Document downloaded from:

http://hdl.handle.net/10251/182471

This paper must be cited as:

Ferrer, A. (2020). Discussion of "A review of data science in business and industry and a future view" by Grazia Vicario and Shirley Coleman. Applied Stochastic Models in Business and Industry. 36(1):23-29. https://doi.org/10.1002/asmb.2516

The final publication is available at

https://doi.org/10.1002/asmb.2516

Copyright John Wiley & Sons

Additional Information

Corresponding author:
Alberto Ferrer
Multivariate Statistical Engineering Group, Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Camino de Vera s/n, 7A, 46022, Valencia, Spain

Discussion of "A review of data science in business and industry and a future view" by Grazia Vicario and Shirley Coleman

Grazia Vicario and Shirley Coleman are to be congratulated for writing this inspiring paper on such an important topic: clarifying the role of Data Science in business and industry.

I am more involved in problem solving in industry than in business, so it is expected that my comments are biased from my industrial experience.

In the following I would like to comment on some of the topics addressed by the authors and complement others.

What has triggered Data Science?

Data Science has emerged to cope with the so-called data tsunami. Modern industry and advanced technology is adopting the Industry 4.0 paradigm fostered by the Industrial Internet of Things (IIoT) connecting intelligent physical entities to each other and allowing complex equipment units to have embedded sensors and special modules (agents) providing connection to the monitoring center. Smartphones and Internet are contributing in a similar way in social networks, marketing, sales and finance. Processes not only produce goods or provide services but also data, and for the first time in history we have data everywhere. This is leading to the so-called Big Data environment, characterized by the four V´s: volume, variety, velocity and veracity. We live in a new era of digitalization where there is a belief that data contain useful information that has to be mined for helping the decision-making process.

But data sets are so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications (Wikipedia). For example, a large data set, such as a petabyte of data (10^{15} bytes), will not fit into an Excel spreadsheet, and cannot be stored on a standard laptop, so alternative approaches are needed. Most commercial analytics software programs are unable to process data sets of this size. In terms of analysis, standard scatterplots with such data sets will generally produce a large black blob—too much data to see what is going on. Many statistical hypothesis tests will show statistical significance simply because of the massive sample size—even if the differences are not relevant in practice. So clearly, alternative approaches and analysis methods are needed for processing and analyzing Big Data (Hoerl et al 2014).

To address these challenges a new discipline has emerged, and this is Data Science. There are many different issues to be solved in this new environment. These can be arranged in two groups: infrastructure (e.g., data collection, warehousing, integration, and cloud services) and analytics (e.g.*,* data analysis tools for visualization, processing and interpretation). Although Big Data environment has posed some new challenges in the analytics (with respect to traditional data analysis), most of the Big Data projects are nowadays focused on the infrastructure. But investing just in infrastructure is not synonymous of big business success. The focus is on data analytics to effectively extract such information to give organizations new insights about their products, customers and services and drive the decision-making process. This can be particularly valuable when it is critical to maintain quality and uptime, such as in process monitoring applications, by quickly detecting and diagnosing abnormal activities, predicting the time-to-failure of equipment units, or when rapid new products development is critical for company survival.

Data Science proposition: from Data to knowledge

Vicario and Coleman state that the Data Science proposition is the *transition from data to knowledge*. In fact some people call Data Science as Knowledge Discovery from Databases. I would like to stress two things: i) this proposition is not new, ii) in most Data Science applications knowledge is not synonymous of scientific knowledge (i.e., process understanding) but feature extraction or pattern recognition, and iii) the transition from data to (scientific) knowledge is not one-shot but an iterative approach.

As commented by Box (1976), R.A. Fisher devoted his entire life to the interpretation of empirical data by using the scientific method. So, the above proposition is not original of Data Science but comes from the use of statistics as a catalyst to learning.

Regarding my second comment, most of the Data Science applications are predictive oriented (i.e., obtaining predictors or classifiers with best performance). They are mainly focused on extracting features or discovering patterns from databases that, used as predictors in the algorithms developed, yield good predictions or classifications. Given the "black-box" nature of the algorithms used and the usual correlation of predictors, model interpretation to gain process understanding (i.e., scientific knowledge) of the problem addressed is not possible.

