ROCKET SCIENCE

PHASE 1

# Assess your company's maturity level in data valuation with our Data Performance Score Card

Before starting your journey in the world of Data Science and Machine Learning, you need to assess existing assets, understand where your weaknesses are as an organization, see where the value in terms of projects is, and so on. Assessing your level of maturity in terms of data valuation inside your company will help you understand where you should put your efforts before or while investing in Data Science and Machine Learning.

This interactive tool will give you an overall score based on your answers, it will come with comments about what to improve and what to do next. Keep in mind that this is a qualitative tool to aid decisions. The questions are divided into three groups: "Data", "Data Science", "Business and culture".

Your data assets and how they are structured reflect how ready you are to create complex solutions in Machine Learning that will rely on those data. Your existing Data Science projects or efforts reflect the appetite of your organization for such solutions. Your culture and the business side will reflect how much support you will have later in your journey and how ready your people's organization is for Data Science and Machine Learning.

# Data Questions

1 Do you have any data lineage** to recognize your data sources and data flow?

*** Data lineage is the process of understanding, recording, and visualizing data as it flows from data sources to consumption.*

Yes

2 Do you have a proper tool for data cleansing and transformation (ETL tool)?

Yes

3 Do you have a Business Intelligence Practice with reporting tools and resources?

Yes

4 Do you have a data warehouse?

Yes

5 If you have a Business Intelligence Practice, how successful do you feel it is from 0 to 10? 0 would be not successful at all, 10 is a tremendously successful practice.

0

6 If you have a Business Intelligence Practice, how long this practice has been in place?

Less than 6 months

7 Do you have a Data Governance program in place that defines who can take action on data?

Yes

## Data Science Questions

8 Have you performed any Proof of Concept in Machine Learning?

Yes

9 Have you purchased any solution that is powered by Machine Learning (including self-service tools)?

Yes

10 How many Machine Learning projects have you identified in your organization?

10+

11 Do you have a data analyst in the company that performs your data analysis?

Yes

## Business and Culture Questions

12 Would you qualify your company as a data-driven company?

Highly data driven

13 Is your top management supporting future investments in Machine Learning?

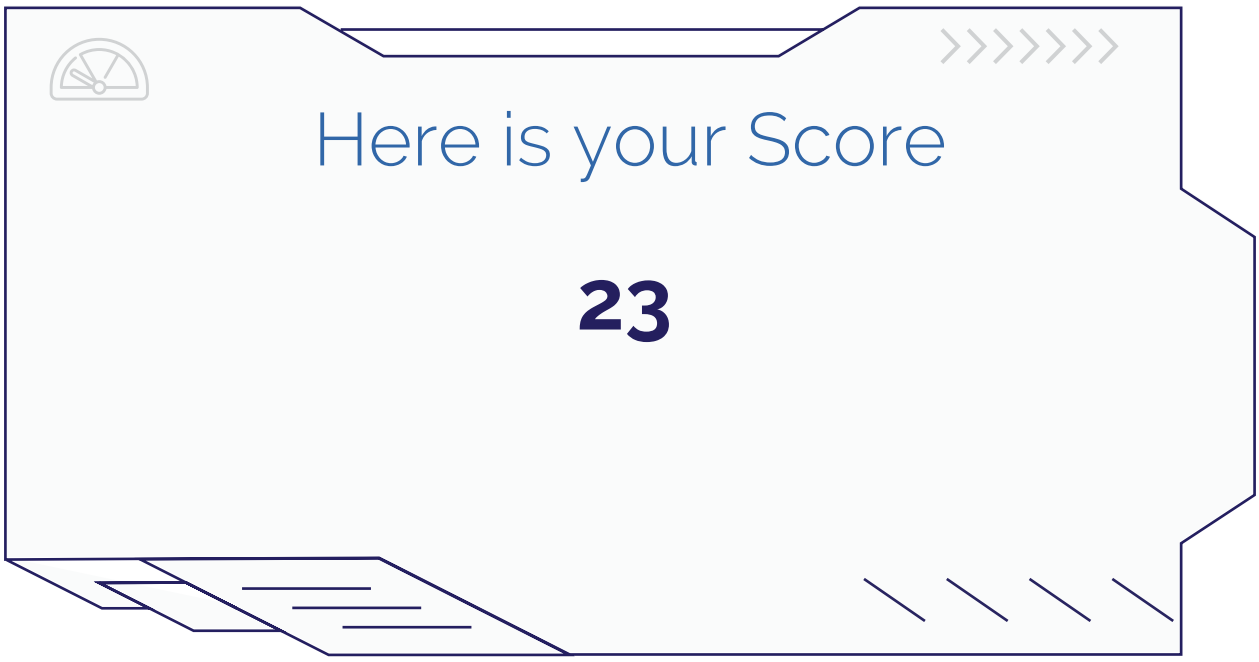
Yes, totally

14 Do you have the support of IT in your process?

Yes

15 Are there internal departments or units that have an appetite for Machine Learning solutions?

No



# Matrix result


If the score is under 15

Becoming a data-driven organization is a hard process and implementing a culture where you create value through data is a long path. Your level of maturity reflects the need to invest in building a stronger practice in Business Intelligence with the proper tools. Looking for support from top management will also help you gain traction inside your organization.

