



2024 EDITION



# GUIDE TO DATA LABELING

# TABLE OF CONTENTS

INTRODUCTION TO IN-HOUSE DATA LABELING	04
--	----

## CHAPTER 1:

HOW TO BUILD A SOLID DATA ANNOTATION STRATEGY	12
---	----

## CHAPTER 2:

HOW TO MAINTAIN HIGH QUALITY OF LABELED DATASETS	24
--	----

## CHAPTER 3:

HOW TO KEEP THE ML DATASETS SECURE	31
------------------------------------	----

## CHAPTER 4:

HOW TO HIRE DATA ANNOTATORS	41
-----------------------------	----

## CHAPTER 5:

HOW TO TRAIN DATA ANNOTATORS	51
------------------------------	----

## CHAPTER 6:

HOW TO CHOOSE BETWEEN IN-HOUSE VS. OUTSOURCED	57
---	----

WHY CHOOSE LABEL YOUR DATA	63
----------------------------	----

“

*AI/ML teams often struggle to find the perfect labeling setup for their data pipelines. We've been there.*

*Over 4 years, we've seen everything from open-source tools with API integrations to commercial solutions with human-in-the-loop workflows. In this guide, we dive into our best labeling practices for ML engineers and AI researchers wishing to make their data pipeline more efficient.*

*From exploring key labeling strategies and quality metrics to building an in-house team from scratch, here's everything you need to know to get started with dataset labeling for ML.*

”

**Karyna Naminas,**  
CEO of Label Your Data ❤️



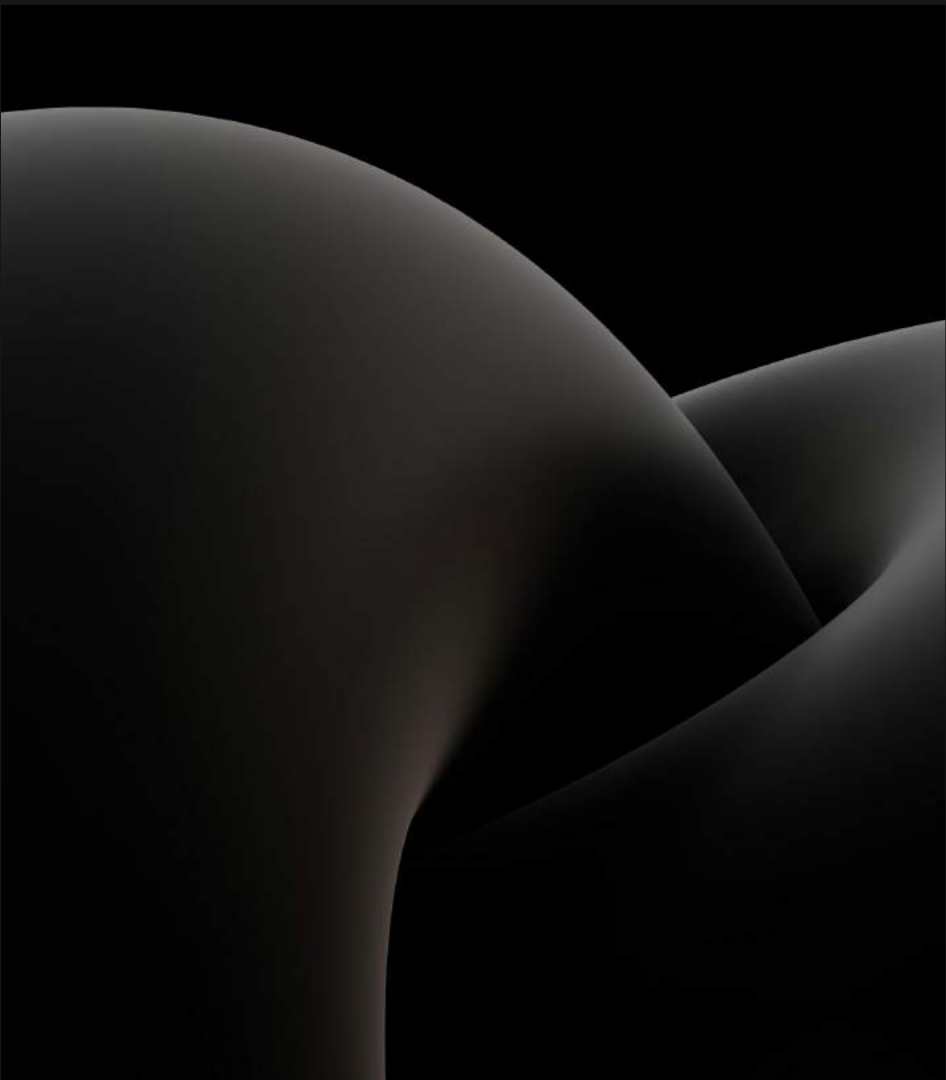
**Need feedback  
on your ML data  
annotation  
setup?**

**FREE CONSULTATION**

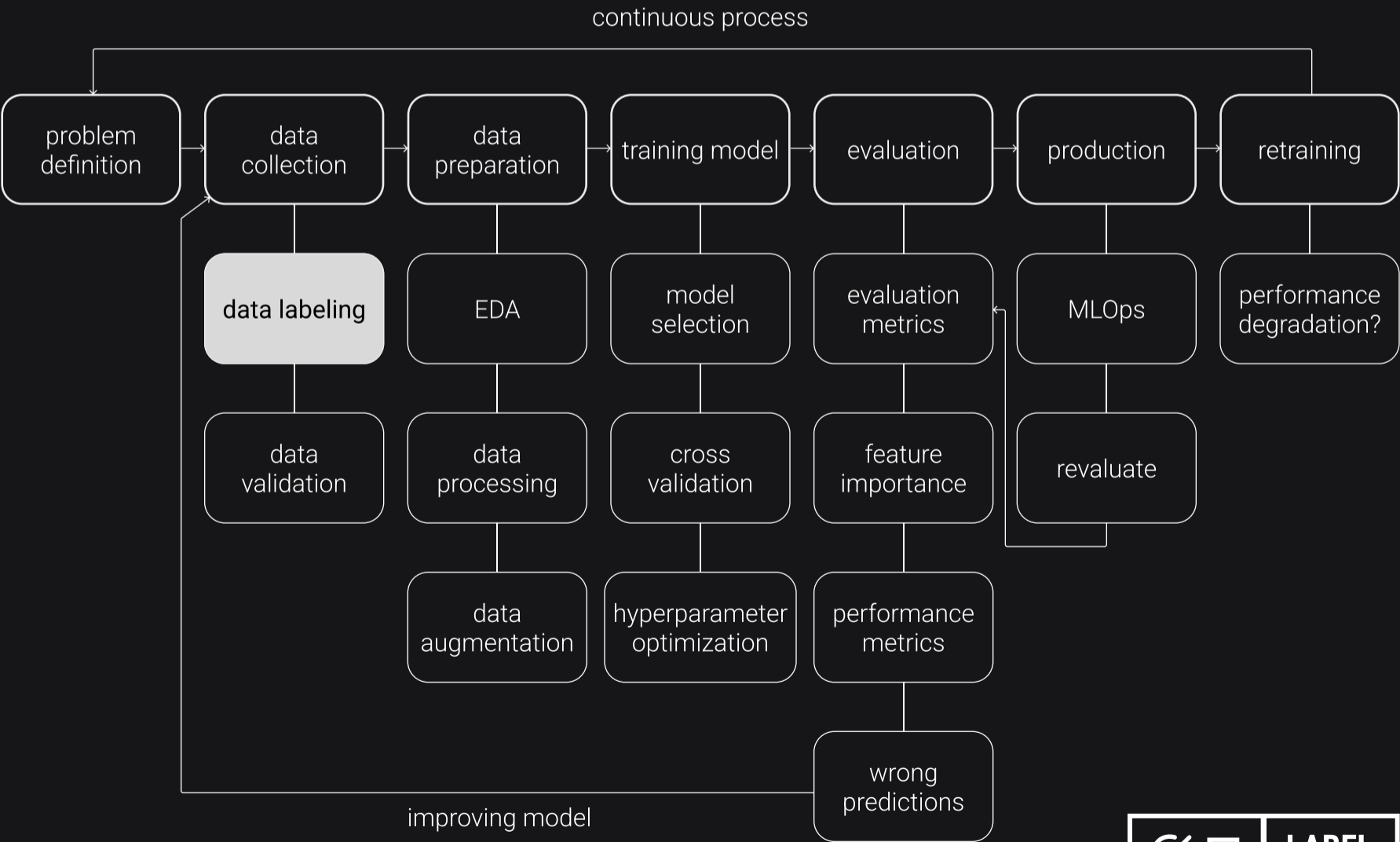


# Introduction to In-House Data Labeling: Where to Start?

Data annotation, often referred to as data labeling, is a cornerstone of the machine learning pipeline. It acts as the bridge between raw data and a functional ML model. During this step, human annotators or automated tools add labels or tags to the data, helping the model understand the underlying structure and meaning of the data.



## ML Project Stages



# Data Labeling in the Machine Learning Pipeline

Here's a breakdown of how data labeling fits in the ML pipeline:

## > Data collection

The pipeline begins by gathering the raw data you want your model to learn from. Data collection implies gathering raw, unstructured data (images, videos, text documents, or audio files) that needs to be labeled. The more data you have, the more precise your model will be.

Here's where you can gather data for your ML project:

- **Freelance fieldwork:** If you require specific data that isn't readily available online, hiring freelance data collection specialists can be a valuable option.
- **Public datasets:** There's a wealth of free data available online, with a few top resources to explore, such as [Kaggle](#), [UCI Machine Learning Repository](#), and [Data.gov](#).
- **Paid Datasets:** For highly specialized data or access to exclusive information, investing in paid datasets can be worthwhile.

## > Data cleaning

The next step is preparing data for supervised ML by cleaning it. That is, eliminating irrelevant, duplicate, or corrupted files to uphold data quality, as well as identifying and correcting (or deleting) errors, noise, and missing values. Data cleaning is an ongoing process that happens throughout the development and potentially even deployment of your machine learning project.

The final step here is storing your collected data the right way and in the right format. Data is usually stored in a data warehouse (traditional data warehouses like Oracle Exadata, Teradata, or cloud-based services like Amazon Redshift) or data lake (cloud-based solutions like Amazon S3 with AWS Glue or Azure Data Lake Storage with Azure Databricks), for easier management. We suggest choosing the storage system able to meet the needs of your model as the data increases.

## > Data labeling

Here, the data is labeled with relevant information to create a labeled training dataset. Let's start with data labeling for Computer Vision models. If you're building a computer vision system, you deal with visual data, such as image,



videos, and sensor data. Here, you can use several types of data annotation:

- Image Categorization
- Semantic Segmentation
- 2D Boxes (Bounding Boxes)
- 3D Cuboids
- Polygonal Annotation
- Keypoint Annotation
- Object Tracking

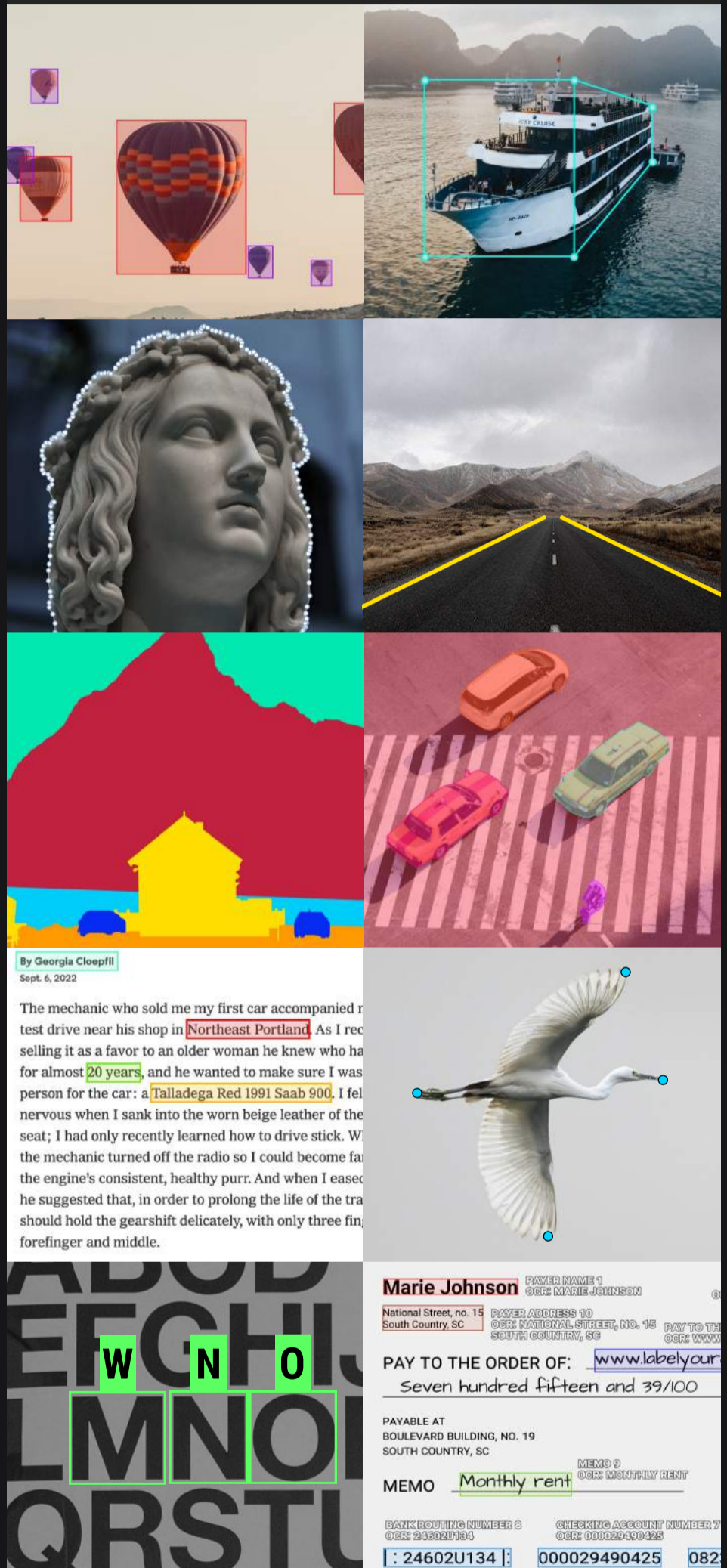
**For Natural Language Processing (NLP)** models, data labeling requires annotators to possess linguistic knowledge for handling the following types of text and audio data annotation:

- Text Classification
- Optical Character Recognition
- Named Entity Recognition
- Intent/Sentiment Analysis
- Audio-To-Text Transcription

## > Model training

Once you've labeled data in machine learning and checked the quality and consistency of the performed annotations, it's time to put the labeled

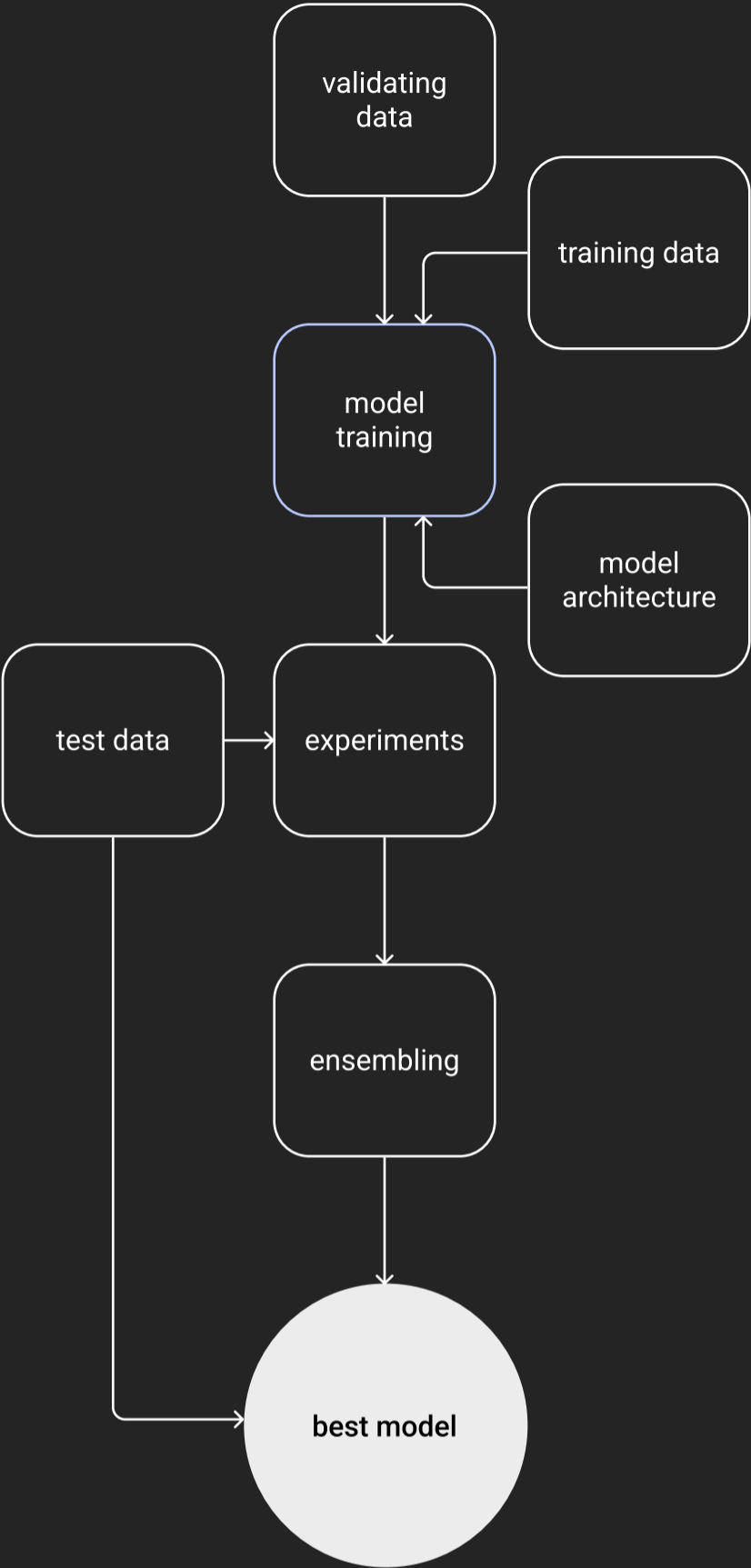
## Data labeling



dataset to use. By analyzing the labeled data, the model learns to identify patterns and relationships between the data and the labels.

More specifically, the dataset can now be split for model training, testing, and validation, respectively, following this useful rule of thumb:

Labeled data, %



> Model evaluation & deployment

Once trained, the model’s performance is evaluated on a separate dataset. If successful, the model can then be deployed for real-world use.



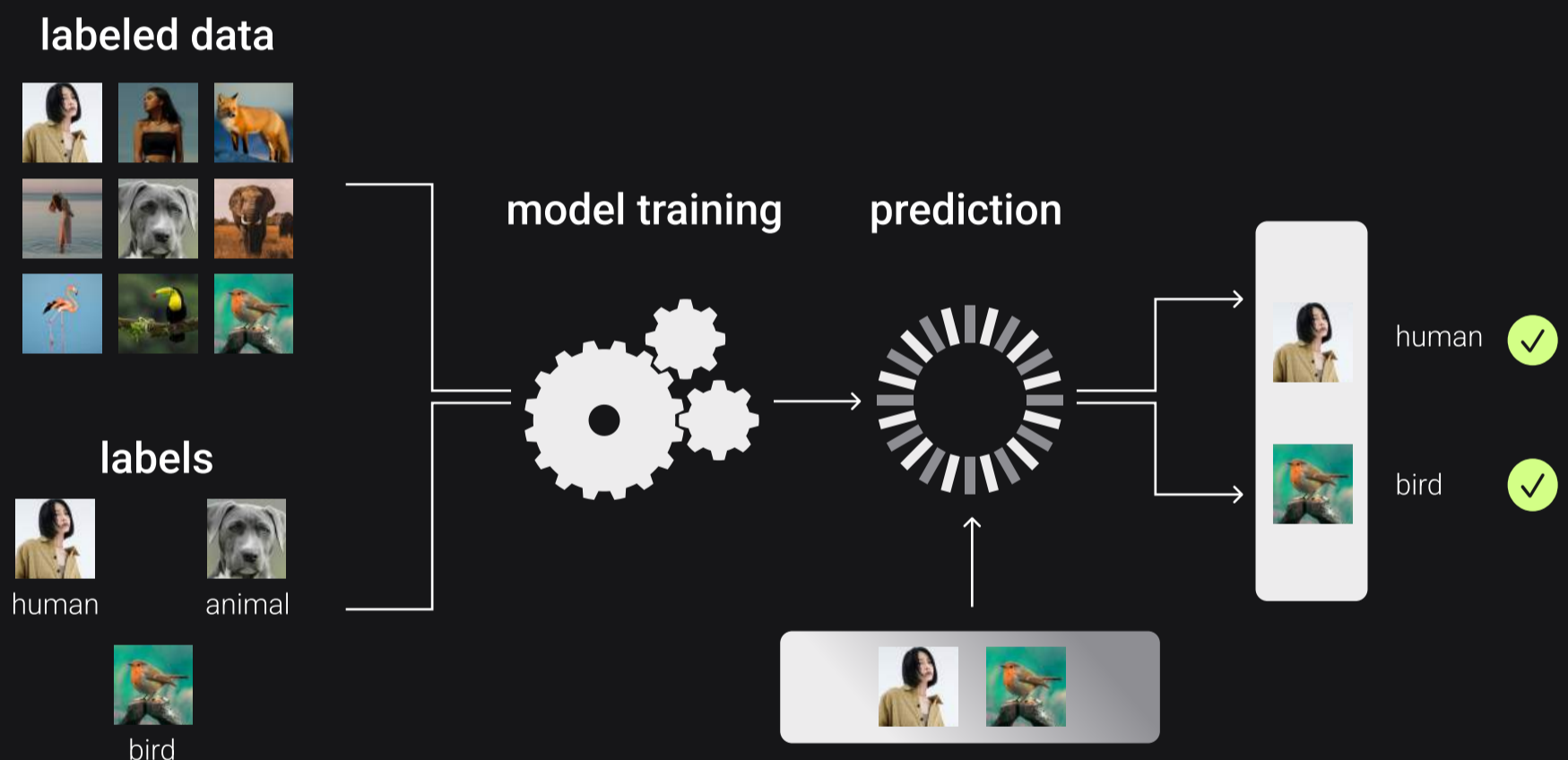
# How Does Data Labeling Work?

Most ML models use **supervised learning**, where an algorithm maps inputs to outputs based on a set of labeled data by humans. The model learns from these labeled examples to decipher patterns in that data during a

emphasizing the importance of investing time and resources in accurate data labeling.

With high-quality annotations on hand, data scientists can identify the important features within the data. However, common dataset labeling pitfalls can impede this crucial process.

## Data labeling pipeline in ML



process called model training. The model can then make predictions on new data.

Labeled data used for training and assessing an ML model is often referred to as “ground truth.” The model’s accuracy relies on the precision of this ground truth,

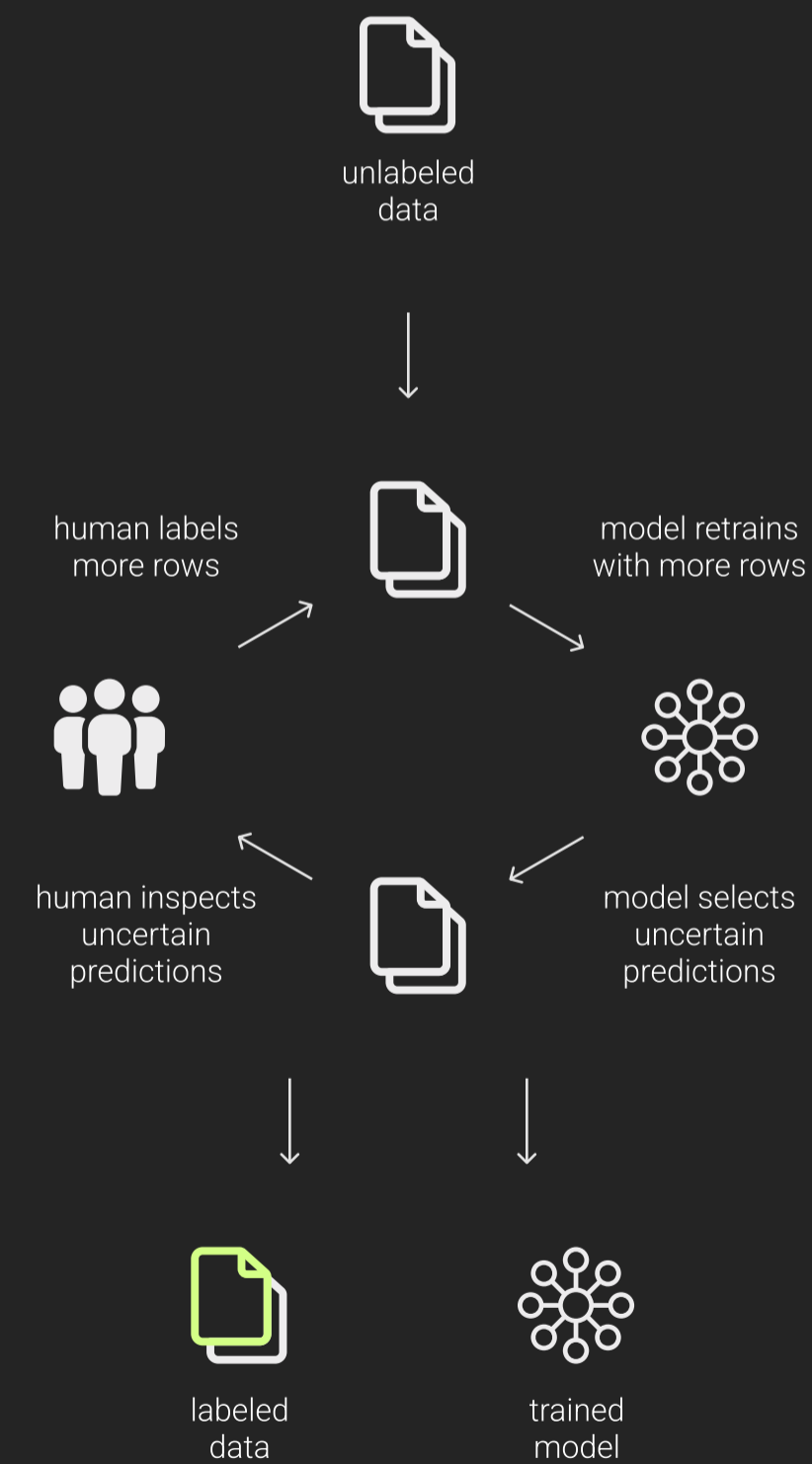
More specifically, public datasets often lack relevance or fail to provide project-specific data, and in-house labeling can be time-consuming and resource-heavy. Automated tools, while helpful,



don't guarantee 100% accuracy or offer all the features you need. And even with automation, human oversight is still a must.

