أكاديمية طويق
Tuwaiq Academy

Introduction to Machine Learning

Where We Left Off:

أكاديمية طويق
Tuwaiq Academy

AI Concepts and Knowledge Representation:

- Knowledge-based Agents and Models
- Ontologies and Knowledge Graphs
- Rule-Based Systems and Semantic Web
- Probabilistic Reasoning and Bayes' Rule

- **Probabilistic reasoning** lays the groundwork for ML’s handling of uncertainty.
- **Knowledge representation** inspires how features are extracted for ML models.

MACHINE LEARNING IS CHANGING OUR WORLD !

أكاديمية طويق
Tuwaiq Academy

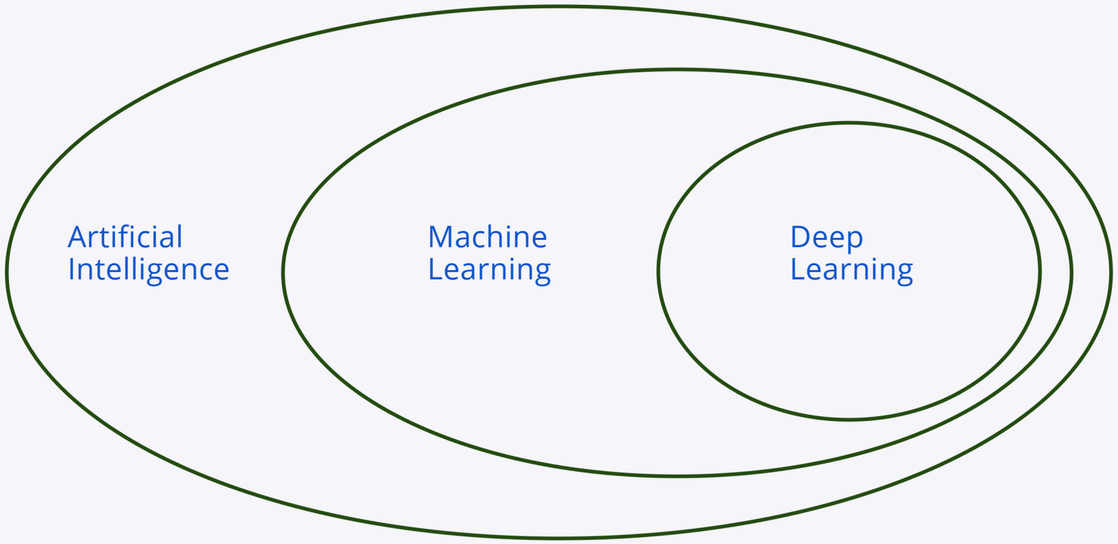
- Search engines learn what you want
- Recommender systems learn your taste in books, music, movies,...
- Algorithms do automatic stock trading
- Google Translate learns how to translate text
- Siri learns to nderstand speech
- DeepMind beats humans at Go
- Cars drive themselves
- Smart-watches monitor your health

THE WORLD OF AI

أكاديمية طويق
Tuwaiq Academy

... and the connections to Machine Learning and Deep Learning
Many people are confused what these terms actually mean.

- **AI** refers to machines trained to perform tasks requiring "intelligence."
- Originated in the 1940s with the invention of computers.
- Key early contributors: Turing and John von Neumann.
- **AI encompasses fields like machine learning, NLP, computer vision, robotics, and more.**
- Often confused with **ML** or basic data analysis in modern usage.

