Small Data, Mid data, Big Data vs. Algebra, Analysis, and Topology

Xiang-Gen Xia

I have been thinking about "big data" in the last a few years since it has become a hot topic. On most of the time I have been confused. Occasionally I thought I was clear but soon, I became confused again. I hope that I am now clear. Whenever one sees data, one may think that it is related to numbers and counting. In fact, nowadays, data is more general than numbers. However, when they are input to computers, they become bits and/or numbers. So, what is big data? When was it started? Where will it lead to? These questions may have different answers to different readers. To me, trained in mathematics, as a signal processing professor in an electrical engineering department, data is quite natural and I would like to try to give my answers to these questions below.

First of all, what is big data? Unfortunately, there is no precise mathematical definition for this concept. Big data or small data is relative. To see what big data is, let us first see what small data is. Everyone eats two apples per day in my family and my family consists of 4 people. So, my family eats 8 apples per day. This is small data and it is accurate. What is next? For example, my whole family including all relatives eat 400 apples per day. My neighbor's whole family including all their relatives eat 500 apples per day. Then, the total number of apples these two families eat will be no more than 900 apples per day. One might want to ask why it is not exactly the sum, 900, of the 400 and 500 apples. The reason is that these two families may have some members in common and some of them from one family may be married to another in the other family. In this case, although the total count may not be accurate but one can have an accurate upper bound. Is this small data or something else? I would like to think about it as mid data. Next it comes to the number of apples consumed in the world. How many apples do the people on the earth eat per day? To see this, one might say, let us make a table of the numbers of apples eaten per day for every country. OK, it is about 300 million for USA, 300 million for Japan, etc. Oops, how many apples do the people eat in North Korea per day? Sorry,

there is no trustable data available. Then, what do we do? Can we still count for the numbers of apples consumed per day for the whole world? No, but we may use some colors to mark the levels of the numbers for all the countries on a map. In this case, I would say it is big data, i.e., it is so big that no one can even estimate its volume but only get some high level indices.

In mathematics, there are mainly three subjects: algebra, such as high school algebra and abstract algebra etc.; analysis, such as real analysis and functional analysis etc.; and topology and geometry, such as algebraic topology and differential geometry. In my opinion, all these subjects are about counting and calculation, which is, of course, all mathematics is about. In algebra, one can count exactly. In analysis, one may not be able to count exactly but roughly or just estimate. One might want to ask where probability and statistics are. They belong to analysis since they belong to measure theory that belongs to real analysis. In topology, one is not able to count the whole thing that is not countable but one still wants to count. In this case, what can one do? OK, one can think of the whole thing as consisted of several “pieces” and then, one just counts for the number of “pieces”. The real question is what a “piece” is, which is topology and geometry about. It is kind of index and one may get it in the limiting case. If I am asked to make an analogy between mathematics and data classification, I would say that algebra corresponds to small data, analysis corresponds to mid data, and topology/geometry corresponds to big data.

Small Data and Algebra

As I said earlier, mathematics is about counting and calculation. In fact, calculation is a kind of counting. In many calculations, finding solutions of equations is always one of the most important tasks. Among finding solutions of equations, finding roots of polynomials is probably the most important. The fundamental theorem of algebra tells us that any non-constant single variable polynomial has at least one complex root. This means that any single variable polynomial equation can be solved with possibly complex numbers as solutions/roots. As one

knows that roots of a polynomial of degree less than 5 have closed forms in terms of the coefficients of the polynomial. However, for a polynomial of degree of 5 or higher, its roots may not have closed forms in terms of its coefficients, which was first mathematically proved by Galois and it is called Galois theory. In order to do so, Galois invented the concepts of group, ring and field, which led to the modern mathematics. The smallest field is the binary field {0, 1} and the largest field is the complex field **C** that is the set of all the complex numbers. The reason why **C** is the largest field is because every polynomial equation over the complex field can be solved already by the fundamental theorem of algebra. There are many kinds of subfields and extended fields, such as algebraic number fields etc. by including, for example, some roots of unity, i.e., $\exp(-2\pi \, \mathrm{j}/m)$ for some positive integer m, in the middle of {0,1} and **C**. After the complex field, mathematicians generalized **C** to quaternionic numbers that form, in fact, a domain, and also octonionic numbers. For example, a quaternionic number can be equivalently written as $\begin{pmatrix} x & y \\ -y^* & x^* \end{pmatrix}$, where x and y are two complex numbers. With these generalizations, mathematicians found that the most important property from all these structures is the norm identity:

