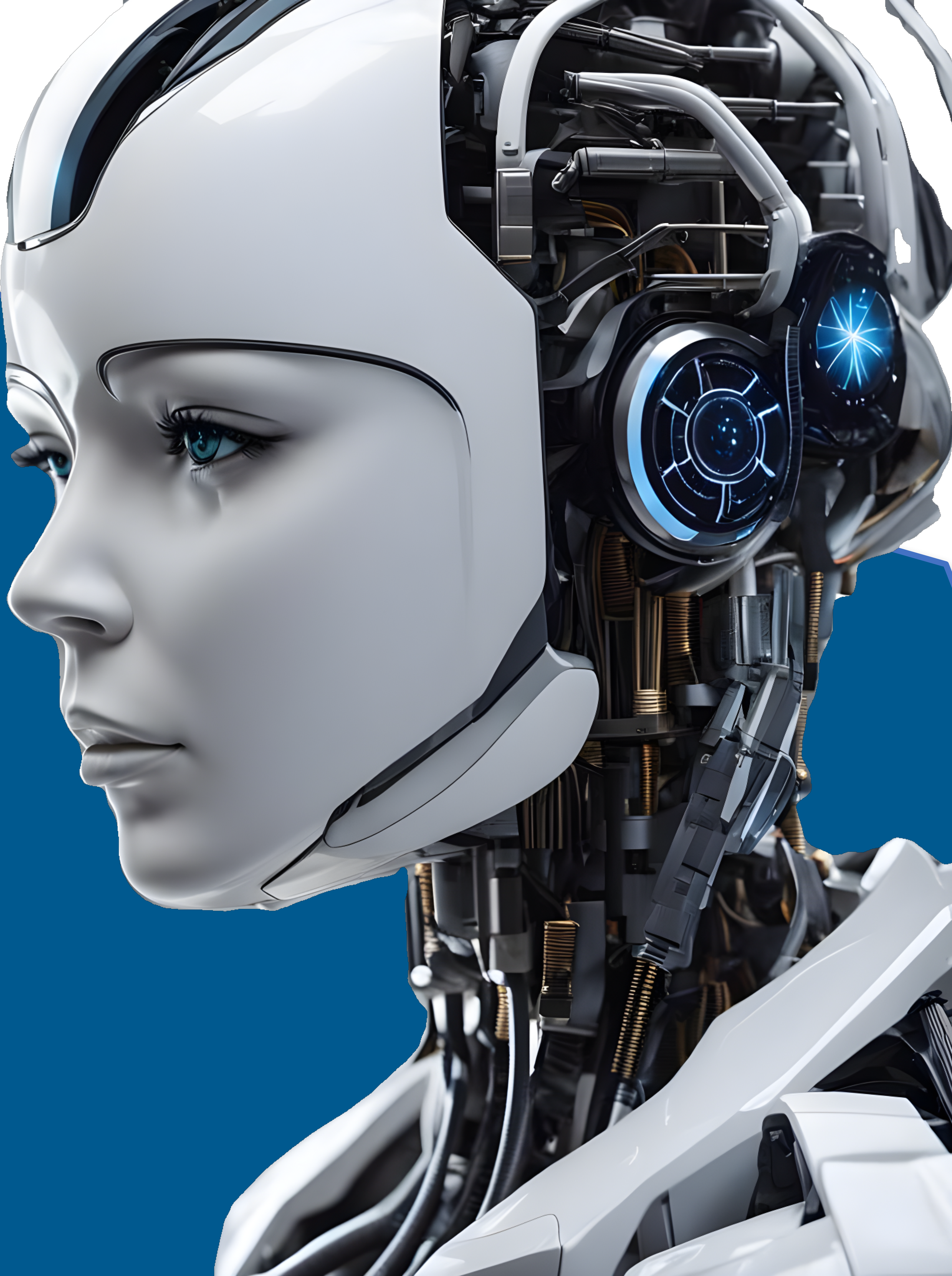
Tishk International University
IT Department
Course Code: IT-344/A

Introduction to Machine Learning

The Fundamentals of Data in ML



Lecture 3

Spring 2024
Hemin Ibrahim, PhD
hemin.ibrahim@tiu.edu.iq

Outline

- What is Data?
- Why data in ML?
- Types of data
- Data representation
- Using Data in Machine Learning
- Feature Extraction

Objectives

- Understand the fundamental concept of data
- Distinguish between the core categories of structured and unstructured data.
- Recognize the importance of choosing suitable data representations for effective machine learning algorithms.
- Understand the idea that machine learning algorithms fundamentally rely on data patterns to learn
- Understand the crucial step of splitting data into training, validation, and testing sets to ensure accurate model evaluation
- Understand that Feature extraction extracts essential data points for better machine learning.

Data

The concept of machine learning as algorithms "learning" from “data” rather than being explicitly programmed for everything.

What is Data?

“data”

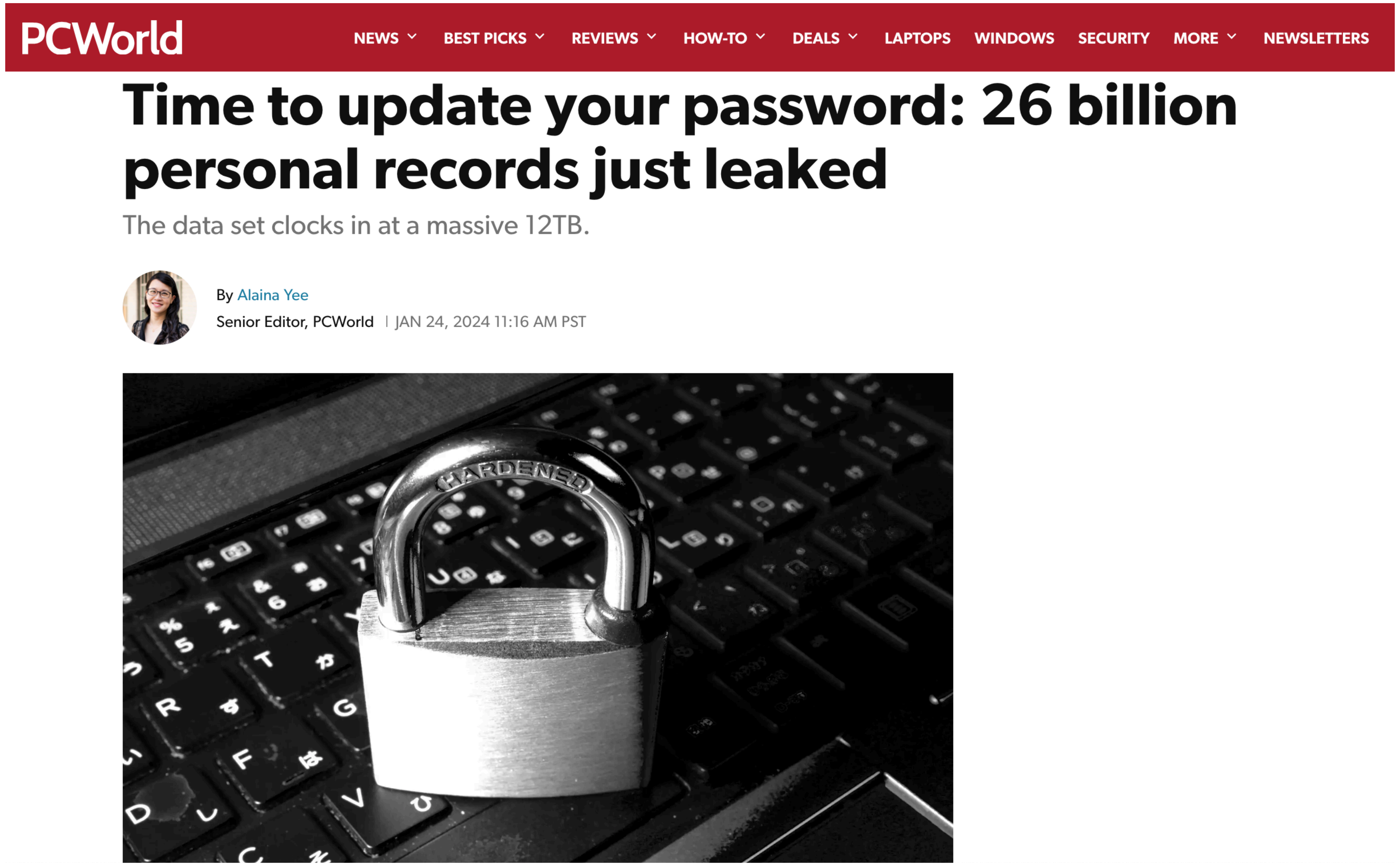
Data is a collection of raw facts, figures, measurements, or observations. In machine learning, it's the information used to train the algorithms.

- Numerical values (e.g., house prices, test scores)
- Text (e.g., product reviews, tweets)
- Images (e.g., medical scans, photographs)
- Audio (e.g., speech recordings, music)
- Video (eg., movie, a tutorial, a recorded live event)

Data

"data"

is **power, control, money, dominance, ..**



Time to update your password: 26 billion personal records just leaked

The data set clocks in at a massive 12TB.

By Alaina Yee
Senior Editor, PCWorld | JAN 24, 2024 11:16 AM PST

Why Data in ML?

- Data is the foundation of machine learning algorithms.
- **Learning from patterns:** Machine learning methods can find complex patterns and correlations in data that may be difficult for humans.
- High-quality, relevant data leads to accurate models.
- **Making predictions:** Based on learned patterns, models predict outcomes **unseen data** points (e.g., weather, future sales).
- **Driving decisions:**Data guides decisions, making industries more efficient and effective.

Types of data

- **Structured Data:** highly organized and follows a specific format.
  - Common in business applications, financial records, health data, and relational databases.
- **Unstructured Data:** lacks a predefined data model or structure.
  - Text documents (e.g., emails, articles, social media posts).
  - Multimedia files (e.g., images, audio, video).
- **Semi-Structured Data:** it has some organizational properties but doesn't conform to a strict structure like structured data.
  - XML, API, and JSON

Data Collection and Preparation

Data collection and preparation are crucial steps in the data analysis and machine learning pipeline.

