**The Necessity of Theory in Science**

**Or**

**Big Data is Anti-Science**

**John Day**

**2015**

There has been considerable hype surrounding Big Data of late as if it were something really new. The histrionics have gotten quite deafening. I have characterized the current fad as the 6th Generation of Big Data, starting with first generation in the 1830s. After a carriage accident made it impossible to go back to sea, Matthew Maury was made head of the US Naval Observatory in Washington D.C. There using the logbooks returned by Naval Captains after each voyage, he collated data and was able to discover previous unknown currents in the Atlantic and patterns in the wind that allowed shorter sailing times by days or even weeks. The second generation was Sears Roebuck, who in 1910 built a facility on the outskirts of Chicago to fill orders. At the time, Sears had no brick and mortar stores. The catalog was their Web. They were filling 100,000 orders a day and moving a million pieces of merchandise a day in 1910! (You could order everything from nails, clothes, carriages to an entire house! The houses (for which there was a whole separate catalog of styles) were pre-cut (not pre-fab) and shipped in several installments to give you time to build each phase.)

The 3rd Generation would be 1940s Bletchley Park and the advent of the computer. Von Neuman’s interest in computers was to get more data to see the patterns in differential equations he was working on. The 4th Generation would be the 1960s with Illiac IV and the advent of supercomputers, and the 5th Generation would be the 1980s and the establishment (in the US) of supercomputer centers. And now we turn the Moore’s Law crank yet again and we are at the current fad with racks of machines filling huge buildings and with millions of sensors spread around us.

But this last generation is the most dangerous, the greatest threat. Some have even called it a “new science” (if it is then so was the microscope was a new science) or the end of science (all we have to do crunch all this data and we will get the answers). It is closer to the latter than the former, but not for the reasons they think. Big Data is accelerating us toward stagnation. Let me explain:

As a grad student, I discovered Joseph Needham’s magnum opus, **Science and Civilization in China**. It isn’t just a book. It is a multi-volume (with some volumes having multiple books) encyclopedia of science and technology in China up to about 1750, when it becomes too difficult to determine what was purely Chinese and what was influenced by Western contact.

Why was I reading such things?

First of all, it was interesting! What other excuse does one need!?

Second, any system designer or architect must collect models to avoid the “If all you have is a hammer, . . .”[[1]](#footnote-1) syndrome. And the models and the accomplishments, I found in Needham were fascinating: a very different approach to many problems than found in the West.

A couple of examples will illustrate what I mean:

In ship design Needham points out that both East and West used nature as a guide. The West used fish; China used waterfowl: Much more appropriate for something at the interface of air and water. Fish are a good model for submarines, but ducks are a better model for boats.

In China, the axis of a windmill is vertical and the vanes hang down. Not only is the gearing simpler, but it is always in the wind. It doesn’t have to turn into the wind.

China had Pascal’s Triangle centuries before Pascal.

Seventy years before Vasco DaGama in 1497 clawed his way down the African coast and rounded the Cape of Good Hope to put into Mombasa on the East African coast, the Chinese Admiral Zheng He paid several visits to Mombasa with a large fleet of huge ships with water tight compartments and other advancements, just out on a good will tour to say the Emperor thinks all of you are wonderful and if you would like to send back tribute to the Great Ming Emperor that would be fine.

Interesting isn’t it?

It is appropriate that we are meeting in Portugal, which had such a major role in the Age of Discovery. Henry the Navigator’s great accomplishments earlier in the 15thC had left a legacy for Da Gama to build on that the Chinese didn’t have. As Needham points out, there was one thing missing in Chinese technology: There was no scientific theory.

It is all technique, technology; it is an artisan tradition, craft. What do I mean by scientific theory? Robert MacArthur, one of the founders of biogeography, distinguished Natural History from Science in that Natural History *describes* but Science *predicts*. The Chinese had certainly achieved a critical mass of knowledge that should have lead to theory. But for some reason (still debated by scholars), there was no theory. Some say, it was because they were so practically minded. However, because there was no theory, there was a tendency to lose knowledge: When Matteo Ricci, the first Jesuit into China in 1600, he initially thought they had brought the knowledge that the earth was round.[[2]](#footnote-2) When quite to the contrary, the Chinese had known it centuries before but the knowledge had been lost. But the lack of theory had another far worse consequence. By the late-Ming dynasty (16thC), stagnation had clearly set in. Artisan traditions are predicated on doing what had been done before; improvements come by trying things, not by using theory to point the way.