$$\|x \bullet y\| = \|x\| \bullet \|y\| \tag{1}$$

for any two elements x and y in the domain of interest, where the dot stands for the multiplication in the domain or the real multiplication, and $\| \ \|$ stands for the norm used in the domain. In other words, the norm of the product of any two elements is equal to the product of the norms of the two elements. This is clear when x and y are two complex numbers but less obvious for other cases. A general design satisfying (1), as generalizations of complex numbers, quaternionic numbers and octonionic numbers, is called compositions of quadratic forms [1]. A [*k,n,p*] Hermitian composition formula is

$$\left(|x_1|^2 + \cdots + |x_k|^2\right)\left(|y_1|^2 + \cdots + |y_n|^2\right) = |z_1|^2 + \cdots + |z_p|^2, \tag{2}$$

where|| stands for the absolute value, $X=(x_1,\cdots,x_k)$ and $Y=(y_1,\cdots,y_n)$ are systems of indeterminates, $z_i=z_i(X,Y)$ is a bilinear form of X and Y. As an example, let *k=n=p=2* and $z_1=x_1y_1-x_2y_2,\quad z_2=x_1y_2+x_2y_1$. This corresponds to the following case. The product of the absolute values of two complex numbers is equal to the absolute value of the product of the two complex numbers: if $x=x_1+\mathrm{j}\,x_2$ and $y=y_1+\mathrm{j}\,y_2$ for real valued x_1,x_2,y_1,y_2, and $z=z_1+\mathrm{j}\,z_2=xy$, then $|z|=|xy|=|x||y|$. More designs on compositions of quadratic forms can be found in, for example, [2], which has found applications as space-time coding in wireless communications with multiple transmit antennas.

With this in mind, I would say that algebra is with the norm identity, where one is able to count precisely, the same as the first apple example mentioned earlier, where |2 4|=8=|2| |4| and also |500+400|=|500|+|400|, when the dot sign in (1) is the real multiplication and the real addition, respectively. This, in my opinion, corresponds to small data.

Mid Data and Analysis

In most cases, the norm identity (1) does not hold. Instead, it is the following inequality:

$$\|x\bullet y\|\le\|x\|\bullet\|y\| \tag{3}$$

for any two elements x and y in a set called space. This leads to the concept of a norm space, i.e., if there is an operation $\|\ \|$ on a set that satisfies (3) for any two elements x and y in the set, this set with some additional scaling property is called a norm space. It is the key for functional analysis or analysis, including measure theory and/or probability theory and statistics. In (3), the dot sign is the addition + in this case and (3) is correspondingly called the triangular inequality. In my opinion, the difference between algebra and analysis is the difference between the norm equality and the norm inequality shown in (1) and (3), respectively. It is the same as the second apple example mentioned before, where

$$\begin{aligned}
&\left\|\{400 \text{ apples in one family}\} \cup \{500 \text{ apples in another family}\}\right\| \\
&\le \left\|\{400 \text{ apples in one family}\}\right\| + \left\|\{500 \text{ apples in another family}\}\right\| \\
&= 400 + 500 = 900,
\end{aligned}$$

where the dot sign in (3) corresponds to the union of two sets and the real addition, respectively. This, in my opinion, corresponds to mid data.

Another observation about the above norm inequality is that the dot operation in (3) for two elements x and y can be thought of as a general operation as we have seen above for different cases of the dot sign. The norm inequality (3) becomes the triangular inequality when the dot is + as mentioned earlier. When the dot is a true product of two elements, such as matrix multiplication of two matrices, the inequality (3) is the conventional norm inequality. The norm inequality (3) becomes the Cauchy-Schwarz inequality when the dot is the inner product:

$$\left|\int_a^b f(t)g(t)\,dt\right| \le \left[\int_a^b |f(t)|^2 dt\right]^{1/2}\left[\int_a^b |g(t)|^2 dt\right]^{1/2}, \qquad (4)$$