Therefore, this type of predictive applications is not suited for troubleshooting, and process improvement and optimization, critical goals in industry and technology. As discussed in the following, for fulfilling these goals the use of the scientific method (i.e., iterative inductive/deductive approach) is required.

What science is in Data Science?

Some people consider that the term "data science" does not have a broadly accepted definition; it is a "buzzword" that conveys little specific information. If we look at what people who claim to be data scientists actually do, data scientists are people with the following skills: i) processing of large data sets; ii) programming (Python, SQL, R, etc.); and iii) application of machine learning methods (e.g., neural networks, support vector machines, random forests, etc.). These are all existing skills; hence we could conclude that Data Science is not really a new science, per se, but rather an integration of existing sciences. Data Science motivation is to build something of importance to society from this skills list. Therefore, Data Science would be more accurately named "data engineering". Note that while many claim otherwise, statistics (skills in inference and probability-based models) are not included in this list (Hardin et al 2015).

In the Big Data community there are several messages that are creating a big debate in the scientific world. One comes from Chris Anderson 2008 paper entitled “The End of Theory: the Data Deluge that Makes the Scientific Method Obsolete” stating that because of the quantity and speed of data production, new technologies could now solve major scientific and industrial problems solely through empirical data analysis, without the use of scientific models, theory, experience, or domain knowledge. I guess that if this were true, this would imply that hereinafter there is no need on any kind of training but data scientist, and by “pushing the bottom” we would expect good results from analytics. A second worrying statement is that we no longer have to be fixated on causality and the world is shifting from causation to correlation (Mayer-Schönberger and Cukier 2013).

I guess that both statements are severely biased due to the typical problems that Data Scientist have traditionally addressed from information technology companies such as Google, FaceBook, YouTube, Amazon, and so on. These companies have been traditionally focused on building efficient algorithms for getting good predictive models to forecast the future. Given the abundance of data and the powerful computing resources available, hundreds of complex predictive models can be fitted at a low cost. Overfitting is addressed by ensemble methods that resample the data, fit several models and average the output of the fitted models. Models degradation are fought by updating them at the appropriate pace. For these predictive goals only correlation (not causation) is needed, and good predictions may be obtained just by playing with the (abundance) available data (by trial and error). In industry this approach is used for building soft sensors that forecast the value of a response variable difficult or expensive to measure, based on process variables sampled with high frequency and cheap to measure. In those situations there is really no new scientific knowledge (i.e., process understanding) acquired, only patterns and good predictions (i.e., the so-called high-level knowledge extracted from low-level data), but not science.

But as Breiman (2001) explains, in addition to forecasting, there is another culture of data analysis: modeling. This is focused on understanding the real world (i.e., process understanding) and assumes that the data are generated by a given stochastic model. For fulfilling this goal, theory, experience, domain knowledge and causation really matters, and scientific method is essential. This is the approach required for industrial troubleshooting, process improvement and optimization (and also for discovering new knowledge in basic sciences). Therefore, I strongly disagree with Anderson´s statement that faced with massive data, this approach to science (i.e., hypothesize, model, test) is becoming obsolete (Anderson 2008). What I consider in some way obsolete is the parametric statistical distribution-based (i.e., theoretical) approach of classical statistical tools. This will be discussed later on.

In fact, realizing the benefits of modeling even in typical forecasting applications can generate significant economic revenue for companies. For example, yet besides obtaining a good model to know the probability that a customer returns a loan on time (passive attitude), an attempt is made to understand the loan repayment process and discover the customer characteristics that increase the probability of repayment, then, instead of waiting for potential customers, you can actively search for them.

George Box (1976) wisely illustrated the inductive (questioning) / deductive (testing) iterative process of the scientific method. For this learning process to be effective there has to be two critical ingredients: subject matter knowledge (to ask relevant questions, formulate hypotheses and models, and deduce plausible consequences) and data (to evaluate the discrepancy between what hypotheses and models suggest should be so and what practice says is so), and a catalyst: statistics (Box and Liu 1999).

