THE VALIDITY AND RELIANCE OF BIG DATA PROJECTS

G. Donald Allen
Department of Mathematics
Texas A&M University
College Station, TX 77843
gdonaldallen@gmail.com

Abstract. We put forth questions as to the potential problems with big data, their applications, and their implications in our world. Our particular interest is with distributed big data projects. The increase in available data has been occurring for centuries. One of the earliest big data practitioners, Johannes Kepler (1571-1630), used decades of Tycho Brahe's planetary measurements to render his laws. Modern projects now use masses of data that must be rendered with statistical or modeling tools by data engineers. These are among the latest innovative applications of computing. Of course, bigger and faster computers can and will push to new limits ordinary and well explored topics, and this will so continue for centuries. We are entered into a discussion about the use of computers to solve new, even revolutionary, problems of this world, particularly those involving big data. The problems themselves have become more complicated, often with no clear answer provided, and just as often without the problem even being well defined. We live in an age of wicked and impossible problems. We do what we can, but one clear problem, the one discussed here, is that offered solutions may lead us in false directions from which recovery may be difficult.

1. Big Data Successes. These involve the search of massive data bases for patterns, predictive analytics, analysis, and comprehension. Some seem rather mundane, but for a company to increase sales or decrease costs by as little at 5% could mean corporate survival. The investment in the big data enterprise seems to be a necessity in every competitive environment, these days. Yet many experts caution the importance of having a clear goal in mind. They advise us to use the right data, to be aware of issues with third party data, not to build a black box – wherein there comes to pass a trust and reliance on learning algorithms to make sense of the data – but without understanding. The black box exacerbates the natural asymmetry between people and data. The examples below are but a mere sample among hundreds. (Hassler 2015, IBM, 2011, Laskowski 2015, Henschen 2013)

- Customer profiling for targeting merchandise and understanding customer buying patterns. - Amazon, Macy's, and many others
- Detecting fraud for insurance claims – Metlife Auto and Home. As well, Infinity Property & Casualty Corp used dark data to detect fraud in adjustor's reports, realizing $12m in subrogation recoveries.
- Flight analysis – Using "Flightscope," a software tool to detect indicators of mishaps – US Navy

- Airline booking system – SABRE (one of the earliest big data systems)
- Facial recognition - for a variety of applications mostly at airports and other public venues. (a large mathematical research venture as well)
- Predictive Policing –PredPol, a software tool that enables law enforcement agencies to better prevent crime in their communities by generating predictions on the places and times that future crimes are most likely to occur – multiple police agencies
- Machine performance, a supermarket chain collected 70 million data points from its multiple refrigeration units, looking for clues as to when machines may need maintenance – TESCO
- Customer loyalty – looking for indicators predictive of customers dropping their account, AMEX.
- Analyzing student success and study – a variety of methods to examine students time on task and visual representation of this information. This is an example of the use of soft data, or evaluative, or imprecise data to be discussed in Section 4.
- In the NBA, every team has special cameras that create 4 million data points for every game. The hope and promise is that team managers will see better ways to manage their resources. Sports analytics is an entire college major.
- In medicine, new self-monitoring devices will help collect our vital stats into powerful databases, allowing us to prevent or predict health problems as well as treatment success or failure

Big data successes have spawned an optimism in their value; it has generated an avid exploration and belief that big data applications can and will ferret out solutions to problems, finding even the slightest edge in a highly competitive world looking for the most elusive solutions. Entire collegiate programs have been created to produce experts in this redoubtable art. Or is it a science?

2. Problem solving. The human brain is a marvelous organ. It is designed for but one thing: survival of the body, and survival means successfully solving a non ending stream of problems. Ever thought about how you figure things out? Your marvelous brain has it covered. Indeed, you use seven separate systems to make conclusions, resolve questions, solve problems, and just about everything else. These systems, with the acronym BRAIPE, are beliefs (B), random (R), analytic (A), intuition (I), programs (P), and emotions (E). The seventh is innovation, the brass ring for business. But what is it? We could say an entity is self-aware and conscious if it can innovate on the basis of itself or on stimuli it received. Innovation is beyond mere problem solving. Stimuli or thoughts occur and the entity can somehow innovate a new idea or recourse over and beyond the expected response. It implies a higher order of thinking, well beyond the single-celled organism, and well beyond the comprehensive climate control systems, well beyond the dog or cat, and well beyond other species. The climate control system includes only the "P" above. Dogs and cats may respond to four of them, the "I", "R", "P", and "E." Yes, they are emotional, too, and think emotionally at times!

