Week 4 Paper

The biggest three things I took away from chapter 4 are:

1. There are a huge number of ways to handle missing or incorrect data, and while few of them are perfect, the options are varied enough so that you can pick and choose methods in order to make sure the data is as optimal as possible for what you need it for.
2. Even after handling incorrect and missing data, there are many transformations that have to be made before the data is ready to work with, be it normalizing and scaling, categorizing/converting to discrete values, or doing transformations to fix a distribution.
3. Being able to easily sample data and splitting data into testing and training portions is hugely important to any project. For a long time I know that I would think of computers working with huge data sets in full all the time, because they are capable of it. However, this is not the best way to do it because a random sampling works just as well, and it is easier for us as the programmer to manage/comprehend the smaller slices of data to make things work before attempting to work with the whole set.