Education and ...
Big Data versus Big-But-Buried Data

forthcoming in Lane, J.E., *Building a Smarter University*

Elizabeth Bringsjord • Selmer Bringsjord*

0201141900NY

Abstract

The technologized world is buzzing about "big data," and the apparent historic promise of harnessing such data for all sorts of purposes in business, science, security, and — our domain of interest herein — education. We distinguish between big data *simpliciter* (BD) on the one hand, versus big-but-buried (B^3D) data on the other. The former type of data is the customary brand that will be familiar to nearly all readers, and is, we agree, of great importance to educational administrators and policy makers; the second type is of great importance to educators and their students, but receives dangerously little direct attention these days. We maintain that a striking two-culture divide is silently emerging in connection with big data: one culture prudently driven by machine-assisted analysis of BD; and the second by the quest for acquiring and bestowing mastery of B^3D, and by the search for the big-but-buried data that confirms such mastery is in place within a given mind. Our goal is to introduce, clarify, and contextualize the BD-versus-B^3D distinction, in order to lay a foundation for the integration of the two types of data, and thereby, the two cultures. Along the way, we discuss the future of data analytics in light of the historic Watson system from IBM, and the possibility of human-level machine tutoring systems, AI systems able to teach and confirm mastery of big-but-buried data.

*The second author acknowledges, with deep gratitude, generous support provided by IBM to think about big data systematically, in connection with the seminal Watson system. The second author is grateful as well for (i) data and predictive analysis (of the big *simpliciter* variety) regarding student performance in calculus at RPI, provided by IR expert Jack Mahoney, and (ii) enlightening conversations about big-but-buried data and (differential and integral) calculus with Thomas Carter.

Contents

List of Figures

List of Tables

1 Introduction

The technologized world is of course buzzing about "big data," and the apparent promise of harnessing such data for all sorts of purposes in business, science, security, and — our domain of interest herein — education. We distinguish between big data *simpliciter* (BD) on the one hand, versus big-but-buried (B^3D) data on the other. The former type of data is the customary brand that will be familiar to nearly all readers, and is, we agree, of great importance to educational administrators and policy makers; the second type is of great importance to educators and their students, but is dangerously overshadowed by attention paid these days to the first type. Part of this danger derives from the fact, explored below, that while big-but-buried data is elusive, and while technology to exploit it is expensive and still primitive, B^3D is absolutely central to first-rate learning and teaching.

One of the hallmarks of big data *simpliciter* is that the data in question, when measured against some standard yardstick (e.g., the **byte**, which is eight **bits** of data, where each bit is 0 or 1), is exceedingly large. For instance, internet traffic per month is known to now be well over 20 exabytes ($= 20 \times 10^{18}$ bytes); hence an attempt to enlist software to ascertain, say, what percentage of internet traffic pertains directly to either student-student or student-teacher communication connected to some formal course would be a BD task. Or, more tractably, if one used R[1] to ascertain what percentage of first-year U.S. college students in STEM disciplines graduate in those disciplines as correlated with their grades in their first calculus course, one would be firmly focused on BD. We find it convenient to use a less pedantic yardstick to measure the size of some given collection of data. One nice option in that regard is simply the number of discrete symbols used in the collection in question. We are sure the reader will immediately agree that in both the examples of BD just provided, the number of symbols to be analyzed is staggeringly large.

Big-but-buried data is very, very different. What data does one need to master in order to thrive in the aforementioned calculus course, and in those data-intensive fields (e.g., macroeconomics) that make use of calculus (and, more broadly, of real analysis) to model vast amounts of BD? And what data does a calculus tutor need in order to certify that her pupil truly has

[1] R is by far the dominant software in the world used for all manner of statistical computing, stands at the heart of the "big-data" era, and is free. R can be obtained at: http://www.r-project.org. To start having fun with R in short order, we recommend (Knell 2013). With R comfortably on hand, those wishing an introduction to basic statistical techniques essential for analytics of BD, can turn to the R-based (Dalgaard 2008).

mastered elementary, single-variable calculus? In both cases, the answers exhibit not BD, but rather B^{3}D. For example, one cannot master even the first chapter of elementary calculus unless one has mastered (in the first few pages of whatever standard textbook is employed) the concept of a **limit**, yet — as will be seen in due course — only 10 tiny symbols are needed to present data that expresses the schematic proposition that the limit of some given function f is L as the inputs to that function approach c.[2] Students who aspire to be highly paid data scientists seeking to answer BD problems (for Yahoo!; or for massive university systems like SUNY; or for those parts of the U.S. government that take profound action on the basis of BD, e.g, the U.S. Department of Education and the Federal Reserve; etc.) without truly understanding such little 10-symbol collections of data, put themselves, and their employers, in a perilous position. This is confirmed by any respectable description of what skills and knowledge are essential for being a good data scientist (e.g., see the mainstream description in Minelli, Chambers & Dhiraj 2013). In fact, it may be impossible to know with certainty whether the results of analytics applied to BD can be trusted, and whether proposed, actionable inferences from these results are valid, without understanding the underlying B^{3}D-based definitions of such analytics and inferences. Of course, the results by BD analytics, and indeed often the nature of BD itself, are probabilistic. But to truly understand whether or not some proposition has a certain probability of being true, at least the relevant data scientists, and perhaps also the managers and administrators ready to act on this proposition, must certainly understand what probability *is* — yet as is well-known, the nature of probability is expressed in none other than big-but-buried form.[3]

While we concede that there is some "crossover" (e.g., some pedagogy, to be sure, profits from "analytics" applied to BD; and of course some educators are themselves administrators), nonetheless we maintain there is a striking two-culture divide silently emerging in connection with big data: one culture driven by machine-assisted analysis of BD, and the fruit of that analysis; and the second by the quest for acquiring and bestowing mastery of B^{3}D,

[2]The limit of the function that takes some real number x, multiplies it by 2, and subtracts 5 (i.e., f is $2x - 5$), as x approaches 3, is 1. This very short statement, which also appears in Figure 2, rather magically holds within it an *infinite* number of buried datapoints (e.g., that 2 multiplied by 1, minus 5, is not equal to 1). But no high-school student understands limits without first understanding general 10-symbol-long schematic statements like this one. We return to this topic later (§4).

[3]While invented by Pascal, probability was still fundamentally obscure until Kolmogorov (1933) used precious few symbols to provide a classic big-but-buried axiomatization of all of probability.

and by the search for the big-but-buried data that confirms such mastery is in place within a given mind. Our chief goal is to introduce, clarify, and contextualize the BD-versus-B^3D distinction, in order to lay a foundation for the further integration of the two cultures, via the integration of the two types of data around which each tends to revolve. The truly effective modern university will be one that embodies this integration.[4]

The plan for the sequel is straightforward: We first present and affirm a serviceable account of what data is (§2), and specifically explain that, at least in education, *information* is key, and, even more specifically, *knowledge* is of paramount importance (in the case of *both* big data *simpliciter* and big-but-buried data). Next, in the context of this account, we explain in more detail the difference between BD and B^3D, by presenting two competing sets of necessary conditions for the pair, and some informal examples of these sets "in action" (§3). In the next section (4), we turn to the example of teaching calculus in the United States, in order to further elaborate the BD-versus-B^3D distinction, and to illuminate the importance of uniting data-driven effort from each side of the distinction.[5] Readers can rest assured that they will not need to know *any* calculus in order to understand what we say in this section, but we do explain that without a rudimentary understanding of calculus, human experience of even the simple motion of everyday objects quite literally makes no sense (from which, as we point out, the untenability of recent calls to drop any traditionally required pre-college math courses follows). We next (§5) briefly discuss the future. We first discuss the future of BD analytics in light of the historic Watson system from IBM. We then

[4] A sign the integration is missing is perhaps that there continues to be widespread tension between administrators and faculty, since the former live and die, these post-"Great Recession" days, by how well they obtain, analyze, and act on BD in the increasingly tight-money environment of today's Academy, while the latter, if still providing face-to-face instruction to physically co-located students, must be focused on teaching mastery of B^3D.

