



D-LAMA

The nice and easy Data Labeling Manager

Project Outline

ZHAW School of Engineering

02.04.2023

PM4

Feuereissen, David | Grand, Joel | Güntert, Gianmarco | Lichtenecker, Noah |
Mohammad, Schazad | Nobel, Gabriel | Sigrist, Stefanie | von Wartburg, Rebekka

Table of contents

1	Introduction	2
2	Idea	2
3	Customer benefit	3
4	Competitive analysis	4
5	Main sequence	5
5.1	Project admin view	5
5.2	Data labeler view	8
6	Additional requirements	9
7	Resources	9
8	Risk analysis	10
9	Rough planning	12
9.1	Roadmap	12
10	Economy	13
11	Source List	14

1 Introduction

Thanks to artificial intelligence (AI), many processes are currently being simplified, and many new projects are being implemented. Supervised learning plays a big role in this, as it is a common use case or at least a component of most AI networks. Last semester, the team members of this project have visited Machine Learning and Data Mining (MLDM) and realized that labeling data takes a lot of time and is not very satisfying. According to KOBOLD.AI many companies fear data labeling as the bottle neck of machine learning (ML) in the future. [1] This is something D-LAMA is trying to change.



Picture 1: Different attempts of finding a logo for the D-LAMA labeling application.

To make manual labeling data for AI projects exciting and efficient, D-LAMA shall be a simple web application that makes labeling easy and accessible to everyone. With the swipe mechanism, labeling is simplified, and through gamification, quality as well as quantity are measured in a fun way. Thanks to configuration options, clients should also have a great experience and get their data more efficiently.

A modern and smooth design as well as an eye-catching logo shall tempt the data labeler to use this application. Therefore, four different logos were designed, as can be seen above in picture 1, before choosing the one.

2 Idea

The idea of developing a software tool for data labeling to train supervised neural networks is very promising. Data labeling is an essential task in supervised machine learning that requires a significant amount of time and effort. Automated labeling comes with many obstacles, for example are algorithms prone to malfunctions such as date errors or edge cases. This means the effort must be made by humans. To keep them interested and increase their quality and quantity, gamification and a simple user experience are essential.

The benefits of such a tool are clear. It can help businesses develop ML models faster and more efficiently, which can ultimately improve their competitiveness in their respective industries.

However, as with any technological product, there are also risks and challenges that need to be addressed. One of the main risks of data labeling is the potential for inaccuracies and inconsistencies in the labeling process. This can lead to the development of inaccurate or unethical ML models, which can be costly for businesses.

To mitigate this risk, there should be a control mechanism and a point system that motivate the data labelers to deliver the best quality they possibly can. A person who wants to have data labeled needs to be able to upload different data types such as video, picture, text and audio with ease and choose the necessary labels. The person who wants to label a dataset shall be able to do that by swiping in the direction of a previously set label (inspired by tinder [2]) with a maximum of four labels at a time (up, down, left, right).

3 Customer benefit

The software will help businesses and research laboratories by reducing the time and resources required to develop ML models. It will allow them to label datasets quicker by a decentralized method and at the same time get a good mix of different options, which in the intended case leads to a more stable dataset. Decentralized labeling means the labeler does not have to be part of the company, making the process itself more accessible and flexible. Every user the company enables to can easily label data from everywhere in the world at any time.

Following are three of the most important advantages of a good and decentralized data labeling system:








- 1. Faster development of ML models:** Data labeling is a time-consuming and labor-intensive process. By decentralizing the data labeling process, businesses can significantly reduce the time and effort required to develop ML models. This can help businesses get their products and services to market faster, giving them a competitive edge in their industry.
- 2. Increased accuracy and consistency:** Different quality control measurements and the mixture of labelers can help reduce the potential for error and bias in the labeling process. This can lead to more accurate and consistent labeling and can therefore improve the overall quality of the ML models. Especially in industries where accuracy is critical, such as healthcare, finance and law, this is of great importance.
- 3. Scalability:** Decentralized data labeling can be scaled easily, allowing businesses to process large amounts of data quickly and efficiently. This can be especially important for businesses that need to process large datasets on a regular basis.

4 Competitive analysis

Due to the increasing popularity and importance of machine learning, tools have been developed on the market to simplify and improve the quality of manual data labeling. ML engineers must invest more than 80% of their time in data preparation and labeling in order to have the data ready to work with. [3]

To determine the current possible market position of D-LAMA in comparison to existing products, the best-known competing products have been analyzed in table 1:

Table 1: Comparison of individual competitor products based on their main features as well as the positive and negative aspects.

