Affairs**

Social Determinants of Health  
Early Treatment Identification and Patient Compliance  
Bias Detection and Mitigation  
Improve Patient Behavior Insights With ML  
Balancing Insights and Privacy Protection With LLMs  
Pharmacovigilance Safety Analytics and Signal Detection  
Generate Targeted and Actionable Internal Medical Insights

## ● **Human Resources**

Management Effectiveness Analysis  
Operationalize AI for Workplace Analytics

## ■ **Putting Data in Motion at Scale in the Pharma industry**

Streamlining Analytics and Machine Learning  
Enterprise-Level Data Democratization  
Developing Widespread AI Literacy  
Leveraging Generative AI Across the Value Chain



# **Accelerating Discovery and Development**

# Drug Repurposing Graph

The decline in R&D returns faced by the Biopharmaceutical industry due to protracted development timelines, stringent regulations, and complex patient and disease landscape, combined with the rising uncertainty in pharmaceutical R&D have acted as a catalyst for pursuing label extensions and product repurposing for new indications.

Using AI and data analytics on a multitude of data sources spanning chemical structures, molecular targets, pathways, and patient reported outcomes, businesses can explore potential indications, patient populations, and even new lines of therapies for their approved products. This can significantly shorten development times, reduce costs (by over 80%), reduce regulatory uncertainty (by 150% compared with a novel drug), expedite market entry, and tap into new revenue sources.

But this is far from being a simple exercise, with no clear starting point and the need for thorough analysis to identify hidden patterns in massive amounts of data from disparate sources. Leveraging graph analytics approaches provides a powerful accelerator to simplifying complex data structure representations and facilitating the understanding of intricate relationships between pharmaceuticals, targets, side effects, genes, diseases, and more, acting as a perfect starting point for research questions.

Dataiku's drug repurposing knowledge graph solution includes the following highlights:

- Automate ingestion of key public data sources on drugs, diseases, and genes via FTP and Postgres from NCBI, DrugCentral, and related gene ontologies and pathway databases.
- Clean and prepare input data leveraging visual recipes and python code to extract graph nodes and relationships for drugs, diseases, symptoms, genes, pathways, and more.
- Quickly develop your capacity to deep dive into complex relationships between drugs with a biomedical knowledge graph and push the graph to your choice of graph technologies such as neo4J in a matter of a few clicks.
- Integrate further data, visual graph exploration, and graph-based analytics to discover novel insights for drug repurposing. Create scenarios for new data ingestion, and run Cypher queries to boost relationship understanding through graph analytics.

[Learn More](#) ↗

# Novel Drug Target Identification

Pharma companies are accelerating early discovery and research by applying natural language processing (NLP) methods to mine the massive stores of patent applications related to biological processes, as well as other sources (gene ontologies, proteins, disease terminology, etc.). This allows for the identification, classification, and ranking of patents based on the relevance and novelty of a therapeutic target (e.g. the gene or protein) to inform computational chemist teams that fuels the discovery and direction of new therapies. Applying generative AI techniques further deepens the insights this use case can deliver, where either foundational large language models or LLMs tailored to the complexity of biology and scientific research can be trained to extract and structure the key components of the patents that align to relevance and novelty as well as generate key patent summaries to further inform the scientists working on new therapeutic targets.

R&D teams harnessing such novel drug target identification capabilities will experience increased efficiency, by reducing the number of patents to review for novelty, and improved drug discovery, by exploring and filtering historical patent and target data with multiple criteria (e.g., strength of evidence, target and disease, and target and mechanism-of-action associations) to speed research and development of new therapies aligned to their pipeline development.

[Learn More ↗](#)



# Scientific Publication Relevance

With the vast amount of scientific content published on a daily basis, the attempt at keeping up with industry trends let alone extracting or generating tangible insights and benefits can be extremely time consuming and costly. Nevertheless, it's a necessary process for researchers and leaders both for healthcare providers and life sciences companies, whether it is to support R&D in target identification, regulatory submissions for post marketing surveillance, market access, HEOR, or simply to stay up to date on new therapeutics and findings.

With advanced analytics, data teams can use NLP to predict article relevance based on abstract and other metadata. Articles can be scored for relevance relative to one another, and this process can be automated, allowing business leaders to consistently focus their attention on the most urgent articles in the most efficient manner. Further applications of large language models (LLMs) can further summarize relevant findings for broader insights and consumption to drive new research objectives.

[Learn More ↗](#)

# Molecular Structure Property Prediction

Lead and target selection are by far the biggest challenges in accelerating pharmaceutical research & development, with the abundance of targets and the vast libraries of molecules to select from, the process becomes costly, inefficient, and uncertain. Leveraging advanced analytics and machine learning, biotech and pharma companies can predict molecules' characteristics, including their affinity to targets of interest, as well as their chemical properties including absorption, distribution, metabolism, excretion, and toxicity, using their chemical structure (molecular fingerprints). In addition to selecting the best molecule for specific targets, computational chemists and biologists can speed and prioritize experimentation by predicting whether certain molecules in an organization's library could have the necessary affinity to said targets.

Dataiku allows you to create API connections to common publicly curated chemical databases as well as ingest proprietary and public datasets to build, test, and deploy statistical and ML models to streamline and expedite your lead and target selection. More recently, these traditional molecular structure prediction problems are also being flipped by Generative AI applications to change the paradigm to AI-generated potential novel compound (molecule, protein, etc.) designs based on a desired property instead of testing and scoring either existing or proposed molecular structures.

[Learn More ↗](#)

# Optimal Medical Device Utility

For medical device manufacturers, it's essential to improve patient outcomes, device utilization, and device performance. A leading medical equipment supplier wanted to improve the utility of image-guided instruments employed by surgeons during medical procedures. Their primary objective was to meticulously analyze the data generated by these advanced instruments, ultimately aiming to optimize their functionality, guarantee their correct deployment, and facilitate improved patient outcomes during surgery.

To accomplish this, the team leveraged advanced data analysis techniques and cutting-edge technology. They collected and scrutinized vast datasets obtained from these instruments, spanning multiple data points extrapolated from videos of image-guided device usage, to uncover patterns that could shed light on optimal usage techniques and inform tailored device settings to improve utility.

The ultimate goal of this data-driven endeavor was twofold. Firstly, they sought to identify best practices and optimal techniques for utilizing these image-guided surgical instruments, which could significantly enhance the precision and efficacy of surgical procedures. Secondly, by understanding the ideal device settings and usage protocols, they aimed to provide valuable and precise product recommendations to surgeons, thus improving patient outcomes and empowering them to make more informed decisions during surgery.

