Big Data, Data Loss and Observation Analysis

GERALD FRIZELLE, University of Cambridge
PHILIP WOODALL, University of Cambridge

With the increasing amount of data available to organisations, it is becoming problematic to know how and what data to perform an analysis on in order to obtain intelligent insights for the business. This paper proposes a new approach to address the problem of turning big data into something useful for the business, which utilises information theory. As a core part of the approach, we show how four types of data loss play a major role in providing a way to ensure that comparisons within a system or between systems are put on a common basis. Two example situations are presented where the approach was trialled in order to identify operational improvements in manufacturing-related organisations.

1. INTRODUCTION

Big data commonly refers to the increase in volume, variety, and velocity of data in today's organisations and society in general. The terms veracity and value are sometimes also included in the definition. It is the volume and value aspects, and the associated difficulty in extracting value from the ever increasing volumes of data available to organisations, which are the focus of this paper. In a review of big data Hilbert [2016] points out that three developments have allowed for the expansion of data: increasing telecommunication bandwidth, centralized and decentralized data storage systems, and digital computation capacities. He goes on to show the consequences in terms of data flow, information stock and information computation. In the case of data flow he notes that Google received 2,000,000 search queries per minute in 2012. Storage capacity is even more impressive; as long ago as 2012, a hard disc costing just $600 could store all the world's music [Kelly 2011]. As for computing power, this has grown two to three times faster than the capacity to store data. He concludes that there have been five key consequences: any digital network creates data almost as a by-product, random sampling has been supplanted by big data as the entire data set is already to hand, data are accessible in real time and, fourth, come in many guises. Finally, big data allows (*or should allow*) for various types of analysis to be carried out that were hitherto impossible. However the biggest obstacle is 'lack of understanding of how to use (big data) analytics to improve business' [De Mauro et al. 2014].

This paper addresses the problem of sifting and converting Big Data into something useful. Our approach, which we have chosen to call Observation Analysis (not to be confused with Observational Analysis [Rosenbaum 2010]), is to use ideas taken from Communication Theory to assess the information content of data sets; something greatly facilitated with the advent of Big Data. The approach assumes that we have an independent observer, observing the evolution of a system, and from his/her observations, conclusions about the system's behaviour can be drawn based on the data he/she has recorded. This contrasts with Communication Theory where the emphasis is in assessing the information content of messages passed between a sender and a receiver.

As an example of a system, this could be a manufacturing line containing a series of machines which produce the products. And a typical behaviour we may want to observe would be whether one of the machines is a bottleneck in the line.

One of the main findings from this research is that data loss plays a major role. In particular in providing a way to ensure that comparisons within a system or between systems are put on a common basis. That is not to imply that two systems with equal loss of data are comparable but rather that the converse is always true: that two systems with unequal loss of data are not comparable. By contrast, noise and its elimination, is one of the main targets of Communication Theory.

Four forms of data loss are revealed, three arising from the structure of the record ("record" being the term hereon used to refer to the data recorded by the system observer), in decreasing order: volume loss, range loss, granularity loss, and transactional losses (from errors in recording the data).

2. INFORMATION THEORY

Before looking at Observation Analysis, it is worth understanding some of the ideas behind Information theory [Shannon 1948] as it provides the underpinning for all that follows. The name of Claude Shannon is scarcely known and yet his work has probably had a greater impact on the lives of ordinary people than any of the other discoveries made by the 20th century giants of science and technology. He published a paper in 1948*, A mathematical Theory of Communication*, that a colleague described as '*probably no single work in this century has more profoundly altered man's understanding of communications..'* Slepian [1974]. All types of communications are built on his foundations. He had three great insights: that the average information content of a message—in fact the entropy content of the message (sometimes called Shannon entropy)—puts a lower limit on the extent to which a message can be compressed; that the upper limit on what can be transmitted is the bandwidth of the channel and that, provided there is sufficient bandwidth, transmissions can be almost noiseless.