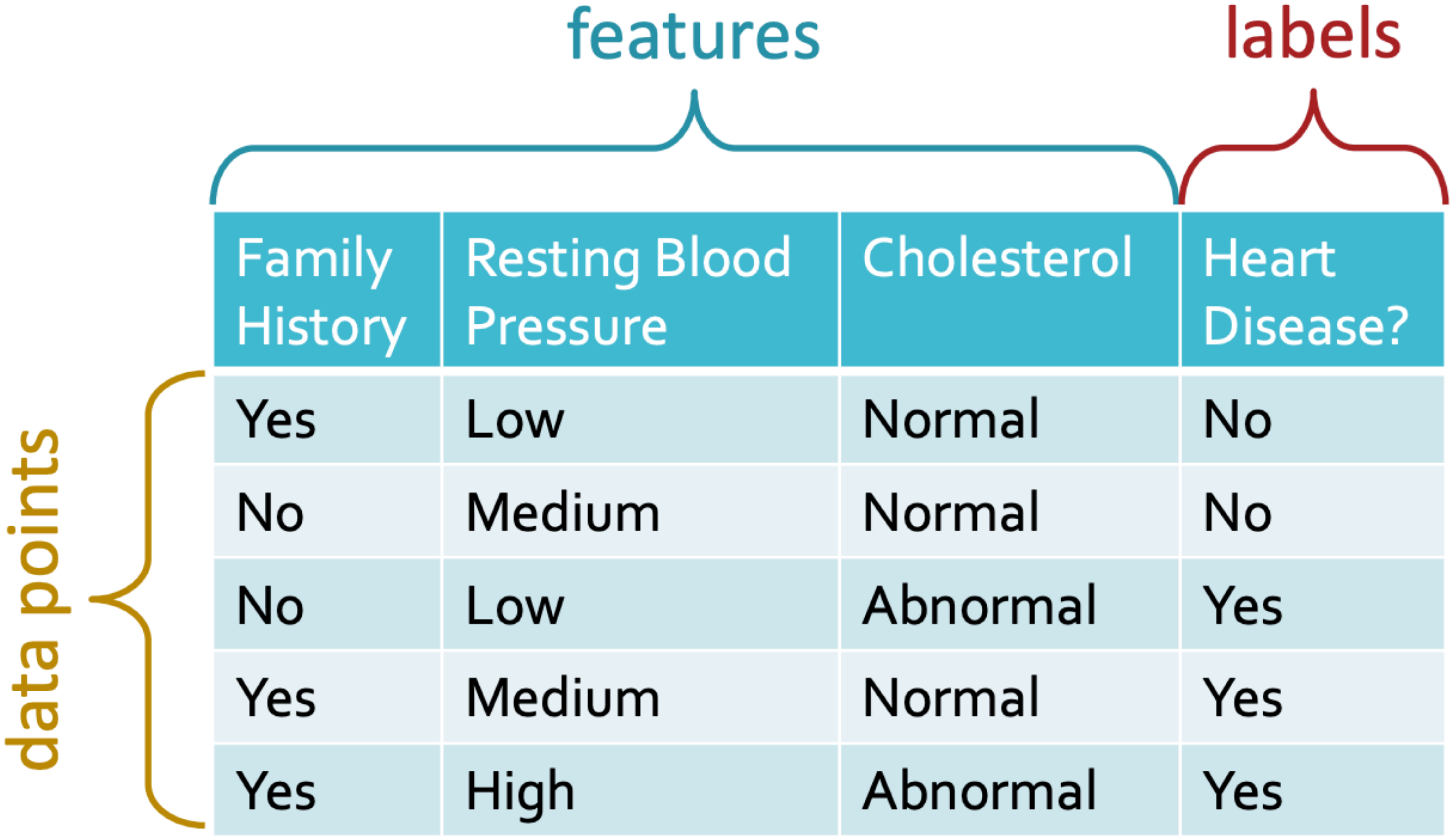
- **Collecting data from various sources:** sensors, databases, APIs, web scraping, surveys, logs.. Each source may have its own format and structure.
- **Data preprocessing**: cleaning, normalization, handling missing values, and feature engineering (generate new features)
- **Why it is a challenge?**
  - availability
  - quality
  - relevance
  - privacy

Data Representation

- Tabular data: Rows and columns
- Text data: Bag-of-words, TF-IDF.
- Image data: Pixel values.
- Sound data: Waveforms, spectrograms.

Data Representation

Tabular data: Rows and columns



features

labels

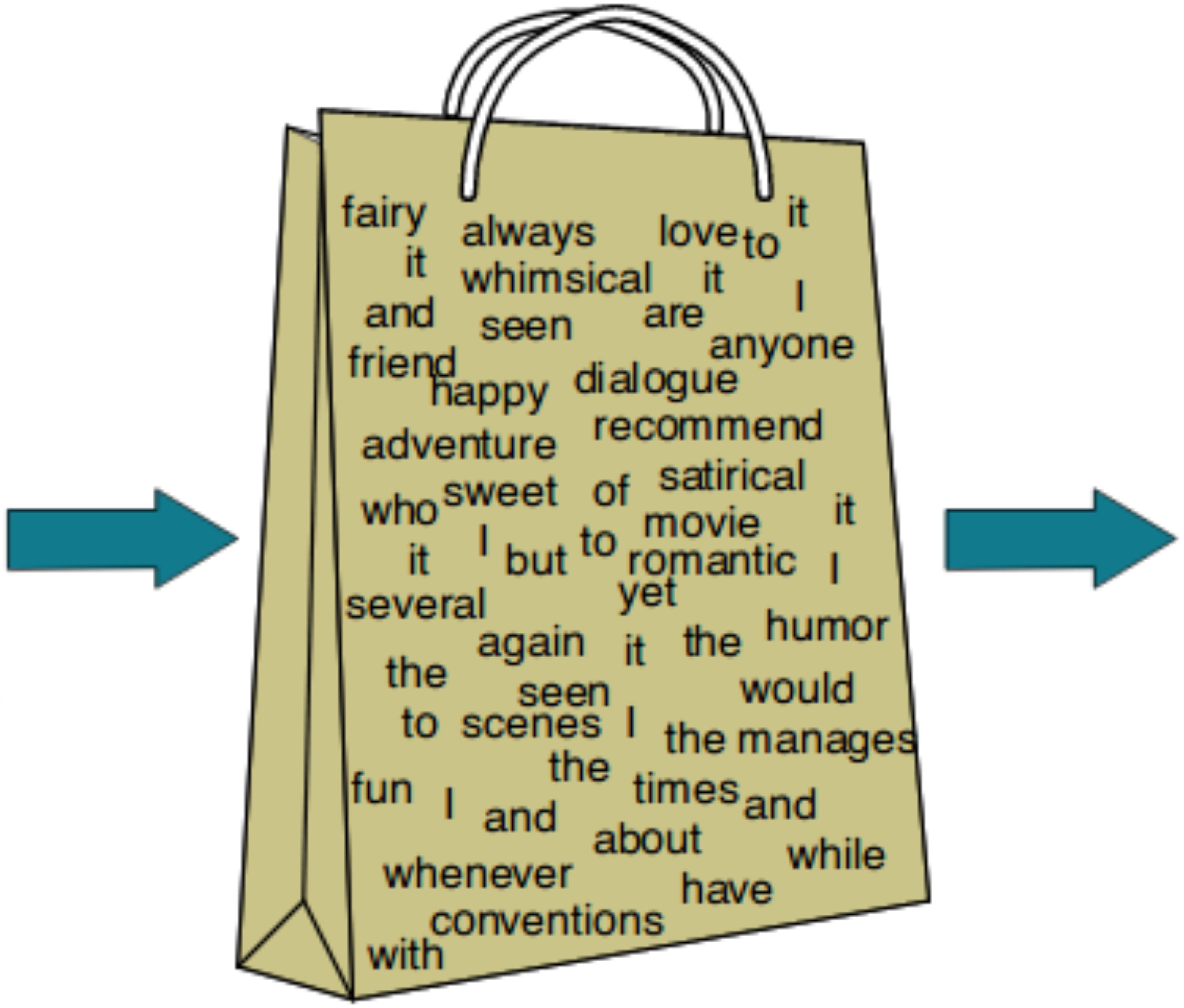
data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Data Representation

Text data: Bag-of-words, TF-IDF.

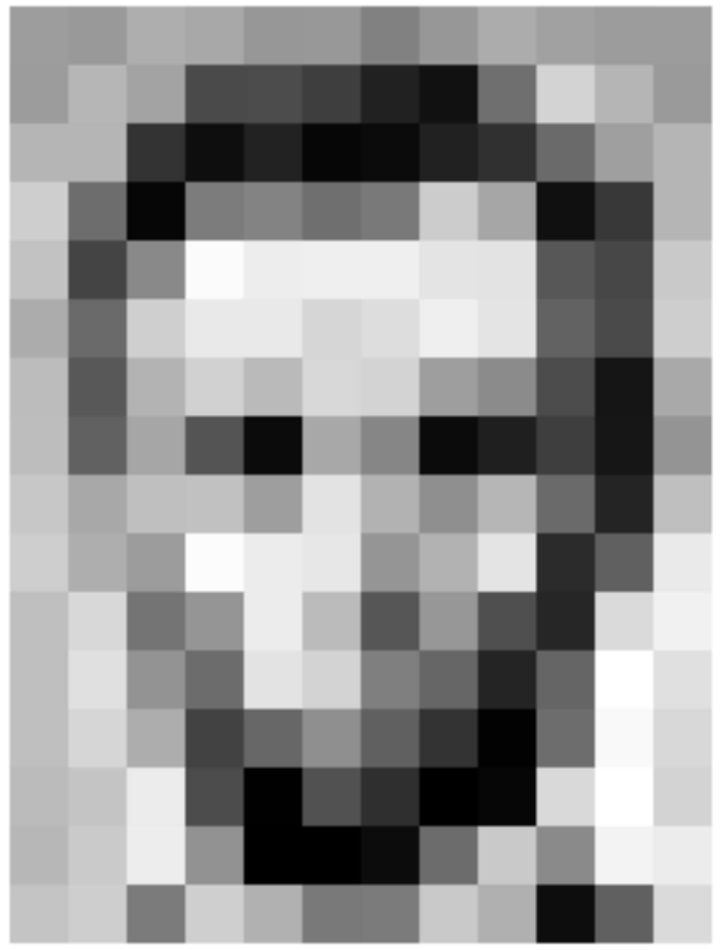
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

