scimcon

Machine Learning and Lab Informatics

Where to Start?

SCIMCON WHITEPAPER

FEBRUARY 2024

Lab Informatics

Where are we now?

The delivery of informatics systems within lab-based companies has always been a very active space. Traditionally this was focused on single systems or multiple connected systems covering a specific set of workflows. The vast majority of labs now operate at least some form of LIMS, ELN, LES, SDMS, CDS etc.

These lab informatics projects are challenging as they generally affect a large portion of the companies' scientists and often cover a wide range of differing workflows.

However, over recent years driven by success in other industries and accelerated by the demands to change working practices during the pandemic, Digital Transformation (DT) has become an area of increasing focus. DT dictates a much wider scope to laboratory systems with greater depth of functionality, increased benefits and a longer delivery time frame associated with the implementation of a programme of changes.

It feels like many lab-based companies and their lab informatics teams are just starting to understand the challenges associated with DT. However, with DT still very much a work in progress we are already faced with preparing for the 'Next Big Thing'.

The 'Next Big Thing' is of course how to use laboratory informatics systems to deliver data that supports Artificial intelligence and Machine Learning (AI/ML).

What does ML mean for laboratories?

Let's start by defining a high-level objective for ML in the laboratory domain:

Informing/streamlining scientific experimentation based on the data and findings of past experimentation.

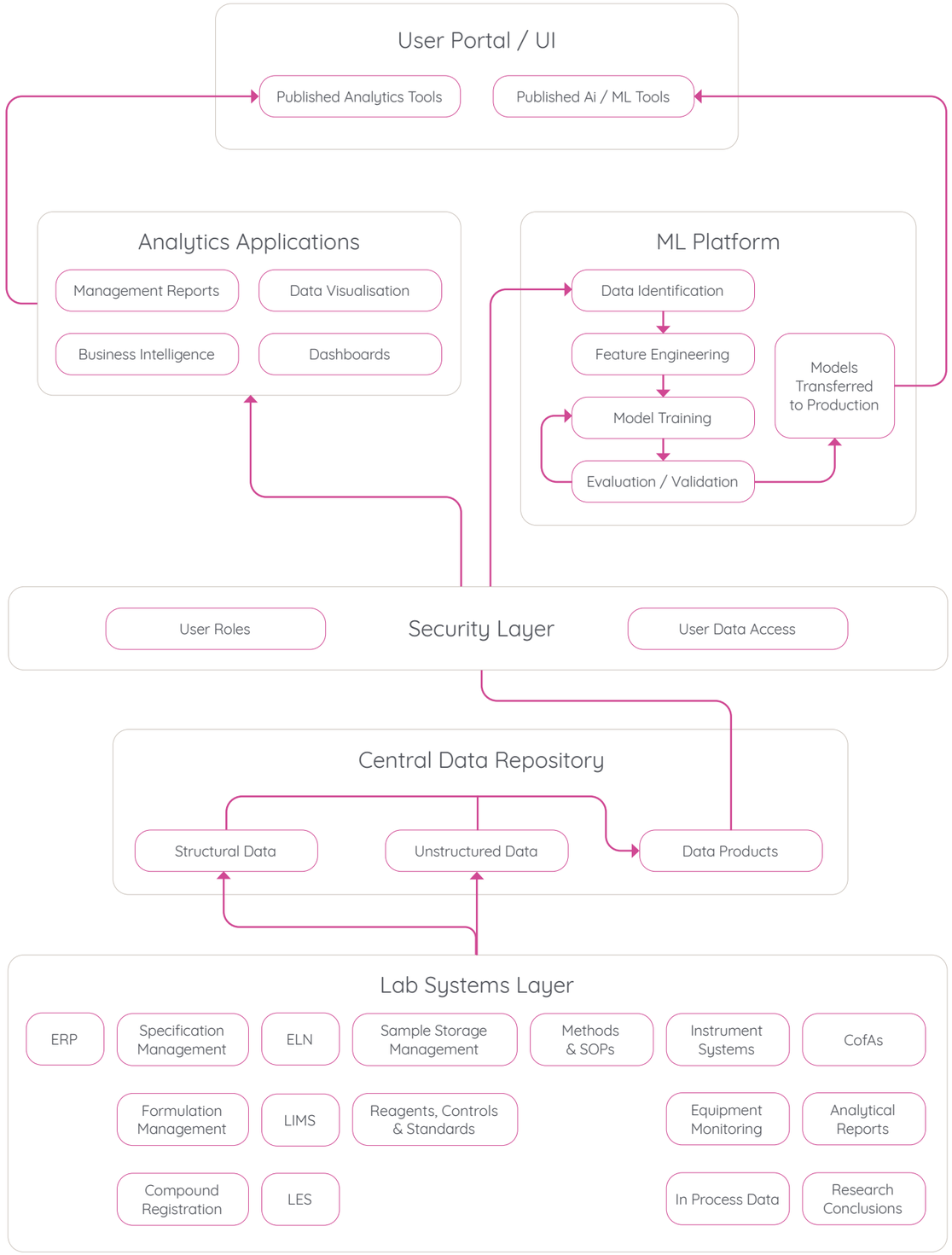
The specific use cases for each organisation or lab group within organisations will differ. Below are a number of example use cases where AI/ML may be used to:

