

WHY BIG DATA IS A SMALL IDEA ... AND WHY YOU SHOULDN'T WORRY SO MUCH

Stephen P. Needel

INTRODUCTION

I recently presented a paper (Needel, 2013) in which I, somewhat facetiously, looked at the questions surrounding Big Data through the prism of religion. My reason for taking this perspective was twofold. First, it let me be funny, and it's always more satisfying when you can amuse an audience while making a point. Second, and much more important, is that there is a loud, vocal, swarm of converts to the Big Data movement. This group is almost apostolic in their fervor of telling us why Big Data will fundamentally change how we do marketing research, how we think about marketing research, and how marketing research is viewed in the corporate world. As they go on and on, which I'll try not to do today, I'm usually thinking, "not so much".

The proselytizing began, in part, at a panel discussion at last year's ESOMAR Congress (Passingham et al, 2012). While ostensibly about data privacy, much of the talk and the follow-up social media posts focused on claims that have since become a mainstay of Big Data adherents' manifesto. In particular, there is the belief that the marketing research world will be overtaken by interlopers who are coming in through the back door using Big Data as their entry pass. Many, including Greenbook blog's editor, Lenny Murphy (2012), have suggested that companies like IBM, SAS, and Google will be taking control of the industry (indeed, he hints at times that maybe they already have).

The attack on marketing research is two-pronged. The invader theme represents one side of the "death of marketing research" prediction. The other point of attack depends on the 3Vs of Big Data – Volume, Velocity, and Variety (Laney, 2001). The argument many make is that the amount of data, the speed at which the data becomes available, and the variety of types of data all will redefine marketing research. In 2012, P&G took the 3Vs to the extreme, saying they wanted this data analyzed as soon as it was collected and available on marketing teams' desks instantly, so that they could react instantly. The absurdity of this proposition should be obvious to anyone in marketing or marketing research. Let's ignore the processing problem, for which there is a whole separate body of literature on hardware and software. Instead, let's point out (a) the inherent dangers of looking at sales data in real time and trying to take action on that and (b) whether anyone actually has the ability to take action on a real time basis in the CPG world. My view, again, is, "not so much." And there is little evidence, according to conversations I've had with Chris Murphy of InformationWeek magazine, that P&G has made any progress in this direction.

Here's where Big Data preachers would like you to think we are going. We will shift to an ideographic (rather than traditional nomothetic) approach to marketing. They believe that Big Data contains enough information about a person to predict their behavior with a significant level of statistical rigor. They espouse modeling at the individual level as the promise of Big Data today, envisioning a world where all data is connected and massive computers with magical powers reveal patterns unseen by the naked eye or even the cleverest of researchers. These patterns, combined with mobile devices, will one day allow marketers to deliver messaging person by person at just the right time. This is the dream.

My big point today, in keeping with the Congress' theme of "Think Big", is that Big Data's promise relies on a set of assumptions, none of which are valid. In fairness, these assumptions might prove to be valid in the future, but today, "not so much". Let's look at these assumptions:

- The world is deterministic.
- All causal factors and end results (purchases) are known
- A paradigm exists for investigating and interpreting
- We are competent to deal with the results

We'll consider each of these assumptions in turn.

DETERMINISM

Big Data, as do most versions of science, assumes a deterministic view of the world, that a person's behavior is predictable. We all believe this to some extent – otherwise we wouldn't be in the research business. Science, by definition, makes the assumption that phenomena are explainable and predictable, and marketing research is principally science. But let's go deeper into the implications of this assumption. There are two criteria necessary to build a deterministic model; consistency and homogeneity.

In order to build models of shopping behavior, we need to assume that a shopper's behavior is consistent over time. When it is not consistent over time, then we need to assume the change in behavior has a recognizable and measurable cause. Here's an example: I buy Cheerios almost every time I buy cereal because they are [supposedly] heart healthy – they are very low in cholesterol. Every few months, though, I switch to a box of Cocoa Puffs, because they go on sale and I really like them, even though they have no nutritional value whatsoever. My shopping behavior is consistent (I buy Cheerios) and my deviations from consistency (I buy Cocoa Puffs) are explainable. Here's where the problem comes in; occasionally, I buy Apple Jacks, and for no discernible reason other than I didn't buy them the last time and they are really tasty. In situations like this, where variety seeking may be a standard driver of category purchasing, modeling my behavior will produce lots of error.

