LABEL
YOUR
DATA

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.

”

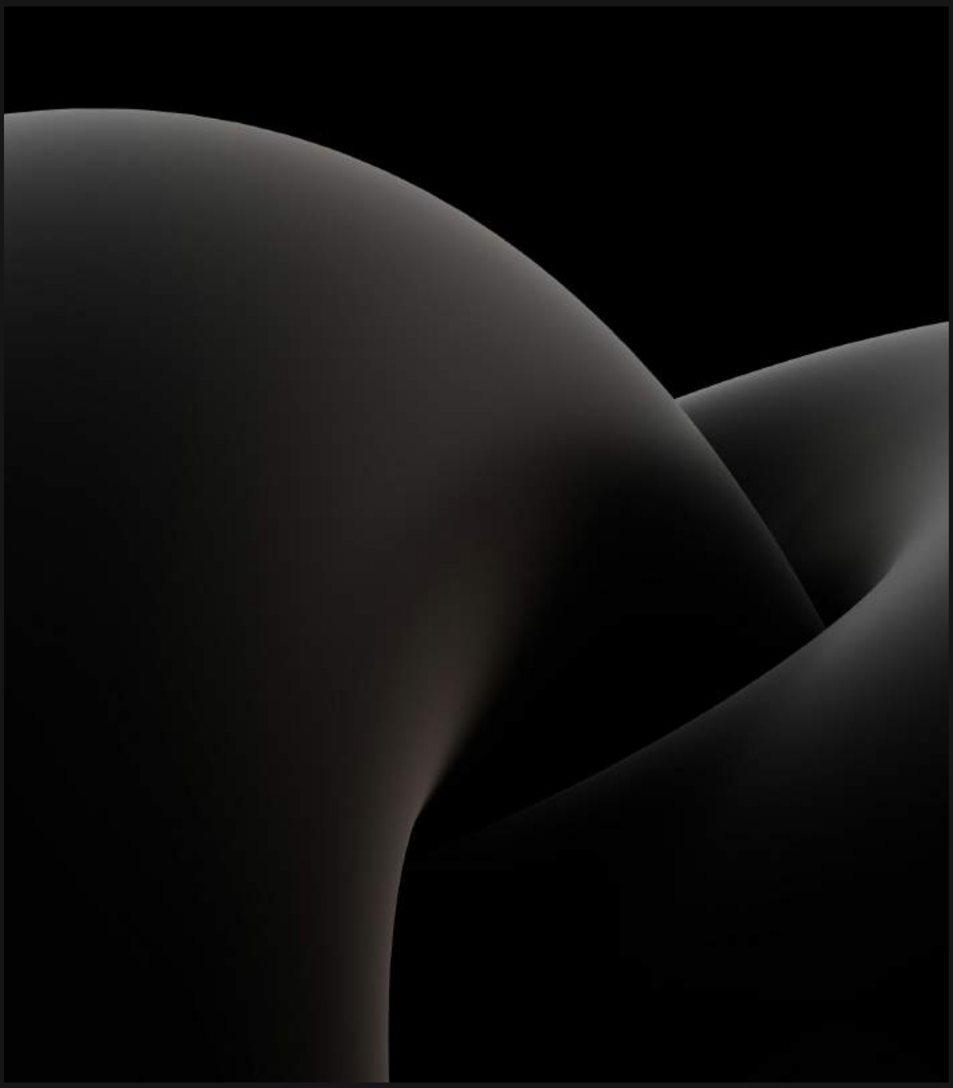
Karyna Naminas,
CEO of Label Your Data 🤍

Need feedback on your ML data annotation setup?