[Learn More ↗](#)

# Scientific Research Imaging Classification

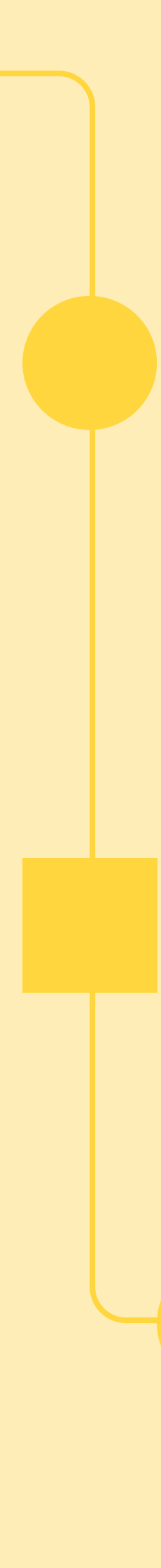
Biotech companies are innovating by creating robust high-throughput processes for the detection of subvisible particles (SVPs) in injectable formulations to prevent contaminants (both synthetic and biological) affecting both development operations and product safety. This ensures that stringent regulatory quality inspection criteria are consistently met throughout formulation development and manufacturing phases. It also prevents potential patient adverse immunogenicity reactions in product usage.

The primary objective of this endeavor revolves around improving the ability and speed to process and utilize high resolution microscopy images. This process is essential as it forms the foundation for the development of advanced deep learning models intended for automation in anomalous SVP limit detection in quality inspection.

The value added by this research is multifold. First and foremost, it significantly strengthens the client's quality assurance efforts, reducing the risk associated with producing medicines with subvisible contaminants that could lead to patient safety or efficacy issues. Furthermore, the efficiency gains from this initiative are noteworthy, translating into a faster time to deliver data products and realize their inherent value. This streamlined approach also reduces operational risk, resulting in less waste – both in terms of formulation and materials – when producing injectable products.

Research in scientific imaging not only addresses a critical need for regulatory compliance, but also promises substantial benefits to our clients. By harnessing the power of advanced imaging and machine learning with a platform like Dataiku, biotech organizations can enhance the quality, efficiency, and safety of their injectable products, ultimately reinforcing their commitment to delivering excellence in the biotechnology sector.

[Learn More ↗](#)

A decorative graphic on the left side of the page consists of a vertical yellow line. At the top, it curves to the right and then down to a yellow circle. Below the circle, it continues down to a yellow square. From the bottom of the square, it curves to the right and then down to a yellow circle. This circle is connected to another yellow circle by a horizontal line. From the second circle, a horizontal line goes to a yellow square. From the bottom of this square, a vertical line goes down, then curves to the right and then up, ending at the top right of the page.

# Optimizing Clinical Research and Operations

# Predict Clinical Site Risks and Impacts With NLP

A leading US based biopharmaceutical company with a significant global footprint in terms of clinical research, manufacturing, and market was only one of many businesses struggling to predict risks to their global operations, risks impacting their entire value chain with delays, shortages, and business disruptions.

The business decided to use Dataiku to enable the Global Security Operation Center (GSOC) to create custom risk identification models using NLP and data analytics combined with automated flows and triggers to analyze global news cycles as well as other data sources to issue early alerts through emails, flagging events and trends such as natural disasters and geopolitical developments that had the potential to impact supply chains, disrupt clinical operations, and propagate upstream throughout their business, all in time to enable sites and partners to adjust and adapt to minimize impact and mitigate the risks.

The goal of this solution was to supercharge the GSOC team's efficiency in detecting, planning, and communicating proper responses to these potentially disruptive events by generating and extracting actionable insights with natural language processing and understanding data and trend analytics.

[Learn More ↗](#)

# Predicting the Need for an IDMC in Clinical Studies

With global pharmaceutical companies engaging in dozens of clinical trials simultaneously across both internal and outsourced R&D initiatives, there are multiple responsible bodies for keeping track of 100s of thousands of generated documents, one of which is the Independent Data Monitoring Committee (IDMC) which is responsible for assessing the progress, safety data and, if needed critical efficacy endpoints of the study. Despite this pivotal role, these committees are not always timely formed, which poses the challenge of predicting when they are needed, and initiating the process of forming them in a timely fashion to avoid any unnecessary delays to the trial.

Roche, a global pharmaceutical and diagnostics company, addressed the challenge of establishing the need for independent data monitoring committees (IDMC) by developing a predictive model to determine the necessity for an IDMC, streamlining the process, and ensuring timely committee establishment.

The solution created with Dataiku ingests CTMS and other trial data, while approximating missing data using Natural Language Processing (NLP) techniques to extract relevant information from protocols, alongside potentially incorporating external data sources such as clinicaltrials.gov and health authorities' data. Although challenges related to missing data persist, this comprehensive approach enhances the model's accuracy and reliability.

The model was able to generate accurate predictions for a trial's need for an IDMC, enabling the trial teams to initiate the process of forming an IDMC on time, reducing delays and improving the team's efficiency.

[Learn More ↗](#)

## RWE Patient Insights

One of the most common burgeoning use cases for efficient data operations and predictive analytics in healthcare and life sciences is leveraging the power of massive stores of real world data sources (electronic health records, claims, patient registries, etc.), or RWD, to uncover deeper hidden patient insights that can inform healthcare and research & development decisions, ensure optimal and inclusive clinical trials designs, augment clinical research and outcomes evidence generation, improve market access and reimbursement, and inform deeper patient outcome insights that fuel the market launch and post-market safety of new therapeutics.

One real world evidence (RWE) oncology team wanted to provide its scientists with an efficient research, visualization, and analysis tool for insights on the outcomes of patients. Many of their existing RWE projects used the same raw data sources (EHRs/Claims/Registries) and variables, but there was no reusability, collaboration, or consistency in data pipelines for analysis. Leveraging data ops, collaboration, and extensibility with Dataiku, the team was able to solve this problem to build reusable common data pipelines fueling various analysis requests. They benefited from an interactive dashboard that allows them to define the relevant patient cohorts and run powerful analyses on patient trends, treatment patterns, and longitudinal outcomes.

[Learn More ↗](#)



A decorative yellow line starts at the top left, goes down, then right, then down again, ending at a square. Another yellow line starts from the bottom right, goes up, then left, then down, ending at a square. A horizontal yellow line connects two circles in the bottom middle section.

# Digital Manufacturing and Resilient Supply Chains

# Predictive Process Modeling

In pharmaceutical companies, batch processes form a critical part of the manufacturing value chain where inefficiencies cost billions of dollars each year. At a time when supply chains are unpredictable and stressed by shortages, the need to maximize equipment utilization by reducing downtime and to improve yield by reducing unnecessary waste becomes even more critical.

