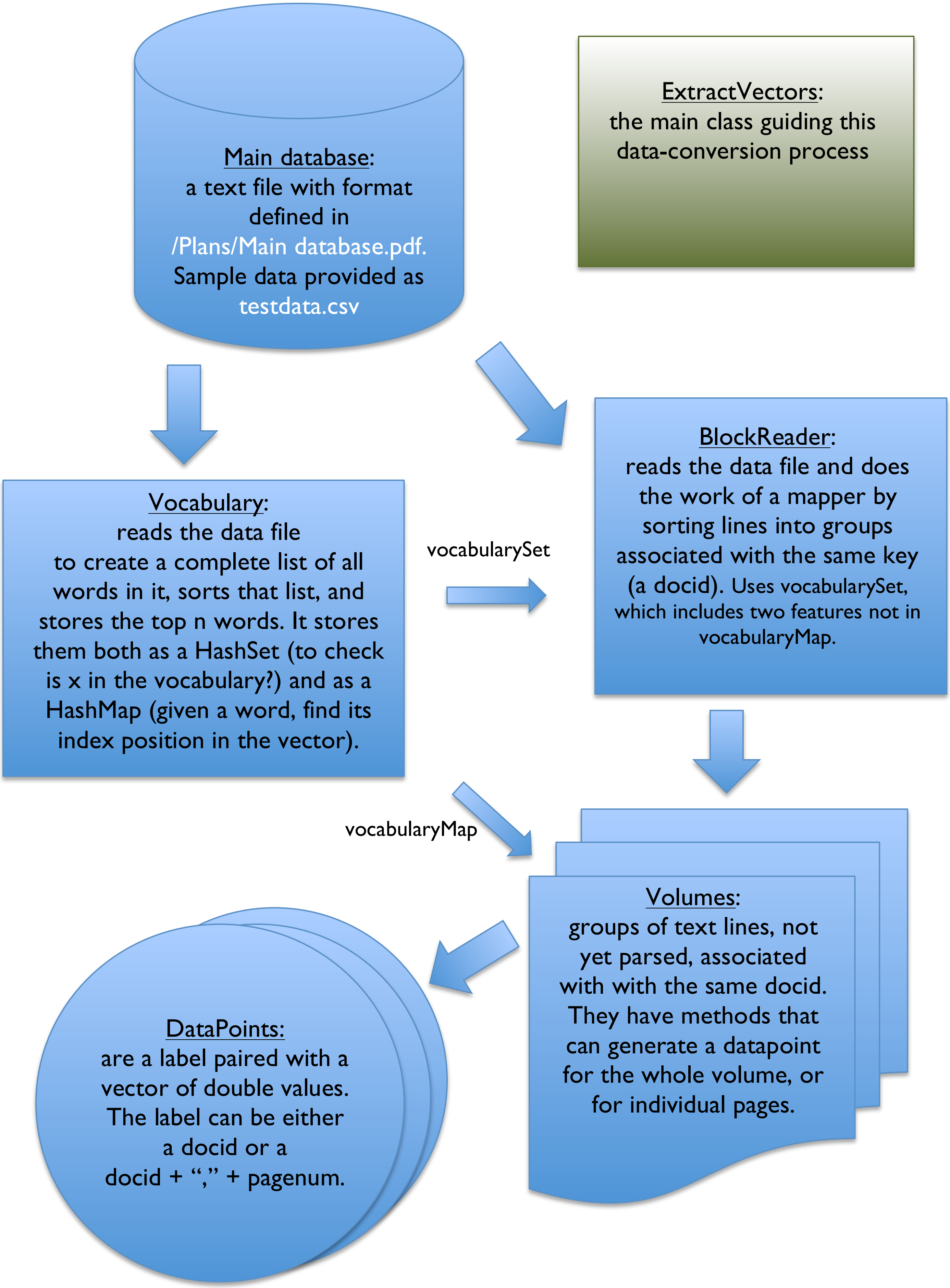

Main database:
a text file with format defined in /Plans/Main database.pdf. Sample data provided as testdata.csv

ExtractVectors:
the main class guiding this data-conversion process

Vocabulary:
reads the data file to create a complete list of all words in it, sorts that list, and stores the top n words. It stores them both as a HashSet (to check is x in the vocabulary?) and as a HashMap (given a word, find its index position in the vector).

vocabularySet

BlockReader:
reads the data file and does the work of a mapper by sorting lines into groups associated with the same key (a docid). Uses vocabularySet, which includes two features not in vocabularyMap.

vocabularyMap

Volumes:
groups of text lines, not yet parsed, associated with with the same docid. They have methods that can generate a datapoint for the whole volume, or for individual pages.

DataPoints:
are a label paired with a vector of double values. The label can be either a docid or a docid + “,” + pagenum.