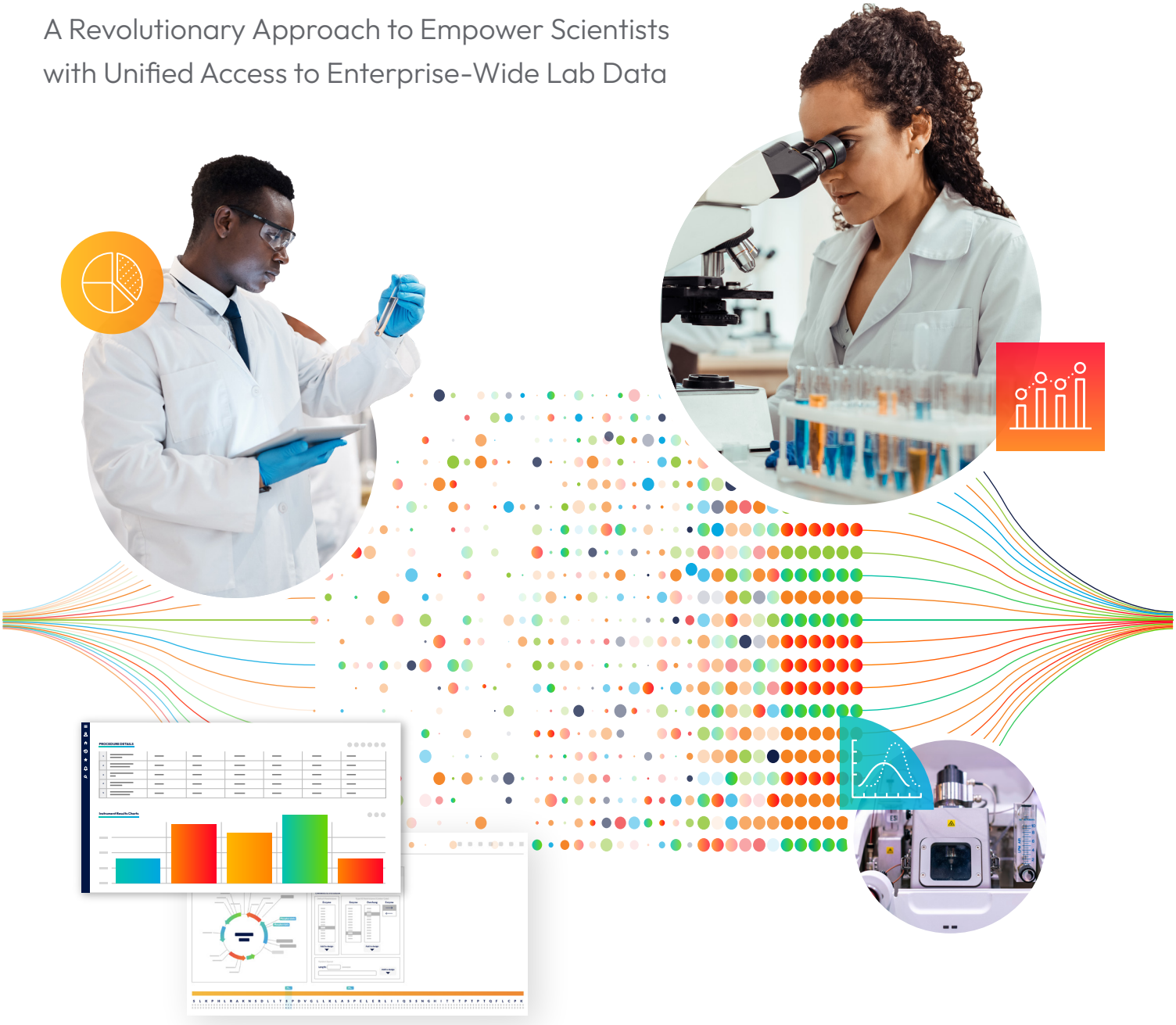
WHITEPAPER

sapio sciences

From Data Silos to Data Synergy

A Revolutionary Approach to Empower Scientists with Unified Access to Enterprise-Wide Lab Data



Introduction

In the last two decades, tools like electronic lab notebooks (ELNs) and laboratory information management systems (LIMS) have achieved what Gartner, Inc. calls "the Plateau of Productivity." These transactional systems have proven so popular that multiple instances of LIMS and ELNs, often from different vendors, have proliferated across and within research organizations. Organizations routinely capture experiments and data with good compliance, but their data now resides in isolated data silos that impede scientific decision-making.

