DATA SCIENCE INITIATIVE

Manual for Data Science Projects

Victor Pereboom
CTO at Dutch Analytics

Sascha van Weerdenburg
Machine Learning Engineer at Dutch Analytics

This manual for AI projects within the government has been developed in cooperation with and based on the AI project from Ariea Vermeulen and Arie van Kersen from Rijkswaterstaat (the Ministry of Infrastructure and Water Management). In 2019, they have started with a project to research whether the combination of drone images and machine learning could lead to a better inspection of bridges and infrastructure.

AUTHORS

Victor Pereboom
CTO at Dutch Analytics

Sascha van Weerdenburg
Machine Learning Engineer at
Dutch Analytics

ABOUT DUTCH ANALYTICS

The software platform of Dutch Analytics, called Xenia, supports enterprises to operationalize data science models. Xenia also manages and monitors Artificial Intelligence (AI) algorithms after development and deployment. The idea behind the software originates from the experience that many companies struggle to operationalize the results of data science projects. Many algorithms are never put into use after the first proof-of-concept, which wastes time, effort and money. But more importantly: missed value. By using Xenia, companies can easily take the step from data science model to successful scalable end product with related business revenues.

More information about Dutch Analytics and Xenia at
www.dutchanalytics.com

EDITORS

Victor Pereboom
CTO at Dutch Analytics

Sascha van Weerdenburg
Machine Learning Engineer at
Dutch Analytics

Ariea Vermeulen
Rijkswaterstaat

Arie van Kersen
Rijkswaterstaat

Francien Horrevorts
Rijkswaterstaat, Fran&Vrij Communicatie

Koen Hartog
DSI

Marloes Pomp
DSI

In the past five years Data Science has proven itself as a domain with major socialimpact.

We see increasingly more applications appearing in our society: modern smartphones are equipped withimage and speech recognition technology. Self-driving cars are no longer part of only fictional filmscenarios. These developments are not only at the local level. Companies worldwide are investing ininnovative technologies to develop data-driven solutions and new products.

However, only a small percentage of these initiatives grows into a fully-fledged solution. Surprisingly, if you look at how much effort is invested to make these projects successful. And that raises some follow-up questions.

What are the factors of a successful data science project? In which steps do you go from an idea to a good solution?

This white paper aims to provide more insight into the life cycle of data science projects and is based on experiences gained during the development of various complex data-driven solutions.

What is Data Science?

> *So what is exactly data science? How does it relate to AI, for instance Machine Learning and DeepLearning? We have put all the answers together for you.*

Data Science

Data science is an area where data is examined for patterns and characteristics. This includes a combination of methods and techniques from mathematics, statistics and computer science. Visualization techniques are frequently used in order to make the data understandable. The focus is on the understanding and usage of the data with the aim of obtaining insights which can further contribute to the organization.

Artificial Intelligence

Artificial Intelligence is a discipline which enables computers to mimic human behaviour and intelligence.

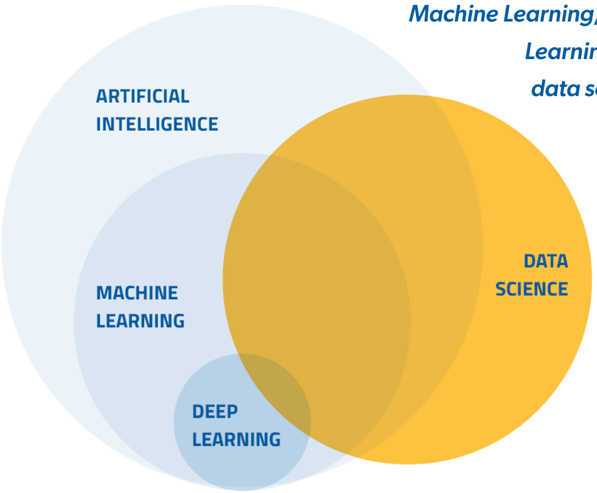
Machine Learning

Machine Learning is part of Artificial Intelligence which focuses on 'learning'. Algorithms and statistical models are being designed for autonomous learning of tasks from data, without giving explicit instructions upfront.

Deep Learning

Deep Learning is part of Machine Learning that uses artificial neural networks. This type of model is inspired by the structure and function of the human brain.