A paradigm change

In the DMAIC (i.e., Define, Measure, Analyze, Improve and Control) cycle, characteristic of Six Sigma process improvement methodology (Snee and Hoerl 2003) the Measure phase begins not by measuring but by asking questions, once the problem to be solved have been previously defined in the Define phase. The focus is on the problem to be solved and data is just one of the resources used to help accomplish our goals. Data have no meaning in themselves; they only have meaning within the context of a conceptual model of the phenomenon under study (Box et al 2005). In fact, data analysis without a problem is pure waste (Kempthorne 1980). The engine of the DMAIC cycle is the scientific method.

This is the classical approach in Statistics, following the Question-Data-Analysis (QDA) paradigm (Cao, 2019). Once the objective of the study is defined, questions give rise to collect (or generate) data and analyze it in an iterative scheme of several steps: model selection, model estimation and model validation, until the model is considered satisfactory and some answers can be found from the questions posed. The outcome of the analysis leads to new questions and so on.

Statistics as a scientific discipline was created in a data-scarce context: low number of variables with small sample sizes (usually more observations than variables that fit in a small Excel spreadsheet) that in most of the cases had to be measured (or generated) (i.e., there were no data available at the beginning of the project).

However, this has nothing to do with the data-rich context, typical of the Big Data era. The previous paradigm has changed to a Data-Questions-Analysis (DQA). Most of the time, a lot of data are already available before the questions are posed. This is the result of the launching of digitalization projects that many companies are undertaking. In this context all eyes are focused on the (huge) data sets, relegating to the background other very important aspects of the scientific method that are critical for success when the interest is not just prediction but process understanding and improvement.

Even in the new DQA paradigm, due to lack of statistical thinking, the fundamental principles well known to experienced statisticians are often ignored in some data science applications: i) critical evaluation of data quality; ii) integration of sound subject-matter knowledge; iii) development of an overall strategy for attacking the problem; and iv) sequential approaches versus “one-shot studies” (Hoerl et al 2014). These oversights have sometimes produced embarrassing errors like the breakdown of Google Flu Trend model (Lazar and Kennedy 2015), the Duke Genomics Center Debacle (Kolata 2011) or the bankruptcy of Lehman Brothers on September 15, 2008 (Wikipedia).

Big data vs Value data

The aim of any big data project must be to add business value – by enabling cost reductions, productivity gains or revenue increases (White 2016). That is the reason why a new fifth V: value, should be added to the previous four V´s list.

But do big data mean value data? Some people consider that massive data sets always reflect objective truth. This is what Crawford (2013) calls Data fundamentalism, and it is aligned with the idea that with enough data, the numbers speak for themselves. But this is not necessarily true because big data may be biased data (i.e., large-sized samples not representing the whole picture of the population under study). An illustrative example of biased big data is the Twitter data generated by Hurricane Sandy. The greatest number of tweets about Sandy came from Manhattan. This makes sense given the city's high level of smartphone ownership and Twitter use, but it creates the illusion that Manhattan was the hub of the disaster. Very few messages originated from more severely affected locations due to extended power blackouts, drained batteries and limited cellular access.

Therefore, quantity does not mean quality (i.e., more data do not mean better). Moreover, the larger the quantity and diversity of data the harder quality assessment is. Data Scientists must be involved in the data generating/collecting process and not adopt a passive role waiting for the data to come no matter: i) the science, engineering, and structure of the process or product from which the data were collected; ii) the collection process used to obtain data and prepare for analysis; and iii) how the measurements were physically obtained. These three issues summarize the *data pedigree*, a key information for the critical evaluation of data quality and relevance for solving a particular problem (Hoerl et al 2014). In fact, this issue is related to one of the four Big Data V´s: veracity, and it should not be ignored in Data Science projects.

