February 2024

initiative for applied artificial intelligence

ML Skill Profiles: An Organizational Blueprint for Scaling Enterprise ML

Contents

Executive Summary 4

Introduction 5

Organizational challenges to scale AI in the Enterprise 6

The Machine Learning Skill Profiles Framework 8

- Benefits 8
- Deep Dive 9
  - Framework Overview 9
  - Embedding the Skill Profiles within the organization 10
  - ML Roles - An expert's opinion 24
- How to apply the framework 26
  - Considerations specific to your organization 26

Case study: How ML teams are composed at Uber 30

Authors 32

Acknowledgements 32

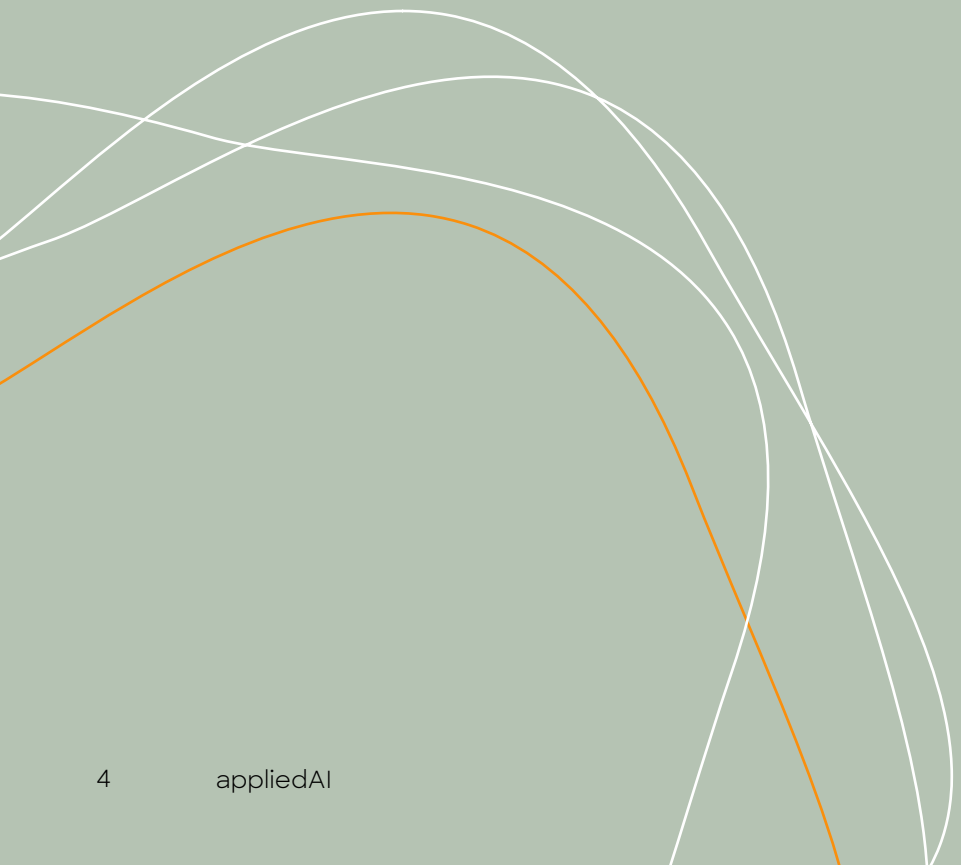
Contributing Partners 32

About appliedAI 33

Executive Summary

In this whitepaper, we introduce the Machine Learning Skill Profiles framework as a comprehensive organizational blueprint to scale machine learning in enterprises. Scaling enterprise ML refers to the efforts that aim to create high-quality and reliable systems that provide value in production environments while adhering to trustworthiness and compliance principles.

With interviews and multiple working group sessions with practitioners in leading corporations in Germany, we identified ten different roles that contribute to a machine learning project throughout its lifecycle. We name these skill profiles and discuss their organizational embedding, responsibilities, skills, and educational requirements. With these insights, we propose a framework that - we hope - helps managers and practitioners who are building machine learning teams in their organizations do so more effectively.



Introduction

Machine learning has found its way into many industries. Organizations that employ machine learning in production can create significant value and achieve a competitive advantage. However, it's not only one ML model that will make the difference, but hundreds or thousands of models in production. While enabling new use cases and business models, **organizations are currently struggling to scale the amount of self-developed machine learning models in production** with a high rate of project failure[1].

One major obstacle that companies face is the lack of guidelines to define the right skill composition of projects that go to production. Additionally, AI projects in production are inherently interdisciplinary with evolving skill requirements throughout the lifecycle. As the complexity and scale of ML projects grow, so does the need for robust infrastructure, seamless deployment, and efficient management of data and models in production environments. As a result, successful projects require domain knowledge, expertise with data, algorithms, infrastructure, and software engineering capabilities, among others. This complexity might result in many blind spots in practitioners and a lack of knowledge on what skills are necessary to successfully scale ML in the enterprise.

To address these challenges, this paper introduces the ML Skill Profiles framework. This enables the targeted composition of teams capable of facilitating the scale of ML in the enterprise. The framework covers the ten most common skill profiles found in successful machine learning teams. The framework is based on insights from professionals in the industry and is intended for managers who seek to create or scale their machine learning teams.

This report is the result of the appliedAI MLOps Working Group and is based on the experiences of leading experts from appliedAI's partner companies. This working group consisted of multiple sessions with representatives of ten large companies in Germany who followed a guided exchange about their MLOps practices. Additionally, we conducted expert interviews with working group participants and also with external experts.

The paper starts with laying out the challenges that occur when companies establish their machine learning teams. Subsequently, we discuss how these challenges are mitigated by introducing Machine Learning Skill profiles. We then proceed with an in-depth discussion of the Machine Learning Skill Profiles and considerations to implement the framework within your specific organization.

1 https://www.gartner.com/en/newsroom/press-releases/2020-10-19-gartner-identifies-the-top-strategic-technology-trends-for-2021

Organizational challenges to scale AI in the Enterprise

Companies face three main organizational challenges when scaling out their machine learning capabilities. These challenges are (1) the inherent interdisciplinarity of machine learning in production, (2) the lack of detailed guidelines with the specific skill sets to cover all the important ML lifecycle aspects, with the resulting possibility of skill gaps among the participants in the project and (3) the evolving required skill set per phase in the lifecycle of a machine learning project.

Machine learning projects are inherently interdisciplinary

AI can only bring value if it is successfully deployed to production. A major challenge is that researchers and educational programs focus mostly on the development of AI projects under research conditions: They often have a fixed standardized dataset, they work alone or with few researchers, they focus on iterating over multiple model architectures, and after the model is performing as expected there is often no actual model deployment and monitoring.

Unlike AI prototyping, in production systems, the datasets are rarely fixed, but they are continuously changing and often big. In addition, you rarely work alone but in teams where actual deployment and monitoring of AI models take place. As a result, AI in production requires a large number of interdisciplinary activities in addition to model training, such as robust data and compute infrastructure, efficient data, model and deployment management, seamless deployment, and continuous monitoring, among many others. Moreover, machine learning projects usually include large experimental components that demand expertise in project management methods that mitigate the risks and uncertainties of AI projects. As a consequence, project contributors need to be well versed in domain knowledge, the applicable algorithms expertise with data, algorithms, infrastructure, project management, and software engineering capabilities, among others.

Lack of detailed guidelines to define the right team composition

In the existing literature regarding roles in AI projects, we identified two clusters of publications. On the one hand, we have articles providing a good overview but a high-level description of the different roles. For example, the Azure MLOps Accelerator[1] provides a short overview of the key roles and responsibilities in an ML project. The

1 https://microsoft.github.io/azureml-ops-accelerator/0-GettingStarted/2-KeyProjectRoles.html

same is true for Domino Data Labs[1], The Analytics Club[2], MLOps: Overview, Definition and Architecture[3], and Krystian Safjan's blog[4]. These articles provide a high-level overview of roles in an ML project, but lack considerations for all the factors at play, such as what skills are needed per role, what are the responsibilities along the ML lifecycle and what constitutes the required technical knowledge.

Similarly, the RACI matrix from the book What is MLOps[5] provides a nice framework for an overview of responsibilities per role, but does not provide enough granularity about the separation of tasks and roles.

On the other hand, there is literature that provides more details. For example, in the book Fundamentals of Data Engineering[6], Joe Reis provides some more details in the descriptions of the technical profiles associated with data engineering, i.e. its upstream and downstream counterparts. Neptune's article on how to build ML teams[7] lists relevant skills, responsibilities, and technical knowledge for five roles in an ML team. However, it leaves out important roles related to data governance, data architecture, and AI architecture.

Overall, the existing literature provides high-level descriptions for a subset of roles and responsibilities relevant for AI in production. However, it does not establish a detailed guideline with the skills, efforts and responsibilities along the ML lifecycle for all the relevant AI in production roles. In addition, there is little information on how to apply these guidelines to your organization, where factors such as AI maturity, company size, and desired degree of centralization are important.

Due to the lack of guidelines about the right skill composition, it becomes more difficult for managers to hire the right people to cover all the important ML lifecycle aspects, leaving practitioners with blind spots when it comes to building ML teams. Lack of experience and skills might result in wrong decisions, delaying or impacting a project's chances of success.

The evolving skill requirements throughout the lifecycle of a machine learning project

Throughout the project lifecycle, the machine learning project traverses different phases. It starts with project scoping and then continues with data engineering, model development, deployment, and monitoring. You can find a detailed discussion of the machine learning lifecycle here[8].

Each of the lifecycle phases demands different skill sets and mindsets from different disciplines to be completed with a suitable quality standard. For example, initial exploration phases require fast iterations and outside-of-the-box thinking to demonstrate the feasibility of a project. Later productization requires maintainable and documented code that follows best practices.

To mitigate these challenges, avoid blind spots and enable staffing that corresponds to the needs of the respective project phases, we establish the **Machine Learning Skill Profiles framework.**

1 https://domino.ai/blog/7-roles-in-mlops
2 https://www.the-analytics.club/mlops-roles-and-responsibilities/
3 https://arxiv.org/ftp/arxiv/papers/2205/2205.02302.pdf
4 https://safjan.com/roles-in-mlops/
5 https://www.oreilly.com/library/view/what-is-mlops/9781492093626/ch04.html
6 https://go.redpanda.com/fundamentals-of-data-engineering
7 https://neptune.ai/blog/how-to-build-machine-learning-teams-that-deliver
8 https://appliedaiinitiative.notion.site/The-ML-lifecycle-396fadfaf2f14ae3b8765860c9f61ef7

The Machine Learning Skill Profiles Framework

We have established that machine learning projects are interdisciplinary and that few people possess all the skills necessary to complete a project end-to-end without help from others. In this whitepaper, we propose a framework of ten Machine Learning Skill Profiles to mitigate the organizational challenges to scale AI in the Enterprise.

Benefits of ML Skill Profiles Framework

These clearly defined Machine Learning Skill Profiles help you to build, scale, and operate machine learning projects and project teams in three distinct ways:

- ✓ Separation of Responsibilities
- ✓ Project Staffing
- ✓ Upskilling and Hiring

Separation of Responsibilities
Well-defined machine learning skill profiles and associated responsibilities are reference points during machine learning projects. They establish a **clear separation of responsibilities** within the team, delineating the responsibilities of each team member. By assigning distinct responsibilities to specialized roles, organizations can ensure that every aspect of the machine learning lifecycle is handled efficiently and effectively. This does not mean siloed contribution to the project but clearly defined ownership of different tasks that occur within a machine learning project.

Project Staffing
Having a standardized process to structure AI projects enables organizations to scale their ML activities. When an organization establishes a machine learning project with the aid of defined skill profiles, it gains an overview of the organizational blueprint, which enables efficient and effective project staffing.

Having a standardized approach across projects enables efficient communication, promotes reusing best practices in the enterprise, and enhances the overall operational efficiency of machine learning projects.