Our model has rated your organization around **25% mature** in terms of data valuation.

Don't give up! Your data journey is still in its infancy, and you have room for improvement!

**Here are some actionable items that you can use to start improving your maturity:**

- Start by optimizing your BI practice or create one if not already the case. Make sure your user's needs are fulfilled the right way.
- Invest in tools for data cleansing, data lineage, and a data warehouse (if not the case).
- Start thinking about data governance inside your organization to help track data and secure access to data.
- Look for potential projects based on data across your organization.
- Get closer to the LoBs, and try to position projects based on data.
- Create internal workshops, webinars, or demo sessions inside your organization. This will support the drift to a data-driven organization and people will start thinking of ways how to use data.
- Attend conferences and webinars to help you meet other leaders like you and learn from their successes and failures.

**The next step is to download the framework: Identify Machine Learning projects with high value to your organization.**

START YOUR JOURNEY

# Matrix result


If the score is between 15 – 25

Your organization seems to have good assets in terms of data valuation. We can definitely improve this process by investing in proper tools and creating best practices. You probably have a supportive top management with a good ear from different departments and LoBs. Your journey in Data Science and Machine Learning can start on the right track if we optimize some of your existing assets.

Our model has rated your organization around **50% mature** in terms of data valuation.

Keep going! Your journey in Data Science and Machine Learning could start today, but let's optimize your existing assets to help you be on solid ground!

**Here are some actionable items that you can use to start improving your maturity:**

- Invest more seriously in data governance to help you gain control over your data flow.
- Move your Business Intelligence practice from a "demand and do" approach to a more proactive approach where you foresee your user's needs and create replicable work (dashboards, data requests and reports).
- Invest more in a data engineering team that will clean your data and help you have more reliable and clean data for BI, but for Data Science as well.
- Create internal workshops, webinars, or demo sessions inside your organization. This will support the drift to a fully data-driven organization and people will start thinking of ways how to use data.
- If this is not already the case, start identifying projects in Data Science using our framework (2—Identify projects with high value to your organization)

**The next step is to download the framework: Identify Machine Learning projects with high value to your organization.**

START YOUR JOURNEY

# Matrix result


If the score is Over 26

Your organization has great maturity in terms of data valuation. It seems that you are in control of the situation. You have good support internally and your BI team is doing a good job. The probabilities of having success in Data Science and Machine Learning are high.

Our model has rated your organization at least **75% mature** in terms of data valuation.

You are definitely ready for the great jump in Data Science and Machine Learning!

**Here are some actionable items that you can use to consolidate your existing strengths:**

- Optimize your data governance to help you gain control over your data flow.
- Create internal workshops, webinars, or demo sessions inside your organization. This will support the drift to a fully data-driven organization and people will start thinking of ways how to use data.
- If this is not already the case, start identifying projects in Data Science using our framework (2—Identify projects with high value to your organization).

**The next step is to download the framework: Identify Machine Learning projects with high value to your organization.**

START YOUR JOURNEY

## Looking for support in your journey?

GET IN TOUCH NOW

ROCKET SCIENCE

PHASE 2

# Identify projects with high value to your organisation

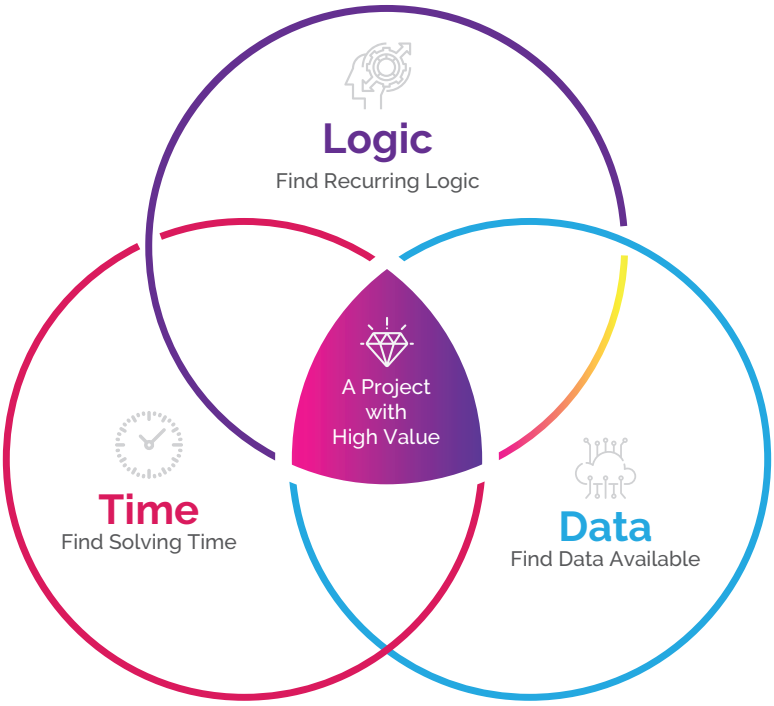
In phase 1, we have assessed the level of maturity for data valuation. Now we have enough insights to improve our foundation and data management in order to get ready for Data Science and Machine Learning projects.