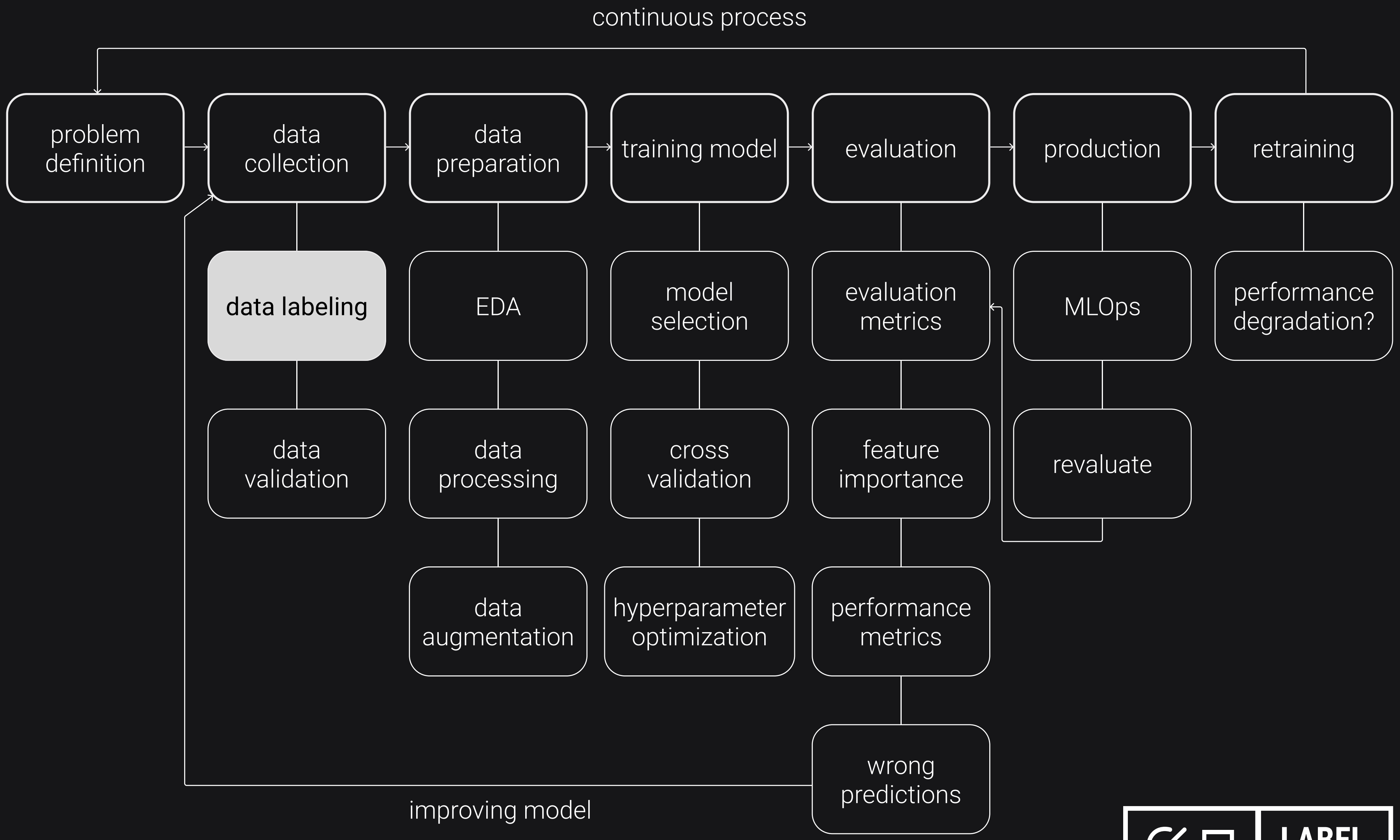
LABEL YOUR DATA

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



LABEL YOUR DATA

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

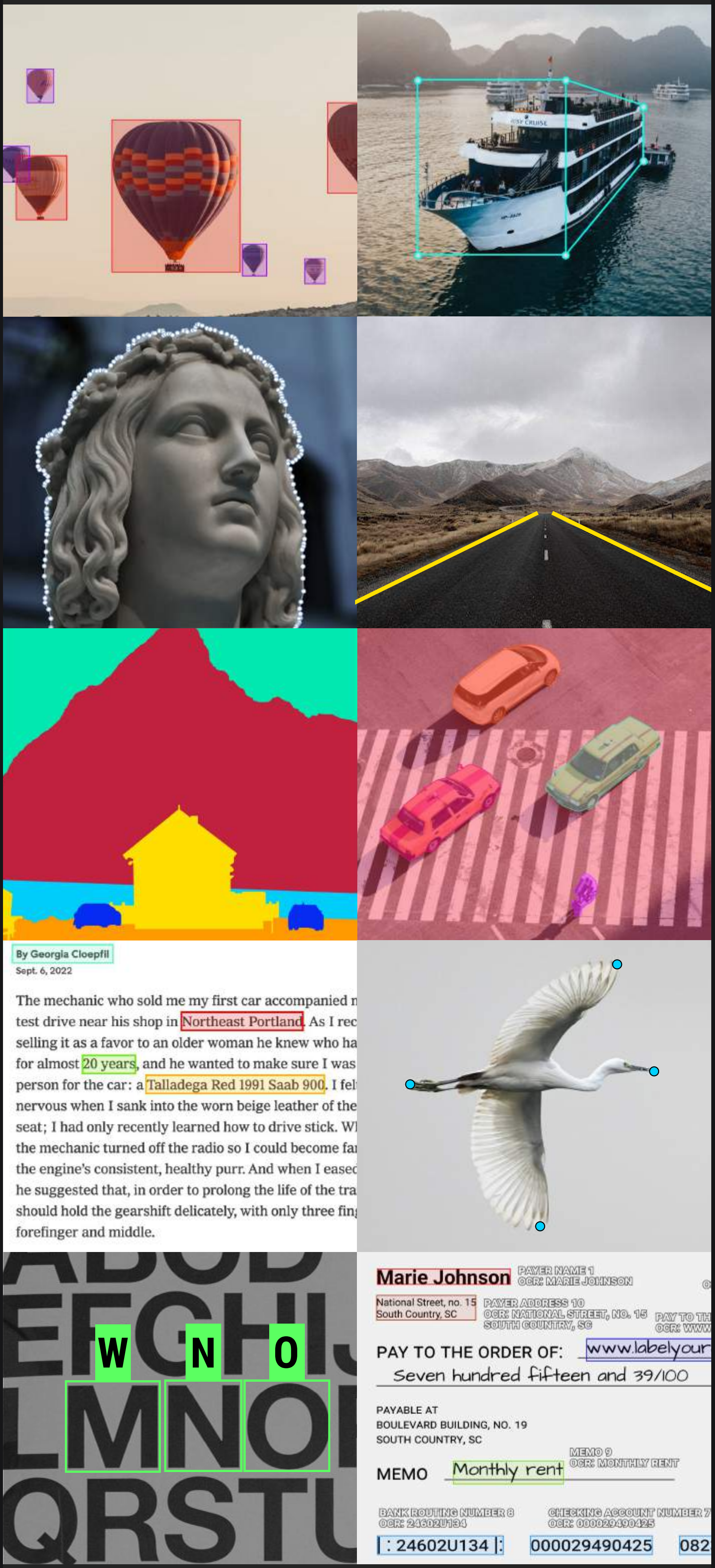
For Natural Language Processing (NLP) models, data labeling requires annotators to possess linguistic knowledge for handling the following types of text and audio data annotation:

- 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



By Georgia Cloepfil
Sept. 6, 2022

The mechanic who sold me my first car accompanied n
test drive near his shop in Northeast Portland. As I rec
selling it as a favor to an older woman he knew who ha
for almost 20 years, and he wanted to make sure I was
person for the car: a Talladega Red 1991 Saab 900. I fel
nervous when I sank into the worn beige leather of the
seat; I had only recently learned how to drive stick. Wl
the mechanic turned off the radio so I could become fa
the engine's consistent, healthy purr. And when I easec
he suggested that, in order to prolong the life of the tra
should hold the gearshift delicately, with only three fin
forefinger and middle.

Marie Johnson
PAYER NAME 1
OCR: MARIE JOHNSON

National Street, no. 15
South Country, SC
PAYER ADDRESS 10
OCR: NATIONAL STREET, NO. 15
SOUTH COUNTRY, SC