Simultaneously, laboratory automation, the pace and scale of research, and clinical data science have put in-silico augmented research tantalizingly out of reach. Organizations possess the data that can help them explore complex biological mechanisms of action, accelerate the discovery of novel targets, rapidly discover and optimize therapeutics, and fuel artificial intelligence (AI) and machine learning (ML) algorithms. But that data resists interpretation because it isn't findable, accessible, interoperable, and reusable (FAIR), and most of all, it lacks the scientific context that gives raw data meaning.

This white paper reviews the challenges labs face in empowering scientists to capitalize fully on data, R&D's most precious resource. It also presents a solution to those challenges, a scientific data cloud called Sapio Jarvis℠. Jarvis automates data collection, syncing, and parsing from instruments and contextualizes it with data housed in an organization's scientific informatics systems. Jarvis is designed to increase research efficiency, provide a centralized location for visualizing and analyzing scientific data, and generate accurate, contextualized FAIR data to drive informed, innovative decision-making and fuel the AI revolution. Above all, Jarvis gives scientists unified access to enterprise-wide data in a simple, familiar, and readily usable way.

What we'll cover

Current Barriers to Unlocking Scientific Data

Since 2003, when the completion of the Human Genome Project launched genomics and its related subdisciplines (proteomics, transcriptomics, epigenetics, etc.), there's been a revolution in drug discovery, development, and delivery. This revolution has completely reshaped the way scientists explore biology, study disease mechanics, identify and pursue targets, and test the impact of their discoveries in the clinic. Personalized medicine is finally a reality, providing new, targeted treatments to patients and opening new lines of inquiry for future drug candidates.

This revolution has also presented research organizations with two broad data management challenges. First, organizations now generate unprecedented quantities of varied data types. While this provides significant opportunities to inform all aspects of scientific and business decision-making, a critical prerequisite is collecting, parsing, and presenting that data to scientists in actionable ways.

Second, capitalizing on this revolution's fruits requires data access, especially for scientists seeking to conduct in-silico augmented research such as model-informed drug development (MIDD), ML, and AI. There is no lack of algorithms and models to accept data. The challenge for organizations is finding and preparing all their data for efficient and effective consumption by these algorithms and models. Most organizations don't have a model problem preventing them from running advanced analytics; they have a data problem.

> We are spending a lot of our energy just trying to get all of our data harmonized so that some algorithm could maybe find anything of use.
>
> **—Vas Narasimhan, CEO of Novartis**
> (Quoted in Forbes)

Most organizations don't have a model problem preventing them from running advanced analytics; they have a data problem.

Data Integration Does Not Equate to Data Adoption

Current scientific informatics landscapes keep biopharma, research, and clinical labs from fully exploiting the data they possess. Many organizations currently utilize a patched-together assortment of tools and custom connectors to get raw data off instruments and into the disparate ELNs and LIMS that contain context about that data. They must then deploy another set of systems and one-to-one integrations to pull data into places where scientists can use it to assess scientific outcomes and decision-making.

While many ELN and LIMS providers offer applications to facilitate data import and export, these integrations are limited. Organizations often try to build their own integrations, but these can be difficult to implement and expensive to maintain. As a result, crucial organizational assets remain fragmented and dispersed. Without a centralized reference point for easily discovering and utilizing data, data lacks vital scientific context and meaning. Most importantly, it isn't FAIR—findable, accessible, interoperable, and reusable.

Why Data Warehouses and Data Lakes Have Failed Scientists

Organizations have used various informatics solutions to bridge the gap between disparate transactional applications and FAIRify their data. Scientific data management systems (SDMSs) are used to capture and store instrument data. Data warehouses are constructed to contain structured, filtered data processed to be queried en masse.