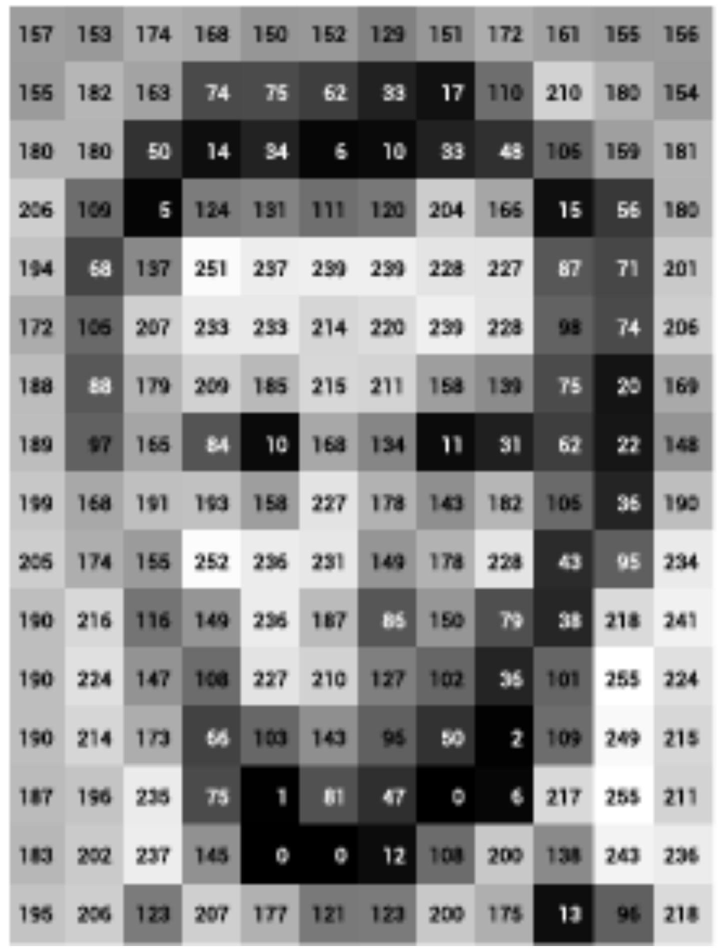


it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

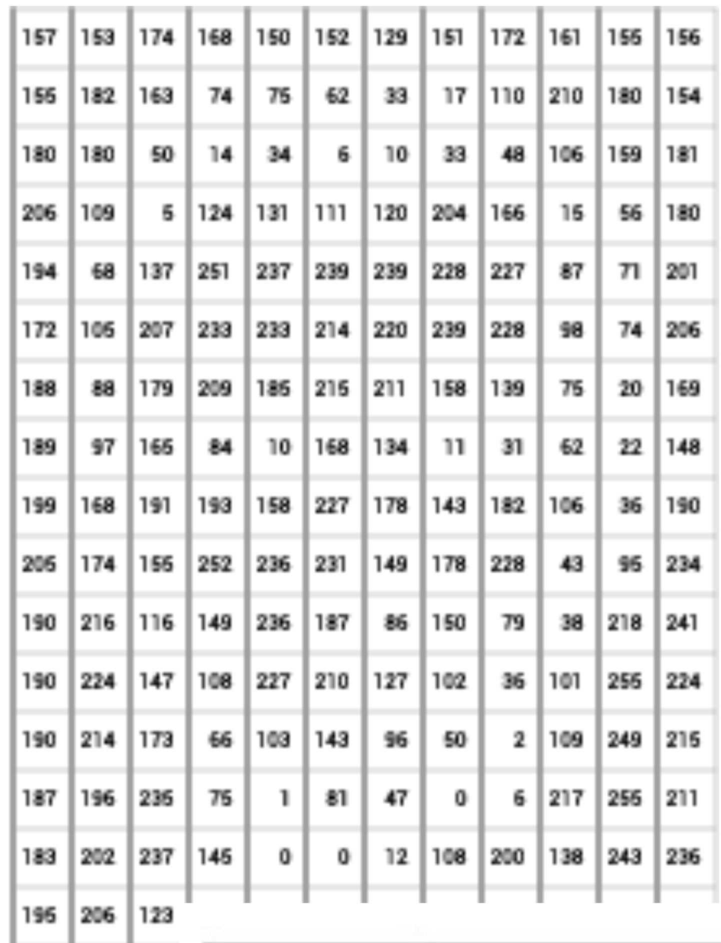
Data Representation

Image data: Pixel values.