Needham attributes both the lack of theory and the stagnation to the fact that merchants had very little status in China, and virtually no power.  All power was with the Emperor. In other words, Needham saw commerce as the driver of technology and the reliance on government funding as leading to stagnation.  But as we have seen more recently, the short ROI of commerce can also lead to stagnation. Everyone is looking for a technology enhancement that will yield a quick result, rather than delving deeper for more fundamental results that could yield far more but may take longer and also threaten to undermine existing investment.

I would tell historians, that another reason there was no theory in China was that there was no Euclid. They would give me blank stares as if to say, “Huh!?” But historians don’t see Euclid as we do. Clearly not Euclid for geometry, the Chinese understood geometry quite well, *but as an example of an axiomatic system.*

As *we* all know, Euclid’s accomplishment is the Holy Grail of science.[[3]](#footnote-3)  The ultimate goal in any field is to be able to reduce it to a small number of assumptions from which all else can be derived. Newton did it for mechanics; Maxwell, for electricity and magnetism; CERN and the physicists are trying to do it for everything. Not that any scientist sets out to do that, but every scientist worth his salt is always open to that flash of insight that points toward a unification.

That of course begged the question: Why did the West have Euclid!? Why did Euclid do what he did? What pushed him to create such an elegant edifice!?

Of course, we are lucky just to have Euclid’s Elements let alone know anything about who Euclid really was, how the Elements came about, What made him want to organize things that way? Why was he looking for such an elegant solution? etc. Lost in the sands of time.

As it turned out, I didn’t need to know why Euclid did it to understand why. The insight to Euclid came, reading a Geometry book by Heilbron, where he notes that while several civilizations developed mathematics, only the West developed the concept of *proof*. The others have recipes, examples that are used as patterns, (dare I say algorithms?) but not *proof*. It is clear that the Babylonians had the Pythagorean Theorem, but they didn’t have a proof for it.

This answers the question! How?

What do you do when you challenge a proof?  You question the assumptions. Continually challenging the assumptions leads to the minimum set of assumptions that will suffice.  Hence, an axiomatic system. Then it is a short step to ask what results can be derived with just the assumptions, and then what do those enable and so on and you have the Elements!  (BTW, my favorite disposition of the development of meta-mathematics is Chapter 1 of Bourbaki’s Theory of Sets. It is delightful!)

Why is theory important to science? Bear with me.

I had been asked to write a review of a book for Imago Mundi, the premier history of cartography journal. Over the 2014 holiday break, I decided to knock it out.  The book was on Jesuit Mapmaking in Early 18thC China. (I have published a bit on this period.)  The book is primarily about the first major scientific mapping effort anywhere instigated by the Emperor Kiang-Xi and the resulting Atlas. But the book also discussed one of two well-known incidents in the late 17thC where the Jesuits had been pitted against the Court astronomers to see which could most accurately predict three astronomical events: a lunar eclipse, the length of a shadow cast by gnomon at a given time of day, and the relative and absolute positions of the stars and planets on a given day.  The Jesuits produced more accurate results than the Chinese Court Astronomers, resulting in their being put in charge of the Court Observatory in Beijing.

Why were the Jesuits’ calculations more precise? It certainly wasn’t because the Chinese couldn’t do the math to the proper precision. After all, the Chinese had been using the decimal system for centuries. (When discussing surds, Needham notes that the Chinese had adopted the decimal system so early it wasn’t clear they noticed that there were irrational numbers.)

Then why?

Because the Jesuits were using techniques developed with and backed by theory.  They didn’t develop the techniques or the theory. Others in Europe had done that. But the “theory” behind it had forced the Europeans to be more precise to back up what they knew, to look more critically at their work, to think more deeply about it, improve their arguments. Hence creating more precise techniques.

The Chinese, on the other hand, had a procedure to follow. They didn’t understand why it was correct other than it had always worked “well enough,” so why look further? *(Hmmm, where have I heard* ***that*** *before!)* They had been trained that it was the way to do it.  They just knew it worked. And, the procedure didn’t really indicate directions that would lead to how to improve it. (Needless to say, respect for authority and ancestor worship didn’t help in this regard.)