But innovation is also well beyond many humans. All this is a gentle attempt to resolve an ancient question. The question remains unsolved. It seems that the definition of consciousness is rather undefined (because there are so many definitions), and that is if the question of consciousness is well-posed in the first place. We identify four types of innovation.

- Device – save time, money, health, increase food, decrease pollution
- Process – way to think, way to organize, way to produce
- Concept/paradigm (paradigm shift) – a new idea to affect how society or its institutes function
- Principle – a statement given as true, upon which structure is built

The first two are clearly within the domain of computing, while the second two remain only within the realm of human thought.

A newer type of problem is now present. Called wicked problems, they often have no clear solution, no clear statement, and no stopping time for arrival at the solution. Big data is sometimes used to approach them, or to find a process to produce results. For more information on wicked problems see (Allen, 2014).

3. The Current State. Computing furnishes us with a generalized tool for doing new things, though only things for which quantification, numbers, and information are at play. Practitioners have now spent several decades devising ways that quantitative data and information can be made qualitative. We say *some decades* because that is the currently longevity of computing machines. As usual, it was the real world that spawned the need for machine encoding and processing of large quantities of data. Beginning with the complexities of processing census data, Herman Hollerith (1860–1929) devised the punched-hole card that could be machine read and input data analyzed using mechanically complicated but pre-electronic simple machines. These were years before the computer was commercially available. The IBM 740 machine from the 50-60's is one example, it with programmable hard-wired boards. Indeed, the most striking aspect of computers is the concept of a stored program originally, and more recently the adaptive program.

Our mission here is not to recount the history of computing, but rather to suggest where it is going, and importantly where and when it may reach an end – data-wise. Yet it is important to note that the idea of computers evolved from the works of Turing, with the Turing machine leading to programmable computers, Cantor, with set theory, and Gödel, on the incompleteness of our mathematical systems which concluded there are problems within any mathematical system that cannot be solved. This is to say, that within any logical system, there must be impossible problems solvable within it. The implication is the system must be expanded, but then the expansion will lead to new impossible problems. This could be a spiral situation. Let's look are where we are.

3.1 Numerics. The world functions on mathematical models of reality and essentially what they predict, what they can show, and importantly what they can solve. Astonishingly successful, numerics generate a wealth of information and lead to solutions of problems formerly impossible to consider, if not pose, well beyond the scope of all problems even two centuries ago.

3.2 Data. The accumulation of population, climate, weather, and other statistical data has amassed at such a rate and have exceeded all estimable bounds that it cannot be any longer processed without computers to help. Just imagine a large bank trying to keep track of its customers with index cards and phone numbers on a Rolodex. Below, we take up the similar-sounding topic of "big data," quite another issue.

3.3 Modeling. The most successful technique for the explanation of phenomena, following Sir Issac Newton's theory of planetary motion and Euclid's *Elements*, has been the model. Clearly, these were mathematical models, but the idea has caught on – despite philosophical debate. Models and making models is the active task of almost all innovators in almost every human endeavor. We now have all sorts of model s including statistical models applied to data sets. Most mathematical models have no tight, closed solution coming by way of a formula. Most problems involving these models require intense numerical computing. Indeed, in the past four decades an entire field of numerical analysis has evolved simply to help provide solutions to such problems. Statistical models are models generated from small and more recently massive data sets to correlate and thus to explain cause and effect. However, there remains the fantastic misuse of models, data, and patterns recognized to derive causes.

> *Modeling has become so successful we have come to accept models and their results without really understanding either.*

Statistics, a key component of both numerics and modeling, has played a pivotal role in bring to heel the massive data channels confronted with analysts. Statistics has also allowed a multiplicity of simulations to study particularly troublesome data by isolating various factors. Simulations in polling, testing and sports have emerged as significant applications.

3.4 Big data. This is the newest and a most exciting application for computing. The premise is we can make discoveries far beyond the efforts of the individual investigator. We have information available that just may solve problems of hunger, climate, medical issues, and education. They involve massive volumes of data with one goal being to determine patterns within. Big data is characterized by the three Vs: volume (too much data to handle easily); velocity (the speed of data flowing in and out makes it difficult to analyze); and variety (the range and type of data sources are too great to assimilate). A classic definition is that "data is big when its size becomes part of the problem." A variety of taxonomies exist for big data types. The following is a composite of a few.