Competitive product	Main characteristics	Positive aspects	Negative aspects
Labelbox [4] 	A comprehensive toolset for data labeling, collaboration, and project management.	<ul style="list-style-type: none"> easy to use automated workflows adaptability 	<ul style="list-style-type: none"> relatively expensive
Hasty [5] 	A quick and easy app for labeling data based on deep learning technology.	<ul style="list-style-type: none"> quick and easy to use deep learning support 	<ul style="list-style-type: none"> less comprehensive functions than some of the other tools limited user definability
Scale [6] 	A powerful and flexible data labeling platform that supports a wide range of applications.	<ul style="list-style-type: none"> powerful functions flexibility 	<ul style="list-style-type: none"> relatively expensive more difficult to use less adaptability
SuperAnnotate [7] 	A user-friendly platform for labeling data with a focus on computer vision and natural language processing applications.	<ul style="list-style-type: none"> user-friendly fast processing times community support 	<ul style="list-style-type: none"> limited functions
V7 [8] 	A data labeling platform specifically designed to meet the requirements of computer vision applications.	<ul style="list-style-type: none"> specialized in computer vision applications customizable functions adaptability 	<ul style="list-style-type: none"> not very user-friendly
Dataloop [9] 	A comprehensive data labeling platform with a focus on deep learning models.	<ul style="list-style-type: none"> comprehensive functions fast and precise processing times adaptability 	<ul style="list-style-type: none"> relatively expensive
appen [10] 	A platform for labeling data especially for companies and research projects.	<ul style="list-style-type: none"> high quality of results adaptability 	<ul style="list-style-type: none"> relatively expensive non-automated workflows longer processing times

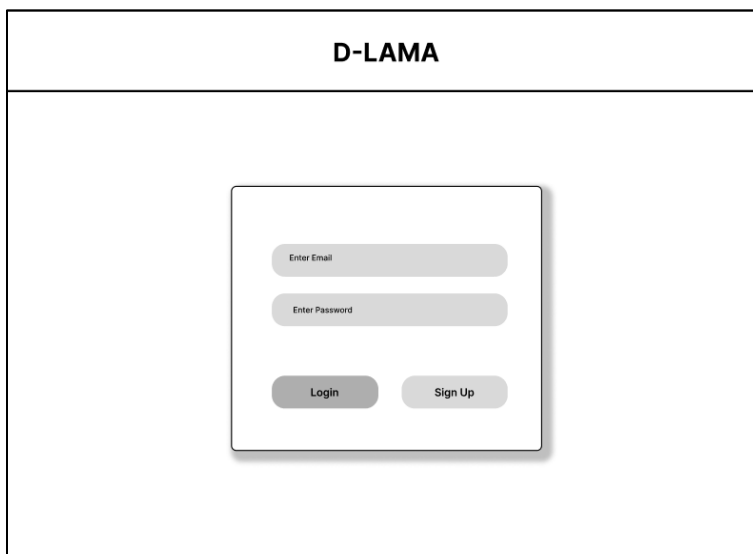
Although there are now many competing products, D-LAMA is clearly different from them. D-LAMA stands out for its simplicity. It is designed to make data labeling as simple and efficient as possible. Most existing applications only offer a web view. D-LAMA, on the other hand, has a web view for the administrator, but a mobile view is available for the data labeler. This should make labeling possible anywhere and at any time as mentioned in chapter 3, customer benefit. In addition, D-LAMA offers the data labeler a bonus system as an incentive to achieve the optimum in terms of both quality and quantity. Furthermore, the included swipe mechanism for labeling the data creates a kind of gamification.

5 Main sequence

In this chapter, the main sequence for the D-LAMA web app will be outlined from the perspectives of two types of users: The data labeling project admin and the data labeler. The project admin is responsible for creating and managing labeling projects, whilst the data labeler is responsible for labeling the datasets.

5.1 Project admin view

1. The project admin opens the D-LAMA web app and logs in with their admin account credentials in the login view (picture 1).



Picture 2: A mockup of the login view.

2. After logging in, the project admin is directed to the admin dashboard, where they can view a list of all labeling projects.
3. The project admin selects the option to create a new labeling project.

4. In the project creation view (picture 3), the app prompts the project admin to enter a name and type (text) for the new project. They then define the labels or categories for the dataset, as well as upload the dataset to be labeled.