Shannon's view of information is quite the reverse of what we commonly think of as information. He saw the information content of a message as the data that a sender of the message has and that the receiver requires. It is a little like a potential difference. Essentially there are five stages involved: the sender formulates the message, he/she then encodes it (in Shannon's day this was probably in Morse code), he/she then transmits it to the receiver who then decodes it and reads it. Encoding and decoding require a code book which both the sender and receiver possess.

Shannon defines information mathematically as :

$$\text{Information} = \log_2 1/p \qquad (1)$$

There are four questions to be asked about such an unusual formula:

(i) Why is a probability involved? The answer is that it represents the receiver's position at the point of transmission. If we recall that we are involved always in binary notation, then a single piece (bit) of information is either a zero or one. Since the receiver does not know which the sender has transmitted (and has no external source of information), he/she can only assume that each is equally likely to arrive. Thus putting this into equation (1) generates a value of 1, called a *bit.*

(ii) Why is it a reciprocal? This is because the rarer the event, the more information it contains. It is more newsworthy. Thus 'man bites dog' is news but the converse is not - attributed to John Bogart, the editor of the New York Sun [Bartlett and O'Brien 2012]. Also think of a scoop, i.e. publishing a rare event that competitor newspapers are unaware of.

(iii) Why is it a logarithm? Taking logarithms is a way of counting bits. If there are two possible outcomes - 0 and 1, then $\log_2 2^1 = 1$(bit)

(iv) Why do we take logarithms to the base 2? This is simply because we transmit bits. However, there is a more fundamental consequence: this is the point where qualitative and quantitative systems overlap. Thus the approach applies equally to qualitative data as to quantitative.

The most important concept underpinning Information theory is the average information or entropy. This is given by:

$$H(S) = \mathscr{E}(\log_2 1/p)$$

Where H is the entropy of some system S and $\mathscr{E}$ denotes mathematical expectation. H(S) has the following forms for discrete systems and the continuum, the latter known as differential entropy:

$$H(S) = -\sum_S p_i \log_2 p_i \text{ for countable states and } h(S) = -\int_S f(x)\ln f(x)\,dx \text{ for the continuum}$$

When we develop a model for analysing a system S, we will use the differential form. This is because the discrete form can always be derived from this but the converse is not true. This is particularly the case where we deal with qualitative systems, where the continuum has no meaning. The final record will be a quantised version of the continuum.

3. OBSERVATION ANALYSIS

The information content of observations.

Observation Analysis is about assigning quantitative values to categories based on what can be gleaned from observations made on systems. It is based on the ideas of Information Theory but with one important difference: instead of a sender and a receiver of messages there is an observer recording what they see. More importantly, the observer must not interfere with the system so as not to disrupt the very variety of behaviour they wish to observe. They eavesdrop on the system. Frizelle and Suhov [2008] call this 'treating the system as though it were *secretive*'. As a consequence, the observer may only draw conclusions from the record they make. We call this the Record Sheet $\mathscr{R}$.

However, the idea of using entropy as a measure of systems' performance has been around for some time. The work centred primarily on queuing behaviour. Several researchers have developed and applied the ideas to various applications [Efstathiou et al. 1999; Wu et al. 2013], mostly in supply chains and manufacturing.

Mutual information and taking observations

A second consequence of *secretiveness* is the absence of a pre-agreed codebook. For the same reason, the observer cannot consult the system about that either. In fact, the system may not yet be in existence! Instead the observer must construct his or her own 'codebook' i.e. the Record Sheet $\mathscr{R}$. Its structure is dictated by the transfer of Mutual Information between the system and the observer, as shown in equation (3). Mutual Information is the reduction in the uncertainty about a random variable *Z*, as a result of knowledge provided by a related random variable *Y*. A good example is searching for something. On 1 June 2009 flight AF447 crashed in mid Atlantic en route from Rio de Janeiro to Paris. Its last known position was a waypoint off the Brazilian coast. This gave an indication of the area to be searched. On June 2 wreckage was spotted in mid Atlantic pinpointing the location more accurately. The resulting reduction in the search area is a measure of the mutual information provided by the wreckage.