- Reduce the number of test formulations needed to produce, test and analyse results to determine the optimum formulae for a new or revised product.
- Predict the efficacy and toxicity of a molecule based on its structure or its genetic markers.
- Plan instrument runs to best utilise expensive instrument availability while delivering analytical results in a timely manner.
- Reduce instrument or equipment downtime by collating a wide range of operating conditions, parameters and test results to predict when maintenance will be required.
- Combined ML with digital twins to predict the outcome of changing parameters in complex reactions.

A layered approach

Before we look at the question, “Where should lab organisations start when planning to create an ML environment”, it will help to discuss what such an ML environment entails.

The following diagram is a schematic to aid the discussion of some of the key components that make up an AI/ML ecosystem.



The diagrams within this document are not intended to be an architectural design and neither are they intended as a complete illustrations of what is required.

Lab Systems Layer These are the day-to-day transactional workhorses of your laboratory informatics systems; you probably won't have, or need all of them but they are the tools your scientists use to get work done. The key players are LIMS, ELN and instrument systems supported by various additional systems depending on the types of lab workflows and data flows required.

Central Data Repository (CDR) The purpose of this layer is to consolidate data produced by the lab informatics systems together with unstructured data such as key MS-Word documents, PowerPoints, meeting minutes etc.

A key advantage of this two-layer approach is that it separates the systems, data and users of day-to-day operations of the lab informatics systems from the data cleaning, formatting, normalising and analytics data provision workload of the central data repository.

In addition to housing internally generated data the Central Data Repository can also act as a landing zone for externally generated data from 3rd parties such as CROs, CDMOs etc.

The Central Data Repository can be built on several different technologies.

Enterprise Data Warehouses (EDW) are built to store highly curated and structured data as a result they support the fastest query results. EDW's can be thought of as a 'Database of Databases' in that they can store data exports from many operational systems all in one place. EDWs have high levels of integrity when transactions such as writing, reading and updating are executed. In addition, the high degree of structure imposed by EDWs facilitates user access and data security.
For example, this makes EDWs a good match for receiving LIMS – Sample, Test, and Result data, together with product data from an ERP and compound registration data etc.

Enterprise Data Lakes (EDL) are more flexible in that they can store unstructured, semi-structured and structured data. The EDL stores incoming data in the format it arrives in without adding metadata or indexing. If the data is structured, the structure is preserved. If the data is unstructured such as a Word Document, PDF or PowerPoint slides they are stored in their native format. Any data can be loaded directly, and transformations can be performed post-loading or only when required.

EDLs can be used for example to receive batches of data from third parties such as several different CROs each with different data export and reporting formats.

The downside of this flexibility is that retrieving data from an EDL is more complex. Whereas EDWs have a defined database schema and indexes, EDLs can be likened to searching through folders to locate the data required. In addition, applying security and access roles within an EDL is more challenging than in an EDW.

To improve search utility and performance EDLs are often used in conjunction with Data Catalogue software. The Data Catalogue can create schemas and add metadata as the data is uploaded to the EDL. Searching the Data Catalogue will give the relevant results for the given metadata or features.