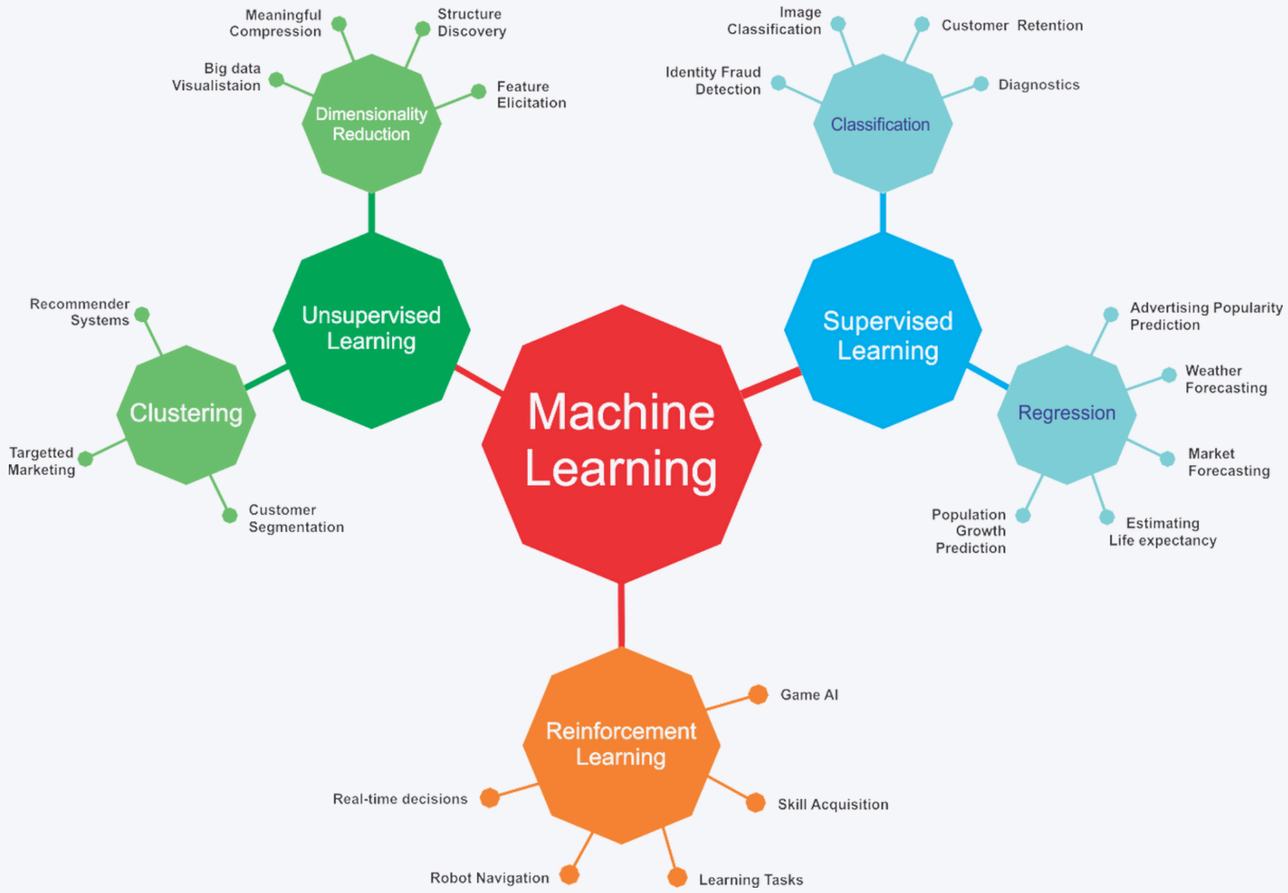


Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

MACHINE LEARNING

أكاديمية طويق
Tuwaiq Academy

- **ML** focuses on **mathematically** well-defined, narrow tasks.
- Constructs **predictive/decision** models from data rather than explicit programming.
- **Learning** is defined as improvement in task performance (T) based on experience (E), measured by performance (P).
- Tom Mitchell's 1998 definition highlights this concept.



https://www.oreilly.com/library/view/java-deep-learning/9781788997454/assets/899ceaf3-c710-4675-ae99-33c76cd6ac2f.png

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

ML VS. STATS

أكاديمية طويق
Tuwaiq Academy

- **ML and Statistics** originated in different fields but share equivalent mathematical foundations.
- ML models focus on **precise predictions**; statistical models emphasize pattern interpretation and sound inference.
- **ML and predictive modeling** in statistics address similar problems using similar tools.
- Communities remain divided with **inconsistent terminology causing confusion.**
- **ML** can often be viewed as nonparametric statistics combined with efficient numerical optimization.

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

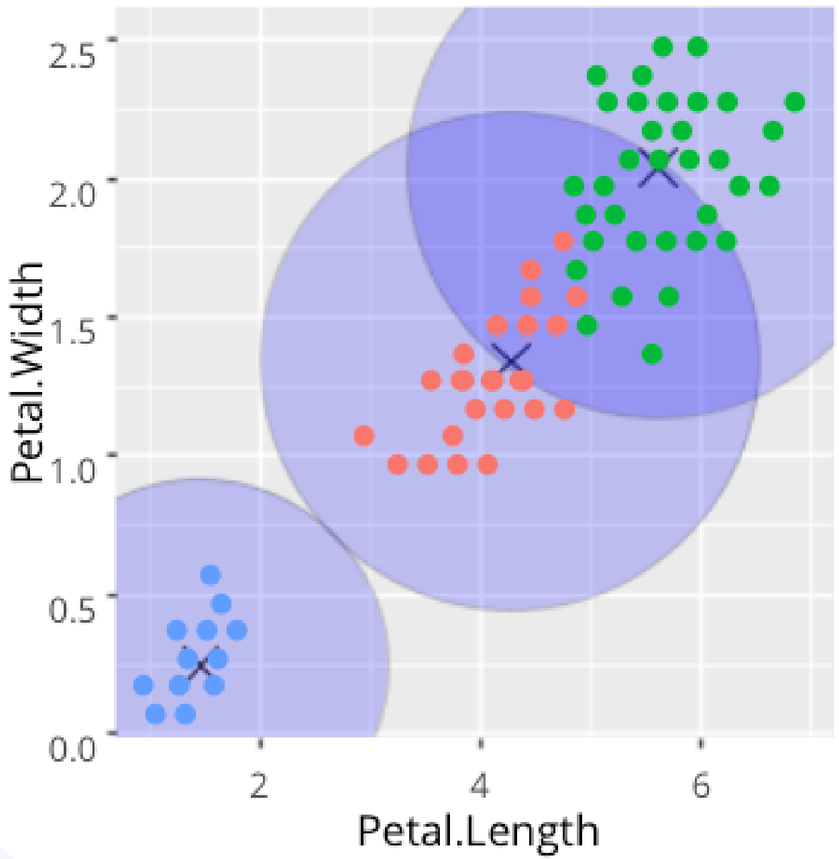
UNSUPERVISED LEARNING

أكاديمية طويق
Tuwaiq Academy

- **Unsupervised learning** involves data **without** labels (no "true" output to optimize against).

