Equilibrium and Extreme Principles in Discovering Unknown Relationships from Big Data
Part 1: Methods of Advanced Data Analytics in Light of the Equations of Mathematical Physics

By Pavel Barseghyan, PhD

Abstract

The further development of the methods of Advanced Data Analytics requires using the extreme principles and balance equations for identifying relationships hidden in the data. These fundamental methods and principles are widely used in the classical areas of quantitative science.

Besides, any area of the classical quantitative science, including mechanics and electrodynamics, can be represented as an area of Big Data. For doing that, it is sufficient to collect a large amount of data of electrodynamic or mechanical nature from the surrounding environment.

Having the scientific representations of some quantitative area in the form of fundamental equations on the one side, and in the form of Big Data on the other side, a natural question arises of whether there is a correspondence and an agreement between these two views?

Also, if there is such a correspondence between them, are unambiguous mutual transitions between the fundamental equations and the Big Database possible?

In particular, is it possible to obtain the well-known fundamental equations or some of their equivalents for a quantitative field of knowledge from its Big Data in a statistical or semi statistical way?

This paper discusses the mutual relationships and the possible transitions between the fundamental equations of a quantitative science and their corresponding Big Data bases.

The purpose of the paper is to show that the experience of classical quantitative science can be effectively used in the development of contemporary Big Data Analytics in order to create more reliable methods for detecting unknown patterns and relationships hidden in Big Data.

The second part of this work will be devoted to the applications of the principles and approaches proposed in this paper, for analyzing Big Data in the project management area.

Introduction

Imagine electromagnetic phenomena and processes that taking place in our environment all the time. By collecting the results of measurements of even a tiny part of these processes one can have a typical Big Data base of electromagnetic nature.

But we do not make such measurements because we have a fundamental means of describing these phenomena, such as Maxwell's equations and their simplified versions in the form of the wave equation, Kirchhoff's laws, Ohm's Law, and other laws [1].

The same situation we have, for example, for the relationships and patterns of water flow. Imagine for a moment the infinite variety of ways water is flowing, spilling, dropping, and so on every second. Obviously, the measurements of these forms of water flow parameters can also result a typical base of Big Data. But again, there is no any need to do that because we have fundamental means to describe the different forms of water flows - the mechanics of fluids with its equations by which the interpretation of any data on the movement of water is a straightforward task [2].

Obviously, in similar manner one can gather Big Databases in various other fields of human activity and natural processes. We can divide all of these fields into two big groups:

1) Areas that have fundamental means of description in the form of mathematical equations (such as mechanics and electrodynamics);

2) Areas that do not have this capability (such as marketing, or drug development, etc).

For the areas in the first group, it is not necessary to have Big Databases for estimation or forecasting purposes because basic equations can be used to do that.

For the areas of the second group, the absence of fundamental quantitative descriptions makes it necessary to collect Big Databases and develop special methods of data analysis and interpretation.

Therefore, a natural solution to the problems of the areas of the second group is the gradual development of fundamental methods for the analysis and interpretation of Big Data.

Good examples of developing fundamental mathematical theories by gradual generalization of the results of empirical research are the thermodynamics [3] and the electromagnetic theory [1].

In the mid-19th century, both of these areas, by today's terminology could be considered as typical areas of Big Data.

For these classical areas of science is very characteristic the fact that their development path was different from how today's Big Data Analytics is being developed. They have been using the way of gradual generalization of theoretical and mathematical character. And this is the case when investigating the thermodynamics object inherently has a purely statistical nature (due to the velocity distributions of molecules). In this sense, the problems of thermodynamics mid-19th century are very similar to the problems and current situation in the area of marketing and its Big Data.

In turn, this means that the area of marketing, and other areas facing Big Data problems, in addition to their existing semi statistical path of development, which is the central methodology of modern Big Data Analytics, have also an alternative path of development. It is the step-by-step building of adequate mathematical models and equations and creation of fundamental theories of these areas by means of gradual quantitative generalizations.

In this sense, every area that faces the Big Data challenge should have its own fundamental theory, since the purely statistical or semi statistical means are not able to adequately reflect the essence of the behavior of the studied objects.