Because data warehouses can be rigid and hard to adapt to accommodate new data types, organizations have turned to data lakes, which can store data in its native format, allowing greater flexibility in integrating it with various applications, including advanced analytics tools. Warehouses, lakes, and lakehouses, though, are merely repositories for data and aren't equipped with interfaces and tools suitable for scientists. Extensive (and expensive) coding and spot integrations are necessary to get data out of these systems and into the hands of scientists to support their work.

The Shortcomings of Informatics Architectures

1. Data Warehouses

Structured data, rigid and hard to adopt.

2. Data Lakes

Unstructured data, little usability without code.

3. Data Lakehouses

Structured and unstructured data, no science-awareness.

	Data Warehouse	Data Lake	Lakehouse	Jarvis℠
Cost (TCO)	High cost to build/maintain	Low	High cost to build /maintain	Low
Performance (e.g. query)	High	Poor	High	High
Scalability	Costly	Scales to any amount or type of data	Scales to any amount or type of data	Scales to any amount or type of data
Adaptability (e.g. new datatypes)	Low; costly to add new data types/sources	Requires data ingestion, transformation, and governance processes	High	High
Intended Users	Data scientists. Requires custom integrations for use by scientists.	Data scientists	Data scientists. Requires custom integrations for use by scientists.	Chemists, molecular biologists, bioinformaticians, data scientists
Source Data Types	Structured transactional systems only	Structured to unstructured	Structured to unstructured	Structured to unstructured
Scientific Data	Requires extensions for scientific search; separate UI apps for scientific visualization	Requires extensions for scientific search; separate UI apps for scientific visualization	Requires extensions for scientific search; separate UI apps for scientific visualization	Scientific search and visualization built-in
Instrument Data	Requires separate SDMS/LIMS to feed & process	Requires separate SDMS/LIMS to feed & process	Requires separate SDMS/LIMS to feed & process	Captured and parsed automatically
Data Context	Yes, but hard-coded in schema	Discovered on read	Discovered on read	Automatically captured in flexible knowledge graph
Analytics	Send data to bioinformatics and reporting apps	Send data to data science/ML apps, bioinformatics and reporting apps	Streaming analytics, send data to bioinformatics, ML apps	Built-in statistics, data analytics, and reporting
Scientific Analytics	Separate applications consuming data from the warehouse	Separate applications consuming data from the data lake	Separate applications consuming data from the lakehouse	Built-in scientific analytics (flow cytometry, bioinformatics, etc.)

Understanding the Scientific Data Maturity Model

When managing and capitalizing on scientific data, too many organizations are stuck at the bottom of an adoption maturity model. They have acquired extensive infrastructure for organizing and managing data. Still, they cannot consolidate and unify access to that data in an actionable way for scientists and research and clinical project team leaders.

Without this access, scientists are unable to query all their organization's data intelligently and thoughtfully, asking questions ranging from simple queries ("Which studies ran with which subjects?") to complex interrogations crossing multiple siloed source systems ("What are all the experiments we've run involving samples of particular oligonucleotides?" or "What assays have been run on my project compounds and which ones show my desired selectivity and ADME profiles?").

Without a holistic view of their data, organizations struggle to use it in enterprise decision-making, including resource planning, utilization and management, compliance, and quality control.

Managers can struggle to answer simple questions such as "How many flow cytometry runs did we do last week across the whole company, and what was the average turnaround time?" And, at the highest level, data that isn't FAIR can't be quickly and effectively used to fuel advanced analytics.

Few solutions exist to address these issues and advance organizations along the adoption pathway. Some applications focus on gathering and collating data. Others serve as a background infrastructure to pass data through to ELNs and LIMS rather than capturing data so that it can be deployed to inform scientific and business decisions. Many systems fail to provide an intuitive place for scientists to view and interact with data and lack a scientifically AWARE™ analytics layer. Those that do supply analytics often call out to third-party analytics, adding more data processing hurdles for users to jump through.