We are seeing the same thing in the systems side of computer science today and especially in networking, where it has been a badge of pride for 30 years that they do not do theory.  In 2001, the US National Research Council lead a study of stagnation in networking research, one quote from their report sums up the problem:

“A reviewer of an early draft of this report observed that this proposed framework – measure, develop theory, prototype new ideas – looks a lot like Research 101. . . . From the perspective of the outsiders, the insiders had not shown that they had managed to exercise the usual elements of a successful research program, so a back-to-basics message was fitting.” [1]

It must have been pretty sobering for researchers to be told they don’t know how to do research. Similarly, the recent attempt to find a new Internet architecture has come up dry after 15 years of work. The effort started with grand promises of bold new ideas, new concepts, fresh thinking, clean-slates, etc and has deteriorated through ‘we should look outside networking for ideas’ (a sure sign they don’t have any ideas when, in fact, the answers were inside as they always are); to ‘the Internet is best when it evolves,’ (they have given up on new ideas) to ‘we should build on our success’ (It is hard to get out of that box). When I asked my advanced networking class to read recent papers on the 6 efforts funded by NSF on Future Internet, after a chance to read some of the papers, their first question was, “These were written by students, right?” Embarrassingly, I had to reply that, they had been written by the most senior and well-respected professors in the field.

This is a classic case of confusing economic success with scientific success. They were focused on *what to build*, not asking the much harder and more dangerous question: *what didn’t they understand*. They didn’t question their basic assumptions. Even though, fundamental flaws were introduced as early as 1980 and made irreversible by 1986 and compounded in the early 90s.

On the other hand, our efforts, which have questioned fundamentals and forced us (me) to change long held views, have yielded new and often surprising result after new result: that a global address space is unnecessary; reducing router table size by 70% or more; recognizing that a layer is a securable container, greatly simplifying and improving security; that decoupling port allocation from synchronization yields not only a more robust protocol but more secure; etc.

Of course they have also shown that connectionless was maximal shared state not minimal; that of the four protocols we could have chosen in the 1970s TCP/IP was the worst, that of the two things IP does (addressing and fragmentation) both are wrong, that the 7 layer model was really only 3 (well, by 1983 we knew it was only 5), and much of what has been built over the past 30 years is questionable. Of 9 major decision points in the Internet, they have consistently chosen the wrong one, even though the right one was well known at the time.

There are many examples from networking, where not doing theory has missed key insights. A few examples should suffice:

1. It is generally believed and taught in all textbooks that establishing a connection requires a 3-way handshake of messages. However, this is not the case. In 1978, Richard Watson proved that the *necessary and sufficient* condition for synchronization for reliable data transfer is to bound three timers: maximum packet life-time, maximum time to send an ack, and maximum time to exhaust retries. The three messages are irrelevant. They have nothing to do with why synchronization is achieved. There are three messages exchanged, yes, but there are always 3 messages. They aren’t the cause. Watson then demonstrated the theorem in the elegant delta-t protocol. By not doing theory they missed the deeper reason that it worked and missed that the resulting protocol is more robust and more secure.
2. Many people will tell you that network addresses name the host. That naming the host or device is important. (Several of the projects noted above among them.) As it turns out, it may be useful for network management, but not for communications. In fact, it is irrelevant for communications. If you construct an abstract model and carefully look at what has to happen, you see that what the address names is the “process,” the locus of processing, that strips off the header of the packet carrying the address. The host is merely a container. Well, you might say, there are places where there is only one “process” stripping off that header, so it and the host are synonymous. Yes that case exists and in large numbers. But it is not required to exist in all cases and doesn’t in some very significant ones. By not doing the theory, they missed this insight, which made dealing with VMs very messy.
3. In 1972, we first realized that in peer networks, the “terminals,” now computers, could have multiple connections to the network. In all previous networks, the “terminals” were very simple and having only had one was all that was possible. The advantage of a computer having more than one link to the network is obvious: if one fails it still has connectivity. However, addresses in the ARPANET like all previous networks named the wire to the “terminal,” i.e. the interface. If one interface went down, *the network had no way to know that the other address went to the same place.* To the network, it appeared to be two different hosts. One could send on both interfaces, but not receive on both (not without re-establishing a connection to use the other address). Addressing the interface made addresses route-dependent. *Addresses had to be location-dependent but route independent*. The solution is apparent if there is a theory: Which we had in Operating Systems. In operating Systems, Application names designate what program and are location-independent, Logical addresses provide location-dependence, but route-independence (independent of where in physical memory), and physical addresses are route dependent addresses (dependent on accessing the memory). Naming the node, not the interface solved this problem. Not only did it not cost anything but it is significantly less expensive, because it requires between 60% and 90% fewer addresses and router table size is commensurately smaller. All other network architectures developed in the 1970s/80s got this right, only the Internet, which doesn’t do theory, got it wrong.
4. But we still thought that addresses could be constructed by concatenating an (N-1)-address with an (N)-identifier. This seemed natural enough. After all, files were named by concatenating directory names down through the tree with the file name as the leaf. That was until in 1982 when we started to look at the detailed theoretical model of what would happen if we did that. It quickly became apparent that it defined a path up through the stack. Concatenating the addresses made them route dependent. *Precisely* what we were trying to avoid. The address spaces at each layer have to be independent. Of course it was obvious once you remembered what Multics called a filename: a *path*name! But it was doing the theory that lead to recognizing the problem. So why does IPv6 embed MAC addresses in the IPv6 address? Because they don’t do theory.