Detecting projects that have value across an organisation may seem to be easy. By following the type of data available you could detect projects here and there. But how can we know whether a project will have a high value for an organisation?

Our framework for detecting projects with high value will help you select projects, which will ease the process of building a roadmap for your Data Science practice. This framework is based on three dimensions that filter down projects.

# Our framework T.L.D.:

The framework T.L.D (or Time, Logic and Data) focuses on the underlying business process of a specific project to understand if the project is suitable for Machine Learning. Moreover, it will help understand if a Machine Learning system will help optimize the business process targeted by a specific project.




# What is the resolution time for the underlying business process?

The resolution time defines how long it takes to complete the business process or a subprocess in a business process. The higher the resolution time, the more valuable becomes to use Machine Learning to solve the problem in the underlying business process. For example, in a factory, the process of scanning a box or item is using a bar code scanning system (not powered by Machine Learning). This process is fast as the time for reading the bar code is low. Then the resolution time of this process is very low. Meaning that machine learning is less valuable as a project for this process.

## Is there a recurrence of logic in the way of solving the underlying business process?

The recurrence of logic is critical in the framework, and it represents how predictable the resolution of problems is in a specific business process. The recurrence of logic reflects patterns and logical schema that we can cross when solving a problem. For example, when you are trying to understand if a candidate in your organisation will become a good employee, you are following logic and extracting patterns from your previous experience with other candidates to understand if this one will become a good employee. When there are rich patterns in a way of solving a business process, this means that the project is highly suitable for machine learning.

**The recurrence of logic is almost everywhere, but some business processes have richer and more clear patterns than others.**

## Is there any data generated that reflect the business process?

The last dimension is data. Data doesn't necessarily have to be already numeric or structured. It can be unstructured (text, image, etc.). For example, creating a virtual assistant would require historical data of conversations which is raw text. Large quantity of data is not necessary as it depends on how complex the problem we are trying to solve is; some problems require large amount of data, others don't. Having data collected is most important, and this will increase the value of a Machine Learning project.

# Examples

## Stock market prediction

| Path | Score | Comments |
|---|---|---|
| Resolution time | High  | *When a human predicts the stock market, to take an efficient decision, he does research, asks another expert, etc. This can be a long process.* |
| Recurrence of logic | Medium/Low  | *Sometimes the stock market movement can be hard to understand. So, predicting the future can have an obscure logic (there is still logic, but it is hard to capture)* |
| Data collected | High  | *Many data sources are available, like historical prices, news, people communication and comments, companies' financial statements, etc.* |

## Legal artificial intelligence assistant

| Path | Score | Comments |
|---|---|---|
| Resolution time | High  | *When a human legal assistant is trying to help a lawyer build a case, this process can take time, while doing the research, reading about the case.* |
| Recurrence of logic | High  | *The logic followed when doing research is recurrent as the legal assistant will use his/her experience to know what to look for.* |
| Data collected | High  | *Law is a field with rich data from various sources- old cases, law books, etc.* |

## Predicting a customer's behavior

| Path | Score | Comments |
|---|---|---|
| Resolution time | High | *For a human to analyze and be able to predict a customer's behavior, it takes a deep understanding of the customer and the product and also, some previous experience in customer marketing. This makes the resolution time very high.* |
| Recurrence of logic | High | *The logic that drives a customer's behavior can be captured and is clear. A customer tends to follow specific patterns.* |
| Data collected | Medium/Low | *Predicting the customer's behavior requires having data about past customer's behavior, but also a complete customer's profile. Customers' profiles (including socio-demographics) are hard to collect for many companies.* |

**The next step is to download the framework: Identify Machine Learning projects with high value to your organization.**



Looking for support in your journey?

GET IN TOUCH NOW

ROCKET SCIENCE

PHASE 3

# The feasibility of a Machine learning project

Now that we know how to detect projects with high potential, we can move on and learn how to evaluate the feasibility of a project. Note that having a project with high potential doesn't mean that it is a feasible project.