In this guide, we reveal the 6 steps to overcome these challenges when labeling datasets for machine learning.

## Active learning



## 6 Steps to Overcome Data Labeling Challenges



In the [study by Hivemind](#), a managed annotation workforce demonstrated a 25% higher accuracy rate compared to crowdsourced annotators, who made over 10 times as many errors.



This means that building and managing your own in-house data labeling team can significantly improve the quality of your training data. To help you achieve efficient in-house labeling, we've gathered our time-proven steps to help you navigate the main challenges in data labeling:

## 1. Build a solid data annotation strategy

Data annotation projects usually fall under one of the categories: data labeling for initial ML model training, data labeling for ML model fine-tuning, and human-in-the-loop (HITL) and active learning.

Your data annotation process must be scalable, well-organized, and efficient. It's an iterative step in the entire ML pipeline, involving constant monitoring, feedback, optimization, and testing.

Check out the first chapter to learn more about 6 approaches to building your data annotation strategy.

## 2. Maintaining high quality of labeled datasets

Regular QA procedures are crucial to verify label accuracy and consistency. This includes reviewing random data samples or employing validation techniques. The labeling process also follows an iterative loop, with initial results reviewed and feedback incorporated for further label refinement. Check out the second chapter to read about the ways of

achieving the top quality of your annotations.

## 3. Keeping the machine learning datasets secure

To ensure labeled data security, you should prioritize a multi-layered approach encompassing:

- **Physical security:** Secure facilities with manned security, access restrictions, video surveillance, ID badges, and limitations on personal belongings in sensitive areas.
- **Employee training & vetting:** Consistent training on data security risks, phishing, password management, and ethics. Background checks and requiring adherence to security policies and NDAs.
- **Technical security:** Strong encryption (AES-256), secure annotation software, multifactor authentication, role-based access control to limit data exposure, and restricted internet access.
- **Cybersecurity:** Proprietary communication tools, penetration testing, and external security audits.
- **Data compliance:** Adherence to industry regulations like GDPR, CCPA, and ISO 27001 with ongoing updates to maintain compliance.



Check out the third chapter to find out more about legal issues associated with private data and how to avoid them.

#### **4. Hiring data annotators**

Inconsistent data annotations can cripple the model's performance. To tackle this, you need to hire skilled data annotators. You can build your team by leveraging your network through job boards and social media, or look beyond it by partnering with data annotation companies or targeted online ads. By choosing the right hiring approach, you'll assemble a strong data annotation team to fuel your ML project's success.

Check out the fourth chapter to learn about 6 steps to building your data annotation dream team through expert hiring strategies.

#### **5. Training data annotators**

Despite the level of automation we've reached so far, data labeling cannot do without human intelligence. Always make sure to have human experts on your team. They bring the context, expertise, experience, and reasoning to streamline the automated workflow. Training a team of annotators to use a specific labeling tool and follow the project guidelines. When dealing with a specific type of data and edge cases in data labeling, you need to hire subject-matter experts

(SMEs) for complex domains, like healthcare, finance, scientific research, or for multilingual tasks in NLP.

Check out the fifth chapter to explore the top training steps for data annotators.

#### **6. Choosing between in-house vs. outsourced data labeling**

Choosing between in-house vs. outsourced data labeling depends on your specific needs and priorities. Consider the size and complexity of your dataset, the turnaround time required, and the level of control you need over the labeling process.

In short, outsourcing works for projects involving large datasets with simpler labeling tasks and a focus on faster turnaround times. However, this strategy might pose potential quality issues. A dedicated in-house team, in contrast, is suitable for those looking for a balance between cost, quality, and scalability, especially for projects requiring domain expertise.

Check out the sixth chapter to understand the difference between an in-house labeling team and outsourcing and see what fits your ML project best.



# CHAPTER 1:

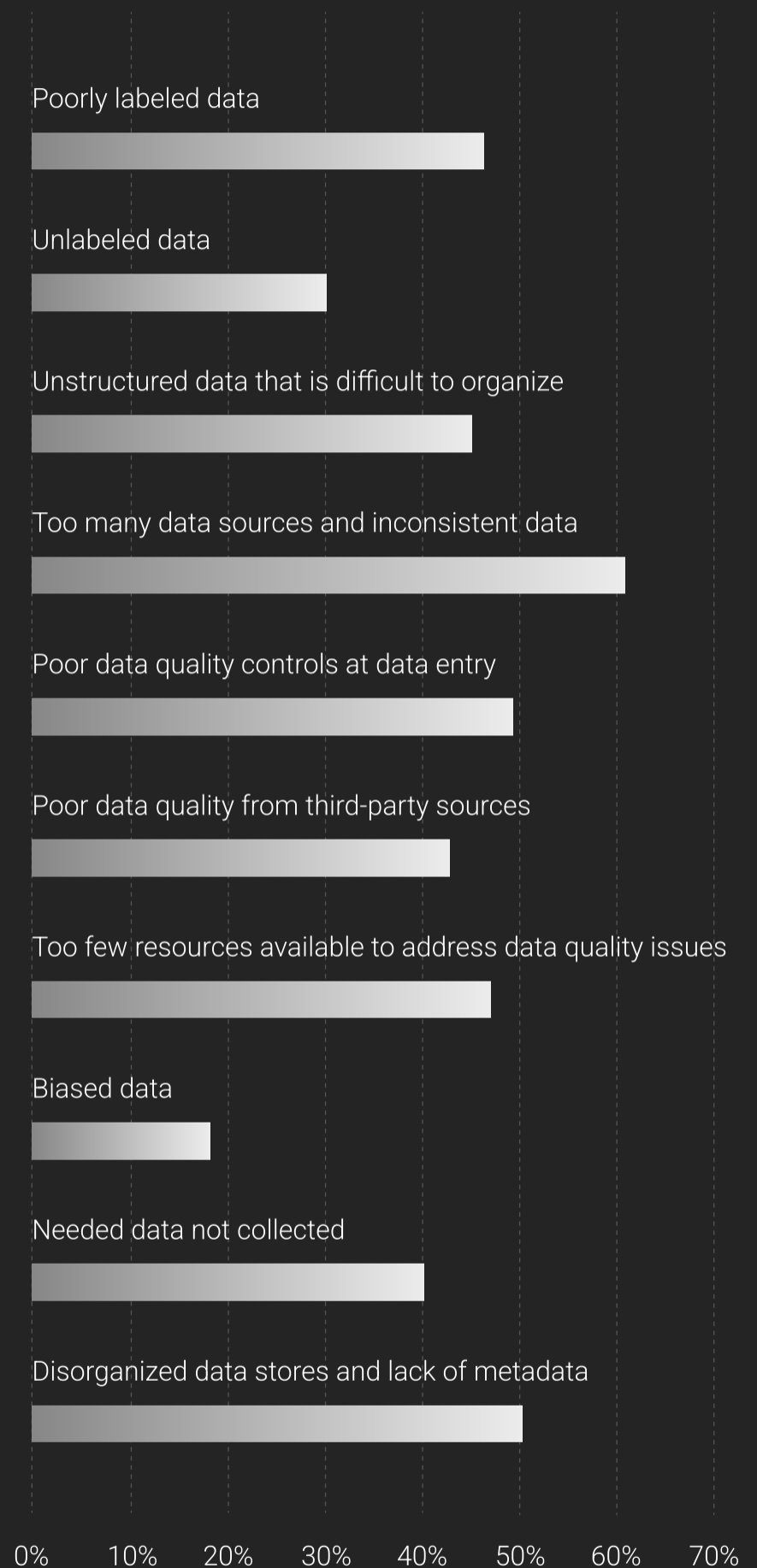
## HOW TO BUILD A SOLID DATA ANNOTATION STRATEGY



Machine learning helps 48% of businesses to get the most out of their large datasets. Yet, poorly labeled data, unstructured data, the abundance of data sources, biased and inconsistent data are among the primary data issues that these businesses face.


The solution here is a sound data annotation strategy, so that ML models can be trained on clean, organized, and representative datasets, unlocking the true potential of large datasets for better decision-making and business outcomes.

## What are the primary data quality issues your organization faces?



# How to Measure the Scope of the Dataset Volume to Label

To optimize annotation workflows, AI engineers and operations managers require precise dataset calculations and monthly new data generation rates. This information helps the annotation team:



Plan for the initial annotation cycle: Knowing the total dataset volume allows for efficient resource allocation and task scheduling.

Identify bottlenecks and staffing needs: High monthly data generation rates might necessitate streamlining the annotation process or hiring additional personnel.

## We suggest taking these steps:

### ✓ Count the number of instances

Determine the total number of data points or instances in your dataset. This could be the number of rows in a table, documents in a corpus, images in a collection, etc.

### ✓ Evaluate data complexity

Assess the complexity of the data. Consider the variety and types of data and the diversity of labels or categories needed.

### ✓ Consider annotation granularity

Understand the level of detail required for annotation. Finer granularity may require more effort (i.e., annotating each word in a document versus annotating the document as a whole).

### ✓ Understand the difficulty of the labeling task

Annotation tasks vary in complexity. Labeling images, for instance, can include object detection, segmentation, or classification, each with differing levels of difficulty. Assess the complexity of annotating each instance, as some may be straightforward while others demand more nuanced judgment.

### ✓ Analyze time requirements

Estimate the average time required to label each data point. This can depend on the task and the expertise needed for accurate annotation.

### ✓ Use sampling techniques

If the dataset is large, you might



consider sampling a subset to estimate the annotation effort required. Ensure that the sampled subset is representative of the overall dataset.

✓ **Consult domain experts**

*“Seek input from expert data labeling service providers like Label Your Data to understand the context and intricacies of the data. They can provide valuable insights into the annotation process. In addition, only industry experts can help ensure the quality and consistency of your labeled data.”*



**Ilyas El Alj**

Customer Success Manager at Label Your Data

# Top 5 Data Annotation Tactics in Machine Learning

## Raw data



Manual labeling vs. Automated labeling



In-house labeling vs. External labeling



Open-source vs. Commercial labeling tools



Public datasets vs. Custom datasets



Cloud data storage vs. On-premise Storage

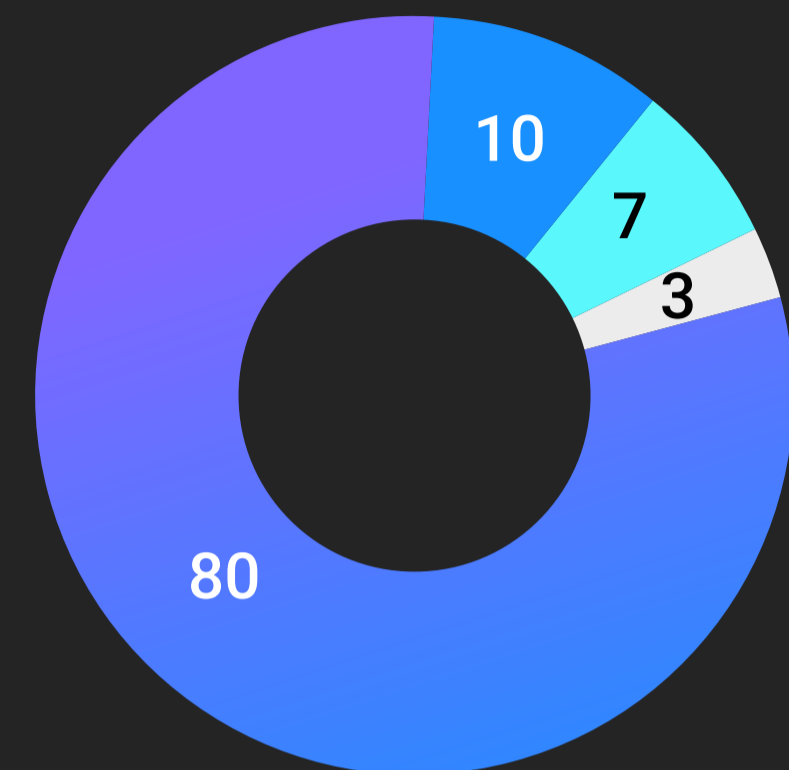


## Labeled datasets

Since about 80% of the time spent getting an ML algorithm ready involves collecting, cleaning, and annotating data, a well-defined annotation strategy can streamline the entire machine learning pipeline and save you significant time on your project.



## What activity takes most of data scientists time?, %



- Collecting, labeling, cleaning and organising data
- Building and modelling data
- Mining data for patterns
- Refining algorithms

Building a robust data annotation strategy is crucial for the success of your ML project. This strategy essentially defines a set of tactics that determine how you'll label your data. Selecting the right tactics depends on various factors specific to your project.

Here's a breakdown of the key data labeling tactics to help you decide which ones work best for your project:

## 1. Manual Labeling vs. Automated Labeling

### MANUAL LABELING

This involves human annotators identifying and assigning labels to specific elements within your data points.

#### PROS:

Ensures high accuracy, allows for complex labeling tasks, and provides greater control over quality.

