Abstract: Modern networks can be such that complete inputs and outputs for computational problems on them are *so large*, that there is not time to read or write them in their entirety. However, if one is only interested in small parts of the output at any given time, is it really necessary to solve the entire computational problem? Can one avoid the need to view the whole input? Or, is there a way to take advantage of certain "locality properties" of the problem?

We describe recent work in the model of "local computation algorithms" which for a given input, supports queries by a user to values of specified bits of a legal output. The goal is to design local computation algorithms in such a way that very little of the input needs to be seen in order to determine the value of any single bit of the output. Though this model describes sequential computations, techniques from local distributed algorithms have been extremely important in designing efficient local computation algorithms. We describe results on a variety of problems for which sublinear time and space local computation algorithms have been developed.