Turning Data Into Information

Making Sense of an Avalanche of Supply Chain and Production Data
Table of Contents

Introduction ........................................................................................................................................................................... 3

Data Rich, Information Poor .................................................................................................................................................... 3
Case study: whitespace in a biomanufacturing facility ........................................................................................................... 4

The Devil is in the Detail .......................................................................................................................................................... 4
Collect data ................................................................................................................................................................................... 5
Parametrize data ........................................................................................................................................................................... 6
Build and validate models ........................................................................................................................................................... 7
Multi-model / design of experiment analysis ........................................................................................................................... 8
Case study: facility fit model ....................................................................................................................................................... 8

Conclusions ................................................................................................................................................................................... 8
Introduction

As biopharmaceutical manufacturing has evolved from small-scale lab production in small flasks to massive commercial production in deep-tank bioreactors, an avalanche of production data has evolved to support it. From individual unit operation skids to distributed control systems and enterprise resource planning tools, biomanufacturing today is buried by the sheer quantity of data coming in from the manufacturing floor.

The challenge is to see this data in a way that allows its manipulation and use by as many groups as possible, and to turn the data into information that can be used to understand current as well as future system states. In this whitepaper, we discuss ways of integrating data from a number of disparate sources, and how to turn this data into a comprehensive information model that will be of value to the whole organization.

Data Rich, Information Poor

Biopharmaceutical manufacturers are unique in their need to understand the patterns of data across many different systems. Biologics manufacturing requires considerably higher levels of measurement and control than most production systems since the process itself is regulated (rather than just the end product). This means the process measurement systems must ensure the process is under control and show proof that the product was produced according to the data submitted to regulatory agencies.

In addition, the operating conditions for biologics manufacturing are often very close to those that are ‘ideal’ for a variety of organisms that can contaminate or otherwise affect production. This has created the need for highly controlled and sterile processes, and measurement systems around those processes. It is not uncommon for hundreds of measurements to be taken every second in a large biomanufacturing facility operating at full capacity.

Figure 1: Utilization of a cip skid in the facility.

Getting enough data is clearly not the issue in such biologics facilities: rather, the crush of data and process measurements are overwhelming. Many biotech facilities have multiple systems controlling process execution, multiple GMP-validated systems of record, and multiple measurements of process parameters. This means that data is stored in many places and is typically not directly accessible to engineers in the facility. This makes it difficult to find bottlenecks and key resources, perform sensitivity analysis, and make process changes with a high degree of confidence.
Case study: Whitespace in a biomanufacturing facility

One of the critical issues in a biotech drug substance (bulk) manufacturing facility is the need to find downtime when scheduled maintenance can be performed on tanks or other key equipment. Bulk manufacturing facilities are typically 24/7 operations and in a busy facility it is not uncommon for key pieces of equipment to have high utilizations of 80-85%. The challenge for planners is to identify periods of time when the resource is not being used, as well as the length of time that the resource could be down without affecting operations. Such a period is called ‘white space,’ and the identification of such white spaces is essential for the efficient operation of all biomanufacturing facilities.

One of the major difficulties with white space analysis is that the amount of time a resource spends ‘idle’ (and thus available for periodic maintenance or other activities) is a function of all of the other activities that are happening in the facility. CIP and SIP skids are a classic example; since they are utilized on an ‘as-needed’ basis and clean a number of pieces of equipment, delays or other events can create a highly unpredictable operating pattern. Figure 1 shows actual data for the CIP skid utilization in one facility (non-colored sections indicate periods of white space).

Emerson’s analysis showed that in order to get the whitespace for this CIP skid, it was necessary to obtain CIP durations for the 65 pieces of equipment that needed to be cleaned by the skid. In addition, the likelihood of those pieces of equipment running to schedule also needed to be obtained by the toolset, as well as the relationship of any delays in particular processing steps to the overall cleaning.

The conclusions of the study were compelling: at the proposed run rate of the facility, the likelihood of enough white space to perform maintenance was essentially zero. This was despite the fact that the overall utilization of the CIP skid was relatively low compared to other areas of the facility (80%); this time, however, was composed of many small ‘buckets’ of time rather than one single period that could be used to schedule maintenance. The analysis revealed that by purchasing additional CIP skids and some other equipment, the run rate at the facility could be increased.

The devil is in the detail

The main characteristic of the case study above was the need to firstly collect a large set of data, then to assimilate that data into a model framework, and finally to use that model to generate information about the business process. Implementing such an approach with a high degree of accuracy is critical in the biopharmaceutical sector where small errors have significant impacts.
Figure 3 shows an integrated approach for creating models that accurately reflects the processes, facilities, and supply chains they represent. The advantage of such a framework is that it not only builds models on real data, but also parametrizes the raw data into a format that allows it to be used by the model in a meaningful way.

Collect data

In the first stage of the process, enterprise data is collected (or access to the data is provisioned if it is in an SQL database or other repository). If multiple historians are available that store similar data, data is collected as close to the generating source of that data as practically possible. As an example, consider a highly automated facility with a number of systems of record. The batch historian contains information about recipe phases and instructions from the automation system. The automation system also passes information about process execution to a planning system, say SAP. People involved in the process also write the process times on paper tickets.