There are many more examples. All cases where not doing theory lead to missing major insights and improvements. But notice that today we are doing the same thing the Court Astronomers were doing. Our textbooks recount how things work today, which students take as the best way to do things. We teach the tradition, not the science. We don’t even teach how to do the science. We don’t teach what needs to be named and why. Watson’s seminal result is not mentioned in any textbook. (One young professor asked me, why he should teach delta-t if no one is using it. (!) I almost asked him to turn in his PhD! We aren’t teaching the fundamental theory, we are teaching the tradition.)

For a talk on this problem a few years ago, I paraphrased a famous quote by Arthur C. Clarke to read:  Any sufficiently advanced craft is indistinguishable from science.”  (Clark said, “Any sufficiently advanced technology is indistinguishable from magic.”) We are so dazzled by what we can do; we don’t realize that we are doing craft, not science.

Big Data is the same thing only worse.  Big Data is accelerating the move to craft and is sufficiently sophisticated to appear to be science to the naive. *Correlation is not causality.* We create algorithms to yield results, but do we have proofs? Big data is supposedly telling us what to do without telling us why or contributing to a framework of theory that could lead to deeper more accurate results and likely even deeper insights.

Even Wired Magazine called Big Data the End of Science. Although as usual, they didn’t realize what they were advocating stagnation. Of course, it is always the case that every field goes through a period of collecting a lot of data before it becomes clear what is important and the theory is. This has happened before. But what hasn’t happened before is to advocate that we don’t need theory. It is putting us in the same position as the Court Astronomers in 17th C China. And the rate of change and adoption is far faster now than then.

There are those who claim it is a \*new\* science!  When it is actually the greatest threat to science since the Catholic Church found Galileo guilty of proving a heathen (Aristotle) wrong. (I never have understood that one!)  The Scopes trial was more circus than threat, though that may have changed in the backward US.

We are taking on the same characteristics seen in Chinese science in the 17th C.  It isn’t pretty and it isn’t just networking.  Read the last 5 chapters of Lee Smolin’s The Trouble with Physics.  He sees it there!  And others have told him they are seeing it in their fields as well.

Big Data has us on the path to stagnation, if we are not careful. Actually, we are a long way down that path.

1. **Looking over the Fence at Networking,** Committee on Research Horizons in Networking,

National Research Council, 2001.

2. Needham, Joseph. **Science and Civilization in China**, Cambridge University Press, (Vol 1- Vol. VII, Book 1) 1965-1998.

3. Smolin, Lee. **The Trouble with Physics: The Rise of String Theory, the Fall of Science, and What Comes Next**, Houghton-Mifflin, 2006.

1. As Mike O’Dell says, “When all you have is a hammer, everything looks like your thumb!” [↑](#footnote-ref-1)
2. This is not the huge discovery we generally think it is. Since ancient times among the educated classes, it was well known. One merely has to watch a ship sail over the horizon and notice that the hull disappears before the sails do (the origin of the phrase “hull down”) to know the earth is round. No one funded Columbus not because they thought they could sail off the edge of the world, but because they knew it was round, knew its circumference and knew they didn’t have ships with sufficient range to make the voyage. Columbus fudged the numbers to make them look feasible, found someone with money that believed his math, and then got very lucky when there was a continent in the way! [↑](#footnote-ref-2)
3. A carry over from when the distinction between math and science was not as clear as it is now. [↑](#footnote-ref-3)