Feasibility means that a project is technically doable. Normally it requires deep technical knowledge with good industry experience to know if a project is really feasible. And due to the uncertainty in Machine Learning, a feasible project on paper can end up being a not possible project during development.

This framework aims to simplify your decision regarding how feasible a project is. We recommend validating the feasibility by a specialist in Machine Learning that will guide you around the feasibility of your projects.

# Question to ask when analyzing the feasibility

**Is the data available internally or does it need to be acquired externally?**

Very frequently, what limits a project is the availability of data. Companies and leaders end up discovering that internal data is not enough to conduct a specific project, and that no third party can provide this data, or that buying third-party data is too expensive. The most popular use case is in marketing, where companies want to conduct Machine Learning projects on their customer data (like predicting customer's behavior) and they realise that they are missing data about their customers (such as age, sex, and any socio-demographics). When turning to third party providers, this can become very costly, especially when you have millions of customers with many characteristics desired.

This reflects if the company collects the data related to a specific project. Sometimes collecting data is not possible for marketing or ethical reasons, and sometimes it can be time-consuming as it can take a long time before having a decent amount collected.

> So, the availability of the data can play a critical role in the feasibility analysis.

## Does the project have existing use cases on the web or from other companies?

If a project is already documented on the web as a use case, this means that it is probably a feasible project. For example, when you google "customer churn prediction", you will find many use cases on the web and even companies explaining how they implemented this Machine Learning project in their organisation. As well by attending conferences in your industry or AI, you might hear about use cases related to your potential projects that explain how it has been implemented.

All these are signs that your project might be feasible. But there is no guarantee that your project will be feasible technically in reality as every company and project are unique. However, it is still a good indicator of the feasibility.

**How hard would it be to deploy and maintain the model in production?**

Another factor that can play a certain role in the feasibility analysis is how this model will be deployed. And how costly it is to maintain in production. For example, a telecommunication company would like to create a Machine Learning model that can detect any anomaly on their network data. This project requires creating ingestion pipelines in production. Then the pipelines are streamed with the machine learning model deployed. Creating such streams and pipelines for all network data is very costly in terms of hardware infrastructure, but streams also require efforts to be maintained and updated.

Defining how hard it is to deploy and maintain can get quite technical which might require you to speak with a specialist (a Machine Learning Engineer or MLOps specialist).

## Scenarios

A retail company wants to analyse if a customer churn project is feasible. This company has assessed that a customer churn prediction project is highly valuable.

First, the company looked at the data collected and available. They realized that they have transactional data about the customers, as well they have basic information about their customers, like the address. Another data could be available, and that is the data about calls to customer support, including duration of calls, number of calls per month, etc. Having access to additional data about customer socio-demographics would have been great to support the project, but they don't want to ask their customers to provide this information and are not willing to purchase it from third-party providers. All this makes the project quite feasible technically from a data perspective.

One of the leaders decided to google "customer churn prediction", and realized that churn prediction is a well-documented project with many use cases available online. This makes the project feasible from a tracking record perspective.

From deployment and production perspective, the models could be deployed as a simple script against the databases, and this would perform pretty well. Other more sophisticated deployments are possible like deploying in Microservice which would allow integration with other tools and software. This makes the project feasible from a deployment perspective with average to low complexity.

Overall performance

| | |
|---|---|
| Data | Enough to start first POC/Prototype |
| Tracking record | Highly available |
| Deployment and maintenance in production | Average to low complexity |

## What do I do if the project is not really feasible?

Well, a project that is qualified as not feasible can still be conducted depending on what aspect is not feasible. If the project is not feasible due to a lack of data, the project can be delayed waiting to collect enough data. If the project is not feasible due to a lack of tracking record (no existing use case found), the project can be conducted in an R&D mode. If a project is not feasible from a production perspective, the company can start implementing on a smaller scale or in a simulation environment. Working with Machine Learning Architects can help design an optimized and engineered environment for production that will lower production costs and efforts.



Looking for support in your journey?