Visualization of the difference between Artificial Intelligence (AI), Machine Learning, Deep Learning and data science



Types of Machine Learning and Deep Learning

There are different types of Machine Learning approaches voor different types of problems. The three most known are:

Supervised Learning

Supervised Learning is the most widely used type of Machine Learning in data science. While inspecting damages on concrete material, photos of previous examples are being used of which it is known whether they show damage or not. Each of these photos has been given a label: 'damaged' or 'not damaged', which helps in further classification.

Unsupervised Learning

In Unsupervised Learning labels are not used, but the model itself tries to discover relations in the data. This is mainly used for grouping (clustering) the examples from the data. For example, creating different customer groups where customers with similar characteristics are in the same group. Beforehand it is unknown, which groups of customers there are and which characteristics they meet, but the algorithm can distinguish the different groups with enough data available.

Reinforcement Learning

Finally, a Reinforcement Learning model learns on the basis of trial and error. By rewarding good choices and punishing bad ones, the model learns to recognize patterns. This technique is mainly used in understanding (computer) games (such as Go) and in robotics (a robot which learns to walk through falling and standing up again). This type of Machine Learning usually falls outside of data science, because the purpose of 'learning' a task is the goal and not understanding and using the underlying data.

Life cycle of a Data Science project

> *Now that the different terms have been explained, we focus on the data science projects. What do you need to pay attention to in such projects, what does it require and which best practices and learnings can we provide you? We start with a little more background about the life cycle of these projects. The life cycle of a data science project consists of two phases, which contains 7 steps.*

FOCUS ON THE BEST MODEL: FROM BUSINESS CASE TO PROOF-OF-CONCEPT

In this phase the focus is on the development of the best model for the specific business case. This is the reason why the definition of a good business case is essential. The data science will then work towards a working prototype (proof-of-concept).

The first phase consists of 4 steps:

1. A good business case
2. Obtain the correct data
3. Clean and explore the data
4. Development and evaluation of models

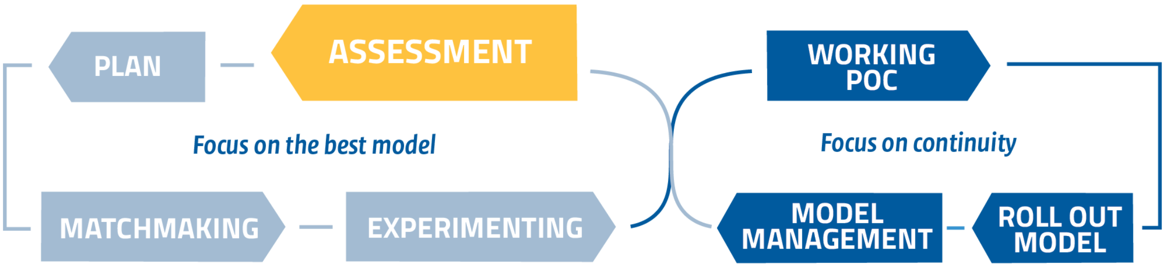
FOCUS ON CONTINUITY: FROM PROOF-OF-CONCEPT TO SCALABLE END PRODUCT

In the second phase the focus is on continuity and a working prototype will be developed to an operational end product.

This phase consists of 3 steps:

5. From proof-of-concept to implementation
6. Manage model in operation
7. From model management to business case

Together, these 7 steps (in two phases) form the life cycle of a data science project



How to effectively tackle this process as an organisation? In the next chapters we detail the 7 steps of the life cycle and we have put all things together which you need to pay attention to. Every chapter will close off with a short summary.

FOCUS ON THE BEST MODEL: FROM BUSINESS CASE TO PROOF-OF-CONCEPT

Step 1: A good business case

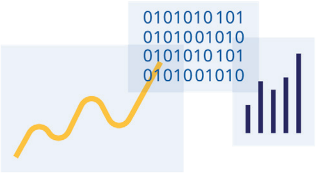
The ultimate value of a data science project is dependent on the clear business case. What do you want to achieve with the project? What is the added value for the organisation and how will the information from an algorithm eventually be used? Here are some guidelines for defining and testing the business case.

