Journal of Business Chemistry

Practitioner's Section

Successful data applications: a cross-industry approach for conceptual planning

Robert Jenke*

* Dr. Jenke Consulting, Process Automation & Data Science, www.jenke-consulting.de, mail@jenke-consulting.de

DOI: 10.17879/38129709602; URN: nbn:de:hbz:6-38129709797

Data-based solutions powered by artificial intelligence (AI), and especially its subdomain of machine learning, are a key driver of today's fast-paced technological evolution. In the process industry, barriers for many organizations to apply this technology are missing know-how for conceptual planning as well as lack of economic feasibility studies. However, companies risk to lose their competitive position by not applying this technology. In this article, we describe the CRISP-DM model as a conceptual planning approach. In addition, we provide practical advice based on experience in other industries how this technology can be applied in the process industry. Here, the steps process analysis and data understanding are key success factors in order to develop economically viable use cases. An implementation strategy should include an agile environment to develop ideas fast and with little risk before transferring working solutions to the requirements of the operational business. A combined bottom-up / top-down approach of knowledge distribution and pilot projects can help organizations to successfully embrace this technology in their operational businesses, overcome associated fears and organically seize company-individual opportunities that arise.

1 Introduction

The fast-paced technological evolution of our time leads to equally fast changes in the business world, from efficiency gains through process improvements to the redefinition of long-established business models. In this context, a key technology, which accelerates this technological evolution, is artificial intelligence (AI) including its subdomain of machine learning. Machine Learning methods allow the training of models from data examples instead of tediously having to define the input-output relationship of the models manually. For instance, such models can be applied to make decisions and predictions. While this technology has been cost- and time-consuming in the past, it has now reached a level of maturity where broad industrial application is feasible in many scenarios (VanThienen, 2016).

As AI constitutes a universal technology – like the steam engine or electricity in the past – it is bound to influence many sectors of industry well beyond its origin (Brynjolfsson, 2017). Consequently, also the process industry is already being influenced by this trend and therefore evaluates how to harness the arising opportunities. For example, predictive asset management has been adopted by major players in order to maximize asset utilization or minimize unplanned downtime. Furthermore, the monitoring of a plant can be made more efficient by the digitization of a plant's control data. Additionally, monitoring of production processes can be enhanced with pattern recognition in order to assess influences on batch consistency or to predict deviations from the production process before they could occur. Finally, downstream data from points-of-sale can be used to forecast the demand of customers and to plan a responsive schedule in production (VanThienen, 2016).

However, at the same time several barriers are

currently hampering the use of digital technologies like machine learning, as Stoffels and Ziemer point out in an article of an earlier issue of this journal (Stoffels and Ziemer, 2017): Among those barriers are

1) Unclear benefits and/or lack of economic evaluations and
2) Missing know-how on methods for analyzing and adapting processes.

Yet, it is imperative for companies to understand the impact that this important technological change has on their business model, their operations and their competitive landscape. As with any new technology, identifying and defining use cases is not straight forward. At the same time, they need to fulfill at least two preconditions:

1) the application must be technologically feasible and
2) backed by a solid business case.

If these preconditions are not fulfilled in the beginning, many companies eschew investments due to risks involved. However, there is also considerable risk associated with ignoring the changes ahead and thereby risk becoming the next Kodak (Anthony, 2016).

In this article, we aim to make the following contributions:

1) We present methods and approaches concerning data applications that have been successfully applied to other industries. More importantly, this contribution shows what can be learned and transferred to the process industry. In this, we introduce the cross-industry standard process for data mining (CRISP-DM) model, which is a generic framework how to approach data mining problems. We discuss key issues in depth based on experience from other industries.
2) We discuss possibilities to overcome the previously mentioned risks and barriers, i.e. implementation strategy and integration into the organization on a strategic level, in order to increase the overall ROI of investments in this field.

2 CRISP-DM model

The CRISP-DM describes a generic framework for data mining and knowledge discovery projects. It is used by the majority of experts in the field of Data Science (Marban, 2009). The approach captures the essence of what is needed to successfully carry out a data mining project, but at the same time it is easily transferable to any industry. The six steps of the CRISP-DM are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Shearer, 2000). It is worth noting that the steps are not necessarily carried out in a linear manner, as depicted in Figure 1. The single steps are explained below (for details, see (Shearer, 2000)):

1. Business understanding

Before touching any data, it is very important to understand the business context in which this data will be used. This phase concerns itself with questions around the business objectives, success criteria, relevant business processes, and an assessment of the overall situation, in order to derive goals and develop a first plan of the project. The better this step is carried out, the more valuable insights and outcome can be expected from the remaining steps.