where the equality holds if and only if functions $f(t)$ and $g(t)$ are linearly dependent, i.e., $f(t)=cg(t)$ or $g(t)=cf(t)$ for some constant c. From this observation, almost all inequalities can be derived from the norm inequality (3). Many fundamental results are derived by the Cauchy-Schwarz inequality (4), i.e., the norm inequality. For example, the Cauchy-Schwarz inequality leads to the conclusion that the optimal linear time-invariant (LTI) filter to maximize the output signal-to-noise ratio (SNR) is and only is the filter that matches to the signal, i.e., the matched filter. It has been extensively used in radar and communications etc. Another application of the Cauchy-Schwarz inequality is the proof of the Heisenberg Uncertainty Principle (HUP). It says that the product of the time width and the bandwidth is lower bounded by ½ and the lower bound is reached if and only if when the signal is Gaussian, $a\exp(-bt^2)$ for some constant a and some positive constant b. As a simple consequence of the HUP, one is not able to design a signal that has infinitely small time width and infinitely small bandwidth simultaneously. Otherwise, one would be able to design as many orthogonal signals as possible in any finitely

limited area of time and frequency, i.e., it would have infinite capacity for communications over any finite bandwidth channel. One can see that both results have played the key roles in science and engineering in recent history.

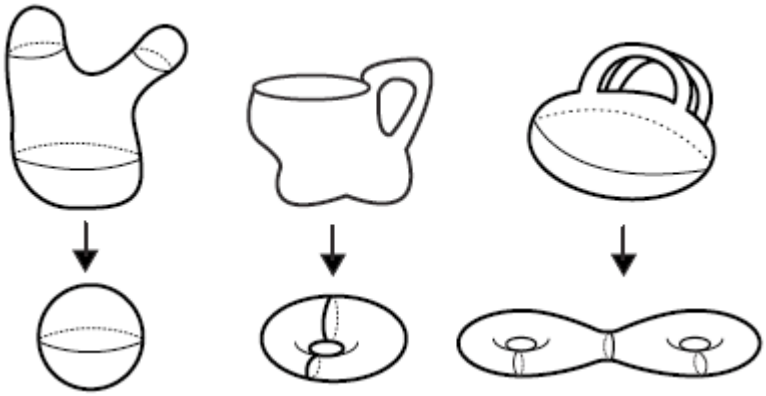
Big Data and Topology

Fig. 1. Massive groups of fish [4]

Fig. 2. Massive groups of birds [5]

When one sees many huge groups of fish moving in the ocean as shown in Fig. 1 (or huge groups of birds flying on the sky as shown in Fig. 2), one may not be able to count exactly or estimate approximately how many fish out there. One may just count how many disconnected groups of fish. If one treats each group as a visible hole of the ocean, it is the concept of genus, one of the key concepts in topology, where the number of holes (or fish groups in our case here) on an object (ocean) is the genus of the object [6]. More precisely, the genus of a connected, orientable surface is an integer representing the maximum number of cuttings along non-intersecting closed simple curves without rendering the resultant manifold disconnected [7]. In the above definition, "cutting" is understood as the conventional cutting by a knife. Some simple examples are shown in Fig. 3. Another simple but more mathematical way to understand it is as follows. If any loop (a simple closed curve) on a surface (it could be a solid object, such as a solid ball), such as the sphere shown in Fig. 3 (a), can be continuously (on the surface, or inside the solid object) contracted/tightened (also called continuously transformed) to a point on the surface, then the surface has genus 0. For the torus shown in Fig. 3 (b), it is impossible to do so because if one picks up a simple loop around the hole, then this loop cannot be

continuously, on the surface, contracted to any point on the surface. However, if the torus is cut in the middle (with one cut) as shown in Fig. 3 (b) (note that there are two cuts total shown in Fig. 3 (b)), then it is not possible to have a loop around any hole and thus any simple loop can be continuously contracted on the surface to a point. In this case, the torus has genus 1, i.e., one and only one cut is used/needed to do so. As shown in Fig. 3, genus is a topologically invariant variable in the sense that two shapes may look totally different but they have the same genus, where the objects in the first row have 0, 1 and 2 holes, and are topologically equivalent to those in the second row, respectively.



(a) Genus 0 (b) Genus 1 (c) Genus 2

Fig. 3. Genus of an object [6]

A possible application of the above concept of genus in topology would be in the current investigations of big data representation that plays an important role in big data analysis. One of the efficient ways to represent a big data is to use a proper tensor product [8]. When a big data is too big and its tensor product representation is properly used, it may be treated as a multi-dimensional massive object. In this case, its topological properties, such as genus, may become simple but an important feature.