In biopharmaceutical manufacturing, few things are more disruptive than process interruptions or inefficiencies. When processes fail, everything from overall equipment effectiveness to labor utilization to revenue and profitability can be negatively affected. Process inefficiencies alone can cost billions of dollars. In general pharmaceuticals, for example, the costs are as much as \$50 billion per year; on top of this, around 30 billion drug doses are lost due to underutilization.



Dataiku's batch performance optimization solution provides an adapt-and-apply application to empower data teams to dissect vast volumes of production process data. The solution allows those teams to easily develop actionable insights for technicians, operators, and reliability and process engineers to understand the root cause of failures and to predict process outcomes, thereby accelerating the move from reaction to anticipation in batch manufacturing.

The solution comes with a ready-to-use dashboard that allows you to easily follow your production process, identify the source of process upsets with explainable ML, and better predict the failure possibilities for your next batch.

Here are some of the solution highlights in more detail:

- Easily ingest production process data from Manufacturing Execution Systems (MES), Pi Historians, or other IoT sources and transactional batch data through an intuitive Dataiku App.
- Visualize insights on batch process historical performance, key failure patterns, and their impact on future probability of success.
- Perform root cause analysis by analyzing sensor data per batch and recipe or product.
- For each recipe or product, predict risk of failure via an explainable and transparent machine learning model.
- Deploy a web application to deliver insights and provide decision support to production and operators.

[Learn More ↗](#)

# Real-time Product Defect Detection on the Production Line

In the age of AI, smart manufacturing means sourcing hard, material data from machines and building models in digital space that, in turn, help manufacturers make better use of (and decisions about) those machines. A flagship case in point is defect detection on the production line.

One world-leading medical equipment company learned this first hand. Being one of the world's leading producers of lenses, the company prioritizes maintaining quality above all. Thanks to their lens-making machines, they had the relevant data needed to take the step into AI future: error mapping data and images of the glasses under manufacture. With the help of Dataiku, a labeling plugin was used to successfully detect defects on the glass and modify the machine parameters. The suggested (and improved) parameters allowed the organization to correct for defects and improve product quality — all of which, in turn, increased brand and image popularity, and reduced costs for defective products.

[Learn More ↗](#)

# Improving Manufacturing Processes With Predictive Maintenance

Unplanned maintenance is costly. By some estimates, unplanned downtime can cost plants upwards of \$100 million (depending on size and sector). The desire to chip away at downtime and recapture some of that lost capital is understandable, even crucial. And predictive maintenance, as has been proven time and again this past decade, can certainly help with that.

In its simplest form, predictive maintenance combines data from a myriad of sources (MES, IoT, and so on) and uses machine learning techniques to anticipate equipment failure before it happens, minimizing repair costs, down and response times, and speeding up maintenance. While IoT data plays an important role, other data sources can be included such as external data from APIs (like weather), geographical data, manual data from human inspection, and much more.

Platforms like Dataiku allow pharma companies to bring their manufacturing operations into the AI future by collating their varied manufacturing data into a single place, and by using that data to build machine learning models that can prevent downtime and improve process efficiency. The results — time and labor saved, revenue secured, and peace-of-mind ensured — are too good to pass up.

[Learn More ↗](#)

# Improving Data Collation Processes

Not all data analytics projects need to be complex. In fact, some of the simplest solutions are the most powerful. Many pharma companies share the problem of navigating cumbersome manual manufacturing reporting processes to meet stringent regulatory oversight requirements. At one such company, for example, the Industrial Operations & Product Supply (iOps) team wanted to address the challenge of the manual efforts required to send manufacturing quality control reports to labs.

Working with Dataiku, they solved this problem handily. The objective was to automate the reporting process, improving the ability to access, format, and visualize data in an easy and non-code method. By feeding Excel files from lab machines into their Dataiku instance (with its native connections, scenarios, and built-in metrics and checks), the company was able to develop an automated reporting process with powerful data visualization capabilities. Projects like this help organizations save time and refocus their resources on high-value tasks.

[Learn More](#) ↗

# Inventory Visibility

In healthcare, effective inventory management is a multifaceted challenge that can profoundly influence both costs and the quality of patient care. To address this challenge, data science teams commonly embark on a mission to empower warehouse managers with enhanced visibility into their inventory operations, ultimately aiming to optimize the process in three key areas:

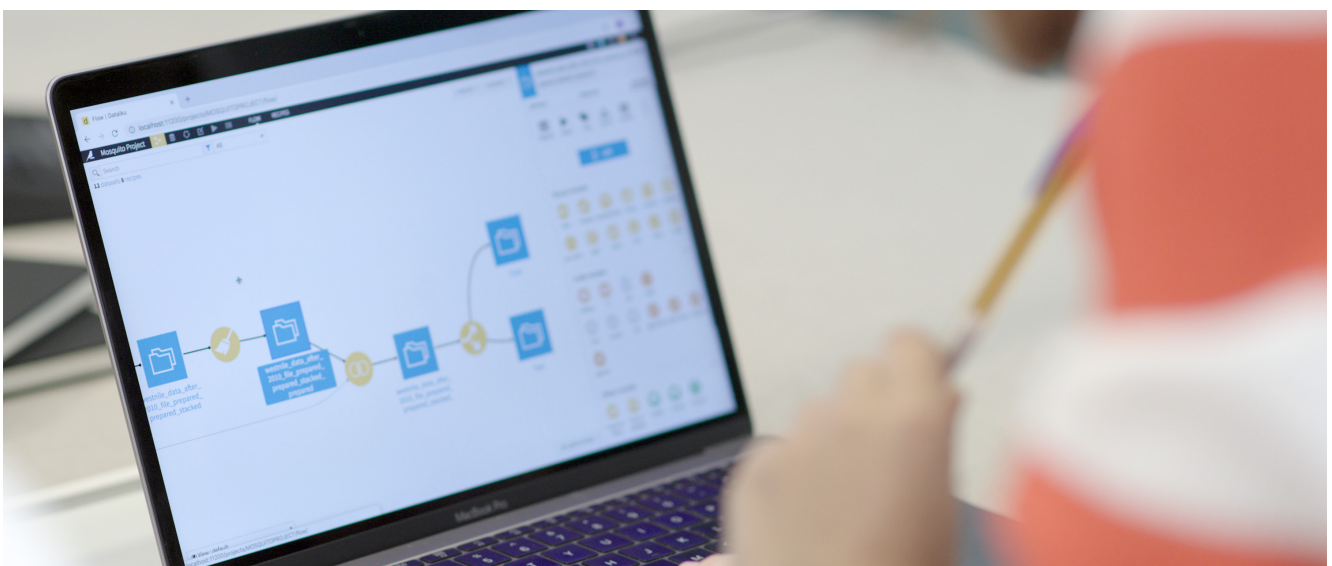
- 1. Prediction of customer orders and consumption:** One of the central facets of this initiative is to develop predictive models that anticipate customer orders and consumption patterns. By analyzing historical data and trends, data teams seek to provide warehouse managers with valuable insights, enabling them to prepare for demand fluctuations more effectively. This not only reduces the risk of stockouts but also ensures that patients receive the essential medical supplies they require in a timely manner, directly contributing to improved patient care and satisfaction.
- 2. Comparison with actual inventory levels:** Accurate real-time inventory tracking (comparing projected with actual inventory levels) is critical to ensuring that resources are allocated efficiently. It's essential to implement robust tracking mechanisms that allow for precise comparisons between predicted and actual inventory levels. This transparency enables proactive decision-making and swift corrective actions when discrepancies arise, saving time and resources while bolstering customer satisfaction.
- 3. Recommendation of supply orders:** To optimize costs without compromising on quality, data teams use platforms like Dataiku to develop algorithms that recommend supply orders based on anticipated demand and current inventory levels. These recommendations strike a balance between avoiding excess inventory, which can be costly, and preventing shortages, which can disrupt patient care. The result is cost savings, streamlined operations, and a reduction in the risks associated with stock outs or expired products.