[5] Our points in this section could be based on any of the crucial big-but-buried data future data scientists ought to master (e.g., decision theory, game theory, formal logic, search algorithms, R, programming languages and theory, etc.), but calculus, occupying as it does a pivotal place in STEM education within the Academy, and — for reasons we herein review — in a general, enlightened understanding of our world, is particularly appropriate given our objectives. In addition, calculus provides the ultimate, sobering subject for gauging how math-advanced U.S. students are, or aren't, now, and in the future. We assume our readers to be acquainted with the brutal fact that, in math, K–12 U.S. students stack up horribly against their counterparts in many other countries. A recent confirmation of this longstanding fact comes in the form of the PISA 2012 results, which reveal that of 34 OECD countries, the U.S. is below average, and ranked a dismal 26th — and this despite the fact that the U.S. spends more per student on math education than most countries. See http://www.oecd.org/unitedstates/PISA-2012-results-US.pdf.

confront the acute problem of scalability that plagues the teaching of big-but-buried data, and point to a saving future in which revolutionary AI technology (advanced **intelligent tutoring systems**) solves the problem by teaching big-but-buried data in "sci-fi" fashion. A short conclusion wraps up the paper.

2 Data, Information, and Knowledge

It turns out that devising a rigorous, universally acceptable definition of 'data'[6] is surprisingly difficult, as Floridi (2008), probably the world's leading authority on the viability of proposed definitions for these concepts (and related ones), explains. For example, while some are tempted to define data as collections of *facts*, such an approach is rendered acutely problematic by the brute truth, routinely exploited in our "data age," that data can be **compressed** (via techniques explained e.g., in Sayood 2006): How could a fact be compressed?[7] Others may be inclined to understand data as knowledge, but such a view, too, is untenable, since, for example, data can be entirely meaningless (to wit, "The data you sent me, I'm afraid, is garbled and totally meaningless."), and surely one cannot know that which is meaningless. Moreover, plenty of what must be pre-analytically classified as data seems to carry no meaning whatsoever; Floridi (2005) gives the example of data in a digital music file. Were you to examine any portion of this digital data under the expectation that you must declare what it means, you would draw a blank, and blamelessly so. Of course, when the data is processed, it *causes* sound to arise, and that sound may well be eminently meaningful. But the data itself, as sequences of bits, means nothing.

In the interest of efficiently getting to the core issues we have targeted for the present paper, we affirm without debate a third view of what data is, one nicely in line with the overall thrust of the present volume: viz., we adopt the *computational* view of data, according to which data are collections of strings, digits, characters, pixels, discrete symbols, etc., all of which can be processed by algorithms unpacked as computer programs, which are in

[6] Or 'datum', a definition of which could of course be used to define the plural case.

[7] That which *expresses* a fact is of course readily compressible. This is probably as good a place as any for us to point out that the hiding that is part and parcel of big-but-buried data has nothing to do with data compression. In data compression, some bits that are statistically redundant are removed; by contrast, in B^3D, nothing is removed and nothing is redundant: usually all the bits or symbols, each and every one, is indispensable, and what's hidden is not found by adding back bits or symbols, but rather by human-level semantic reasoning.

turn executed on modern high-speed digital computers.[8] Affirmation of this view would seem to be sensible, since after all the big-data rage is bound up inextricably with computational analytics. When the IR office at university U is called upon by its Provost to bring back a report on what percentage of (undergraduate) transfer students from community colleges at U graduate, versus what percentage do who come in the customary first year to U from high school (aka. pipeline students), invariably their work in acceding to this request will require (not necessarily by the IR professionals themselves) the use of algorithms, programs regimenting those algorithms, and the physical computers (e.g., servers) on which the programs are implemented. And of course the same tenor of toil would be found outside of academia: If Amazon seeks to improve the automated recommendations its browser-based systems make to you for what you are advised to consider purchasing in the future given your purchases in the past, the company's efforts revolve around coming up with algorithmically smarter ways to process data, and to enticingly display the results to you.

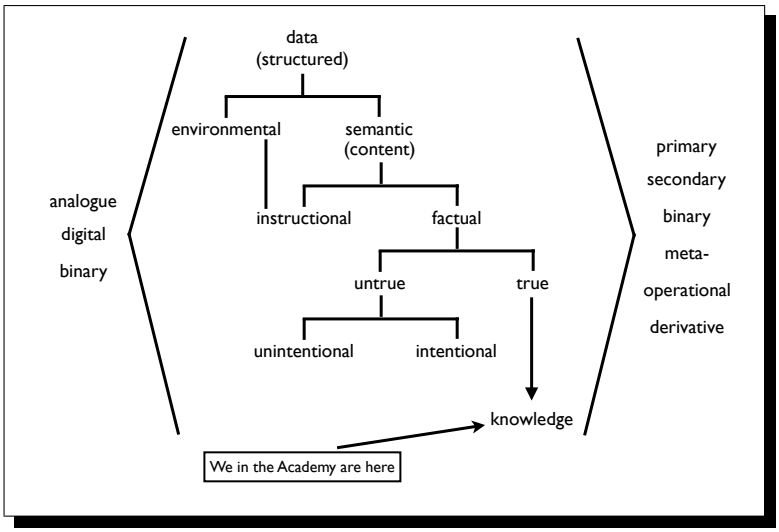
But we need a crisper context from which to move forward. Specifically, it's important to establish at the outset that universities and university systems, and indeed the Academy as a whole, are most interested in a specific *kind* of computational data: data that is both well-formed and meaningful. In other words, administrators, policy makers, analysts, educators, and students, all are ultimately interested in *information*. An elegant, succinct roadmap for coming to understand what information, as a special kind of data, is, and to understand the various kinds of information that are of central importance to the Academy and the technologized world in general, is provided in (Floridi 2010).[9] This roadmap is summed up in Figure 1. The reader should take care to observe that in this figure we pass to a kind of data that is even more specific than information: we pass to the sub-species of data that is a specific form of factual and true semantic information: we pass, that is, to *knowledge*. (Hence, while, as noted above, data isn't knowledge, *some* data does indeed constitute knowledge.) We make this move because, as indicated by the "We in the Academy are here" comment that

[8] Alert readers may protest that, technically speaking, there is such a thing as **analog data** and **analog computation**. But this quarter of modern information processing is currently a minuscule one, and students trained in data science at universities, as a rule, are taught precious little to nothing about analog computers and analog data. A readable, lively overview of computation and intelligence, including the analog case, is provided in (Fischler & Firschein 1987).

[9] Those wanting to go deeper into the nature of information are encouraged to study (Floridi 2011).

we have taken the liberty of inserting into Figure 1, the cardinal mission of universities is the pursuit and impartation of knowledge. From this point on, when, following common usage (which frames the present volume), we refer to data, and specifically to the fundamental BD-vs.B^{3}D dichotomy, the reader should understand that we are referring, ultimately, to *knowledge*. In the overarching world of data, data analysis, and data science, it is knowledge that research is designed to produce; knowledge that courses are designed to impart; and knowledge that administrators and managers seek out and exploit, in order to enhance the knowledge that research yields and classrooms impart.