Is it possible to imagine a statistical method, or a structured semi statistical method (similar to the methods of contemporary Big Data Analytics), which is able to reflect the richness of the four Maxwell's equations? This is an impossible dream even for the most advanced statistical methods.

It also means that the non-statistical way of gradual theoretical and mathematical generalizations for the areas with problems such as Big Data has no alternative.

Mutual links between Big Data and basic equations of an arbitrary area of knowledge

In various fields of human activity there are rapidly increasing bases of Big Data, and there is a strong business need for rapid and high-quality analysis and interpretation of this data by means of contemporary methods of data analytics.

For further development of this area of knowledge it is important to use the rich experience of physics in order to find adequate answers to the challenges of Big Data Analytics.

In this sense, it is very important to find out the mutual links between Big Data and the corresponding mathematical equations for the areas, where there are well developed means of quantitative description.

If we consider the analysis and interpretation of data by using equations of mechanics, electrodynamics and other well-developed theories, then in this area all issues are well studied. This is the way from the equations to data, where all the issues have clear and precise answers.

But the issues are not clear and do not have unambiguous answers on the other way: from Big Data to the equations, which is very important for the further development of the methods of Advanced Data Analytics.

To discuss this issue let’s consider an example of theoretical mechanics and conduct a thought experiment with many varieties of mechanical processes taking place in our vicinity. Imagine also that we have the ability to measure mechanical properties of these movements or mechanical processes.

Clearly, by accumulating data in this way, we will have a typical Big Data base, which contains a wealth of information about the moving masses, their velocities and accelerations, the path traversed by them and other parameters of mechanical nature.

Continuing the thought experiment let’s try to find out what can provide this Big Data on the laws of motion of bodies.

On the one hand, we know that each separate movement can be fully and comprehensively described by the equations of mechanics [4].

On the other hand, we have the typical Big Data base of mechanical nature, which basically contain the same mechanical laws as the equations of theoretical mechanics.

In such a situation, the question is whether we can obtain adequate and reliable relationships in the field of mechanics by processing our Big Data by means of statistical or semi statistical methods?

Or, in the ideal case, can we obtain the well-known equations of mechanics, or some their equivalents from the Big Data by means of modern Data Analytics?

In view of the apparent impossibility of obtaining this kind of results, we can simplify the problem as follows: Taking the equations of mechanics as completely accurate means of describing the motion of bodies, let us try to figure out how much accuracy can we achieve when trying to extract the fundamentals of theoretical mechanics from Big Data.

One can also come to similar conclusion that the probability of obtaining Maxwell's equations from Big Data of electromagnetic nature by means of modern Data Analytics is negligible.

Another conclusion of this mental experiment could be that instead of trying to re-obtain the laws of mechanics or electrodynamics by unreliable means of statistical or semi statistical methods of Data Analytics, perhaps we need to improve the methods of Data Analytics itself. Namely, perhaps the Data Analytics itself needs to utilize methods of classical quantitative sciences to overcome its own difficulties.

This is a top-down approach that will allow building mathematical theories for different Big Data bases built on specific principles, similar to the equations of mechanics that are totally derived from a single principle of least action [5].

Solution of the problem is the creation of a new Data Analytics methodology similar to the methods of theoretical physics

In all likelihood, the judgments made in respect of mechanics and electrodynamics are also applicable to many other equations of the fundamental nature and corresponding Big Data.

This state of affairs logically leads to the following fundamental question: What prevents us from directly applying the experience of classical quantitative science to solving problems in the areas of Big Data?

It is well known that the methods of classical quantitative science are based on:

1. Balance equations, or equations of state, which reflects the equilibrium of systems, and are based on data from specific areas, and

2 . Variational, or extreme principles, which are also reflections of reality [6].

But the main feature of classical quantitative methods is that they are not based directly on the data, but on the **generalizations** of data. Therefore, quantitative theories built on the basis of such a methodology in general are based on the data, but at the same time they are invariant with respect to specific data.