Upskilling and Hiring
The ML Skill Profiles framework aids in uncovering organizational skill gaps and helps in developing a necessary hiring plan. Having standardized ML Skill Profiles is useful for structuring your ML teams, writing job descriptions, getting people interested in what you do, and hiring. Moreover, it supports your specific upskilling and development opportunities for existing employees.

Deep Dive into the Framework

Now that we discussed how the machine learning skill profiles help you manage your machine learning projects, we will dive into the framework. Therefore, we will first discuss an overview of the skill profiles, and then we will dive into each skill profile in detail. Finally, we will briefly cover the methodology used to extract the results and the framework limitations.

Framework Overview

The Machine Learning Skill Profiles Framework is a comprehensive organizational blueprint to scale machine learning in enterprises. It consists of a set of skill profiles with well-defined responsibilities and skill sets along the ML lifecycle. Scaling machine learning in enterprises refers to the efforts that aim to create high-quality and reliable systems that provide value in **production environments** while adhering to trustworthiness and compliance principles.

Projects that will not be deployed into **production systems** and have different targets, such as proof-of-concept or projects in early start-ups, might only need a subset of the skill profiles described here.

For organizations seeking to scale up the number of projects that will be deployed into production systems, we identified ten skill profiles that contribute to a machine learning project throughout its lifecycle. These skill profiles are the following:

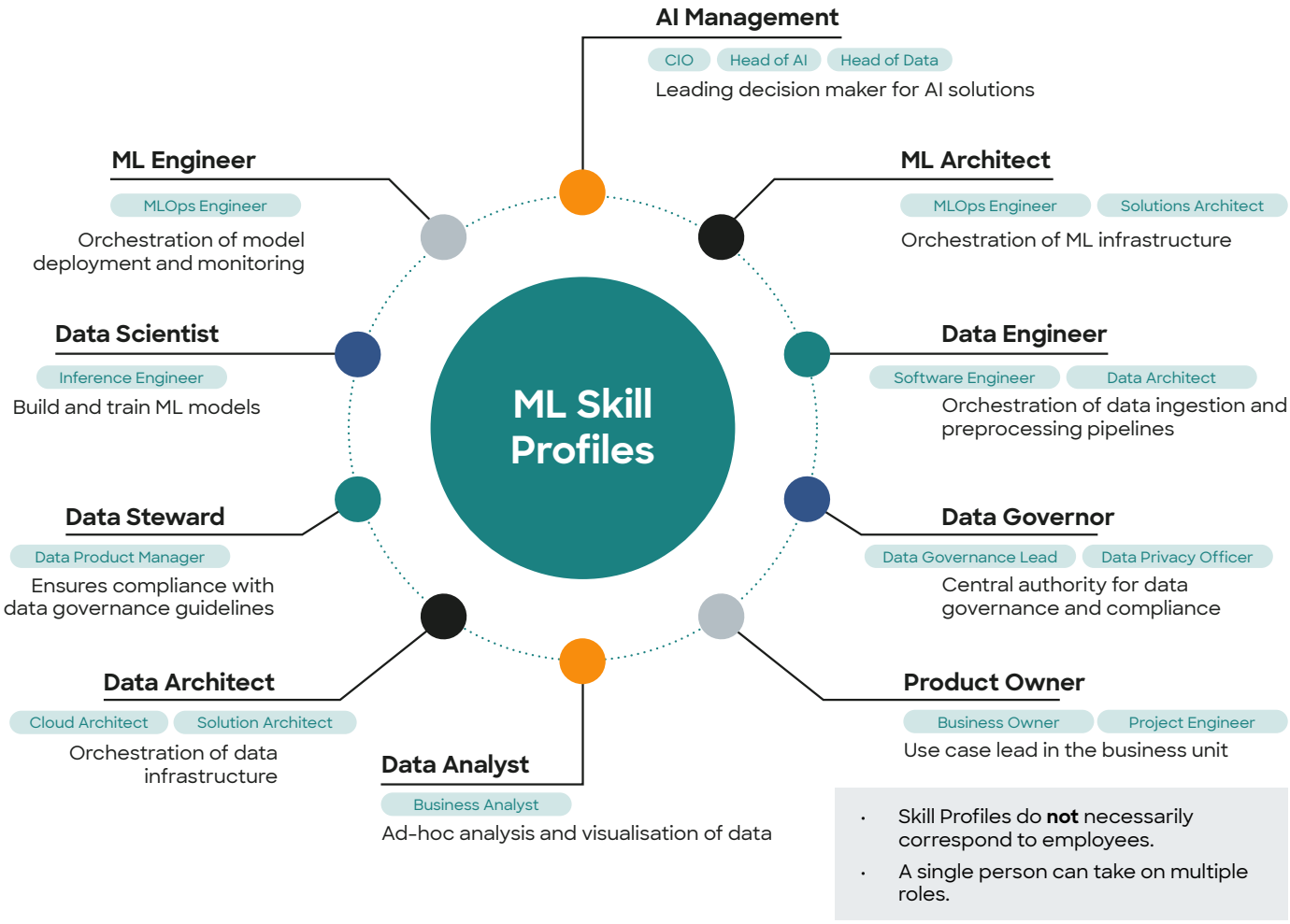


Figure 1. Overview ML Skill Profiles in an organization: Exemplary Skill Profiles, alternative titles and task descriptions.

AI Management defines the vision and ambitions of machine learning projects in the organization. They organize and greenlight the necessary resources and ensure the alignment of individual projects with the guiding vision.

Data Governors establish and maintain the organizational guidelines for the handling of data in the organization to ensure the data is usable, of high quality, and compliant with regulations.

Data Architects design and manage the organization's central data infrastructure, enabling data engineers and other data practitioners to use it effectively for their specific use cases.

Machine Learning Architects create and maintain the central machine learning infrastructure, providing tools for model management and ensuring that the infrastructure can support various deployment requirements.

Product Owners are pivotal in managing specific machine learning projects. They own the task prioritization and coordinate a swift delivery of value of the project by aligning the different stakeholders and mitigating the arising difficulties.

Data Engineers work within project teams to integrate and prepare data according to the standards set by data governance and using the infrastructure designed by data architects.

Data Scientists are tasked with developing predictive models by translating business problems into machine learning challenges, performing data analysis, feature engineering, and model training.

Machine Learning Engineers are responsible for deploying models into the production environment, implementing CI/CD/CT pipelines, and integrating automated monitoring of model performance.

Data Stewards are responsible for enforcing the governance guidelines on a project level. They manage the data as a valuable asset via the application of product management methodologies and ensure compliance with data governance guidelines.

Data Analysts focus on ad-hoc analysis and statistical examination of data, providing insights through visualizations and reports to support business decisions. Their work is often ad-hoc and not integrated into a production pipeline.

As mentioned above, these skill profiles contribute to a machine learning project at different stages within its lifecycle. An overview of the relative effort by project phase is displayed in **Figure 2**. Our findings corroborate and extend what is found in existing literature.

Embedding the Skill Profiles within the organization

The different skill profiles are embedded within the organization in different ways. We can characterize them by their degree of centrality within the organizational structure. Some skill profiles are more centralizable than others. That means they can provide central guidelines and direction to other actors in the organization. Moreover, we also have less centralizable skill profiles. They are typically located within a business unit and work on a specific use case or project.

On the one hand, AI Management, Data Governance, Data Architect, and ML Architect tend to be **more centralizable** functions that do not solve a direct customer problem but interact with the use case teams. Namely, they provide strategic guidance and resources, compliance guidelines, or infrastructure. However, depending on the organizational needs, company size, flexibility, and reusability requirements, these functions can be central for a whole organization (implemented via a centralized or hub-and-spoke model) or decentral for each business unit.

On the other hand, the Product Owner, Data Steward, Data Engineer, Data Scientist, Data Analyst, and ML Engineer are **less centralizable** functions and are usually deployed in a business unit as they work on a specific project (use case) with a specific goal. They take care of the actual implementation of a use case based on its individual needs.

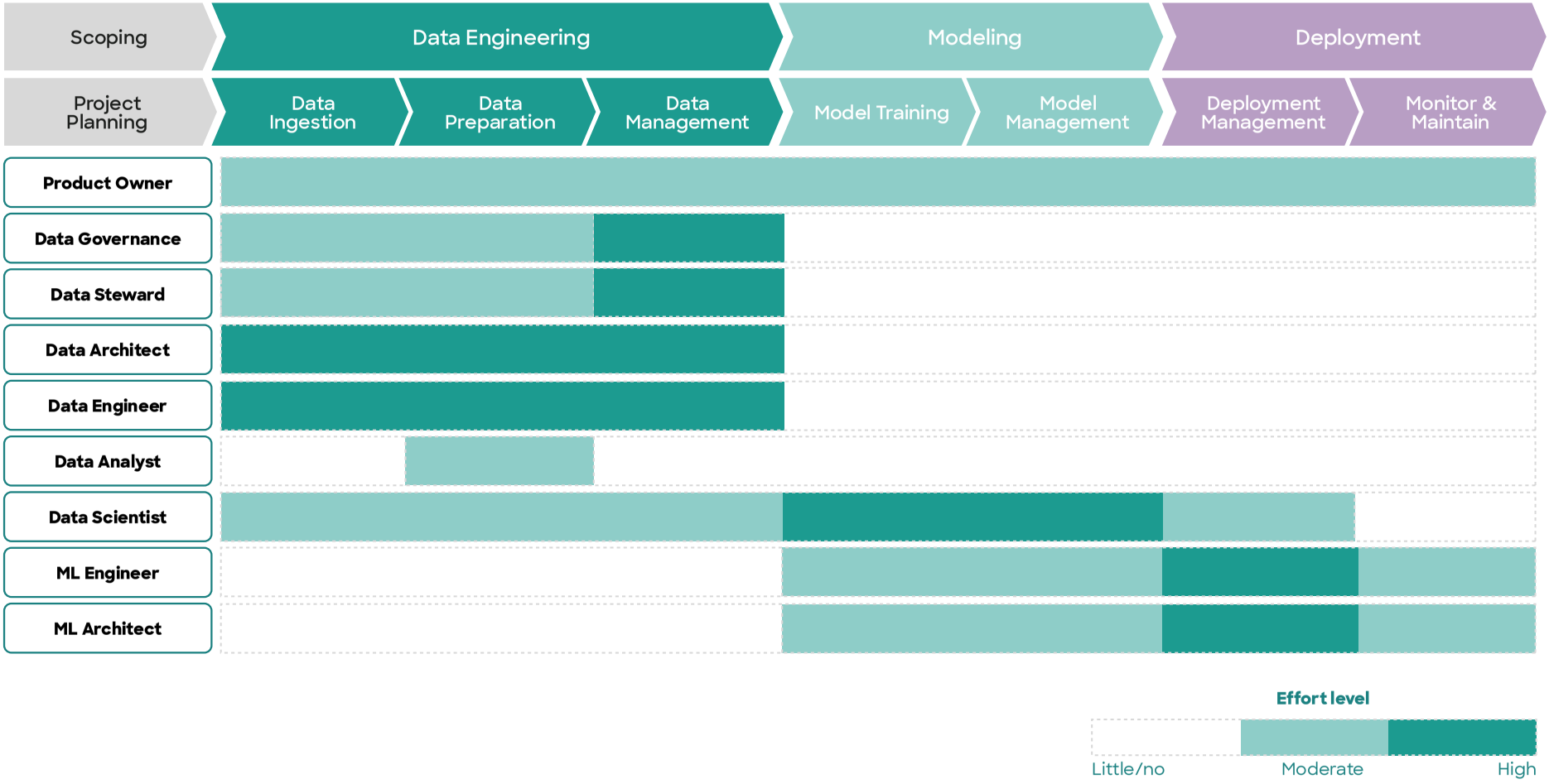


Figure 2. Relative Effort of the skill profiles along the machine learning lifecycle. AI Managers are not included as the time and intensity of their involvement are highly specific to each organization.

Mapping skill profiles
One can map the dependencies between the more centralizable and less centralizable skill profiles depending on their respective responsibilities in the machine learning lifecycle. For example:

- AI Management defines a central imperative strategy. The Product Owner is typically located within a business unit, manages the use case team, and ensures they are aligned with the strategy developed by the AI management.