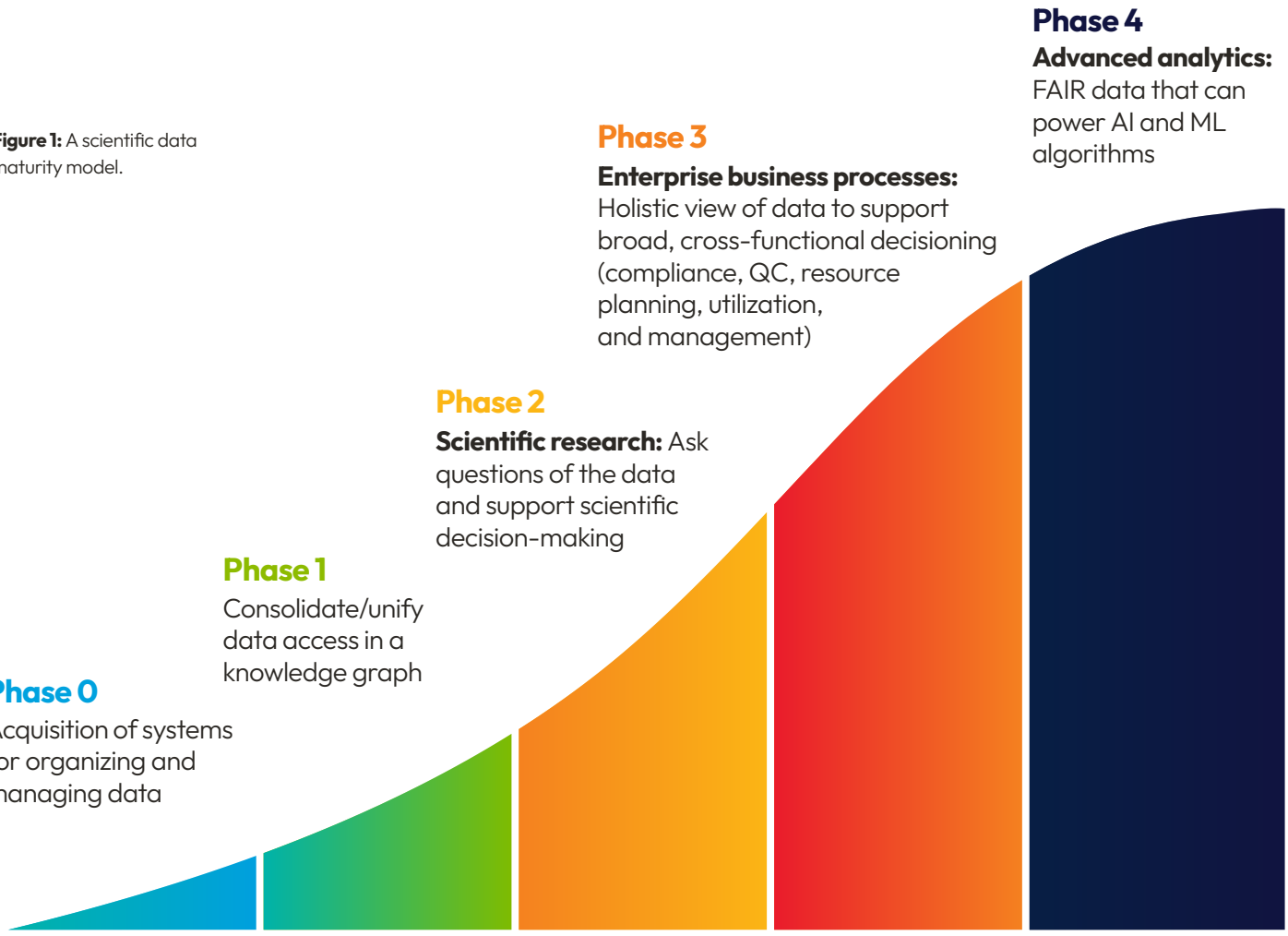


Figure 1: A scientific data maturity model.