Central Data Repositories often combine EDWs and EDLs to give the required functionality.

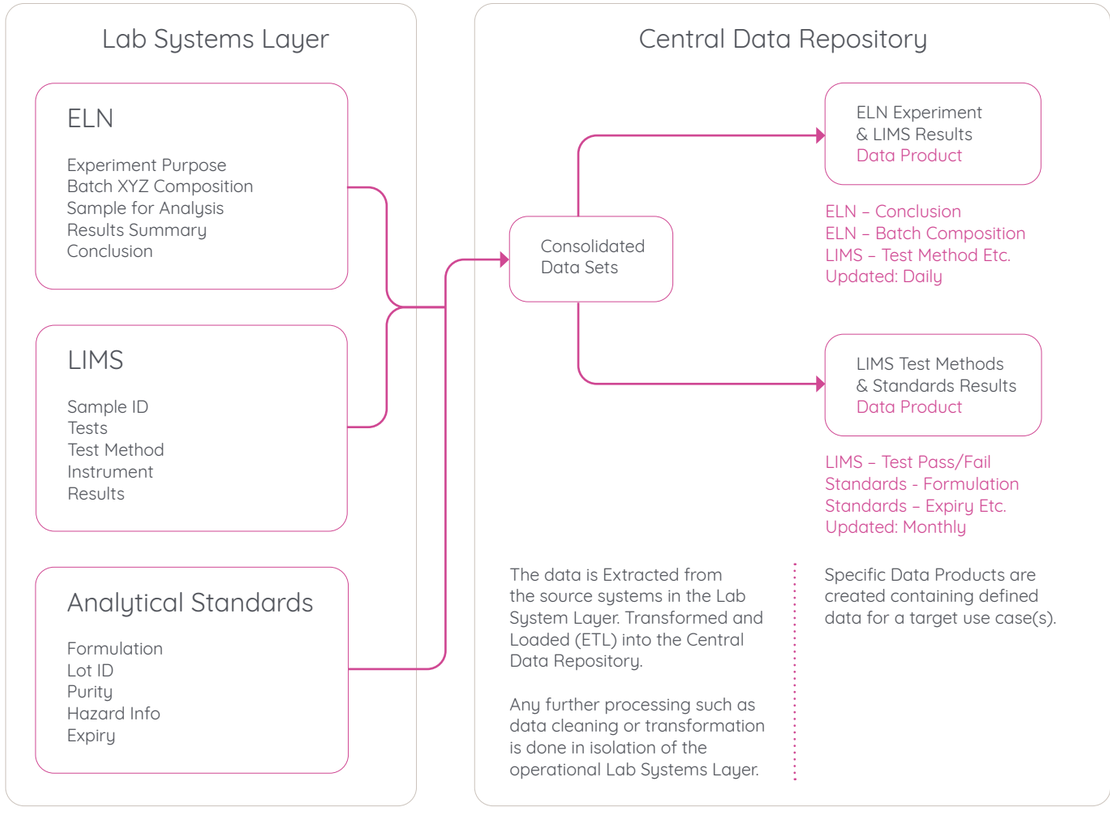
Enterprise Data Lakehouses (EDLH) are a modern data platform which themselves combine the benefits of EDWs and EDLs. EDLH can store unstructured, semi-structured and structured data. Their advantage is that they also benefit from improved metadata linking, searching, data integrity, performance, and security control similar to that of a traditional EDW but across all types of data. Although there is a limited number of off-the-shelf vendors providing these platforms their numbers are set to increase.

There are also new data architectures becoming available such as Data Fabric that present a central virtualised data platform that supports the use of a distributed data storage infrastructure. This means data can be 'left in situ' and accessed across multiple environments for example within a vendor's AWS cloud account, the organisation's virtual private cloud account and on-premises lab informatics systems. Similar to EDLHs the Data Fabric provides a holistic view of all the data and embeds data governance and data security / access controls.

Data Products bring together data from various sources into a defined dataset.

The data has been cleaned, it is of a known quality and the frequency of data refresh is clear. The Data Products are structured in a manner that supports its targeted use case(s). Data Products have an owner and intended data consumer(s).

The diagram below illustrates the journey of the data from source systems to become reusable, accessible and efficient data products.



