COST DEPLOYMENT WITH PROCESS MINING

Summary

At GlaxoSmithKline (GSK), we use Data Science, Artificial Intelligence (AI) and classical continuous improvement to drive productivity in manufacturing. GSK took quality management data from Enterprise Resource Planning (ERP) systems and combined continuous improvement methodologies with process mining and machine learning to understand how variations of a process affect cost.

The learning from the activity was profound. A resulting model with various algorithms, together, now enable an understanding of the cost of a process instance. Process mining and the use of the Disco software was essential to understand the variation in the process.

We quickly learnt that by taking process mining, applying additional Data Science techniques such as Latent Dirichlet Allocation, Gradient Boosted Trees, and a whole stack of Structured Query Language (SQL), we could transform the way we understand process costs in manufacturing. The model gave an understanding of how variation affects the costs of drugs and medicines.

Furthermore, within Data Science projects, organizational learning is gained on the requirements for digitisation investments, a pathway

- **Process mining case study with new approach for cost deployment in manufacturing**
- **Quality management process improvement of processing time of 202%**
- **Key success factor was the involvement of the SMEs and the initial segmentation of data**

for being Data Science & AI enabled. Running Data Science projects characteristically means some form of feature engineering, fixing data gaps, modelling, and bringing in additional data sources to broaden the model. Learning on how both structured and unstructured data works together benefits in a wider understanding of business performance.

Company

We are a science-led global healthcare company with a special purpose: to help people do more, feel better, live longer.

Our goal is to be one of the world's most innovative, best performing and trusted healthcare companies [1]. Our strategy is to bring differentiated, high-quality and needed healthcare products to as many people as possible. We do this with our three global businesses, the scientific and technical know-how, and our talented people [2].

At GSK one of the performance objectives as described by the CEO is to "remain focused on controlling costs and cash generation…" [3]. *"In 2018, we set out new commitments to build Trust with a strong focus on three principal areas: using our science and technology to address health needs, making our products more affordable and available, and being a modern employer."* [4]

With the three GSK businesses Pharmaceuticals, Vaccines, and Consumer Healthcare we have a wide product range and technologies. For the Cost Deployment with Process Mining project, we focused on the Pharmaceutical business.

Process

The Continuous Improvement function had seen use cases and analysis of process mining projects within Digital, Data and Analytics (DD&A) and presented some challenges to solve, can process mining can give a "Cost Deployment" view by using transactional data? Can we understand if a standard process is performed out of the normative path, is there an affect on cost? What is the differentiation of standardisation? Can process mining identify further waste and non-value-added activities.

Cost Deployment is a method from World Class Manufacturing, where an industrial engineering approach is taken to understand the cost of losses within an organization based on 100% of the cost. To do this, costs are allocated to processes and a distinction is made between value-adding and non-value adding tasks, causal and resultant.

"Cost Deployment is a method from World Class Manufacturing, where an industrial engineering approach is taken to understand the cost of losses within an organization based on 100% of the cost."

-KEVIN JOINSON, DIRECTOR OF DATA SCIENCE & AI COE AT GLAXOSMITHKLINE (GSK)

With the focus on more affordable and available products in GSK, these challenges are aligned with our priorities. With the clarity of understanding process cost, we can draw attention to the right areas that will enable the best result in making products more affordable.

For the pilot of this project, we chose quality management processes. Improvements in quality management can improve product quality, help products to be more affordable, improve services and availability to patients and consumers.

Data

80% of the effort with this Data Science project was related to data preparation and processing. These data preparation tasks included extraction, modelling, storing, transforming data, dealing with missing data, and the dealing with process changes that affected the data composition. Most of the data was taken from our ERP systems. Furthermore, some external data was added to support the model. Transforming the data took the most effort: More than 70 SQL statements were needed to take the raw data from the ERP and transform the data to enabling the analyses to begin.

One challenge for this project was that the drug and medicine product range is broad. Although the processes are the same in the ERP, the cost related to the tasks can be significantly different. Therefore, we needed to apply a distribution model.