2. Data understanding

In the second step, data is collected and an understanding is developed. This includes describing the data and what kind of information it contains. Also, quality of data should be assessed, e.g. consistency and completeness. First hypothesis can already be derived from data. Often, it is necessary to go back to step one multiple times, until a clear picture of the interaction between business processes, objectives, and data is gained.

3. Data preparation

Once a specific use case is defined, data must be selected based on the relevance to the project's goal as well as the technical limitations. Here, the data must be usable. Data sets must be provided, which can be used to develop models. Quality can be increased, for example, by removing outliers, handling missing values, and giving a proper structure to the data.

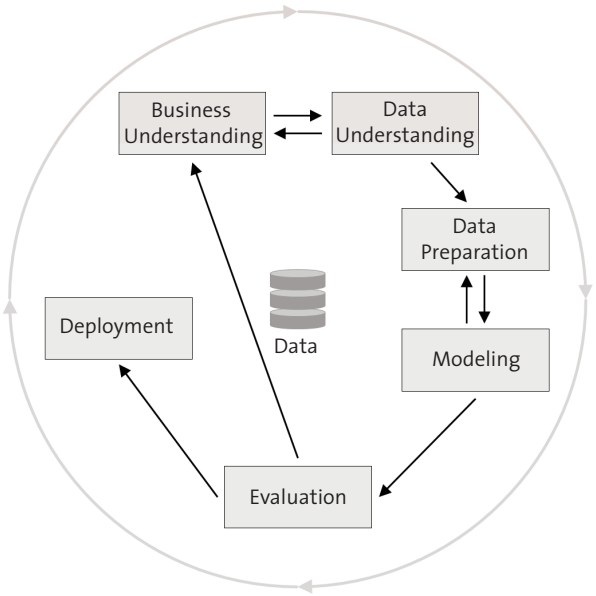
4. Modeling

In this phase, tools like machine learning methods are employed to train one or several models on the selected data. Usually, more than one method is available. This phase aims to identify the best fit and optimize any free parameters. This also includes testing the quality of the model, e.g. its generalizability measured by the error rate on test data.

5. Evaluation

While the model itself is verified in the fourth step, this step evaluates the suitability of the developed processing pipeline and model with respect to business application. Only when there are no critical issues overlooked, the model can be deployed in the next step. Too often, false assumptions require

Figure 1 CRISP-DM model (source: Shearer, 2000).



going back to step 1 and to revise the business understanding.

6. Deployment

When the suitability for real-life business application has been shown, the deployment phase aims to transfer the findings from the data mining project to day-to-day operations, i.e. actually make use of the created models. In some cases, this might be a simple report, in others the implementation of software to track and analyze real-time data in order to support the decision-making process of the organization.

In Data Mining more than anywhere else, a famous quote, that may or may not be attributed to Albert Einstein, can be literally applied for good results: "If I had one hour to save the world, I would spend 55 minutes understanding the problem and 5 minutes trying to find a solution." Challenges frequently arise in the definition of use cases, i.e. truly understanding the problem. Therefore, understanding business and understanding data need to be carried out thoroughly and will have to go back and forth. We collected learnings from experience gathered in other industries and summarize these in the following together with some practical examples.

2.1 Process modeling and analysis

The step of understanding business is very much about understanding processes, i.e. identify tasks and steps that data can improve, enhance or carry out more efficiently and reliably. Therefore, modeling tools like Business Process Modeling Notation 2.0 (BPMN 2.0 (OMG, 2013)) are helpful to model processes and develop a common understanding. Although the BPMN 2.0 notation consists of a large number of available elements, a handful of the basic elements are already sufficient to significantly increase process understanding.
Approaches to generate use cases can be versatile. A high-level approach is to focus on areas of the value chain that have an impact on operating and growing the business, as van Thienen et al. suggest (VanThienen, 2016). On a more practical level, it should be looked out for the following cues to identify high-potential areas for data and/or AI usage when analyzing processes:

1. High-volume tasks:
Naturally, task that are frequently carried out often provide a greater leverage for potential improvements. In these cases, even small time saving measures or quality improvements can lead to a viable business case for applying machine learning.

2. High-value decisions:
Similarly, decisions with a high inherent potential to influence large areas of the organization are promising candidates for intensive data-usage. In these cases, additional efforts may be well justified to make the right business decision.

3. **Media discontinuity**:
Despite increasing efforts to digitize, many processes still involve media discontinuities, e.g. printing of documents or form, switching from online workflows to phone calls or emails. While sometimes this is obviously not avoidable, e.g. sending a hardware product to the end-customer, steps within the process that go back and forth between digital and analog workflows are good candidates for applying automation.