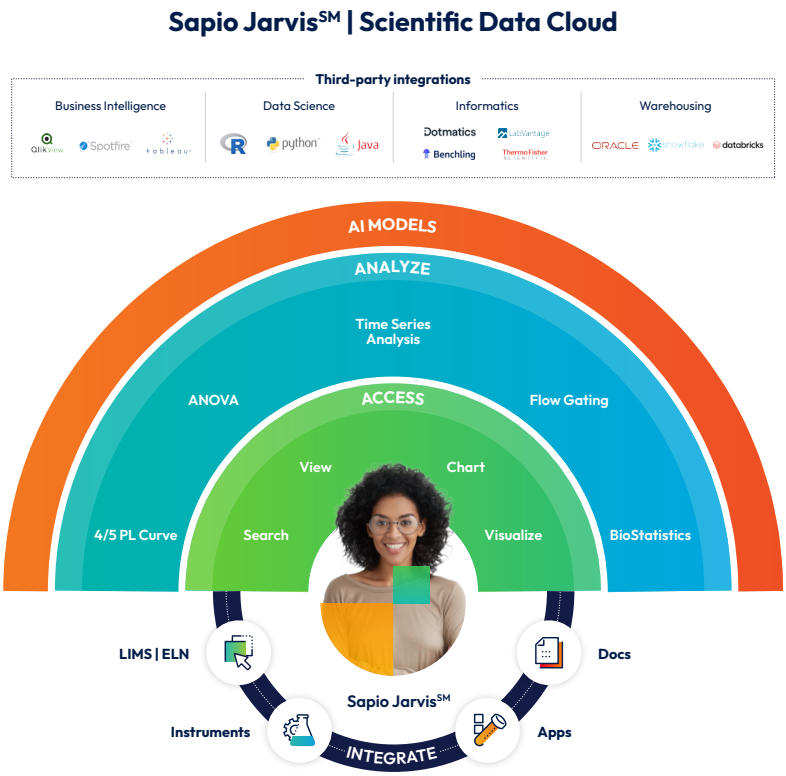
A Scientific Data Cloud for Scientists, Not Coders

A truly scientific, science-aware data cloud does more than collect, store, and parse raw data. It automatically syncs data with information stored in transactional applications such as ELN and LIMS to connect and provide valuable context to instrument results, which can help drive decision-making.

A scientific data cloud shows science entities in a meaningful way. Compounds render as compounds, plasmids show annotations, and proteins display as 3D objects that can be interrogated. Science-based searches should enable scientists to search for molecular substructures as well as chemical properties or assay results. Most importantly, a scientific data cloud should erase boundaries between systems and present all relevant context on a scientific entity, including information about the samples, experiments, and projects in which data was obtained, used, and consumed. And that information and context should be accessible regardless of in which LIMS, ELN, or instrumentation file format data is stored.

With a scientific data cloud, it's not about getting data in, which is the transactional focus of ELN and LIMS applications and many analytics tools. It's about providing a single place to get insights out—capitalizing on current and historical data to empower scientists in making data-driven scientific and enterprise-wide decisions. Notably, a scientific data cloud does this in a way that flexibly adapts to constantly evolving data as science progresses.

Figure 2: Jarvis sits at the center of an iterative process that enables scientists to collect, contexualize, explore, and analyze data, resulting in informed decision-making.



Biopharma decision-making is impeded when organizations and their scientists do not have a unified place to store, access, visualize, and analyze critical life science data in context. Even if advanced analytics algorithms could work across this data, scientists are unlikely to use systems that aren't scientifically AWARE®, where AWARE stands for:

ACCESS

Access to all scientific data, all in one place

WORKFLOWS

Configurable scientifc workflows that adapt to novel science without coding or extensive bioinnformatics engineering

ANALYTICS

Rich analytics built into a living knowledge graph

REUSABLE

Reusable objects and data to power productivity and scale

EXPERIENCE

A unified experience that delivers tools and applications in one place, including capabilities for CRISPR editing, plasmid/small molecule design, flow cytometry analytics, next-generation sequencing (NGS), and live ideation sessions to enable scientific collaboration and decision-making.