GET IN TOUCH NOW

ROCKET SCIENCE

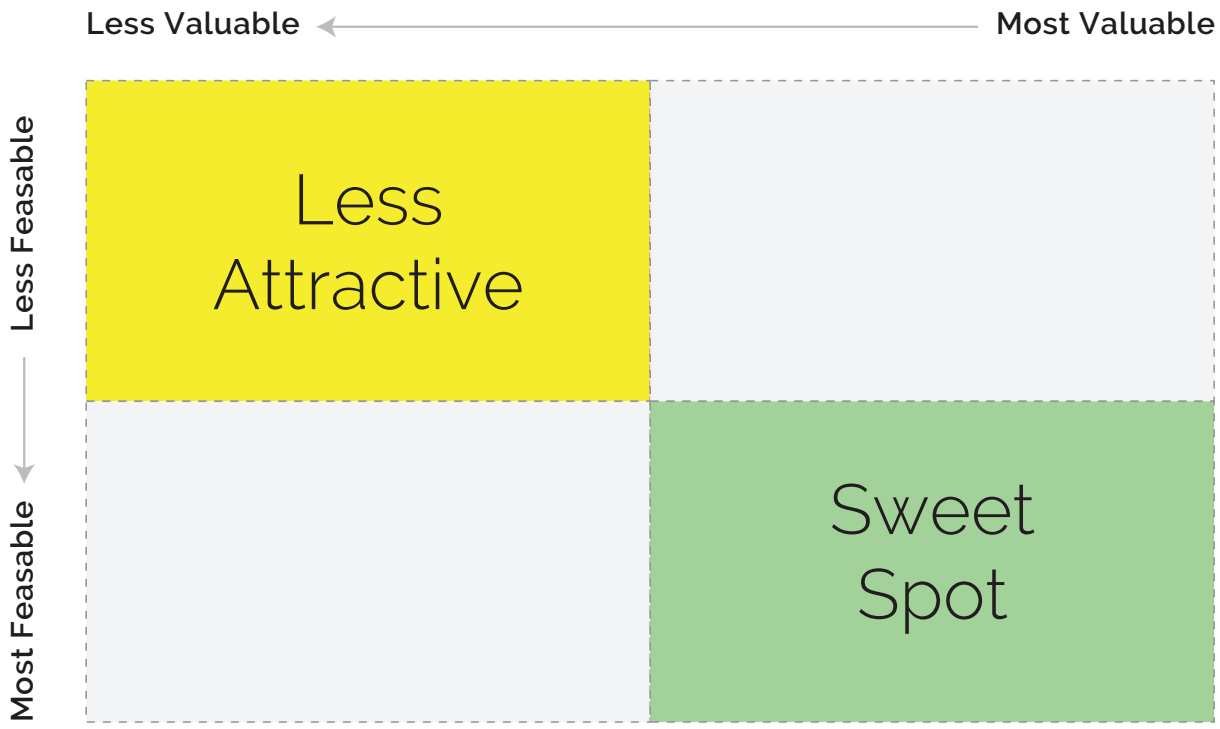
PHASE 4

# Prioritize machine learning projects framework

Now that we learned how to assess if a project is feasible, you probably assessed many projects. How can you prioritize all those projects and decide which ones should be your priorities in the short term of your Data Science practice?

In the previous phases, we assessed the feasibility and the value of a project. Those two factors will become our decision vectors in prioritizing projects. We will use quadrants that cross the feasibility with the value to understand which projects should be a priority.