The data generated is for a transactional purpose:
- Record the experiment
- Manage the samples and tests
- Track reagent lots.

Multiple individual lab systems, designed to focus on transactions and not prioritized for data reuse.

→

Consolidated data from multiple source systems. Additional data cleaning or further transformation is done in isolation of the operational Lab Systems Layer.

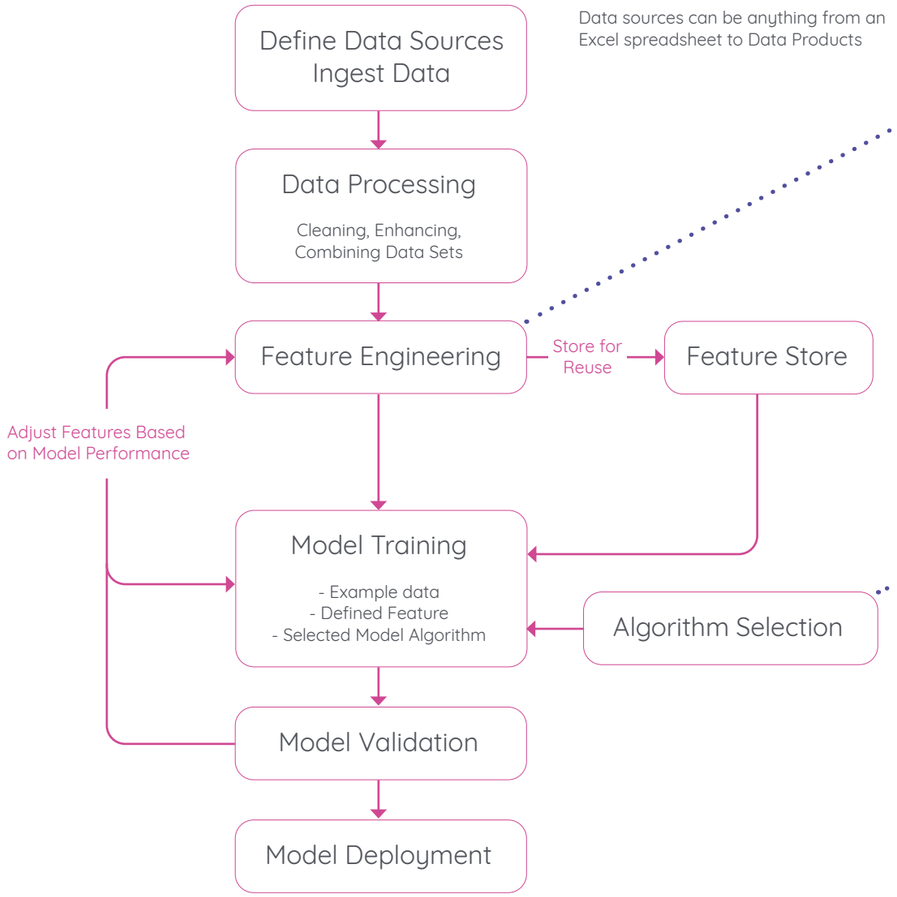
→

Data Products are high-quality, cleaned data extracts optimized for efficient reuse. They are ideal data sources for analytics and ML use.

Machine learning platforms

Machine learning platforms (MLP) provide tools that support the ability to define data sources, the iterative development of ML algorithms and the deployment of the models for use. Machine learning platforms can be built in-house with the assistance of open-source tools, purchased as a commercial off-the-shelf suite, or can utilise a combination of commercial and in-house tools.

The diagram below shows a typical ML workflow and highlights some of the key considerations.



Data sources can be anything from an Excel spreadsheet to Data Products

Feature Engineering involves creating rules to classify or weight data items.

Example ML Task – Examine various data elements and how they influence a customer's car buying decisions.

- Data Element 1 – Buyer's Salary
- Data Element 2 - Car's Cost
- Data Element 3 – Buyer's Age
- Data Element 4 – Car's Top Speed
- Target Attribute – Did Buyer Purchase Y/N

How should the machine learning algorithm interpret:

- A £50 difference in Salary
- A £50 difference in Cost
- A 50-year difference in age
- A 50 mph difference in top speed

Without setting weighting, scales or classifications for each data element the ML algorithm would not recognize the number 50's context. It would view a 50-year difference in buyers' age the same as a £50 difference in the car's cost.