Figure 1: Floridi's Ontology of Information



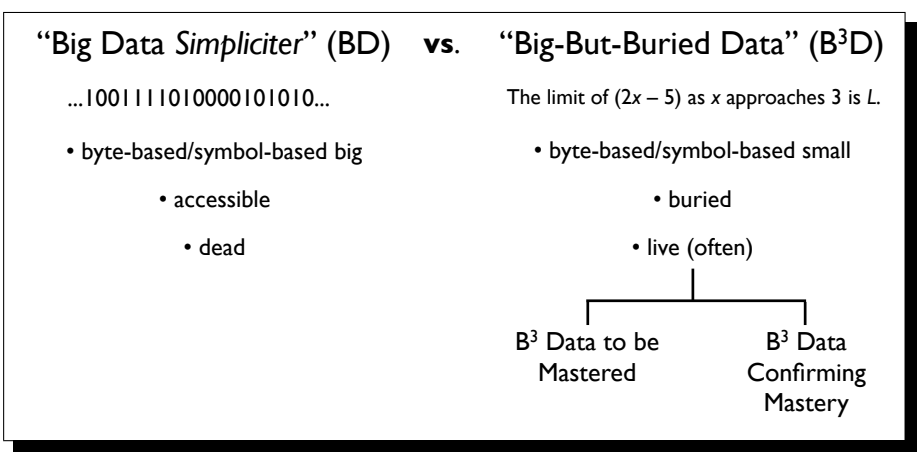
3 Big Data *Simpliciter* (BD) vs. Big-But-Buried Data (B^{3}D)

We provided above (§1) a provisional account of the difference between BD and B^{3}D. Let's now be more precise. But not *too* precise: formal definitions are outside the scope and nature of the present chapter. In the present context, it suffices (i) to note some necessary conditions that must be satisfied by any data in order to qualify it specifically as *big* in today's technology landscape (i.e., as BD), or instead as big-but-buried (i.e., as B^{3}D); and (ii) to flesh out these conditions by reference to some examples, including exam-

ples that connect to elementary calculus as currently taught in America's educational system. The "calculus part" of the second of these steps is, as planned, mostly reserved for the next section (4).

For (i), please begin by consulting Figure 2, which sums up in one simple graphic the dichotomy between BD and B^3D. Obviously, BD is referred to on the left side of this graphic, while B^3D is pointed to on the right. Immediately under the heading for each of the two sides we provide a suggestive string to encapsulate the intuitive difference between the two types of data. On the left, we show a string of 0's and 1's extending indefinitely in both directions; the idea is that you are to imagine that the number of symbols here is staggeringly large. For instance, maybe there are as many symbols as there are human beings alive on Earth, and a '1' indicates a male, whereas a '0' denotes a female. On the right, we show a simple 12-symbol-long statement about a certain limit. The exact meaning of this statement isn't important at this juncture (though some readers will perceive this meaning): it's enough to see by inspection that there are indeed only 12 symbols in the statement, and but to know that the amount of data "buried" in the statement is much, much bigger than the string of 0's and 1's to its left. This is true because the 12-symbol-long-statement is making an assertion (given in prose form in footnote 2) about every single real number, and while there are indeed a lot of human beings on our planet, our race is after all finite, while there are an infinite number of real numbers in even just one "tiny" interval, say the real numbers between zero and .5. Now let's look at the remainder of Figure 2.

Figure 2: BD vs. B^3D



Notice three attributes are listed under the BD heading, and a different, opposing trio is listed under the B^3D heading. Each member of each trio is a necessary condition that must hold of any data in order for it to qualify, respectively, as BD or B^3D. For example, the first hallmark of BD is that (and here we recapitulate what has been said above), whether measured in term of number of bytes or in terms of number of symbols, the data in question is large. The second necessary condition for some data to count as big data *simpliciter*, observe, is that it must be "accessible." What does this mean? The idea is simple. BD must be susceptible of straightforward processing by finite algorithms. To see this notion in action, we pull in here the suggestive string for BD given on the lefthand side of Figure 2:

$$\begin{array}{ccccccccccccccccccccc} \ldots & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & \ldots \\ & & & & & & & & & & \uparrow & & & & & & & & & \end{array}$$

Suppose we wanted to ascertain if the data here contains anywhere a substring of seven consecutive 0's. How would we go about answering this question? The answer is simple: We would just engage a computation based on a dirt-simple algorithm. One such algorithm is:

> Moving simultaneously left and right, starting from the digit pointed to by the arrow (see immediately above), start a fresh count (beginning with one) for every switch to a different digit, and if the count ever reaches seven, output "Yes" and HALT; otherwise output "No" and HALT when the digits are exhausted.

It should be clear that this algorithm is infallible, because of the presupposition that the data in question is accessible. Sooner or later, the computation that implements the algorithm is going to return an answer, and the correct one at that, for the reason that the data is indeed accessible. This accessibility is one of the hallmarks of BD, and it is principally what makes possible the corresponding phenomenon of "big analytics." The techniques of statistical computing are fundamentally enabled by the accessibility of the data over which these techniques can operate.[10] Things are very different, though, on the other side of the dichotomy: big-but-buried data is, as its name implies, buried.

[10] Of course, we give here an extremely simple example, but the principles remain firmly in operation regardless of how much BD one is talking about, and regardless of how multi-dimensional the BD is. The mathematical nature of BD and its associated analytics is in fact ultimately charted by working at the level of running algorithms over binary alphabets, as any elementary, classic textbook on the formal foundations of computer science will show (e.g.,, see Lewis & Papadimitriou 1981).

Here's a simple example of some B^3D:[11] Suppose we are given the propositional datum that (a) everyone likes anyone who likes someone. And suppose as well that we have a second datum: (b) Alvin likes Bill. The data composed of (a) and (b) is how big? Counting spaces as separate characters, there are only 58 symbols in play; hence we certainly are not in the BD realm: we are dealing with symbol-based small data; which is to say that the second hallmark of B^3D shown in Figure 2 is satisfied. Or at least the reader will agree that it's satisfied once the hidden data is revealed.

Toward that end, then, a question: (Q) Does everyone like Bill? The answer is "Yes," but that answer is buried. Most people see that data composed of (a) and (b) imply that (c) everyone likes Alvin; few people see that (a) and (b) imply that (d) everyone likes Bill. Datum (d), you see, is buried. And notice that (d) isn't just buried in the customary sense of being extractable by statistical processing (so-called "data mining"): No amount of BD analytics is going to disclose (c), accompanied by the justification for (d) on the strength of (a) and (b).[12] If you type to the world's greatest machine for answering data queries over BD, IBM's historic *Jeopardy!*-winning Watson system (Ferrucci et. al 2010), both (a) and (b), and issue (Q) to Watson, it will not succeed. Likewise, if you have R running before you (as the second author does now), and (a) and (b) are represented in tabular form, and are imported into R, there is no way to issue an established query to express (Q), and receive back in response datum (d) (let alone a way to receive back (d) plus a justification such as is provided via the proof in footnote 12). To be sure, there is a lot of machine intelligence in both Watson and R, but it's not the kind of intelligence well-suited for productively processing big-but-buried data.[13]

It is crucial to understand that the example involving Alvin and Bill has been offered simply to ease exposition and understanding, and is not representative of the countless instances of big-but-buried data that make possible the very data science and engineering heralded by the present book. It is student mastery of B^3D that is cultivated by excellent STEM education,

[11] The example was originally given to the second author by Professor Philip Johnson-Laird as a challenge.