# Value/Feasibility Quadrants



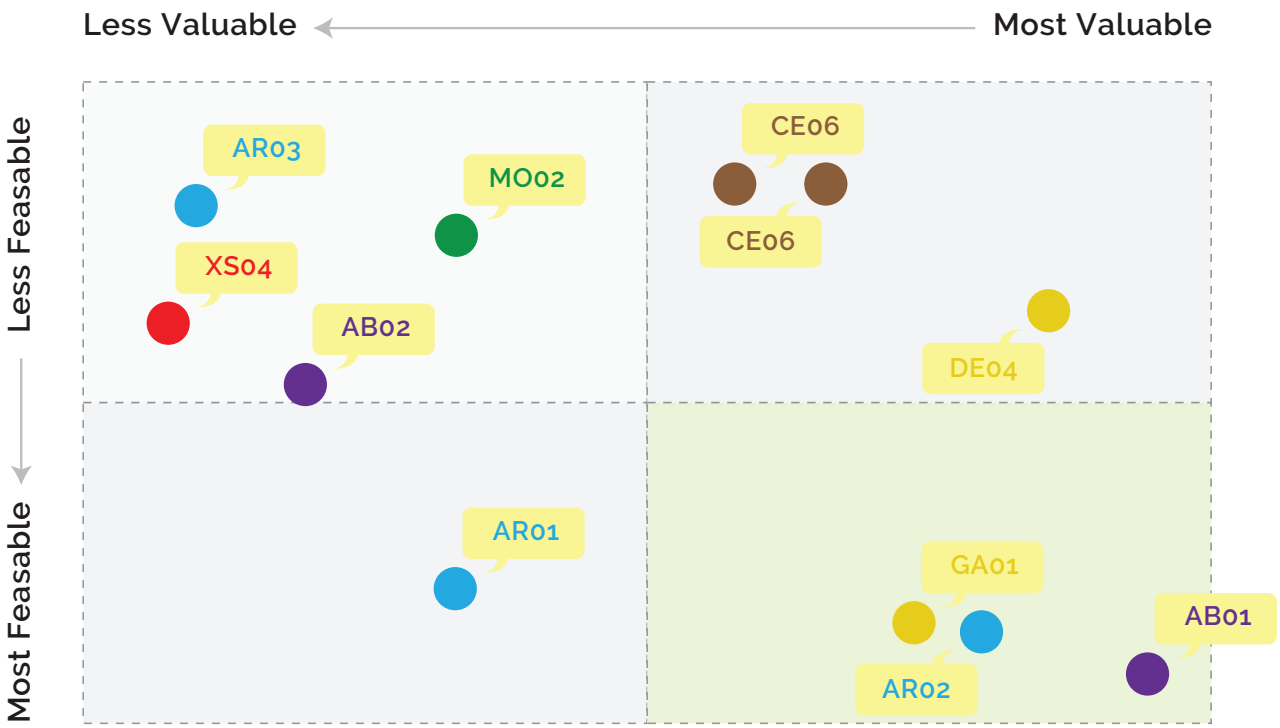

Every project must be added to the quadrant based on its value and feasibility. The square in green should regroup all projects with the highest feasibility and value. Note that it is important to compare the projects in terms of feasibility and value. A project is always valuable and feasible compared to other projects.

This is a qualitative approach so we advise conducting this analysis with your team members as multiple opinions will increase your chances of being accurate.

A little advice is to group projects into categories (like sales boost, marketing content, customer experience, etc.). If you have a large number of projects detected, you can break them into smaller quadrants by category before running a final quadrant.

# Examples

After evaluating the value (framework 2) and feasibility (framework 3) you can create your quadrant by comparing feasibility and value between projects. In the following example, 3 projects are in the sweet spot, so they will become the priority for your short-term roadmap. Keep in mind that the other projects are not disqualified or left behind. They can be added to the mid/long-term roadmap based on their order of priority.




Looking for support in your journey?

GET IN TOUCH NOW

ROCKET
SCIENCE

PHASE 5

# Scope of Work (SOW)

After prioritizing the projects, we know the projects that we need to start with. To start our first initiative, we need to define the scope of work per project.

# Scope of Work

The following document is a template for drafting the scope of work.

| Path | Details | Comments |
|---|---|---|
| Business context | *Provide an introduction about the business process underlying the project and how it will help the business.* | *The business context will help the team keep in mind the objectives of the project.* |
| Problematic | *What problem are we trying to solve?* | *The problem must be well defined to avoid leading to wrong results.* |
| Technical overview | *Explaining what is technically required and what type of technics are needed (supervised learning? Classification? Computer Vision?).* | *Has to be done with the technical team (a data scientist or a machine learning engineer)* |
| Data required | *Define what data sources are needed to perform the project. Detail as well variables available with the data type.* | *Has to be detailed with the data source.* |
| Assumptions | *Any assumption that can impact the delivery.* | *The type of assumptions can be the collaboration of a specific resource outside your team. Or that the project will be conducted in Research mode, etc.* |
| Deliverables | *What is expected as an outcome?* | *Do we need to build software? Or do we need to deliver a report? Present our results?* |
| Estimated effort | *How much time is required?* | *Can be in hours or days.* |
| Required expertise | *What type of skills are needed? And what type of resources?* | *Do we need skills in computer vision? Or a good knowledge of SQL? Do we need software developers to put the model in software?* |
| Exclusions | *What is excluded from the project?* | *We can exclude deployment if it is not a part of the scope and it will be conducted later. Or exclude a specific data analysis phase, etc.* |

ROCKET SCIENCE

PHASE 6

# Necessary profiles to build your Data Science team

Now that we know which the priority projects are. We have scoped our projects, and we can move on to building the team and get talents. The data science world is complex and there are more than just data scientists out there in the market. Many different profiles are available, and a proper team is composed of complementary profiles from different backgrounds.

Your ability to identify the necessary profiles and understand what skills are needed to complete the projects you have is very important. An R&D project doesn't require the same profiles as business-driven projects. Additionally, deploying models and managing machine learning life cycle require special skills. We will start by identifying the different roles available in the market.