- The Data Governance Officer sets central guidelines for data quality and compliance, while the Data Steward ensures compliance with respect to the centrally defined data governance guidelines within a project.

- The Data Architect provides the infrastructure to get access to the relevant data and work more efficiently. Data Engineers, Data Scientists, and Data Analysts are downstream users of the infrastructure provided by the Data Architect.

- The ML Architect provides central infrastructure to deploy and monitor the production models. The ML Engineer uses this infrastructure to deploy and operate AI models in production.

These relationships are displayed in **Figure 3.**

Decentralized team structure
However, it is important to note that while this is the most common organizational pattern, other structures can make sense depending on the specific situation of the organization. Generally, centralizable functions streamline the operations with central management of infrastructure and guidelines. This helps people in the business units with guided instructions. However, this comes at the cost of less flexibility. Thus, organizations with vastly different use cases or limitations to sharing knowledge among business units might value flexibility and the individual choice of tooling and infrastructure in the business unit more than increased efficiency due to a more uniform solution landscape throughout the organization. These organizations tend to follow a decentralized team structure, where some or all of the centralizable functions are located in the business unit

Pool of experts
Another common deviation from the pattern described above is the establishment of centralized pools of Data Engineers, Data Scientists, and ML Engineers who are ad hoc deployed on projects as needed. This way, companies may develop a greater number of use cases in a specific time. However, managing the ownership of production maintenance is more difficult, and the constant shuffling of personnel introduces friction to their ability to develop use-case-specific domain knowledge.

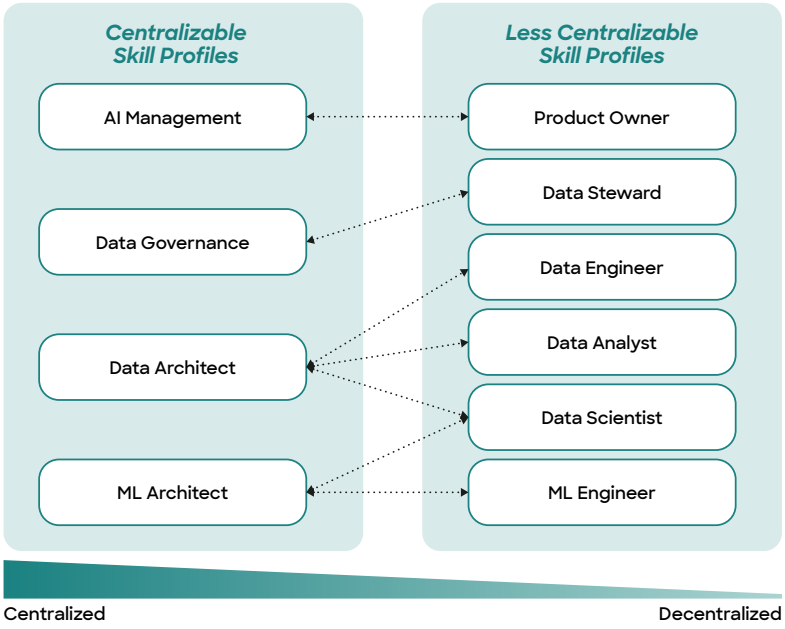
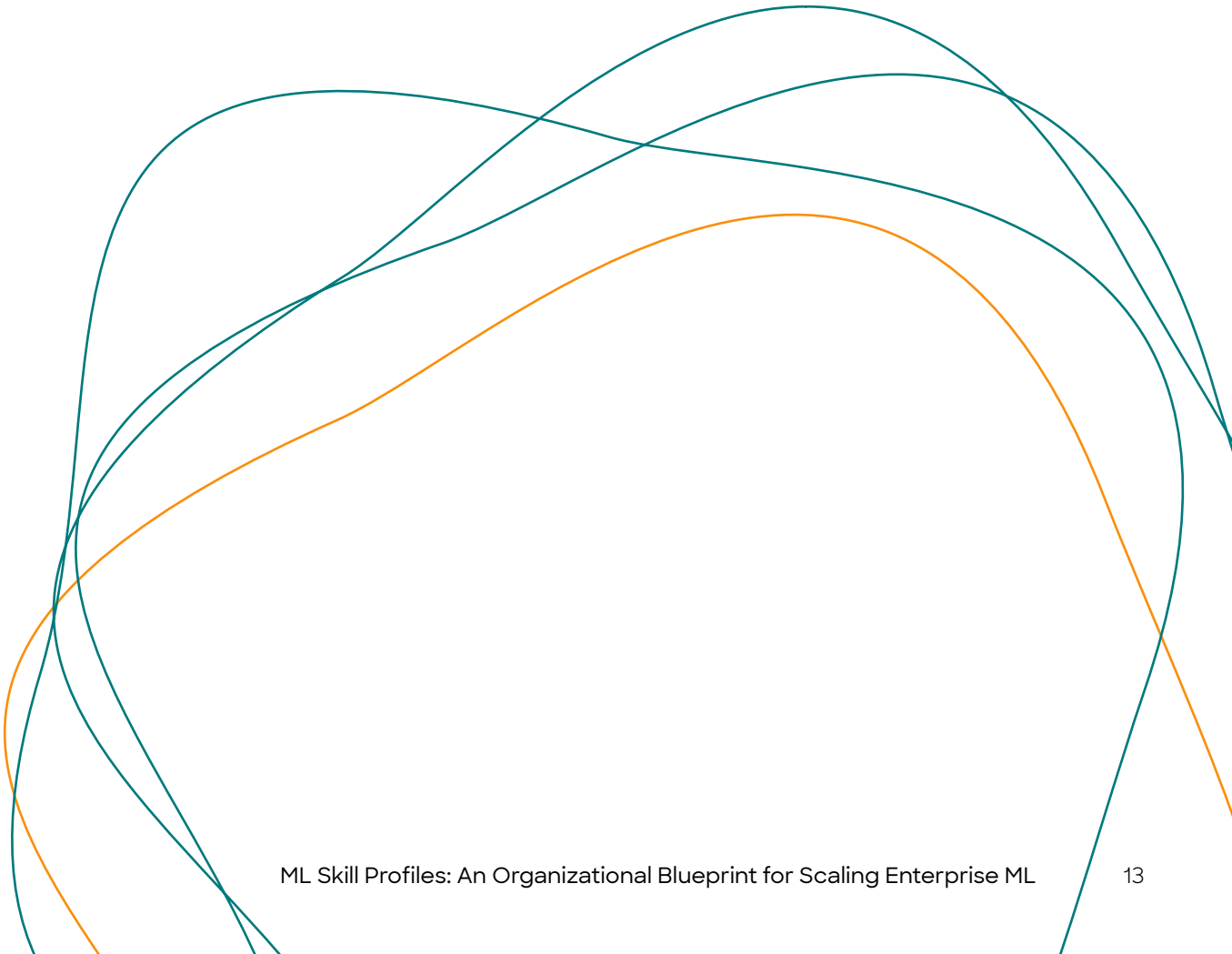


Figure 3. Centralizable Skill Profiles

We will base the following discussion on the individual machine learning skill profiles on the organizational division with centralized architect, governance, and management skill profiles and with the decentralized business unit skill profiles we presented above.



Centralizable Machine Learning Skill Profiles

AI Management

AI Management are executives in the respective organizations responsible (and accountable) for successfully defining the company data and AI strategy, which serve as guidelines for downstream decisions. Their main goal is to enable value creation through the efficient implementation of the use case suite in production systems. They are responsible for identifying and systematically mitigating bottlenecks and defining the best AI and MLOps team structure for the given organization constraints (central, decentral, hub-and-spoke, among others). Some organizations implement this role through the Head of Data, Head of AI, or CIO (Chief Information Officer) role.

AI Management is involved in the use case prioritization, identifying and strategically committing resources to relevant business opportunities to fulfill the strategy and company vision. This includes the oversight and budget authority for talent acquisition and development. They provide insights and guidelines for downstream skill profiles (such as the data architect) in order to design the best-fitting data architecture, data persistence strategy, and ETL vs. ELT, among others. Additionally, they supervise use cases and MLOps KPIs jointly with the product owner (and individual contributors) to ensure that any of the KPIs align with the company vision.

They also act as management-level evangelists for AI's promises (and limitations). They ensure that, on the one hand, relevant use cases in different branches of the organization are identified. On the other hand, they manage overinflated expectations about the capabilities of AI.

People in AI Management typically hold an advanced degree either in business, technology, or science. They have had an extensive tenure as an individual contributor before they took on management roles in which they developed expertise in data, AI, people management, and the development and execution of organization strategy.

Typical skills for AI Management

Domain knowledge, professional experience in AI industry applications, data quality understanding, innovation, roadmap development, agile project management, business understanding, sales

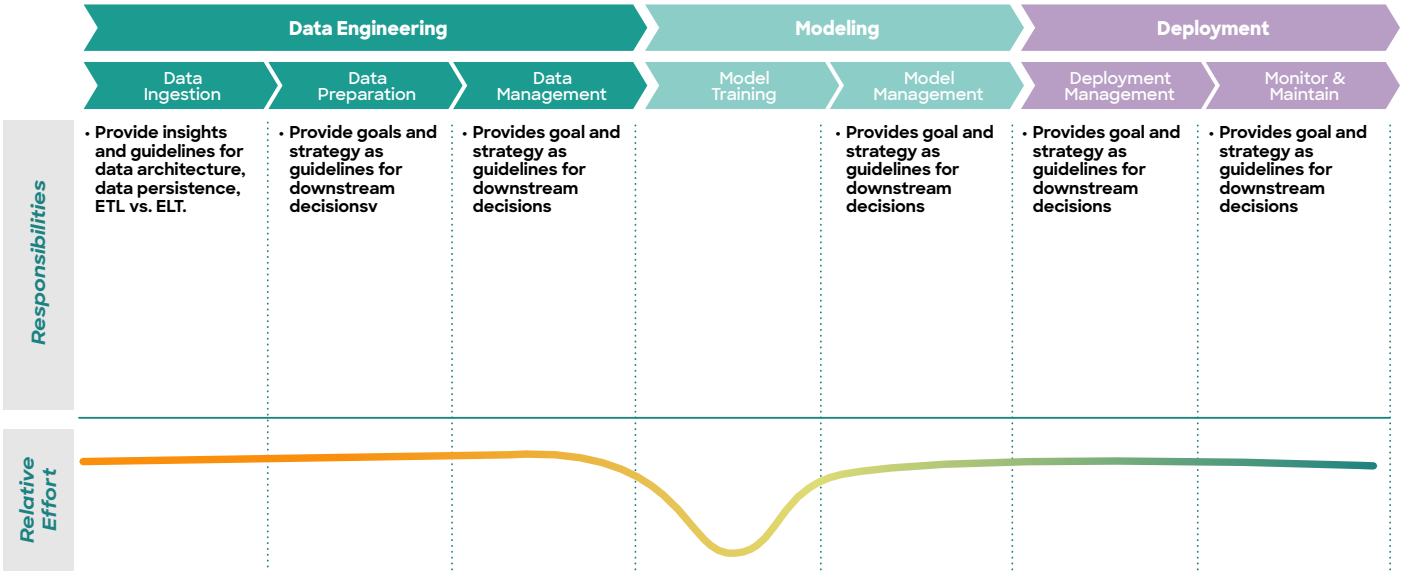


Figure 4. Objectives of AI Management per ML Lifecycle Phase

Data Governor

The Data Governance Officer or Data Governor is an authority within an organization responsible for ensuring compliance with the standards, regulations, and policies regarding data. They are responsible for establishing and maintaining the organization-wide data governance framework. This framework outlines the policies, guidelines, and procedures that enable efficient and responsible data management, treating data as a product, enhancing data quality, and ensuring data privacy, protection, and security.

On the most basic level, this entails setting central standards for data quality and versioning conventions to ensure consistency and interoperability. It also concerns data access, data retention, and data sharing policies. These guidelines serve as input for the data architect in designing and developing the department or company data platform.