The value added by this data-driven approach is substantial. It includes heightened customer satisfaction due to the consistent availability of critical medical supplies, significant time savings through automation and proactive decision-making, and cost savings achieved through efficient inventory management practices.

[Learn More](#) ↗

## Self-Healing Supply Chains

With advanced analytics, data teams at HLS companies can facilitate the ranking and intervention of supply chain exceptions that could not be caught with previous automation; rapidly and optimally reduce losses in the supply chain and increase throughput to deliver life saving medication to patients; and upskill & educated the supply chain IT and manufacturing teams on machine learning and AI project scoping and fundamentals.



One leading global pharmaceutical company deployed such a use case and reaped the benefits: \$1.5-3 million savings potential, a model that could be used across their plants globally, and a marketing spend optimization of \$1 million per year.

[Learn More](#) ↗



# AI-Driven Drug Stocking and Transportation

Supply chain leads responsible for distributing new investigational products during clinical research operations commonly struggle with drug inventory management for their global multi-center later phase trials.

There is a fine balance that must be struck between narrow expiration windows for investigational active product ingredient manufacturing, custom packaging for double-blind clinical trials, and site activation, patient enrollment, and visit scheduling to characterize drug supply for clinical site operations. For these reasons, it often results in ordering excess stock, which can expire; cost escalations; drugs sitting idle at clinical sites; mismatches between drug expiration and transit timeline; among other issues.

Life sciences companies can leverage their historical drug usage data, patient visit and dropout data and visit scheduling from IRT systems, depot and site inventory data, and historic transportation route data to develop AI models for forecasting and route optimization.

Dataiku and the services partners who support them have helped organizations do exactly that, resulting in accurate inventory and transportation predictions, system centralization (by way of a handy central dashboard), automated email alerts before drug expiration dates, and optimal route identification for transportation.

[Learn More ↗](#)

A yellow decorative line starts at the top left, goes down to a circle, then right to a square, then down to a circle, then right to a circle, then right to a square, and finally up to the right edge of the page.

# Market Launch and Commercialization

# Optimizing Omnichannel Marketing in Pharma

With global market demand coupled with diminishing returns on development pipeline investments and local market regulations and constraints, pharmaceutical companies depend on strategic marketing campaigns to increase the reach and knowledge of their products and ultimately therapeutic availability.

Effective marketing engagement benefits from an omnichannel approach of both personal and non-personal messaging to inform HCPs in a variety of ways. The challenge for companies is to evolve from the use of multiple channels in a siloed manner to true omnichannel marketing: Targeting prospects/clients (here, HCPs) with the right channel (emailing, phone calls, online ads), with the right content (product vs. informative), and at the right time.

Defining an omnichannel marketing strategy requires pharma organizations to know:

- Who the target audience is (specialty therapeutic areas of HCPs, sensitivity to specific contents or products, associated hospital systems and supported patient populations, etc.)
- What are their ideal engagement channels (preferred communication channel, format of content, best time, etc.)
- What are the KPIs used to measure the success of this strategy and, ultimately, optimized cycles for the sales rep or marketing content engagement

The omnichannel marketing optimization solution for pharma provides a reusable project wireframe to accelerate development of analytics tailored to your data and business structure. It is a Dataiku application that eases development of brand-specific tactics.

With this solution, organizations can:

- Easily integrate their brand sales data, HCP characteristics, and marketing outreach data, and explore.
- Train machine learning models to understand channel engagement and HCP drivers influencing sales deviations and brand activation with a step-through Dataiku application.
- Evaluate marketing campaign effectiveness with business-friendly descriptive analytics dashboards across analysis drivers: brands, drugs, hospitals, individual drug prescribers, etc.
- Score physicians and other individual drug prescribers for brand adoption with explainability via SHAP importance and channel engagement attempt/success summaries.

[Learn More](#) ↗

# Physician Scoring for Adopting New Prescriptions

Pharmaceutical organizations looking to increase brand adoption by healthcare professionals (HCPs) could pursue few better avenues than finding experienced physicians who would be enthusiastic about adopting and prescribing their brands. This is, of course, easier said than done. It takes time and resources to speak with physicians about the benefits and applications of a given drug — meaning the question of whom to approach and with what products is a matter of efficient and selective marketing.

Dataiku has worked with U.S.-based pharmaceutical companies to improve their marketing engagement methods with physicians by developing a physician scoring model. The model is designed to predict whether a new physician will adopt and prescribe a given company's brand. And it works with just a few simple inputs:

- Physician Profiles
- Promotion Data
- Sales
- Patient Safety Reports

An ML model learns from and converts these inputs into predictive analytics that score physicians based on the likelihood of them adopting the brand. This is all then fed into a dashboard with reports and logs providing an easy interface for users, and the ability to train or score models and perform custom calculations.

[Learn More ↗](#)

# AI for Personalized Marketing Segmentation

Healthcare Professionals (HCPs) are the linchpin of the pharmaceutical industry, helping to connect patients with the right drugs and treatment plans. But how can pharma organizations know which HCPs to approach with which products? The glut of data on HCP profiles does not, on its own, help companies decide whom to market to and how. Understandably, pharma companies would benefit from a tool that helps them better profile and segment the HCPs in their marketing ecosystems.