- Focuses on finding patterns within input data (x).

- Common methods in unsupervised learning:
  - **Dimensionality reduction** (e.g., PCA, Auto encoders) > Compress information.
  - **Clustering**: Group similar observations.
  - Outlier and anomaly **detection**.
  - Association rules.



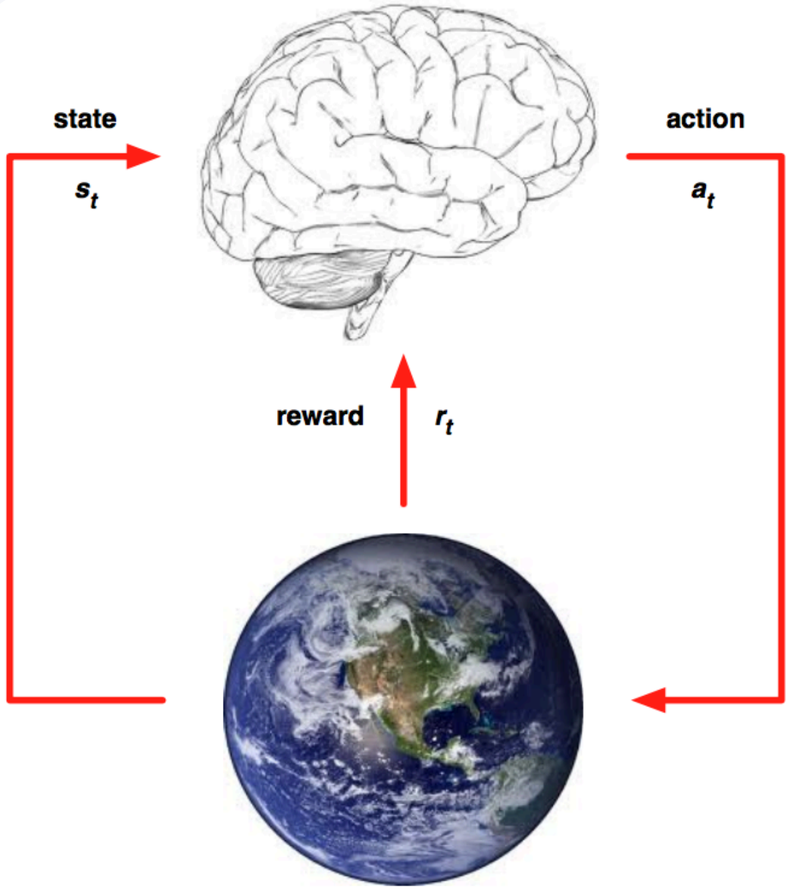
Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

REINFORCEMENT LEARNING

أكاديمية طويق
Tuwaiq Academy

- **Reinforcement Learning (RL)** is a general-purpose AI framework.
- Interaction with the environment involves:
  - Observing the state.
  - Receiving a reward.
  - Executing an action.

- **Goal**: Maximize future rewards.
- **Challenges**: Reward signals can be sparse, noisy, and delayed.



Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

SUPERVISED LEARNING

أكاديمية طويق
Tuwaiq Academy

- Supervised learning for **Regression and Classification.**
- **Predict** labels (y) based on features (x) by learning patterns from **labeled data.**

- **Key foundational concepts in supervised ML:**
  - Types of data used for learning.
  - Formalizing the learning goal.
  - Understanding prediction models.
  - Quantifying predictive performance.
  - Defining learning algorithms.
  - Operationalizing the learning process.