[12] But we supply this here: Since everyone likes anyone who likes someone, and Alvin likes Bill, everyone likes Alvin — including Bill. But then since Bill likes Alvin, and — again — everyone likes anyone who likes someone, we obtain: (d) everyone likes Bill. QED

[13] Our purposes in composing the present essay don't include delivery of designs for technology that can process BD and/or B^3D. Readers interested in an explanation of techniques, whether in the human mind or in a computer, able to answer queries about big-but-buried data, and supply justifications for such answers, can begin by consulting (Bringsjord 2008).

in general.[14] And we are talking not just about students at the university level; B^3D is the key part of the 'M' in 'STEM' education much earlier on. For instance, just a few hundred symbols are needed to set out the full quintet of Euclid's Postulates, in which the entire infinite paradise of a large part of classical geometry resides. The data composing this paradise is not just very large; it's flat-out infinite. Exabytes of data does make for a large set to analyze, but Euclid, about 2.5 millennia back, was analyzing datasets much bigger than the ones we apply modern "analytics" too. And the oft-forgotten wonder of it all is that the infinite paradise Euclid (and Aristotle, and a string of minds thereafter; see e.g. Glymour 1992) explored and mapped can by crystalized down to just a few symbols that do the magical "hiding." These symbols are written out in about one quarter of a page in every geometry textbook used in just about every high school in the United States. And geometry is just a tiny exhibit to make the point.[15] The grandest and most astonishing example of big-but-buried data in the realm of mathematics is without question the case of axiomatic set theory: it is now agreed that nearly all of classical mathematics can be extracted from a few hundred B^3D symbols that express a few basic laws about the structure of sets and set operations. (Interested readers can see for themselves by consulting the remarkably readable and lucid (Potter 2004). A shortcut for the mathematically mature is to consult the set-theory chapter in (Ebbinghaus, Flum & Thomas 1994).)[16]

Finally, with reference again to Figure 2, we come to the third hallmark of BD ('dead'), versus the corresponding opposing hallmark of B^3D ('live'). What are we here referring to? A more hum-drum synonym in the present context for 'dead' might be 'pre-recorded.' In the case of BD, the data is pre-recorded. The data does not unfold live before one's eyes. The *analysis* of BD is of course carried out by running processes; these processes are (by definition) dynamic, and can sometimes be watched as they proceed in

[14] This is perhaps the place to make sure the reader knows that we know full well that mastery isn't always permanent. Re-education is very important, as is the harnessing of mastery in support of ongoing work, which serves to sustain mastery. In fact, the sometimes fleeting nature of mastery only serves to bolster our case. Due to space limitations, we leave aside treatment of these topics herein.

[15] As even non-cognoscenti will be inclined to suspect, Euclid only really kicked things off, and the B^3D-oriented portion of the human race is still making amazing discoveries about plane geometry. See the positively wonderful and award-winning (Greenberg 2010).

[16] Lest it be thought the wonders of B^3D are seen only in mathematics, we inform the reader that *physical* science is increasingly being represented and systematized in big-but-buried data. For instance, half a page of symbols are needed to sum up all the truths of relativity theory. See (Andréka, Madarász, Németi & Székely 2011).

realtime. For example, when Watson is searching BD in order to decide on whether to respond to a *Jeopardy!* question (or for that matter any question), human onlookers can be shown the dynamic, changing confidence levels for candidate answers that Watson is considering — but the data being searched is itself quite dead. Indeed, big data *simpliciter*, in and of itself, is *invariably* dead. Amazon's systems may have insights into what you are likely to buy in the future, but those insights are without question based on analysis of "frozen" facts about what you have done in the past. Watson did vanquish the best human *Jeopardy!* players on the planet, but again, it did so by searching through dead, pre-recorded data. And IR professionals at university U seeking for instance to analyze BD in order to devise a way to predict whether or not a given first-year student is going to return for her sophomore year will analyze BD that is fixed and pre-recorded. But by contrast, big-but-buried data is often "live" data.

Notice we say *some* B^3D is live. Not all of it is. This bifurcation is explicitly pictured in the bottom right of Figure 2. What gives rise to the split? From the standpoint of education, the split arises from two different cases: on the one hand, situations where some big-but-buried data is the target of learning; and on the other, situations like the first, *plus* the live production of big-but-buried data by the learner, in order to demonstrate she has in fact learned. Accordingly, note that in our figure, the bifurcation is labeled to indicate on the left that which is to be mastered by the student, and on the right, the additional big-but-buried data which, when generated, confirms mastery.

For a simple example of the bifurcation, we have only to turn back to this trio

(a) Everyone likes anyone who likes someone.

(b) Alvin likes Bill.

(Q) Does everyone like Bill?

and imagine a student, Bertrand, say, who in a discrete-mathematics class, during coverage of basic boolean logic (upon which, by the way, modern search-engine queries over BD on the Web are based) is given this trio, and asked to answer (Q). But what sort of answer is Betrand specifically asked to provide? Suppose that he is asked only for a "Yes" or "No". Then, *ceteris paribus*, he has a 50% chance of getting the right answer. If Bertrand remembers that his professor in *Discrete Math* has a tendency to ask tricky questions, then even if Bertrand is utterly unsure, fundamentally, as to what the right answer is, but perceives (as the majority of college-educated people

do) that certainly from (a) and (b) it can be safely deduced that everyone likes Alvin, he may well blurt out "Yes." And he would be right. But is mastery in place? No. Only the live unearthing of certain additional data buried in our trio can confirm that mastery is in place: viz., a proof (such as that provided in footnote 12) must be either written out by Bertrand, or spoken.

3.1 Two Questions, Two Answers

Some readers are likely thinking: "But why do you say the 'frozenness' of big data *simpliciter* is a necessary condition of such data? Couldn't the very systems you cite, for example Watson and Amazon's recommender systems, operate over vast amounts of data, while that very data is being generated? It may be a bit creepy to ponder, but why couldn't it be that when you're browsing Amazon's products with a Web browser, your activity (and for that matter your appearance and that of your local environment) is being digitized and analyzed continuously, in real time? And in terms of education, why couldn't the selections and facial expressions of 500,000 students logged on to a MOOC session be collected and analyzed in real time?"

This is an excellent question. Eventually, perhaps very soon, a lot of BD will indeed by absorbed and analyzed by machines in real time. *Today*, however, the vast majority of BD analytics is performed over "dead" data; Figure 2 reflects the current situation. Clearly, BD analytics is not *intrinsically* bound up with live data. On the other hand, confirmation of the kind of mastery with which we are concerned *is* intrinsically live. Of course, we do concede that a sequence in which a student produces conclusive evidence of mastery of some B^3D could be recorded. And that recording is itself by definition — in our nomenclature — dead, and can be part of some vast collection of BD. (A MOOC provider, for instance, could use a machine vision system to score 500,000 video recordings of student behavior in a class with 100,000 students.) But the instant this BD repository of recordings is relied upon, rather than the live generation of confirming data, the possibility of cheating rears up. *If* one assumes that the recording of live responses is fully genuine and fully accurate, then of course the recording, though dead, conveys what *was* live. But in addition to the cheating issue, there's what can be called the "follow-up" problem in the case of recordings: You can't query a recording on the spot in order to further confirm that mastery is indeed in place. In sum, then, there is simply no substitute for the unquestionably authentic live confirmation of deep understanding; and, accordingly, no substitute for the confirmatory power of oral examina-

tion, over and above the examination of dead data, even when that dead data is a record of live activity.

We also anticipate some asking: "But why do you say that the kind of data produced by Bertrand when he gives the right rationale is big-but-buried? I can see that (a) and (b) together compose a simple instance of B^3D. But I don't see why what is generated in confirmation of a deep understanding of (a) plus (b) is *itself* a simple case of big-but-buried data."

The answer is that, one, as a rule, when a learner, on the spot before one's eyes, generates data that confirms mastery of big-but-buried data, she has extracted that data from the vast and often infinite amount of big-but-buried data that is targeted by the teacher for mastery; *and*, two, because the data that is unearthed is *itself* big-but-buried data: it's symbol-wise small, yet hides a fantastically large (indeed probably infinite) amount of data. In the immediate case at hand involving Bertrand, if the correct rationale is provided (again, see footnote 12), what is really provided is a reasoning *method* sufficient for establishing an infinite number of results in the formal sciences.[17]

4 The Example of Calculus

We now as promised further flesh out the BD-vs.B3D distinction by turning to the case of elementary calculus.