Involve end users

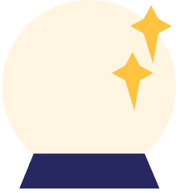
A business case is generally strong when it comes from practice and from end-users / domain experts, as they are usually the people who have to use and rely on the information from the models. It is therefore important that they especially understand the added value of the data science solution. Where does the need of the user lie and how can the end product create value for the user?

Does AI give the best solution for the problem?

After defining the business case, it is wise to assess whether AI, in the form of self-learning algorithms, is the best solution for the problem. AI is very suitable for finding patterns in data which is too large and complex for people to oversee. AI has proven itself valuable in a number of areas, such as image and speech recognition. AI helps to automate these tasks, however in some cases, AI is less useful:

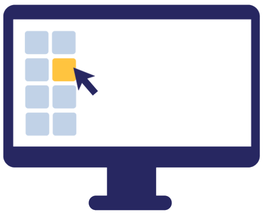


AI models learn from (large amounts of) data. If little relevant data is available or if necessary contextual information is missing in the data, it is better to not use an AI solution.



AI is not good in unpredictable situations which require creativity and intuition to solve them.

This also applies if transparency of the algorithm is of great importance, because models (in particular Deep Learning) are often difficult to understand why they show certain behaviour / give certain results.



There may simply be more effective and/or cheaper solutions than data science solutions for a specific business case, such as traditional softwares.

NAME STAKEHOLDERS

To make the data science project a success, three stakeholders are required:

THE MANAGER

(someone from company policy/strategy)
The manager is responsible for checking the added value of the business case.

AN END-USER / EXECUTIVE EXPERT

(someone from operations)
The executive expert is responsible for end-users acceptance. Without a clear point of reference and support for the end-users, it is likely that a data science model will not be adopted in practice.

SOMEONE FROM THE DATA SCIENCE TEAM

The data scientist is responsible for assessing the project's probability of success, given the potential solution and available data.

These three stakeholders must be aligned in advance about the objectives of the end product. When there is no internal data science team available, one can also choose to outsource the project. In that case, the manager is the coordination point between end-user/executive experts internally and the external data science team.

STEP 1 SUMMARIZED

- Define the business case together with the executive experts.
    - *What is the need?*
    - *What are the requirements that the solution must offer to be able to generate significant added value for the company?*
- Assess whether applying AI is the best solution for the business case defined.
- Appoint the relevant stakeholders: from the perspective of policy and strategy (the manager), the implementation (the end user) and the data scientist.

Step 2: Obtain the correct data

Developing an AI algorithm often requires large amounts of high-quality data, since a flexible model will extract its intelligence from the information contained in the data. How do you obtain this good data and information?

Make sure that data is available

Before a data science team can start building a model, the correct data must be available. Data science teams depend on data engineers and database administrators, since these employees have the appropriate permissions to access the required data. Executive experts/end-users do not always know how the data is stored in the databases. Finding the required data is an interaction between the executive experts, the data science team and the data engineers and/or database administrators. It is also possible that the required data is not available within the organisation. In that situation one can choose to collect data from external sources, if available.

Data Dump for the first phase

In the first phase of the data science lifecycle, a one-off collected data dump is usually sufficient.

This will partly be used to "train" the model and the other part to "validate the model".

When multiple data science teams compete for the same business case, all teams should have the same data set. This is to ensure that all teams have an equal chance of finding the best model. The performance of the various models can be compared.

It is wise to separate an extra 'test dataset', which has not been shared with the data science teams beforehand. Based on the predictions of each model on this test data set, it can then be compared which model performs best. A good example of this is how the platform ***Kaggle*** operates, which organises public competitions for data science teams on behalf of companies.

Automatically retrieving data in the second phase

> *In the second phase of the data science lifecycle, once a model becomes operational, a one-off data dump is no longer sufficient.*

The data must then automatically reach the model for computing. In practice, this is often rather complicated, because there are so-called data silos: closed databases which are difficult to integrate into an application. This is due to the fact because many systems are not designed to easily communicate with each other. An internal IT security measure makes communication more difficult. That is why it is recommended to think about the second phase already in the first phase.

Start setting up infrastructure in time

Investing in a good infrastructure for the storage and exchange of data is essential for a data science project to become a success. A data engineer can facilitate this process, where a robust infrastructure is set up. Start with this process early and keep security, access rights and protection of personal data in mind.