$$\max(I((X,R)) = h(X) - \min h(X \mid R) = \max h(R) - \min\ h(R \mid X) \qquad (3)$$

where I is the mutual information, which we wish to maximise, h(X) is the entropy of the system under observation and h(R) is the maximum information we can retrieved from the record. The conditional entropies represent data loss and noise

respectively. We wish to minimize both of these. The significance of the two conditional entropies becomes clear when we rearrange the terms of equation (3) and impose the secretiveness condition that we may only consult our record $\mathscr{R}$.

$$\max h(R)) = h(X) + \min h(R \mid X) - \min h(X \mid R) \quad (4)$$

This says that our record will comprise the information we retrieve from the system plus noise picked up in transmission less any data lost. Since secretiveness requires us to construct the Record Sheet in advance, the entire measurement process is in four stages: construct the record, observe the process, record the system's states and estimate the entropy rates. Figure 1 shows the model. Here the 'bandwidth' is the total number of observations N made on the system.

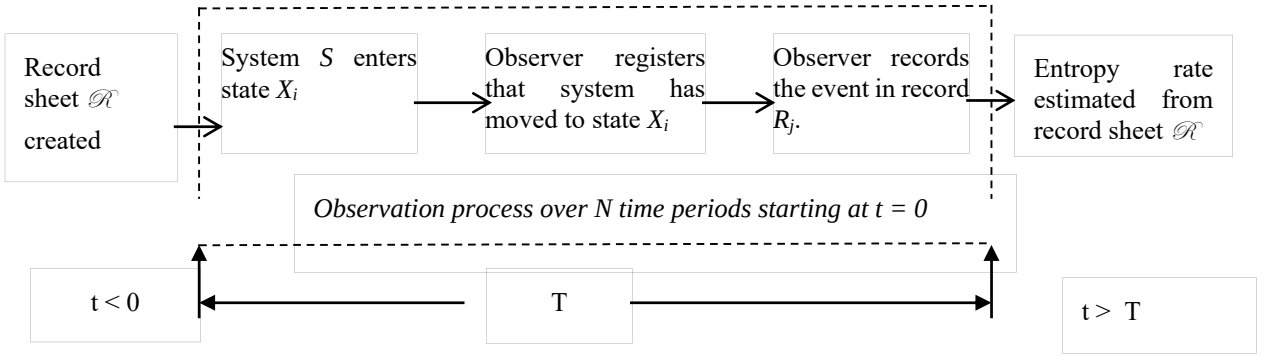

Figure 1 – The stages of the measurement process [Frizelle and Suhov 2008]

Where T is the elapsed time up to the point where the observer has sufficient data to estimate probabilities, starting at t = 0 and ending at t = T.

4. STRUCTURAL DATA LOSS

Constructing the Record Sheet t < 0

We use a modified version of equation (4) for the construction stage, because we want a record that unambiguously identifies events. Thus the Record Sheet has J columns, one for each state we wish to record and N rows for the observations. Only one of the I states of the system must ever appear in one of the J columns, recorded as a 'one' (with zeros in the other columns). Thus there is no noise at the construction stage and equation (4) simplifies to equation (5). Speaking technically, each row entry representing the state of the system at the nth observation, as a row vector of length J with a 'one' in column j representing the observed state, and zero elsewhere

$$\max h(R)) = h(X) - \min h(X \mid R) \quad (5)$$

Examining in detail how data can be lost at this stage, it turns out that three possibilities exist: in increasing order of magnitude they are *granularity losses, range losses* and *scope losses*. We call these three, structural data losses.

Granularity losses t < 0

Granularity losses simply represent the width Δ of the columns in the Record Sheet. When dealing with real variables, we reflect the support V of the system under observation, with a distance d in the Record Sheet, assuming both are finite. We then partition the distance d into J columns of equal width Δ: choosing columns of

variable widths would imply prior knowledge on the observer's part. For the same reason the probability of an observation being entered in one column must be equally likely as in any other column. This implies that the uniform distribution will best describe the situation, with entropy $\log_2 d$.