- Facts, (i.e. exact numbers, e.g. medical records, demographic data).
- Subjective (e.g. boredom in class, job performance, polling) – Evaluative data is an important subclass. It is subjective data cast as quantitative. As such it has associated confidence intervals – difficult at best to even estimate and a perennial problem with polling. Their meanings can change over time; this is significant. Also, evaluative data suffers many of the same problems as testing unfairness.
- Social networks (e.g. blogs, comments, personal docs, searches, Facebook, Twitter) – also subjective and analysis often requires language processing tools.
- Sensor data (e.g. traffic, computer logs, mobile, images, body cams).
- Electronic files
- Broadcast files

Big data applications usually have a couple of objectives in mind. These include all the various forms of data mining plus prediction, causality, identification of significant factors such as patterns, and courses of action.

All of these apply to educational problems, the pathways to which may range from modeling to simulations to visualization, and more. The fact, to be developed in Section 4, is that along the pathways, there is someone that needs to understand the mechanisms of machine activity, and how it fits with the models, and the statistics.

There are specialized software tools (RapidMiner, Gephi, SAS, etc) designed to do just that. Remarkable results have been achieved. Patterns in medicine, in finance, in education are just a few of the bigger topics that have been rendered to previously undiscovered and even unconsidered conclusions. Perhaps this justifies big data projects, or is it just the beginning? The tools for analysis follow rather standard constructive models. When one is found there is celebration and delight. It is then used for predictive ends. Often the models are used to predict backwards in time to validate their predictions toward the future. Herein lies an issue. Unless the reality is time-reversible, validation of the model is suspect. Indeed, it can lead to conclusions that are simply false. We could call this a *modeling a false positive*. Having tools to process big data (Hadoop, et.al) does not imply resultant analyses have any meaning.

Here is one new example. In a Cornell study of Facebook (Wynick 2013) pages of more than one million participants, there has evolved though big data analysis a predictive of relationship breakups. The study suggests that their models can predict a breakup before even the participants are so aware. They show, though big data modeling, that if you both all have the same set of friends, this is an indicator of a possible breakup. But this is just a model – NOTHING MORE. In fact, there may be other factors of the personality types of people having same sets of friends. What conclusions should be determined? A typical method is to find some correlation between two data set, and then to construct a causality of one to the other. In fact,

one politician recently noted a positive correlation between climate change and the rise of ISIS, concluding the former is the cause of the latter.

We would be remiss in not mentioning the US government big data project of collecting all phone call meta-information to determine clandestine threats to the homeland. However, it isn't clear it is known how to do this effectively. It seems clear that all such analyses are overwhelmed by each of the three Vs. This is somewhat reminiscent of early attempts at weather forecasting using models by Lewis Fry Richardson (1881-1951). Using mathematical models, forecasting just the next day's weather would require about 300 years of computing time (computing time included people-work). By 1950, this was improved to 24 hours of computation for a 24 hour forecast with the use of the Eniac (Lynch 2008). Happily, things have improved.

4. Distributive Big Data. This is another form of big data, only now emerging. We consider the idea of the *distributed big data project*, where the distribution is in disparate data sources, disparate experts who understand and manage them, the project director who more-or-less knows the area, the programming tools to process the information, and the nature of conclusions sought. As we will see, this chain has many weak links, the weakest link being at the end of the chain, i.e. the user.