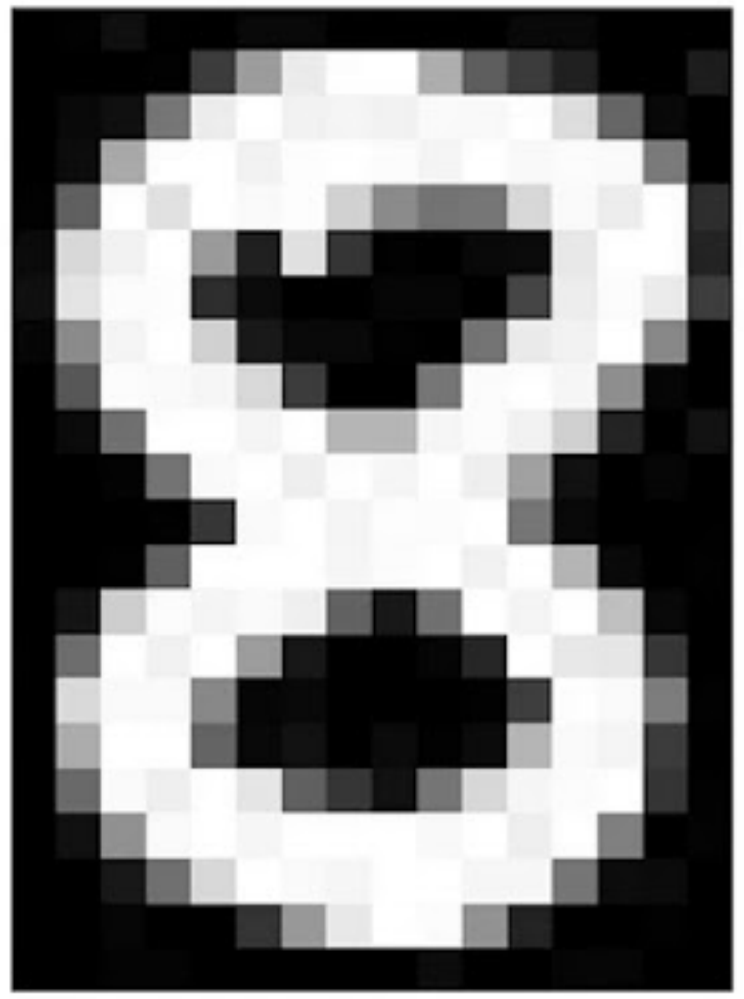


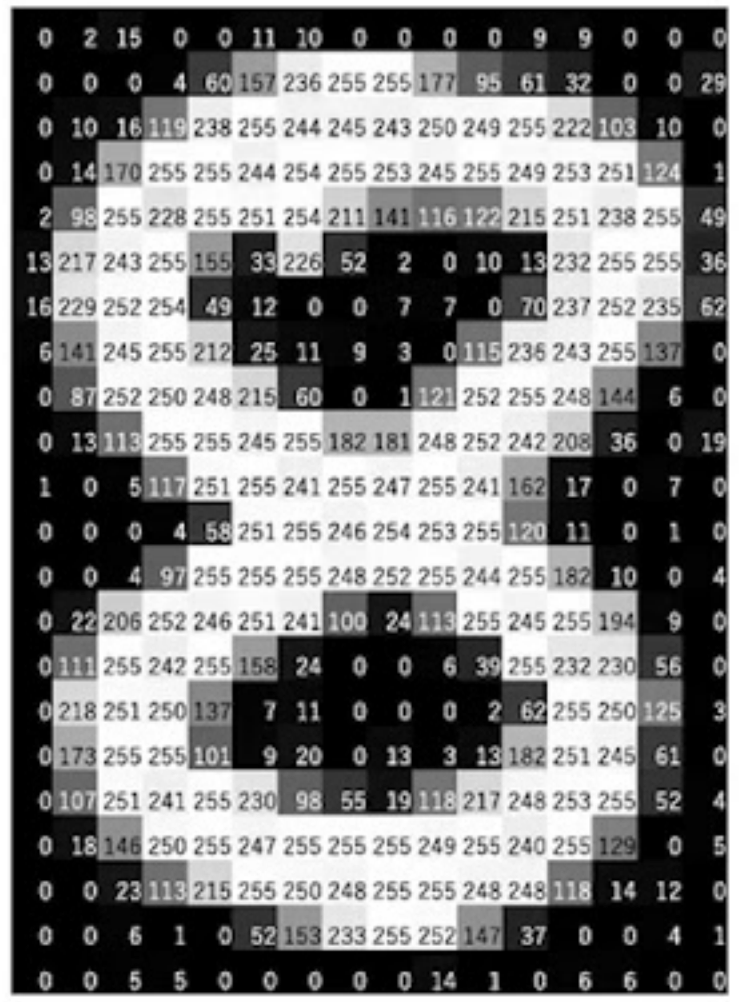


157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

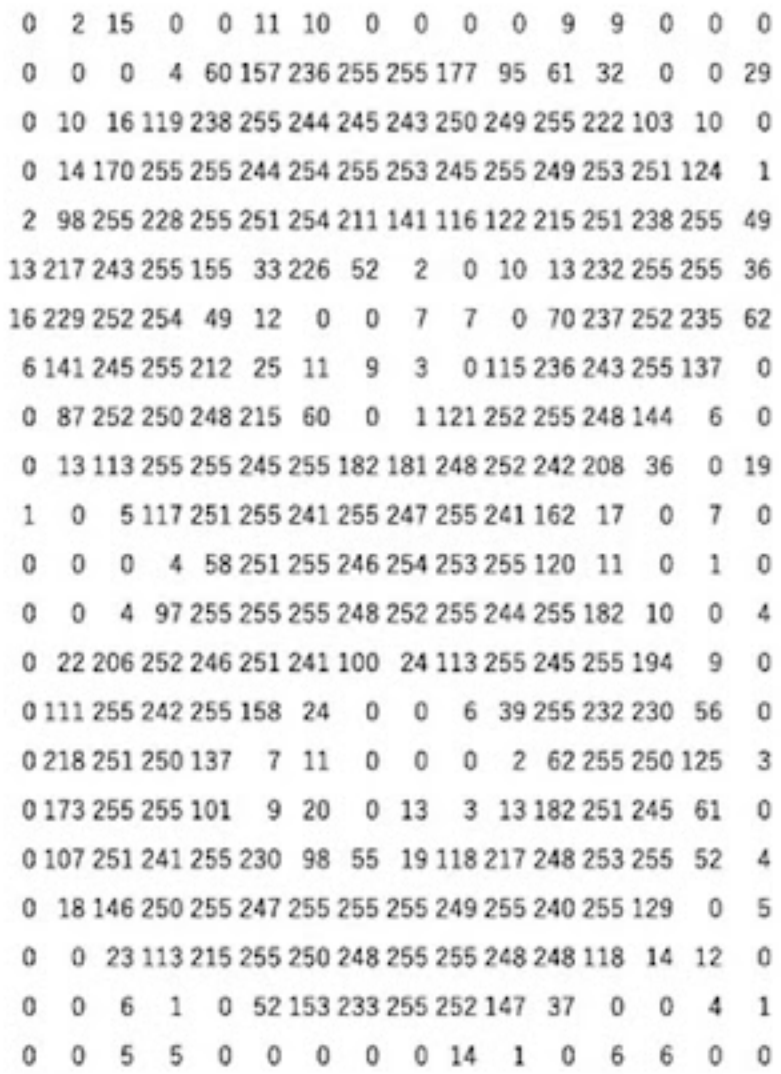


157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123									





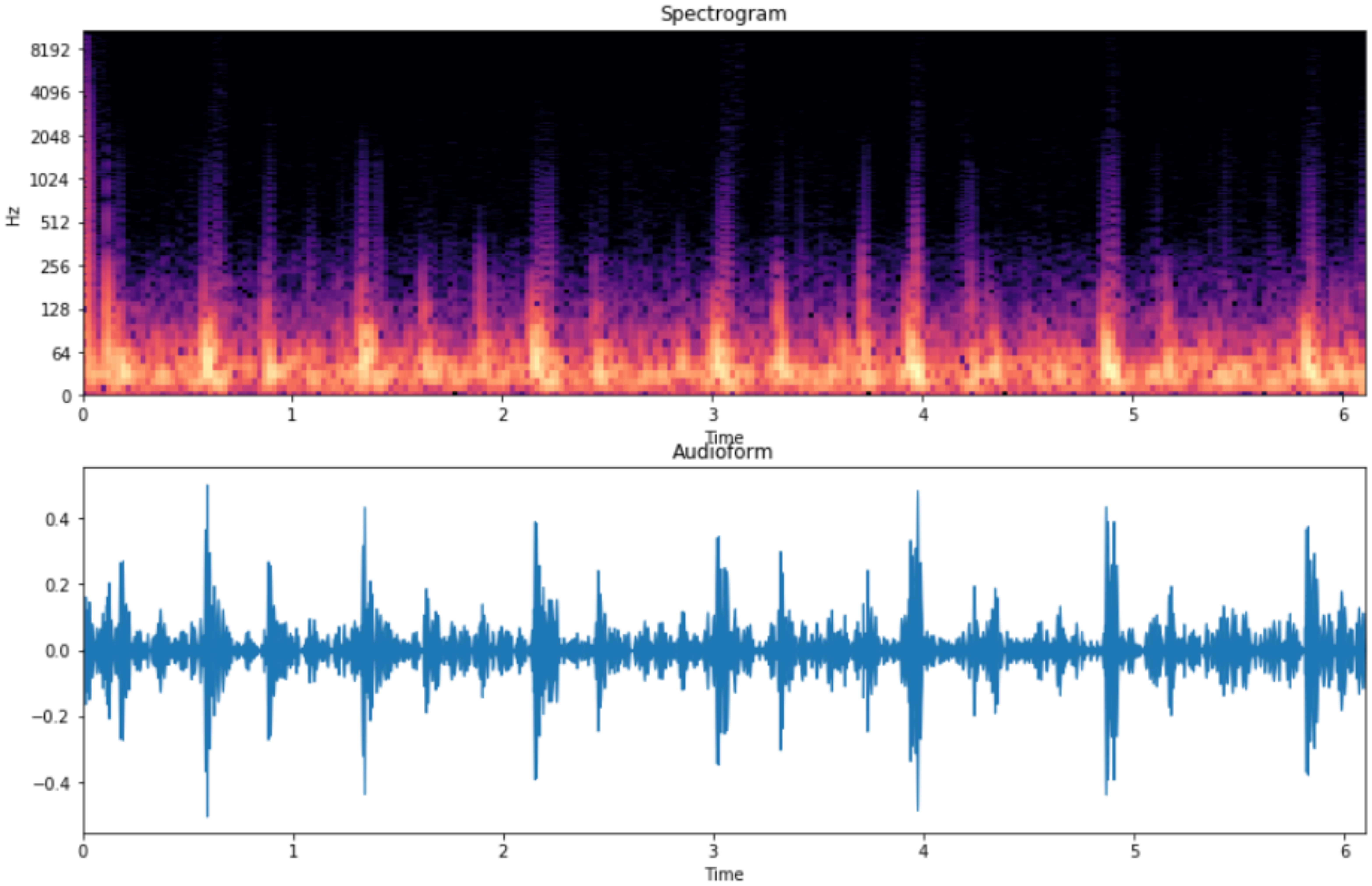
0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0



0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

Data Representation

Sound data: Waveforms, spectrograms.



Using Data in Machine Learning

- Splitting data into training, validation, and testing sets.
- Cross-validation for model evaluation.
- Handling imbalanced datasets.

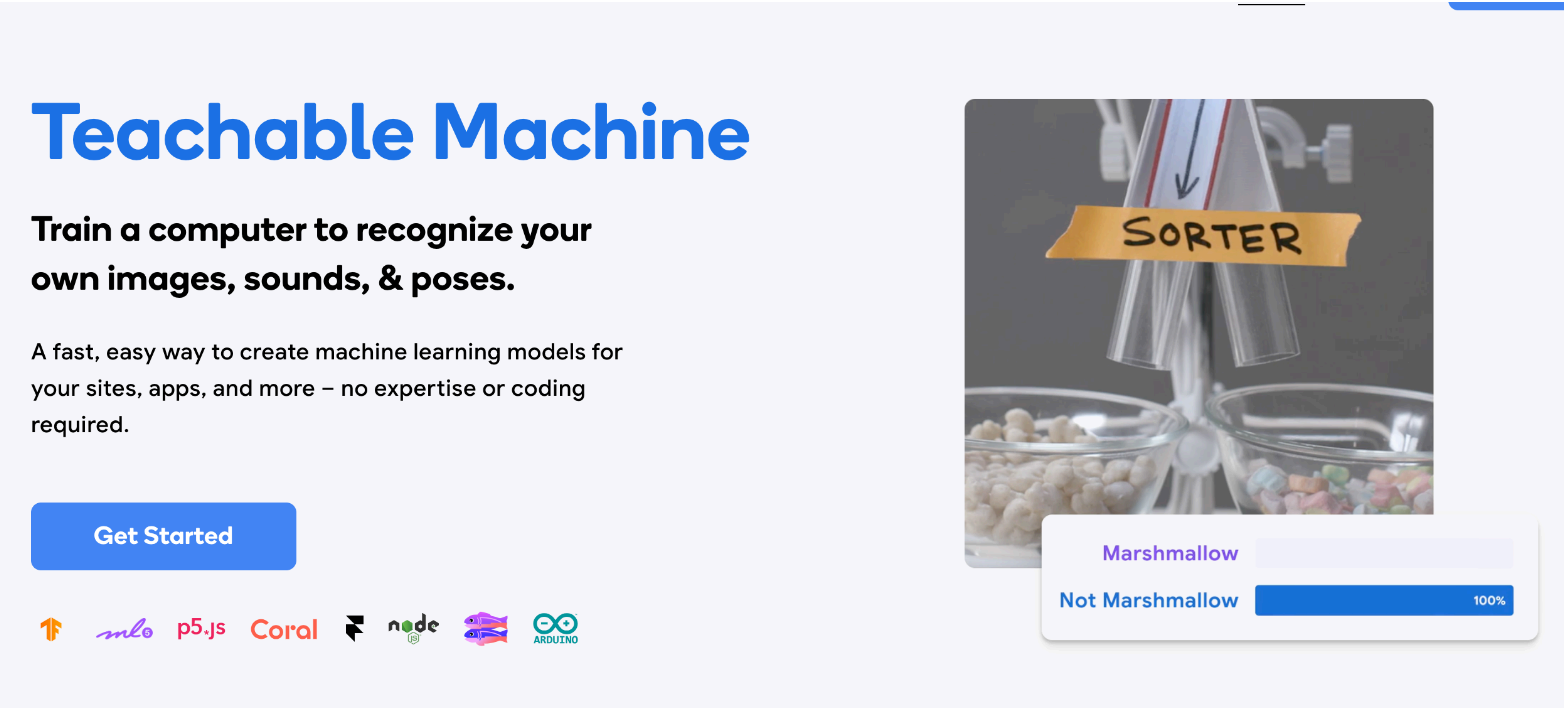
Training, Validation and Testing Data

- **Training data**: Used to train machine learning models.
- **Validation data**: Used to fine-tune model parameters.
- **Testing data**: Used to evaluate model performance.

Test
Train
Validation

Training, Validation and Testing Data

https://teachablemachine.withgoogle.com/



Teachable Machine

Train a computer to recognize your own images, sounds, & poses.

A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.

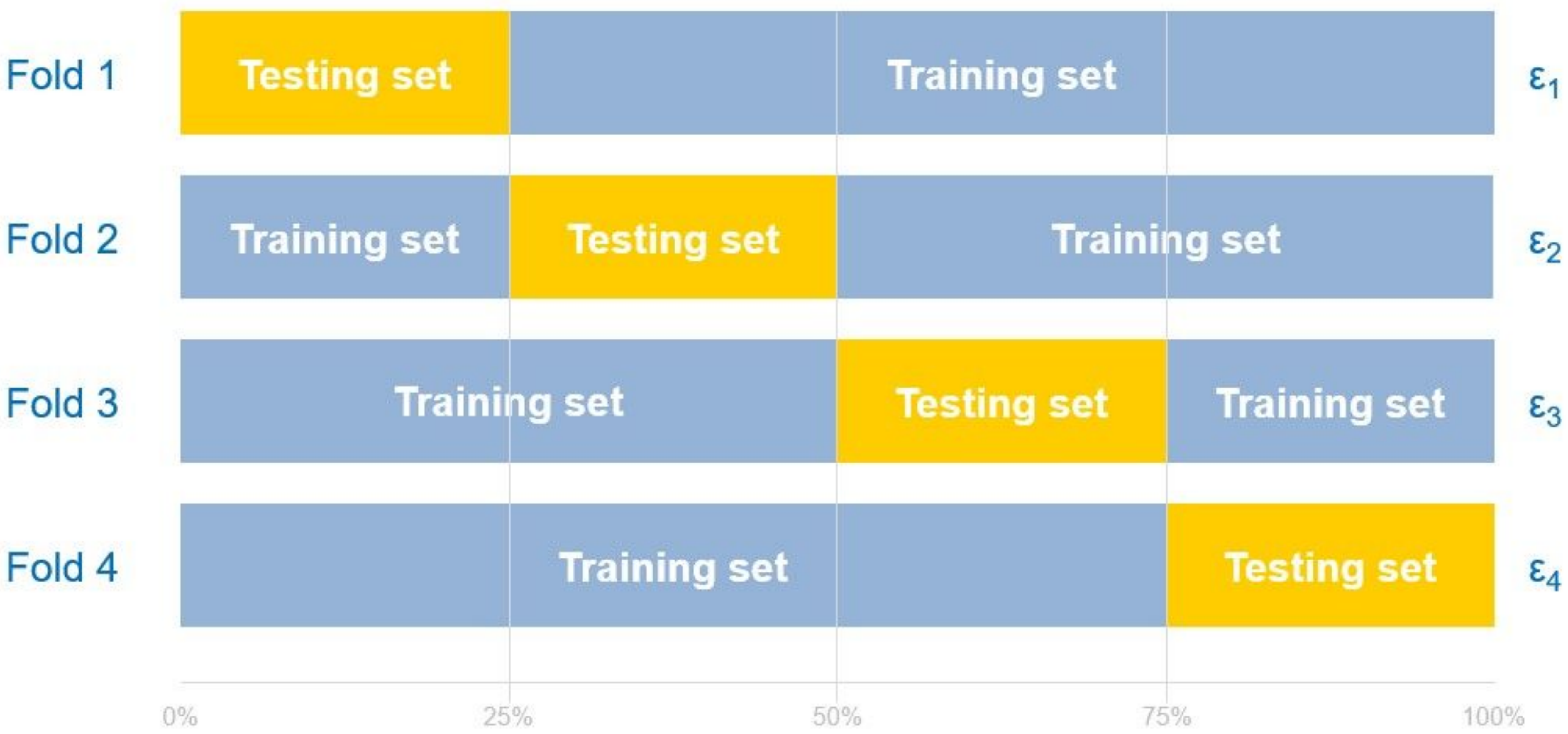
Get Started

Cross-validation

Cross-validation is a method to check how well a model works by repeatedly testing it on different parts of the data.

K-Fold Cross-Validation: The dataset is divided into k subsets/folds of approximately equal size.