Big Data: State of the Art

First of all, it should be noted that the realm of Big Data has critical importance for business. At the same time it must also become a field of scientific research of a fundamental nature, as it is important for the success of the businesses in many areas of human activity.

With this regard, it is clear that the traditional statistical approach to the problem of Big Data cannot be justified for several reasons. Most importantly, the reliability of the results obtained in this way is always questionable, which makes them unacceptable from the business point of view. The cause of this unreliability is that in essence Big Data is a result of collecting information from corresponding realms of knowledge, and not a result of a planned process in the sense of the theory of experiments [7].

From the very beginning the results of a planned experiment are divided into groups based on the conditions in which these results are obtained. That means each group contains results of experiment conducted in the same conditions. This cannot be said about Big Data obtained randomly from different sources and in different conditions.

This in turn means that to increase the reliability of the statistical analysis of Big Data, the data points must be rationally grouped by some features of similarity. Otherwise, they cannot be compared with each other, and therefore cannot be processed together [8].
Therefore the problem of data similarity becomes a central issue in analysis and interpretation of Big Data.

The Big Data bases usually are multivariable systems, so here we are not talking about absolute similarity of conditions as in the planned experiments. Here we talk about relative similarity by "important" parameters.

Thus, the bottom up statistical analysis of Big Data cannot give us reliable relationships that are stable with respect to changes in data. To obtain reliable relationships we need to have a top-down methodology to group the data by "similarity", and then combine it with the bottom up statistical processing procedures [8].

This means that depending on what kind of top-down hypothesis of data "similarity" will be used during the grouping of data, its processing results may significantly differ from each other.
This may make an additional element of unreliability in the mixed methodology of top-down (hypothesis of data "similarity") and bottom-up (statistical processing within the group) processing of Big Data.

This state of affairs in the analysis and interpretation of Big Data indicates the need for a critical review of the methodology of Advanced Data Analytics in terms of risk assessment associated with it.

A possible way to test the methods of Data Analytics and estimate associated risks

Since the adequacy and accuracy of the known equations of physics do not cause any doubt, they can be used as an ideal means of testing the adequacy of the methods of modern Data Analytics. This approach creates fundamental possibilities of testing the methods of modern Data Analytics and objectively measuring their accuracy.

In particular, this idea of testing can be carried out using the equations of physics as the standards of accuracy and comparing their solutions with the results of Data Analytics.

From this perspective, the analysis of the methods of Data Analytics shows that they contain serious risks associated with their qualitative adequacy and quantitative accuracy.

It's one thing to compare the accuracy and the usefulness of the methods of Data Analytics with the results of data processing by purely statistical methods, and quite another thing to compare them with the solutions of the fundamental equations, which have much better accuracy.

It is natural that the average accuracy of the methods of Data Analytics will be higher than the accuracy of purely statistical methods and it is this fact that lies at the basis of their successful business applications.

The fact that it is practically impossible to obtain an adequate fundamental description of the phenomena and processes from Big Data by using methods of Data Analytics indicates that these methods may have greater risks in those areas of knowledge that have no means of quantitative descriptions of a basic nature.

This question can be discussed also in terms of the uniqueness of the fundamental description of the phenomena and processes, arguing that the same phenomenon may have a few basic descriptions, which is unlikely.

Conclusions

1. Purely statistical bottom-up methods, because of their unreliability, are not suitable for uncovering the rich inner content of Big Data in the form of functional dependencies between the parameters of systems;

2. This circumstance stimulated the penetration of analytical methods into the area of Big Data in the form of top- down and bottom-up mixed technologies;

3. Along with positive qualities the methodology of Advanced Data Analytics has significant limitations that are inherent to all statistical methods;

4. The relationships obtained in this way strongly depend on the specific data used, and depending on the circumstances they may contain a substantial portion of unmanaged subjectivity, which may lead to prediction errors;

5. This approach does not allow obtaining all necessary functional relationships on the basis of one basic hypothesis or principle. This is fraught with potential dangers due to the fact that functional relationships obtained on the basis of different local hypotheses later may come into conflict with each other;

6. At the same time, Data Analytics methods based on the local top-down hypotheses pave the way for the penetration of variational methods of theoretical physics in the area of modern data analysis;