Dataiku has worked with leading global pharmaceutical companies to address the challenge of collecting data from various sources, like prescriber insights, HCP promotions, patient longitudinal data, demographic data, and to build HCP behavior profiles for personalized marketing content. This has historically always been difficult, especially due to the large variety of data sources and data provider dependencies, and the frequent updates to that data.

The use case allows companies to build robust marketing segments and HCP profiling and brings them quicker and continuous updates when data changes for specific time periods. The value added is clear: greater transparency, time savings, end-to-end project deployment, and easy cross-team collaboration.

[Learn More ↗](#)

# Optimizing Content in Email Marketing Campaigns

A major pathway to getting healthcare professional (HCP) buy-in to one's brand or a specific suite of drugs is to engage them via email marketing. Whether you're slotting thought leadership blogs or customer success stories into their inboxes, you can magnify trust in and respect for your brand by demonstrating your influence and presence within the industry.

The challenge is targeting the right professionals with the right content: to whom should you reach out? And what should you send them? Dataiku has worked with leading global pharmaceutical companies to develop inference-based content curation and personalization for HCPs. Specifically, we built a recommendation engine that could predict content and brand preferences for each HCP based on previous purchases.

The resulting recommendation engine plugin has the benefit of being reusable across projects, reproducible, governable, and scalable. And the value add has been tremendous: time saved, improved content performance, increased demand and prescriptions, and eased cross-team collaboration (with flexibility to control and access data by both coders and clickers).

[Learn More ↗](#)

# Enabling 360° Healthcare Professional Customer-Centricity

Customer insights teams at many pharmaceutical organizations — including some of the leading global organizations that Dataiku has worked with — routinely want to get a full understanding of healthcare professionals (HCPs) across all interaction channels and prescription habits. Historically, each interaction channel has usually been measured separately, often pulling data from separate and siloed sources. In other words, cumbersome manual effort has been required to get a 360° analysis on HCPs — a process which is very time intensive and not scalable with channel/interaction volume increase.

Platforms like Dataiku can help organizations develop a complete and automated integration of all the relevant HCP data, packaging everything together into a 360° view of the HCPs that dot the landscape of the pharmaceutical industry. With real-time results displayed across a customized dashboard, customer insights teams can save valuable time, benefit from end-to-end project deployment (allowing for automation, governance, and scale), and optimize HCP interaction and digital planning.

[Learn More ↗](#)

# Next Best Action (NBA) Recommendation Engine

Selecting the optimal approach, content, and the right time to engage healthcare professionals and organizations is at the core of optimizing commercial operations at pharmaceutical companies. From increasing field force efficiency and performance to ensuring HCP engagement satisfaction, the ability to reliably predict said approach and when to engage HCPs and HCOs is critical.

Global pharma organizations are leveraging the Dataiku platform to develop recommendation engines that employ machine learning techniques to predict the next best action within a specific marketing or engagement sequence. This advanced system relies on data inputs from both standard customer interactions and digital influences, allowing it to model and reconstruct the complete sequence of past events, including historical events and customer responses, such as website visits and digital engagements. Data teams utilize Dataiku's powerful "What if analysis" tool to enable a comprehensive evaluation of various action strategies, helping businesses fine-tune their marketing sequences for maximum effectiveness. The solution enhances HCP engagement, loyalty, and campaigns' efficiency, while increasing revenue and achieving significant savings in cost and time, with the potential to achieve better results when using the flow in combination with LLMs to generate personalized engagement scripts and materials and ML models to select the best scientific and marketing materials to achieve the best results.

These organizations further rely on Dataiku to scale and productionalize such recommendation systems with global, reusable data pipelines for dynamic sales engagement strategies that are used as the common backbone for extended tailored local market engagement models, customized to regional market and/or brand constraints. These pipelines often ingest and sift through a massive amount of both traditional and digital data sources to build actionable (and horizontally scalable) recommendation systems.

[Learn More ↗](#)



# Automated Reports for Sampling Operations

One of the simplest and most effective ways that life sciences companies can use their data is by automating the myriad reports their data teams routinely need to produce in order to drive decision making among their peers in the line of business. One American biotech company worked with Dataiku to do just that, specifically hoping to:

- Reduce sample shortages
- Leverage ML to better predict the likelihood for a given prescribing physician to use drugs packaged as samples
- And standardize marketing mix data

Platforms like Dataiku help data teams gather all of their shipment, sample utilization, branded prescription, and sample inventory data in one place, and then transform, calculate, and compile that data into powerful visualizations and automated reports. All of this reduces the possibility of human error and the number of points of error, brings report production time down from hours to minutes, and gives users repurposable resources (such as project templates that can be reused across brands). These insights into physician sample utilization and effectiveness has the added benefit of feeding back into HCP characteristics and profiles that can be leveraged in broader omnichannel marketing and targeting strategies.

[Learn More](#) ↗



# Patient Engagement and Medical Affairs

# Social Determinants of Health

Research tells us that health inequities account for approximately \$320 billion in annual United States healthcare spend and could reach \$1 trillion by 2040 if left unaddressed. Furthermore, social determinants of health (SDoH) are among the most significant health status predictors and may contribute up to 90% of health outcomes variability. Despite that, fewer than 25% of today's hospital care or screening models consider social needs.

Equity-focused approaches should recognize and incorporate how SDoH factors (poverty, social/racial/ethnic discrimination, and housing/transportation conditions) contribute to health and disease status. Multiple stakeholders can use those insights to ensure that health services and therapeutics are accessible, affordable, and culturally competent or appropriate for diverse populations. Particularly for drug manufacturers, these insights can ensure therapeutic access equity is considered both in investigational product research and market access and launch.

Dataiku's SDoH solution focuses on leveraging local area data (U.S. Census and CDC surveys at the county and tract level) with goals to optimize regional health outreach, services, and therapeutic access programs by discovering community-level SDoH impacts on chronic disease prevalence patterns.

This plug-and-play solution helps to address challenges in understanding how social factor vulnerabilities underpin chronic disease prevalence to promote equity through:

- Drug and device manufacturers in effective market access, HCP engagement, risk surveillance, and inclusive clinical trial design to remove social factor bias that could contribute to poor outcomes or adverse reactions
- Hospitals and health services organizations in practice and outreach programs tailored to social needs
- Health insurers in promotions and care coverages to reduce the burden on the U.S. healthcare system from preventable disease

[Learn More](#) ↗

# Early Treatment Identification and Patient Compliance

As we move to increasingly patient-centric care, there is a growing shift from HCP engagement to proactive patient identification and tailored patient care around both life saving and life improving therapies. Pharmaceutical companies that strive to identify patients in need of treatment as early as possible, as well as keep them adherent over the course of treatment, typically achieve this by way of intelligent and strategic targeting and messaging combined with tailored treatment recommendations. But this is easier said than done: with the increasing availability of consumer and patient level datasets, it can be quite challenging to navigate the noise and tap into the full potential of the available information.