PAY TO TH
OCR: WWW

PAY TO THE ORDER OF: www.labelyour
Seven hundred fifteen and 39/100

PAYABLE AT
BOULEVARD BUILDING, NO. 19
SOUTH COUNTRY, SC

MEMO Monthly rent
MEMO 9
OCR: MONTHLY RENT

BANK ROUTING NUMBER 8
OCR: 24602U134
CHECKING ACCOUNT NUMBER 7
OCR: 000029490425

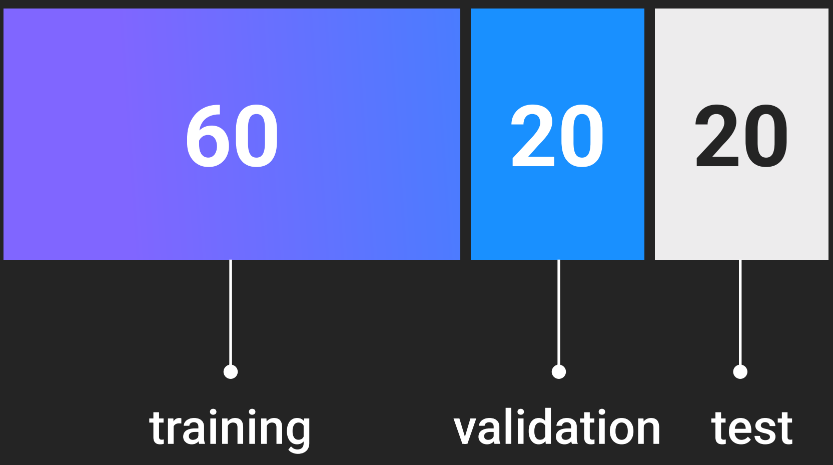
|: 24602U134 |: 000029490425 082

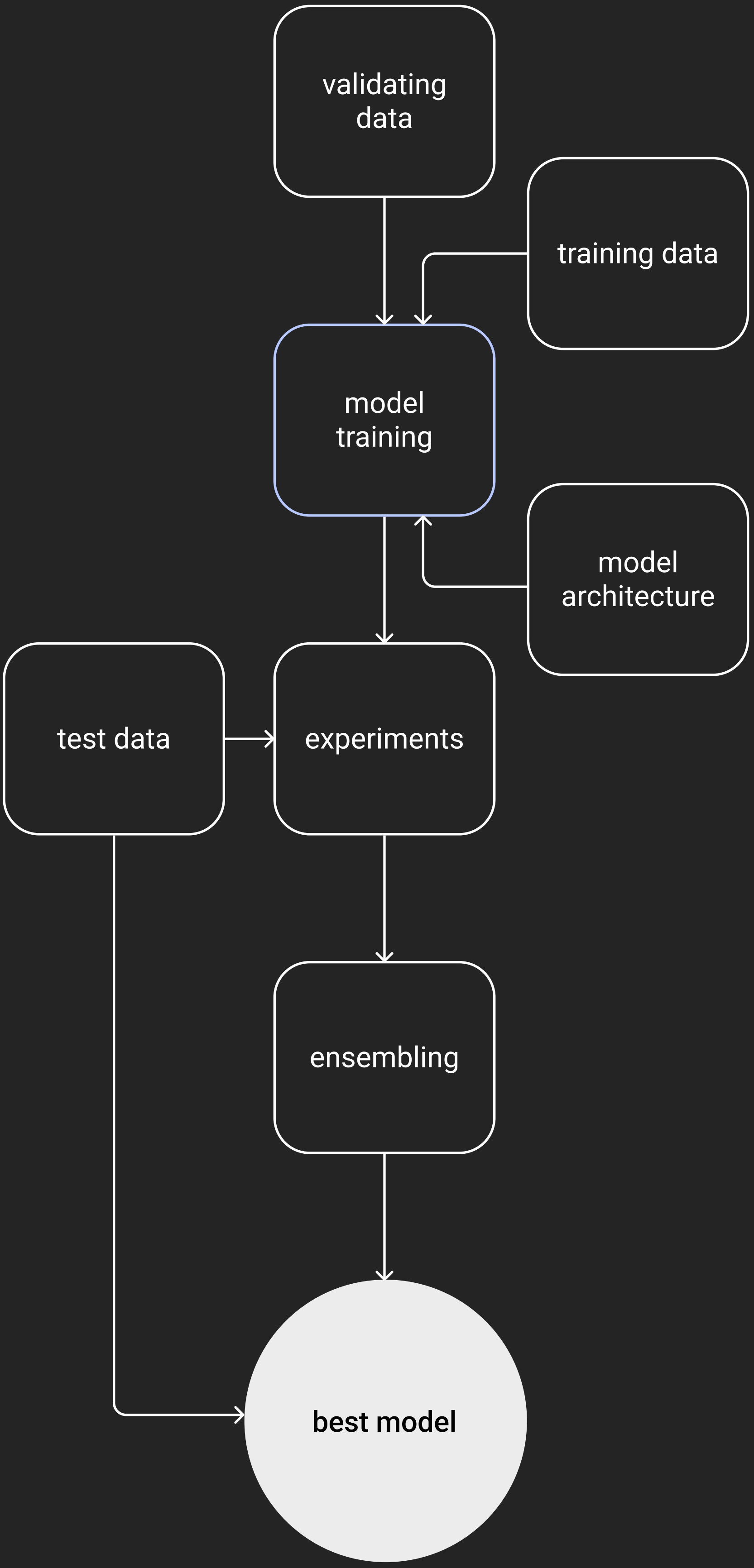
LABEL
YOUR
DATA

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?