Introducing Jarvis[SM]

In the Marvel Cinematic Universe, Tony Stark, the industrialist inventor who becomes Iron Man, creates an AI that serves as a virtual assistant. Its primary duties are running systems across Stark's business and controlling Stark's Iron Man armor suit. He calls this AI J.A.R.V.I.S, which is said to stand for "Just A Rather Very Intelligent System."

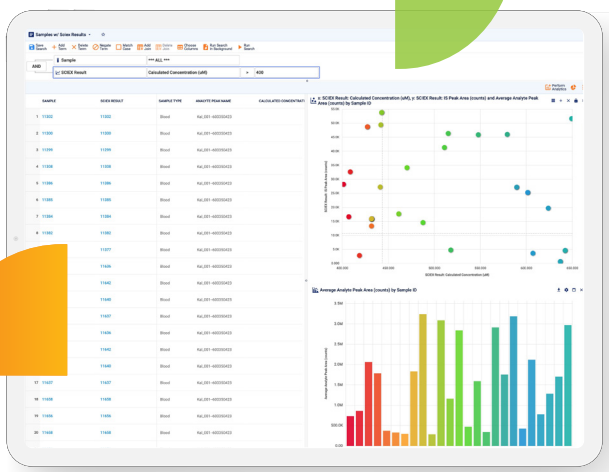
Like its cinematic namesake, Jarvis from Sapio Sciences manages and assists scientists in gaining access to and control of the scientific data they need to make decisions. Built on Sapio's low-code/no-code platform, Jarvis's primary innovation comes in its ability not just to collect and parse data off instruments but also to sync that data automatically with vital contextual data on samples, specimens, experiments, and projects contained in disparate ELN and LIMS applications—any applications, not just the ones developed by Sapio. This is accomplished through no-code pipeline rules built into Jarvis that are easily configured to detect new raw data and load that data to Jarvis, where it is parsed and synced to proper context about that data's use in projects, studies, subjects, experiments, and samples. Data is not only usable within a scientific team's target applications, but accessible within the Jarvis knowledge graph and through configurable dashboards. It's easy to track new data types—Jarvis comes with over 200 parsers for common instrument files, and new parsers can be rapidly developed right in the Jarvis user interface. Using the same rules interface, contextualized instrument results can even be inserted back into the source ELN or LIMS connected to the samples for which they were generated.

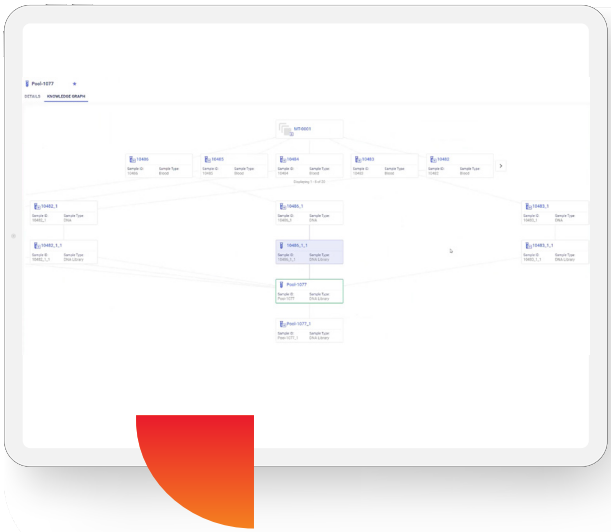
Importantly, Jarvis is designed to be used by all scientists, not just data scientists. In addition to capturing structured and unstructured text and numeric data and making it searchable within the Jarvis knowledge graph, Jarvis offers ways to visualize scientific data objects such as compounds (2D), biologics (3D), and annotated plasmids. Charts and tables are built into the system, and Jarvis provides meaningful statistical and scientific analytics and AI directly in the platform where the data resides—no more moving data in and out of data warehouses or data lakes. With Jarvis, scientists have all the data they need, all in one place. And it's easy to provide access to new data and capabilities straight from a common, scientifically friendly user interface–no coding needed. With built-in science tools for a range of analytics, organizations can empower their scientists to ask powerful questions of their data and, most crucially, find answers that may lead to the next research breakthrough.