Platforms like Dataiku can ingest multiple sources of longitudinal patient-level characteristics and integrate them with population and behavioral data sources to develop a deeper insight into patient needs and behaviors. One such use case with a global pharmaceutical company demonstrated the benefits of significantly reduced time to ingest, explore, and synthesize longitudinal patient data to both identify patients in need of therapies earlier, and based on their behavior characteristics, offer treatment plan recommendations that keep them adherent to medications to improve outcomes. Dataiku users can create easy visual interfaces for business users to interactively explore patient patterns, sub-groups, and behaviors leading to disparate outcomes. And they are empowered with the ability to quickly plug-in new machine learning models without integration costs.

[Learn More ↗](#)

# Bias Detection and Mitigation

It's crucial for many pharmaceutical companies that their data teams are able to identify and reduce bias in their models. The data science team at Pfizer, for instance, starts at the core by asking questions about representation or lack thereof in the data. A critical first step in bias detection is identifying the various factors that lead to unrelated outcomes which stray from original targets. After identification, the next step is to understand the structural or institutional reasons that these phenomena may occur.

For example, you could observe patients missing from a certain protected demographic subgroup in your healthcare data regarding propensity and risk for a disease. This leads to questioning whether they are naturally less likely to acquire a particular disease or if it is that the group has lower access to healthcare leading to selection bias or survivorship bias which ultimately affects how the algorithms learn. Ideally, this bias should be detected early in the exploratory phase to prevent long-term negative impact on patient care models.

The data science team at Pfizer utilizes Dataiku's interactive statistics for exploratory data analysis (EDA) to paint a broad picture of what is going on in the data from beginning to end. The team has also developed Dataiku plugins such as odds ratio visualizations and disparate impact ratio calculators to quantify target outcome differences in the data.

They look at things like odds ratios, which measure the strength of association between two events (i.e., the strength of association between target outcomes and different feature values of their data). Disparate impact ratio compares the relative likelihood of the target outcome occurring across different groups, usually to compare privileged groups versus underprivileged groups. These measurements allow data scientists to understand patterns of target behavior and outcomes long before training a model, and Dataiku's platform makes it easy to explore with the team's chosen plugins.

[Learn More](#) ↗

# Improve Patient Behavior Insights With ML

In any industry, having a lot of data does not necessarily mean that you have a lot of insights. Advanced analytics can help life sciences companies make better and more intelligent use of the data they ingest from customer behavior surveys.

At one leading global pharmaceutical company, the internal commercial and medical research groups wanted to find actionable insights from the surveys commissioned for characterizing patient behavior patterns between vaccination rates across their different vaccine lines. To better understand patient sentiment or acceptance of new vaccines, the team leveraged Dataiku to build a model combining internal data with other publicly available data, and end-user focused visualization to assist in decision making. The project gave the organization a model that includes a weighted decision tree, extracted paths in readable format, and interactive visualizations.

All told, the organization was able to gain unique customer insights, including knowledge about potential customers for a pneumonia vaccination based on past vaccination behaviors that were not uncovered before. And they gave their research analysts, data scientists, and business analysts real experience in using AI and working collaboratively on advanced analytics.

[Learn More ↗](#)

# Balancing Insights and Privacy Protection With LLMs

Pharmaceutical companies are on a quest for insights that can transform patient care and drive innovation in the industry. Data and business teams across the industry seek to harness patient-level data at increasing degrees of granularity to make informed decisions. The insights they're after could range from understanding population health economic outcomes to tailoring individual patient care plans.

But this journey is riddled with complex challenges that demand creative and secure solutions. Where is data stored, and who has access to it? How can it be integrated into models without compromising the privacy of the individuals whose details make up that data?

One approach, pioneered by Sarus — a tech company focused on AI approaches to data privacy — and platforms like Dataiku, involves privacy-preserving data generation: patient-level data is synthetically generated with the help of LLMs and differential privacy theory, all while retaining the key signals required for analysis. (More specifically, it preserves population-level signals while removing individual-level signals.)

This approach seeks to offer actionable insights without exposing PHI. And because differential privacy does not require us to make any judgment on what is identifying and what is not, the process is fully automatic, cutting manual operations and reducing time-to-data.

Once trained, the LLM can be used to generate a new synthetic dataset that preserves the population patterns of the original dataset, and can therefore be used for annotation or machine learning, but with the mathematical guarantee that nothing private will ever come out of the LLM.

The risk that a LLM may output private information from its training dataset may seem abstract and remote, but it is very concrete and real. When prompted with “John Doe suffers,” a LLM trained without differential privacy would produce completions revealing the disease of the actual “John Doe” from the private training dataset: “John Doe suffers from a severe form of pancreatic cancer.”

The generation of privacy-preserving synthetic data therefore carries enormous benefits relative to the de-identification of data. It allows data teams to preserve a much higher level of detail, specificity, and data-richness than simple de-identification. And the differential privacy guarantee means that sensitive PHI is kept under lock.

[Learn More ↗](#)

# Pharmacovigilance Safety Analytics and Signal Detection

The human cost of adverse drug reactions (ADRs) is staggering: ADRs are a leading cause of mortality or morbidity, with an estimated 197,000 deaths annually in Europe. They pose a severe economic expense as well, accounting for at least 5% of all hospitalizations with a global economic cost burden over \$1 trillion. Regulatory mandates requiring drug manufacturers to report potential adverse drug reactions have led to the curation of large public databases, such as the FDA Adverse Event Reporting system (FAERS), that provide a rich source for data mining to discover novel signals.

Dataiku's pharmacovigilance solution — designed with these goals and strategies in mind by industry experts — provides a ready-to-use application to ingest adverse drug reaction safety reports, process and clean the data, and generate safety insights of potential signals that inform drug risk profiles.

The solution enables data managers to seamlessly connect to FAERS and FDA drug data sources and allows drug safety analysts to select the safety report filters and analysis cohort parameters they require. Analyses can be run quickly and dynamically, and the outputs can be consumed via intelligent and easy-to-use data visualizations on the main dashboard.

