**The Problem of Large Heterogenous Data**

I encountered the problem of large heterogenous data when working on my Kickstarter Analysis project. And it’s probably often the case with heterogenous data that it’s kind of *hidden* heterogenous data in the sense that at first you assume the data is homogenous and normally distributed.

Here are the challenges:

1. **Size** - When the data set is large, in my case almost 380,000 sets, you can’t get an intuitive feel of the data. Even plotting it becomes a challenge in itself because you lose a lot of the details if you don’t restrain the data in some kind.  
   Also, many statistical measures like the mean and median are of little use because the standard deviation is very large.
2. **Not normal** – When the data is not normally distributed, many of the standard tools don’t work.
3. **Many cases** – In my case, there were 15 main categories and 159 sub categories and 6 different states the projects could be in.

We can move this problem over to another topic that may be easier to handle. Let’s say we needed to analyze mammals. Mammals go from the Etruscan shrew (3.5 cm and 1.8 g) to the mighty Blue Whale (30 m and 173 t). And within this class, we also have special cases like the dog which has a huge variety in itself due to artificial breeding.

How do we deal with this and how much time should we denote to analyzing the data and build up domain knowledge?