Abstract:

Today's technologies produce unprecedented amounts of data, characterised by extreme volume, speed, heterogeneity and uncertainty. This Big Data in the raw form is merely is a cost, in terms of storage and collection, but the information hidden within has real value. With the dramatic increase in the availability of data, however, also comes unrealistic expectations of being able to transform this sheer volume into huge advantage. There has appeared notions such as "the end of theory", essentially claiming that "data can speak for itself" without the need for an underlying model.

This has caused some well-deserved skepticism towards much of the Big Data movement. The fact is that the same pitfalls that has existed in statistics, and has been well known in that field for decades, still exist in the context of massive data sets, possibly just on a larger scale: a model, whether implicit or explicit, is needed to make sense of the data; the generative process must be taken into account, especially with regard to noise, bias, and uncertainty; correlation must not be mistaken for causality; and the more data you search for patterns, the more spurious ones you will find.

Nevertheless, there is a huge demand to be able to process the large amounts of streaming and heterogeneous data produced today, and make use of the information that it actually contains. In CAISR we focus our research on the design of systems that, as autonomously as possible, can construct knowledge from real life data created through the interaction between a system and its environment; systems able to handle events that are unknown at the time of design. At the same time, the human role and interactions with the systems is an important research question: clues on interesting data representations; additional relevant data sources; feedback on suggested structures; and categorisations of events are all things that can be provided.

What is important is that machine and human create knowledge together, not like in the traditional Artificial Intelligent or Machine Learning settings where humans provide expertise that the computer is expected to replicate. Such joint human-machine learning leads to aware systems that are curious and never stop exploring. The amount of data available today allows us to focus on more descriptive and explanatory analysis. Users no longer pose well-formulated, concrete questions, but instead require a system capable of highlighting interesting aspects such as deviations, anomalies, relations and co-occurrences.

It is almost effortless to generate data, while the cost of analysing it does not change. We aim for continuous learning model, where the training and usage is not easily separated, and both the system and the user learn and improve their performance all the time, taking advantage of the new data as it arrives.