Nature of data in Industry 4.0

The information technology revolution in Industry 4.0 has not caused only a change in the *number* of the variables (that in some cases is even higher than the number of observations) but also a change in the *nature* of the registered data. Nowadays it is possible to register data from customers, quality, process and even from equipment. There are a huge variety of different kind of sensors providing different type of signals: spectra (chemical signals), pressures, temperatures, flows, etc. (physical signals), pH, conductivity, dissolved oxygen, etc. (biochemical signals), electronic eyes (digital images), electronic noses and tongues (potentiometric signal), electronic ears (acoustic signals), and so on. These data mostly collected from daily production often exhibit high auto and cross correlation, rank deficiency, low signal-to-noise ratio, multi-stage and multi-way structure, and missing values. They are happenstance data (i.e., not generated from an experimental design) and, therefore, correlation does not mean causation.

Classical statistical assumptions as normality or independence, and the parametric statistical distribution-based (i.e., theoretical) approach of classical statistical tools are no longer appropriate in these data-rich environments. For example, in Statistical Process Control (SPC) the hypothesis that assignable causes of variation result in shifts of a particular subset of the parameter vector characterizing the statistical distribution of the monitored variables is hard to believe in practice. A holistic role of SPC for process improvement, and not only for monitoring, must be considered. The focus must be the process, not the model (Ferrer, 2014).

Process data in industry, although shares many of the characteristics represented by the four V´s (i.e., volume, variety, veracity and velocity), may not really be Big Data in comparison to other sectors such as social networks, sales, marketing and finance. However, the complexity of the questions we are trying to answer with industrial process data is really high, and the information that we wish to extract from them is often subtle. This info needs to be analyzed and presented in a way that is easily interpreted and that is useful to process engineers. Not only do we want to find and interpret patterns in the data and use them for predictive purposes, but we also want to extract meaningful relationships that can be used to improve and optimize a process (García-Muñoz & MacGregor 2016).

Latent variables models, such as principal component analysis–PCA (Jackson 2003) or partial least squares–PLS (Wold et al 2001), are especially suited for successfully addressing the characteristics 4 V´s of Big Data. They are compressing tools that easily handle the dimensionality and collinearity issues of the high volume of data. They can cope with the variety of data by using multiblock methods (Westerhuis 1998) for integrating data from different sources (data fusion). Latent variables models can be updated in real time to cope with the speed of data acquisition (velocity). Finally, these multivariate models are specially suited for outlier detection, missing and noisy data, typical issues for checking the veracity of the data.

Latent variable methodology exploits the correlation structure of the original variables by revealing the few independent underlying events (latent variables) that are driving the process at any time. This is done by projecting the information in the original variables down onto low-dimensional subspaces defined by a few latent variables. The multivariate scores are mathematically orthogonal and optimal summaries of the measured variables. The scores are also less noisy than the measured variables, because they are weighted averages (linear combinations) of the measured variables. Classical statistical assumptions (independency, normality and so on) can be reasonable for the scores, and therefore classical statistical tools are appropriate to analyze them. We could conclude that in the latent space, dealing with big data is easier than in the original variables space.

All models are wrong, some are useful

As commented by MacGregor (2018), to analyze historical data, one needs to make use of models, usually empirical – such as regression, machine learning (deep learning neural networks, support vector machines, random forest, etc.) or latent variable models (e.g., PCA or PLS). All models are wrong but some are useful (Box 1976). But all empirical models are not equally useful. Whether a model is useful depends on three issues: i) the objectives of the model (passive vs active); ii) the nature of the data used for the modeling (historical operating data vs data from design of experiments–DOEs); and iii) the regression method used to build the model (machine learning and classical regression vs latent variable models (PLS).

Regarding the objective, there are two major classes of models – those to be used for passive use and those to be used for active use. Models for passive use are intended to be used just to passively observe the process in the future. Such passive applications include predictive modelling and maintenance, pattern recognition and classification, and process monitoring, fault detection and diagnosis. For such passive uses one does not need or even want causal models, rather one wants to just model the normal variations common to the operating process. Historical data are ideal for building such models. On the other hand, models for active use are intended to be used to actively alter the process. Such active applications include gain causal information (i.e., process understanding) from the data, trouble-shooting, optimization and control. For active use one needs causal models. Causality implies that for any active changes in the adjustable variables in the process, the model will reliably predict the changes in the output of interest (MacGregor 2018).