The Data Governor also defines and conceptualizes data cataloging guidelines. They oversee the implementation of a central data catalog for data discoverability and lineage. Following other relevant stakeholders and the organizations' AI strategy, they decide on a cataloging architecture (push vs. pull), the necessary maturity of a dataset to be cataloged, and how to include lineage.

Many companies tend to centralize this skill profile in order to have a consistent data governance strategy across business units. However, it can also be implemented per business unit, with the drawback of having different guidelines within the organization.

Concerning the lifecycle, data governance is most prominent in the data engineering stage, in particular in the data management phase.

Data governance personnel typically hold an advanced technical or regulation-focused degree. They are well versed in the regulatory landscape as well as the basics of data analysis to gauge the implications of the policies on the usability of the data.

Typical skills for Data Governor

Ability to interpret regulation and derive operational requirements, responsibility, information system management, experience in product/data lifecycle management, analytical skills, communication skills

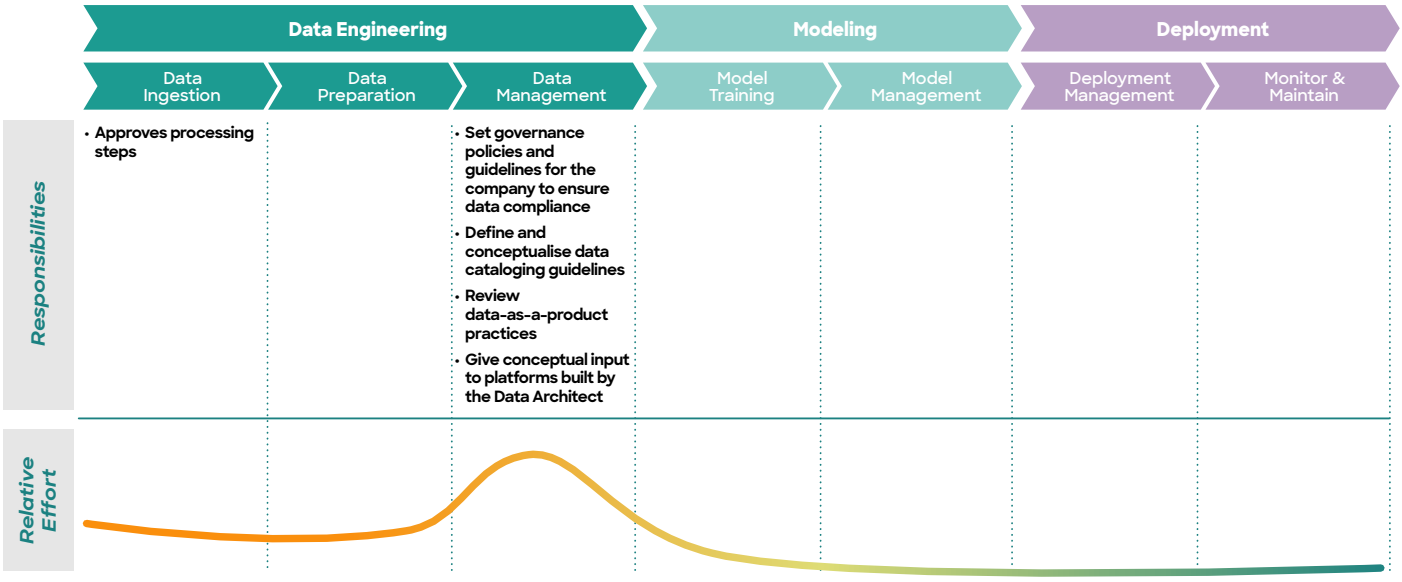


Figure 5. Objectives of Data Governor per ML Lifecycle Phase

Data Architect

The main objective of the Data Architect is to plan, design, and provide the central data infrastructure of the organization, keeping in mind the organizational requirements and governance guidelines defined by the Data Governor. This infrastructure is used by the data engineers and data scientists in the use case teams.

The Data Architects determine the appropriate data storage systems (e.g., Data Lake, Data Warehouse, Data Mesh, among others) and location (on the cloud, on-premise or hybrid). They design a suitable data ingestion pipeline (e.g., ELT vs. ETL), keeping data reusability in mind. They design and implement scalable data pre-processing infrastructure. Additionally, they are responsible for selecting the specific tool stack, supporting their colleagues in utilizing the provided data capabilities, and ensuring the data governance principles are technically embedded. They work closely with Data Engineers and Data Scientists in individual projects to ensure they can work effectively.

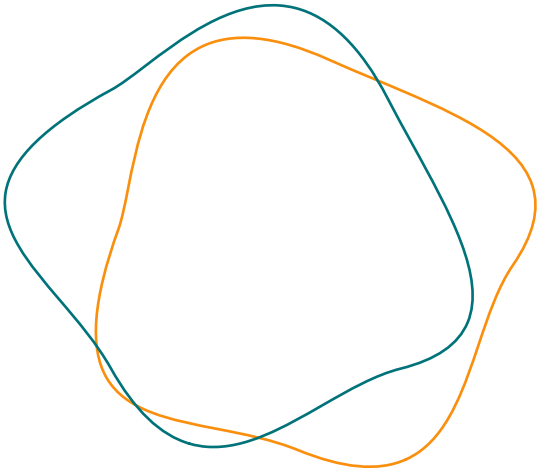
Regarding data management, they define processes and select tools for data quality validation, data versioning, data cataloging, data lineage, data as a product, and data lifecycle management (e.g., deletion at some point due to regulations). They work closely with the data governance team and data stewards to ensure data governance guidelines are followed.

The Data Architects allocate their contributions in the data engineering stage of the lifecycle, with a similar high effort during data ingestion, data preparation, and data management.

Data Architects usually hold a degree in a quantitative or information technology discipline. They have previous experience in data and software engineering, which enables them to centrally design and implement reusable, effective, and scalable data infrastructure to be used by the business unit teams.

Typical skills for Data Architect

Experience in design, set-up, maintenance, and documentation of IT systems and (big) data architectures and processes, data mining, data and metadata management, containerization, agile project management, analytical and communication skills



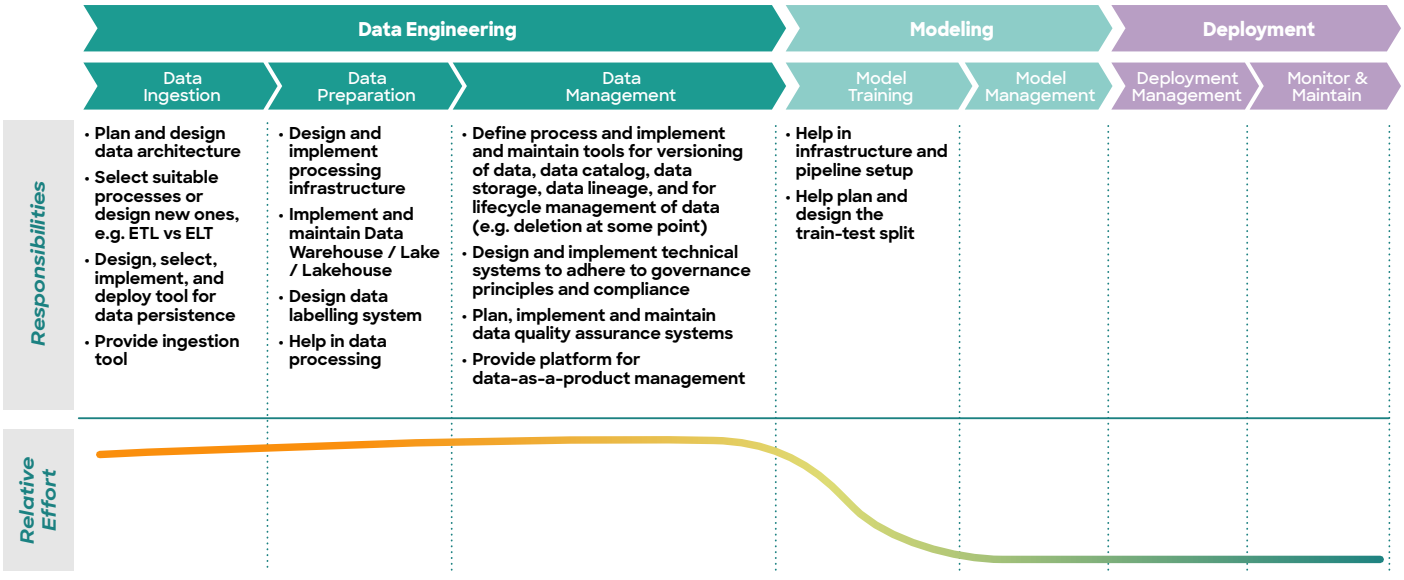


Figure 6. Objectives of Data Architect per ML Lifecycle Phase

Machine Learning Architect

The Machine Learning Architect is a centralized counterpart to the Machine Learning Engineer. They carry the responsibility to design, build up, and serve scalable machine learning infrastructure that encompasses model training, model management, model deployment, and monitoring. This infrastructure is typically used by Data Scientists, Machine Learning Engineers, and others in the use case teams.

The Machine Learning Architects design and implement scalable training infrastructure and define its location (on the cloud, on-premise, or hybrid). They design an easy way for it to be consumed by downstream users such as Data Scientists and ML Engineers. They are a sparring partner for downstream users.

In addition, their tasks include setting up and maintaining processes and toolings for model management, including experiment tracking, model versioning, and model registry, among others. They provide processes and tooling to validate models before they are deployed in production systems through staging environments (e.g., development, QA, production).

On top of that, they implement the central infrastructure for serving the models. Therefore, they ensure that requirements from the use case teams, such as batch processing, API serving, or real-time serving, can be met. They define processes, guidelines, and tooling that implement deployment strategies such as AB tests, canary deployments, and shadow deployments, among others. They also provide guidelines and maintain the infrastructure that enables the use case teams to perform continuous integration, deployment, (re-)training of their models (CI/CD/CT), workflow orchestration, and endpoint security.

Machine Learning Architects also provide the infrastructure for the monitoring and maintenance of deployed models and ensure the safety of the model endpoints. They implement and maintain processes and tools for data drift monitoring, model performance monitoring, system monitoring, business metric monitoring, system monitoring, alerts, and reports, among others.

While a Data Architect provides their work in the data phase of the lifecycle, the machine learning architect helps during the modeling and deployment stages. They provide reusable components and provision and maintain platforms for multiple business case teams.

Machine Learning Architects usually hold a degree in either an information technology or scientific discipline. They have extensive experience in software engineering, machine learning development, and the design of efficient computer systems. Additionally, they know the best practices in the field as well as relevant tooling and design patterns.