As we have discussed above, when an object is too complicated or too massive, the indices and/or the topologically invariant variables, such as the genus, i.e., the number of holes, and/or disconnected pieces, come to the picture. These topologically invariant variables may be obtained by taking a limit when some parameters go infinity, which may “smooth” out all the

uncertainties/unknowns caused by the massiveness and make the calculations possible. In other words, taking a limit may simplify the calculation. One simple example is the calculation of the integration of a Gaussian function. For any finite real values a and b and a positive constant α, $\int_a^b e^{-\alpha t^2} dt$ does not have a simple closed form, while $\int_{-\infty}^{\infty} e^{-\alpha t^2} dt$ does. Another example is the diversity and multiplexing tradeoff (DMT) obtained by Zheng and Tse for multi-input and multi-output (MIMO) antenna systems in wireless communications, which becomes a necessary parameter in designing a MIMO wireless communication system. Let R be the transmission rate in bits per second per Hz. Let r be the normalized rate $r=R/\log(SNR)$, where SNR is the channel SNR. When SNR is huge, one may expect that R is huge as well by Shannon's channel capacity formula that is about $\log(SNR)$, i.e., massive data (or big data) can be transmitted through the channel. In this case, counting R may be not possible, while counting r becomes more reasonable, where r is called the multiplexing gain. Let P_e be the error probability at the receiver of a MIMO modulation scheme with transmission rate R. Let

$$d(r) = -\lim_{SNR\to\infty} \frac{\log(P_e)}{\log(SNR)}. \tag{5}$$

Then, $d(r)$ is the index of the negative exponential of the error probability P_e and called the diversity gain:

$$P_e \approx SNR^{-d(r)}, \tag{6}$$

when SNR is large enough. Zheng and Tse [3] obtained the following well-known DMT tradeoff:

$$d(r)=(m-r)(n-r),$$

where m and n are the numbers of transmit and receive antennas, respectively. One can see that both r and $d(r)$ are sort of indices and they are only meaningful when SNR approaches infinity, i.e., in a massive transmission rate case, or big data case. This is the case, when it is impossible to count one element by one element for a massive data, one needs to sort out its index, such as, exponentials and/or genus, in some way to describe and/or extract features from the massive/big data. It, I think, belongs to topology in mathematics. Thus, in my opinion,

topology in mathematics corresponds to big data, where it is impossible or not necessary to count one element by one element.

Summary and Discussion

In summary, I would like to say that, small data corresponds to algebra, mid data corresponds to analysis, and big data corresponds to topology in mathematics. Was big data started when "big data" was named? Of course not. Big data has existed for a long time, as massive groups of fish move in the ocean, massive groups of birds fly on the sky, and/or a massive number of people on ground travel around the world. Nowadays, massive bits are transmitted through both wired and wireless channels called internet. The key is how to get some indices/trends/patterns from these massive data or/and how to find a needle in the ocean. What will big data lead to tomorrow? Or in other ways to ask: how deep can we go towards infinity tomorrow? Or how fast will a computer be tomorrow?

References

[1] D. B. Shapiro, *Compositions of Quadratic Forms*, New York, De Gruyter, 2000.

[2] K. Lu, S. Fu, and X.-G. Xia, "Closed-form designs of complex orthogonal space-time block codes of rate (k+1)/(2k) for 2k-1 or 2k transmit antennas," *IEEE Trans. on Information Theory*, vol. 51, pp. 4340–4347, Oct. 2005.

[3] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. on Information Theory*, vol. 49, pp. 1073–1096, May 2003.

[4] http://cir.institute/collective-intelligence, Collective Intelligence Research Institute.

[5] http://becausebirds.com/2014/07/29/how-do-bird-flocks-work/, becauseBirds.com.

[6] https://www.learner.org/courses/mathilluminated/units/4/textbook/03.php, Mathematics Illuminated, Annenberg Learner.

[7] https://en.wikipedia.org/wiki/Genus_(mathematics), Wikipedia.

[8] A. Cichocki, "Era of big data processing: A new approach via tensor networks and tensor decompositions," arXiv: 1403.2048v4 [cs.ET], http://arxiv.org/abs/1403.2048v4, 2014.

Author

Xiang-Gen Xia (xianggen@gmail.com) is Charles Black Evans Professor in the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA. His main research interests include wireless communications and radar signal processing. He is a Fellow of the IEEE.