We also need to assume that the determinants of shopping behavior operate in a homogenous manner. We can take homogeneity at many different levels; across categories within individuals, across individuals of some cohort within categories, across categories and individuals. At some level of specification, we need to build a model of response to stimuli. For example, we might build a model of exposure to advertising that says I won't switch to a new cereal until I've seen the TV ad at least three times. Or we might determine that I don't switch to anything new without seeing three ads. Or we might build a model that says nobody switches to a new cereal without at least three exposures to a TV advertisement. Or, we may build a model that says nobody switches from their current anything without at least three exposures. The point is that whatever the determinants are, they operate consistently and homogeneously enough at some level so that you can actually build a model of that behavior and use it for prediction.

We have little reason to believe that either consistency or homogeneity holds at the individual level. People are inconsistently consistent – in some areas they are extremely brand loyal and in others either not at all or loyal to a set of products. Anyone who has done "shop-alongs" has heard the answer to why a product was chosen, "gee, it looked good/interesting/tasty" repeatedly. All of that is indicative of inconsistency, which will keep models of individual choice forever unsatisfying. We also have no reason to believe in these conditions holding at any homogenous level. Our industry has found prediction to be well beyond our reach for most FMCG products. When 80% of new product introductions in the U.S. fail and we cannot predict trade promotion impact from the reams of scanner data we have, we need to question whether we will ever have models that predict consumer choice sufficiently at any analysis level.

This does not mean research is doomed – but it may mean that modeling (rather than experimentation) may not be as fruitful as Big Data proponents would like you to believe, and certainly not at the individual level.

DATA AVAILABILITY

Big Data assumes sufficient acquisition of all the causal factors involved in making a decision and this is unlikely to be true. Brian Singh, in his 2013 Greenbook blog, is excited by the fact that we may now have all the behavioral data for "the desire-awareness-intention-engagement-purchase-loyalty cycle". Like most True Believers, he skips over the evidence of this claim – we don't have to be so hasty. In building a model of consumer (or shopper) behavior, we need to know:

- What media has been received, whether via broadcast or search
- What search processes precede the purchase decision
- What has been purchased

Media Exposure

We might know the external antecedents of a purchase decision, such as media exposure and media engagement, if someone were to a) completely abandon all personal privacy options and b) only used an internet-connected device to interact with media. Yes, we understand there are people with tablets or smart phones who watch TV, listen to the radio, read magazines and newspapers and surf the web only on this one wireless device. And yes, these people may not care that their every touch is recorded somewhere. For these people, we may have a full reckoning of their exposure to various persuasive messages, including social media. Such a complete record (and complete is a requirement for Big Data to be accurate) is unlikely to exist for most of the population we might want to model.

Search Processes
The cycle assumes that shoppers are involved in a monitored search process – otherwise, how would we know what desire, attention, and engagement levels exist. I use Singh's path-to-purchase, but there are others with the same concepts (e.g., Court et al, 2009). For those of you in our industry who are involved in research on durables or services, pardon me for a moment while I talk to the rest of us. The idea that we spend a lot of time searching is not so true for much of what we shop for in our daily lives. We make many more decisions about grocery products than we do about all other categories combined. With that in mind, think about how often you are involved in search behavior for toothpaste, toilet paper, and breakfast cereal. This is not an activity which many people often engage in. So for much of what we purchase, we don't know the intra-personal aspects of the decision.

Purchasing
Purchase and loyalty are known to those outside the purchaser-retailer relationship only when the purchaser has consciously given up the right to privacy. The major retailers, both online and off-line, are not telling anybody else what you are buying – you, the purchaser, have to choose to make that data public. Yes, I know they may be aggregating data or making it anonymous – for the purpose of Big Data and the path-to-purchase, that's not very useful.

Even within a retailer, critical mistakes can be made by assuming a level of loyalty that isn't there. While there is much to be learned from retailer loyalty programs, they only reflect the purchasing from a sub-group of shoppers from one chain. In the US, it is common for people to shop two or three different grocery chains, especially if you consider Wal-Mart and Target to be grocers in their super-centers. There is a lot of data that is not going into a model under these circumstances.

Mashing the Data
Here's the final part of the problem that I see with today's conception of Big Data. To model my behavior, you would have to have a radio meter in my car, the ability to hack into my cable provider's files to see what I'm watching on TV (assuming they even keep this), a website tracker on my three computers that I use regularly, eye-tracking data from my morning newspaper reading, access to my Amex and Visa cards and my frequent shopper cards at three different grocery stores and two pharmacies in order to have any shot at modeling my FMCG shopping behavior. Big Data only works if it has all the relevant data, and it is unlikely that will ever occur.