Another challenge with ERP transactional data is that there is seldom a ‘start’ timestamp for an activity as well as the ‘end’ timestamp. Usually, a person receives a task, then goes and performs the task (the actual service time), then logs on to the system and executes the task completion activity in a contemporaneous manner. So, only the ‘end’ timestamps are available in the data.

Approach

We took the standard “Cross-industry standard process for data mining” (CRISP-DM) approach to the project, along an agile path to identify value and deliver fast.

Step 1: Business Understanding

We first worked with a single manufacturing site that produced a broad range of medicines and drugs. To understand the necessary data objects, we worked with Subject Matter Experts (SME) of the quality processes. They explained their internal and external business processes. This enabled us to model all aspects of the process cost in the data (see Figure 1 for a schematic, anonymized view).

Step 2: Data Ingestion

We then extracted raw data from ERP by exporting all the fields that were related to the process objects and generated a data schema. It is important to work with the process owners to understand what changes occurred in the process and when the change occurred, this is to

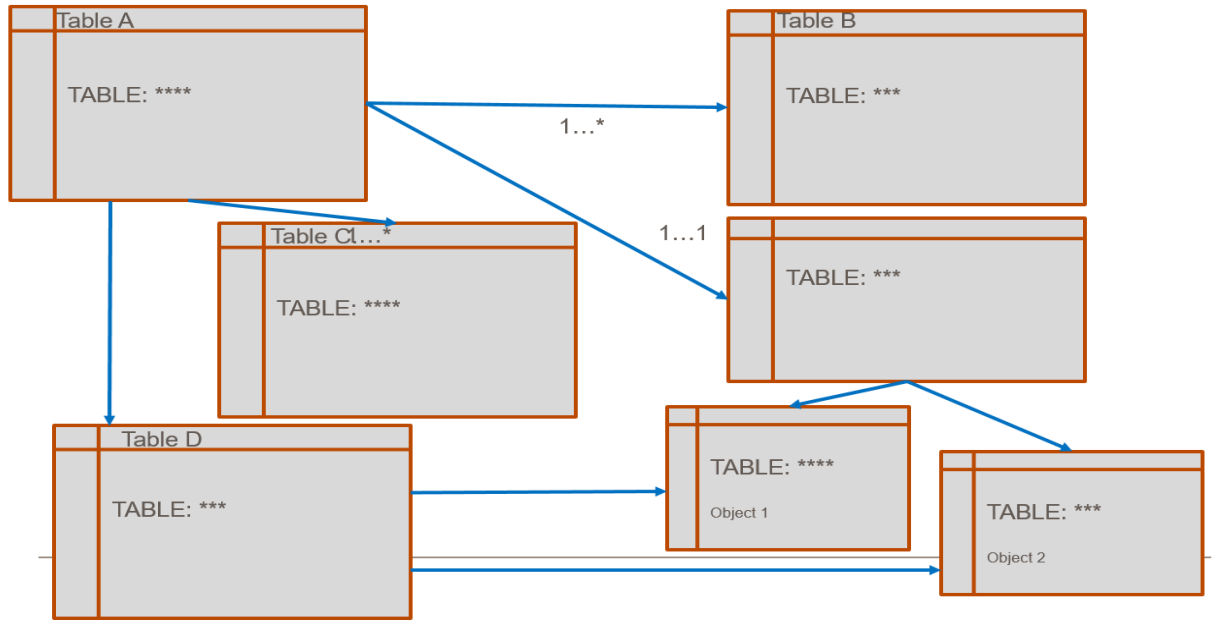

Figure 1: Data models and extraction of tables for cost deployment scope

have consistent data sets. In an example of this, we could see in Disco when a process was simplified by looking at the activities over time.

Step 3: Data Preparation

The data preparation was done in several steps and took a lot of trial and error and an iterative approach in search for a consistent model. The first step was to build an event log with the fields that are required for the process mining analysis: Case-ID, Activity, Resource, and Timestamp (CART). Then we needed to classify the quality management processes, because the origin of the process was widely varied.

One challenge was that some of the activities had an unstructured (“free text”) activity name. This resulted in many different activities in the process map and, therefore, a Spaghetti model view (see Figure 2).