7. In classical quantitative science the major development path goes through gradual non-statistical generalizations of methods for description of phenomena and processes, with the ultimate goal of creating a fundamental mathematical theory for a certain area of expertise. This approach should dominate in all areas that are facing Big Data problems;

8. Similar to the methodologies of classic fields of physics, the proposed new methodology of Data Analytics enables deriving stable relationships that are present in the data. The derivation is done in a purely analytical way and without the participation of specific data, and the derived functional relationships are invariant with respect to changes in the data;

9. One important feature of the new methodology of Data Analytics is that here even a single point of data can get interpretation. This, by the way, is a usual thing when interpreting data using the fundamental equations of physics;

10. This last fact is fundamentally changing the policy of processing and interpretation of sparse and scattered data. It is known that selection of "similar" data in even very powerful databases often results in a very small sample size, which needs special analytical methods of interpretation.

Future research

The experience of applying the principles and approaches of theoretical physics to analyze and interpret data in the field of project management for the past ten years have shown their effectiveness and prospects for such areas as marketing, and other areas with Big Data.

In particular, in the field of project management numerous functional relationships between project parameters are obtained in a pure analytical way. Despite the fact that project data was not directly used during the derivation of these relationships, all this derivation is based on generalizations of the data from the practice of project management.

These and other issues will be discussed in the second part of the article.

References

1. Maxwell's Equations, http://en.wikipedia.org/wiki/Maxwell's_equations.
2. Fluid mechanics, http://en.wikipedia.org/wiki/Fluid_mechanics
3. Thermodynamics, http://en.wikipedia.org/wiki/Thermodynamics.
4. Classical mechanics, http://en.wikipedia.org/wiki/Classical_mechanics
5. Principle of least action, http://en.wikipedia.org/wiki/Principle_of_least_action
6. Variational principle, http://en.wikipedia.org/wiki/Variational_principle.
7. Experiment, http://en.wikipedia.org/wiki/Experiment.
8. Pavel Barseghyan (2010) "**Similarity of Projects: Methodology and Analysis with TRANSCALE Tool**". *PM World Today* – July 2010 (Vol. XII, Issue VII). 14 pages. http://www.scribd.com/doc/114755077/Similarity-of-Projects-Methodology-and-Analysis-With-TRANSCALE-Tool

About the Author

Pavel Barseghyan, PhD

Yerevan, Armenia
Plano, Texas, USA

Dr. Pavel Barseghyan is a consultant in the field of quantitative project management, project data mining and organizational science. Has over 40 years' experience in academia, the electronics industry, the EDA industry and Project Management Research and tools development. During the period of 1999-2010 he was the Vice President of Research for Numetrics Management Systems. Prior to joining Numetrics, Dr. Barseghyan worked as an R&D manager at Infinite Technology Corp. in Texas. He was also a founder and the president of an EDA start-up company, DAN Technologies, Ltd. that focused on high-level chip design planning and RTL structural floor planning technologies. Before joining ITC, Dr. Barseghyan was head of the Electronic Design and CAD department at the State Engineering University of Armenia, focusing on development of the Theory of Massively Interconnected Systems and its applications to electronic design. During the period of 1975-1990, he was also a member of the University Educational Policy Commission for Electronic Design and CAD Direction in the Higher Education Ministry of the former USSR. Earlier in his career he was a senior researcher in Yerevan Research and Development Institute of Mathematical Machines (Armenia). He is an author of nine monographs and textbooks and more than 100 scientific articles in the area of quantitative project management, mathematical theory of human work, electronic design and EDA methodologies, and tools development. More than 10 Ph.D. degrees have been awarded under his supervision. Dr. Barseghyan holds an MS in Electrical Engineering (1967) and Ph.D. (1972) and Doctor of Technical Sciences (1990) in Computer Engineering from Yerevan Polytechnic Institute (Armenia). Pavel's publications can be found here: http://www.scribd.com/pbarseghyan and here: http://pavelbarseghyan.wordpress.com/. Pavel can be contacted at pavelbarseghyan@yahoo.com