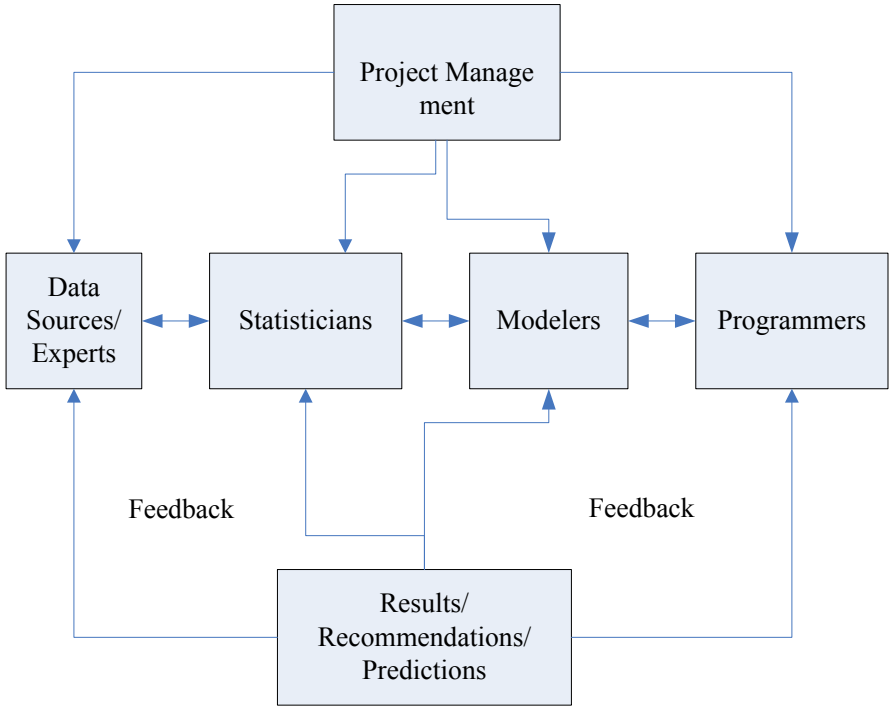

Figure 1 – Distributed Big Data Project

It is a consequence of an even greater complexity that no one in the loop has operational much less theoretical command of the input, the models, the statistics, and

the mathematics involved. It is important as it posits problems that will be difficult for computers to handle and for humans to comprehend. Many times when confronted with the incomprehensible, we often accept that offered with little basis. As Frederich Hess stated, "Educators have made great strides in using data. But danger lies ahead for those who misunderstand what data can and can't do." (Hess 2008).

We focus on a big data project pertaining to the schools, indicating that distributed big data applications may channel institutions in false directions. These comprise big data applications where the contributors are experts at only a single phase of the project. The key challenge here is in getting people with "the triangle of skills" - the right combination of data knowledge, computer science, and statistics, says Terence Siganakis, data scientist at Growing Data Consultancy. This is the beginning of distributed big data projects. Indeed, this triangle may be better visualized as a polygon or even polytope. The schools have all these components but seldom is there a viable triangle of skills activated toward any end.

It may be best to see this as a flow schematic as depicted in Figure 1. Understanding these types of applications often involves a *data lake* of information. This means there are multiple data bases, about each of which there may be only single expert on the design staff, if that. The distributed big data project not only has distributed data sources, but distributed and often orthogonal, even conflicting, expertise.

As is apparent, there are many possible flaws in the general schema. Examples: How well is the feedback loop completed? How effective are the inter- and intra-team/component communications? Do the results make sense? Are results consistent with known information? Do the project component sub teams understand each other? Most of these seem somewhat dependent on the project management[1].

However, other questions and issues appear. We list only a few.

a. How do we know the units are compatible or blend in any way with evaluative data? For example, suppose student boredom levels are mixed with teacher efficacy measures. Each has a value of some sort in the data base. How may these values to be related?
b. How is the model checked for validity? For what types of circumstances are results unreliable?
c. How much does management know about the technical components of the project? When a commercial firm is involved results based projects, it must be sold. The research must be concluded and the model made permanent. The product must be rolled out.

[1] Note one reason the high energy physicists have been so successful is that the practitioners are expert at almost all aspects of the projects, from the theory, to the evidentiary processing, to the coding. Physicists are some of the best coders in all of science.

d. How does the school manage all this data when the sales engineers have left the site? Does the administration understand what they have signed on to?

Other problems pertain to the more technical issues such as logical, sustainability, reliability, relevancy, statistical, linearity, validity and more. Examples:

1. What happens when the model based on big data is applied and there results changes the population? For example, students become aware that time on videos is measured and can affect their grade. They adjust, of course. But is the model adjusted? Reliability issue.
2. What happens when people apply the model without a basic understanding of what data is used? Application issue
3. Does the project use "red-herring data," i.e. extraneous data that distract or divert attention from the emerging story line contained in the data? Relevance issue.
4. How do confidence intervals change within the model as data sets are included? Can confidence intervals even be developed? How does the end user reconcile multiple confidence intervals for various components of the project? Reliability issue
5. What happens when assumed independence is actually not true? Statistics/model issue
6. Are the data compatible? That is do units match up, or scales of measurement. (Recall the spectacular NASA Mars mission failure where two expert contingents used different, English and metric, units with project failure the result.) Validity issue.
7. How does the value of the data change with time? In many cases, it is the time changes we look for. But when the data becomes part of a time independent pool, there is the question. Time evolution and sustainability issues.
8. Do correlations within and between data structures become a substitute for truth? Logical issue.