4.1 On Big Data *Simpliciter* and Calculus

We begin by reviewing some simple but telling BD-based points about the AP (= Advanced Placement) calculus exam, in connection with subsequent student performance, in the United States. These and other points along this line are eloquently and rigorously made in (Mattern, Shaw & Xiong 2009), and especially since here we only scratch the surface to serve our specific needs in the present paper, readers wanting details are encouraged to read the primary source. We are specifically interested in predictive BD analytics, and specifically with the question: Does performance on the Calculus AP exam, when taken before college, predict the likelihood of success in college? And if so, to what degree?[18]

[17] Bertrand, if successful, will have shown command over (at least some aspects of) what is known as **recursion** in data/computer science, and the rules of inference known as **universal elimination** and **modus ponens** in discrete mathematics.

[18] Analytics applied to non-buried data generated from relevant activity at individual universities is doubtless aligned strikingly with what the College Board's AP-based anal-

The results indicate that AP Calc performance is highly predictive of future academic performance in college. For example, using a sample size of about 90,000 students, Mattern et al. (2009) found that those students scoring either a 3, 4, or 5 on the AP Calc (AB) were much more likely to graduate within five years from college, when compared to those who either scored a 1 or a 2, or didn't take the test. With academic achievement identified with High School GPA (HSGPA) and SAT scores, the analysis included asking whether this result held true when controlling for such achievement. In what would seem to indicate the true predictive power of student command of calculus, even when controlling for academic achievement (SAT and HSGPA run as covariates), the result remained: those earning a 3, 4, or 5 were much more likely to graduate from college.

But *why* is the cognition cultivated in calculus apparently so powerful and valuable? This is something that BD will not reveal, for the simple and widely known reason that correlation doesn't explain causality. An administrator or policy maker could thus see in the analysis of BD evidence that such cognition highly *correlates* with desirable outcomes (graduate rate, e.g.), but would not see what underlying, buried data define what calculus is, and would not see what mastery of the subject consists in. This brute fact is of course perfectly consistent with the real possibility that the administrator is herself a calculus wiz: the limitation is in the nature of BD, not in the mind of those analyzing BD. Likewise, even if an administrator had further correlation data (e.g., showing that achievement in economics and physics correlates stunningly well with high performance in calculus courses, which happens to also be true), no deep understanding of why the correlations hold is on the table. Indeed, an administrator could, for all that the BD analytics tells us, view calculus as simply some kind of magical black box — but a black box to be advocated for. We thus now look at calculus from a B^3D perspective.

ysis shows. For instance, at Rensselaer Polytechnic Institute, grades in the first calculus course for first-year students (Math 1010, Calc I) is highly predictive of whether students will eventually graduate. In fact, such grades have more predictive power at RPI than HSGPA and SAT scores. (Of course, RPI is a technological university, and STEM disciplines dominate. Perhaps a grade in calculus would not be more predictive than Verbal SAT scores with respect to students entering a university with the intention of securing, say, a BA in Classics.

4.2 On Big-But-Buried Data and Calculus

Calculus[19] is a tremendously important subject in the modern, digital economy — for many reasons. One reason is that, as the sort of BD analysis visited above indicates, apparently the cognition that goes hand in hand with learning calculus in turn goes hand in hand with academic success in STEM.[20] A second reason why calculus is crucial is that real analysis (of which the calculus is a part, and to which, in our K–16 educational system, calculus is the gateway) stands at the heart of many important approaches to the analysis of BD. Contemporary macroeconomics is for instance based on real analysis; it's for instance impossible to understand the most powerful macroeconomic arguments in favor of generous Keynesian spending by the U.S. government, despite budget deficits and debt, without an understanding of calculus.[21]

[19]By 'calculus' here we have meant and mean elementary versions of both the differential and integral calculi, invented independently three centuries ago by Leibniz (whose ingenious and elegant notation is still used today in every calculus course) and Newton, which are united by the Fundamental Theorem of Calculus, a result traditionally presented to students in their first calculus course. (While today calculus is taught to the world's students through the starting "portal" of a the concept of a limit (a contemporary tradition echoed, of course, in the present chapter), this pedagogical approach is historically jarring, since, instead, **infinitesimals** (infinitely small numbers) formed the portal through which Newton and (especially) Leibniz seminally passed to find and provide calculus to humanity. Today, we know that while Leibniz was long lampooned for welcoming such a fuzzy thing as an infinitesimal, his approach has been fully vindicated, through the groundbreaking work of Robinson (1996), who continued the seminal work of Norwegian logician Thoraf Skolem (1934), and one can even find an occasional textbook today that gives an infinitesmal-based approach to teaching calculus.) There are many other calculi of great importance in our increasingly digital world; for instance, the λ-calculus, introduced by Church (1936), occupies a seminal and — often through much-used-today formalisms to which it is mathematically equivalent — still-central place in the history of data science.

[20]Of course, some of the natural sciences aren't all that intimately bound up with calculus; biology would be a case in point. We are saying that the *cognition* required to learn and apply calculus is what transfers across learning in data science and STEM, not all of the B^3D particulars of calculus. By the way, while largely ignored, the idea that biology itself can be expressed in just a few symbols in an axiomatic system was rather long ago seminally presented by Woodger (1962).

[21]E.g., see the intriguing case in favor of Keynesian spending articulated in (Woodford 2011), in which economies are modeled as infinitely-lived "households" that maximize utility through infinite time series, under for instance the constraint that the specific, underlying function u, which returns the utility produced by the consumption of a good, must be such that its first derivative is greater than zero, while its second derivative is less than zero. Without understanding the differential calculus, one couldn't possibly understand Woodford's (2011) case. And note that, in how Woodford models an economy,

To illustrate the prudence of a focus on B^3D at the present juncture in our discussion, consider the case of Johnny, who, upon arriving as a first-year student intending to major in math at university U, boldly announces to Professor Smith, at orientation before the start of classes, that he (Johnny) should leapfrog over the three math-major calculus courses (I, II, III) in the department, straight into *Advanced Analysis*.

Professor Smith: "You know, Johnny, our *Calc III* requires not just what some of our students call 'plug and chug,' but proofs. One must be able to prove the formulas that are used *for* plugging and chugging."

Johnny: "Not a problem, Sir."

Dr. Smith, looking down at his desk, observes that Johnny received an A in pre-college (single-variable) calculus, and scored a 5 on the Calculus AB Advanced Placement test. Smith knows that this record is good enough, by a decision tree generated from analysis of relevant BD, to skip *Calc I* for math majors; but many students with super-high SAT scores don't even do that. We make two claims:

Claim 1 Even if Dr. Smith has at his beck and call all the BD in the world, and even if by some miracle he had the time right here on the spot to issue a hundred queries against this data while Johnny waits in silence, he can instead find out whether Johnny is all bluster, or the real deal, by asking one or two single-sentence questions, and by then sitting back to see whether the young man writes out the one or two key proofs requested, or not.[22] In short, it will be live big-but-buried data that settles the question, on the spot.

Claim 2 The best classroom teaching arguably proceeds by way of the teacher ascertaining directly, in decidedly low-tech oral-exam fashion, whether a "golden," buried datum of true mastery or understanding is there or not, and then striving to get such understanding to take root if not, and then testing in like manner again, and ... so the cycle continues until learning is confirmed.

This pair of claims can be put into action for teaching even very young students. For instance, by using visual forms of big-but-buried data one can quickly make serious headway in explaining the concept of a limit to even

he is hardly idiosyncratic, since he follows a longstanding neoclassical approach articulated e.g. by Barro & King (1984).