#### CONS:

Time-consuming, expensive (especially for large datasets), and prone to human error.

### AUTOMATED LABELING

This tactic utilizes ML algorithms to label your data points, reducing the need for manual human intervention. Automatic annotation greatly assists when dealing with large datasets. For instance, AI models like [Grounding-DINO](#) and [Segment Anything Model \(SAM\)](#), when combined, can be a powerful tool for object detection and segmentation in images. For text data, you can opt for AI models like [BERT](#), [Longformer](#), [Flair](#), and CRFs that offer functionalities more tailored to the labeling tasks.



## PROS:

Saves time and cost for vast datasets, reduces human error for simpler tasks.

## CONS:

Accuracy can be lower than manual labeling, might not be suitable for complex tasks, requires high-quality training data for the automation tool itself.

## 2. In-House Labeling vs. External Labeling

In-house labeling involves building and managing your own team of annotators within your organization. In external labeling, there are two main approaches to external labeling: crowdsourcing and a dedicated labeling service. For a more in-depth comparison of in-house and external labeling, as well as tips on choosing the best approach for your needs, see the final chapter of this guide.

## 3. Open-Source vs. Commercial Labeling Tools

### OPEN-SOURCE TOOLS

Freely available software with the underlying code accessible to the public. Anyone can contribute to their development and improvement.

### Pro tip:

*“Choose manual labeling for small datasets, critical tasks requiring high accuracy, or projects with complex labeling needs. However, when you have large datasets but simpler tasks, it’s better to use automated labeling. This tactic can also be used as a pre-labeling step to improve efficiency in manual labeling workflows.”*



Karyna Naminas  
CEO of Label Your Data



## PROS:

Free to use and modify, allows for customization to specific needs.

## CONS:

They can't cover specific use cases, lack bulk data import/export via API, and require relying on community forums for help instead of dedicated support. Additionally, their functionality might be limited by the size of your dataset, and any customization often requires developer assistance.

## COMMERCIAL TOOLS

They are developed and offered by private companies and typically require a subscription or license fee for use. Some popular options include [Labelbox](#), [SuperAnnotate](#), and [Amazon SageMaker Ground Truth](#) data labeling tools.

## PROS:

Wide range of features for various labeling tasks, user-friendly interfaces, often include data security and quality control measures, provide technical support.

## CONS:

Can be expensive, might have limitations on customization.

## Pro tip:

*"If you have the technical expertise and a small project with specific needs, open-source tools can be a good option. For most projects, especially those involving large datasets, complex labeling tasks, or requiring user-friendliness and support, commercial data labeling tools are a better choice."*



Ivan Lebediev

Integration Specialist at Label Your Data



## 4. Public Datasets vs. Custom Datasets

### PUBLIC DATASETS

These are pre-labeled datasets readily available online from various sources like research institutions or government agencies. There's a wealth of free data available online, with a few top resources to explore, such as [Kaggle](#), [UCI Machine Learning Repository](#), and [Data.gov](#). They can be a valuable resource for getting started with ML projects.

#### PROS:

Readily available, free to use, can be a good starting point for initial training.

#### CONS:

Might not perfectly match your project's use case, may have quality or bias issues, might not be suitable for all tasks (e.g., privacy-sensitive data).

### CUSTOM DATASETS

A custom dataset is a collection of data specifically curated and prepared for a particular task or project. Unlike public datasets that are readily available online, custom datasets are tailored to address the unique needs of your model.

#### PROS:

Tailored to your specific project requirements, higher quality and relevance to your task.

#### CONS:

Require time and resources to collect and label data.

#### Pro tip:

*"Leverage public datasets to get started quickly, test your models, or for tasks where a perfect fit isn't crucial. Invest in building a custom dataset when public datasets can't cover your use case."*



**Karyna Naminas**

CEO of Label Your Data



## 5. Cloud Data Storage vs. On-Premise Storage

### CLOUD STORAGE

This tactic implies storing your data on remote servers managed by a cloud service provider (CSP) like Google Cloud Platform, Amazon Web Services (AWS), or Microsoft Azure. These providers offer scalable and readily accessible storage solutions, accessible from anywhere with an internet connection.

#### PROS:

Scalable storage capacity, easy access from anywhere, eliminates hardware management needs, often has built-in security features.

#### CONS:

Relies on internet connectivity, potential security concerns depending on the provider, can be more expensive for very large datasets over time.

### ON-PREMISE STORAGE

This involves storing your data on physical servers located within your organization's infrastructure. You have complete control over the hardware and its maintenance.

#### PROS:

Provides greater control over data security, eliminates reliance on external

providers, potentially lower ongoing costs for massive datasets.

#### CONS:

Limited scalability, requires in-house hardware management and maintenance, can be less accessible for remote collaboration.

#### Pro tip:

*“Choose cloud storage when your project requires scalability, easy collaboration, or limited in-house infrastructure. Be mindful of internet connectivity needs and ongoing costs for massive data volumes. Pick on-premise storage if*



*your project involves highly sensitive data, strict security requirements, or predictable and large storage needs where cloud costs might outweigh benefits.”*



**Ivan Lebediev**

Integration Specialist at Label Your Data

## Additional Tips

### 1. Ask questions before taking actions

Before you start the project, grasp the specific issues it's trying to solve:

- What does your ML project aim to achieve?
- How much and what type of data is needed?
- What sources will you use to gather data?

- How much time do you need to finish the project?
- What results do you expect?
- Is the budget sufficient for the results you want?

After answering these questions, you can set up a team and a data annotation process.

### 2. Plan, document, and secure your workflows

To enhance the scalability of data operations, document annotation workflows to establish standard operating procedures (SOPs). This not only protects datasets from theft and cyber threats but also ensures a transparent and compliant data pipeline according to data labeling and data privacy guidelines.

Before project commencement, make sure to:

- Establish clear processes,
- Obtain necessary labeling tools,
- Set a comprehensive budget covering tool expenses, human resources, and QA,
- Gain expert support,





Secure resources, including operating procedures.

### 3. **Treat data annotation as an iterative process**

To establish an effective strategy for data labeling ops, start with a small-scale approach. By doing so, you can learn from any minor setbacks that may arise, make necessary improvements, and then gradually expand the process. Starting small allows you to invest less time initially compared to starting with a larger dataset.

Regularly monitor annotation progress and be prepared to adapt the strategy based on feedback, challenges, and evolving project needs. Once you've achieved a smooth operation, including the integration of appropriate labeling tools, you can move forward with scaling up the entire operation.

### 4. **Communication and consistency are your cornerstones**

*"Data annotators are more effective when they understand the purpose behind their tasks. Summarize the*

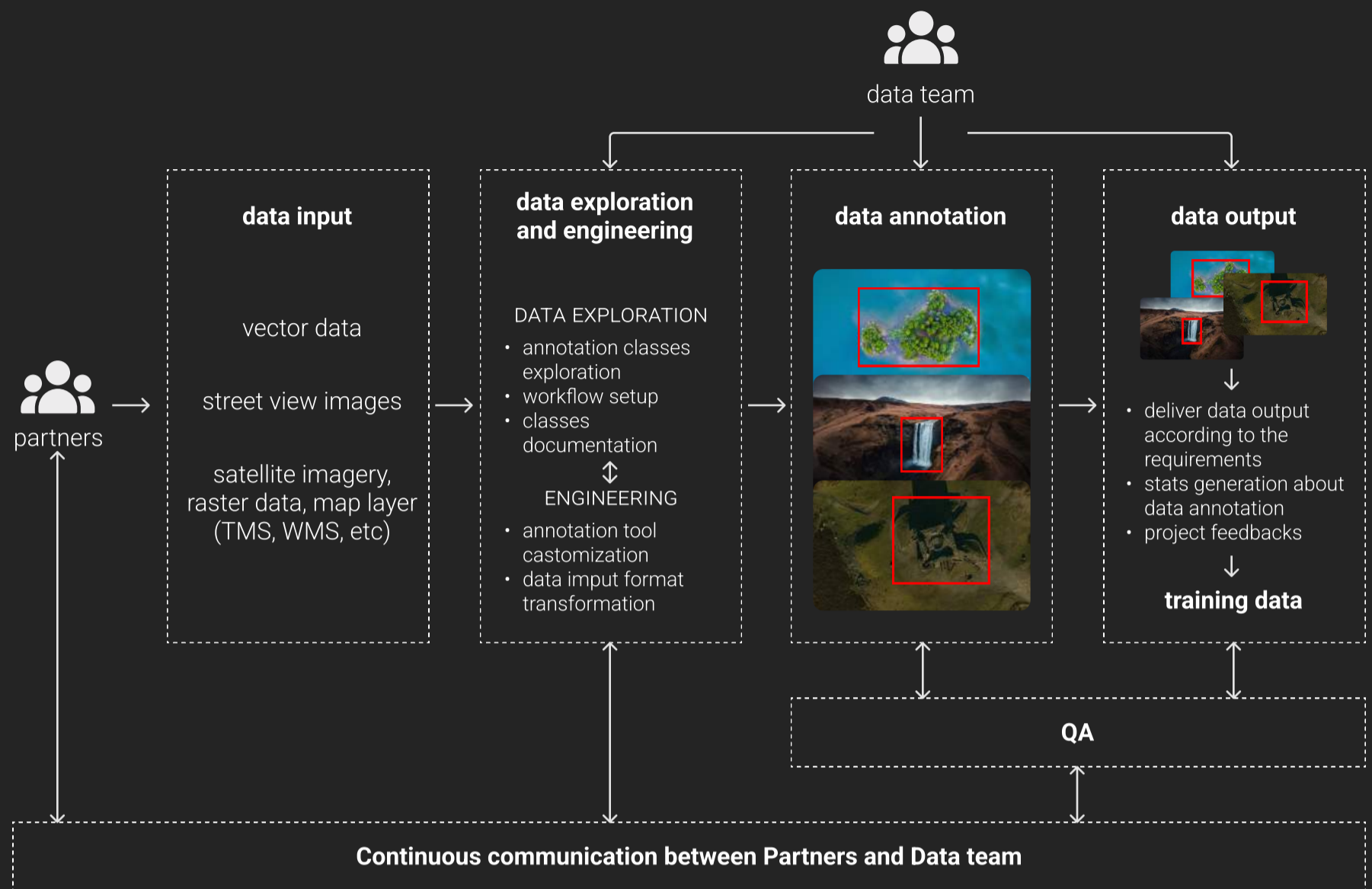
*guidelines by providing examples of a "gold standard" to assist in understanding complex tasks. Highlight edge cases and errors to minimize initial mistakes. Clearly communicate the evaluation criteria to annotators, preventing potential issues during reviews. Implement version control for guidelines to adapt to the ML project lifecycle."*



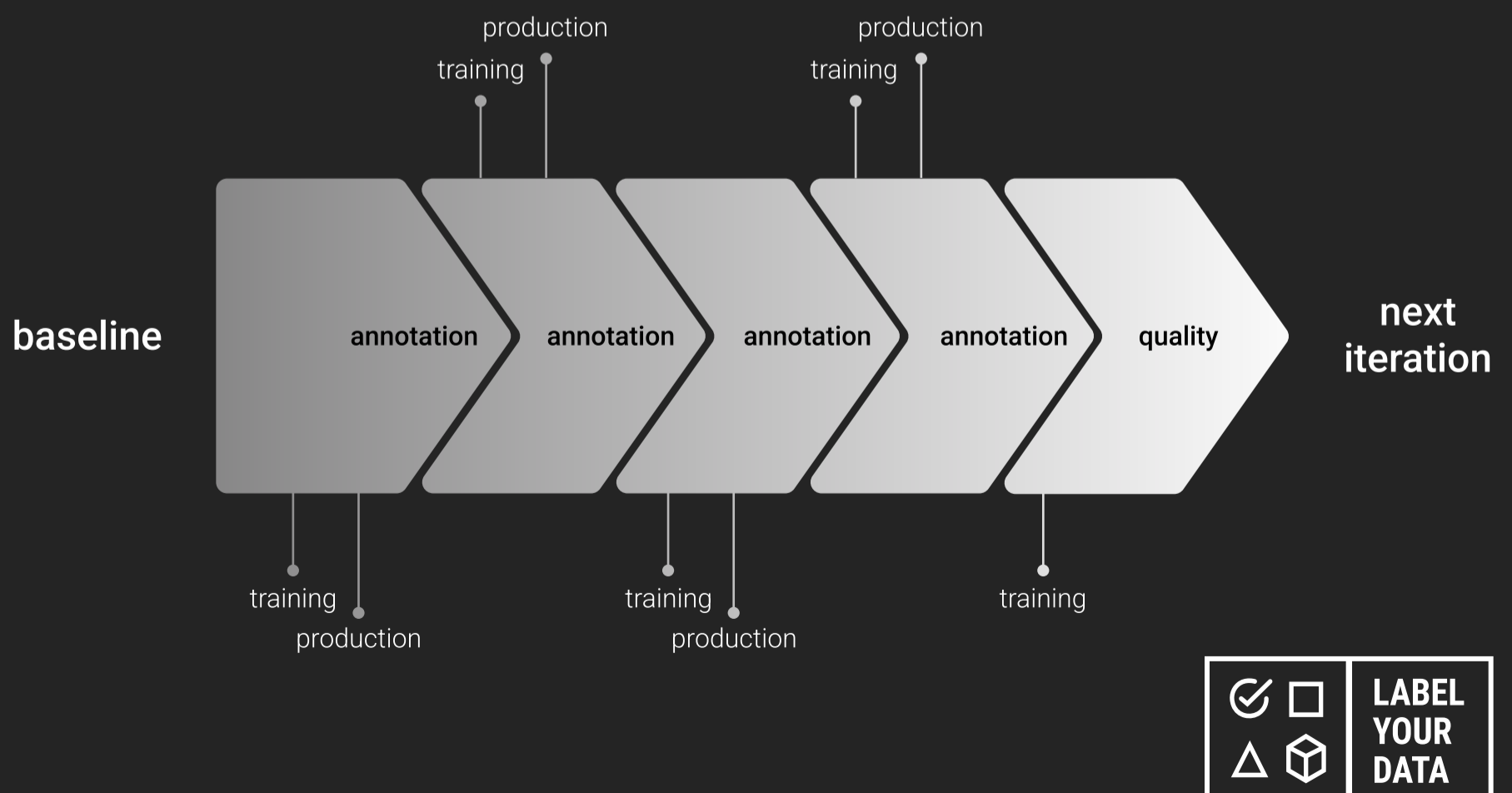
Karyna Naminas  
CEO of Label Your Data



# Communication within a data annotation team



## Data annotation is an iterative process



# CHAPTER 2:

## HOW TO MAINTAIN HIGH QUALITY OF LABELED DATASETS



VentureBeat states that about 90% of data science projects don't reach the production stage. The main reason for this, as found in [this report](#), is that 87% of employees blame data quality issues.