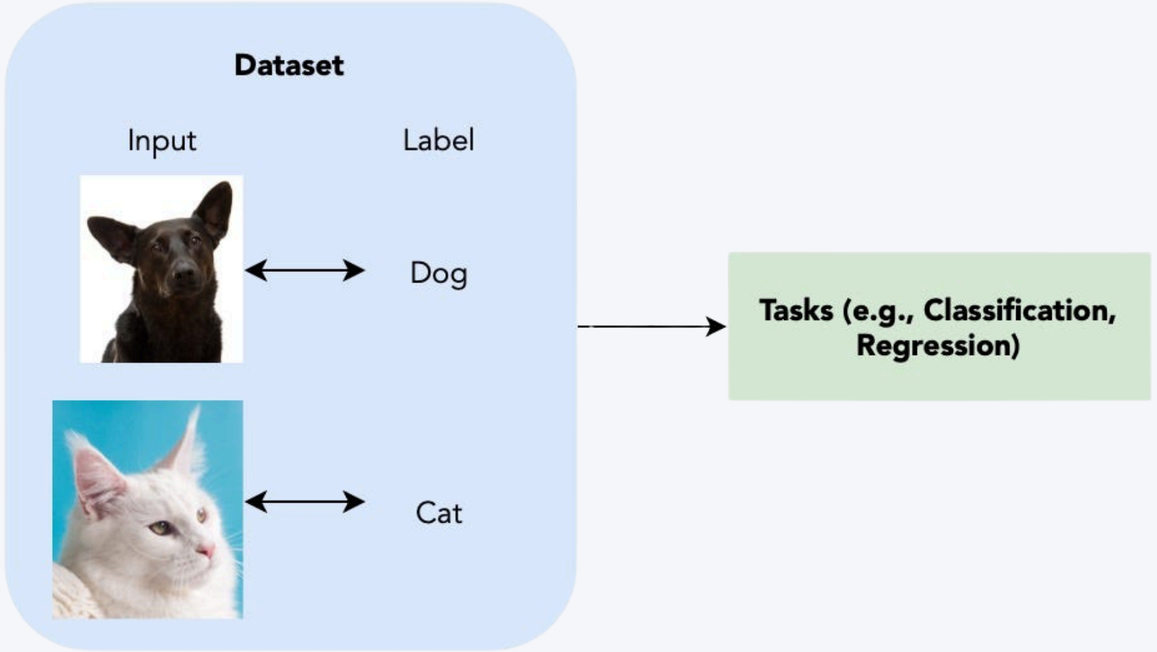


Image from

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

SUPERVISED LEARNING

أكاديمية طويق
Tuwaiq Academy

Regression:

- Predicts a **continuous numerical value.**
- Example: Predicting house prices based on features like size, location, and number of rooms.

Classification:

- Predicts a **categorical label (class)** from predefined categories.
- Example: Determining if an email is "spam" or "not spam."

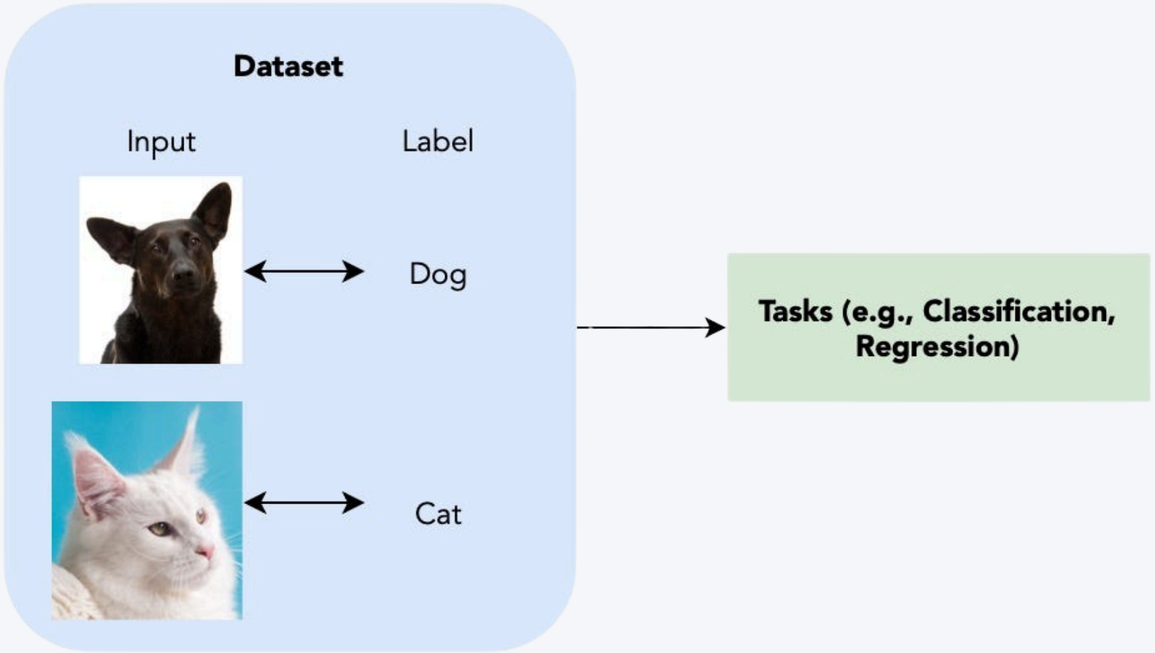
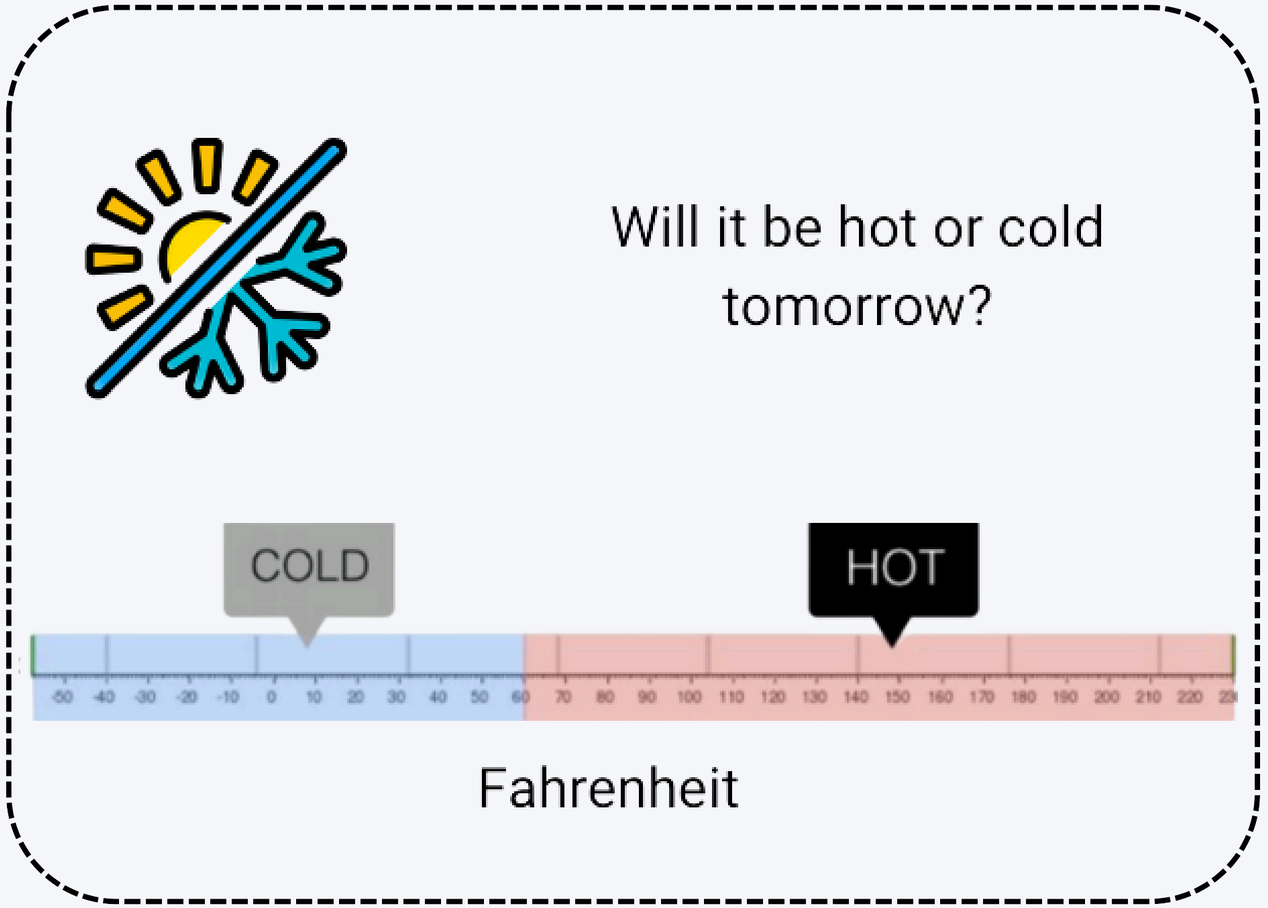