[22] Any of the theorems explicitly presented and employed in early calculus courses (where students are typically *not* asked to prove them) would do. In his NSF-sponsored, seminal approach to engineering computers able to assist humans in their learning of calculus, Suppes (see e.g. Suppes & Takahasi 1989) asked students to e.g. prove the Intermediate Value Theorem.

middle-school students, and thereby build a substantial part of a path to full-blown calculus for them. For example, see Figure 3, which is taken from page 268 of (Eicholz et al. 1995), a middle-school textbook. Imagine that Alexandra, in the 7th grade, is asked to determine the "percent pattern" of the outer square consumed by the ever-decreasing shaded squares. The pattern, obviously, starts at $\frac{1}{4}$, and then continues as $\frac{1}{16}, \frac{1}{64}, \frac{1}{256}, \ldots$. When asked what percentage the shaded square would "get down to" if someone could forever work faster and faster, and smaller and smaller, at drawing the up-down and left-right lines that make each quartet of smaller squares, Alexandra announces: "Zero." That is indeed none other than the limit in the present case: the percent "in the limit" the shaded square consumes of the original square is indeed zero. The figure in question is tiny, but hides in gem-like fashion an infinite progression.

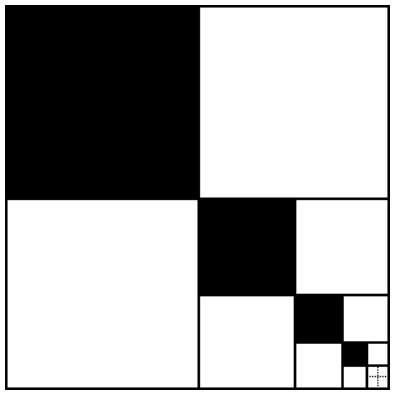


Figure 3: B^{3}D-based Representation of a Limit in Seventh-Grade Math

Of course, asking for and assessing the kind of live big-but-buried data that Johnny and Alexandra are here asked to produce, if in fact such techniques can scale to millions of students (an issue we take up in §5.2), is an expensive proposition, to put it mildly. Skeptics will pointedly ask why something as recherché as calculus would ever warrant the expenditure of all the time and money it would take to ensure mastery in this manner. Unfortunately, the mistaken view that a deep understanding of calculus is a needless luxury is shared by many.

In fact, even among many members of the Academy in our day, the view that calculus has narrow value is firmly afoot. Many university professors are under the impression that calculus has value in fields like engineering, math itself, and the like, but doesn't reach across the human experience. Unfortunately, this view belies an ignorance of intellectual history, and specifically

of the fact that without calculus, everyday concepts like motion are literally incoherent. One quick antidote to this ignorance, and the even worse ignorance behind strident calls to block federal educational standards for higher-level mathematics in high school, is to turn to some of Zeno's paradoxes of motion, for instance to the Paradox of the Arrow, to which we now briefly turn our attention.

4.2.1 The Paradox of the Arrow

Here then a summary in our words of Zeno's reasoning:[23] "Time is composed of moments, and hence a moving arrow must occupy a space filled by itself at each moment during its supposed travel. Our arrow is thus at a particular place at each moment during its supposed travel. Assuming for the sake of argument that an arrow (supposedly) travels only a short distance, the picture given in Figure 4.2.1 should be helpful. But there is no motion here whatsoever. After all, places certainly don't move. Hence, if, as shown, the arrow is at each moment at a particular place, occupying a space equal to its volume, the arrow cannot possibly ever really move: it is not moving at any of the moments m_i, since at each such moment it is simply at the place where it is, and there are no other moments at which it can move! The reasoning here can be effortlessly generalized to show that the movement of anything is an illusion."

The quickest way to reveal to an intelligent person in the modern information age the centrality and indispensability of calculus for understanding the world in more than a mere child-like, hand-wavy manner is to ask whether motion is real; and upon receiving an affirmative, to then ask how that can be in the light of the Zenoian reasoning given here. (It is not a cogent response to simply shoot an arrow or throw a baseball and say "See?", since Zeno's claim is precisely that while things certainly *seem* to move, they actually don't.) All cogent responses must include appeal to calculus, and all the big-but-buried data that calculus at bottom is.[24] We might mention

[23]The vast majority of Zeno's direct writings are unfortunately not preserved for us living in the big-data era: We know of Zeno's reasoning primarily via Aristotle's (certainly compressed) presentation of it. The Paradox of the Arrow is presented by Aristotle in *Physics*, 239b5-32, which can be found in (McKeon 1941). The titles given to Zeno's paradoxes (with 'Paradox of the Arrow' no exception) have been assigned and affirmed by commentators coming after him. Zeno himself wrote in the fifth century B.C. Aristotle about two centuries later. Would-be scholarly detectives with an interest in intrigue, we promise, will be nicely rewarded by searching out what is written/known about both Zeno the man and his work, beyond Aristotle as source.

[24]Put with brutal brevity, one learns in calculus that the escape from Zeno's otherwise

Table 1: Zeno's Framework for Paradox of the Arrow

Moments	**Places Where Arrow Located**
m_1	$\longrightarrow$ @ place 1
m_2	$\longrightarrow$ @ place 2
m_3	$\longrightarrow$ @ place 3
m_4	$\longrightarrow$ @ place 4

that in light of this, it is quite astonishing that, in response to Common Core Math Standards urged by the U.S. Department of Education and most States (the main rationale for which is of course based upon analysis of BD showing that U.S. students, relative to those in other countries with whom they will be competing in the global, data-driven economy, are deficient), some maintain that mathematics should be simply an elective in high school. For instance, Baker (2013) stridently advances the claim that even a dedicated high-school algebra course is, for most, downright silly, and downright painful; and, accordingly, no such course should be required.[25] Needless to

valid reasoning is that motion is formally defined in terms of what occurs at "nearby" moments. An arrow simply can't be motion in or during a particular moment, but thanks to calculus, we know precisely that it can certainly and easily have **instantaneous velocity** (formally defined early in a first calculus course using derivatives), since a traveling arrow is at different positions at moments before or after the instant in question. Zeno's reasoning stood rock-solid and (assuming honesty on the part of those courageous enough to confront it) compelling, despite rather desperate attempts to refute it (Aristotle struck out first), for millennia, until the advent of calculus.

[25] Among the many fallacies committed by Baker (2013) is this prominent one: *reductio ad absurdum* deployed in the absence of any absurdity. All serious students of mathematics are taught that when deploying this rule of inference, one must obtain the absurdity or contradiction in question, at which point one is then free to reject the proposition that implies the absurdity. Baker, apparently having never been taught this, blithely quotes (out of context, by the way) snippets from algebra textbooks, taking it for granted that the absurdity is thereby made plain (so that, in turn, the required teaching of these textbooks is shown to be a very bad idea). For instance, we are supposed to instantly perceive the absurdity in the following, which is word for word in its entirety a specimen of what Baker confidently presents and assumes to be self-evidently absurd:

> A rational function is a function that you can write in the form $f(x) = \frac{P(x)}{Q(x)}$, where $P(x)$ and $Q(x)$ are polynomial functions. The domain of $f(x)$ is

say, if the ordinary motion of everyday objects makes no intellectual sense without at least a fundamental conception of calculus, without mastery of even algebra one quickly advances toward lowering a definition of the human from — to use Aristotle's phrase — *rational animal* to just *animal*. And of course it is impossible that our universities produce the data scientists our economy needs without taking in students who know algebra, and who can then build upon that knowledge out to knowledge of valuable analytics, including techniques requiring calculus.

5 The Future

As promised, we now briefly touch upon the future, in connection, respectively, with IBM's Watson system, and following naturally on that, with so-called **intelligent tutoring systems** (ITSs), AI systems able to tutor individual students in various disciplines.[26]

5.1 Watson, BD, B^{3}D, and the Future

Most people, at least those in the U.S., are aware of the fact that Watson, an AI system engineered by IBM, triumphed to much fanfare in 2011 over the best (at the time) human *Jeopardy!* players. Most people are also aware of the fact that this victory for a machine over humans expert in a particular game follows an entrancing pattern that IBM established and pulled off previously, when, in a 1997 rematch, its Deep Blue, a chessplaying computer program, with the world watching move by move, beat Gary Kasparov, at that time the greatest human chessplayer on the planet. Yet the pattern isn't quite the same, for there is a big difference between the two AI systems in question: Whereas Deep Blue had narrow expertise and no capacity to process data expressed in so-called **natural languages** like English and

> all real numbers except those for which $Q(x) = 0$. (Quoted on p. 32 of (Baker 2013).)