4. **Time consuming steps:**
A strength of AI methods is to process large amounts of data fast without any errors. This often complements human abilities and can lead to significant time savings. Thus, it is worth focusing on the annoying and tedious tasks.

5. **Bottlenecks:**
Restructuring processes around bottlenecks can increase overall efficiency. A solution can be parallelization of sub-tasks carried out simultaneously by a machine in the background.

As with any new product or service, it is important to take on a "Customer centric" mind-set and develop the solution this way. Here, the customer could also be an internal one. In the following, we like to illustrate this approach with an example from practice.

Example: In one project, we analyzed the internal processes of the sales department. One of the planning steps consumed half a day on average and was carried out regularly. The organization was growing substantially, so that the process needed to be rethought for scalability. In a workshop, we modeled the current state of the process and analyzed possibilities for restructuring including the automated processing of data. We were able to eliminate all media discontinuities by fully digitizing the planning process through a web-application. Large parts of the data were then gathered and processed automatically. Overall, this led to significant time savings of up to 80% and further increased quality and transparency of this process.

2.2 Data understanding and preparation

The second important step is to understand and prepare the available data. Both technical and economic feasibility of a use case depend strongly on the data available, which makes use cases highly individual according to a company's data situation. Therefore, several topics need to be taken into account when aiming to use data successfully. Put in simple terms, the following factors influence the potential of your data and how well you can build models from it. They can be considered as levels of data needs building on each other. Each level requires a certain maturity of the previous level:

1. **Data sources and types**
First, where does your data come from and since when are you recording it? Is it a structured SQL-Database, a sensor-stream, a manually filled spreadsheet-file, or something else? Data can be structured or unstructured. For example, a free text document like a project report or an invoice in pdf-format is considered unstructured data. If you have a form like a contact form on a website, then each field (e.g. name or email address) has a particular meaning. This represents structured data. Unstructured data is much harder to process and requires more steps to prepare the data.
Data types regard to the kind of data. For example, this can be text, numerical, or categorical. The latter refers to a fixed set of values that the data can take on, e.g. 'True' or 'False' or colors of a product 'red', 'green', or 'blue'.

2. **Data quality**
Second, it is important to assess and understand the quality of the data available. In particular, the following questions are of importance: Are the values complete over the whole data set or are there missing values? How should they be handled and how does this affect the quality of the application? Can the root problem be fixed or can the data be processed even when values are missing? Are outliners present in the data? If so, the data might need to be cleansed. Many of these rather technical issues can also point to issues in the process setup itself.

3. **Data content**
Third, it is important to understand what information does this data represent? With numerical data, descriptive statistics can be used to describe the data (i.e. data distribution, histograms, etc.). Although this step may seem tedious and without a clear goal at first, every minute spent is well invested. It is a prerequisite to understand the data available very well. Then, a judgment of the models concerning their suitability and application is possible at later stage of the process. When digging into the data, ideas for potential use cases can be generated. In most cases, insights about operational and organizational processes can be gained, which are not expected at first.

Example: At one of our customers, production controlling data was recorded in both Enterprise Resource Planning (ERP) system and Manufacturing Execution System (MES), but was not intensively utilized. Before creating use cases to apply AI, the available data was inspected for its types, quality

and content. This exercise proved to be highly educational for everyone involved. In fact, untapped information was identified, and a new detailed report personalized for each foreman was designed and introduced as a result. This example illustrates, that effective solutions to enhance data usage can also be quite simple.

2.3 Modeling and evaluation

In the data modeling step, the actual building of models using machine learning methods takes place. Models are a representation of relationships. As shown in Figure 2, they map input data to an output. In the past, models were often created manually. For example, if-than-that-rules can make up the relationship between input and output. The strength of machine learning methods is that they can extract relevant information like rules from data automatically. By doing this, the model can be learned from existing data, i.e. input-output examples from the past. Against common belief, this step consumes only a relatively small part of the time – thanks to the versatile tools like software packages available today. These tools facilitate the implementation of machine learning methods to train and evaluate models.
Machine learning methods can be summarized in three categories:

1) Supervised Learning, where the model output is provided to the machine learning method for training. For example, Classification, Regression, and Feature Subset Selection methods belong to this category.
2) Unsupervised Learning, where the model output is unknown. That means, the methods have to find structure and make connections by themselves. Clustering and Dimensionality Reduction are examples here.
3) Reinforcement Learning, where the model is trained independent of the output. Instead a trained rule is reinforced by rewarding / punishing depending on whether the output was right / wrong. Artificial Neural Networks and Deep Learning can be assigned to this category.

