Xiang-Gen Xia

Small Data, Mid Data, and Big Data Versus Algebra, Analysis, and Topology

Big data has become a hot topic during the last few years. But at times, its meaning has been quite confusing. I hope that through sharing my thoughts in this article, we can have a better understanding of what big data is.

Whenever you see data, you may think that it is related to numbers and counting. In fact, today, data is more general than numbers. However, when data are input in computers, they become bits and/or numbers.

So, what is big data? When was it started? Where will it lead? 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 so in this article, I provide my answers to the aforementioned questions.

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 look at what small data is. Each person in my family, which consists of four people, eats two apples per day. Therefore, my family eats eight apples per day. This is small data and is accurate. What is next? For example, my whole family, including all relatives, eats 400 apples per day. My neighbor's whole family, including all of their relatives, eats 500 apples per day. Then, the total number of apples these two families eat will be no more than 900 apples per day. You might want to ask why it is not exactly the sum, i.e., 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, the total count may not be accurate, but you can have an accurate upper bound. Is this small data or something else? I would like to think of 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 find out, one might say, let us make a table of the numbers of apples eaten per day for every country. It is approximately 300 million for the United States, 300 million for Japan, etc. Oops, how many apples do the people eat in North Korea per day? Unfortunately, there is no trustworthy data available. So, what do we do? Can we still count 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 of the countries on a map. In this case, I would consider it big data, i.e., it is so big that no one can even estimate its volume but can only get some high-level indices.

In mathematics, there are mainly three subjects: algebra, e.g., high school algebra and abstract algebra; analysis, e.g., real analysis and functional analysis; and topology and geometry, e.g., algebraic topology and differential geometry. In my opinion, all of these subjects are about counting and calculation, which is, of course, all that mathematics is about. In algebra, you can count exactly. In analysis, you may not be able to count exactly but roughly or just estimate. You might want to ask, where are probability and statistics? They belong to analysis since they belong to measure theory, which belongs to real analysis. In topology, you are not able to count the whole thing, but one still wants to count. In this case, what can be done? You can think of the whole thing as consisting of several pieces and then just count for the number of pieces. The real question is: what is a piece, and what is topology and geometry about? It is a kind of index that you may get 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.

In mathematics, there are mainly three subjects: algebra, e.g., high school algebra and abstract algebra; analysis, e.g., real analysis and functional analysis; and topology and geometry, e.g., algebraic topology and differential geometry.

Small data and algebra

As discussed previously, mathematics is about counting and calculation. In

Digital Object Identifier 10.1109/MSP.2016.2607319
Date of publication: 11 January 2017

fact, calculation is a type of counting. In many calculations, finding the solutions of equations is always one of the most important tasks. Among finding the solutions of equations, finding the roots of polynomials is probably the most important. The fundamental theorem of algebra tells us that any nonconstant single-variable polynomial has at least one complex root, which means that any single-variable polynomial equation can be solved with possibly complex numbers as solutions/roots. We know that roots of a polynomial of a degree lower than five have closed forms in terms of the coefficients of the polynomial. However, for a polynomial of a degree of five or higher, its roots may not have closed forms in terms of its coefficients, which was first mathematically proven by Galois and is, therefore, called the *Galois theory.* To do so, Galois invented the concepts of group, ring, and field, which led to modern mathematics. The smallest field is the binary field {0, 1}, and the largest is the complex field C that is the set of all 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, by including, e.g., 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 as well as 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 of these structures is the norm identity

$$\| x \bullet y \| = \| x \| \bullet \| y \| \qquad (1)$$

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 is 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

$$\begin{aligned}(|x_1|^2 + \cdots + |x_k|^2)(|y_1|^2 + \cdots + |y_n|^2) \\ = |z_1|^2 + \cdots + |z_p|^2 \qquad (2)\end{aligned}$$

where | | stands for the absolute value, $X = (x_1, \ldots, x_k)$ and $Y = (y_1, \ldots, y_n)$ are systems of indeterminates, and $z_i = z_i(X, Y)$ is a bilinear form of X and Y. As an example, let $k = n = p = 2$ and $z_1 = x_1 y_1 - x_2 y_2,\ \ z_2 = x_1 y_2 + x_2 y_1$. 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, i.e., if $x = x_1 + jx_2$ and $y = y_1 + jy_2$ for real-valued x_1, x_2, y_1, y_2 and $z = z_1 + jz_2 = xy$, then $|z| = |xy| = |x||y|$. More designs on the compositions of quadratic forms can be found in [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 you are able to count precisely (the same as the first apple example mentioned previously), where $|2 \cdot 4| = 8 = |2| \cdot |4|$ and $|500 + 400| = |500| + |400|$, 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:

$$\| x \bullet y \| \leq \| x \| \bullet \| y \| \qquad (3)$$

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 this case, in (3), the dot sign is the addition +, 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 previously, where

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

where the dot sign in (3) corresponds to the union of two sets and the real addition, respectively. I feel that this 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 previously. When the dot is a true product of two elements, such as the 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

$$\begin{aligned}&\left| \int_a^b f(t) g(t)\, dt \right| \\ &\quad \leq \left[\int_a^b |f(t)|^2 dt \right]^{1/2} \left[\int_a^b |g(t)|^2 dt \right]^{1/2}, \qquad (4)\end{aligned}$$