To guarantee causality when using data-driven approaches, however, independent variation in the input variables is required (Box et al 2005). But, even if nowadays large amounts of historical data are available in most production processes, the variation in the inputs is commonly not independent (i.e., data are not obtained from a DOE that guarantees this independent variation in the inputs). Therefore, input-output correlation does not mean causation.

In this context classical predictive models (such as linear regression and machine learning models), proven to be very powerful in passive applications, cannot be used for process optimization (active use). They cannot be used for extracting interpretable or causal models from historical data for active use. With historical data, there are an infinite number of models that can arise from any of these machine learning methods, all of which might provide good predictions of the outputs, but none of which is unique or causal. Because the process variables are all highly correlated and the number of independent variations in the process is much smaller than the number of measured variables, one can get many of those models all using different variables and having different weights or coefficients on the variables that give nearly identical predictions. This does not allow for meaningful interpretations, even more so if the results come from averaging or voting on many models, such as in random forest (MacGregor 2018).

Latent variables methods, such as PLS regression, allow the analysis of large datasets with highly correlated data. Since they assume that the input (X) space and the output (Y) space are not of full statistical rank, they not only model the relationship between X and Y (as classical linear regression and machine learning models do), but also provide models for both the X and Y spaces. This fact gives them a very nice property: uniqueness and causality in the reduced latent space (this is the only space within which the process has varied) no matter if the data come either from a DOE or daily production process (historical data). By moving the latent variables one can reliably predict the outputs (Y) (i.e., the definition of causality in the latent space). But to move the latent variables one cannot just adjust individual X variables, but rather combinations of the X variables that define the latent variables as defined by the X space model (i.e., respecting the correlation structure of the model). This property makes them suitable for finding the combinations of input variables that guarantee appropriate outputs (if possible), such that the desired values for the quality attributes of interest for the final product are met (Jaeckle and MacGregor 2000, MacGregor et al 2015). Furthermore, since the initial number of variables involved is reduced to a smaller number of uncorrelated latent variables, the computational cost of any optimization problem in the latent space will decrease in comparison to the equivalent problem in the original space (Palací-López et al 2019, Tomba et al 2012).

Teamwork in interdisciplinary groups

The time to solve relevant problems just by one person has ended; complex problems require interdisciplinary teams and good communication between data scientists, process knowledge people (technical and operational staff) and managers.

This has tremendous implications in the way data scientists are trained. Data scientists have to understand the application domain of the projects they face, therefore in addition to the mathematics, statistics and computer science already present in their curricula, they have to be trained on the basics of different branches of science (chemistry, biology, physics, economy…) and technology (i.e., basic engineering). There are different ways to fulfill this: working in interdisciplinary projects with chemists, biologists, economists or engineers, or enrolling in a Bachelor in Science, Economy or Engineering, and later on a Master in Data Science.

Data analysts are not people who just analyze the given data; they have to be involved in all the process of collecting data. They are not people who just plug the data into the algorithms to get a prediction/classification. The problem to be solved and the relevant questions really matter. As commented by Box (1976), Fisher´s work made clear that the statistician´s job did not begin when all the work was over – it began long before it was started.

What is required for Data Science to be successful in Industry 4.0?

I would like to summarize in the following bullets what I consider should be the requirements for Data Science to be successful in Industry 4.0:

- Powerful storage and computational facilities, parallel processing management and display infrastructure.
- Statistical thinking strategy.
- Teamwork and basic domain knowledge of the problems to solve.
- Focus change:
  - From deductive (theoretical) to inductive (exploratory) reasoning. Be more exploratory minded. Do not just try to prove your a priori assumptions with hypothesis testing. Let yourself be surprised by what data can teach. Be open minded to discover something new from the data.
  - From parametric statistical distribution-based (i.e., theoretical) to distribution free methods by resampling methods such as bootstrap or jackknife, and permutation testing.
- Let the data speak for themselves. Do not select in advance the variables to be used but compressed, and prune later if desired.
- Programming skills and good knowledge of optimization algorithms, machine learning techniques and classical statistical tools (for passive applications) and latent variables models (for passive and active applications).
- Real-time analytics (fast decision making): perform time-critical analysis almost immediately after the data are generated.