Hence;

$$\text{Log(Uniform)} = \log_2 d = \log_2 J + \log_2 \Delta \tag{6}$$

There are three things to be said about this equation (6) The first is that when J = 1, all data are lost for then $h(X|R) = h(X)$ in equation (5). The second observation is that the smaller the values of Δ, the smaller the data loss, with two caveats. The first is that the value of Δ must be uniform not just within the states of a single variable but, more importantly, when comparing variables or where more than one observer is involved. Otherwise one can have the anomalous situation where two observers arrive at different values for the same variable. The second caveat is that Δ must always be finite, hence the result only applies to finite supports. The final point is that there is a level of ambiguity about the term $\log_2 \Delta$, as it changes sign depending on whether it is greater or less than 1. We can circumvent the problem because the partitions have to be of equal width. We this achieve equality by repeatedly dividing each partition into two equal sub-partitions. In that case $J = 2^m$, where m is the number of times we carry out the partition process, starting with no partitions = 2^0. Thus we have

$$d = \Delta \log_2 m \qquad m \in \mathbb{N}$$

Taking logarithms to the base 2 of both sides yields:

$$\log_2 d - m = \log_2 \Delta \tag{7}$$

Equation (7) suggests that granularity is more about structure than distance. Sometimes we need to 'lose granularity'. For example, one exercise carried out by one of the authors involved locating lorries using GPS. This has an accuracy in meters with the result that no lorry ever appeared in the same place twice! As we only wanted to know which depot or road the lorry was taking, we had to 'lose granularity' to obtain meaningful results.

Range losses t < 0

Range losses reflect the range of each random variable i.e. the value of J for each random variable. While this is likely to be in a one to one correspondence with the I states observed for finite systems, this will, of course, not be so for systems with a semi-infinite number of states. In the latter case, the requirement is to choose a value of J such that data losses are minimised. This is helped by the fact that for a system with semi-infinite support, the entropy is maximised by the Pareto distribution for real numbered variables if the entropy is to remain finite.

The entropy of the Pareto distribution is given by:

$$H(\text{Pareto}) = \ln\left(\frac{V}{a}\right) + \frac{1}{a} + 1 \tag{8}$$

where V is the lower limit, equivalent to the upper value of V and corresponds to 'd' in the record. 'a' determines the shape of the curve and needs to be greater or equal to 3 in our case (as we require the mean and variance to be finite [Cover and Thomas 2006; Johnson and Kotz 1976]. Note that here the entropy is reduced by

taking ever higher values of 'a'. If we now express equation (8) in its quantized form, as required for the record we can write this as:

$$\begin{aligned} H^{\Delta}(\text{Pareto}) &= \log_2 d + \frac{4}{3} - \log_2 3 \\ H^{\Delta}(\text{Pareto}) &\approx H^{\Delta}(\text{Uniform}) - 0.25 \end{aligned} \tag{9}$$

It is clear that Pareto losses and Uniform losses are intimately connected and that the latter are slightly smaller than the former, becoming progressively so the greater the vale of 'a'.

Scope losses t < 0

In any practical exercise we deal with more than one random variable. Suppose we have K random variables, then we call K the scope of the exercise. Thus the relationship between the variables is a logical AND, the relationship between the states for any single random variable is an OR relationship. From this it should be clear that dropping the scope of the problem will cause a greater loss of data than dropping a state from the range - the former loss is polynomial, the latter logarithmic. Arguably, the most important role of data loss is to provide rules that ensure different data sets can be compared on a common basis; the losses generally cannot be eliminated because the observer cannot intervene, but they can be equated between and within data sets. Table 1 summarises the major findings on data loss and why they play such a central role, given that the need to treat the system as secretive means the losses cannot be recovered.

Observing the System 0 < t ≤ T – Transactional losses.

The observation process starts as soon as possible after the Record Sheet has been constructed and continues until sufficient data have been collected for the observer to have confidence in what he or she has found. This is straightforward sampling theory. However observation also incurs a fourth form of data loss; noise. At first sight this might be seen as a data gain from equation (4). However provided it can be spotted, the noisy entries must be dropped from the analysis. Secretiveness prevents us asking the system 'what it means'.