Typical skills for Machine Learning Architect

Experience with infrastructure as code, git, CI/CD, cloud development, test-driven-development, Docker/Kubernetes, MLOps processes and tools, experience in AI development, agile development

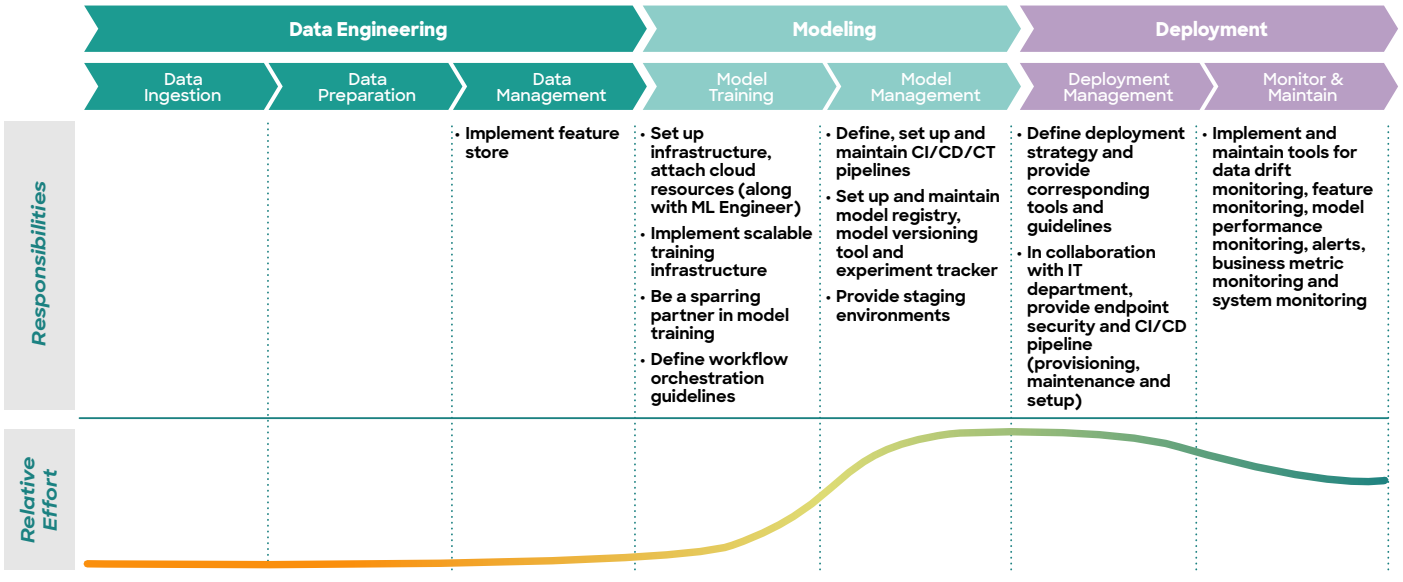


Figure 7. Objectives of ML Architect per ML Lifecycle Phase

ML Skill Profiles in the Business Unit

Product Owner

First and foremost, the Product Owner coordinates the efforts for an individual use case project in the context of the larger organization and its AI strategy. To this end, domain expertise is paramount and has to be developed if it is not already present.

Next, the Product Owner is responsible for gathering and prioritizing requirements. Therefore, they work in close alignment with the project's different stakeholders. Here, the final users are the most important (which could be in-house or external). In the case of different user types and profiles, it makes sense to characterize the different archetypes with idealized personas. Other important stakeholders are the AI Management for the alignment with the organization's AI strategy and the central IT services for infrastructure usage. Ultimately, Product Owners guide the development efforts and manage the project.

For agile projects, this means developing and prioritizing the user stories (feature requests) that make it into the project backlog. They are also responsible for sprint planning, backlog refinement, and developing acceptance criteria for user stories.

Throughout the project, the Product Owner is responsible for ensuring smooth operations. They can achieve this by contributing their domain knowledge (for example, when labeling data), collecting stakeholder feedback, mitigating conflicts, and other means necessary to align the team on a vision for the product and a strategy to get there.

Product Owners can have varying backgrounds. It is important that they have relevant experience in managing projects using agile frameworks. They know the company structure and can identify and align relevant stakeholders. Additionally, they need knowledge in the application domain and machine learning fundamentals. They should know their way around common project management software as well.

Typical skills for Product Owner

Business domain and ML expertise, track record of successful projects, project management software (Jira/Confluence), prioritization, agile development, communication skills

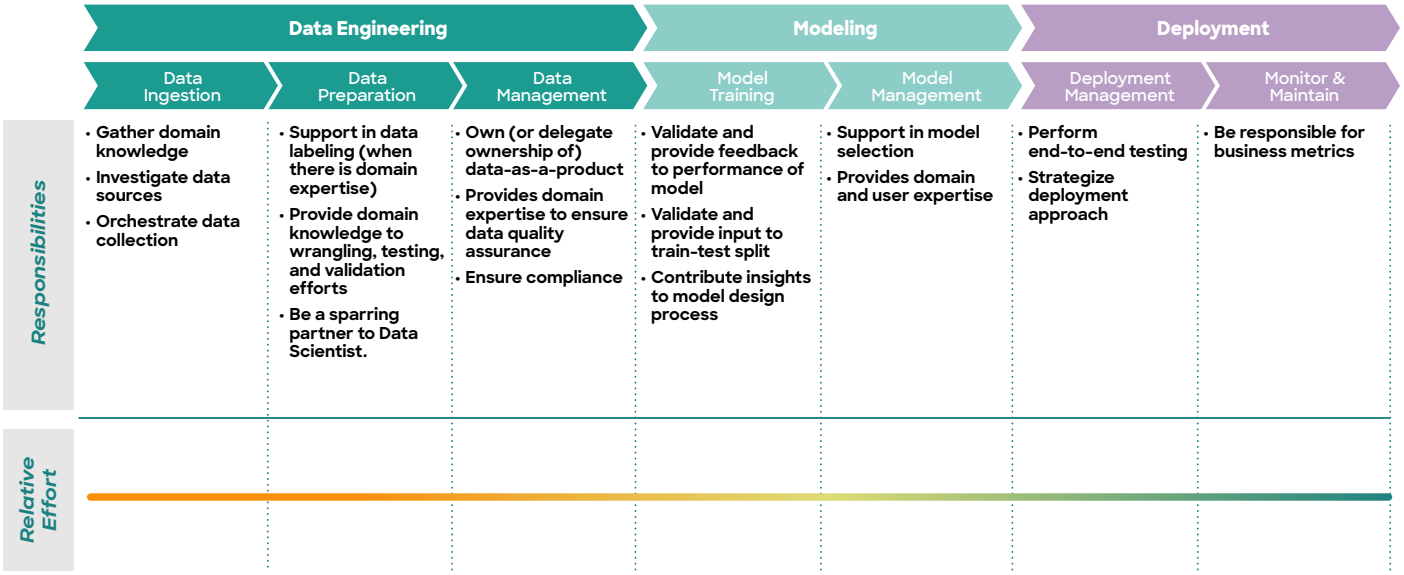


Figure 8. Objectives of Product Owner per ML Lifecycle Phase

Data Engineer

The main objective of the Data Engineer is to integrate and prepare relevant data to be used in a project. They do so by utilizing the centrally commissioned resources provided by the Data Architect, which were designed to adhere to the applicable guidelines and regulations imposed by the Data Governance team.

They perform the ingestion of datasets into a centralized data architecture by connecting to different heterogeneous data sources. They make sure datasets are persisted according to the guidelines and they engineer the data schema to be worked with.

They learn basic domain knowledge to perform general data cleaning transformations, data validation, and other preparation tasks before the data is used in a given use case. They process data in a scalable way by using the resources provided by the architect. They set up workflow orchestration pipelines to transform data sets.

Data Engineers follow governance guidelines. They are responsible for implementing and operating data versioning, data lineage, data lifecycle management, data storage, and data quality. In order to mitigate data silos, they register the mature datasets into discoverability tools such as data catalogs and lineage dashboards. When features can be reused by multiple projects, they implement and register features into feature stores. They hand over prepared datasets to data scientists.

Lastly, they might implement data drift monitoring once the use case is deployed such that diminishing performance of a served model due to underlying concept drift can be addressed early.

Regarding their efforts in the ML lifecycle, most of their contributions are in the early stages of data preparation and data ingestion phase.

Data Engineers usually hold a degree in a quantitative discipline, have experience with big data technologies, and may hold certifications for database management technologies.

Typical skills for Data Engineer

Cloud infrastructure, workflow automation, containerization, code control (git) and CI/CD, database management systems, Python, SQL, ETL/ELT design, orchestration tools, data modelling and data mining, data product design, agile development, analytical skills

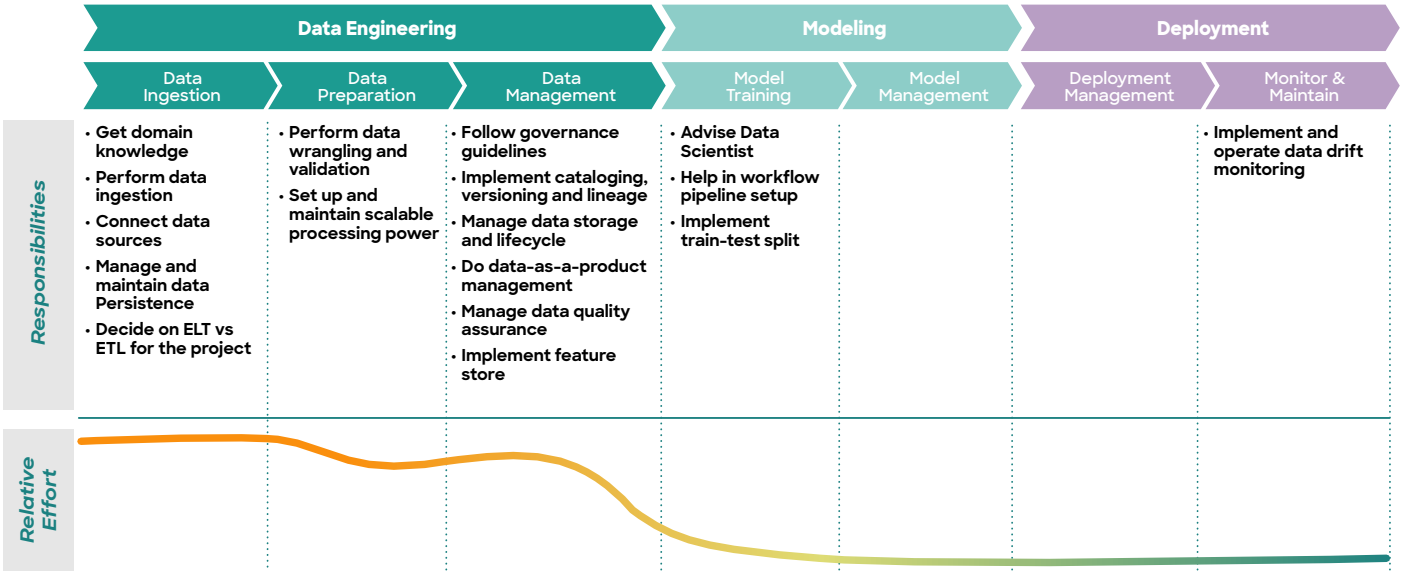


Figure 9. Objectives of Data Engineer per ML Lifecycle Phase

Data Scientist

Data Scientists are responsible for model development within the context of machine learning projects. For this, they translate the business problem into a machine learning problem, prepare datasets for the specific use case, and address model development with their machine learning toolbox.

To do so, they first gather domain knowledge and discover and examine relevant data sources. They perform exploratory data analysis, data wrangling, and data validation, followed by feature engineering and feature selection. For this, they work closely with the data engineer to define the shape and form the data has to be in to be effectively utilized for model training. For the produced use-case-specific datasets, they do data versioning, data quality assurance, data lineage, data catalog, and data storage.

Subsequently, they perform the model training. They are responsible for hyperparameter optimization, train-test split, score model performance, and selecting the best-performing model. They are also responsible for devising a retraining strategy and implementing it jointly with the Machine Learning Engineer.

Regarding model management, they use the platform or template provided by the ML Architect to perform model versioning, model registry, experiment tracking, and CI/CD/CT pipelines. They implement model inference and integrate it into the inference pipeline. They perform data drift monitoring, feature monitoring, model performance monitoring, and business metric monitoring.

The Data Scientist is often embedded in a business unit. With respect to the lifecycle, they are most active in the modeling stage. Nonetheless, they have close contact with the skill profiles involved in preparing the data (Data Engineer) and facilitating the deployment (Machine Learning Engineer).

Data Scientists usually hold an advanced degree in a quantitative discipline with a strong focus on data analysis and machine learning. They are well versed with the data science technology stack (mostly in the Python or R ecosystem) and have experience with one or more machine learning frameworks.