To overcome this challenge, we applied the Latent Dirichlet Allocation a Natural Language Processing algorithm to provide relevant topics from the text sentiment. This method comes with a challenge of the right distribution of alpha and beta. The outcome was a topic classification based on the analysed free text activities. This classification allowed us to distinguish the different types of the quality management processes that related to different topics.

By further applying Gradient Boosted Trees, we could detect that some topics had a very high confidence of their predicted duration and, therefore, we were already able to provide predictions on their end time. Some topics, however, had a low confidence and we looked at these Case-IDs for further analysis and understanding (see Figure 3).

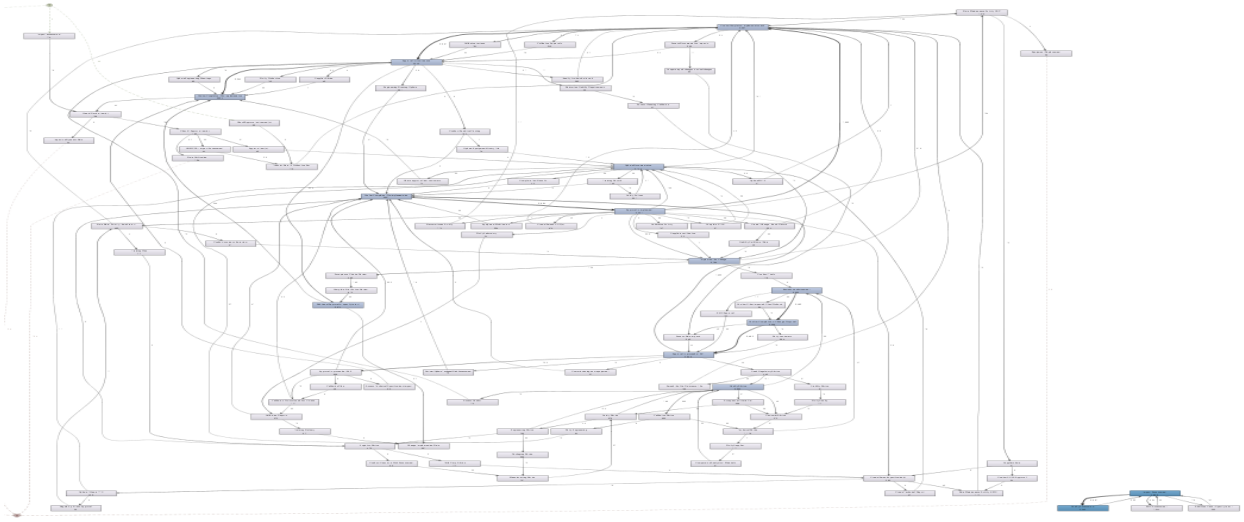

Figure 2: Spaghetti model

Topic id	Concatenate(Term)
topic_17	engineering, utility, interruption, sequence, activities
topic 16	flask, added, process, procedure, split
topic_47	check, range, process, phase, prior

TopicNumber	Confidence
topic_42	1
topic_33	0.999999998
topic_38	0.999999998
topic_8	0.999999991
topic_17	0.999999979
topic_8	0.999999991
topic_33	0.999970846
topic_30	0.999964217

Figure 3: Topic detection and confidence interval

We validated the topic detection with the business process SMEs to ensure that the grouping is logical and well represented. After this validation, we had a respectable link between a CaseID and its classification. We then added these new classifications to the source data and analysed the variants in Disco to understand the process variants of these classes (see Figure 4).

After analysing the process for the different classes, we understood the activities that took place. In many situations, process mining can be performed with just one timestamp, however, in our situation we needed the service time dimension for our cost analysis, so we needed to add this missing data. We went through the classes and activities with the process SMEs and assigned a service time to each task based on their domain knowledge. We then added the missing start timestamps based on this estimated service time to the source data and had now a complete event log that could be used for our process mining analysis in the Disco software.