Summary of losses

Table 1 below summarizes the types of loss encountered when observing a system and recording what has been observed. As noted earlier it supplants noise as the key issue in this application of Information Theory.

Table I. Summary of data losses

Parameter	Data loss	Nature of loss	Comments
Granularity	Granularity loss	Structure of data set. Insufficient detail in state description	Loss from partitioning real number support set. Smallest structural loss. Key to ensuring comparability between records
Range	Range loss	Structure of data set. Values falling outside of range of a particular variable	States of a particular variable not included. Always happens for semi-infinite support.
Scope	Scope loss	Structure of data set. Variables not included in analysis	Key variable excluded from data set(s). Biggest structural source of loss
Data Transfer	Transactional loss	Data lost in recording	Errors in record. Found to be biggest source in practice. Usually appears as blanks or meaningless entries.

Estimating entropy - Operational complexity, goals, Tolerance and Relative turbulence t > T.

With sufficient data collected we can start to analyse the findings. This means estimating a value for H(R) - which may be a quantised version of h(R). We look at a measure on the system we call Operational Complexity and denote by $H^{(O)}$. This is the term h(X) in equation (4). To understand its structure, we introduce three other concepts that turn the measure from a rather abstract concept into a tool of practical utility. The first is the goal of the system. We assume that all the systems of interest are goal-seeking. This goal may be a single point or, more frequently, time related and is established when the Record Sheet is created. For quantitative random variables we have the simple and elegant result that, for a system whose goal is g, the associated entropy h is given by equation (10): where g is by its very nature deterministic:

$$h(X - g) = h(X) \tag{10}$$

The term Tolerance is a simple extension of the idea of a goal. Thus where the goal is a real number, there will always be a level of tolerance, representing the accuracy with which we can measure it. However there is value in extending the idea to both quantitative and qualitative systems. Now we split (6) into two categories, Tolerated (T) and Non- Tolerated (NT) states and redefine X as the joint random variable h((X-g), T/NT)

$$H^{(O)}\left\{\left[\left(X-g\right),\left(T/NT\right)\right]\right\} = -\left\{p(X^{T})\log_2 p(X^{T}) + p(X^{NT})\left[\log_2 p(X^{NT}) + \int_{V-g} f(x-g)\ln f(x-g)\,dx\right]\right\} \tag{11}$$

For technical reasons, the overall value of $H^{(o)}$ {[(X-g),T/NT]} is only of marginal interest. Primarily because it is a concave function. More important are the elements. First the term $p(X^{NT})$ is the percentage time the system is in Non-Tolerated states; typically when it may be running outside of its control limits. Next we may write the term under the integral sign as:

$$\int_{V-g} f(x-g)\ln f(x-g)\,dx = \mathscr{T}\int_{V-g} f^{(max)}(x-g)\ln f^{(max)}(x-g)\,dx \text{ where } \mathscr{T} = \frac{\int_{V-g} f(x-g)\ln f(x-g)\,dx}{\int_{V-g} f^{(max)}(x-g)\ln f^{(max)}(x-g)\,dx} \tag{12}$$

the function $f^{(max)(}(x)$ is the function that maximises the entropy for the particular form of support; the Uniform distribution for finite support and the Pareto for semi-infinite support. $\mathscr{T}$ is called Relative Turbulence and can be expressed as a percentage of the worst case. Finally the other key term is $p(X^{T})$, the proportion of time it is behaving predictability. This is usually interpreted as the efficiency of the operation.

Cost

One consequence of equation (12) is that it provides us with a way to cost non-compliance. For this we look at $p(X^{NT})$ from relative frequency calculations. This will either be based on the number of samples N or the total time T. If we are able to put a value on either of these, then the cost penalty follows. In the case of manufacturing, these costs are often the cost of holding either unwanted or unnecessary products. Other costs are the time lost to the enterprise or the cost of waste or recycling. Similar unwanted costs are found in supply chains. Added to these are hidden costs. One obvious example is paying too much to outside carriers to ensure prompt delivery. In order to ensure they deliver to time, they may claim excessive time for journeys to be on the safe side.