When we consider our real-world data supplied by the Lab Systems Layer it is clear for feature engineering to be successful, good laboratory data domain knowledge is essential.

Factors affecting algorithm selection include

Supervised Learning is used to predict an outcome based on learning from previous data sets of example data. The data within the example data sets have the correct outcome identified (worked examples). The following types of algorithms can be used :

Logic Regression
Used where the Target Attribute is one of two values
- Yes or No
- Up-regulated or Down-regulated

Multinominal Logistic Regression
Used where the Target Attribute is one of several values
- Soluble, Insoluble, Sparingly Soluble, Miscible, Immiscible, Hydrophilic or Hydrophobic

Linear Regression
Used where the Target Attribute is a numeric value
- Compounds where activity is > 80%?
- What is the ideal amount for the new ingredient in this formulation?

Semi-Supervised Learning is used where there are insufficient worked examples or producing worked examples requires too much effort. In this type of learning the use of both worked examples and non-worked examples are used.

Unsupervised Learning used where there are no datasets with worked examples available. The algorithms are used to detect patterns and predict outcomes based on those patterns. Algorithms can use several techniques including:

Clustering
- Reduces the complexity of the dataset by grouping similar data together into clusters. Clusters can then be further analysed for patterns.

Dimension Reduction
- Reduces the dataset complexity by eliminating data elements from the algorithm that don't affect other data items.

Machine Learning and Lab Informatics

Where to start?

As is often the case with complex integrated platforms there is not a straightforward answer to ‘where should we start’. It will very much depend on the answers to questions, such as:

- How much of the target ML environment is already in place?
- What are the business drivers for the individual components that make up the ML dataflows?
- What is the state of the current systems in the various layers?
- What are the business drivers and priorities of the system implementations, replacements or upgrades that fall within the scope of the ML environment?
- How is the project likely to be funded, will it cover a complete solution, or will you need to demonstrate value on a smaller scale to secure funding?

For most organisations, their answers to the above question will suggest directing efforts at one of the three main areas, the Lab Systems Layer, the Central Data Repository or the Machine Learning Platform.

The lab systems layer

Start here?

There is a good reason this layer is at the base of the ML environment diagram: the lab systems generate data which forms the foundation of the ML dataflows.