Let's take a look at some of the Pharmacovigilance Solution highlights in more detail:

- Quickly ingest data files extracted from the FDA Adverse Event Reporting System (FAERS) database.
- Easily process the data, detect duplicate reports, and filter on demographic, drug, reaction, and report characteristics.
- Identify and visualize patterns in safety data with Dataiku's descriptive analytics and charts.
- Calculate common disproportionality metrics for statistical inference and signal detection.
- Achieve immediate insights with a user-friendly Dataiku App to upload new quarterly data files, filter reports, generate cohort signals, and run drug/reaction analytics to ensure patient safety and increase regulatory compliance with early detection of potential ADR signals.
- Adapt and extend to other public (such as Vigibase or Eudravigilance) databases or privately curated drug safety data sources.

[Learn More](#) ↗



# Generate Targeted and Actionable Internal Medical Insights

Having great stores of patient and drug-related data is useless if you don't have the means to leverage, transform, and model that data. Platforms like Dataiku help data teams clean and transform data from free-form texts, medical reports, doctors' notes, hand written notes from the field, and more. With a consolidated workbench for pharmacists, analysts, and data engineers, the platform can provide timely insights using large language models (LLMs) with sentiment analysis and natural language understanding that can drive infectious disease intervention and education.

For Moderna, a leading global pharmaceutical and biotechnology company, this type of use case allowed them to:

- Leverage Generative AI
- Drive improved efficiency within the data team (process time was reduced from 10 hours per month to fraction of it, thanks to automation).
- See a cost savings of about 40 hours per month.
- Reuse models across a variety of natural language processing (NLP) use cases.
- Discover new insights by way of sentiment analysis trends.

[Learn More ↗](#)



# Human Resources

# Management Effectiveness Analysis

Human resources (HR) teams looking to identify which employees are likely to have success as managers and in senior roles can leverage advanced analytics for that purpose. Drawing on performance ratings, 360 degree feedback, organizational health survey data, recognition data, and hiring, terminations, and movement data, platforms like Dataiku can transform these sources of information and test multiple machine learning models for the best predictions and accuracy.

HR departments can visualize manager performance at a glance to help them make decisions that improve manager performance, optimize development investment, and promote greater engagement among managers and teams.

[Learn More](#) ↗

# Operationalize AI for Workplace Analytics

With talent in life sciences scarce and in high demand, C-suite executives at pharmaceutical companies across the globe are interested in understanding and optimizing collaboration and productivity among internal teams. Platforms like Dataiku can help with this, turning existing, unused data into valuable insights for HR departments.

For example, to gain better insights by mining Outlook calendar data, a topic modeling engine can be developed to analyze the subjects of calendar events and employee data. What results, quite simply, are workforce productivity insights, presented to the user in the form of powerful visualizations in the Dataiku instance. With these, HR departments can properly operationalize AI in their day-to-day work and, more importantly, institute new policies for more productive meetings, or for driving collaborative practices and cultural change.

[Learn More ↗](#)



# Putting Data in Motion at Scale in the Pharma Industry

# Streamlining Analytics and Machine Learning

Like in many pharma companies, the business team at Novartis has a weekly task of updating data in Excel to generate important metrics. This process involves repetitive manual calculations of various key performance metrics and decisions are made based on the outcomes.

The team faced some obstacles, such as modifying key parameters in the existing process. Additionally, there was a lack of real-time data refresh and ineffective data tracking, leading to discrepancies due to human error. This also affected the team's ability to identify risks in budget forecasts and field-force allocation.

To solve these issues, Novartis's data engineering and data science teams came together and developed an automated solution using Dataiku. With this solution, the team can now avoid repetitive manual calculations and make more informed decisions based on accurate and real-time data. Key performance indicator (KPI) reports are refreshed automatically, and pre-built templates are used for visualization and reporting on the custom-built forecast models.

[Learn More ↗](#)

# Enterprise-Level Data Democratization

At Merck, a market leader in biopharmaceutical research, manufacturing, and supply, collective efforts from various departments, regions, and teams are pivotal for maintaining the quality and accuracy of data projects.

While digitalization and continuous medical advances have helped the team meet higher expectations and needs, it has also generated increasingly large data stores. Democratization of this data at an enterprise level was identified as a transformative business solution that would break down data silos, promote further collaboration, and empower employees with data-driven insights. But going “enterprise” with data democratization would make the challenge multifold due to the associated cross functional demands, policies, and scaling efforts.

Merck explored various product options, and most of them posed yet another barrier of “stiff learning curve.” Irrespective of user personas, many products required technical integrations and coding language capabilities. So the critical business challenge Merck contemplated was how to enable people, processes, and technology at scale and speed to achieve their enterprise-level data democratization and advanced analytics goals.

Leveraging Dataiku met the four key components of success for these goals: data accessibility, self-service (no/low code) analytics, governance, and enablement. The value of this new enterprise data strategy is reflected in hyper productivity gains, enhanced decision-making, increased collaboration, and improved business agility.

[Learn More ↗](#)

# Developing Widespread AI Literacy

As a 10-year-old, digital-native, research-stage biotech, Moderna exploded into the biopharmaceutical scene with approval of one of the earliest and most effective COVID-19 mRNA-based vaccines.

The “AI Applied” program was born in 2022 to promote widespread data and AI literacy, particularly geared toward those without a data science or coding background. Dataiku is an integral part of the program, empowering all profiles and skills to create data insights.

The program included:

- Tool training to introduce Dataiku for self-service analytics
- Use case discovery
- Operationalization to move use cases to product and track key metrics
- Employee satisfaction surveys to assess utility of the program to improve both skills and AI literacy
- Culture building to bring common understanding around future opportunities of data science

Key value additions include upskilling and collaboration across diverse teams (previously with limited interaction), transparency and centralization for data access, project and data pipeline reusability across workflows, and speed to accelerate new use cases by empowered participants of the AI literacy training.