While the labeling process is done according to the set benchmarks of the project, measuring data quality is an inevitable step before the annotation is completed. Labeled data has a direct impact on the final performance of your model, which is why you should know the key practices for measuring its quality in the annotation process.

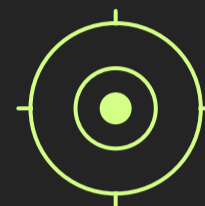
## The Key Methods for Measuring Labeled Data Quality

Since the initial ML dataset can differ in number and complexity, a few people can take part in the quality assurance.



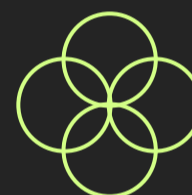
### data volume

more data for efficient model training



### accuracy

accurately annotated dataset



### relevance

data relevant for the ml project use case

To avoid inconsistency in final labels, the accuracy of data labeling is controlled on all stages, using one of these metrics:



**1.**

## Inter-Annotator Agreement (IAA) Metrics

To ensure that the approach of every annotator is consistent across all categories of the dataset, we refer to Inter-Annotator Agreement (IAA) metrics. They can apply for the entire dataset, between annotators, between labels, or per task.

### The most commonly used metrics include:

- Cohen's Kappa
- Krippendorff's Alpha
- Fleiss' Kappa
- F1 Score
- Percent Agreement

High level of IAA between annotator 1 and annotator 2 on all common tasks, %



● agreement ● disagreement

**2.**

## Consensus Algorithm

As the name suggests, the algorithm serves as a consensus between annotators on which label to use for defined datasets. Basically, every annotator provides their labels for the defined data, and the consensus algorithm applies to determine the final label and measure data quality. This method is considered one of the simplest, as annotators can choose the final label even by simple majority voting.

**3.**

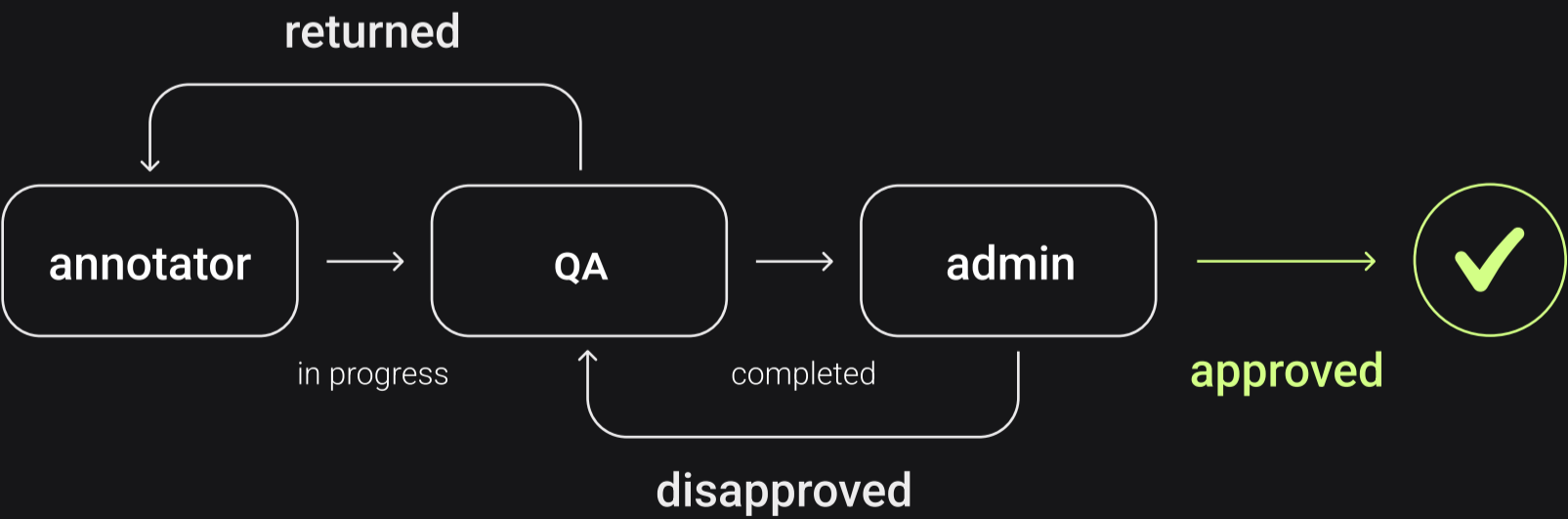
## Cronbach's Alpha Test

Another statistical method to assure the final annotation sticks to defined data labeling standards. It helps to check consistency and reliability of annotation across the dataset. The reliability coefficient is usually marked with 0 for completely unrelated labeling and 1 for high similarity among the final labels. As a result, the more alpha is closer to 1, the more final labels share in common.



METHODS	USE CASE
Inter-Annotator Agreement (IAA) Metrics	Ensure consistency across annotators
Consensus Algorithms	Simple and fast for choosing final label among annotators
Cronbach's Alpha Test	Measure consistency and reliability of annotations

# How to Set Up QA Procedures for Data Labeling



Developing a systematic quality assurance (QA) process significantly improves labeling quality. This process also follows an iterative cycle. Sometimes, to reduce human disadvantages, you may employ automated tools. Choosing the one that seamlessly integrates into your quality control workflow is crucial for faster resolution of annotation bugs and errors.

We at Label Your Data have honed a proven QA process that delivers consistent results for our clients:

**STEP 1.**

Gather all the instructions for the data to be annotated. They can contain requirements for further ML training, as well as ready examples, which we later use as a benchmark.

**STEP 2.**

Organize the training for all annotators involved in the project to ensure that the final labels meet the expectations and require no to minimum changes. At this stage, annotators receive all the instructions on how to label that particular dataset.

**STEP 3.**

Launch a pilot, which usually consists of a small part of the project. Check its quality and compare it to the initial instructions. If it's approved on the client's side and the data quality is high, continue with the rest of the dataset.



There are two more QA techniques for you to consider:

### Cross-reference QA

This method ensures the final labels are consistent by involving multiple experts performing annotations for further comparison and verification. The main result is to reach consensus between all annotators, especially in matters regarding subjectivity. We had cases where two or three annotators were doing the same task. These are usually the projects that contain datasets of text and maps.

### Random sampling

By randomly selecting multiple labels, we check that the quality corresponds to the project requirements. This approach is more relevant to smaller projects and is used as an extra step to regular quality control checks.

*"When dealing with large datasets, divide it into smaller milestones and tasks. Accomplish data quality control after*

*every task, and not only at the end of the project. This helps to save time on corrections and ensure all team members move in the right direction."*



Ivan Lebediev

Integration Specialist at Label Your Data

## Consequences of Poor Data Labeling Quality

With poor data, incorrectly trained models can lead to adverse consequences, especially in such areas of AI implementation as medicine or finances. The most common consequences of bad data labeling quality include:



## Biased models

Biased models are ML models that produce unfair results, often reflecting prejudices found in the data they're trained on, or the choices made during development. For instance, an algorithm trained to predict loan riskiness might deny loans to qualified individuals from certain neighborhoods based on historical biases.

## Incorrect performance metrics

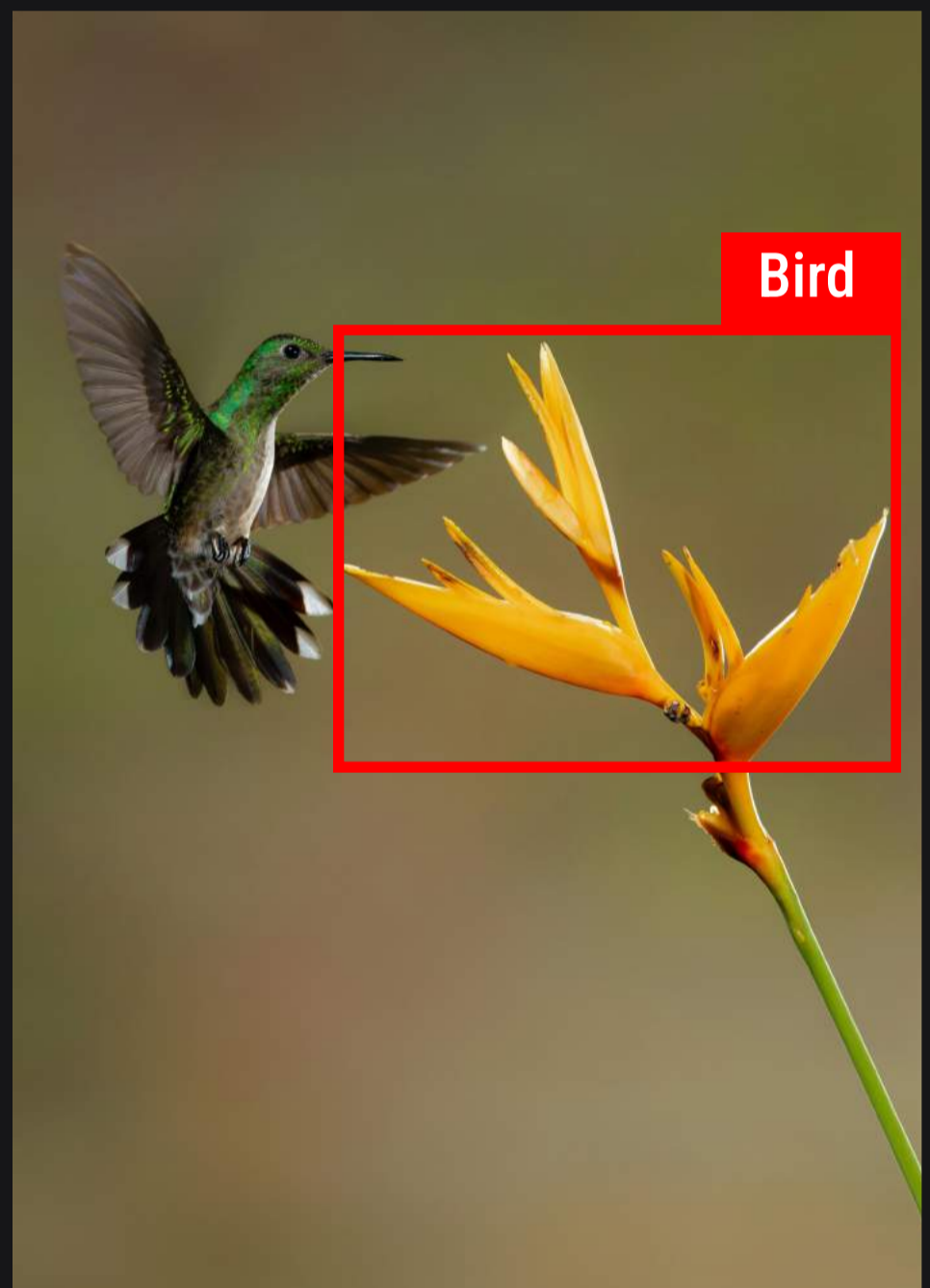
In machine learning, the evaluation of a model's performance hinges on the quality of data annotation. Inaccurate or inconsistent labels can significantly skew metrics like accuracy, rendering them misleading.

## Inefficiency of model development

Inefficiency in ML development arises from poor data labeling quality. In essence, the model is trained on inaccurate ground truth, leading it to learn faulty patterns. Consider an image recognition system: if poorly labeled training data confuses cats and dogs, the model will underperform, requiring significant time and resources for correction.

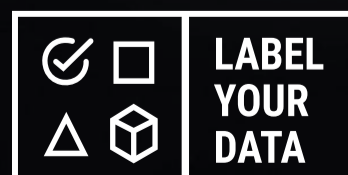
## Constraints of AI adoption

Poor data labeling quality hinders AI adoption. Inaccurate labels confuse the AI, leading to models that underperform or make biased decisions. Imagine if facial recognition software is trained on poorly labeled photos: it might misidentify people, raising privacy and security concerns.



# CHAPTER 3:

## HOW TO KEEP THE ML DATASETS SECURE?



On average, it takes 50 days to discover and report a data breach. During that time, businesses risk significant harm, including unauthorized access, financial losses, and reputational damage.