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

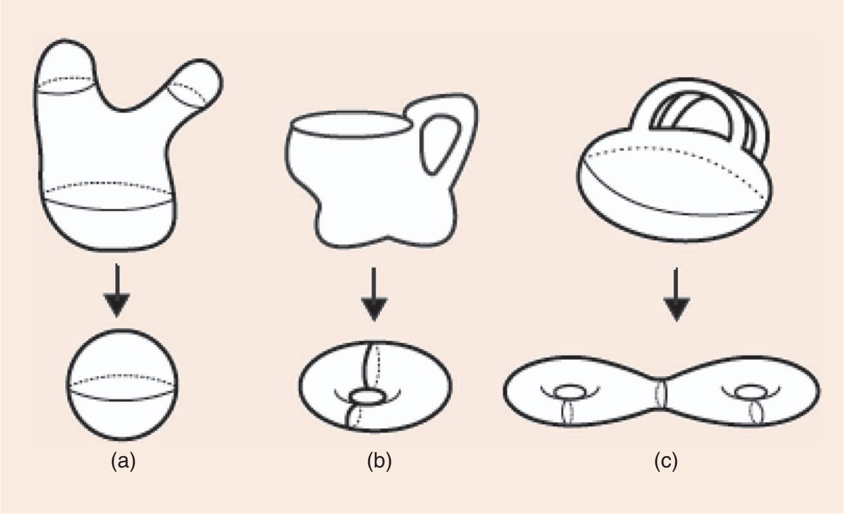

FIGURE 1. The genus of an object; (a) genus 0, (b) genus 1, and (c) genus 2.

(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 filter to maximize the output signal-to-noise ratio 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. 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 one half, and the lower bound is reached if, and only if, the signal is Gaussian, i.e., $a \exp(-bt^2)$ 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 an infinitely small time width and infinitely small bandwidth simultaneously. Otherwise, a person 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 key roles in science and engineering in recent history.

Big data and topology

When a person sees several large groups of fish moving in the ocean (see http://cir.institute/collective-intelligence) or large groups of birds flying in the sky (see http://becausebirds.com/2014/07/29/how-do-bird-flocks-work), he or she may not be able to count exactly or estimate approximately how many fish or birds are there. One may just count how many disconnected groups of fish. If a person treats each group as a visible hole of the ocean, it is the concept of genus, i.e., one of the key concepts in topology, where the number of holes (or fish groups in this case) in an object (i.e., the ocean) is the genus of the object. More precisely, the genus of a connected, orientable surface is an integer representing the maximum number of cuttings along nonintersecting, closed simple curves without rendering the resultant manifold disconnected [4]. In the aforementioned definition, cutting is understood as the conventional cutting by a knife. Some simple examples are shown in Figure 1. Another simple, but more mathematical, way to understand it is as follows. If any loop (i.e., a simple closed curve) on a surface (a solid object, such as a solid ball), such as the sphere shown in Figure 1(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 Figure 1(b), it is impossible to do so because, if one picks up a simple loop around the hole, this loop cannot be continuously contracted to any point on the surface. However, if the torus is cut in the middle with one cut, as shown in Figure 1(b) [note that there are two cuts total shown in Figure 1(b)], then it is not possible to have a loop around any hole; 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 Figure 1, 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 zero, one, and two holes, and are topologically equivalent to those in the second row, respectively.

A possible application of the aforementioned 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 efficient way to represent big data is to use a proper tensor [5]. When big data is too big and its tensor representation is properly used, it may be treated as a multidimensional massive object. In this case, its topological properties, such as genus, may become simple but is an important feature.

What is big data? There is no precise mathematical definition for this concept. Big data or small data is relative.

As we have discussed previously, 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 to infinity, which may smooth out all the uncertainties or 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 [3] for multiple-input, multiple-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/second/hertz. Let r be the normalized rate $r = R/\log(\mathrm{SNR})$, where SNR stands for signal-to-noise ratio and 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 $\log(\mathrm{SNR})$, 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

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

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

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

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

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

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, and 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. I think this 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.

I consider that small data corresponds to algebra, mid data corresponds to analysis, and big data corresponds to topology in mathematics.

Summary and discussion

In summary, I consider that small data corresponds to algebra, mid data corresponds to analysis, and big data corresponds to topology in mathematics. Was big data started when it 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 in the sky, and/or a massive number of people on the ground travel around the world. Today, massive bits are transmitted through both wired and wireless channels called the *Internet*. The key is how to get some indices, trends, or patterns from these massive data and/or how to find a needle in the ocean. What will big data lead to tomorrow? Or, how deep can we go toward infinity tomorrow? Or, how fast will a computer be tomorrow?

Author

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

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. Inform. 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. Inform. Theory*, vol. 49, pp. 1073–1096, May 2003.

[4] Wikipedia. Genus. (2016). [Online]. Available: https://en.wikipedia.org/wiki/Genus_(mathematics)

[5] A. Cichocki. (2014). Era of big data processing: A new approach via tensor networks and tensor decompositions. arXiv: 1403.2048v4 (cs.ET) [Online]. Available: http://arxiv.org/abs/1403.2048v4

SP

ERRATA

Reference [8] in the "SP Education" column of the November 2016 issue of *IEEE Signal Processing Magazine* [1] was published missing a URL. We apologize for any confusion this may have caused. The corrected reference is shown in [2].

Digital Object Identifier 10.1109/MSP.2016.2636258
Date of publication: 11 January 2017

References

[1] E. Richter and A. Nehorai, "Enriching the undergraduate program with research projects," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 123–127, Nov. 2016.

[2] BCI hand control download [Online]. Available: http://classes.engineering.wustl.edu/ese497/BigFiles/IpsiHand_v5.wmv

SP