# The different roles possible in data science

| Title | Role | Skills | Tools |
|---|---|---|---|
| Data analyst | Analyze data and deal with data requests from internal customers. | Business analysis / Data engineering (SQL mainly)/ Reporting/ Statistics | SQL/ Tableau / Excel / PowerBI / R / SPSS |
| Business analyst | Understand the business requirement for every project. Deal with requests from internal customers. | Business analysis / Reporting / Communication | SQL/ Powerpoint / Excel / PowerBI / Tableau |
| Data Scientist | Develop machine learning models. Perform data analysis if required. | Data Science / Statistics / Machine Learning / Mathematics / Reporting | Python / R / SQL / Jupyter Notebooks / Tensorflow / Pytorch / Git |
| Research Scientist | Search for an innovative approach to solve machine learning problems. | Data Science / Machine Learning / Mathematics / Statistics | Python / C++ / C / R / SQL /Notebooks / Tensorflow / Pytorch |
| Machine Learning Engineer | Create machine learning solutions, from the development of models to deployment. | Data Science / Machine Learning / Statistics / Data engineering / Model deployment / Microservices / Containers | Python / R / SQL / Tensorflow / Pytorch / Git / Cloud providers / Docker / Kubernetes / Go / C# / C ++ |
| MLOps Engineer/ ML Software Engineer | Manage machine learning life cycle and create solutions and maintain them in production. | Data engineering / Model deployment / Microservices / Containers / Software engineering / Infrastructure / ML pipelines | Python / Git / Cloud providers / Docker / Kubernetes / Go / C# / C ++ / CI-CD / Microservices architecture / SQL |
| Data Engineer | Make data available to data scientists/Machine Learning engineers. Make data ready for production | Data engineering / ETL / Data Lake / Big data / Reporting | Python / ETL Tools / Hadoop |

# The tasks in data science life cycle and associated roles

| Task | Typical day-to-day | Associated role |
|---|---|---|
| Understanding customers need | - Meet customers and discuss their projects<br>- Conduct workshops and design sessions to understand where the valuable projects are. | Business analyst<br>Lead data scientist<br>Director data science |
| Scope a project and define requirements | - Evaluate project feasibility<br>- Understand necessary resources<br>- Validate the potential of the project<br>- Evaluate effort | Business analyst<br>Lead data scientist<br>Director data science |
| Collecting data | - Create pipelines for ingestion of the data<br>- Create a data warehouse<br>- Create databases (data marts)<br>- Create a data lake | Data engineer<br>Software engineer |
| Cleaning data | - ETL scripts/flow<br>- Scripting<br>- Validate transformation | Data engineer<br>Data scientist<br>Data analyst |
| Labeling data | - Create a training dataset<br>- Use statistics to estimate labels<br>- Use open data | Data scientist<br>Research scientist |
| Exploring the data | - Visualize the data<br>- Understand the distribution of the data<br>- Correlation<br>- Univariate/Multivariate analysis | Data analyst<br>Data scientist<br>Research scientist |
| Reporting on data and model performance | - Reports and dashboards creation about data quality<br>- Reports and dashboards creation about data sources<br>- Reports and dashboards creation about model performance | Data analyst<br>MLOps engineer (for guidance about model performance) |
| Training machine learning models | - Train models<br>- Data preprocessing<br>- Model selection<br>- Model evaluation | Data scientist<br>Research scientist<br>Machine Learning Engineer |

| Task | Typical day-to-day | Associated role |
|---|---|---|
| Optimizing training phase | - Running parallel computing<br>- Optimizing hardware<br>- Running on GPU | MLOps engineer<br>Machine Learning Engineer |
| Creating new machine learning algorithms | - Develop a new machine learning algorithm<br>- Research for new ML approach | Research scientist |
| Researching new technics in machine learning | - Reading research paper<br>- Attending conferences | Research scientist<br>Data scientist |
| Refactoring machine learning code | - Making machine learning code ready for production<br>- Optimizing code<br>- Rewriting code | Machine Learning Engineer<br>MLOps engineer |
| Deploying machine learning models | - Deploy ML models as microservice<br>- Deploying ML models to model registry<br>- Creating docker images<br>- Deploying scripts and applications | Machine Learning Engineer<br>MLOps engineer |
| Creating retraining pipeline | - Create automated pipelines for retraining models | Machine Learning Engineer<br>MLOps engineer |
| Managing machine learning environment | - Creating development environment<br>- Managing deployment environment<br>- Training and coaching team on how to use tools (MLOps tools) | MLOps engineer |
| Defining production requirements for machine learning models | - Defining best practices in terms of deployment<br>- Meeting data scientists to understand what requirements for ML models | MLOps engineer<br>Machine Learning Engineer |
| Setup environments (development or production) | - Create environment for development and for production | MLOps engineer |
| Create a CI/CD environment<br>Define retraining strategies | - Creating Continuous integration pipelines and tools<br>- Creating Continuous delivery pipelines and tools<br>- Define best practices for retraining | MLOps engineer<br>Machine Learning Engineer |
| Setup ML infrastructure | - Setup hardware (resources) for ML projects | MLOps engineer |

ROCKET SCIENCE

PHASE 7

# Key success in implementing data science practice

Congratulations! You have just structured your Data Science practice. Having a strategy based on the right frameworks is going to help you increase your chances to build a healthy practice. Now we will define additional requirements that can help increase your success with your practice.

# Define objectives with a strategy

To value your practice, you will need to define a proper vision that will dictate your practice and help you create requirements.