The mockup shows the 'D-LAMA' app interface. At the top, there is a header with a hamburger menu icon on the left, the text 'D-LAMA' in the center, and a 'Logout' button on the right. Below the header is a 'Create New Project' form. The form contains the following elements: a 'Project Name' text input field, a 'Project Type' text input field, an 'Add Label' button with a circular indicator showing '3' and a downward arrow, an 'Upload Data' section with a file upload icon and the text 'Drag and drop here or browse files', and a 'Create' button at the bottom.

Picture 3: A mockup of the project creation view.

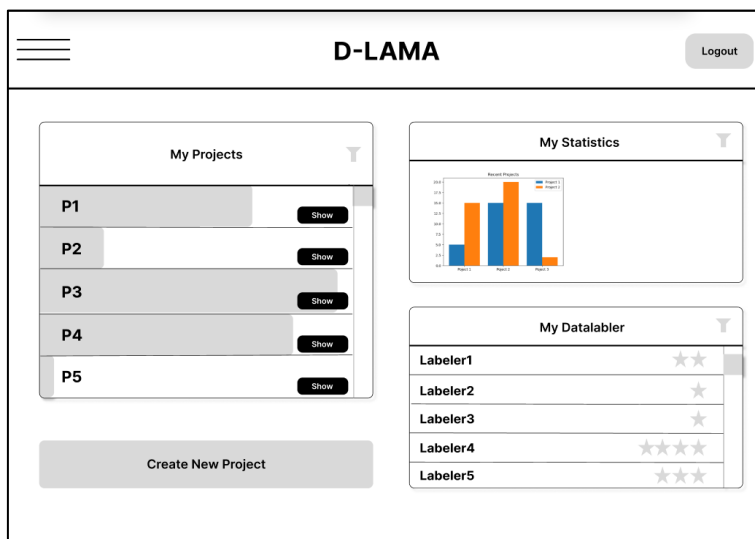
5. After creating the project, the project admin is directed to the project detail view (picture 4), where they can see the just defined details as well as all labelers assigned to it.

The mockup shows the 'D-LAMA' app interface for the 'Project Details' view. The header is identical to the previous view. The main content area is titled 'Project Details' and contains: a 'Project Name' text input field, a 'Project Type' text input field, an 'Add Label' button with a circular indicator showing '3' and a downward arrow, an 'Upload Data' section with a file upload icon and the text 'Drag and drop here or browse files', a 'Make Public' toggle switch, a 'My Datalabeler' section with a search icon and a list of labelers, and an 'Add Labeler' button. Below these is a 'Description of Project' text input field. The 'My Datalabeler' section lists five labelers with their respective star ratings: Labeler1 (3 stars), Labeler2 (2 stars), Labeler3 (1 star), Labeler4 (4 stars), and Labeler5 (3 stars).

Picture 4: A mockup of the project detail view.

6. The project admin can view individual data points in the dataset and see the labels that have been applied to them.
7. If the project admin notices any issues with the labeling process, they can flag specific data points for review and adjust the applied labels as needed.

8. In the project admin dashboard view (picture 5), the project admin can view the overall progress of the project and see how many data points have been collected, how many are remaining, and the overall accuracy of the labels applied.

