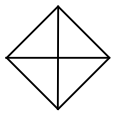

Datasets

1 Why

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

2 Definition

A *dataset* of size n from a non-empty set A is an element of the direct product A^n. In other words, it is an n-tuple. In the case that we speak of a dataset, we call its coordinates the records. So, for example, we call the first coordinate the first record and the second coordinate the second record. Of course the ith coordinate is the ith record, for $i = 1, \ldots, n$.

Although the terminology dataset is standard, datasets are not sets. Datasets, in theory and (usually) in practice, contain repeat elements. Thus we use the "dataset," as a single word. Had we used "data set," the language would suggest that the object involved was a set. But it is not; it is a sequence. A minor other point is that "data" is the plural form of a latin noun, and so is analgous 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. Although we usually denote sequences by $(a_1, \ldots, a_n)$, we will denote datasets by $(a^1, \ldots, a^n)$; notably, using a superscript. Often A is itself a set of sequences, and so we will want to refer to the ith component

of a^1 by a_i^1.