STEP 2 SUMMARIZED

- Make sure all data is available to data science teams in cooperation with data engineers.
    - *For the first phase, a single data dump will be sufficient.*
    - *For the second phase, the data have to be retrieved automatically and fed into the model.*
- Start with the organization of a good data infrastructure and associated access rights in time.

Step 3: Clean and explore data

When the dataset is available, the data science team can start developing the solution. An important step is to first clean and then explore the obtained data. The data must meet a number of requirements to be suitable for usage in developing AI models. The data must be representative and good quality.

REPRESENTATIVE DATA

It is important that the data which is used for developing models is as good a representation of the reality as possible. A self-learning algorithm is getting smarter by learning from examples in the given data. Requirements for the data are the diversity and completeness of the data points. Furthermore, the data must be up-to-date for many business cases, because old data might not be relevant anymore for the current situation. Be aware that no unintended bias is included in the data provided.

EXAMPLE

When developing a model for classifying images, one must take diversity in the images into account. An example for this is searching for damage in concrete through images. Imagine all photos with damage were taken on a cloudy day and all photos without damage on a sunny day. It might be that the data science model will base its choice on the background colours. A new image of concrete damage on a sunny day can therefore be classified incorrectly as undamaged.

Quality of the data

Besides the variety and completeness of the data, the quality is also of great importance. A good data structure supports this: the data should be as complete, consistent and obvious as possible. It is also essential to prevent human (input) errors as much as possible. Mandatory fields, checks and categories instead of open boxes of text can provide a solution, when entering the data.

Confidence

Representativeness and the quality of the data have a positive impact on the confidence in the project and the end solution. This confidence also contributes to the acceptance and usage by end users/executive experts.

Short feedback cycle

When exploring the data, it is crucial that the data is correctly interpreted. This happens through asking feedback from the implementing experts. A short feedback cycle between the data science team and the implementing experts is required for this. This can be executed for instance in every few weeks by giving a presentation of the findings to the implementing experts and /or the manager.

STEP 3 SUMMARIZED

- Make sure that the data is representative. The data must be complete, diverse, recent and free from unintended bias. Make sure that important context which may affect the prediction, are present in the data. Use the knowledge of the executive experts for this.
- Provide high quality data. Prevent errors in the dataset by catching them as accurately as possible when entering the data. Evaluate if the data is complete and consistent.
- Create trust with all stakeholders.
- Provide a short feedback cycle for the right interpretation of the data.

Step 4: Development and evaluation of models

At this stage, the data scientist has a lot of freedom to explore. The goal is to create a lot of business value as soon as possible. As a result, no data science solution is exactly the same and data science has a strong experimental character. What do you have to think of here and what does experimenting involve? On one hand this involves the type of algorithm and its parameters, on the other hand, the variables constructed from the data (the features). Depending on the problem and the available data, different categories of AI models can be used. During the development and evaluation of the models, there are a couple of points which you need to pay attention to.

Labels of the dataset

The most commonly used models are in the category of Supervised Learning, where the model learns from a set of data with associated known annotations or labels. Consider, for example, the aforementioned recognition of concrete damage. The model is trained on a set of photos of concrete structures which are known to be damaged. If trained, the model can be used to classify new photos.

Unfortunately, there is not always a data set available for which these types of labels (in this case 'damage/no damage present ') are known. Then it is necessary to create these labels. This task is also referred to as "labelling" or "annotating" data and is often largely manual work. For the example of the concrete damage, this means that someone manually goes through about 2000 photos and indicates whether damage is visible. There are also methods to speed up this process, for example by only providing a domain expert with photos of which the Machine Learning model is most uncertain. Many models indicate themselves with what certainty a classification was made.

Evaluation of the model

Labels are important for developing the model as well as for assessing the model. It is possible that the actual value will automatically appear in the data when it becomes known. In that case, a direct comparison between the actual value and the model prediction can be made. If the actual value does not automatically appear in the data, as in the example of the concrete damage classification, a feedback loop can be built into the end solution. Then the user will be asked to provide feedback about the correctness of the classification. This information is important for monitoring the quality of the data science model.