[Learn More ↗](#)

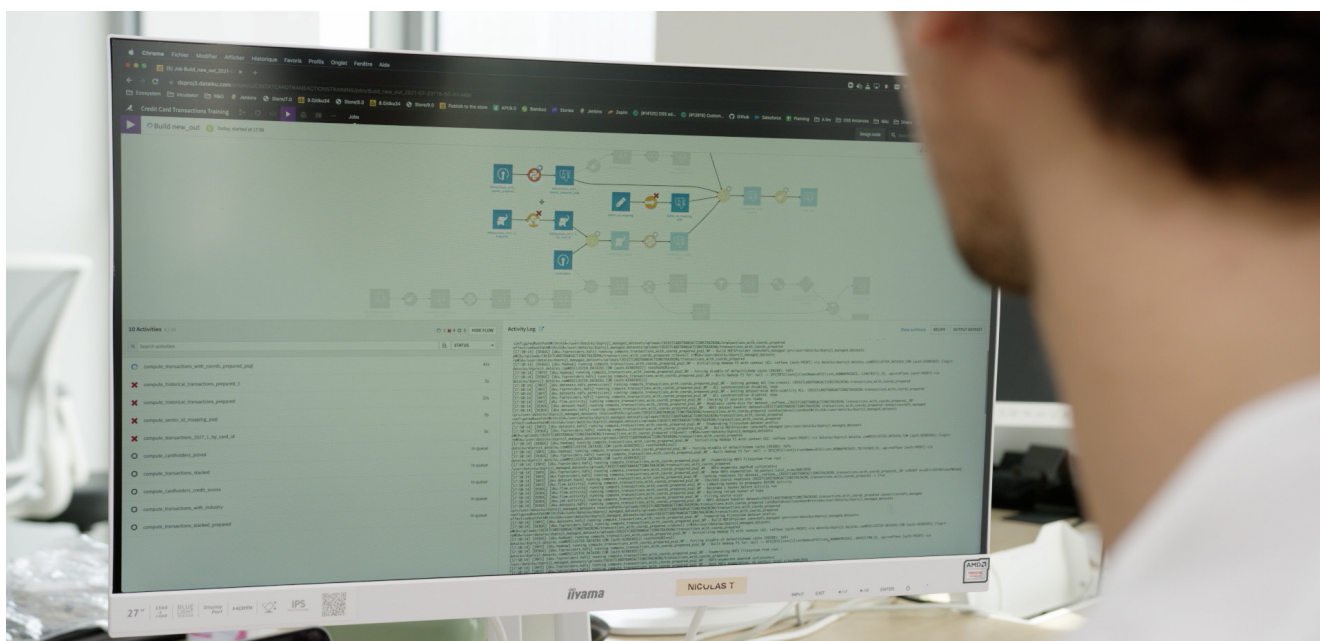


# Leveraging Generative AI Across the Value Chain

Data leaders in the healthcare and life sciences industries are already adopting Generative AI at a rate that outpaces any other industry (a whopping 76% have plans to leverage the technology, according to a [recent survey](#)). The utility of large language models (LLMs) and multimodal models is expansive across all components of the value chain.

There are generally three main application behaviors emerging:

- Creating structured data from unstructured sources, which streamlines self-service analytics and utility of previously locked data insights.
- Generating tailored content from both source data and previously built analytics or machine learning insights.
- Answering key questions from both public and private document stores with conversational AI.



Generative AI is showing promise to:

- Create novel chemical compounds tailored to optimize target molecular responses and designing proteins, antibodies, or other biomolecules for new therapeutic discovery and development.
- Query, structure, and summarize vast stores of scientific research based on relevance to pipeline investments.
- Automate the document-rich area of clinical research and operations, from generation of protocols and other medical documents to informing site management and milestone planning intelligence, or creating initial drafts of clinical submission reports.
- Streamline manufacturing quality reporting and anomaly management plus help prepare for inspections and audits.
- Structure medical safety reporting into analysis-ready data sets to inform early insights into drug adverse reactions to ensure new therapies brought to market are not creating unforeseen harm.

In addition, commercial teams building omnichannel and personalized content engagement with machine learning can enhance the utility of the models by generation of the key recommended marketing assets or engagement plans that ties in interpretability and explainability of model-driven actions. Medical affairs and digital health systems are seeing some of the largest gains in conversational engagement with both healthcare professionals and patients around the digital experience of a therapy or device and the treatment recommendations, particularly in individualized therapies.

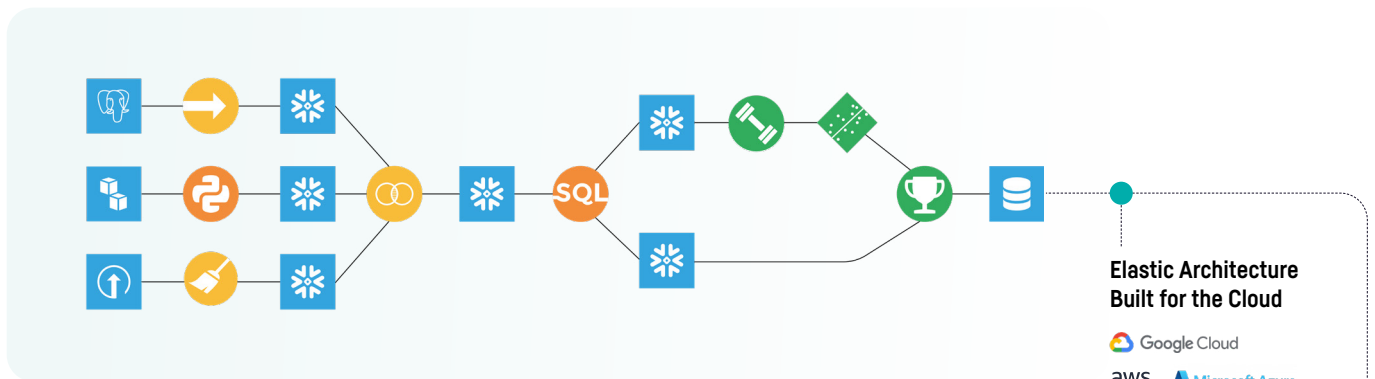
Dataiku allows life sciences organizations to apply a flexible platform approach to develop end-to-end LLM products in a controlled environment. Unlike point solutions, this means secure and safe experimentation and deployment to make Generative AI pervasive across all lines of business.

[Learn More ↗](#)

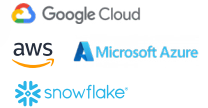


# Everyday AI, Extraordinary People

Dataiku is the platform for Everyday AI, enabling data experts and domain experts to work together to build data into their daily operations, from advanced analytics to Generative AI. Together, they design, develop and deploy new AI capabilities, at all scales and in all industries.



**Elastic Architecture  
Built for the Cloud**



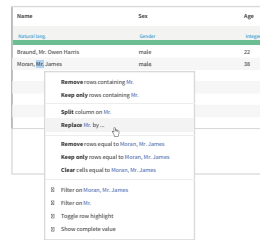
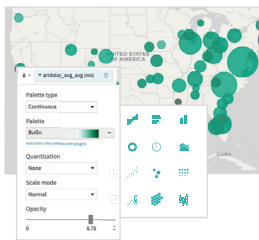
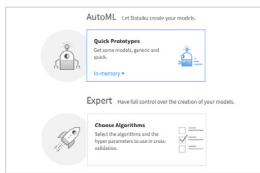
**Machine Learning**



**Visualization**



**Data Preparation**



**DataOps**



**Governance  
& MLOps**



**Analytic Apps**



**Generative AI**