It is easy to see that if this is taken to be self-evidently absurd (simply because some will find it inscrutable?), Baker's project is vitiated by parody, since plenty of people find, say, Dante to be absurd and inscrutable and inapplicable in everyday life. (And if not Dante, then certainly for every chap who finds Baker's specimen absurd, we can find one who regards the altiloquent sentences of Proust to be self-presentingly silly). Euclid, so far as we know the first systematic user of *reductio ad absurdum*, taught us that this pattern of inference requires putting on clear display, for all to uncontroversially see, the contradiction in question.

[26] For a superlative introduction to ITSs, and BD analysis regarding their effectiveness, see (VanLehn 2011).

Norwegian, Watson does have such a capacity (with respect to English, currently). To put it bluntly, despite the fact that a chessplaying machine of the power of Deep Blue realized one of the longstanding and strategically targeted dreams of AI (e.g., see Newell 1973), chess, compared to challenges that involve human language, is easy (Bringsjord 1998). And yet Watson too has some noteworthy limitations.

For example, while Watson is able to return correct answers to many natural-language questions, it does so on the strength, specifically, of its having on hand not simply vast amounts of frozen BD, but specifically vast amounts of frozen **structured** BD. The reader will recall that we defined 'data' for purposes of the present inquiry (§2), but we left aside the distinction between structured and unstructured data. Structured data is data nicely poised for profitable processing by computation. Paradigmatic structured data would for example be data in a relational database, or a spreadsheet; the College-Board data discussed briefly above, for instance, was all structured, and housed in databases. Unstructured data includes what we humans for the most part use for human-to-human communication: emails, narratives, movies, research papers, lectures, diagrams, sketches, and so on; all things that computers cannot currently understand (to any real degree, anyway), not even Watson. Fortunately for fans of BD and BD analytics, and for IBM, this limitation on Watson can be manually surmounted via ingenious human engineering, carried out within a seminal framework that was invented long before Watson.[27] This engineering takes unstructured data in a given domain in as input, and "curates" it to produce corresponding structured data that can be penetratingly analyzed by Watson and its wondrous algorithms.[28]

Can the manual "translation" from unstructured to structured data be automated? IBM recently announced a $100 million expansion in the planned reach and power of Watson (Ante 2014), but that expansion appears to sustain the need for engineers to "translate" unstructured information in some domain (e.g., medicine) into structured data. A profound and open question about the future is whether or not the process of passing from un-

[27] That framework is UIMA; see (Ferrucci & Lally 2004).

[28] It's important to note, and concede, that human communication makes extensive and routine use of diagrammatic information (pictures, videos, diagrams, images, etc.), and that the AI challenge of engineering intelligent machines able to genuine process such content is a severe one. Along these lines, see (Bringsjord & Bringsjord 1996). We used a diagram to represent big-but-buried data above, in Figure 3. There is currently no foreseeable set of AI techniques that would allow a computing machine to understand what even bright middle-schoolers grasp upon study of the remarkably rich diagram in question.

structured to structured data can be automated. Without that automation in place, the cost of providing deep question-answering technology for the university community (and, indeed, *any* community) will continue to carry the labor cost of data scientists and engineers having to configure Watson for deployment. That cost may or may not be surmountable.

But more to the points at hand in the present essay, we remark upon a second limitation that currently constrains Watson: It can only handle questions about BD, not B^3D. Watson, as suitably pre-engineered for *Jeopardy!* competition, would presumably be able to answer, say,

- "Watson, what 'Little Flower' famously ran the Big Apple?"

and this capacity is without question a stunning achievement for AI. But Watson cannot currently handle this (now-familiar-to-our-readers!) question:

- "Watson, what is the limit of the function two times x, minus five, as x approaches three?"[29]

If in the future Watson developed an ability to answer such questions, the consequences for the Academy would be momentous. For then "under one roof," Watson's analysis of BD would be powerful, and deep education centered around B^3D could in theory be provided as well. In other words, Watson would be in position to function as a revolutionary component of an intelligent tutoring system (ITS), a category of intelligent machinery to which we now briefly turn our attention.

5.2 Intelligent Tutoring Systems and the Future

It has doubtless not escaped the reader's attention that the kind of education on which we have tended to focus herein is certainly more akin to one-on-one tutoring than to, for instance, the kind of instruction offered by a professor teaching a MOOC to myriad students spread across the globe. Yet our focus is purely a function of the intimate relationship that undeniably exists between tutoring-style education and big-but-buried data; the

[29]It is possible, subsequent to the publication of the present chapter (since it will then end up being frozen for future consumption and available on the Internet, that the very text you are reading might happen to end up being "digested" by Watson, in which case Watson might in fact return '1.' But obviously a question along the same line, but never asked in the history of our race, could be devised, and posed to Watson. And besides, Watson could be asked, as the aforementioned Johnny was, to prove the answer returned correct.

focus, for the record, is not reflective of any animus on our part to other pedagogical structures. For example and for the record, we both regard peer-to-peer learning to be powerful. Regardless, in the future, why can't ITSs be imbedded within MOOCs? Why can't each of the tens (or hundreds ...) of thousands of students signed up to take calculus in a MOOC, or signed up to watch educational videos from Khan Academy (which offers many excellent ones on calculus), whether they are students at the high-school or college level, *also* have supplementary interaction with an ITS?

If the correct answer to these questions is the sanguine "There's no reason why they can't!", it follows immediately that tomorrow's AI systems, specifically ITSs, will somehow obtain a capacity for understanding natural language, and for understanding infinite sets and structures; but the notion that such capacities will be acquired by tomorrow's computing machines is unsupported by any empirical evidence on hand today. Today, no AI system, and hence no ITS, can genuinely understand the natural-language sentences we routinely use; nor can such a system understand the infinitary nature of even our elementary mathematics. Both plane geometry and calculus, the two branches of mathematics touched upon most above, are irreducibly infinitary in nature, in that the key structure they presuppose is one and the same: the continuum; that is, the reals, which are not only infinite, but breathtakingly so.[30] In light of this daunting situation, we regard it to be much more likely that machine teaching systems of the future will be engineered to simply look for readily observed BD *correlates* to deep understanding of big-but-buried data; for example, perhaps statistical patterns of activity in brain regions that correlate with human genuine understanding of big-but-buried data.

6 Conclusion and Future Work

The truly great universities of the future will leverage both big data *simpliciter* and big-but-buried data. Today, the administrator is boxed by a frustrating conundrum: She knows that big data often doesnt tell her what she really wants to know, but in many cases such data is all she has, and after all, she must be accountable, so, what to do? The individual case, in which big-but-buried data inheres, is left aside, and lots of students are consequently left behind. To escape the box, administrators at future, great

[30] The reals are larger than the natural numbers (0, 1, 2, ...), and larger too than the rational numbers (natural-number fractions). For readable explanation and proof, see (Boolos, Burgess & Jeffrey 2003).

universities will set policies and strategies on the basis of what analysis of BD discloses — but at the same time, educators at these same universities will succeed because their teaching verifiably cultivates relevant mastery in their students. This mastery will be of key B^3D: mastery of the kind of big-but-buried data that is precisely what effective data scientists must understand. And confirmation of this mastery will be provided by the ability of the students in question to generate, in live fashion, additional big-but-buried data. The success of educators will in turn be confirmed by the BD that administrators and policy makers analyze. In addition, as we have shared, in order to among other things address the scalability problem, truly great universities and systems thereof, whether of the conventional bricks-and-mortar or online type, will deploy, or at least seek to deploy, revolutionary AI and neuro-scanning technology able to, respectively, tutor humans, and verify that this tutoring leads to genuine human mastery.