Ensuring data privacy regulations are followed can be tough when labeling personal data. You need systems that keep the data private by not letting people directly interact with it.

### data privacy



governing how data is collected, shared and used

### data security



protecting data from internal and external attacks

## How to De-Risk Yourself from Legal Issues with Private Data

To avoid legal headaches with private data, getting consent upfront from those who generate the raw data is essential for ethical data labeling in ML projects. This builds trust and ensures compliance with data privacy regulations.

### What happens when data is used without consent


Using data without consent can have a number of negative consequences. Perhaps the most significant is the erosion of user trust. When individuals feel their data has been misused, they're less likely to share information in the future, hindering the development of AI and other data-driven technologies.





Additionally, data breaches can expose personal information, leading to identity theft, fraud, and even physical harm. Legal repercussions are also a concern. Data privacy regulations around the world are becoming increasingly strict, and organizations that fail to obtain proper consent can face hefty fines. In the worst-case scenario, misuse of data can perpetuate discrimination or bias, leading to unfair and unethical outcomes.

## What are the examples of raw data


Raw data encompasses any unprocessed information collected from various sources. Here are some common examples you might encounter in data labeling projects:

 **Text data:** This includes emails, social media posts, documents, and chat logs. Text labeling tasks often involve sentiment analysis, topic classification, or information extraction.

 **Images and videos:** This can be any visual data you're working with for training computer vision models. Labeling tasks for this type of data could involve object detection, image segmentation, or scene classification.


 **Sensor data:** Data collected from sensors can include temperature readings, GPS coordinates, or audio recordings. Sensor data


labeling can be used for tasks like anomaly detection, activity recognition, or machine calibration.

 **Audio data:** Music, speech, environmental sounds, or animal sounds are some of the examples of audio data. Speech recognition and sentiment analysis are common types of labeling tasks used for applications like automatic transcription services, music recommendation systems, and real-time voice translation tools.

## How to ask consent from users

Obtaining user consent for data collection and labeling is crucial for ethical and legal reasons. Here are some key principles to follow:

 **Transparency:** Clearly inform users about what data is being collected, how it will be used, and who will have access to it.

 **Granularity:** Provide options for users to choose the specific types of data they're comfortable sharing.



**Control:** Allow users to withdraw their consent at any time and offer an easy way for them to access or delete their data.

**Clear language:** Use concise and easy-to-understand language in your consent forms, avoiding technical jargon.

**No dark patterns:** Here are the things to avoid when you want to ensure that user consent is truly informed, freely given, and reflects a clear understanding of how their data will be used:

- Pre-checked boxes: Don't pre-check consent boxes. Users should actively opt-in to share their data.
- Forced choices: Don't force users to agree to data collection as a condition of service use. Provide a clear "opt-out" option.
- Confusing language: Avoid burying consent requests within lengthy terms and conditions. Present consent information prominently and separately.
- Privacy nudges: Don't use misleading wording or pressure tactics to sway users towards giving consent.

By following these guidelines, you can ensure that your data labeling projects are conducted in a way that respects user privacy and builds trust.

## Data Privacy Protection Laws



It's important to process private data according to data laws. Today, over 120 countries have enacted international data protection laws to better safeguard their citizens' data.

Here's a list of the global data privacy regulations to consider when labeling ML datasets or outsourcing to the vendor:

### GDPR (General Data Protection Regulation)

The most extensive data protection and privacy regulation so far is the General Data Protection Regulation (GDPR), introduced in 2018 in the European Union and the European Economic Area. According to GDPR, people have the right to know how their information is being handled and to have more control over their data online.

Applies to the European Union (EU) and regulates how personal data of EU residents is processed by any organization, regardless of location. It emphasizes

transparency, individual control, and data security.

Key aspects include:

- **Individual Rights:** EU residents have rights to access, rectify, erase, and restrict processing of their data.
- **Lawful Basis for Processing:** Organizations need a legal justification for collecting and using personal data.
- **Data Breach Notification:** Data breaches must be reported to authorities and affected individuals.

### HIPAA (Health Insurance Portability and Accountability Act)

Applies to the United States and safeguards protected health information (PHI) of patients. It focuses on securing medical records and ensuring they are used only for authorized purposes.

Key aspects include:

- **Covered Entities:** Applies to healthcare providers, health plans, and healthcare clearinghouses.



- **Minimum Necessary Standard:** PHI should only be used to the minimum extent necessary for the purpose.
- **Patient Rights:** Patients have rights to access, amend, and request an accounting of disclosures of their PHI.

## CCPA (California Consumer Privacy Act)

While HIPAA focuses on protecting individuals' medical records and personal health information, while CCPA aims to enhance privacy rights and consumer protection for residents of California by regulating the collection and use of personal data by businesses.

CCPA empowers California residents with control over their personal data. It grants rights to know, access, delete, and opt-out of data sales. Businesses must provide clear privacy notices and respond to consumer requests. The CCPA focuses on data sales and has exemptions, but it paved the way for stronger data privacy protections in California and beyond.

## ISO 27001 (International Organization for Standardization)

Not a law, but an internationally

recognized standard for information security management systems (ISMS). It provides a framework for organizations to implement best practices for data security.

ISO 27001 is a global framework for managing information security in a company. Getting certified means your Information Security Management System meets international standards, assuring customers about your system's security. Certification involves evaluating your organization against 114 requirements across 14 security categories. While not specific to privacy, achieving ISO 27001 certification demonstrates a commitment to robust data security practices, which can be helpful for GDPR or HIPAA compliance.

When partnering with a data labeling company, both parties need to establish an agreement that outlines specific labeling details: confidentiality, compliance with laws and regulations, and the deletion or return of data after processing ends.



REGULATION	ML DATASET PROTECTION GUIDELINES
GDPR	Get explicit consent from individuals before using their data for labeling your ML datasets. Anonymize data whenever possible and only store what's necessary for your project. Be prepared to answer data subject requests about their information.
HIPAA	Only authorized personnel can access and label healthcare data. Implement strict security measures to protect patient privacy. De-identify data to the greatest extent possible before using it in your ML models.
CCPA	Allow individuals in California to opt-out of the sale of their personal information used for labeling your datasets. Provide clear privacy notices explaining how their data is used.
ISO 27001	While not a regulation itself, achieving ISO 27001 certification demonstrates a robust information security management system. This can help ensure your ML datasets are protected through access controls, data encryption, and other best practices.

# How to Organize Data Annotation Without Data Leaks

Even with user consent, a data leak during annotation can be a PR disaster. To prove our point, let's take a look at the recent case with Amazon.

## Amazon Ring Case

Ring, the video doorbell company owned by Amazon, [has settled a lawsuit in 2023](#) with the US Federal Trade Commission (FTC) for \$5.8 million over allegations that employees had improper access to customer videos. The FTC claimed Ring

employees had unrestricted access to all customer videos and could download or share them, while Ring maintained the data was encrypted and access was restricted for customer privacy. The settlement required Ring to implement a data security program and disclose employee data access procedures.

This case highlights the importance of data security and privacy regulations in data labeling, especially when dealing with sensitive data. Data labeling often involves human oversight. In the Ring case, the concern was Ring employees potentially having unfettered access to user videos used for training facial recognition or motion detection.



Therefore, it's crucial for data labeling companies to ensure secure storage and access controls to mitigate privacy risks.

For instance, data privacy regulations like the GDPR and CCPA would require limitations on such access. These regulations mandate user consent for data collection and usage, and restrict the retention and distribution of personal data. This can also involve data anonymization techniques to minimize the amount of personally identifiable information (PII) used in labeling.

top concern, followed by compliance (48%) and data access (44%). Hence, your data labeling process must adhere to relevant regulatory standards and security levels required for the data. It should provide a secure environment equipped with appropriate training, policies, and procedures to ensure compliance and data integrity.

**The key factors to consider for secure data labeling process include:**

- > **Annotators security:**  
Ensure that all annotators have

**Data protection**

**SECURITY**

encryption	network security	access control	discovery & classification	DSARs	consents
activity monitoring	breach response	DLP/CASB	3rd-party management	data removal	policies



**SECURITY**

**Best Practices for Keeping Data Secure**

Data quality is critical, with inconsistent collection standards (50%) being the

undergone background checks and have signed non-disclosure agreements (NDAs) or similar



documents outlining your expectations for data security. Managers should closely monitor compliance with these data security protocols.

> **Device control:**

Annotators should surrender any personal devices, such as mobile phones or external drives, upon entering the workplace. The service provider should also disable any features on work devices that could allow data downloading or storage.

> **Workspace security:**

Workers should conduct their tasks in a location where their computer screens are not visible to individuals who do not meet the specific data security requirements for your project.

> **Infrastructure:**

Use an appropriate labeling tool based on your unique needs and security standards, ensuring it offers robust access controls and encryption to safeguard sensitive data.

## The top strategies for securing data annotation:

### Ensure Physical Security:

Maintain secure facilities with manned security and metal detectors.

Restrict access to the building outside office hours.

Use video cameras to monitor the physical security of the workplace.

Require identification badges and biometrics for employee entry.

Prohibit personal belongings and electronics in secure areas.

Monitor access to sensitive data and limit it to authorized project teams.

Utilize polarized monitor filters to restrict data visibility.

Post reminders of critical security measures.

### Implement Internal Security Measures:

Provide consistent training sessions to educate annotators about recent data security risks, phishing, password management, and the importance of security.

Check the backgrounds of the people labeling the data.

Require employees to sign and adhere to various security policies, including codes of ethics and NDAs.



Conduct regular security audits to find weaknesses in security and implement suggestions from security experts.

### Implement Technical Security Measures:

Protect data using strong encryption like AES-256 to prevent unauthorized access.

Choose annotation software with built-in security features and follow standard security practices.

Don't allow the annotation team to use personal devices at work.

Add extra layers of security, requiring both a password and a physical item for login (Multi-Factor Authentication).

Limit access to sensitive data through role-based access control (RBAC) to reduce the risk of data leaks.

### Prioritize Cybersecurity:

Restrict internet access to necessary sites for each project.

Utilize proprietary chat tools for communication.

Conduct regular penetration tests and external audits to identify vulnerabilities.

### Maintain Security Compliance:

Adhere to industry-standard accreditations such as GDPR, CCPA, and ISO 27001.

Stay updated on security protocols and regulations to ensure compliance.

By following these steps, you can effectively enhance data labeling security and mitigate potential risks associated with sensitive data handling, such as:

- Annotators might access your data using an unsecured network or a device without proper protection against malware.
- They could save parts of your data by taking screenshots and sharing them through social media or email.
- Annotators might label your data while they're in public areas.
- Workers might not have enough training, understanding, or responsibility for following security procedures.



# CHAPTER 4:

## HOW TO HIRE DATA ANNOTATORS



While the amount of data continues to explode, finding and retaining skilled annotators can be a challenge. In addition, high turnover due to repetitive tasks can slow progress.

This section dives into the best practices for hiring data annotators, ensuring you build a strong team to effectively support your ML projects.

## 6 Steps to Build Your Annotation Dream Team

The repetitive nature of data labeling takes a mental toll. It demands focus and precision, but can lead to burnout and high turnover. This constant churn disrupts project timelines and increases training costs. Furthermore, it impacts the performance and consistency of dedicated annotators, ultimately driving up overall expenses for businesses.

To address these challenges and ensure high-quality data, effective hiring strategies are pivotal. By identifying candidates with the right skills, temperament, and training them effectively, companies can create a more engaged and resilient annotation workforce. This, in turn, leads to lower turnover rates, improved data quality, and reduced project costs.

### 1. How to Write Job Descriptions for Hiring Data Annotators

Crafting a compelling job description is the first step to attracting qualified data annotators. Here's how to structure your description for maximum impact:

**Grab attention with a catchy opening sentence.** Highlight the importance of the role and what the successful candidate will achieve.

**Clearly outline the responsibilities.** What are the day-to-day tasks of a data annotator in your company? Be specific about the data they will be



## Job Descriptions for Hiring Data Annotators

By writing a clear and compelling job description that highlights the key qualities you are looking for, you can attract top data annotator talent.

Join Label Your Data as a Data Annotator, discover the exciting world of data labeling, collaborate with the AI team, and learn the best practices along the way.  
What you'll be doing:

- Labeling texts and documents using special markup tools;
- Recognize and tag required details with high accuracy according to the rules of the task;
- Analyze data and appoint it to certain categories.

### Requirements:

- B2+ written and spoken English;
- Ability to quickly analyze and achieve required KPIs of performance and quality;
- Confident use of Google services (docs, sheets, slides, drive);
- Understanding of IT terminology.

### We offer you:

- Flexible schedule;
- Enjoy working remotely;
- Intensive training;
- Competitive compensation in USD;
- Great management with no bureaucracy;
- Financial and professional growth;
- Good bonuses for referring friends (referral program).

### Who are we?

Machine Learning (ML) is setting a new milestone for the future of technology development. Data is the most vital resource for any AI-boosted application.

Building on this idea, Label Your Data has been providing high-quality, secure, and flexible data annotation services for any industry, including Retail & E-commerce, Security, Fintech, Health Care, Real Estate, Autonomous Vehicles, Insurance, and Robotics since 2019.

We're a growing company aiming to create state-of-the-art, reliable, and tech-driven data annotation solutions for businesses all over the world. The safety of our client's data is our priority. Grab your chance to join us, and send us your CV in English pointing out your outstanding skills.

