



# Datasets

## 1 Why

We want language and notation for collecting elements from a set.

## 2 Definition

A *dataset* is a finite sequence of elements from a set. We call the elements of a dataset *records*.

We use the single word dataset since “data set” carries the connotation of being a set of data. The terminology dataset is standard. Unfortunately, however, datasets are not sets. Datasets (in theory and practice) contain repeat elements. More linguistically, “data” is a Latin plural and so is analogous to saying something like “records set,” instead of something like “record set.” For these reasons we prefer the single word dataset.

### 2.1 Notation

Let  $A$  be a non-empty set. We usually denote sequences by  $(a_1, \dots, a_n)$ , but for datasets we will use a superscript,  $(a^1, \dots, a^n)$ . Often  $A$  is itself a set of sequences, and we will want to refer to the  $i$ th component of  $a^1$  by  $a_i^1$ .