5. BIG DATA

The advent of so-called Big Data has brought two major benefits and one disadvantage to Observation Analysis. The first big benefit is that now many more databases are accessible for analysis. Moreover, the data are immediately available at almost negligible cost. The second is the sheer size of the files and hence the samples. It is not generally realised how big, samples need to be to have confidence in the value of the probabilities estimated. For instance a sample of around 2500 is needed to be 95% sure that the estimated probability of a binary decision e.g. whether an entry is a one or zero, is within 0.02 of its true value [Parzen 1992]. Big Data can furnish such samples.

The major disadvantage is that the Observer has had no control of either the structure of the data set or of its contents; he/she must make the most of what they have been given. This is the same, of course, for anyone else wanting to use the data. But even here there is an upside as the data have to be transformed into a unified format to allow the analysis to be carried out and to ensure that comparisons can be made between findings. It is therefore worth examining the data set beforehand to see if the exercise is worthwhile or even feasible. There are six fairly simple tests that can be carried out on any data set beforehand, to see if it is useable; Table 2 records the six.

Table II. Six tests for revealing issues with raw data

Nature of test	What it reveals	Comments
Are the required variables currently being measured?	If the answer is NO, then alternatives need to be considered (see next row)	In some situations standard results can be used instead, for example a well-established norm
If not, can required values be derived from other variables	There may be other ways to arrive at a satisfactory answer	One example is where a queue length is not recorded but entries to and exits from the queue, are.
Are variables measured in a consistent way?	If the variables involved, particularly where they come from different sources, are recorded in a consistent way.	All variables have to be in the same format, i.e. showing if a state is occupied or not. The question is then whether or not the existing data can be suitably transformed
Are there sufficient observations to carry out the analysis?	If the sample size is too small then none of the variables can be measured with sufficient confidence.	If the answer is NO, then more observations need to be made before analysis can begin.
Are there any variables for which the sample size is too small?	If the sample size for each variable is large enough to be significant	There will always be variables that fall into this category, with semi-open sets. The question is then whether the variable in question is central to the analysis
Are there obvious gaps in the data?	Lack of data or errors in recording	This has been found to be the most common problem but usually the easiest to spot. Typical examples include blanks in columns or obvious fillers such as 9999 type entries.

6. APPLYING THE ANALYSIS.

We now give two examples to illustrate how the availability of large data sets can open up new possibilities for data analysis, and data loss can facilitate making meaningful comparisons as well as providing a discipline for capturing and analysing the observations. The first was an analysis of a supply chain (for Nestlé Purina) carried out using existing data but before large data sets became readily available. The second, also on a supply chain, used data from a single large source.

Example 1: Nestlé Purina

This exercise was carried out to highlight which areas of a complex supply chain created the most problems, such as late delivery of parts etc. The analysis used data from various sources, but primarily files from a large distribution centre (DC) which acted as a hub in the chain [Robinson 2008]. The depot was supplied from three factories and supplied three supermarkets. The methodology looked at virtual queues represented by the inventory held in the warehouse. Inventory

movements were tracked along with the reasons. The complexity arises when actual movements differ from those planned. Such deviations are costly, either they involve holding too much inventory or, worse still, incur penalties through either being unable to supply what was ordered or not at all.

Figure 2 shows the results. The first thing to notice is that quite different phenomena can be represented on a common scale because the measure is generic. Thus unexplained variations in production at the factories are the biggest source of uncertainty followed by unintended variations in inventory in the distribution centre. By contrast deliveries to the customer were more predictable, unsurprisingly, as these are powerful supermarkets. Amongst the main elements, transportation comes out as the best performer.

