Machine Learning System Design

Suppose a massive dataset is available for training a learning algorithm. Training on a lot of data is likely to give good performance when two of the following conditions hold true. Which are the two?

(i) We train a learning algorithm with a small number of parameters (that is thus unlikely to overfit).

No - If the model has a small number of parameters, then it will underfit the large training set and not make good use of all the data.

(ii) We train a learning algorithm with a large number of parameters (that is able to learn/represent fairly complex functions).

Yes - You should use a "***low bias***" algorithm with many parameters, as it will be able to make use of the large dataset provided. If the model has too few parameters, it will underfit the large training set.

(iii) The features x contain sufficient information to predict y accurately. (For example, one way to verify this is if a human expert on the domain can confidently predict y when given only x).

Yes - It is important that the features contain sufficient information, as otherwise no amount of data can solve a learning problem in which the features do not contain enough information to make an accurate prediction.

(iv) When we are willing to include high order polynomial features of x (such as x_1^2, x_2^2, $x_1 x_2$, etc.).

No - As we saw with neural networks, polynomial features can still be insufficient to capture the complexity of the data, especially if the features are very high-dimensional. Instead, you should use a complex model with many parameters to fit to the large training set.