Case ID	Events	Variant	St
[redacted]1	33	Variant 27	
[redacted]2	29	Variant 28	
[redacted]8	20	Variant 29	
[redacted]9	20	Variant 30	
[redacted]5	18	Variant 31	
[redacted]0	18	Variant 32	
[redacted]3	18	Variant 33	
[redacted]2	10	Variant 11	
[redacted]0	10	Variant 12	
[redacted]3	10	Variant 34	
[redacted]8	10	Variant 35	
[redacted]1	10	Variant 11	
[redacted]6	29	Variant 36	
[redacted]9	18	Variant 37	
[redacted]9	10	Variant 12	
[redacted]2	18	Variant 38	
[redacted]2	18	Variant 39	
[redacted]3	20	Variant 40	
[redacted]1	18	Variant 41	
[redacted]3	18	Variant 10	
[redacted]5	18	Variant 10	
[redacted]0	18	Variant 42	
[redacted]1	18	Variant 43	
[redacted]7	10	Variant 44	
[redacted]4	26	Variant 45	
[redacted]3	31	Variant 46	

Figure 4: Classes of the processes and related variant

Step 4: Modelling

The data was now ready to be modelled towards the objective of process variation. We exported the process information from Disco and loaded it into the database. We then began to shape the cost of the process classes and variants. This modelling was no mean feat: More than seventy SQL statements were needed to provide the correct view of the cost. Part of the modelling was to bring in external financial data, which acts as the cost drivers. This model also meant that an event log was coded in the database, which enabled the data to be refreshed to provide a new log for Disco (see Figure 5).

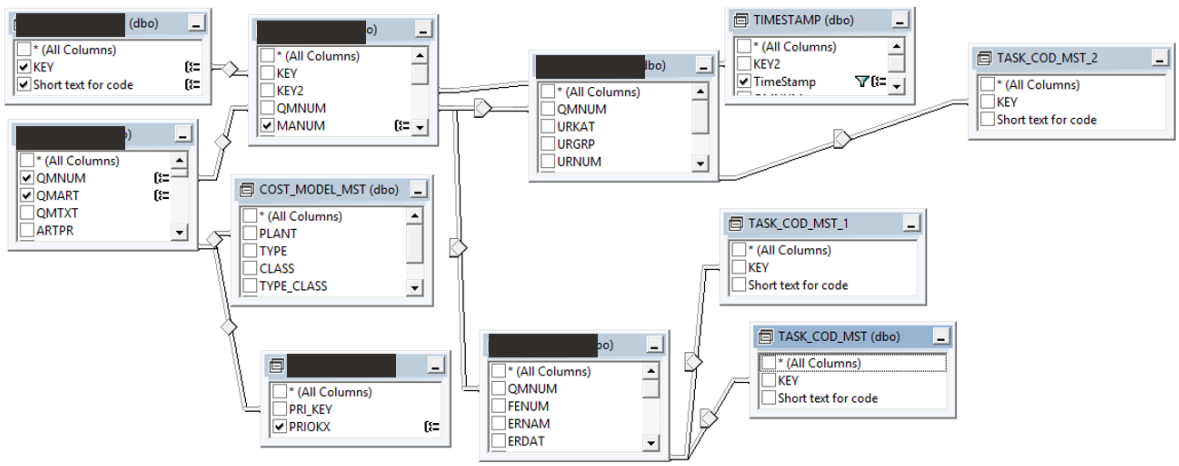

Figure 5: SQL event log for Disco

Step 5: Evaluation

To evaluate the results, we built a research pack with facts and data for the processes. We then took the business process SMEs through the model to understand whether the results aligned with reality. With some further validation, and a few cycles of model adjustments, we had a low standard error margin model. Additional insights on cost, time, and "what ifs" were modelled over time.

"With the same factors being applied across all pharmaceutical sites, we now have a global view of cost opportunities."

-KEVIN JOINSON, DIRECTOR OF DATA SCIENCE & AI COE AT GLAXOSMITHKLINE (GSK)

Step 6: Deployment

The validated model was then industrialised. With the same factors being applied across all pharmaceutical sites, we now have a global view of cost opportunities. Manufacturing sites take advantage of the information and insights from the model. They use Disco to highlight the areas for improvement. Each month the data is refreshed in a few clicks.