References

Box GEP. Science and Statistics. *Journal of the American Statistical Association*. 1976;71(356):791-799.

Box GEP. Statistics as a Catalyst to Learning by Scientific Methods Part II – A Discussion. *Journal of Quality Technology*. 1999;31(1):16-29.

Box GEP, Hunter WG, Hunter JS. *Statistics for Experimenters: Design, Discovery and Innovation*. 2nd ed. Hoboken, NJ: John Wiley and Sons; 2005.

Breiman L. Statistical modelling: the two cultures. *Statistical Science*. 2001;16(3):199-231.

Cao R. Comments on: Data Science, big data and statistics. *Test.* 2019;28(3):664-670

Crawford K. The hidden biases in big data. *Harvard Business Review*, Cambridge. 2013. https://hbr.org/2013/04/the-hidden-bias-in-big-data. (Accessed December 26, 2019).

Ferrer A. Latent Structures-Based Multivariate Statistical Process Control: A Paradigm Shift. *Quality Engineering*. 2014;26(1):72-91.

García Muñoz S, MacGregor JF. Big Data. Success Stories in the Process Industries. *Chemical Engineering Progress*. 2016;112(3):36-40.

Hardin J, Hoerl R, Horton NJ, Nolan D, Baumer B, Hall-Holt O, Murrell P, Peng R, Roback P, Lang DT, Ward MD. Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician*, 2015;69(4):343-353.

Hoerl RW, Snee RD, De Veaux RD. Applying Statistical Thinking to 'Big Data' Problems. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2014;6:222-232.

Jackson JE. *A User's Guide to Principal Components*. New York:Wiley;1991.

Jaeckle CM, MacGregor JF. Industrial applications of product design through the inversion of latent variable models. *Chemometrics and Intelligent Laboratory Systems*. 2000;50:199–210.

Kempthorne O. The teaching of Statistics: Content versus Form. *The American Statistician*. 1980;34(1):17-21.

Kolata G. How bright promise in cancer testing fell apart. *The New York Times*, July 7, 2011. http://www.nytimes.com/2011/07/08/health/research/08genes.html. (Accessed December 26, 2019).

Lazar D, Kennedy R. What Can We Learn From The Epic Failure of Google Flu Trends? *Wired Online*, 2015. http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. (Accessed December 26, 2019).

MacGregor JF. Empirical Models for Analyzing "big" data-what´s the difference. In: Spring AIChE Conf., Orlando, Florida, USA; 2018

MacGregor JF, Bruwer MJ, Miletic I, Cardin M, Liu Z. Latent variable models and big data in the process industries, IFAC-PapersOnLine. 2015;28:520–524.

Mayer-Schönberger V, Cukier K. *Big Data: A Revolution that Will Transform How We Live, Work and Think*. Boston, MA: Eamon Dolan/Houghton Mifflin Harcourt; 2013.

Palací-López D, Facco P, Barolo M, Ferrer A. New tools for the design and manufacturing of new products based on Latent Variable Model Inversion. *Chemometrics and Intelligent Laboratory Systems*. 2019;194:103848.

Snee RD, Hoerl RW. *Leading Six Sigma: A Step by Step Guide Based on experience with GE and Other Six Sigma Companies*. Upper Saddle River, NJ: Pearson Education; 2003.

Tomba E, Barolo M, García-Muñoz S. General framework for latent variable model inversion for the design and manufacturing of new products, *Industrial Engineering Chemistry Research*. 2012;51:12886–12900.

Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*. 1998;12(5):301–321

White D, Big Data. What is it? *Chemical Engineering Progress*. 2016;112(3):32-35.

Wikipedia. Big Data. https://en.wikipedia.org/wiki/Big_data. (Accessed December 26, 2019).

Wikipedia. Lehman Brothers bankruptcy. http://en.wikipedia.org/wiki/Bankruptcy_of_Lehman_Brothers. (Accessed December 26, 2019).

Wold S, Sjöström M, Eriksson L. PLS-Regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. 2001;58:109–130.