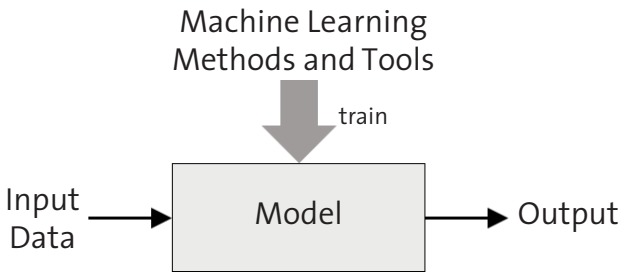
For a given problem, the model can usually be created by means of several methods. The challenge for data scientists is to pick the appropriated method (or a combination of methods) based on their experience in order to develop the best model. Models are trained based on examples. Therefore, it is important to consider the generalization of the model very well, so that unseen examples are also handled well by the model. Thus, the quality of the model is judged based on a so-called "test-set" of data that is put aside in the beginning of the modeling phase. The following example illustrates how pattern recognition can help to improve processes of a plant.

Example: A plant's control data from the past contains information about process deviations and when they occurred. Pattern recognition can be applied to automatically identify such deviations before they occur. This task can be framed as a classification problem with two classes: "deviation" or "no deviation". From past examples of plant operations, relevant data features and rules to detect such deviations are extracted using machine learning methods from supervised learning. Before the final model containing these rules is deployed, its fit is evaluated using a second set of examples from the past.

2.4 Implementation strategy

The understanding of both process and data provides two important results. First, the technical feasibility as well as the potential benefits of a use case can be judged. Second, the effort how to implement the respective use case can be roughly estimated. This provides an indicator of economic fea-

Figure 2 Machine learning methods allow to train models from data examples (source: own representation).



sibility, which significantly lowers the risk of investing disproportionate budgets. However, a considerable amount of ideas might still be disregarded at a later stage due to factors that cannot be assessed at this point. In such cases several key factors help to reach a good ROI. In particular, methods can be used that are already widely applied in software development and the lean start-up approach, which focus on agile and iterative developments (Ries, 2011). Risks of failure can be reduced by allowing multiple possible directions. Then, adjustments can be made based on the learning made within the implementation process. Organizations should start with prototypes at a very early stage to generate this learning. Tackle assumptions with highest risks first, i.e. fail fast to learn fast! This, of course, requires a culture that is open-minded towards failure.

Meanwhile, it is important to keep a clear focus. A strong and rigorous selection funnel is necessary to weed out directions, which are not promising. Only projects should be continued that have proved to be successful by fulfilling the defined evaluation criteria.

The CRISP-DM model aligns very well with this lean approach. This approach will be illustrated by the example below.

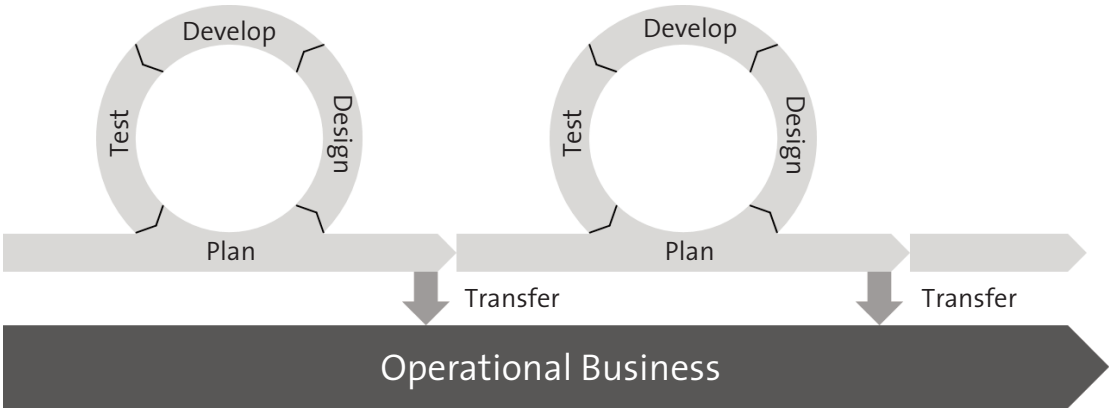
Example: One company holds monthly meetings, in which employees participate that are not involved in the actual project – similar to a "supervisory board". Learning from the project regarding what works and what does not work as well as new ideas are presented to them. The group then judges the value and feasibility of the projects and takes a decision to terminate directions that do not show enough potential with respect to the company's competitive position.