4-fold validation (k=4)



Fold 1 Testing set Training set ε_1

Fold 2 Training set Testing set Training set ε_2

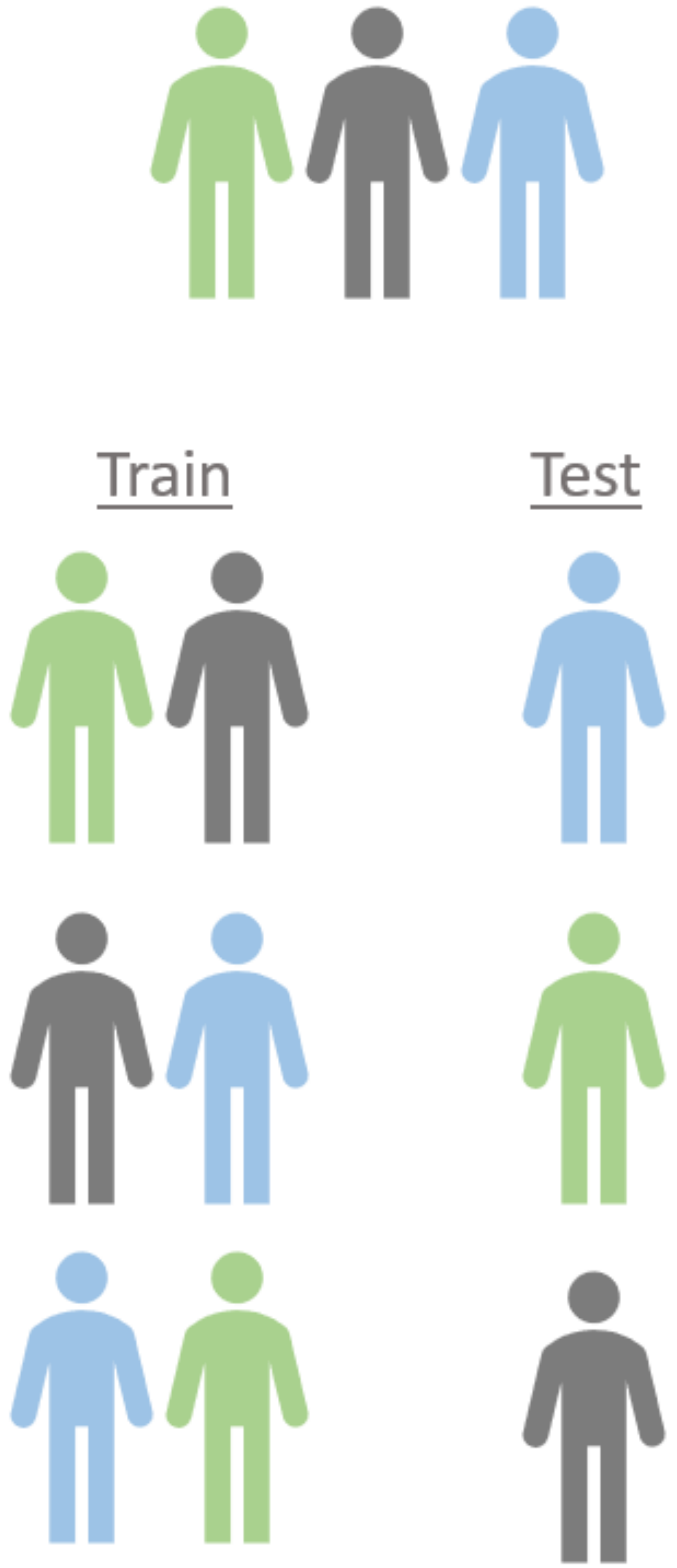
Fold 3 Training set Testing set Training set ε_3

Fold 4 Training set Testing set ε_4

0% 25% 50% 75% 100%

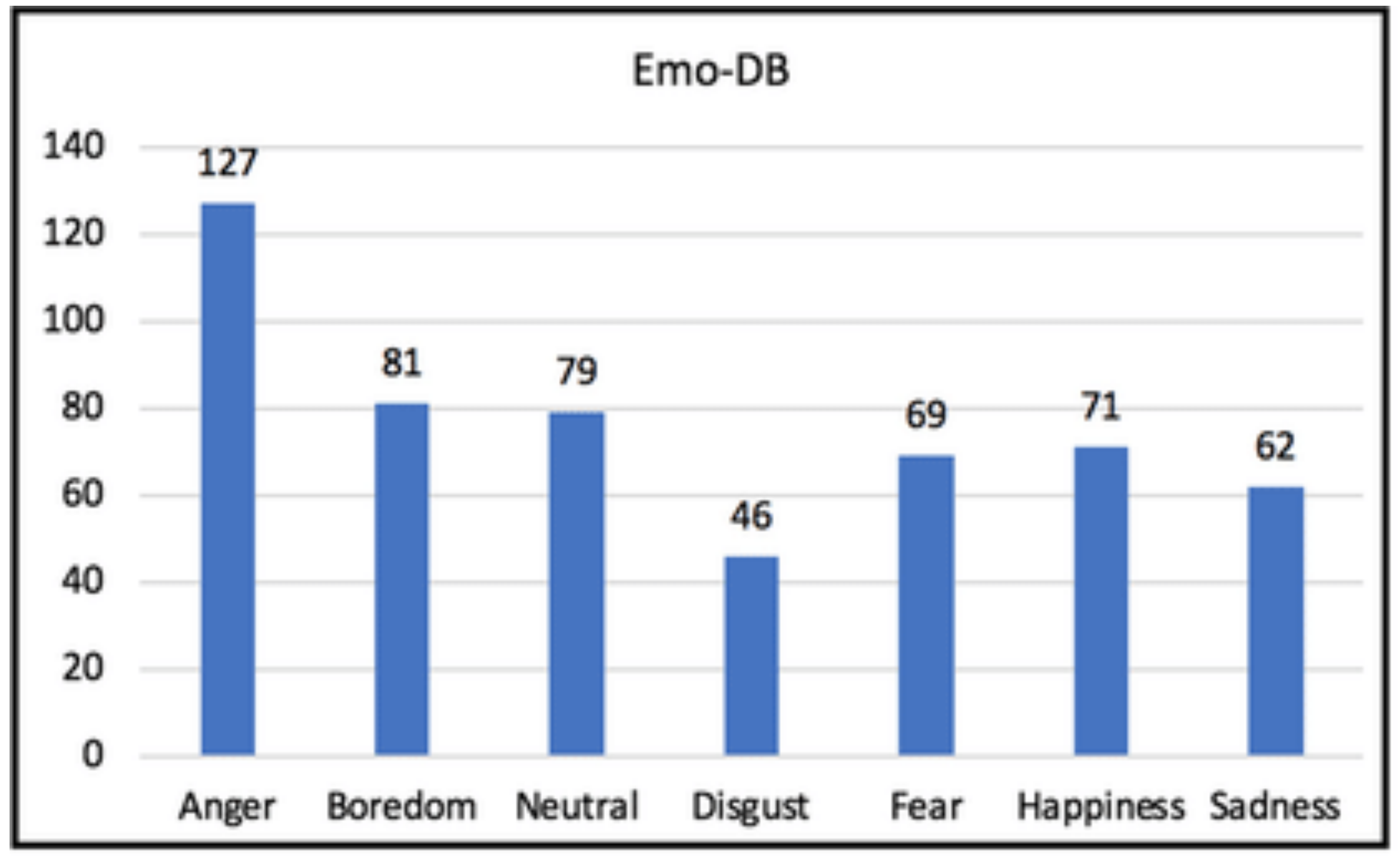
Cross-validation

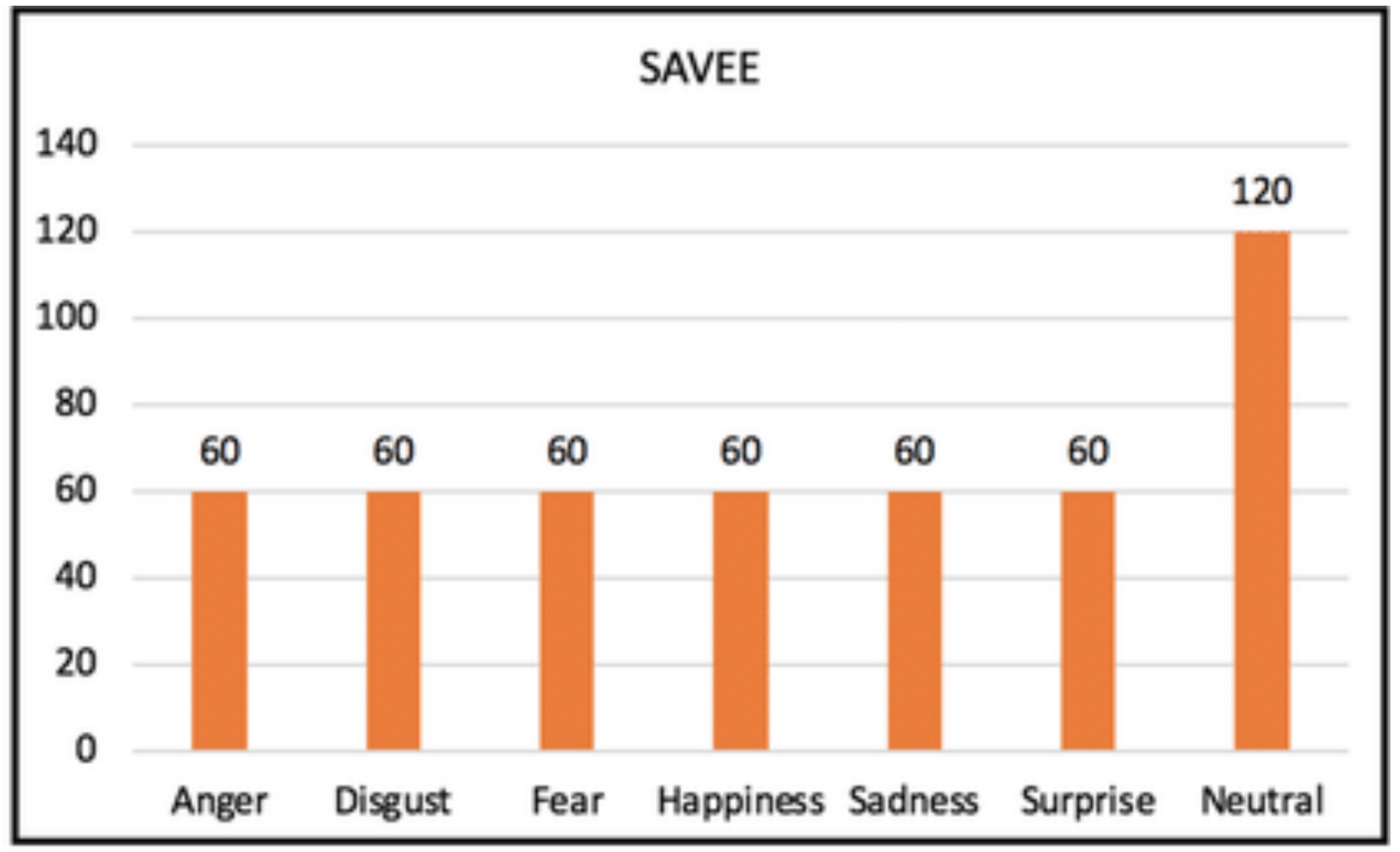
Leave-One-Out: The model is trained on all but one data point and tested on the one left out. This process is repeated for each data point in the dataset.