Finally, there are even philosophical implications, or perhaps better put philosophical manners by which to consider big data applications. First of all, projections, patterns, causality, and the like as developed within the scope of a big data project become themselves a scientific theory, subject to *falsifiability*. According to American philosopher Karl Popper (1902-1994), scientific theories formulated as general laws can never be verified definitively. However, they can be falsified by only one observation, i.e. the counterexample. This implies that a theory may be regarded as 'more' scientific if it is richer in conclusions and provides more opportunity to be falsified. Thus it can be said that the amount of empirical information conveyed by a theory, or its empirical content, increases with its degree of falsifiability" (Popper 1934). Since Popper there has emerged a philosophy of information (Adriaans and

van Benthem 2008a,b; Floridi 2011). Just in its birth is an offshoot, the *philosophy of big data*.

One question about an operational big data project as a theory, such as determining at risk students or ineffective teachers, is how great its degree of falsification. However, this may be impossible because

a. There is no single person that fully understands all the components of a big data project, so as to find any item to falsify. None of the data component managers or contributors can know very much beyond their contributive expertise.
b. The end-user must function on beliefs that the full system is functionally accurate, and in almost no way can challenge its veracity.
c. The manner in which subjective data may be constructed or measured may change over time, or with the evaluator.

Many basic questions about big data are certainly unresolved. New questions are emerging as fast as applications widen. It is one project for a retailer to analyze account profiles to target spending and product preference habits. It is another to use factual and evaluative data to make predictions and prescribe courses of action that impinge a most fragile population. At this point we just don't know if we are confronted with big data, big bad data, or bad big data.

> *The distributive big data project staff is one with actors each having only partial information about components, and with no one fully comprehending them all.*

It is very likely some distributed big data project developed in the near future will fail in a spectacular fashion. Projects that use subjective data or those that use rather diverse data sources are particularly vulnerable.

5. Conclusions. Distributive big data projects seem to suggest big problems are just ahead. Many of the so-called distributive big data projects by design must be of the "black box" variety. As noted earlier, experts have cautioned against using them. When commercial firms are involved, the project must be brought to market. Even with complete instructions provided, the person on site managing the project will know less and understand less about how parameters are measured and used than even the corporate development team. Projects in education suffer such problems more than more analytical users such as hospitals, manufacturers, and others. We see most clearly how the subjective factors enter in when looking as election polling results – not exactly big data. Each of the firms employs ever so slightly different questions, sampling techniques, and analysis methods. Yet published results can be widely different. In fact, there has emerged in recent years a new firm, Real Clear Politics, which reports averages all the polls.

We are fully engaged with big data and the hopes offered. The co-mixture of numerics, modeling, mathematics, and statistics to refine big data may give assurances of a clearly defined reality. Yet, the overall lack of clarity has created obvious internal conflicts and for some true anxiety about validity and reliability. The dark side is what happens if society becomes overly intoxicated with its models, and begins to reinterpret their results as facts, much less truths.

References

1. Adriaans, P.W. and J.F.A.K. van Benthem, 2008a, 'Information is what is does', in Adriaans and van Benthem 2008b.
2. Adriaans, P.W. and J.F.A.K. van Benthem, (eds.), 2008b, Handbook of Philosophy of Information, Elsevier Science Publishers.
3. Allen, G. Donald., (2014), Challenges to Computing, Recent and Innovation Trends in Computing and Communication (IJRITCC), Volume 2, Issue 11.
4. Floridi, L.,(2011), *The Philosophy of Information*, Oxford University Press.
5. Hassler, Susan (2015) Big Data is Transforming Medicine, IEEE Spectrum, 28-May.
6. Henschen, Doug. Big Data Success: 3 Companies Share Secrets, Information Week, October 4, 2013. Online: http://www.informationweek.com/big-data/big-data-analytics/big-data-success-3-companies-share-secrets/d/d-id/1111815?
7. Hess, Frederich M. (2008), The New Stupid, *Educational Leadership*, 66, 4, 12-17.
8. IBM, (2011), *IBM Big Data Success Stories*, Online: ftp://ftp.software.ibm.com/software/data/sw-library/big-data/ibm-big-data-success.pdf.
9. Laskowski, Nicole. (2015). Ten big data case studies in a nutshell, TechTarget - SearchCIO http://searchcio.techtarget.com/opinion/Ten-big-data-case-studies-in-a-nutshell.
10. Lynch, Peter (2008). "The origins of computer weather prediction and climate modeling," Journal of Computational Physics (University of Miami) 227 (7): 3436.
11. Popper, K., 1934, *The Logic of Scientific Discovery*, (*Logik der Forschung*), English translation 1959, London: Hutchison, 1977.
12. Wynick, Alex, 2013http://www.mirror.co.uk/news/technology-science/technology/facebook-can-predict-your-relationship-2659859