



**Why**

We want to use the notation and language of matrices to discuss datasets.

**Definition**

The *data matrix* corresponding to a data set of  $n$  records which are length  $d$  tuples of real numbers is the  $n$  by  $d$  matrix whose  $i$ th row is the  $i$  element of the data set. Some authors refer to the data matrix as the *design matrix*.

**Notation**

Let  $(a^1, \dots, a^n)$  where  $a^i \in \mathbf{R}^d$  for  $i = 1, \dots, n$ . Let  $A$  be the  $n \times d$  matrix whose rows are the  $a^i$ . Then  $A$  is the data matrix of  $(a^1, \dots, a^n)$ . In other words,

$$A = \begin{pmatrix} (a^1)^\top \\ \vdots \\ (a^n)^\top \end{pmatrix}.$$