Three lessons emerge from the exercise. The first was that an enormous effort was needed both to collect all the data and to interpret it, never mind having to then transform it into a useable structure. It was only made possible as a student undertook the project over a period of weeks. It is unrealistic ever to expect such a project to be carried out on a commercial basis. The second lesson was that the units of measure (bits) meant little to the company, even though the recipients acknowledged that it had provided the company with useful conformation of what thy had believed themselves. The last was that lumping together rather different variables, tended to detract from the final results. A more focused approach was needed. Put another way, the scope was too wide to provide anything other than general conclusions.

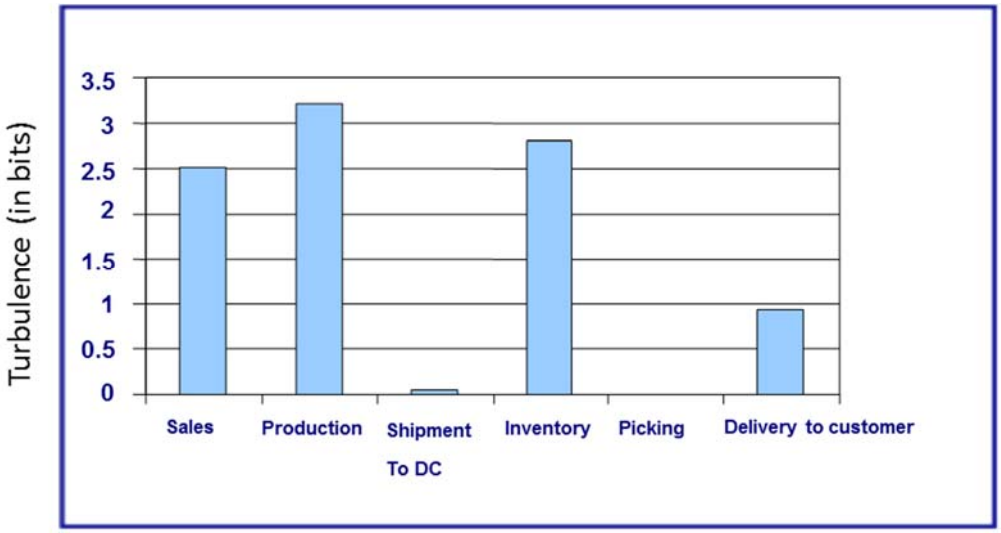

Figure 2: Sources of turbulence in a supply chain

Example 2: Chain Stores

The second example features delivery performance of Chain Stores from a number of stores to customers and the findings are shown in figure 3. The biggest change from the previous example is that here all the data were collected from an existing data source. The most significant cost was that of converting the data into the requisite format. Being a common source it now becomes possible to make meaningful comparisons by equating data losses.

For this example, a number of changes were made, and all to do with the way the findings were presented. The first difference from the earlier example is that *Turbulence* had now been supplanted by *Predictability*. This is given by

$$Predictability = 1 - \mathscr{T} \tag{9}$$

Where $\mathscr{T}$ is, once again *Relative Turbulence*. This change is motivated by the fact that the X axis will run from 0% to 100%, with 100% being totally predictable, making the scatter plot more intuitive. The second change is to introduce *Performance* from equation (7), in this case the proportion of time that the suppliers delivers on time (or within tolerance). Here, 100% means everything on time. The third change is that the size of the circles reflects the size of the supplier. It can be seen that larger suppliers are both less predictable and less punctual than the smaller ones. This may reflect the greater power invested in large suppliers. A final difference, not apparent from the diagram, is that focussing on one particular aspect of the supply chain, in this case punctuality, it is possible to slice the data in different but comparable ways; for example by stores to all customers or vice versa.