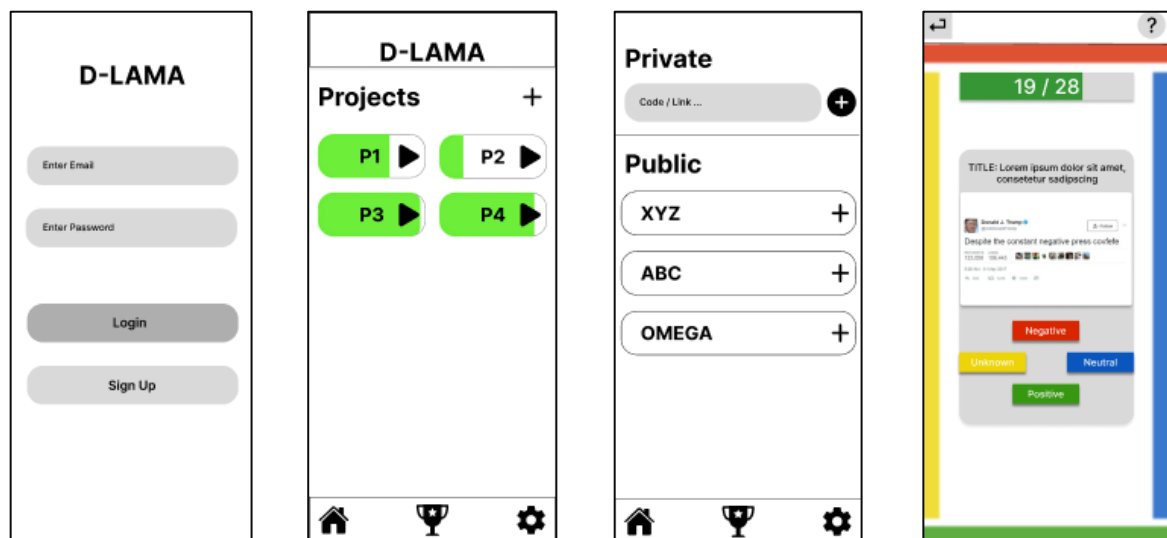


Picture 5: A mockup of the project admin dashboard view.

9. Once all data points have been collected, the project admin can download the dataset for use in ML models or archive the project if it is no longer needed.

5.2 Data labeler view

1. The data labeler opens the D-LAMA web app and is prompted on the login view (picture 6) to log in with their account credentials.
2. After logging in, the data labeler is directed to their home screen (picture 6), where they can view a list of labeling projects that have been assigned to their account.
3. In the project add view (picture 6), the data labeler can open public data labeling projects and assign them to their account or enter a code to do the same with a private project.
4. The data labeler selects a specific labeling project from the list on their home screen (picture 6) to start working on.
5. The selected project opens and displays the first unlabeled data point.
6. In the labeling view, the data labeler swipes up, down, left or right to label the data point or skips if unsure.
7. The app automatically loads the next data point in the sequence, and the user repeats step 6 until all data points have been labeled.
8. When the data labeler finishes labeling all data points in the project or interrupts the labeling process, they are directed back to their dashboard and can view their progress on the project.
9. The data labeler can select another project from their dashboard and start a new labeling process.



Picture 6: Four mockups showing the login view, the home screen of the data labeler, the project add view and the labeling view.

6 Additional requirements

Additionally to the main sequence, the D-LAMA web app should fulfill the following requirements already in the prototype:

- The data labeler should be able to interrupt or pause the process and resume it later.
- The data labeler should be able to add and remove labeling projects.
- Both the data labeler and the project admin should be able to register a new account.
- The project admin should be able to set the datatype of the dataset as an image.

The following extensions can be added in later iterations:

- The project admin should be able to set the data type of the data set as video, text or audio.

7 Resources

Eight developers are available for the implementation of the D-LAMA project. Existing (personnel) resources for implementation:

- Currently, three people are working in application development and have expertise in software architecture.
- Five people have practical basic experience in object-oriented languages.

Additional resources needed for implementation:

- Software developers
- Client
- Development environment
- GitHub as a version control tool [11]
- Rancher Kubernetes cluster of Zurich University of Applied Sciences (ZHAW) [12]
- C4 model for visualizing software architecture [13]
- Figma for sketches and prototypes [14]

The estimated effort for realizing the prototype of the D-LAMA application amounts to approximately 120 hours per project member. Therefore, the total effort for the entire project group is about 960 hours. As the resources used, such as the development environment, GIT, Rancher, C4 model and Figma are provided free of charge, no additional effort is charged for these resources.

8 Risk analysis

When developing an application like D-LAMA, there are risks that affect not only the development of the application, but also the result. The four most important risks of D-LAMA are explained below:

- A. Quality risk:** Although all the team members are experienced in programming and have worked together before, there is still a possible risk that the resulting application does not correspond to the expected quality. Factors such as new technologies, communication issues and errors in the code or the logic itself play a big part in the resulting quality of the program. By ensuring a working testing environment and embedding tests for at least the most important functions of the application, errors can be diminished. Coding and collaboration guidelines help with the same problem and ensure a better way of communication and sharing code. Furthermore, it is of great importance that the customer gets involved from the beginning of the project to get an agile feedback loop.
- B. Resource scarcity:** The most valuable resource in this project is possibly the time of the team members. Since this project is part of the bachelor studies in computer science at the ZHAW and all the involved team members are part-time students, time is a valuable good. With a structured and especially dynamic project plan, the team should be able to schedule the work efficiently and therefore save time. In addition to time, there are also some more concrete resources, which are scarce: All things regarding the needed IT-infrastructure. Fortunately, the ZHAW does provide computer science students with a fault-tolerant IT infrastructure hosted on-site. The team could choose between two virtual machines on an OpenStack cluster or two Rancher Kubernetes namespaces. To be more flexible, both were ordered.
- C. User acceptance:** In the end, the D-LAMA application is a product ordered by a customer. Not fulfilling her acceptance criteria would be a failure. Therefore, user acceptance is a huge risk in many (software) projects. A good communication with the customer from the beginning of the project to its conclusion, as well as a good structured and dynamic method to document and track issues are mandatory.
- D. Communication Issues:** When working in larger teams, especially while working on a customer project, clear and efficient communication is the key. Unstructured or missing communication can cost valuable time and lead to avoidable errors. A good and structured project communication not only involves defined ways of communication between the team members and a structured documentation, but also explicit specified responsibilities. That also includes having defined representatives in case of an outage.