It's easy to track new data types—Jarvis comes with over 200 parsers for common instrument files, and new parsers can be rapidly developed right in the Jarvis user interface. Using the same rules interface, contextualized instrument results can even be inserted back into the source ELN or LIMS connected to the samples for which they were generated.

Jarvis works behind the scenes to assemble raw data from instruments along with context for that data stored in any of an organization's scientific applications, enabling scientists to conduct powerful searches and chart findings without complicated parsing, joining, and cutting and pasting.

Here, Jarvis has pulled mass spectrometry results for all samples with a calculated concentration value greater than 400 mM.



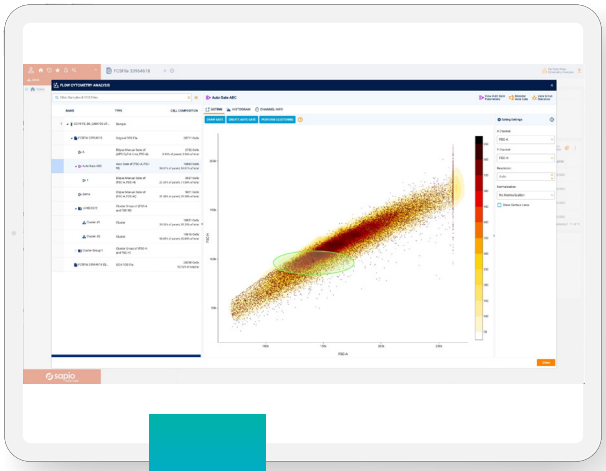


The Jarvis knowledge graph provides a straightforward way to track the connections among Jarvis entities and their use in samples, experiments, instruments, and projects, regardless of the source application storing that data.

Here, for instance, it's easy to see that Pool-1007 is derived from sample 10486 used in study MT-0001.

Jarvis includes an array of built-in tools for advanced scientific analysis, such as flow cytometry. Rather than having to find files, transfer them to their computers, upload files to analysis software, and conduct the analysis, scientists using Jarvis can access FAIR, analysis-ready data, all in a single user interface.

Here is a screenshot of Jarvis' built-in flow cytometry analytics and intuitive user controls.



Are you hunting for a way to FAIRify your data and empower scientific decision-making? Use this checklist to evaluate different solutions.

Ask yourself: Does this product...

	Jarvis	Product 1	Product 2	Product 3
1. **Automatically gather and organize all instrument results and files** from across your organization?	✓			
2. **Parse a wide variety of instrument result files** into searchable structured data?	✓			
3. **Give instrument data context by associating it with related experiments, subjects, studies, samples or projects** from different ELNs/LIMS and other systems used by your organization?	✓			
4. **Store all scientific entities from your organization,** such as molecules, proteins, and plasmids, and show them visually to scientists?	✓			
5. **Use simple rules to pipeline instrument data and results** back into the appropriate ELN experiement or LIMS process/workflow?	✓			
6. **Organize all the data** into an easy-to-navigate knowledge graph?	✓			
7. **Allow scientists to visually build advanced queries** — no technical skills required?	✓			
8. **Allow scientists to perform advanced statistical and scientific analysis (e.g. flow cytometry gating) on all their data** directly without jumping to another application?	✓			
9. **Allow scientists to use science-aware search** such as molecular substructure searching across tens of millions of compounds?	✓			

To see Jarvis in action, contact Sapio to schedule a live demo preview.

Request a Demo

sapio sciences

1. https://www.dcatvci.org/features/top-10-key-trends-in-new-drug-approvals-small-molecules-biologics/
2. https://www.statista.com/statistics/791263/total-r-and-d-pipeline-size-timeline-worldwide/
3. https://www.pharmaceuticalcommerce.com/view/managing-regulatory-intelligence
4. https://www.einnews.com/pr_news/619280785/ai-in-drug-discovery-market-to-hit-usd-8-10-billion-globally-by-2030-at-25-cagr-astute-analytica