### Vision can be defined as

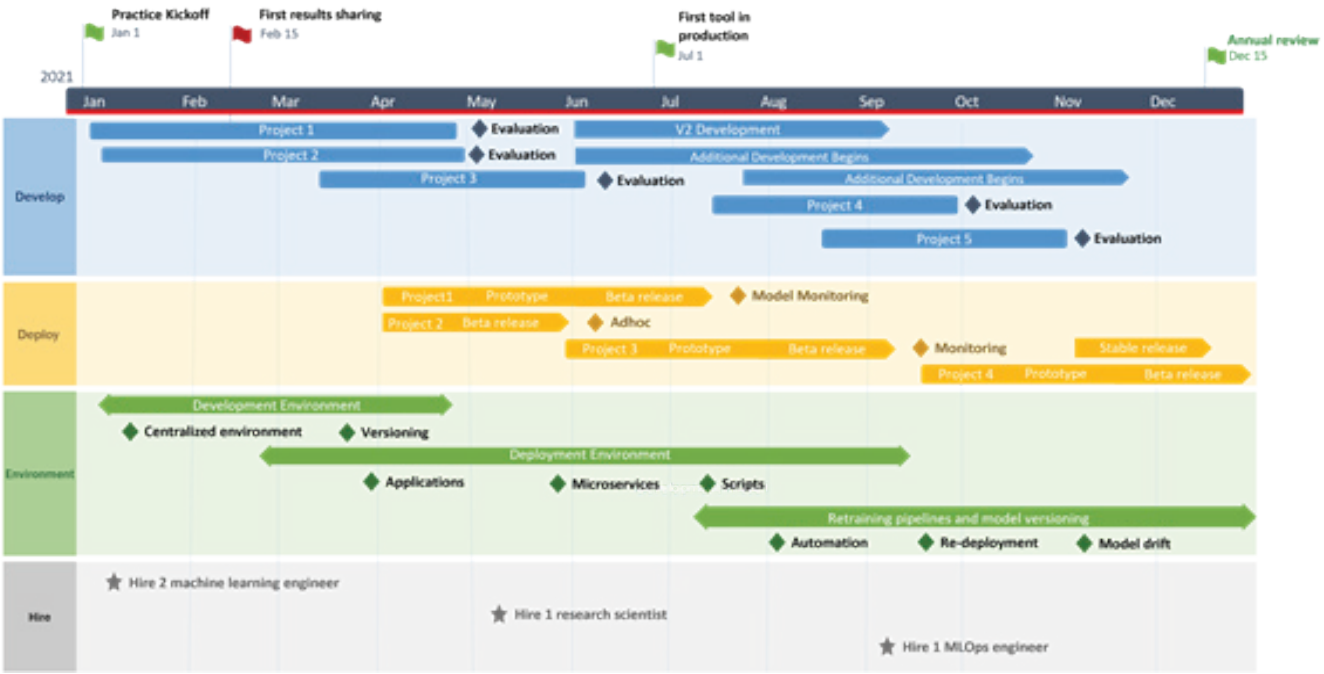
"Drift to a data-driven organisation"

"Monetize data/tools created"

"Generate more revenues through machine learning tools"

"Reduce costs from operations/marketing"

# Roadmap



# MLOps environment

It is very important to design a machine learning environment that will sustain our ML development cycle environment. This will include a sustainable development environment. Note that development can be conducted locally on a resources laptop, but it has to centralize the meta and experimentation metrics and parameters. Additionally, code has to be committed and centralized. Ideally, have a shared development environment that will have versioning control with tracking training performance. Then it is important to have a deployment environment where all assets will be deployed with a decision around the necessary technology to deploy. CI/CD plays an important role in moving from development to production and has to be decided carefully.

# Communication

It is important to communicate internally and externally around created solutions and innovative projects. This will help drive the projects and create an ecosystem around data and data science. Types of communication can be:

- Conferences participation, to share frameworks, success and journey.
- Press release around new tools and solutions.
- Internal communication around findings and new projects.
- Becoming a knowledge center through internal and external channels.

# Change management

Investing in change management will increase adoption for new and innovative solutions created internally in data science and machine learning. Change management contributes to creating a data-driven culture inside the company. This portion is as important as the development of solutions.

**To change the mindset and implement change management strategies, here are some tips**

- Create internal workshops to share knowledge with internal customers.
- Get tools discovery sessions with internal customers.
- Create an internal newsletter to share findings and insights.

## Some additional tips

- Avoid spending too much time in projects development and development in an iterative way.
- Make sure to validate the needs and assumptions with customers.
- In a machine learning project, make sure to have access to a subject matter expert that will have a deep understanding of the industry or domain where you apply your machine learning.
- Adapt deployment to your environment needs.
- Always monitor model performance in production and release a retraining strategy.



Looking for support in your journey?

GET IN TOUCH NOW