Image from

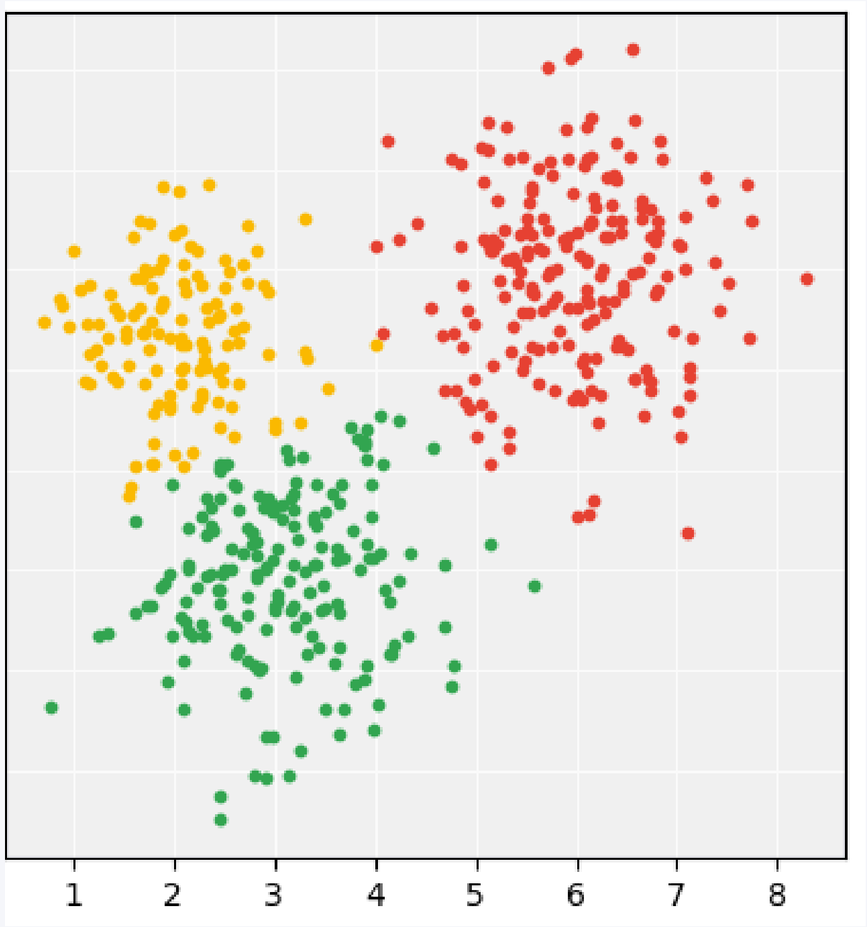
Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