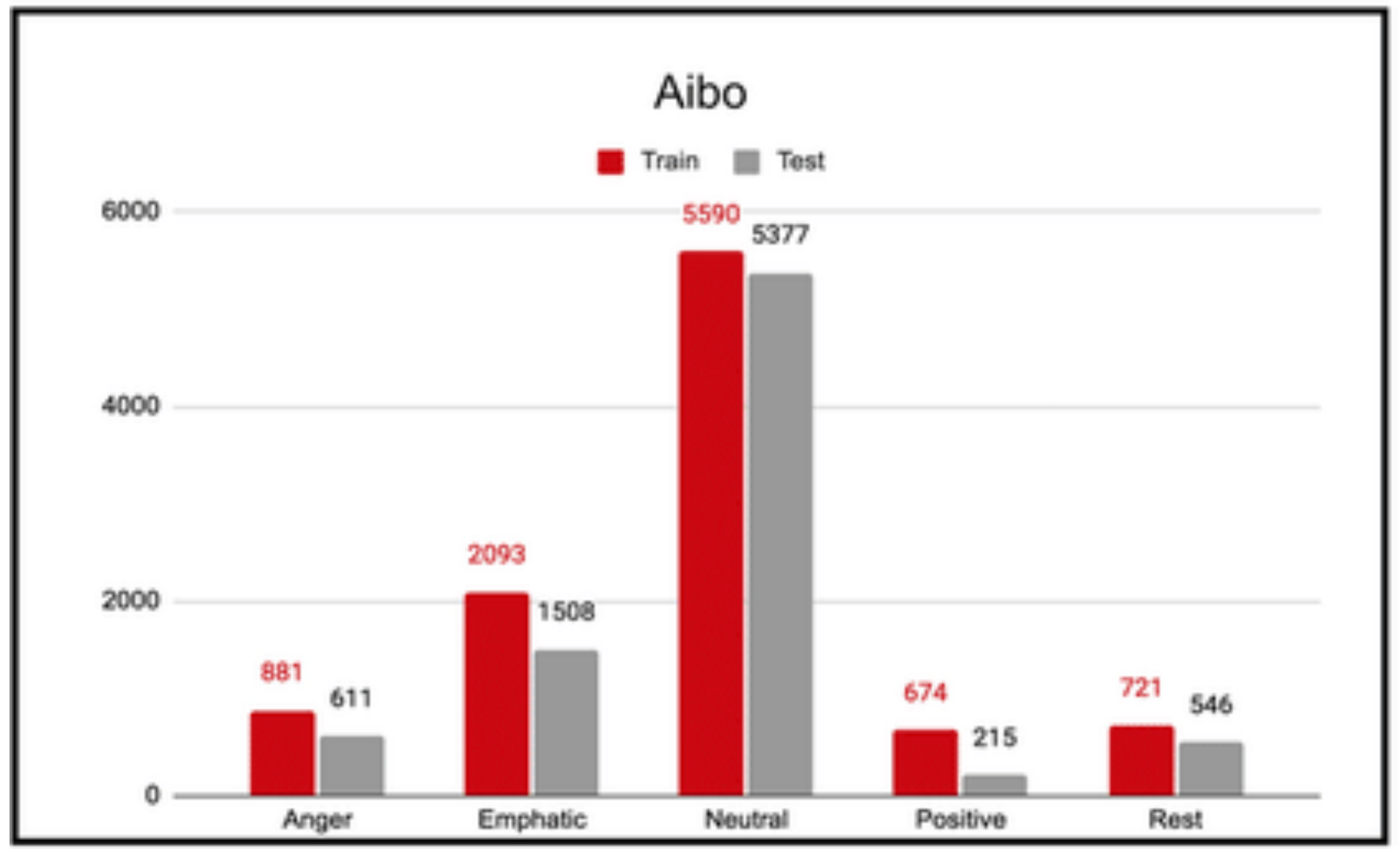


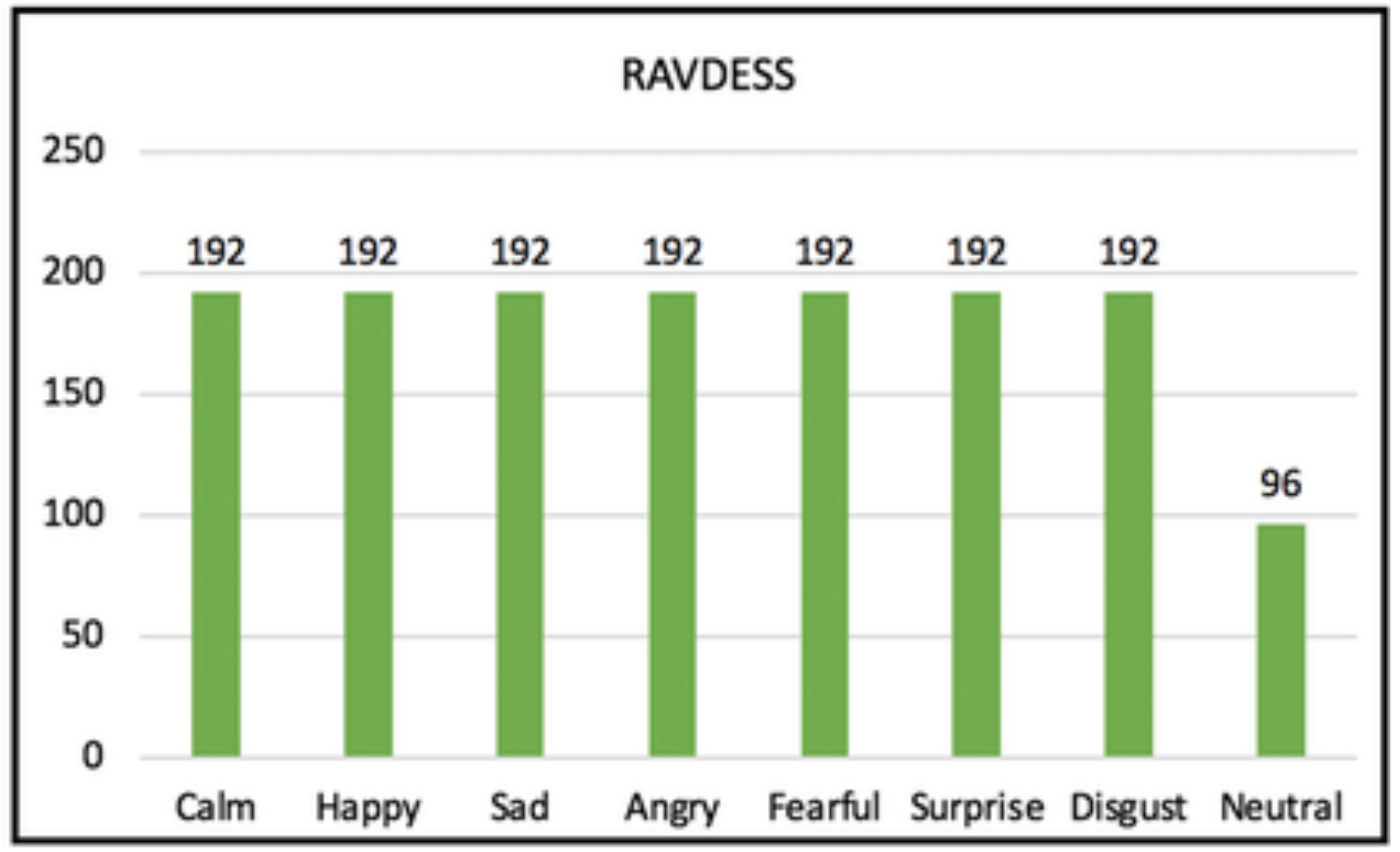
Imbalanced Datasets

Imbalanced datasets refer to datasets where one class has significantly fewer samples compared to the other(s)





