From [CITE SOURCE], we know that there are theoretical limitations on the usefulness of a multilayer perceptron trained with backpropagation. These considerations help explain the limited usefulness of the MLP in classifying the particle collision data.

The analysis begins with a reworking of the error measure in the limit of an infinite training data set. In this limit, we can move from finite sums of input patterns to integrals over the distribution of the data in the input space. Bishops analysis considers the conditional averages of the target data (conditional here refers to conditional on the given input pattern).

With some simplifications (described on p. 202), he achieves the following terse but powerful result: $y_k(x,w*) = <t_k x>$. Put into English, this states that the output of the neural network corresponds to the conditional average of the target outputs. Thus, the trained network is essentially returning the conditional mean of the target outputs given some input pattern. This is not an artifact of a particular network architecture, or even the use of a neural network (p. 203). Rather, the importance of the conditional average comes from the training via the sum of squares objective function.

The relevance of this theory to our project is that certain types of nonlinearity in the distribution of the data will necessarily be ignored by a learning machine trained to minimize a sum of squares error. If, for example, $<t_k | x>$ is an expected value that few (if any) input patterns actually map to (just as the expected value of a die roll is the impossible 3.5), then the network will try to find structure in the conditional average where such structure is not actually present.

Our data and their corresponding target values may suffer from this sort of nonlinearity.