SUPERVISED OR UNSUPERVISED LEARNING



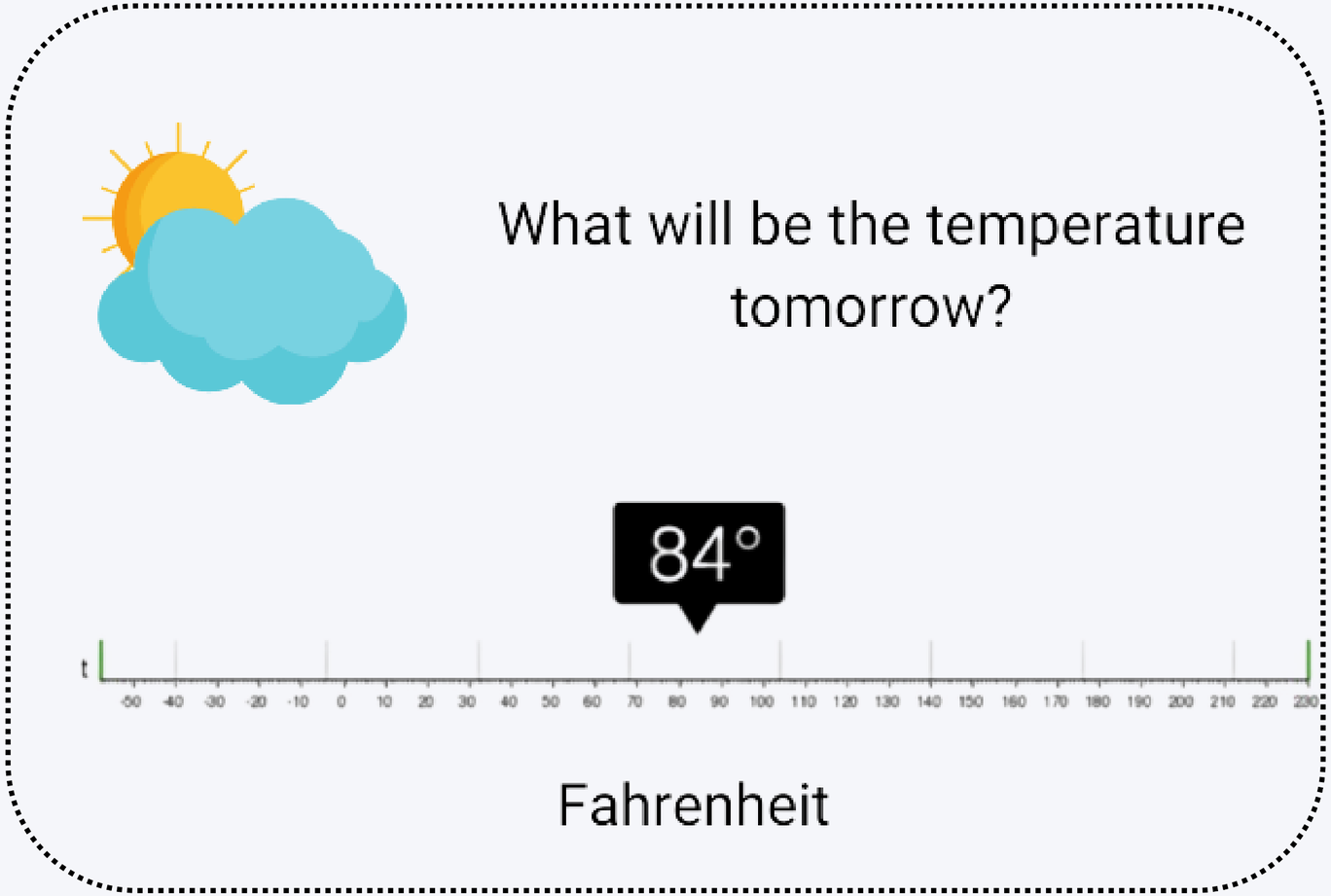
Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