At first, these elements may seem difficult or sometimes impossible to incorporate in the development processes of companies from the process industry, which are traditionally rather conservative and have strict requirements in their quality systems. A method to solve this is depicted in Figure 3 – similar to Google's so-called moonshots[1]: This should be a parallel and isolated track to the operational business. In doing this, it provides a way to develop solutions to a level of maturity without a costly overhead and operational risks. Only when a solution has proven to provide benefits it is transferred to the operational business. This approach greatly decreases costs and risks associated with such endeavors. If a company decides to not establish such a department internally because of its size or budget, an outsourcing of such services to external experts can be considered.

2.5 Integration into organization

For the long-term success, it is vital to find a sustainable way as an organization to embrace the AI technology and to embed it into the organization's strategy. The subsequent adaptation of business processes also brings about change to the employees, which play an important role concerning the integration of such technologies. The human factor and the topic of change are easily overlooked. Thus, human factors such as their behavior must be especially considered. Several fears are associated with AI as the public discourse discloses. For example, its impact on job security and job descriptions. These fears must also be addressed within

Figure 3 Development of innovative applications in a parallel environment (source: own representation).



1 Moonshot refers to projects that are similarly ambitious and unrealistic as landing on the moon seemed back in the early 1960's, but that would constitute a significant advance.

the organization. One way to move forward is to follow a combined bottom-up and top-down approach:

Bottom-up: A key factor to get support from the workforce for technological change is the distribution of facts and knowledge. Employees should be encouraged to identify use cases by themselves since they know their workflows and tasks better than anyone else. This includes the exhausting and repetitive processes, which could be promising candidates for automation. The ideas of employees become a valuable source for process improvement, when they are equipped with the right knowledge about the technological capabilities and limitations. At the same time, they can be mobilized for the change and a momentum can be created, if they are included at the beginning of the transformation process.

Top-down: In parallel, pilot projects initiated by the management serve as vivid examples in a familiar environment help to spread a realistic and clear picture of what the organization's individual needs are and showcase success stories as a generator for new ideas.

3 Summary and conclusions

Advanced and complex technologies like AI require knowledge and a well-structured approach for conceptual planning and implementation. With the CRISP-DM model, an established framework is available, which can be used to guide projects that aim at the utilization of data in order to get valuable insights. In this process, challenges frequently arise regarding the definition of use cases. In this article, insights from practical experience offer approaches how to overcome these challenges.
For the conduction of projects, modern and agile development methods such as the lean start-up approach should be used. Strict evaluation criteria concerning the decision which projects should be further pursued must be in place. Implemented correctly, this will ensure a high ROI across these efforts. Further, a holistic strategy for integrating this change into the organization is necessary to seize opportunities that arise organically.

Many companies are still reluctant to tackle the technological shift towards intensified data usage and artificial intelligence – they remain in a waiting position. Yet, it has to be considered whether the energy and resources spent in observing the market might not be better invested in gaining first-hand experience by diving into the topic. The first option leads to an inevitable time delay. Thus, the second option should be more appealing to companies. Especially, since both feasibility and economic viability are two company-individual factors. First practical experiences can be collected by just diving into the topic and testing assumptions.

References

Anthony, S. (2016): *Kodak's Downfall wasn't about Technology.* Harvard Business Review, available at https://hbr.org/2016/07/kodaks-downfall-wasnt-about-technology, accessed 10 July 2016.

Brynjolfsson, E., McAfee, A. (2017): The Business of Artificial Intelligence, *Harvard Business Manager*, pp. 22-34.

Marbán, Ó., Mariscal, G., Segovia, J. (2016): A Data Mining & Knowledge Discovery Process Model, in Julia, P., Adem, K. (ed), *Data Mining and Knowledge Discovery in Real Life Applications*, pp. 438-453.

Shearer, C.(2000): The CRISP-DM Model: The New Blueprint for Data Mining, *Journal of Data Warehousing*, vol. **5** (4), pp. 13-22.

Stoffels, M., Ziemer, C. (2017): Digitalization in the process industries – Evidence from the German water industry, Journal of Business Chemistry, vol **14** (3), pp. 94-105.

Object Management Group (2013): *Business Process Model and Notation BPMN 2.0*, ISO/IEC 19510.

Ries, E. (2011): *Creating the lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, Inc Magazine, available at https://www.inc.com/magazine/201110eric-ries-usability-testing-product-development.html, accessed 15 October 2011.

van Thienen, S., Clinton, A., Mahto, M., Sniderman, B. (2016): *Industry 4.0 and the chemicals industry: Catalyzing transformation through operations improvement and business growth*, Deloitte University Press.