Visit our website: <https://labelyourdata.com/>



working with and the tools they will be using.

**List essential skills and experience required.** Not all data annotation projects require the same skill set. Tailor this section to the specific needs of your project. For example, some projects may require experience with specific labeling tools like CVAT, while others may require Excel knowledge.

**Highlight the benefits you offer.** Competitive salary and benefits are important, but don't forget to showcase what makes your company unique. This could include opportunities for growth, project variety, flexible employment types, or a positive work environment like we provide at Label Your Data.

**In addition, focus on attracting candidates with the following key qualities:**

- **Attention to detail.** Even small mistakes can have a big impact on the quality of your ML dataset.
- **Ability to handle large data volumes.** Data labeling often involves processing large datasets. Make sure your candidates can work quickly and accurately without getting overwhelmed.

*"The most critical point for annotator job description is working on your EVP. It's an Employee value proposition (EVP) that serves as a magnet for high performers. Write a concise message that tells potential hires exactly what it's like to work at your company. By highlighting what sets you apart, you'll attract, engage, and retain the best talent."*



Liudmyla Boichun  
HR Director at Label Your Data



- **Willingness to work with monotonous work.** Data annotation can involve some repetitive and mundane tasks. While you want candidates who can be meticulous, you also want those who can stay motivated and engaged over time.
- **Analytical mind.** The ability to identify and address inconsistencies or ambiguities in the data is an important skill for a data annotator.

By writing a clear and compelling job description that highlights the key qualities you are looking for, you can attract top data annotator talent.

## 2. Where to Publish Job Vacancies

Once you have a great job description written, it's time to get it in front of the right candidates.

### Target job postings based on location

For a broader international reach, consider using an Applicant Tracking System (ATS) to post your jobs on platforms like Jooble, Startup Jobs, and LinkedIn. Be sure to tailor your postings to the specific countries you are targeting by highlighting relevant skills and experience.

***“Grow your team through a referral program. Your existing data annotators can be a valuable source of new talent. Encourage them to recommend friends and colleagues who might be a good fit for your company. This is a cost-effective way to find qualified candidates who already have some understanding of the role and company culture.”***



**Liudmyla Boichun**

HR Director at Label Your Data



By using a combination of online platforms and a referral program, you can increase your chances of finding the best annotators for your team.

## 3. How to Interview Data Annotators

The interview process is your chance to assess a candidate's skills and suitability for your ML project. Here's a breakdown of the key interview stages to consider:

1.

### **Start with a brief introduction explaining the interview format.**

Briefly outline the topics that will be covered and allow the candidate to ask any questions they may have.

2.

### **Discuss the candidate's experience.**

Even if the candidate's experience is not directly related to data annotation, this can still help you understand their work ethic, transferable skills, and ability to learn new things.

3.

**Gauge their understanding of data annotation.** Ask questions to assess their knowledge of common data annotation tasks and tools to see if they did any research to study the topic.

4.

### **Ensure a company culture fit.**

Discussing your values and work environment to give the candidate a chance to learn about your company culture and see if it aligns with their own values and work style.

5.

**Watch out for red flags.** This can be anything from negativity towards past employers to odd questions that could signal potential issues.

6.

### **Use a presentation to showcase your company.**

Highlight growth, projects, values, and culture in your presentation. This is a chance to impress the candidate and give them a reason to want to work for you. Encourage questions. Let them see the exciting work you're doing and the positive environment they could be a part of.

7.

### **Provide a project-specific test task to assess skills.**

This will give you a firsthand look at the candidate's ability to perform the actual tasks of the job. See how they handle the data, use the tools, and approach any challenges that arise.



Conducting a well-structured interview will help you gain valuable insights into a candidate's qualifications and suitability for the data annotator role.

## 4. How to Choose the Best Data Annotators

Once the interviews are complete, carefully evaluate each candidate and select the best fit for your team. Here are some key strategies:

Evaluate each candidate after the interview. Take time to complete written evaluations that assess their strengths, weaknesses, and overall impression on the team. Consider factors like their technical skills, attention to detail, problem-solving abilities, and cultural fit.

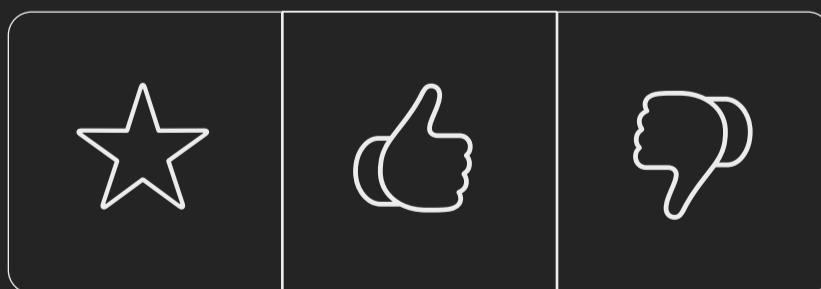
Use Google forms to assess their workplace behavior. Analyze responses from a Google form used during the interview process. This form can provide valuable insights into the candidate's work style, comfort level with communication, and adaptability. For example, how do they handle hypothetical work scenarios or questions about teamwork?

### Look for candidates who demonstrate the following:

- > Strong and timely performance on the test task. This is a clear indicator of their ability to handle the specific demands of the data annotation role.
- > A genuine interest in data annotation and AI. Passionate candidates are more likely to be engaged and motivated in their work.
- > The ability to ask thoughtful questions. This shows their curiosity and desire to learn more about the role and your company.

By carefully evaluating each candidate against these criteria, you can identify the top performers.

### Should this candidate proceed to the next stage ?



## 5. How to Retain Data Annotators

In today's competitive job market, retaining top talent is crucial. Here are some strategies to keep your data annotators happy and engaged:

**Focus on positive employee experiences.** Invest in creating a positive work environment that fosters employee well-being and satisfaction. At Label Your Data, we've created a dedicated People Experience Team that specifically helps annotators with onboarding, communication, and addressing concerns.

**Conduct regular check-ins with annotators throughout their employment.** Schedule regular meetings to discuss their work, address any concerns they may have, and provide valuable feedback. This ongoing communication helps them feel valued and supported.

**Offer clear career paths within the company.** Make sure to provide opportunities for career growth and advancement, such as promotions from Project Annotator (part-time employment) or Dedicated Annotator (full-time employment) to Project Supervisor or even an Account Manager.

**Provide flexible work arrangements for a work-life balance.** Offering flexible hours, remote work options, or compressed workweeks can help your data annotators achieve a healthy balance between their work and personal lives.

*"Offer diversity to keep things interesting. Data annotation can involve repetitive tasks. To maintain annotator engagement and mitigate repetitive work, propose a variety of data types for labeling. Additionally, consider offering part-time arrangements or rotating tasks within*



*projects to prevent annotators' burnout and ensure long-term focus on high-quality annotations."*



**Ivan Lebediev**

Integration Specialist at Label Your Data

This way, you can create a work environment that fosters loyalty and reduces employee turnover, allowing you to retain your top annotation specialists.

## 6. How to Use the Referral Program

Don't underestimate the power of your existing data annotators. A referral program can be a goldmine for attracting top talent.

### Here's why:

#### Proven results:

Referral programs accounted for

25-40% of our main candidate traffic within two years at Label Your Data. That's a significant chunk of qualified applicants coming through the company.

#### Quality referrals:

Candidates referred by current employees already have a certain understanding of the job, expectations, and company culture. They've likely heard positive feedback from their friends, which translates to motivated individuals who are ready to hit the ground running.

#### Cost-effective:

Compared to traditional recruitment methods, referral programs are a budget-friendly way to find qualified candidates. You're leveraging your existing employee network, reducing advertising and agency fees.

### There are two main ways to structure your program:

1.

**Internal referrals:** Encourage employees to recommend friends or colleagues who might be a good fit.



2.

**External referrals:** Open your program to the broader network, allowing anyone to refer qualified candidates.

No matter which approach you choose, make sure to offer attractive incentives to encourage participation. This could be a cash bonus, additional paid time off, or other perks.

## Additional Hiring Tips

No matter which approach you choose, make sure to offer attractive incentives to encourage participation. This could be a cash bonus, additional paid time off, or other perks.

1.

**Leverage your network:** Utilize job boards like Indeed, Glassdoor, and LinkedIn, targeting experienced professionals. Engage with relevant AI communities on social media. Encourage employee referrals for qualified candidates.

2.

**Look beyond your network:** Consider partnering with third-party annotation

companies for leveraging their existing pool of pre-trained professionals. Create targeted online ads to attract qualified individuals, focusing on relevant keywords, such as “data labeling expert” or “AI project annotator.”

3.

**Prioritize quality over cost:** Assess the annotators’ skills, experience, and portfolio to ensure they meet your ML project’s needs. Remember, good data annotation is an investment in your AI’s accuracy.

4.

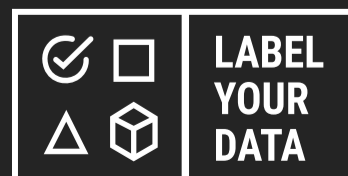
**Choose the right hiring approach:**

**In-house:** Offers control and consistency but can be expensive.

**Freelancers:** Cost-effective with access to specific skills, but managing quality can be challenging.

**Outsourcing:** Turnkey solution with minimal management, but expensive and less flexible.

**Crowdsourcing:** Highly cost-effective, but ensuring consistent quality is difficult.



# CHAPTER 5:

## HOW TO TRAIN DATA ANNOTATORS



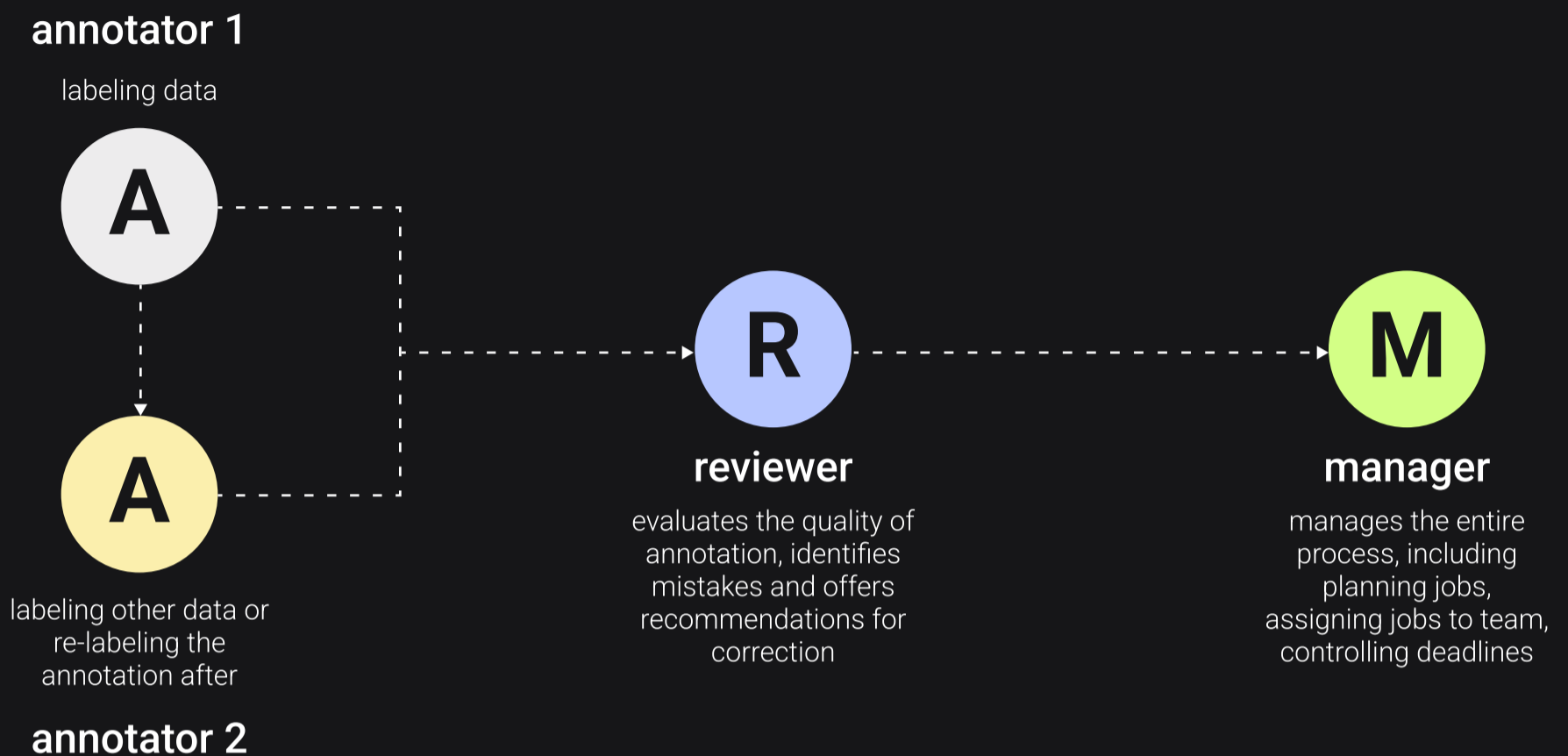
Building a dedicated in-house team can be advantageous, especially for ML projects with highly sensitive data or intricate labeling needs. However, this requires a significant investment in the training process.

The crucial aspect lies in understanding the nuances of data labeling. For instance, you might have a medical data labeling project, which requires training in differentiating benign vs. malignant tumors and recognizing specific organs for accurate AI applications in healthcare. Therefore, the labeling process cannot do without a proper training of annotators and often necessitates specialized skills from subject-matter experts (SMEs) for certain data types and industry-specific knowledge.