Optimization of the algorithm

An algorithm must be optimized and tuned to the problem and the related data. By adjusting hyper parameters, "the model's rules", the algorithm is adapted to the application. This optimization step often follows from a grid search, in which a large set of different values is tried and the best are chosen.

Choosing evaluation method

In order to determine the best model, an evaluation method must be defined which reflects the purpose of the data science model. For example, it is important in the medical field that extreme errors of the data science model do not occur, while in other fields extreme errors may have been caused by extreme measuring points in the data, which people choose to give less value to. Classification of models includes the balance between inclusivity (finding all concrete damage) and precision (finding only concrete damage). The assessment method must correspond to the business case and how the algorithm will be used in practice.

Monitoring the models

Monitoring the quality of data science models is very important. This is due to the dependency of the data.

Data can change over time, as a result of which a data science model may also show lower performance over time. As soon as the outside world changes, the data changes with it. This is how the seasons can have influence on photos and consequently on the results of the model. It can be useful to add an explanation to the prediction of a data science model, for the end user/ executive expert to understand what the prediction is based on. This provides more insight and transparency into how the model reasons.

Explainability of the predictions

Making a data science model mainly consists of iterations: running and evaluating. Also here the feedback from users is important. Do the predictions of the model make sense? Are there essential variables missing which could have an impact? Do the identified relationships also have a causal relationship? The transparency of the algorithm plays an important role in these relationships. Some types of algorithms are very opaque, so that given results cannot be traced back to the input data. This is amongst others, the case for neural networks. More linear methods or traditional statistical approaches are often easier to interpret. The demand for transparent algorithms can be seen in the recent developments regarding Explainable AI. When choosing the model, take the practical requirements of transparency into account with regard to the explainability of the predictions.

Prototype finished?

Is the prototype finished? For example, is there an interface available which shows the results? As soon as the prototype generates value, every step of the first phase has been completed and the second phase begins, the operationalization.

STEP 4 SUMMARIZED

- Make sure the data set is labelled.
- Make sure that the assessment of the model is part of the end solution.
    - *Can the actual value be obtained automatically from the data?*
    - *Can the end user/executive expert give feedback?*
- Is the algorithm evaluated with the correct assessment method?
    - *Is the assessment method used consistent with the problem definition and the data used?*
- Make sure that monitoring the quality of the model is maintained just as the changes in the data. Add for instance an explanation to the predictions of the model, thereby the executive expert/end-user can understand the predictions.
- Make sure that the data science solution is transparent. Discuss the performance of model with the executive expert/end-user.
    - *Do the predictions make sense?*
    - *Does the model meet expectations?*
    - *How can the predictions be used by the executive expert/end user?*
- If the prototype generates value, then the process can proceed to the next step: operationalization.

FOCUS ON CONTINUITY: FROM PROOF-OF-CONCEPT TO A STABLE AND SCALABLE END SOLUTION

We have reached the second phase of the life cycle of a data science project, the operationalization. In this phase a stable and scalable end solution will be developed from the proof-of-concept. This process consists of three steps.

Step 5: From successful proof-of-concept to implementation

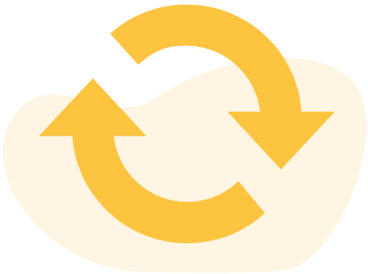
In step 5, the model and the organization will be prepared in order to use the model in practice. We have put all important elements of this step together for you.

Production-worthy code
Data science and software development are two surprisingly different worlds. Where data scientists focus on devising, developing and conducting experiments with many iterations, software developers are focused on building stable, robust and scalable solutions. These are sometimes difficult to combine.

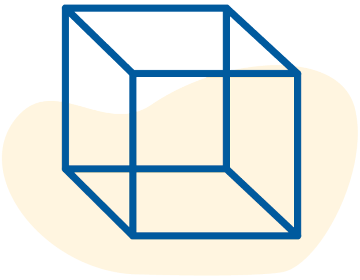
In the first phase the focus is on quickly realizing the best possible prototype. It is about the performance of the model and the creation of great business value as quickly as possible. That is why there is often little time and attention for the quality of the programming code. In the second phase the quality of the code is also important. This concerns the handling of possible future errors, clear documentation and efficiency of the implementation (speed of code).