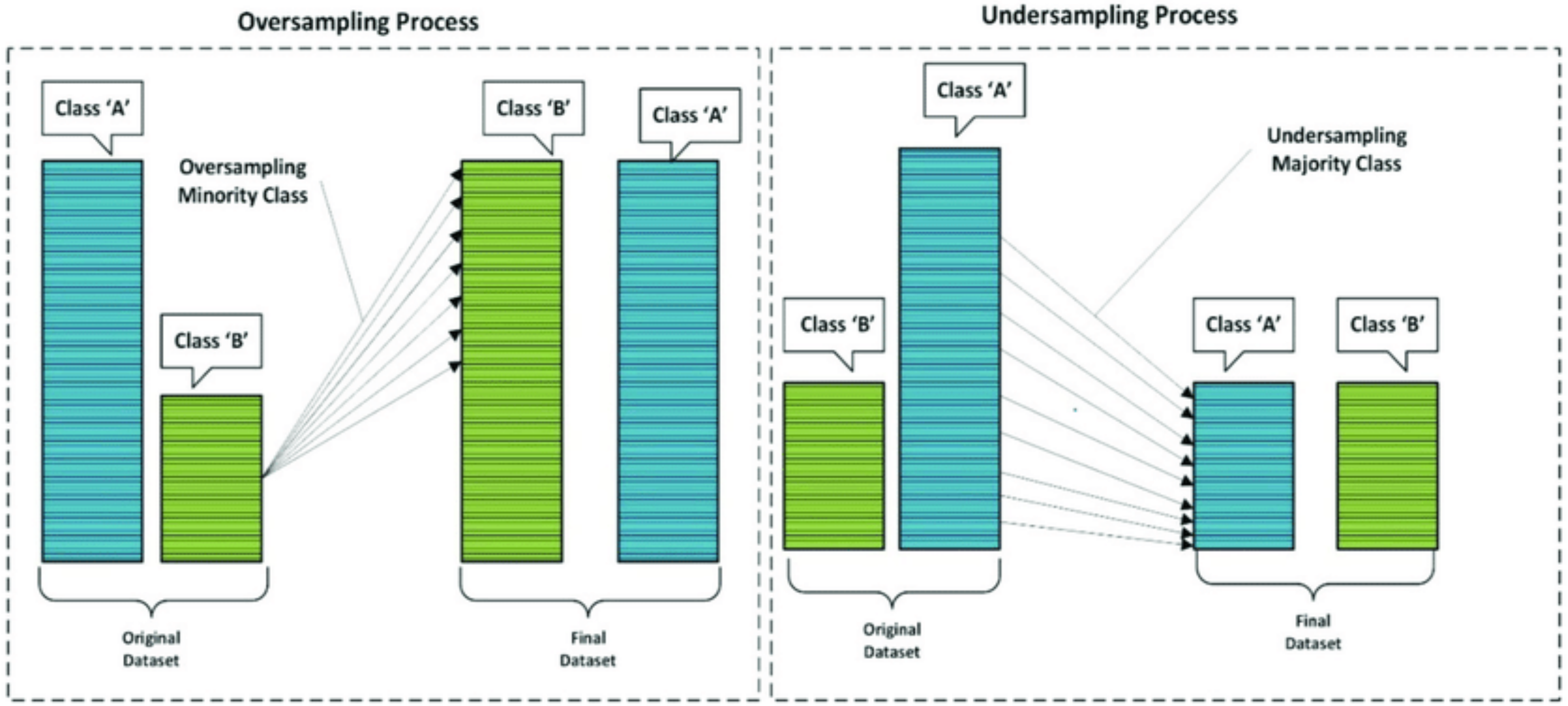




Handling Imbalanced Datasets

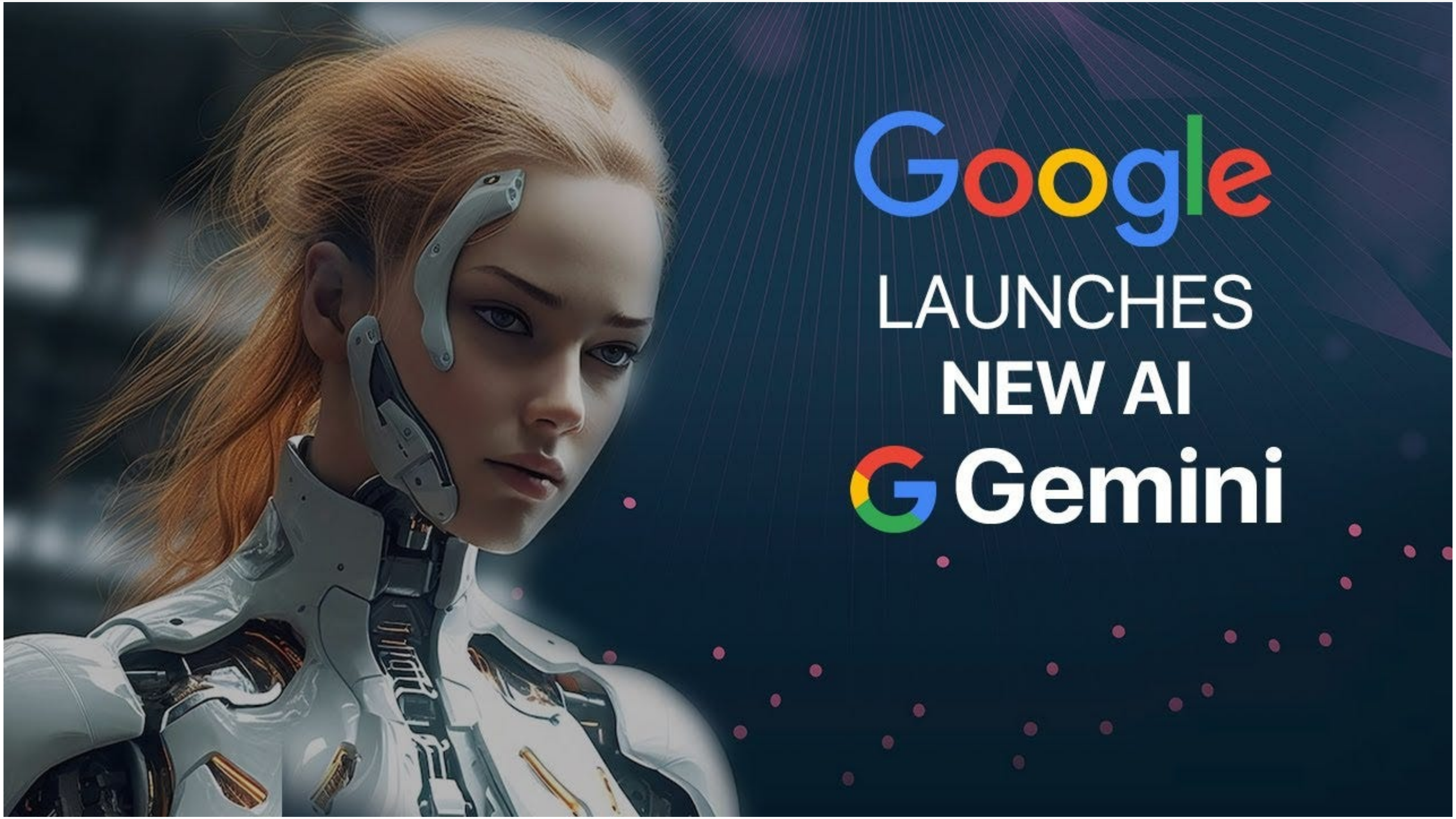
Resampling Methods:

- **Undersampling**: Randomly removing samples from the majority class to balance the class distribution.
- **Oversampling**: Duplicating samples from the minority class or generating synthetic samples to increase its representation.