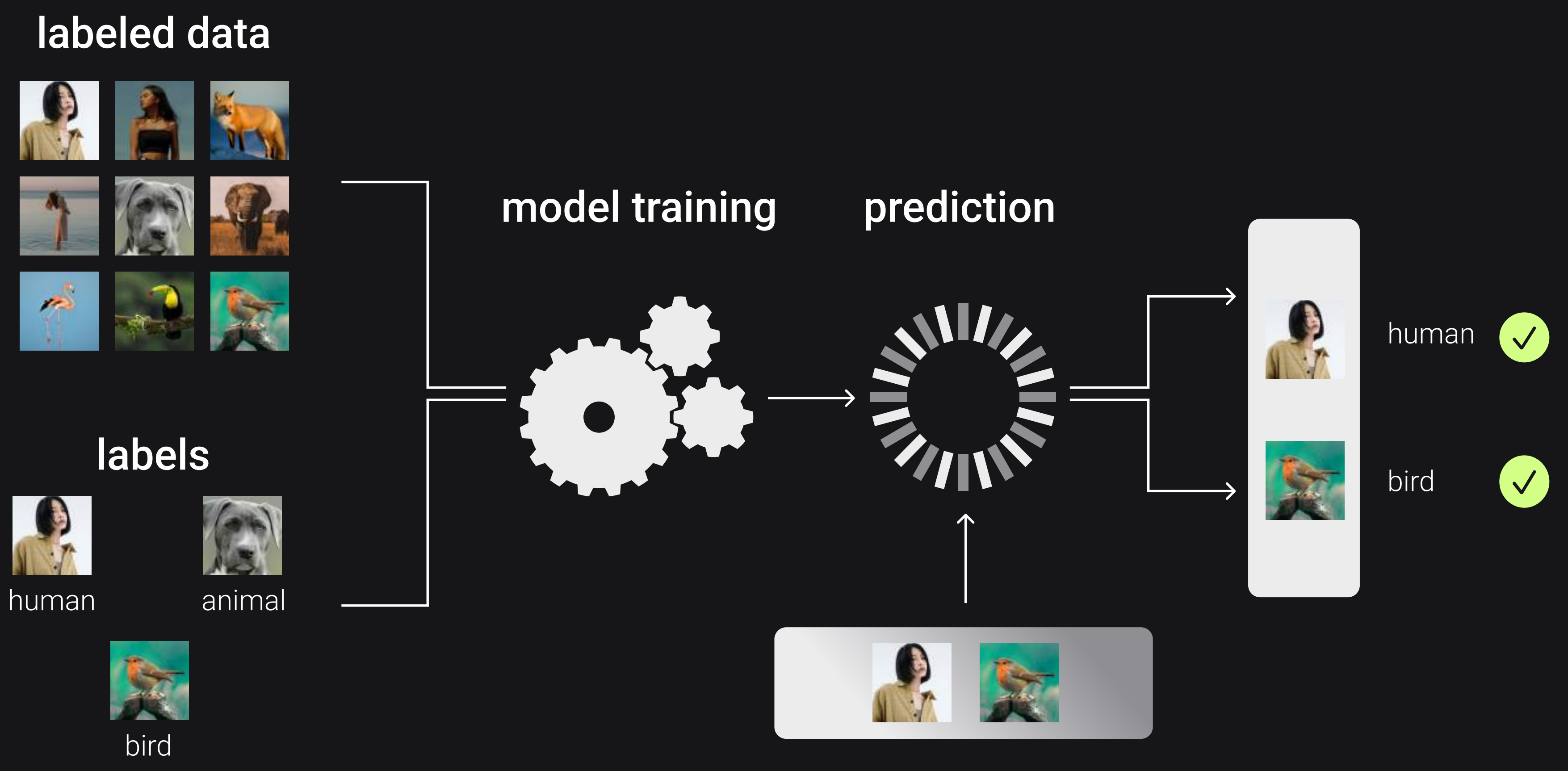
Most ML models use supervised learning, where an algorithm maps inputs to outputs based on a set of labeled data by humans. The model learns from these labeled examples to decipher patterns in that data during a process called model training. The model can then make predictions on new data.

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, 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.

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

Data labeling pipeline in ML



LABEL YOUR DATA