The table below is intended to show the four risks described according to their probability of occurrence as well as their severity. As can be seen three of the four risks are in the red section of table 2. It is of great importance to be aware of them and to take counteractive measures starting at the beginning.

Table 2: The four risks according to their probability of occurrence as well as their severity.

Severity / Probability	insignificant	medium heavy	heavy	critical
unlikely				C
possible				A
probably			B	
almost certain		D		

9 Rough planning

The work on the D-LAMA labelling application started in March 2023. Until July 2023, a first prototype shall be created. Therefore, not only resources were taken and risks were analyzed, but a whole roadmap has been created (diagram 1). Dividing the team into a frontend, backend and organization / infrastructure group, each of them has its own project flow embedded in the overall planning.

9.1 Roadmap

Milestones are achieved if the status of the application fulfills all requirements defined in the milestones description on the date it is set. In the roadmap, four milestones can be seen. Starting with designing different mockups for desktop and mobile view, the first milestone helps the team to start defining the functionality and design of D-LAMA.

By handing in the project outline, the second milestone will be achieved. The project outline includes a rough planning, definition of important parts of the project, such as the main sequence and additional requirements, as well as a risk analysis.

By the end of May 2023, the third milestone shall be realized. It includes having finished the prototype of the D-LAMA app. Shortly after, the product presentation will take place, which is the forth and final milestone in this roadmap designed with roadmunk [15].