We view the present chapter as a prolegomenon to the research needed in order to reach the greatness pointed to in the prior paragraph. There are at least two trajectories such research must take. The first is to climb toward a seamless integration between administrators on the one hand, and on the other educators "on the ground." Making this climb requires that BD and B^3D must themselves be seamlessly integrated. It's not enough to be able to pinpoint that failure to graduate can be predicted by a failure to secure a strong grade in calculus. We must reach a time when, having pinpointed such things, we can in response simulate a range of educational interventions, personalized for each particular student, in order to find those that lead to mastery of big-but-buried data, and are thus wise investments for administrators and policy makers to make. Implementing those interventions will then in turn lead back to improvement signaled at the BD level, for instance higher graduate rates across a university, a university system, a state, or across the United States as a whole. The second trajectory is of course r&d devoted specifically to providing the availability of these implementations; that is, to the design and engineering of intelligent tutoring systems with the kind of unprecedented power we have described.

References

Andréka, H., Madarász, J. X., Németi, I. & Székely, G. (2011), 'A Logic Road from Special Relativity to General Relativity', *Synthese* pp. 1–17.
URL: *http://dx.doi.org/10.1007/s11229-011-9914-8*

Ante, S. (2014), 'IBM Set to Expand Watson's Reach', *The Wall Street Journal* . In the "What's News" section of iPad editon.

Baker, N. (2013), 'Wrong Answer: The Case Against Algebra II', *Harper's Magazine* pp. 31–38.

Barro, R. & King, R. (1984), 'Time-Separable Preferences and Intertemporal-Substitution Models of Business Cycles', *Quarterly Journal of Economics* **99**, 817–839.

Boolos, G. S., Burgess, J. P. & Jeffrey, R. C. (2003), *Computability and Logic (Fourth Edition)*, Cambridge University Press, Cambridge, UK.

Bringsjord, S. (1998), 'Chess is Too Easy', *Technology Review* **101**(2), 23–28.
URL: *http://www.mm.rpi.edu/SELPAP/CHESSEASY/chessistooeasy.pdf*

Bringsjord, S. (2008), Declarative/Logic-Based Cognitive Modeling, *in* R. Sun, ed., 'The Handbook of Computational Psychology', Cambridge University Press, Cambridge, UK, pp. 127–169.
URL: *http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf*

Bringsjord, S. & Bringsjord, E. (1996), 'The Case Against AI From Imagistic Expertise', *Journal of Experimental and Theoretical Artificial Intelligence* **8**, 383–397.
URL: *http://kryten.mm.rpi.edu/case.against.AI.pdf*

Church, A. (1936), 'An Unsolvable Problem of Elementary Number Theory', *American Journal of Mathematics* **58**(2), 345–363.

Dalgaard, P. (2008), *Introductory Statistics with R (2nd ed)*, Springer, New York, NY.

Ebbinghaus, H. D., Flum, J. & Thomas, W. (1994), *Mathematical Logic (second edition)*, Springer-Verlag, New York, NY.

Eicholz, R. E., O'Daffer, P. G., Charles, R. I., Young, S. I., Barnett, C. S., Clemens, S. R., Gilmer, G. F., Reeves, A., Renfro, F. L., Thompson, M. M. & Thornton, C. A. (1995), *Grade 7 Addison-Wesley Mathematics*, Addison-Wesley, Reading, MA.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, W., Nyberg, E., Prager, J., Schlaefer, N. & Welty, C. (2010), 'Building Watson: An Overview of the DeepQA Project', *AI*

Magazine pp. 59–79.
URL: *http://www.stanford.edu/class/cs124/AIMagzine-DeepQA.pdf*

Ferrucci, D. & Lally, A. (2004), 'UIMA: An Architectural Approach to Unstructured Information Processing in the Coporate Research Environment', *Natural Language Engineering* **10**, 327–348.

Fischler, M. & Firschein, O. (1987), *Intelligence: The Eye, the Brain, and the Computer*, Addison-Wesley, Reading, MA.

Floridi, L. (2005), 'Is Information Meaningful Data?', *Philosophy and Phenomenological Research* **70**(2), 351–370.

Floridi, L. (2008), Data, *in* W. Darity, ed., 'International Encyclopedia of the Social Sciences (2nd ed)', MacMillan, Detroit, MI. A preprint is available at:
`http://www.philosophyofinformation.net/publications/pdf/data.pdf`.

Floridi, L. (2010), *Information: A Very Short Introduction*, Oxford University Press, Oxford University Press.

Floridi, L. (2011), *The Philosophy of Information*, Oxford University Press, Oxford, UK.

Glymour, C. (1992), *Thinking Things Through*, MIT Press, Cambridge, MA.

Greenberg, M. (2010), 'Old and New Results in the Foundations of Elementary Plane Euclidean and Non-Euclidean Geometries', *The American Mathematical Monthly* **117**(3), 198–219.

Knell, R. (2013), *Introductory R: A Beginner's Guide to Data Visualisation and Analysis Using R.* A Kindle book available from Amazon. ISBN: 978-0-9575971-1-2.

Kolmogorov, A. (1933), *Grundbegriffe der Wahrscheinlichkeitrechnung*, Ergebnisse Der Mathematik. Translated as *Foundations of Probability*, New York, NY: Chelsea Publishing Company, 1950.

Lewis, H. & Papadimitriou, C. (1981), *Elements of the Theory of Computation*, Prentice Hall, Englewood Cliffs, NJ.

Mattern, K., Shaw, E. & Xiong, X. (2009), *The Relationship Between AP Exam Performance and College Outcomes*, The College Board, New York, NY.

McKeon, R., ed. (1941), *The Basic Works of Aristotle*, Random House, New York, NY.

Minelli, M., Chambers, M. & Dhiraj, A. (2013), *Big Data, Big Analytics: Emerging Intelligence and Analytic Trends for Today's Businesses*, John Wiley & Sons, Hoboken, NJ.

Newell, A. (1973), You Can't Play 20 Questions With Nature and Win: Projective Comments on the Papers of This Symposium, *in* W. Chase, ed., 'Visual Information Processing', Academic Press, New York, NY, pp. 283–308.

Potter, M. (2004), *Set Theory and its Philosophy: A Critical Introduction*, Oxford University Press, Oxford, UK.

Robinson, A. (1996), *Non-Standard Analysis*, Princeton University Press, Princeton, NJ. This is a reprint of the revised 1974 edition of the book. The original publication year of this seminal work was 1966.

Sayood, K. (2006), *Introduction to Data Compression (3rd ed)*, Elsevier, Amsterdam, The Netherlands.

Skolem, T. (1934), 'Über der Nichtcharaterisierbarkeit der Zhalenreihe mittels endlich oder abzahlbar unendlich vieler Aussagen mit ausschlisslich Zahlenvariablen', *Fundamenta Mathematica* **XXIII**, 150–161.

Suppes, P. & Takahasi, S. (1989), 'An Interactive Calculus Theorem-prover for Continuity Properties', *Journal of Symbolic Computation* **7**, 573–590.

VanLehn, K. (2011), 'The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems', *Educational Psychologist* **46**(4), 197–221.

Woodford, M. (2011), 'Simple Analytics of the Government Expenditure Multiplier', *American Economic Journal: Macroeconomics* **3**(1), 1–35.

Woodger, J. H. (1962), 'Biology and the Axiomatic Method', *Annals of the New York Academy of Sciences* **96**(4), 1093–1116.