Typical skills for Data Scientist

Experience in data analytics, data mining, big data, statistics, Feature engineering, ML tools (Jupyter, Pandas, Numpy, TensorFlow, Keras, PyTorch, scikit-learn), MLOps tools (e.g. MLFlow), Python, perl, C/C++, databases, agile development, critical thinking, communication

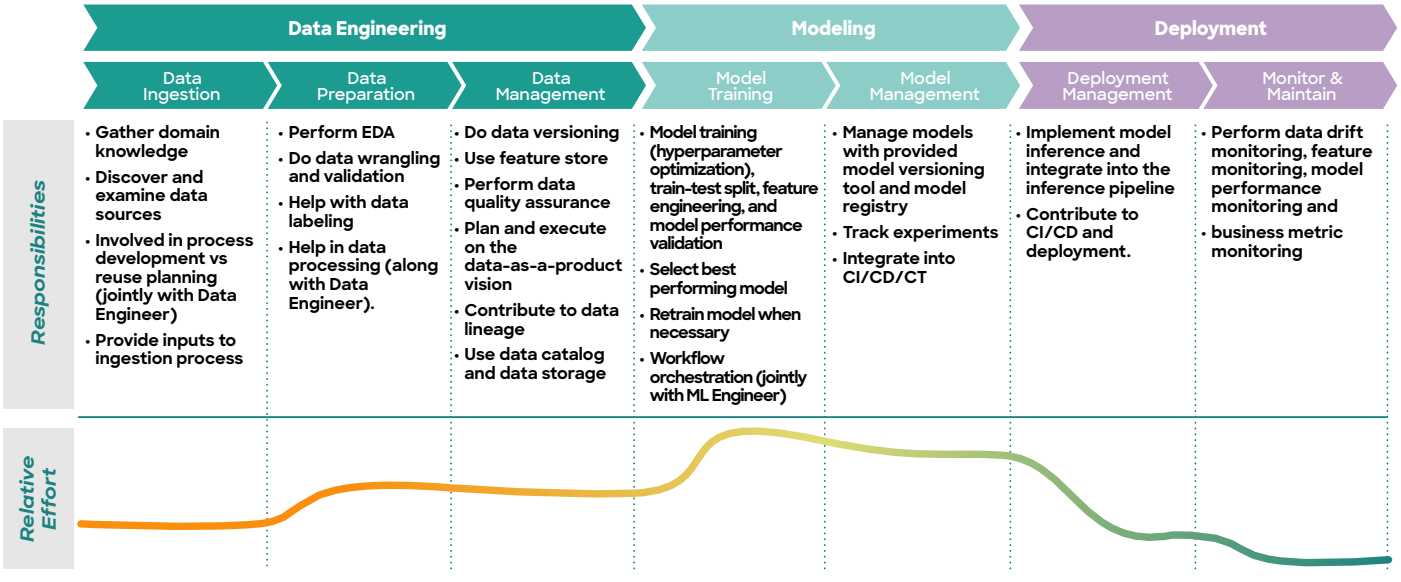


Figure 10. Objectives of Data Scientist per ML Lifecycle Phase

Data Analyst

The Data Analyst specializes in ad-hoc statistical data analysis of static datasets that seek to answer business hypotheses. Often, they also perform the cleaning and preprocessing of the data.

In effect, they primarily work on small-scale projects to provide fast answers to pressing questions. They communicate their findings with visualizations and reports.

To be effective, they usually work closely with management to catch questions immediately once they arise. They must have a solid understanding of their domain and business to embed and communicate their results effectively to executives.

Data Analysts differ from Data Scientists because they do not work on larger-scale automation projects to develop a model capable of predictive inference. Instead, they analyze historical data to find trends and support data-driven support for business decisions. Therefore, Data Analysts are responsible for small-scale projects without a deployment part. Thus, they are active in the data and modeling stage.

Data Analysts hold a degree from a quantitative field and are proficient in applying statistical methods to analyze data. To this end, they are capable of programming with a scripting language for ad-hoc data analysis and visualization..

Typical skills for Data Analyst

Foundational knowledge in maths and statistics, foundations in programming, KPI reporting, SQL, tableau, Excel, visualisation, communication skills

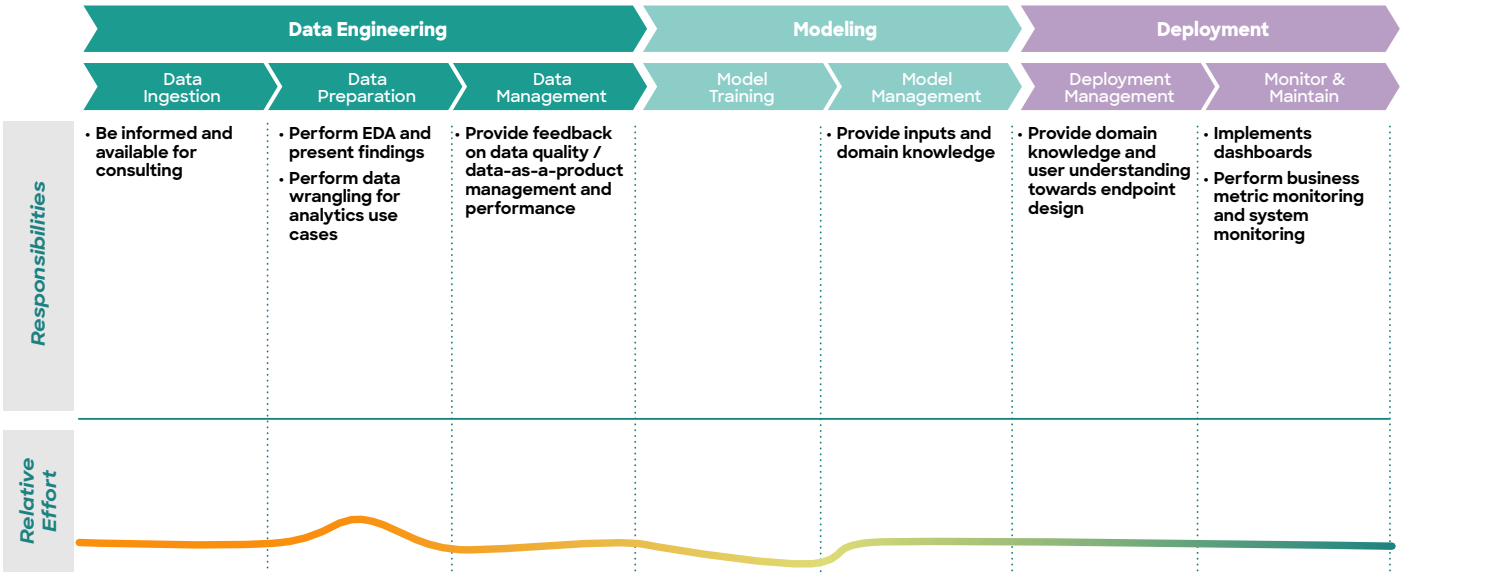


Figure 11. Objectives of Data Analyst per ML Lifecycle Phase

Data Steward

The Data Steward is the organizationally decentralized counterpart of the Data Governor in the use case teams. They are responsible for implementing and ensuring that a specific dataset adheres to the central governance guidelines defined by the Data Governor in terms of quality and metadata management. They can achieve this either by hands-on work or by fostering awareness about governance guidelines in their team.

They are also responsible for supervising the implementation of the data-as-a-product management practice for handling data. This approach aims to establish data as an organization's valuable asset and employ similar methods to product management around it. Namely, one sets a definite goal to answer the "why" for the data collection efforts. Additionally, they define the organization's data lifecycle. Then, data-as-a-product management prescribes guidelines that ensure the data is shaped to deliver the most value for a clearly defined user persona. As such, the process follows a user-centric approach. These principles influence the design of the data (e.g., variables, names) and the communication around it (metadata, organizational awareness about the dataset's existence). It also includes establishing the accessibility of the data for relevant stakeholders in the organization as well as the quantification and measurement of valuable KPIs to ensure the viability of the data as well as its alignment and embedding in the organizational AI strategy.

To continuously improve the dataset's usefulness, the Data Steward defines a roadmap for improving the relevant attributes of the dataset. The specific tasks are either executed by them or relevant colleagues. To achieve these objectives, the Data Steward must be an ambassador for user-centric thinking and educate their peers about practical steps to foster the evolution from within. Usually, Data Stewards are on the same level of hierarchy as their peers on the project.

The Data Steward has an activity profile in the machine learning lifecycle similar to the data governance skill profile. However, Data Stewards are more concerned with the specifics of a particular dataset or product and thus are more decentralized with respect to the organizational structure than the governance skill profile.

Data Stewards need to merge the business and compliance requirements with the technical demands of a dataset. Therefore, they come either from a business or regulatory background and have acquired substantial skills with respect to data analysis or, vice versa, have a background in technology and experience with regulation and product management that they can apply to the data.

Typical skills for Data Steward

Quality awareness, data management, experience in developing and maintaining data quality standards, data ethics, written and verbal communication and presentation skills

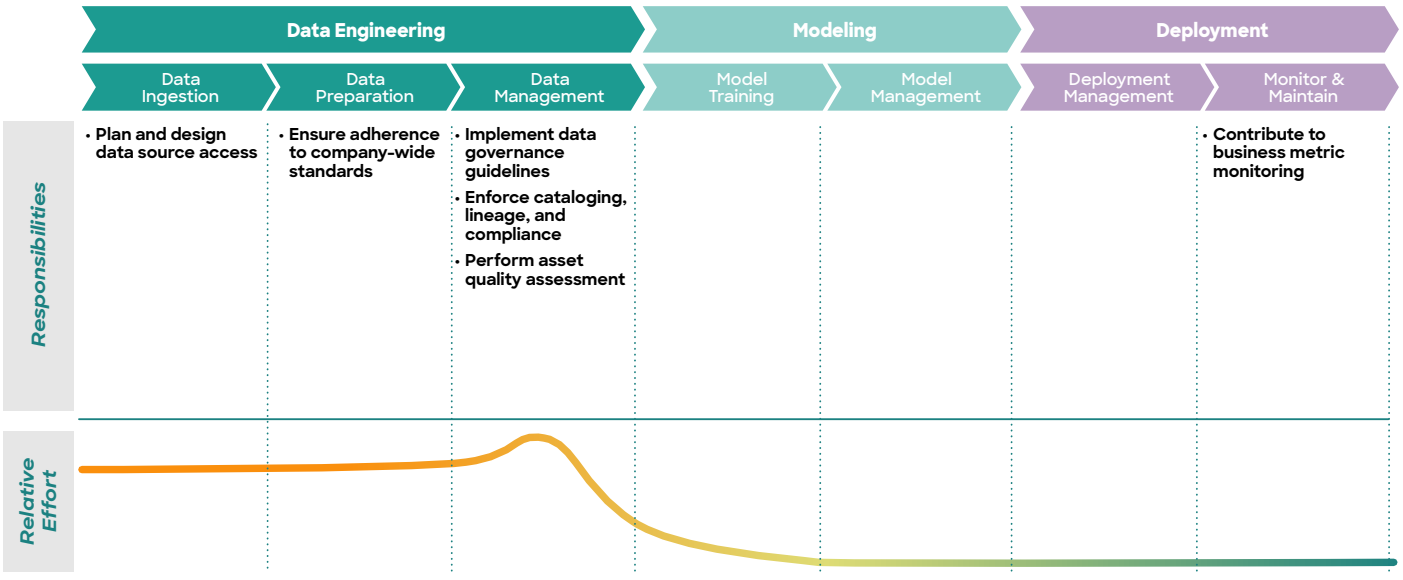


Figure 12. Objectives of Data Steward per ML Lifecycle Phase

Machine Learning Engineer

The Machine Learning Engineer is the counterpart of the Machine Learning Architect in the individual use case teams and is responsible for the software engineering aspects of the machine learning project.

After models are trained by a data scientist, Machine Learning Engineers take over the best-performing model from the Data Scientist and utilize the centrally commissioned infrastructure (provided by the Machine Learning Architect) to model into its productive environment and set up CI/CD/CT pipelines and model monitoring.

During the model development, they also aid the Data Scientist with the setup or attachment to centrally provided necessary infrastructure such as a model registry or a feature store. They also implement workflow orchestration tools to speed up the development.

As such, the Machine Learning Engineer skill profile is located in the business units and becomes active in the later stages of the machine learning lifecycle - mainly during the deployment and monitoring phases.