That is why in practice the second phase usually starts with restructuring the code. The final model is structured and detached from all experiments that have taken place during the previous phase.

Integration into existing processes
The restructured prototype must be integrated into the existing processes in the organisation. This mainly consists of two steps.



Automatization of data flows
The retrieval and writing data back to a database must be done automatically by means of queries. This can be a complex process which can be facilitated by a data engineer.



Setting up an infrastructure hosting and managing the model
The model must be available to all end users/ executive experts and it should scale with its use.

Requirements for the production environment
There are different requirements which you can set for hosting the model in the production environment.

- When a model is to be used intensively, scalability also becomes relevant. This means that a model is started in parallel several times, so that all requests are distributed evenly across the models and all requests can be processed faster. When the number of requests for a model varies greatly over time, it is helpful if the production environment facilitates automatic scaling. This can be compared to a supermarket where several cash registers are opened when it gets busier and close again when things get quieter. The solution is also ready for the future. The model can be used more intensively without additional infrastructure investment.
- Another important condition for the production environment is availability of the solution. The end-users/executive experts must always be able to use the model. Think of the automatic restart if systems fail. Automatic backups can also be part of this.
- A third requirement is authentication and security. People and/or systems which are allowed to view and use the production environment, the model and the data must be accessible for them. This requires a safe and reliable login system.
- Transparency and auditing in the production environment is also a point of attention. You want to exactly know who changed what and when. Changes must be traceable, so that it can be traced back to where things went wrong. Thanks to good logging, you can quickly find out exactly what happened or went wrong, which makes troubleshooting easier and faster. Link the auditing to the monitoring of the models, so that it can be determined whether the performance is related to an update of the model.

> *“When a model is to be used intensively, scalability also becomes relevant.”*

Management of the production environment
When the model produces error messages, they must be corrected. It is important to clearly define who is responsible for the models which run in the production environment. This can be the IT department or the data science team itself. In any case, it will benefit if the code is structured and error messages are described in a clear log.

Dealing with multiple stakeholders: ownership of data and code
It often happens that an organisation cooperates with an external party for the development of data science models, or for the delivery of data. Data handling and data ownership are serious topics and many organisations attach great importance to protecting their data and controlling who has access to it. In addition, this may also be required by legislation, such as the GDPR. The same applies to the model code of a data scientist.

It is possible that discussions arise within collaboration between the parties: who owns the data and/or who owns the algorithm. A neutral model-hosting platform can offer a solution in order to be able to work with multiple parties. This platform can be reached by the data scientist with his model code and by the data supplier with his data. Proper rights can be derived from this neutral platform. When working with a neutral platform, one can test the data scientist's model without having access to its source code. For example, one could make a data test set and compare several model variants.

The legal aspects of an AI project

There are several legal focus points when it comes to an AI project, so get in touch with an internal or external legal expert in the early stages.

Several points which you need to pay attention to:

- Make sure you have an understanding of the marked data, so you are not dependent on one of the developers.
- Fix some specific quality norms for development AI models (supervised learning)
- Intellectual Property (IP): the added value of a model results from the adjustment of the parameters, explicit agreements must be made to have IP rights thereon.

Lawyer office Pels Rijcken in cooperation with Amsterdam City Council developed a ***general terms and conditions for the purchase of AI applications***, which is used specifically for making decisions about citizens. Although purchasing conditions are dependent on the exact case histories, but parts of these conditions might be reused.

Acceptance

Enthusiasm from the users for the new solution does not always come naturally. This is often related to the idea that an algorithm poses a future threat to the employee's role in an organisation. Yet there is still a sharp contrast between where the power of people lies and that of an algorithm. Where computers and algorithms are good at monitoring the continuous incoming flow of data, people are much better at understanding the context in which the outcome of an algorithm must be placed. This stimulates in the first instance to pursue solutions that focus as much as possible on human-algorithm collaboration. The algorithm can filter and enrich the total flow of data with discovered associations or properties, and the expert/technician can spend time interpreting the results and making the ultimate decisions.