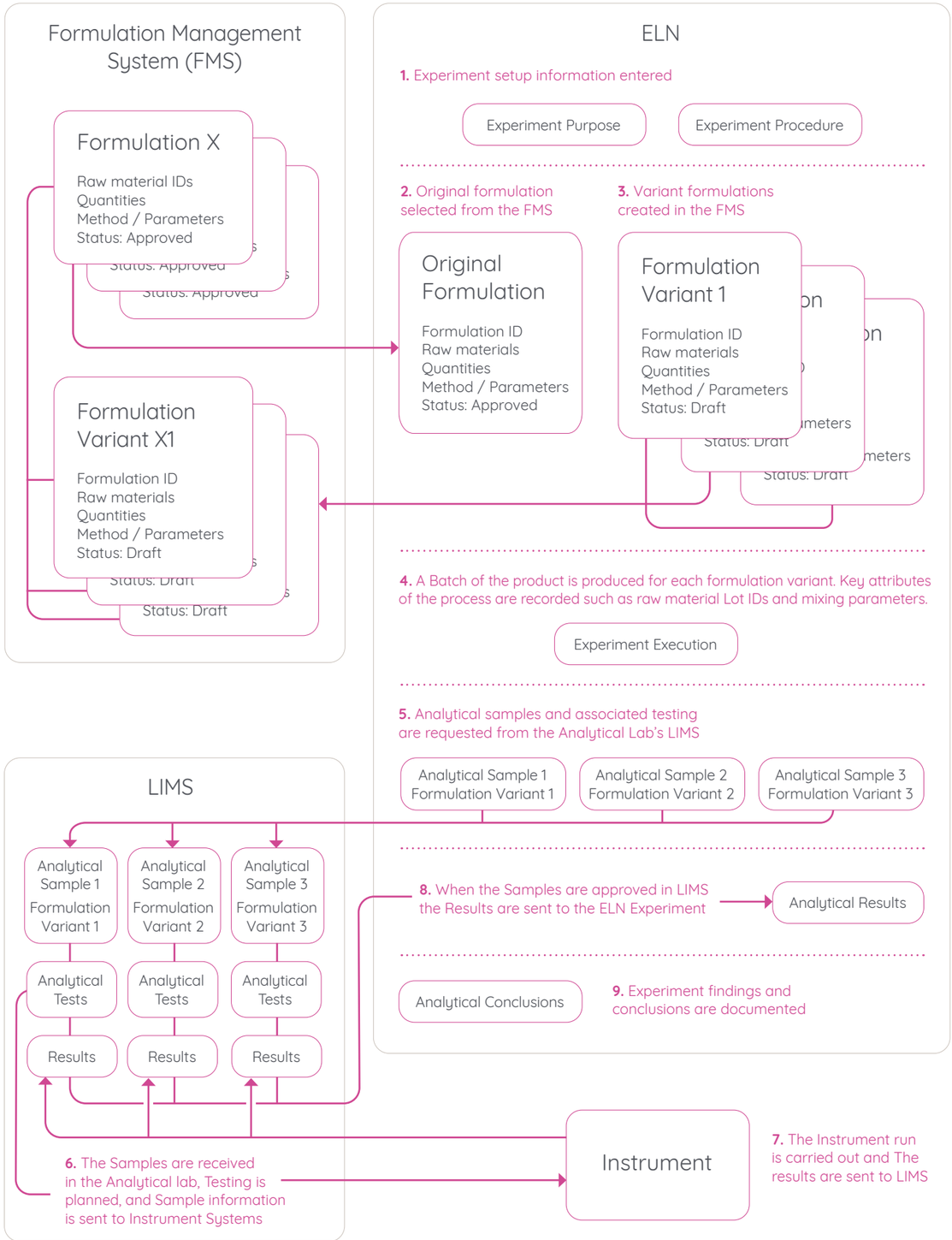
To illustrate the type of Lab Systems, system integration, underlying data and governance the layer can provide the following section will explore what is required to support an example R&D scenario.

The question posed by this scenario is:

‘What changes are needed to a product’s existing formulation to make the product more environmentally sustainable?’

To answer the question posed, experiments will be run to produce and evaluate batches of products created based on various formulation variants. The following Lab Systems layer diagram details the systems, configurations and interfaces that could be deployed to enable the scenario.

Example Lab Systems Layer Diagram



Although at a stretch, it could be argued that it is still possible to operate a laboratory using paper lab notebooks and lots of Excel sheets it certainly isn't efficient. However, without the widespread use of suitable systems to support scientists' day-to-day work and the lab workflow it is highly unlikely that this foundational data will have the quality, integrity, availability and completeness to underpin successful ML outcomes.

The example Lab Systems Layer diagram also shows that implementing isolated systems will not be sufficient, the systems within the Lab System Layer will need:

Consistent reference data: To facilitate categorising similar data together, managed reference data should be extensively used. For example - the keywords describing the experiment purpose would need to be consistent across similar experiments within ELN notebooks. Allowing only the use of free text within the experiment purpose would inhibit consistent categorisation. In addition, where the same reference data is used in multiple systems, there must be governance in place to ensure it is consistently aligned. For example, the same units of measure are often used across multiple systems within the Lab Systems Layer.

Persistent Entity Identifiers: To facilitate the cross-linking of data across systems and creating a holistic data model, data entities created in one system and used in another must be consistent. In the example, Formulations are created in the FMS and are used in the ELN and would be also referenced in the LIMS. Similarly, samples are created in LIMS but will be referenced in the ELN and Instrument systems.

Bi-directional Interfaces: Manual transcriptions between systems should be eliminated if workflow efficiency and data quality are to be maintained. In the example, various interfaces would be needed to share data entities. Example FMS & ELN, ELM & LIMS, LIMS & Complex Instrument Systems and LIMS & Simple Instruments

Scope of Data: Until recently focus has been on recording positive outcomes within lab systems as these tend to drive the decision-making process. To gain the maximum benefit from ML negative outcomes need to be equally well recorded. In our example diagram, there are three variant formulations recorded however, it is entirely possible that the formulation chemist originally made ten batches and only entered the three batches that did not separate within 24 hours. In this case, ML learning would not have access to the data for the seven formulations that failed at the 1st step.