## Data annotator learning path



# How to Set Up a Data Annotation Team



## A data annotation team consists of:

**Data annotation specialists** who prepare high-quality training data for ML models. Their work directly impacts the success of ML projects by guaranteeing the integrity and effectiveness of training data.

**Project manager**, or a data annotator team manager, who coordinates the annotation process, allocates tasks, and manages timelines to ensure project completion.

## QA (Quality Assurance)

**specialists** who verify the accuracy and consistency of annotations, ensuring high-quality labeled datasets.

**Subject-matter experts** that provide domain-specific knowledge to guide annotators in accurately labeling data relevant to the project.





**Data scientists or ML engineers** can sometimes be on the team to oversee the annotation process from a technical perspective, ensuring alignment with the requirements of the ML models being developed.

## Top Training Steps for Data Annotators

Effective data labeling requires a well-trained team. Taking these steps, you are more likely to build one:

1.

### Define your data annotation process clearly

- **Document guidelines:** Establish clear instructions for labeling conventions, training procedures, and quality control measures. Make them readily accessible and regularly updated to ensure everyone is on the same page.
- **Training procedures:** Streamline onboarding for new members and ensure existing members stay aligned. Encourage real-time questions and provide written feedback during training.

*"A clearly defined process ensures clarity, consistency, and enhanced efficiency. First, it fosters a collaborative environment where each of the annotation team members understands roles, expectations, and quality standards. Second, clear procedures minimize confusion and wasted time, leading to faster completion and consistent high-quality data."*



Ivan Lebediev

Integration Specialist at Label Your Data



2.

## Establish effective training procedures

An effective training procedure is crucial for building a data annotation team. Here's how to achieve one:

- Clear communication: Establish clear and consistent guidelines to avoid confusion and maintain data annotation quality.
- Onboarding and ongoing support: Defined procedures ensure efficient training for new members and continued reference for experienced members.
- Consistency in labeling: Consistent application of annotation standards across the team is crucial for reliable data applicable for machine learning models.

The success of training annotators for a project hinges on several factors, including the project's specific needs, time constraints, and the capabilities of the managing team. When planning your project timeline, consider the duration of the training phase. It depends on the experience level of the workforce, the project's complexity, and the chosen method for ensuring data quality.

3.

## Additional considerations for a data annotation team building

Beyond the core training, consider these additional factors:

- **Tagging ontology:** Design for consistency by considering potential edge cases and using contrasting examples to clarify labeling.
- **User experience:** Design task guidelines with ergonomics and collaboration in mind.
- **Language and culture:** Consider variations when setting up tag sets and data collection guidelines.
- **Team diversity:** Create a data annotation team that is diverse, with relevant language skills, and different backgrounds to reduce bias in data models and ensure fair outcomes.
- **Performance monitoring:** Implement a plan to address consistently low performers to maintain data quality.



- **User-friendly tools:** Choose intuitive and user-friendly data annotation tools for efficient data processing and higher-quality results.

By following these steps, you can build a strong annotation team that delivers consistent, high-quality training data. But it's crucial to have a system in place for identifying and addressing consistently underperforming annotators. This could involve either providing additional training for annotators or, if necessary, removing them from the project to safeguard the quality of the ML datasets.

## The Role of Subject-Matter Specialists

Subject-matter experts (SMEs) play a crucial role in building a data annotation team for several reasons:

1.

**Domain-specific knowledge:** SMEs possess deep understanding and expertise in the specific field or domain the data pertains to. This allows them to accurately interpret, categorize, and label data points with the necessary context and nuance.

2.

**Quality assurance:** Their expertise enables them to identify inconsistencies, ambiguities, and potential errors in data annotation.

3.

**Developing guidelines and standards:** SMEs are instrumental in establishing clear and consistent annotation guidelines and standards for the team to follow. This minimizes discrepancies in how different annotators perceive and annotate the data.

While not always feasible to have a team solely of SMEs, their involvement is vital for ensuring high-quality, reliable data annotation.



# CHAPTER 6:

## HOW TO CHOOSE BETWEEN IN-HOUSE VS. OUTSOURCED DATA LABELING



Machine learning requires deep labeling expertise to handle growing data volumes and complex goals. This can be achieved through:

- **Building an internal team:** Hiring data annotators in-house.
- **Outsourcing:** Partnering with a data annotation vendor.

Yet, most AI engineers struggle with the time and resources required for in-house labeling, while also facing concerns about data security, consistency, and potential bias when outsourcing. Here's how to decide which approach fits your labeling needs best.

Having your own team of data annotation specialists is great, but it goes hand in hand with tons of human resources to onboard, finances to spend, and time and efforts to devote. Managing a labeling expert is a commendable effort for the company.

Hiring and training an in-house annotation team is the right thing to do when your ML project is long-term and

includes large datasets. This also ensures that the project is carried out safely, following the highest data security standards.

### In-House Annotation Team, %



- building the ML model
- preparing data for the ML model

- consistent annotation
- better understanding of the task
- the ability to change the instructions anytime
- close cooperation with the development team
- lower error rate
- improved the time-to-market



To build an in-house labeling team, you need to:

- Allocate HR and financial resources
- Develop a labeling tool or use ready-made solutions
- Build a QA team for error risk reduction
- Supervise the annotation team



PROS OF IN-HOUSE ANNOTATION

Consistent, reliable process for long-term success

Continuous improvement through feedback loop

Strong quality control and lower error rate

Choice of existing tools or in-house development

To lead an internal labeling team effectively, the company must strike the right balance between strategic development and high-quality performance of labeling tasks.

CONS OF IN-HOUSE ANNOTATION

Not practical for all data/company sizes

Expensive and time-consuming to set up

Requires investment in finding the right tools

May require a large team for complex or large data



Still, it's not a scalable solution due to operational issues and insufficient training data expertise, unless it's managed by a tech giant.

## Data Annotation Outsourcing



deep data expertise



speed



quality and security



cost-effectiveness



professional experience



technical infrastructure

If you decide to outsource the labeling tasks to a third party, all the burden associated with an in-house option immediately casts off.

Most companies specializing in data annotation have state-of-the-art tools and software that allows clients to review their tasks and monitor the progress. Professional outsourcing partners also provide customized solutions to satisfy different ML projects' needs.

*"Outsourcing works well when there's a clear vision of the rules and standards for the training data used to train the algorithm for a specific use case. However, it inevitably entails the issue of finding a trustworthy vendor whose service*



*is built around the highest data security and privacy standards.”*



**Karyna Naminas**  
CEO of Label Your Data

- When is outsourcing the smart move to make?
- When you want to focus on model development rather than data
  - When you’re looking for guaranteed quality and efficiency
  - When you need to scale the process effortlessly

**PROS OF DATA ANNOTATION  
OUTSOURCING**

- High-quality work through hand-selected workforce
- Cost-effective compared to in-house teams
- Tailored solutions through consultation
- Efficient handling of large, diverse data volumes
- Strong security protocols

**CONS OF DATA ANNOTATION  
OUTSOURCING**

- More expensive than crowdsourcing
- Knowledge transfer limited due to external workforce
- Setup time can be lengthy depending on data complexity
- Professional approach might be overkill for simple projects
-

Both options have their own benefits, but your final decision when choosing between in-house and outsourced data annotation would depend on the following factors:



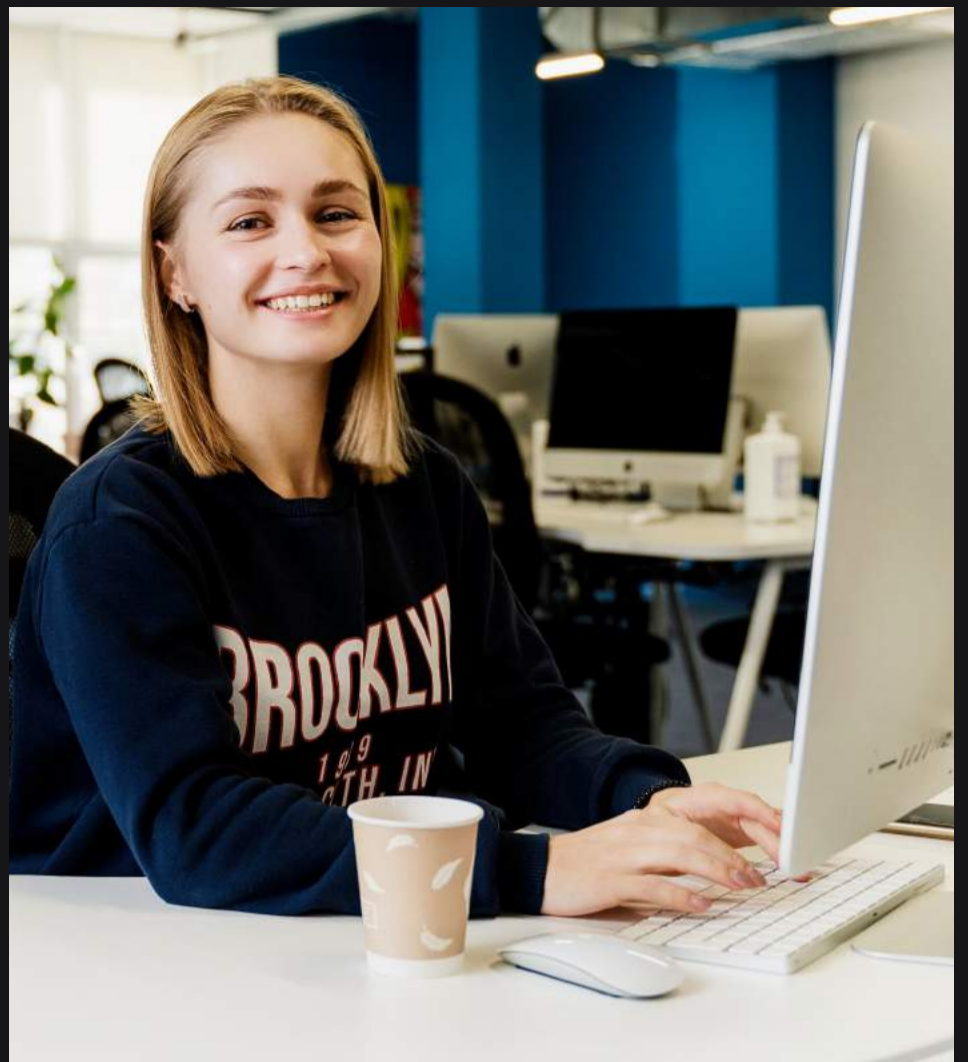
FACTOR	IN-HOUSE LABELING	OUTSOURCED LABELING
Flexibility	Suitable for simple ML projects where internal control and easy communication are crucial. Yet, scaling up for complex projects can be challenging.	Ideal for complex ML projects with specific labeling needs, offering access to a wider range of expertise and diverse datasets. Flexibility might be limited by the vendor's capabilities.
Pricing	Expensive due to infrastructure and training costs, including hiring dedicated staff, procuring software, and maintaining hardware. While cost-effective in the long run for high-volume projects, upfront costs can be significant.	Generally more affordable with various pricing plans based on data volume, complexity, and turnaround time. Finding the right balance between cost and quality requires careful evaluation of vendors
Management	Requires significant investment in time, money, and resources to manage an in-house team, including recruitment, training, performance evaluation, and QA. This can divert resources away from core development activities.	Frees up internal resources to focus on core development activities like ML model development. However, managing a vendor relationship also requires effort, including establishing clear communication channels and monitoring performance.
Training	Requires significant time and money for training annotators on specialized tools, project-specific guidelines, and QA processes. This can delay project timelines and impact initial costs.	No training costs as data labeling service providers have experienced teams that can adapt quickly to project requirements and tools. Yet, ensuring consistency in annotation quality might require additional oversight
Security	Offers higher data security as project details and sensitive data remain within the organization. This is crucial for projects involving confidential information or regulated industries.	Lower inherent security risk as data is shared with a third party. Choosing providers with robust security protocols, data encryption, and compliance certifications is essential to mitigate risks.
Time	Generally slower due to the time required for team training, infrastructure setup, and initial project setup. This can be a downside for projects requiring fast turnaround times.	Faster due to established provider infrastructure and readily available skilled team. This can be great for projects with tight deadlines or ongoing data annotation needs.



# Why Choose Label Your Data

We hope this guide will help set up an internal data labeling workflow for first-timers or improve it for existing teams.

If you choose to delegate the data labeling ops, consider running a free data pilot with us to compare your performance results. Our outsourcing strategy has helped many companies to scale their ML projects.



## Security

Dealing with sensitive data like financial or medical records can be tough. We at Label Your Data can take care of all your data annotation needs, adhering to strict security standards such as PCI DSS Level 1, ISO 27001, GDPR, and CCPA.



## Flexible pricing

Fixed pricing models might hold back your machine learning project. That's why we offer a variety of pricing options at Label Your Data. You can tailor your plan to match your project timeline and allocate resources where they're most needed.



## Tool-agnostic

Struggling with limited tools for data labeling? Whether you prefer our in-house solution or have your own tool of choice, our data annotators seamlessly adjust to meet your specific requirements.



## No commitment

Worried about getting stuck in a long-term commitment? With Label Your Data, you can put those fears aside. We offer a free pilot so you can experience our performance firsthand without any strings attached.

# Run free pilot!