Machine Learning Engineers possess extensive experience in software engineering and the corresponding tools and technologies. Additionally, Machine Learning Engineers are well-versed in the experiment-driven development of machine learning models. As such, they are often either Data Scientists who have gathered additional experience in software engineering or software engineers who have developed an intricate understanding of data and the machine learning development process.

Typical skills for Machine Learning Engineer

Experience working with distributed systems and cloud services, ability to design, integrate, automate, and maintain ML monitoring infrastructure, containerization, scripting (BASH, Python), ML libraries, statistics and data science

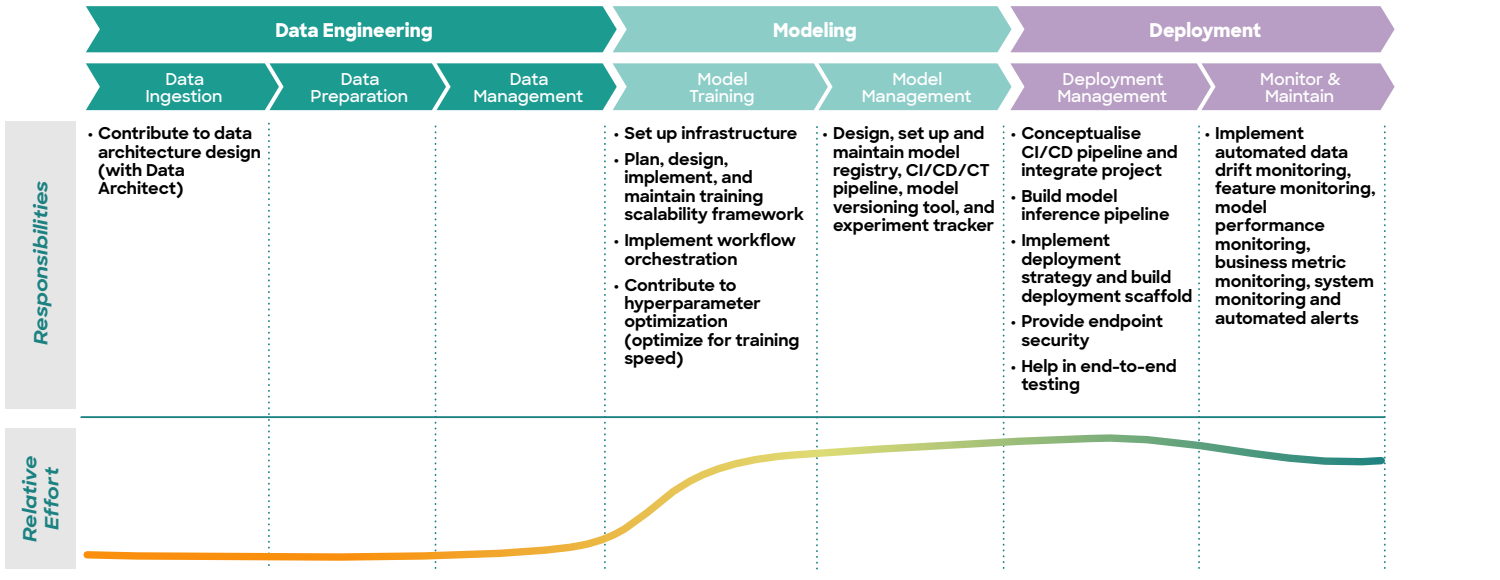


Figure 13. Objectives of ML Engineer per ML Lifecycle Phase

ML Roles - An expert's opinion

The following is an interview with Joe Reis. He is an expert in MLOps at scale with more than a decade of hands-on industry experience, an adjunct professor at the University of Utah, the author of the best-selling book Fundamentals of Data Engineering, and was a guest speaker at appliedAI's MLOps Day in Munich.

His work has been validated by thousands of data practitioners who consider his book to be a must-have for any data professional. We were happy to interview him for this whitepaper to get an additional expert opinion on roles and responsibilities during the lifecycle of an ML project. The article below is a summary essay on our conversation.

Q Why do you think ML roles are important?

A The essential purpose of roles is to establish the separation of concerns. Each team member then knows their responsibilities and also - maybe even more importantly - the boundaries of their work.
Without clearly defined roles, chaos will likely ensue. The result is confusion, frustration, and inefficiency.

Roles are also important in the hiring process. With them, you can attract individuals with the necessary skills and expertise.

Last but not least, roles facilitate effective communication and coordination among team members to allow for smoother project execution.

Q When does it make sense to have ML roles?

A Proper timing of the establishment of roles heavily depends on the nature of the business and its functions. But it's always a good idea to be aware of future needs and to draw the organizational chart early on. For example, for data-driven businesses establishing data-related roles early is crucial. It's a common mistake to early hire data scientists who then have little to no data to work with. The hiring process should be driven by the specific needs of the business.

Joe Reis
Data Engineer & Architect | Author of best-selling book "Fundamentals of Data Engineering" | Professor

Q Are there common pitfalls, tar pits, or anti-patterns you have noticed when designing ML teams?

A One should avoid designing the team without a previous clear understanding of the expected responsibilities. Otherwise, it can result in confusion, inefficiency, and frustration. People expect a certain career development and task spectrum and leave if the role is vastly different than what was promised.
Another problem is utilizing a one-size-fits-all approach. The team's roles and responsibilities should be primarily driven by the company's needs - which you have to understand - rather than any one framework. That being said, frameworks can help you understand your specific needs and avoid blind spots.

With regards to AI in general, one should not apply it just for the hype. It's essential to develop proficiency in analytics first to also understand the potential (or lack thereof) of the data available.

Q Which roles should a company hire first?

A There needs to be a business- and purpose-driven approach. The first one through the door should be somebody who can work with the business and identify applications of data. Next would be a Data Analyst who interfaces with the manager and figures out ways to get the data, systems, and processes to do so effectively. The next role would be the Data Architect. At this point, most of the focus should be on building the data architecture. After that, Data Engineers and Data Scientists in order to implement the use cases. Finally, ML engineers leverage data to train and deploy AI models in production.

Q What are your thoughts on a centralized vs a decentralized team?

A It seems to work both ways. You need leadership support to pull off decentralization, since human tendency is to gravitate towards kingdoms (or centralized teams). Decentralized teams is a newer idea or different philosophy on how you do your business.

The pleasure to work in central companies is smaller than in decentral companies, where you have more freedom to try different approaches. And it is already hard to retain talent.

Nevertheless, having central guidelines is important in both cases. At the end it is about results and trade-offs.

We can see that Joe Reis, a practitioner with vast experience in the field, agrees with many of the points stated in our framework. The importance of having roles is paramount. We also find common ground about when it makes sense to have roles. Lastly, we agree on the fact that both centralized and decentralized team structures can work.

How to apply the framework

Our aim with the machine learning skill profiles framework is to assist practitioners in building or restructuring machine learning teams, encompassing the entire lifecycle of machine learning development. This framework serves as a guide for constructing roles within organizational structures, but it's vital to understand its limitations for effective application.

Consider your Organizational Specifics

While the framework is based on common patterns identified across various companies, it's important to note that it may not capture the unique aspects of every organization. Tailoring the framework to fit the specific context and requirements of your organization is key. The way of application depends on the organization and machine learning team size, the company's maturity regarding machine learning itself, and the desired degree of centralization of certain activities. We discuss this in more detail in the next section.

Avoid excessive discussions on Role Delimitations

Discussions about specific roles should remain productive and aligned with the broader goals of the project and organization. Overemphasis on role boundaries can lead to lengthy and unproductive debates and detract from the main objectives. A trial-and-error implementation of a specific role and its subsequent adaptation based on experiences in real-world scenarios are more grounded in actual circumstances than preceding theoretical discussions.

Additionally, colleagues should (to an extent) outgrow their skill profile boundaries and upskill in adjacent areas to enable effective collaboration and streamlined development.

Beware of Silos and facilitate Cross-Functional Collaboration

In the dynamic field of machine learning, fostering a culture that encourages communication and information exchange is crucial despite having defined roles. Understanding the interdependence of skill profiles in the larger scheme and being consistently mindful of the goals of a project and the mission of the organization is essential for success.

Adapting to Rapid Technological Changes
The field of machine learning is rapidly evolving. Consequently, role definitions need to be flexible to accommodate new technologies, tools, and methodologies. This flexibility ensures that teams remain adaptive and innovative.

Discussions on roles and organizational structures should always add value to the organization rather than create an illusion of control in a field characterized by dynamism and uncertainty. When applied thoughtfully, we believe the machine learning skill profiles framework is a valuable tool for building high-performing teams, adaptable to the dynamic landscape of enterprise-scale machine learning.

Considerations specific to your organization

The application of the Machine Learning Skill Profiles framework depends on different factors such as machine learning team size, AI maturity, and the desired degree of centralization of certain activities. We will discuss how each factor influences how you can apply the framework.

It is important to note that these three factors are not an exhaustive list. The ML Skill Profiles framework should (as any framework) serve you as a guideline and not constitute a dogma you cannot deviate from. You know your company's specific situation best. Tailoring the framework to fit the specific context and requirements of your organization is key.

After all these considerations, managers and practitioners can refer to the machine learning skill profiles discussed above as a "cheat sheet" or guideline and use the adequate building blocks to plan and embed them into their respective company structure.

Figure 14. Organizational factors

Let us now discuss the implications of company and division size, maturity, and degree of centralization.

Company Size and Division Size
A significant factor in how the ML Skill Profiles are implemented is the (often correlated) team and company sizes.

Larger companies, or companies with more people associated with the machine learning department, tend to have teams corresponding to each skill profile. Multiple people are fulfilling the tasks and are specialized more granularly. While team sizes are necessary to handle the workload, most people can or are trained to perform every part of the task associated with the team. While there might be some specialization, individual contributors, for the most part, do similar jobs on different projects.

In this case, the machine learning skill profiles help to define a standard level of expertise to which each team member should conform. If skills are not fully developed, it is easy to facilitate learning from peers.

Smaller companies have a smaller workforce; thus, a single person will likely fulfill multiple skill profiles to facilitate a project end-to-end. Here, one should closely examine the skill profiles to ensure that the individual contributor is equipped with all the necessary skills to bring the project to fruition. If not, one should hire outside experts or begin upskilling early to make the expertise available when required.

Machine Learning Maturity
Another factor that influences the application of the skill profiles is the AI maturity of a company. Generally, increasing the maturity of an organization is the long-term goal. However, sometimes, organizations can generate tremendous value at lower maturity levels. In this case, one should examine the machine learning skill profiles and their specific tasks and responsibilities. From these, one selects the relevant subset. Particular responsibilities may be ignored because the skills are missing, and you are confident the project can succeed without applying the corresponding principles.

For example, when you use small, simple, and interpretable models such as linear or logistic regressions, implementation into the production code is easier as no separate hosting for the ML model is necessary. Additionally, these models are easy to train. Hence, proper experiment tracking of the dependent and independent variables and the resulting coefficients may circumvent the necessity of a model registry.

Another example is if there is only one machine learning team and a static dataset, one can forego data versioning and lineage.

Consequently, one needs fewer skills to get the machine learning operation off the ground, and the limited workforce has the focus to learn the relevant skills and deliver value quickly.

Nonetheless, this is like playing with fire. As increased maturity should be the goal,

conscious omitting of best practices should be and remain conscious. Another use case often succeeds in successful machine learning projects. You collect data on performance in production and decide to try more sophisticated models. Suddenly, you find yourself in a position where your datasets are not static anymore, multiple projects depend on the same data, and on-the-spot retraining becomes annoying or infeasible. If that is the case, you have hopefully started upskilling individual contributors in advance so that the necessary systems can be implemented swiftly.

Degree of Centralization
Another company specificity is the desired degree of centralization. High centralization has the benefit of streamlined operations, centralized support, and the minimization of knowledge silos.