Liability

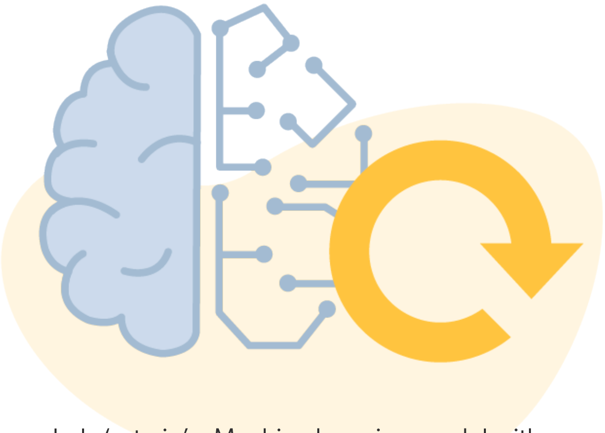
How the model is used in practice, partly determines who is responsible for the decisions which are made. When the data science model is used as a filter, one will have to think about what happens if the model makes a mistake in filtering. What are the effects of this and who is held responsible for this? What about an autonomous data science solution, such as a self-driving car, for example, who can be held accountable if the car causes an accident? These are often tough ethical questions, but it is important to consider these issues prior to integrating a model into a business process.

STEP 5 SUMMARIZED

- Make sure that the quality of the code is good.
- Integrate the model into the existing organizational processes.
  - *Automatize data streams*
  - *Make sure that there is an available infrastructure for hosting and model management*
- Pay attention to the requirements of the production environment: scalability, availability, security and transparency & auditing.
- Specify those responsible for data science model management.
- A neutral platform can help protect IP/ownership of data and model code.
- End users/operational experts must participate in the project.
  - *Encourage human-algorithm collaboration. And make the expert job more challenging with no threat to job loss.*
  - *Help end users/executives use the solution and value the benefits of their contribution.*
- Specify who is responsible for errors that arise from incorrect predictions of the data science model.
- Discuss the legal aspects of the projects in the early stage with a legal expert.

Stap 6: Managing models in operation

A model which runs and is used in a production environment must be checked frequently. It is necessary to agree on who is responsible for this, and who executes these frequent checks. The data science team, or the end user /executive expert can assess whether the model continues to work well and remains operational.



It may be necessary to regularly 'retrain' a Machine Learning model with new data. This can be a manual task, but it can also be built into the end solution. In the latter case, monitoring the performance of the model is essential. Which model performance and code belongs to which set of training data has to be stored in a structural way, so that changes in the performance of a model can be traced back to the data. This is called ' data lineage '. More and more tooling is coming onto the market for this. An example is Data Version Control (DvC).

STEP 6 SUMMARIZED

- Clearly agree who will continue to monitor the model over time when the model is running in the operational environment.
- When a model is frequently 'retrained' on new data, it is important to consistently save: when, which version is used, which performance belongs to which training dataset. This is required for the traceability of model performance.

Step 7: From model management to business case

Frequently checking the performance of a data science model, any degrading model output can be discovered in good time.

This occurs due to the strong dependence between the code of the algorithm, the data used to train the algorithm and the continuous flow of new data, influenced by various external factors. It may happen that environmental factors change, and as a result certain assumption are no longer correct, or that new variables are measured which were previously unavailable.

The development of the algorithm therefore continues in the background. This creates newer model versions for the same business case with software updates as a result. In practice several model versions run in parallel for a while, so that the differences become transparent. Each model has its own dependencies that must be taken into account.

When multiple models and projects coexist in the same production environment, it is of paramount importance that all projects remain transparent and controllable. Standardisation is required to prevent fragmentation. E.g. a fixed location (all in the same environment) and the same (code and data) structure.

Ultimately, data science projects are a closed loop and improvements are always possible. The ever-changing code, variables and data makes data science products technically complex on the one hand, but on the other hand very flexible and particularly powerful when properly applied.

STEP 7 SUMMARIZED

- If possible, run a new model version in parallel with the previous model version for a while so that they can be compared.
- Create standards and make them requirements for new data science projects.
  - *Fixed location (all in the same environment).*
  - *Consistent code structure (if developed internally).*
  - *Uniform data structure.*