Overall Process of Transformation

The overall process of the data transformation is shown below (see Figure 6). As described before, the collaboration with the SMEs was essential in multiple phases: For the initial understanding, the validation of the classified topics, the estimation of service times, and for the validation of the overall results.

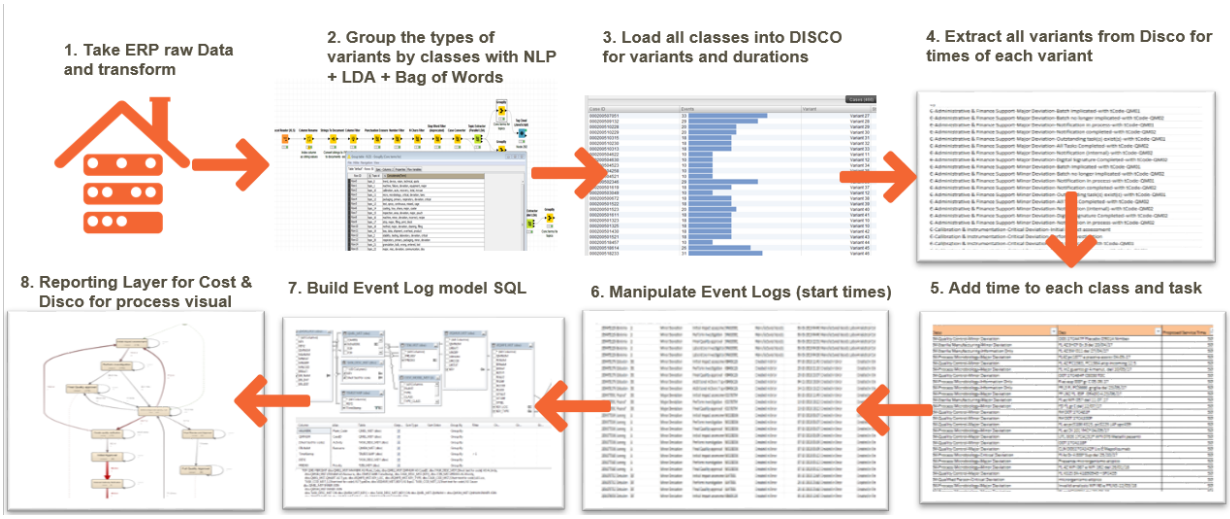

Figure 6: Steps to modelling process costs with Data Science

Impact

Through this analysis, we supported a focused improvement initiative that achieved an improvement of 37% from the baseline performance (see Figure 7). Our method is now being used in the broader manufacturing network across various processes.

Item	Improvement
Overall Improvement	37%
A->B	27%
A->C	42%
A->E	Path Removed
B->D	8%
C->B	29%
D->A	Path removed

Figure 7: Table of improvement summary

For cost reductions it is essential to remove waste from the processes. Process mining is a vital tool in supporting these optimization as shown in this article.

But we go even further, GSK is industrialising Artificial Intelligence to the remaining process duration and uses Deep Learning and Semantic Textual Similarities to support quality professionals. The removal of waste from the process paves the way for the subsequent use of advanced Data Science techniques such as Machine Learning.

Conclusion

Process mining with other data science techniques can use ERP data to understand costs of processes and variations, the challenge is how to bring together both structured and unstructured data to re-use valuable data that is often only used for a primary purpose. With this re-use of data, we were able to better understand how to drive performance, whilst at the same time transforming continuous improvement to be further data driven and empirical.

Author

Kevin Joinson – Director of Data Science & AI CoE at GlaxoSmithKline (GSK)

References

[1] Anon. GSK Annual Report 2018 https://www.gsk.com/en-gb/about-us/ Access online June 2019

[2] Anon. GSK Annual Report 2018 https://www.gsk.com/media/5349/annual-report-2018.pdf Page 2. Access online June 2019

[3] Emma Walmsley. GSK Annual Report 2018. https://www.gsk.com/media/5349/annual-report-2018.pdf Page 3. Access online June 2019

[4] Emma Walmsley. GSK Annual Report 2018. https://www.gsk.com/media/5349/annual-report-2018.pdf Page 4. Access online June 2019