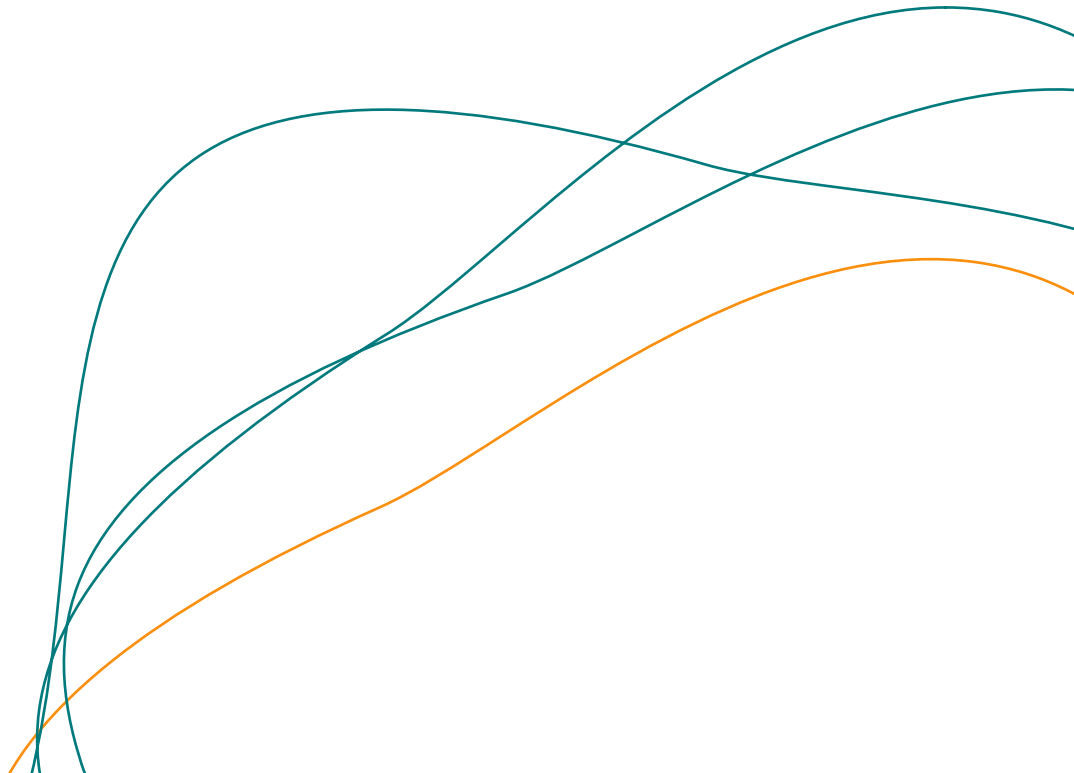
In this case, the architect's skill profiles become prominent. They commission and operate infrastructure and tools centrally and interface with the projects individually. As such, they develop expertise and can help junior people use everything effectively. If one project encounters difficulties, centralized skill profiles know about other team members who might have successfully overcome similar challenges.

The downside of this approach is that specific machine learning projects lose some of their autonomy. They cannot employ the tool they find best for a particular job, but they have to use centrally provided ones. The centrally provided tools tend to be very mature (as they cater to all use cases, and the most mature ones set the benchmark) and may be a considerable overhead for early-stage projects that aim for agility and iteration speed.

Other companies might decide on a high degree of decentralization with highly autonomous business units and machine learning projects. These teams are highly dynamic and independent. Organizations with this approach emphasize the data engineer and machine learning engineer roles instead of their architect counterparts. The people in the project may also fulfill (parts of) the architect roles. They choose and commission the tools that fit the job and the maturity level of the project. This procedure ensures agility and user centricity (as opposed to focusing on meeting organizational standards). However, as there is less communication between teams, double work may occur, and specific knowledge about, for example, efficient processes might get lost in organizational silos.

Companies can counteract these effects by intentional measures.

In either case, centralized or decentralized MLOps capabilities, data governance, and management will probably be centralized roles. There are few benefits to defining guidelines that only apply to a few or strategies that are incoherent with the greater whole.



"Building effective machine learning systems is not a solitary endeavor but a team effort. Collaboration between domain experts, data scientists, and engineers is key to success in the era of AI."

– François Chollet
Creator of Keras Deep Learning Library

Case study: How ML teams are composed at Uber

When Uber set out to bring AI into production, they realized some important lessons when it comes to managing roles and responsibilities that also align with our framework.

We sat down with Melissa Barr and Ben Wang of Uber to go over their experiences and challenges when it comes to ML Roles. Melissa works there as a Senior Technical Program Manager for the ML & AI Platform team and is with the company for more than four years. Ben is also a Technical Program Manager for the AI Platform and Applications team.

Designing ML teams at Uber

In an organization developing AI in production, the number of roles depends on a company's scale, AI maturity, and the complexity of the solutions. Smaller companies will probably combine some of the roles Uber has. However, with increasing scale specialized roles become essential.

We would advise companies to align their roles with the key steps of their ML workflow along the ML lifecycle from data preparation to model deployment. It's beneficial to educate the entire organization on the AI solution and to ensure expertise in all necessary areas.

The roles should be regularly reviewed and adapted as the organization evolves. Throughout this process, a culture of mutual collaboration and support between teams is greatly beneficial. Machine learning is an inherently interdisciplinary discipline.

Existing ML roles at Uber

We employ structured teams with dedicated roles to handle the different aspects of machine learning development. Data Engineers are responsible for data preparation. Data Scientists design models for specific use cases, and Machine Learning engineers handle model training and deployment. We also have Software Engineers in supporting tasks and Managers overseeing the general cohesiveness and direction of the projects.

Apart from that, Product Managers are responsible for the vision and roadmap of a product and the corresponding value-add by machine learning solutions within them. Lastly, Technical Program Managers are responsible for strategy, education, and documentation teams and provide knowledge support.

We created and shaped these roles around the machine learning lifecycle and our company workflows.

Data Scientists and Machine Learning Engineers: are they the same?

The two roles have some overlap, but they have a different focus. Data Scientists provide theoretical insight and decisions about the models. They select the right data sources, model architectures, and performance metrics.

Machine Learning Engineers on the other hand focus on the engineering and operational aspects, work on infrastructure, and write pipelines for training, retraining, deployment, and evaluation of models.

Our engagement with the team at Uber validates our views on the topic. Their experiences match the framework in these respects: dedicated roles are important, the specific roles that are needed are very similar, and the implementation of roles is dependent on a number of factors such as the company's scale and AI maturity.

Authors

Alexander Machado is the Head of Trustworthy AI CoE and former Head of MLOps Processes at appliedAI Initiative. He has a decade of experience in Data Science, Artificial Intelligence, and Data Engineering at appliedAI, the Max Planck Society, and BMW. His work focused on leading, planning, and developing AI solutions from experimentation to production. He has developed multiple MLOps frameworks and published an MLOps online course. Currently, he is further developing processes that address the inherent challenges of production systems and compliance with the upcoming AI Act.

Max Mynter is an MLOps engineer at appliedAI Institute for Europe where he builds open source developer tools for all stages of the machine learning lifecycle. Before he spent some time as a visiting scholar at UC Berkeley's School of Information, and worked as a Data Scientist at an energy-tech start-up and at Allianz Global Investors as a Quantitative Risk Analyst. His academic background is in Physics, Social Sciences and Technology Management.

Acknowledgements

The authors thank Jan Willem Kleinrouweler, Susanne Klausing and Anish Pathak for their contributions.

Contributing Partners

The frameworks in this whitepaper were developed with the support of many companies in Germany that shape the domain as part of the MLOps working group at appliedAI. All partners of appliedAI have contributed in one way or another, and we wish to thank them. Additionally, a number of individuals from these companies have put in extra effort to make this whitepaper available to the public. We want to thank these individuals and their companies.

Elena Zennaro
Senior AI Specialist
Infineon Technologies

Dr. Hendrik Brakemeier
AI Application Development Expert
European Central Bank

Salma Charfi
Cloud Engineer AI
Miele

Eduard Goetmann
Senior ML Engineer
Miele

Benjamin Pohl
Senior ML Engineer
EnBW

Mark Mauerwerk
Senior Portfolio Manager AI
Deutsche Telekom

Natalia Fitis
Data Value Stream Lead
Deutsche Telekom

Matthias Berger
Senior Data Scientist
MTU Aero Engines

Tobias Buchner
AI Strategist
MTU Aero Engines

Jonas Goltz
Data Scientist
Giesecke+Devrient GmbH

Tobias Emrich
ML Team Lead
Snke OS

About appliedAI

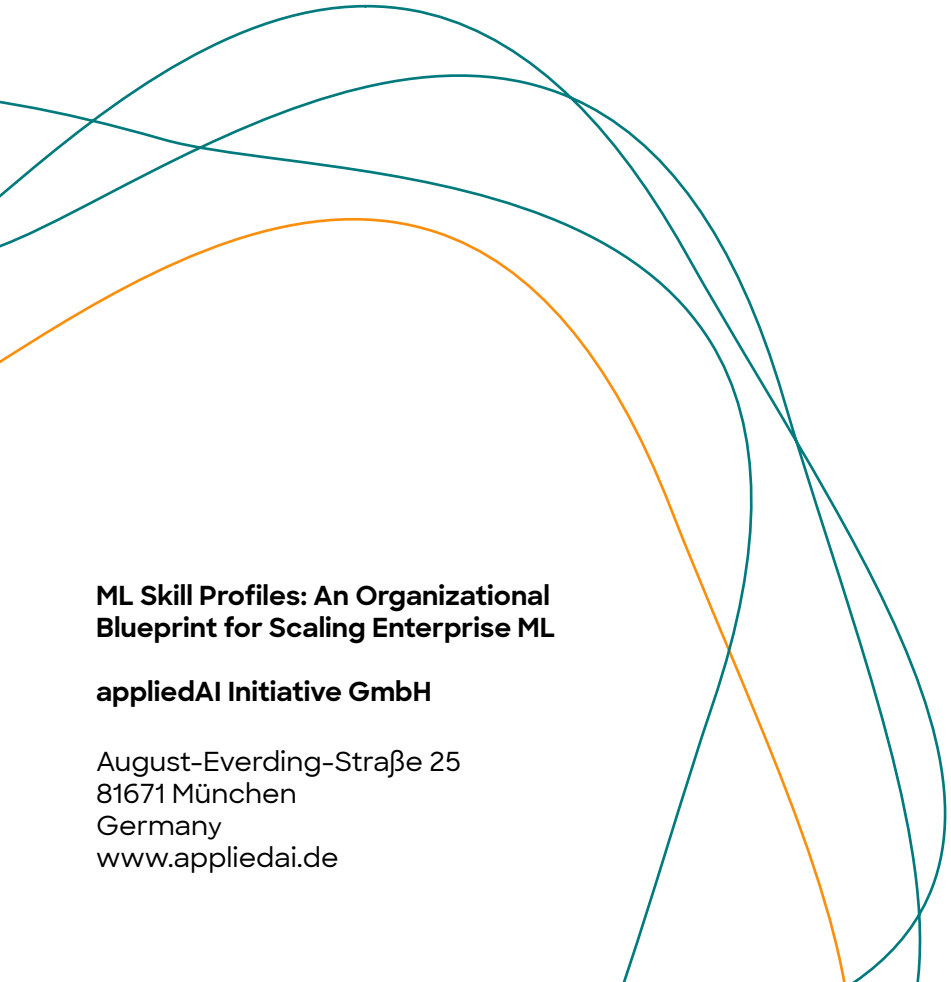
appliedAI is Europe's largest initiative for the application of cutting edge AI. Our vision is to shape the European AI ecosystem as a trusted enabler and innovator.

With partners such as NVIDIA, Google, BMW, Siemens, Deutsche Telekom and many more, we have been strengthening and building the next champions in AI since 2018.

You can find more information about appliedAI at:

www.appliedai.de

initiative for applied artificial intelligence



ML Skill Profiles: An Organizational Blueprint for Scaling Enterprise ML

appliedAI Initiative GmbH

August-Everding-Straße 25
81671 München
Germany
www.appliedai.de