SUPERVISED OR UNSUPERVISED LEARNING



Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

SUPERVISED OR UNSUPERVISED LEARNING

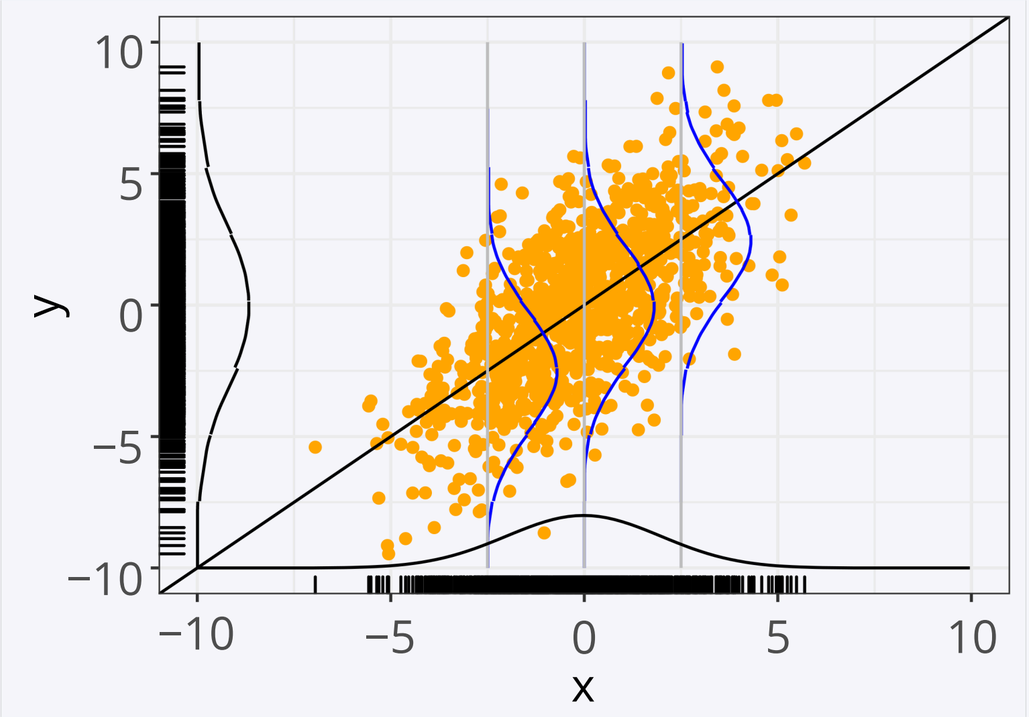


Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

ML-BASICS DATA

You Must Understand ..

- Understand structure of tabular data in ML
- Understand difference between target and features
- Understand difference between labeled and unlabeled data
- Know concept of data-generating process



Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

IRIS DATA SET

أكاديمية طويق
Tuwaiq Academy

Introduced by the statistician Ronald Fisher and one of the most frequently used toy examples.

- Classify iris subspecies based on flower measurements.
- 150 iris flowers: 50 versicolor, 50 virginica, 50 setosa.
- Sepal length / width and petal length / width in [cm].

Source: https://rpubs.com/vidhividhi/irisdataeda

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

DATA IN SUPERVISED LEARNING

أكاديمية طويق
Tuwaiq Academy

- The data we deal with in supervised learning usually consists of observations on different aspects of objects:

  - **Target**: the output variable / goal of prediction
  - **Features**: measurable properties that provide a concise description of the object

We assume some kind of **relationship between the features and the target**, in a sense that the value of the target variable can be explained by a combination of the features.

Features x				Target y
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.3	3.0	1.1	0.1	setosa
5.0	3.3	1.4	0.2	setosa
7.7	3.8	6.7	2.2	virginica
5.5	2.5	4.0	1.3	versicolor

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

ATTRIBUTE TYPES

أكاديمية طويق
Tuwaiq Academy

Feature and Target Variable Types:

- **Numerical variables: Real-valued data (R).**
- **Integer variables: Discrete whole numbers (Z).**
- **Categorical variables: Defined categories (e.g., colors, types).**
- **Binary variables: Two possible values (e.g., 0 or 1, True/False).**

Tasks Based on Target Variable:

- **Regression: For continuous numerical targets.**
- **Classification: For categorical or binary targets.**

Handling Features:

- Most learning algorithms work with numerical features.
- Some algorithms, like decision trees, can handle integers and categorical features directly.
- For other cases, features must be encoded into numerical formats (e.g., one-hot encoding, label encoding).

- **Assumption**: Unless specified, features are considered numerical.

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

ENCODING FOR CATEGORICAL FEATURES

أكاديمية طويق
Tuwaiq Academy

Encoding Methods:

One-Hot Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

- The machine learning algorithm can handle redundancy in input data (e.g., decision trees, random forests, neural networks).
- You want to maintain all category distinctions without losing any information.
- You don't have constraints on the input matrix being non-singular.
- Example: For neural networks or clustering tasks where categorical information needs complete representation.

Dummy Encoding

Places
New York
Boston
Chicago
California
New Jersey

→

[illegible]	Boston	Chicago	California	New Jersey
[illegible]	0	0	0	0
[illegible]	1	0	0	0
[illegible]	0	1	0	0
[illegible]	0	0	1	0
[illegible]	0	0	0	1

- The machine learning algorithm requires non-singular input matrices (e.g., linear regression, logistic regression).
- You need to reduce redundancy in encoded features to improve computational efficiency and avoid multicollinearity.
- Example: In linear regression, removing one column prevents singularity in the design matrix.