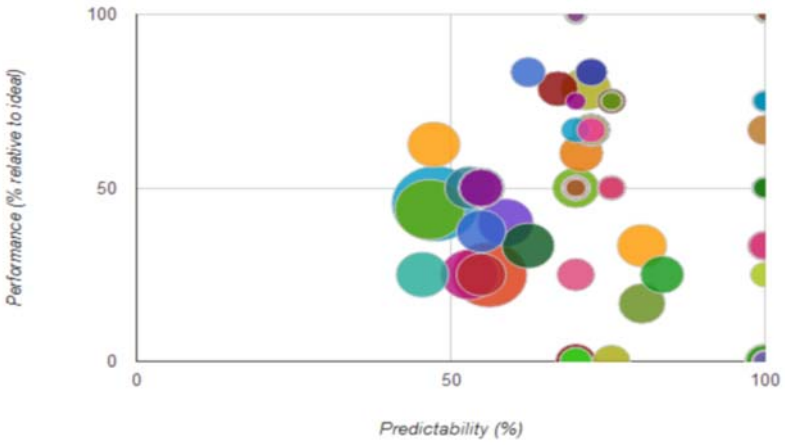

Figure 3: Delivery performance from suppliers to stores

7. CONCLUSIONS

Big data has opened new possibilities for carrying out analyses. However the sheer volume and variety available makes it difficult to decide where to start. Observation Analysis offers one possible solution as it takes a generic property of any data set, its information content, as its starting point. What emerges is that data loss is a key issue, not simply because there will be losses with any set but rather with the nature of the losses. Of the four ways that data can be lost, three—Scope, Range and Granularity—are linked to the structure of the set. The other—Transactional losses—is linked to errors incurred during the observational stage.

Because of the requirement to treat the system as though it were secretive, in general there is no option but to accept the loss. However, given the mathematical structure imposed by Information Theory, it is possible to mitigate the structural losses. However, of more importance are two other considerations that come out of the theory. The first is that the approach provides a basis on which to compare data. The second is that six tests can be carried out to see if a specific data set is worth using in the first place. There is one caveat. By virtue of the fact that the approach is generic it is possible to 'compare anything' and produce valid but meaningless results. Big data allows us to look at a specific aspect of a system, such as queues (including inventories) in manufacturing and punctuality in supply chains and try to evaluate cause and effect e.g. why are inventories so high and where are we performing least well in a particular chain.

REFERENCES

J. Bartlett and G. O'Brien. 2012. Bartlett's familiar quotations: a collection of passages, phrases, and proverbs traced to their sources in ancient and modern literature 18th ed., New York: Little, Brown, and Co.

T.M. Cover and J. Thomas. 2006. Elements of information theory 2nd ed., Hoboken, N.J: Wiley-Interscience.

A. De Mauro, M. Greco, and M. Grimaldi. 2014. What is Big Data? A Consensual Definition and a Review of Key Research Topics. (2014). DOI:http://dx.doi.org/10.13140/2.1.2341.5048

J. Efstathiou et al. 1999. Information Complexity as Driver of Emergent Phenomena in the Business Community. In Proceedings of the International Workshop on Emergent Synthesis. Kobe, Japan, 1–6.

G. Frizelle and Y. Suhov. 2008. The measurement of complexity in production and other commercial systems. Proc. R. Soc. Math. Phys. Eng. Sci. 464, 2098 (October 2008), 2649–2668. DOI:http://dx.doi.org/10.1098/rspa.2007.0275

M. Hilbert. 2016. Big Data for Development: A Review of Promises and Challenges. Dev. Policy Rev. 34, 1 (January 2016), 135–174. DOI:http://dx.doi.org/10.1111/dpr.12142

N. Johnson and S. Kotz. 1976. Discrete distributions, New York: Wiley.

K. Kelly. 2011. Keynote. (2011).

E. Parzen. 1992. Modern probability theory and its applications Wiley classics library ed., New York: Wiley.

M. Robinson. 2008. Complexity observed in supply chains. But can it be diagnosed? In 2008 IET Seminar on Complexity in Business. 1–18.

P. Rosenbaum. 2010. Design of observational studies, New York: Springer.

C.E. Shannon. 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 3 (July 1948), 379–423. DOI:http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x

D. Slepian. 1974. Key papers in the development of information theory, New York: IEEE Press.

Y.R. Wu, L.H. Huatuco, G. Frizelle, and J. Smart. 2013. A method for analysing operational complexity in supply chains. J. Oper. Res. Soc. 64, 5 (May 2013), 654–667. DOI:http://dx.doi.org/10.1057/jors.2012.63