In this case, the ‘closest’ database is the batch historian, since it is the system of record attached to the hardware, and also contains significantly more information than either SAP or paper ticket data. While not all of this information may be utilized in an analysis of the data, it is important to ensure that at least some additional detail is available to support the data being primarily targeted.
Parametrize data

In the second stage of the process, the incoming data sources are analyzed. One of the fundamental steps that is missed in data analysis is the need to ‘parametrize’ the data, or essentially redact the data using curve fitting or parametrization tools to understand the key patterns of data rather than using the raw data itself.

Data parametrization is important for a number of reasons. First, it requires users to examine the data and exclude outliers that have a statistically large effect on the final model. Second, patterns in the data are often readily visible. Figure 5 below shows the processing times for a chromatography step main operation, where each dot represents a batch (and clusters of dots represent a campaign).

![Figure 5: Gradually improving processing times for a chromatography unit operation.](image)

There are two clear patterns in this data: a processing time that is variable between 4.5 and 6.5 hours within each campaign, and a gradually decreasing processing time between campaigns. One or both of these patterns may be important depending on whether the model being constructed is a short-range or long-range one.

Parametrization consists of reducing the amount of data that needs to be considered by the model to the minimum possible dataset. A number of automated toolsets exist to perform these tasks, including curve fitting tools, regression tools, and Markov models. Looking at the chromatography data above, parametrization of the data may involve creating a histogram of the data and then fitting some kind of distribution to that histogram. Figure 6 shows the outcome of this parametrization where the bars represent the actual data, and the line represents the fitted distribution.

The importance of parametrization of the data lies in the fact that the data distribution has hundreds of individual data points, while the fitted distribution contains just four – a minimum, maximum and two ‘shape parameters’ that determine the mean, mode, standard deviation, etc. This allows subject matter experts to quickly validate the accuracy of the dataset, and gives model analysts valuable ‘levers’ to change when analyzing the effect of individual datasets on the overall model.
Build and validate models

Building models is a well-understood process, and there are a number of toolsets that generate excellent models when data has been correctly extracted. However, current solution offerings suffer from a classic biotech shortfall: they either provide an excellent product or a great consulting service, but never both. There is a desperate need in the industry for a vendor with a ‘complete systems’ approach like Emerson provides – both in terms of building tools specifically designed for biotech, and with the deep understanding of the biomanufacturing process that generates accurate models of the process.

One of the other classic issues with building models in biotech is the large impact of relatively small errors in model construction. A summary of Emerson’s research is shown in Figure 7, where we examined the effect of an error of 5%, 10%, and 20% overall magnitude on the net present value (NPV) of a project (Johnston, Zhang, 2009). This study concluded that a 5% change in process throughput decreased NPV by 50%; a 20% change resulted in the project actually having a negative NPV.

This research suggests that model validation is a critical component of model building and should be at least a third of the overall person-hours devoted to a project. Model validation is difficult in biotech since there are so many different centers of excellence that are touched even by simple models. We discuss some approaches that have yielded results in (Yang, Zhang, 2008).
Multi-model / design of experiment analysis

In the final process step, the completed model is used to perform analysis. Analysis typically involves the evaluation of one or more scenarios, or a design of experiment (DOE) space to see a range of interacting variables. If built properly, the toolset can also be used to predict outcomes across a wide range of groups, including process development, supply chain, and manufacturing sciences.

Case study: Facility fit model

Emerson partnered with a large biopharmaceutical manufacturer’s late stage process development and manufacturing sciences organizations to produce a model of their platform process. The process model included a full range of chemical and process calculations and time estimates, but did not include operational data from the facility.

Case study: Facility fit model

Emerson partnered with a large biopharmaceutical manufacturer’s late stage process development and manufacturing sciences organizations to produce a model of their platform process. The process model included a full range of chemical and process calculations and time estimates, but did not include operational data from the facility.

Conclusions

This case study clearly shows the importance of incorporating data into a manufacturing or other process model. Models without real data are simply inaccurate, and generate results that are often dangerously optimistic since they do not incorporate variability or other such ‘real life’ events that affect facility operations. The challenge in this environment is clearly to synthesize data into a framework that allows it to be incorporated into models such as this one.

Biopharmaceutical operations is currently faced with an avalanche of data that overwhelms those looking to perform advanced analysis. It is clear that the data stored by a range of current toolsets like SAP, DeltaV, Excel and other tools data is vast, but is not in a useful format or framework. It is confusing and inefficient to analyze such information, and biotech needs a better way to filter, report, analyze, and learn from the avalanche of data they have today.
The industry is also attempting to move beyond simplistic Excel-based and ‘back of the envelope’ calculations that generate dangerously incorrect results. A true systems-based approach embraces the concept of a ‘single set’ of numbers across the business. This approach uses parametrized data and highly integrated models as the key to taming the deluge of data feeds that threaten to overwhelm decision makers.

Emerson's Crosswalk and Simulation toolsets provide such a modeling platform. By integrating with enterprise data feeds while still remaining flexible enough to integrate with Excel and other planning tools, they provide a way of moving biotech to a truly robust platform for analysis and optimization.