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

OBSERVATION LABELS

أكاديمية طويق
Tuwaiq Academy

We call the entries of the target column labels. We distinguish two basic forms our data may come in:

- For **labeled** data we have **already** observed the **target**
- For **unlabeled** data the target labels are **unknown**

	Features x				Target y
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
labeled data	4.3	3.0	1.1	0.1	setosa
	5.0	3.3	1.4	0.2	setosa
	7.7	3.8	6.7	2.2	virginica
	5.5	2.5	4.0	1.3	versicolor
unlabeled data	5.9	3.0	5.1	1.8	?
	4.4	3.2	1.3	0.2	?

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

DATA-GENERATING PROCESS IN ML

أكاديمية طويق
Tuwaiq Academy

Assumption: Data from a Distribution:

- The observed data (D) is assumed to be generated by an underlying process characterized by a probability distribution:

$$P(x,y)$$

- This distribution defines the joint behavior of **input features (x) and target variables (y).**
-

Random Variables:

- x: Random variable representing input features, sampled from the input space X.
- y: Random variable representing target values, sampled from the output space Y.

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

DATA-GENERATING PROCESS IN ML

أكاديمية طويق
Tuwaiq Academy

True Distribution is Unknown:

- The actual distribution P(x,y) is not directly observable or known.
- It represents the real-world process that generates the data.

Goal of Machine Learning:

- Learning involves uncovering or approximating the structure of P(x,y), or parts of it, to make predictions or understand relationships.

- This approximation is done using:
  - A model f(x) for predictions.
  - Algorithms to estimate parameters or identify patterns in P(x,y).

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

DATA-GENERATING PROCESS IN ML

أكاديمية طويق
Tuwaiq Academy

Why This Matters:

- The assumption of a data-generating process underpins statistical and machine learning methods.

- Models trained on observed data aim to generalize to unseen data, relying on the assumption that the data was sampled from P(x,y)

Key Challenges:

- The true distribution P(x,y) is often complex, high-dimensional, and only partially represented in finite datasets.

- Learning methods must handle this uncertainty and approximate P(x,y) effectively for the task at hand.

Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.

DATA-GENERATING PROCESS IN ML

أكاديمية طويق
Tuwaiq Academy

Assumption: Data is assumed to be independent and identically distributed (i.i.d.).

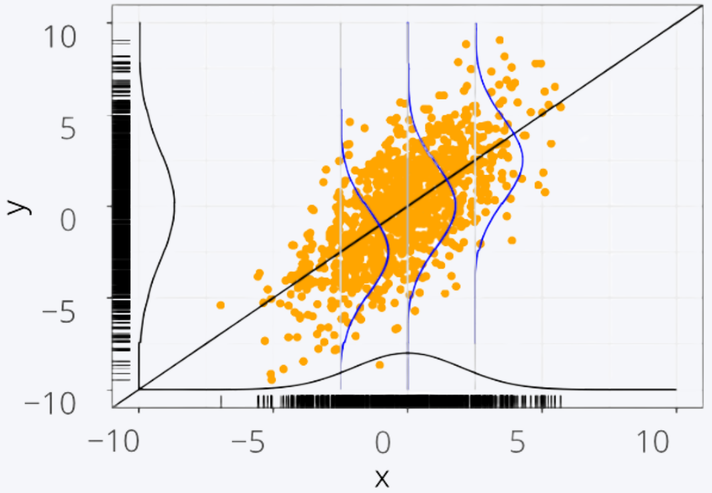
- All samples come from the same probability distribution.
- Each sample is independent of others.

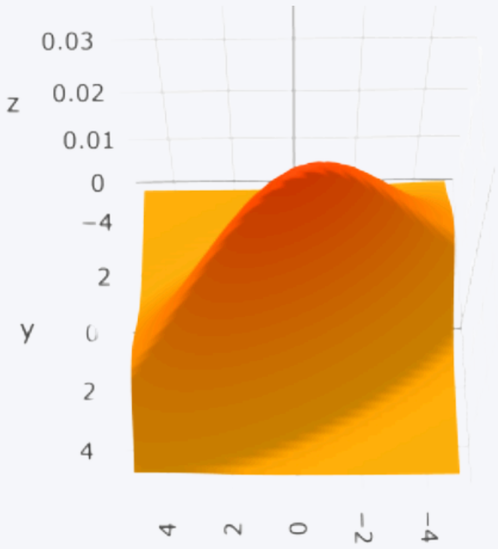
Why This Matters:

- This assumption simplifies theoretical foundations in machine learning.
- It ensures the model learns from a consistent and unbiased data source.

Limitation:

- The i.i.d. assumption may not hold in real-world scenarios (e.g., time series or dependent data).





Bothmann, L., Strickroth, S., Casalicchio, G., Rügamer, D., Lindauer, M., Scheipl, F. & Bischl, B.. (2023). Developing Open Source Educational Resources for Machine Learning and Data Science. Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop, in Proceedings of Machine Learning Research 207:1-6 Available from https://proceedings.mlr.press/v207/bothmann23a.html.