

# dataset

## Authors

Alexander Jung\*, Department of Computer Science, Aalto University, Espoo, Finland

\*Corresponding author: alex.jung@aalto.fi

## Abstract

A dataset is a finite collection of data points on which model training and validation are performed. Strictly speaking, a dataset is a set of distinct data points. With slight abuse of terminology, the term is also used interchangeably with sample, a finite sequence that may contain repetitions. In practice, machine learning (ML) methods often do not have access to the actual dataset. Instead, they must use approximate representations such as tables in the relational model.

## Definition

A dataset is a collection  $\mathcal{D} = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}\}$  of data points. Machine learning (ML) methods use a dataset for model training and validation. Strictly speaking, a dataset is an unordered collection of distinct data points, i.e., a set with no repetitions. In ML literature, however, the term is often used for a sample: a sequence of  $m$  data points, indexed by  $r = 1, \dots, m$ , that may contain repetitions (see Fig. 1).

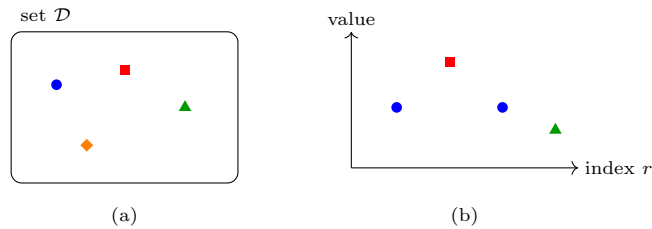


Figure 1: Two readings of the word *dataset*. (a) Strictly, a dataset is a set of distinct data points, with no order and no repetitions. (b) In ML literature, dataset often denotes a sample: a finite sequence of data points indexed by  $r$  that may repeat (positions  $r = 1$  and  $r = 3$  carry the same data point)

The notion of a dataset is broad: data points may represent concrete physical entities (such as humans or animals) or abstract objects (such as numbers). For illustration, Fig. 2 depicts a dataset whose data points are cows.



Figure 2: Cow herd somewhere in the Alps

In ML applications, it is often not possible to directly access the underlying dataset. For instance, accessing the dataset in Fig. 2 would require visiting the cow herd.

In practice, a dataset is represented within some data model, a rigorous formalism for representing and processing data (Abiteboul et al., 1995; Hoberman, 2009; Ramakrishnan and Gehrke, 2002; Silberschatz et al., 2019). For example, the relational model organizes data as a collection of tables, or relations (Codd, 1970; Silberschatz et al., 2019).

A single table consists of rows and columns, where each row corresponds to a single data point and each column represents a specific attribute of a data point. The order of rows is immaterial, and each attribute is associated with a domain that specifies its set of admissible values. ML methods use these attributes as the features or the label of a data point, so the attribute domains correspond to the feature space and the label space. Table 1 shows an example obtained from the dataset in Fig. 2.

TABLE 1  
RELATION (OR TABLE) REPRESENTING A DATASET OF COWS

| Name  | Weight | Age | Height | Stomach temperature |
|-------|--------|-----|--------|---------------------|
| Zenzi | 100    | 4   | 100    | 25                  |
| Berta | 140    | 3   | 130    | 23                  |
| Resi  | 120    | 4   | 120    | 31                  |

A table captures the content of a dataset, but not whether that content is appropriate for a given purpose. A dataset used in ethical and lawful machine learning systems (ML systems) is required to satisfy further properties. For example, the EU AI Act requires that datasets on which high-risk ML systems are trained be “sufficiently representative” of the intended use cases (European Parliament and Council of the European Union, 2024, Art. 10(3)).

Trustworthy artificial intelligence (trustworthy AI) also requires documenting how a dataset was assembled, for example through a datasheet that records the dataset’s motivation, composition, and collection process (Gebru et al.,

2021). When the data points are personal data, this becomes a legal duty: under the purpose limitation principle, the general data protection regulation (GDPR) requires that such data be collected for “specified, explicit and legitimate purposes” that are fixed at the time of collection (European Parliament and Council of the European Union, 2016, Art. 5(1)(b)).

## Cross References

data point, sample, data, relational model, feature, feature space, label space, EU AI Act

## References

1. Gebru et al. (2021). Datasheets for datasets. *Commun. ACM*, 64(12), pp. 86–92.
2. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). *Official Journal of the European Union*, L series, Jul. 12. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.
3. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *Official Journal of the European Union*, L 119/1, May 4. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
4. Abiteboul et al. (1995). *Foundations of Databases*. Addison-Wesley, Reading, MA, USA.
5. Codd (1970). A Relational Model of Data for Large Shared Data Banks. *Commun. ACM*, 13(6), pp. 377–387.
6. Hoberman (2009). *Data Modeling Made Simple: A Practical Guide for Business and IT Professionals*. 2nd ed., Technics Publications, Basking Ridge, NJ, USA.
7. Ramakrishnan and Gehrke (2002). *Database Management Systems*. 3rd ed., McGraw-Hill, New York, NY, USA.
8. Silberschatz et al. (2019). *Database System Concepts*. 7th ed., McGraw-Hill Education, New York, NY, USA.