

Big Data: what's there for Six Sigma?

Marco Reis and Pedro Saraiva

University of Coimbra, Chemical Engineering, Coimbra, Portugal

Abstract

The "Big data" movement is making its own way through different business areas, scientific domains and organizations worldwide. Independently of its meaning being exactly perceived or not, people are receptive to the "nxV's" message, "Volume + Velocity + Variety + Veracity + Value + ..." and many begin proactively searching for the best way to jump in. Data scientists, big software providers and consulting companies have been assuming leadership roles in this process, putting forward an innovation agenda and priorities for the future. Quite unexpectedly, at the same time there is a growing impression that some data-centric professionals, such as Statisticians and Quality Engineers, are essentially passive observers of this big data phenomenon.

Stepping back from all the hassle around "Big data" and trying to define it in more precise terms, we may well come up to the conclusion (as we did), that Big data is not a "new thing" but a "journey". And for many of us this journey started in the earlier 80's, when distributed control systems began to be installed in a massive way in industry. Later on, business intelligence systems, data warehouses, etc., have also emerged and suddenly, a lot of data was being collected and stored. No wonder a large number of professionals (Statisticians, Quality Engineers, Process Engineers) looking at all this movement on the use of data and the primacy of informative data visualization tools, are led to think, with a certain unstated circumspection "hum,... what's really new here?..."

It is our opinion that what makes Big data what it is today is a result of a particular context where the key enablers {Data, Technology, Analytics} are creating unique conditions to bring data to value for the organizations, in a very efficient way, more than what could be achieved in the past. As a data-driven initiative, Six Sigma has all the conditions to take advantage of this opportunity and incorporate it in its DMAIC workflow, as we will discuss.

Big data, as presented today, strongly relies on large volumes of passively collected data. Statisticians know very well the limitations of "happenstance data" and how one can improve with a small set of well-designed experiments or even by dropping out data (sampling). So, what we should also be asking ourselves is, in reverse to the title of this presentation: "Six Sigma: what's there for Big data?"

Drawing from our own experience in addressing difficult sets of data in several different domains, in our presentation we will explore the links between Big data, Six Sigma, Applied Statistics and Quality Engineering.