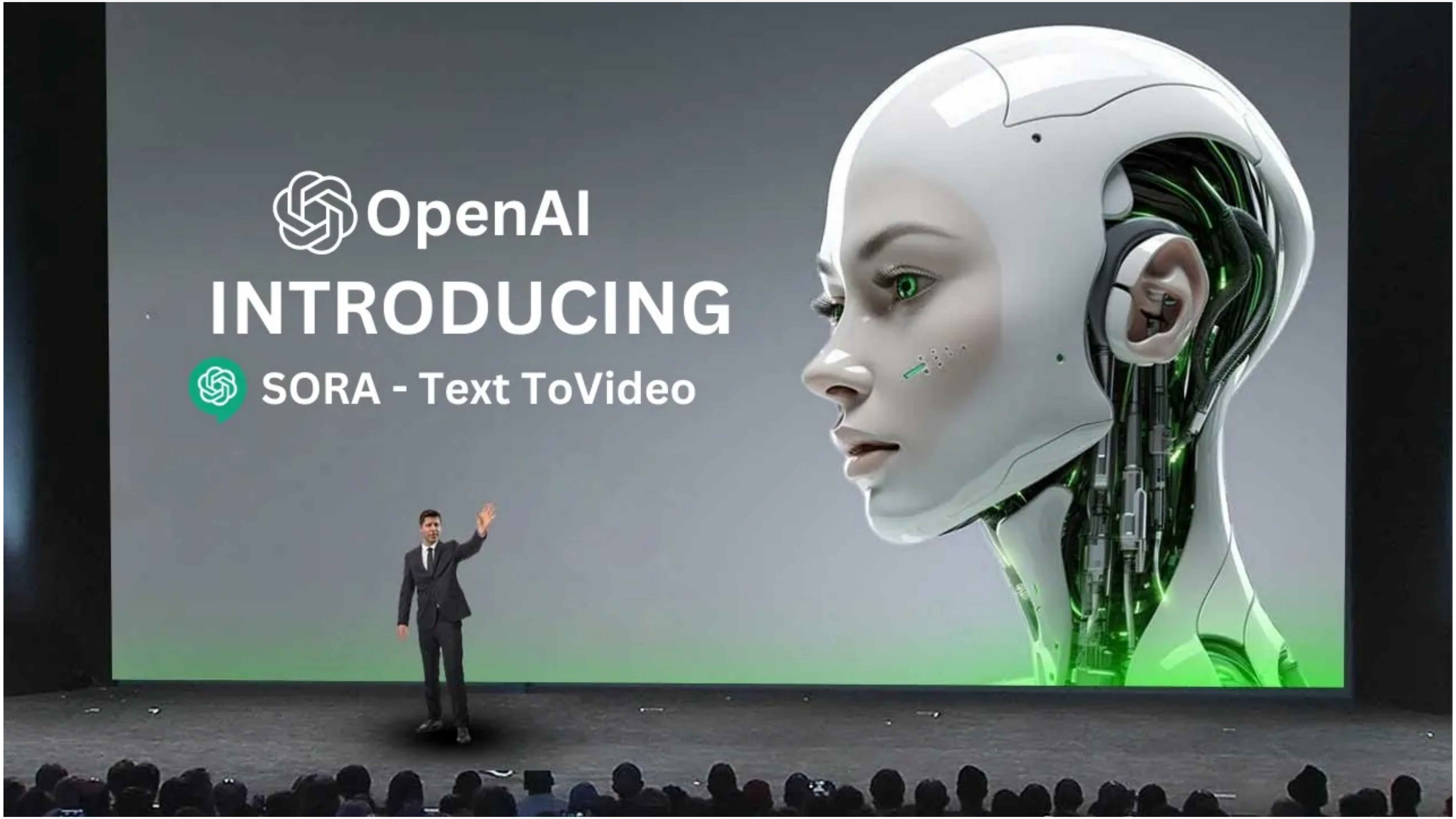


AI State of the art Applications

AI State of the art Applications

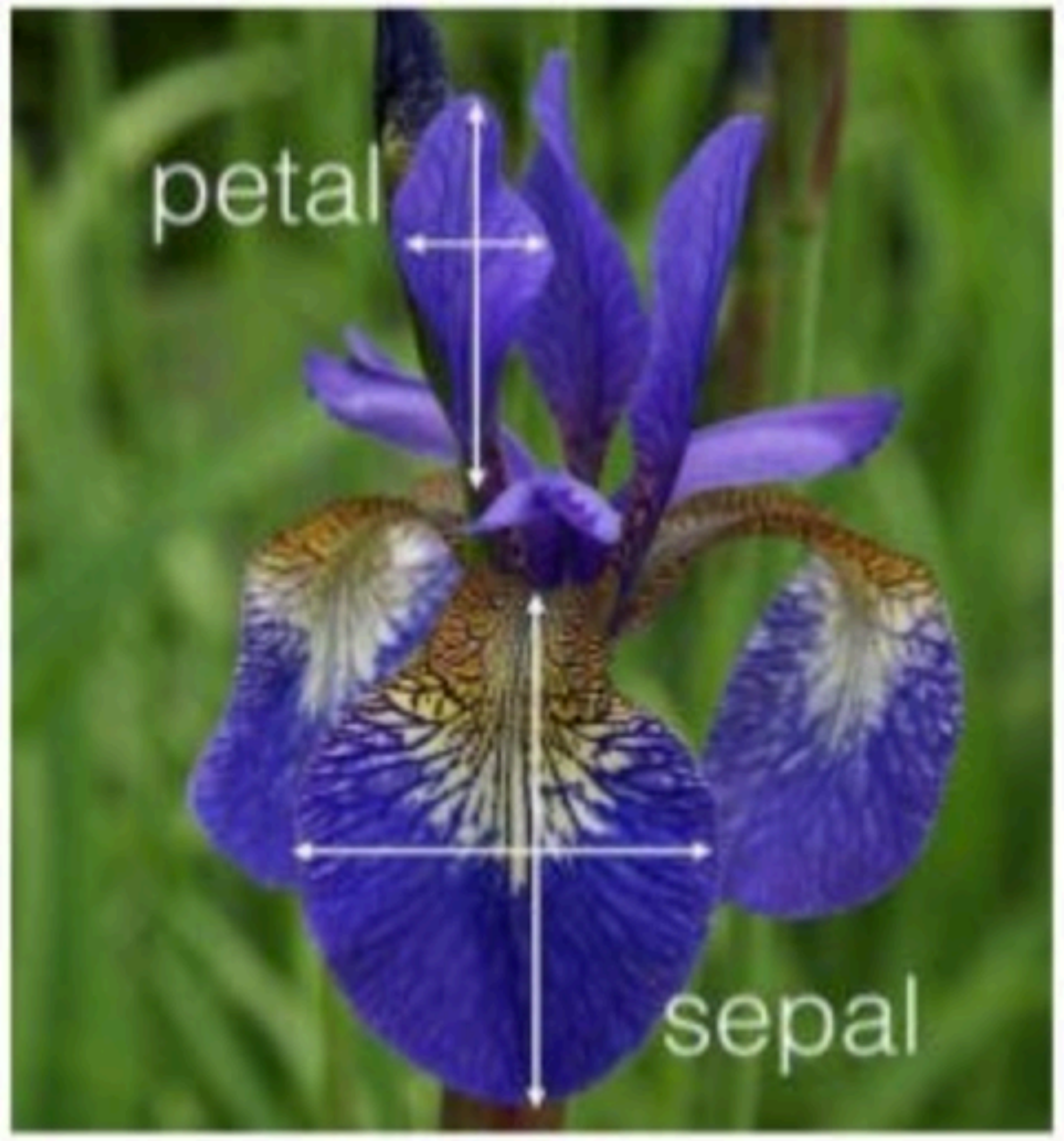


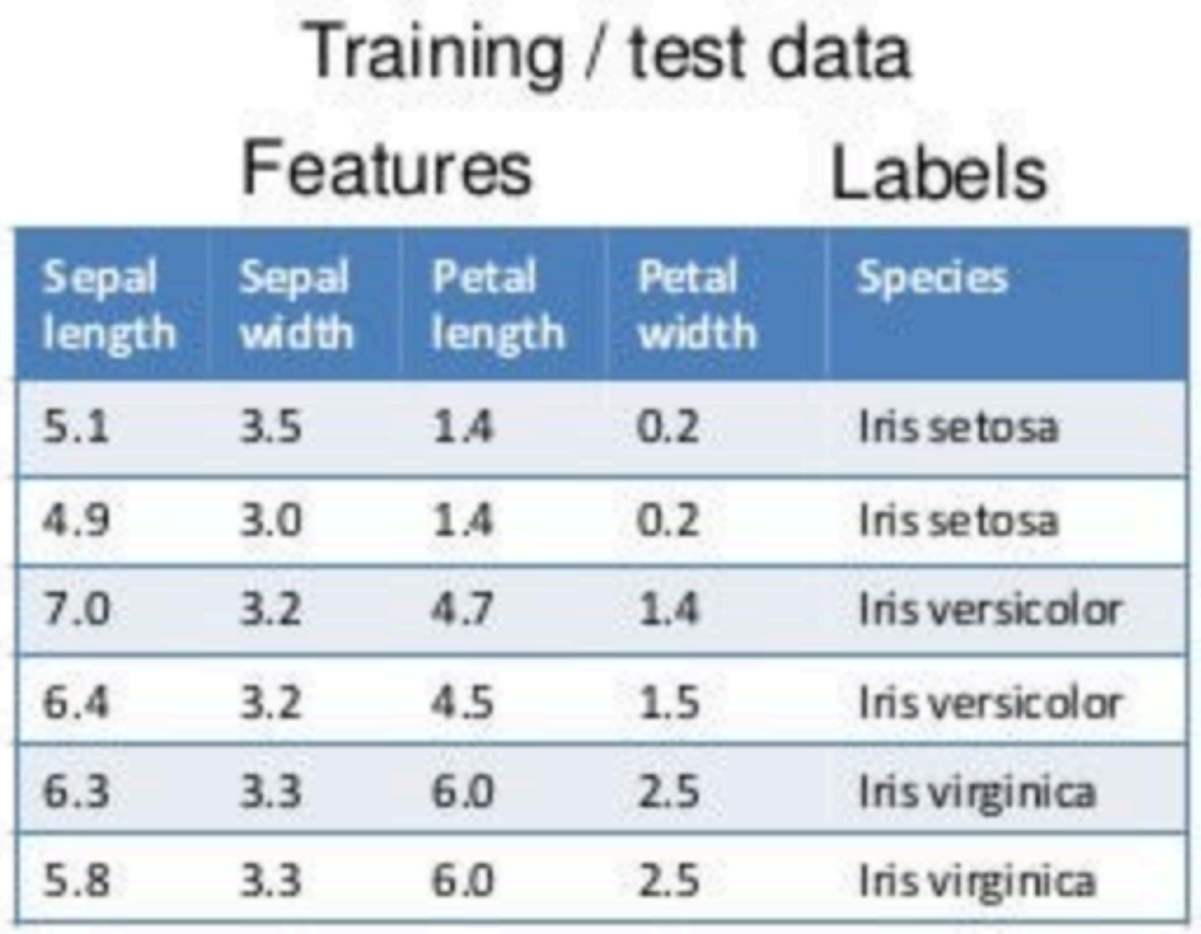
AI State of the art Applications



Features

In the context of machine learning, features refer to individual measurable properties or characteristics of the data that are used as inputs for a predictive model.

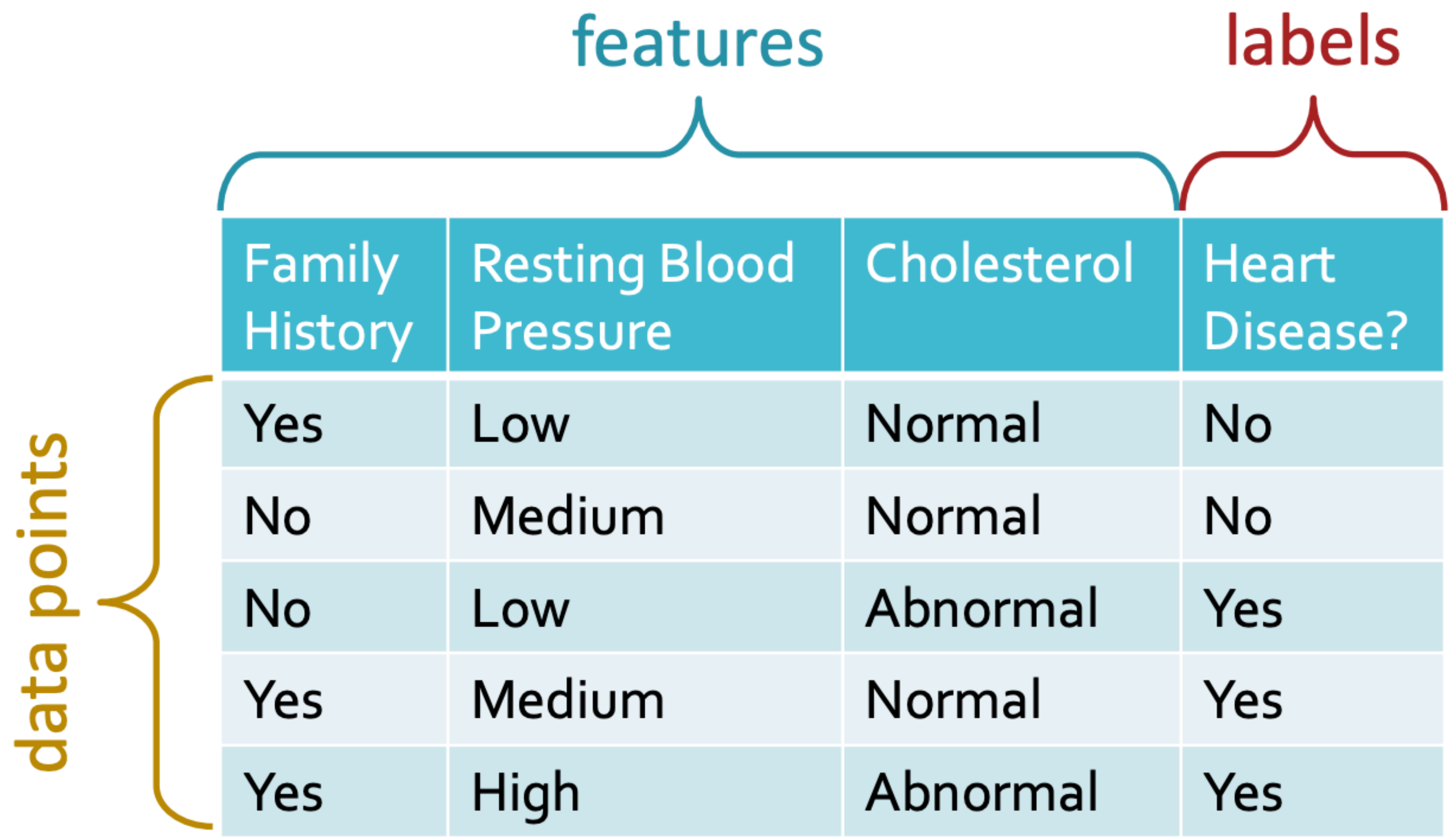




Training / test data

Features Labels

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica

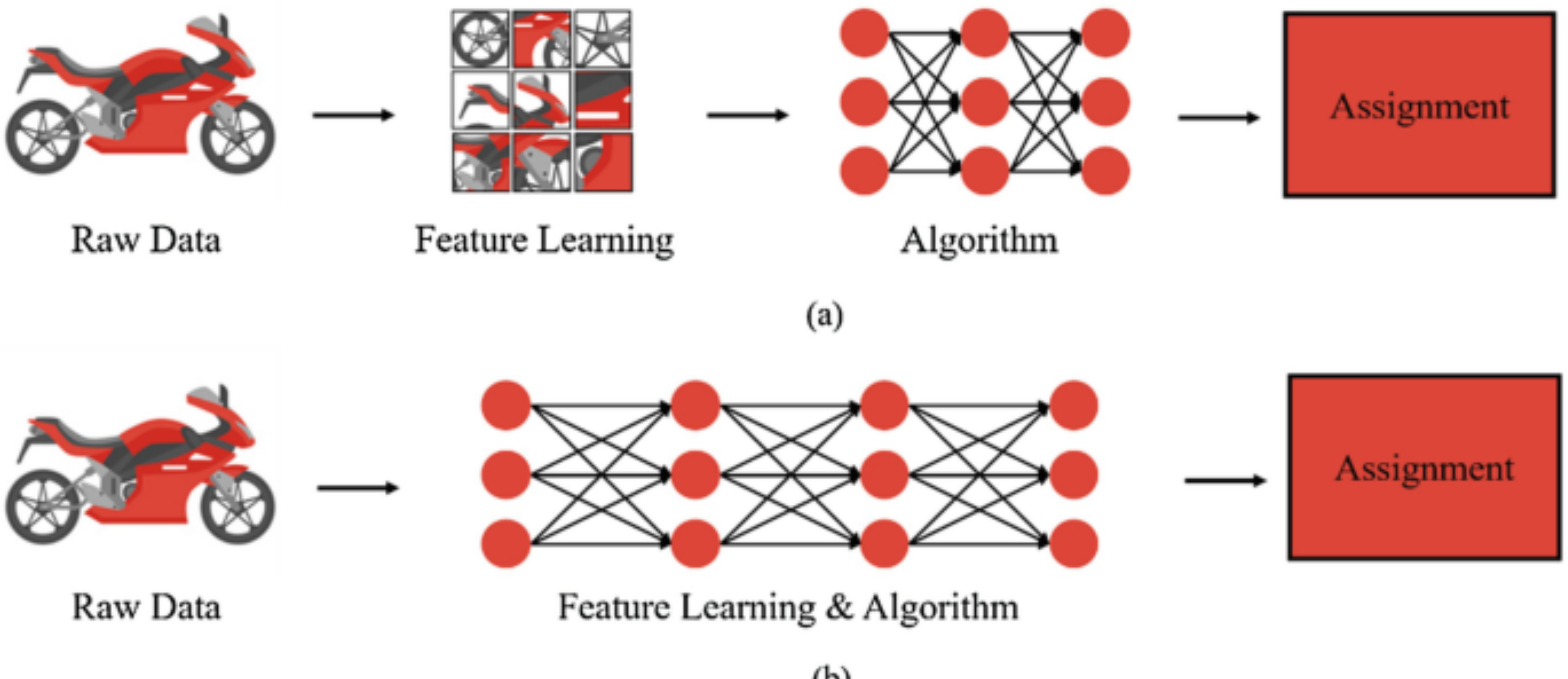


Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Feature Extraction

Feature extraction is a fundamental aspect of machine learning, particularly in tasks involving structured or unstructured data.

It is the process of selecting, transforming, and combining raw data into a set of features that capture relevant information for a given task or problem



Practical - Extract features from image

- Edge Detection: Detects edges in the image using the Canny edge detector algorithm. (Calculates the number of detected keypoints)
- Keypoint Detection: Identifies interest points (corners) in the image.
- HOG Features: Calculates the Histogram of Oriented Gradients (HOG) features. (Calculates the mean and standard deviation of the HOG features)

Thank You

TISHK INTERNATIONAL UNIVERSITY

TIU

ERBIL 2008