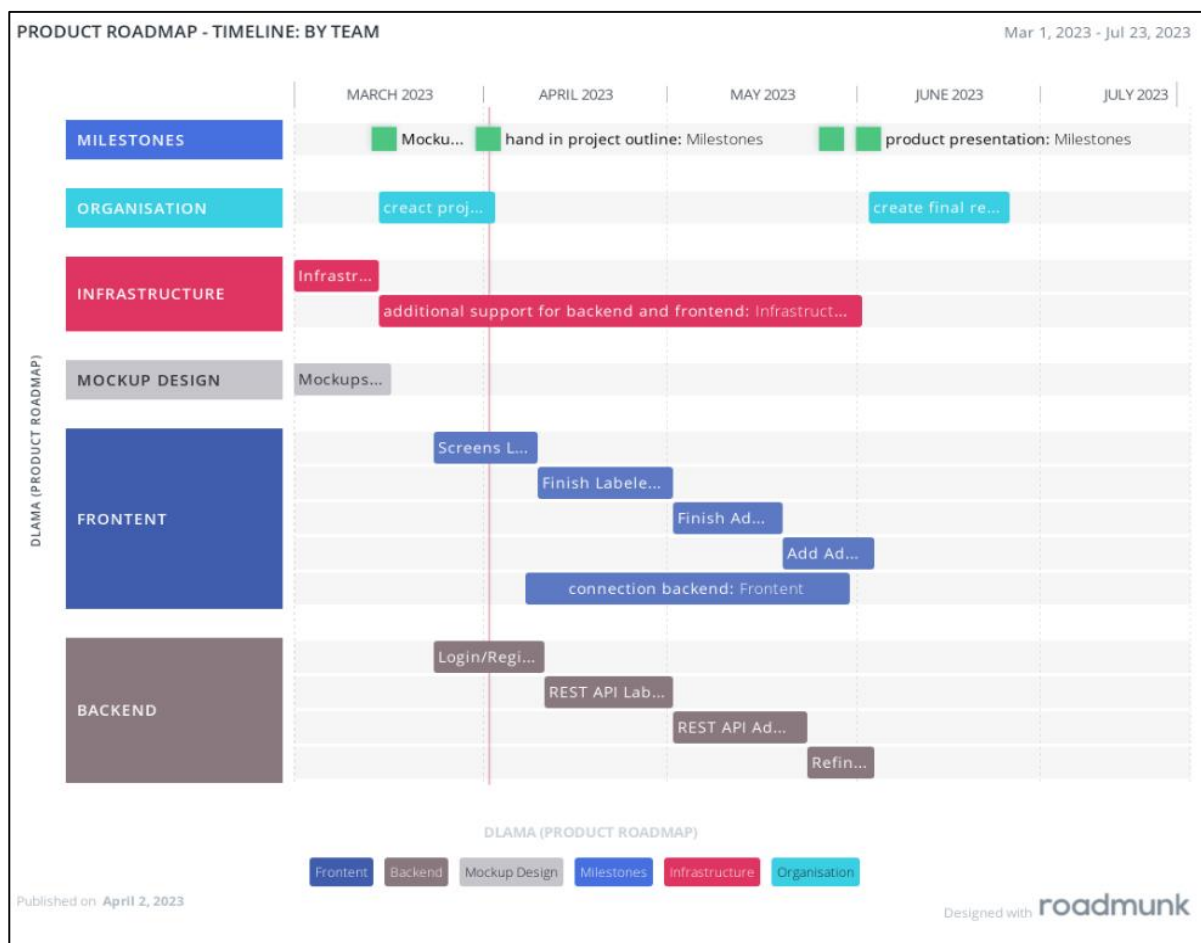


Diagram 1: The roadmap of the D-LAMA project shows all important milestones and bigger tasks.

10 Economy

The idea of D-LAMA to make data labeling possible from everywhere and in a fun way, was initially created to help the data science environment. D-Lama shall be a high-quality system that is as intuitive as possible and allows data science teams to get the most out of their data.

As there are other products on the market, as shown in chapter 4, competitive analysis, one of the biggest challenges will be to establish D-LAMA in this market. To make sure that the application really gets used in the beginning, the ZHAW is allowed free usage for at least one year. This maximizes the opportunity that the application really gets used, and with that it also really gets tested. This will generate as much feedback as possible (from the data labelers and project admins) so that the application can be improved to make it more and more competitive and to establish its reputation on the market.

For further usage or for other companies and universities there will be offered a priced model that depends on the following factors:

- disk space used
- data type used
- amount of labeled data

For each of these factors a fee will be charged that will be similar to the market standard.

11 Source List

- [1] KOBOLD.AI. URL: <https://www.kobold.ai/ml-labels/> [Access: 02.04.2023]
- [2] tinder. URL: <https://tinder.com/de> [Access: 02.04.2023]
- [3] SuperAnnotate - best-data-labeling-tools. URL: <https://www.superannotate.com/blog/best-data-labeling-tools/> [Access: 02.04.2023]
- [4] labelbox. URL: <https://labelbox.com/> [Access: 02.04.2023]
- [5] Hasty. URL: <https://hasty.ai/v2> [Access: 02.04.2023]
- [6] Scale. URL: <https://scale.com/> [Access: 02.04.2023]
- [7] SuperAnnotate. URL: <https://www.superannotate.com/> [Access: 02.04.2023]
- [8] V7. URL: <https://www.v7labs.com/> [Access: 02.04.2023]
- [9] Dataloop. URL: <https://dataloop.ai/> [Access: 02.04.2023]
- [10] appen. URL: <https://appen.com/> [Access: 02.04.2023]
- [11] GitHub. URL: <https://github.com/> [Access: 02.04.2023]
- [12] Rancher. URL: <https://www.rancher.com/> [Access: 02.04.2023]
- [13] The C4 model for visualising software architecture. URL: <https://c4model.com/> [Access: 02.04.2023]
- [14] Figma. URL: <https://www.figma.com/de/> [Access: 02.04.2023]
- [15] roadmunk. URL: <https://roadmunk.com/> [Access: 02.04.2023]

12 List of Pictures

Picture 1: Different attempts of finding a logo for the D-LAMA labeling application.	2
Picture 2: A mockup of the login view.	5
Picture 3: A mockup of the project creation view.....	6
Picture 4: A mockup of the project detail view.....	6
Picture 5: A mockup of the project admin dashboard view.	7
Picture 6: Four mockups showing the login view, the home screen of the data labeler, the project add view and the labeling view.	8

13 List of Tables

Table 1: Comparison of individual competitor products based on their main features as well as the positive and negative aspects.	4
Table 2: The four risks according to their probability of occurrence as well as their severity.....	11

14 List of Diagrams

Diagram 1: The roadmap of the D-LAMA project shows all important milestones and bigger tasks.....	12
---------------------------------------------------------------------------------------------------	----