The potential power of the data model, even in this simple example is clear:

- Comparing similar experiment outcomes and the factors that affect them.
- Examining which types of formulation changes are successful and unsuccessful.
- Comparing the performance of different raw materials across large numbers of experiments
- Predicting advantageous formulations changes

In reality, this example could include several lab systems in addition to FMS, ELN, LIMS, and a single instrument type.

In global organisations, their laboratories often use multiple instances of the same lab system or lab systems from several different vendors.

If your organisation has not yet fully embraced systems in the Lab Systems layer, they are not well-integrated or their reference data is not consistent then generating source data for ML will be problematic. **If this is true of your organisation this would be a good place to start.**

Central data repository

Start here?

The easy option would be to say 'Start Here' if, your organisation's Lab Systems Layer is in a well-developed state and that would be valid. However, that is not the only reason to trigger a CDR implementation project.

Don't wait for the perfect Lab Systems Layer - Implementing systems within the Lab Systems Layer in a way that supports an all-encompassing data model is difficult in isolation. Being able to prototype and develop a lab system's CDR data feeds, validating how the data is cleaned and transitioned into data products during its design and implementation will produce valuable insights. Including the links to CDR as requirements at an early stage can prevent redesign efforts post go-live.

Data analytics and business intelligence - While creating an ML platform may not be an immediate priority, improving data analytics and business intelligence may be more pressing. Making your start by building the CDR environment will facilitate data analytics and business intelligence in the short-term and provide a solid head start when the organisation is ready to consider ML.

A pressing need to handle 3rd party data – Organisations are increasingly using 3rd party companies to support their processes such as contract testing labs, Contract Research Organisations (CROs), Contract Development and Manufacturing Organisations (CDMOs). These collaborations can generate large amounts of data exported from the 3rd parties. There is no set format for the data exports, some will be in the form of PDFs, and some will be in a more technical format such as a form of XML. Having a landing zone inside CDR that will securely store incoming third-party data irrespective of its incoming format can be a sufficient driver to initiate a CDR implementation.

ML platform

Start here?

It may seem counter-intuitive to start with the top layer when the supporting layers are not complete or indeed in place at all. There are good reasons to do this, however.

A strong ML learning use case – Your organisation may be only just beginning the process of creating a solid Lab Systems Layer and CDR. These programmes can take years and a great deal of effort to deliver from start to finish. This does not and should not prevent the creation of a Machine Learning platform. There will be a need for greater manual effort to locate and prepare the source data compared to a fully functioning layered approach, the time to results however, will be significantly shorter.

Proof of Concept / Proof of Value – As mentioned above building a multi-layered environment for ML within the laboratory space can involve significant time and budget. An alternative to a 'big bang' approach is to evaluate ML platform's capabilities together by identifying a small number of good-quality data sets that can be prepared manually or sourced externally. The results from these proving use cases can be used to help build a business case for a solution of a wider scope.

Conclusion

The best time to plant a tree?

ML learning may be 'The Next Big Thing' in the lab informatics space but it is here to stay and will continue to increase in importance. There are many options and drivers for how and where to start building ML learning pipelines. On reflection perhaps where to start is less important than making a start somewhere.

SCIMCON WHITEPAPER

Machine Learning and Lab Informatics

For more information about Machine Learning or any other areas of Lab Informatics, please call one of the Scimcon team today on +44 (1638) 661 631 or email info@scimcon.com

Scimcon is an Information Systems consultancy with over two decades of experience working solely with lab-based companies. We assist with individual information system projects, like LIMS, ELN, CDS and instrument integration, we help redefine lab IS strategy, and we support the 'lab of the future' with digital transformation projects. Our expertise and support covers the complete lifecycle: business case creation, requirements definition, vendor selection, solution design, implementation and validation. Our insight, experience and approach simplifies complex projects hastening delivery and improving outcomes.

scimcon

FEBRUARY 2024