Of course, this points out two bigger problems. First, how does all the data fit together? Let's ignore for now the question of whether the data is structured or unstructured. Unless the data is all collected by one device, there may be no way to put it all together. Do you match on last name, last name and address, and ID number like my Social Security number (which would be illegal for companies I do business with to provide)? We can talk all we like about the volume of big data, but the volume is only relevant if it fits together to provide a bigger picture of the shopper. We don't know how to mash this yet.

The second problem is, I predict, that there will be a backlash by consumers against the collection of all this data. Firefox's plan to block web browser cookies may be the first major indicator that this is a problem from the shopper's perspective. The well-documented story of Target Stores and the Pregnant Teen case (Duhigg, 2012) is a great example of how intrusive Big Data can be; a great question by the marketing team, a great piece of statistical work by the researcher, and a program execution gone horribly wrong. The disclosures this spring of the US's collection of metadata has many outraged and many more shocked at what information exists in our connected world.

PARADIGM
Big Data assumes we know how to ask the proper questions and this may not be true – we don't yet have a paradigm for Big Data to fit into. Because of its lack of scientific structure, Big Data does not impose a set of questions or a way of thinking on its use. This is both its strength and its weakness. Those who do Big Data research are not constrained by many of the thought processes and bound by many of the myths that we've grown up with. On the other hand, a lack of paradigm tends to send us off into areas of investigation that often prove fruitless (Needel, 2008).

Here's what happens when we lack a paradigm that involves basic scientific principals – you get into the 3V discussion rather than the "how to use this tool" discussion. We're very worried about how we're going to manage all this data. The advantages that the "interlopers" we discussed earlier have are that they have a clue as to how to store and manipulate massive amounts of data. To this, I say, "Not such a big deal." The challenge is not to master the physical aspects of Big Data and its analytics, the challenge is to figure out how to use it. Big Data is only as smart as the researcher who is querying the database or creating the analytical models. Therefore, the sky is the limit with what you can achieve and there are no brakes on your skid into ignominy.

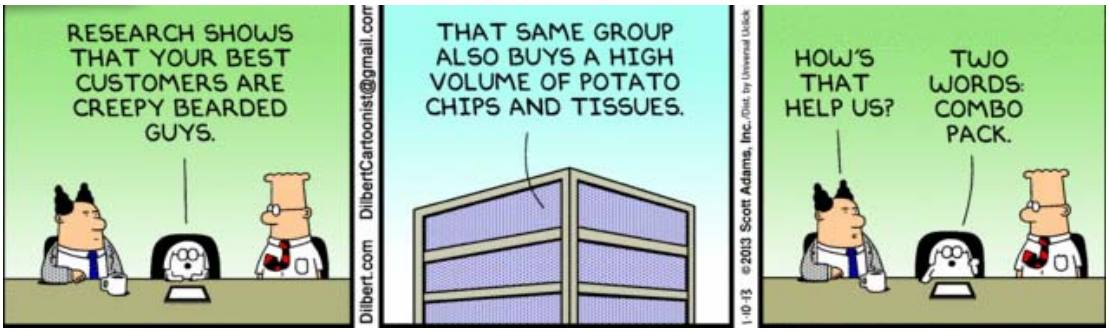
Contrary to some popular conceptualizations, Big Data is not magic. It does not recognize patterns on its own, nor does it create statistical models on the fly. The approach of putting thousands of variables into the mix and creating a correlation coefficient is just about the dumbest thing we've ever come up with. What is the dumbest thing – well, that's another paper in itself. Analyzing Big Data requires extensive data preparation (you didn't really think the data came pre-tagged for merging different sources together, did you?). Models need to be chosen and explicated; whether they are linear, Bayesian, structural equation, or take some other form, it takes a conscious choice. Results need to be checked for rationality; are there collinearity biases, Heywood cases, etc. that make the results suspect. At the conclusion of this paper, I'll give you a paradigm – free of charge.

COMPETENCY

Here's the assumption nobody ever talks about when it comes to Big Data – are we smart enough to know what to do with the outcome of a Big Data analysis. Now, the facetious answer is that marketers may not be, but we researchers certainly are. The real answer is that the marketing actions to take based on outcomes from Big Data analyses are not necessarily intuitive. Just because you are able to create a model for a given behavior or just because you find a high and interesting correlation between two pieces of data does not mean anyone knows what to do with it. This is why the apocalyptic end of marketing research as we know it will not happen. Those who decry the intrusion of outsiders into our province forget about this crucial feature – you need to know what the analysis means in order for Big Data to make marketing sense. Companies like Equifax and IBM may be great at crunching large amounts of data, but when it comes to consumer package goods, they have no clue as to what to do with the answer they get. If you don't understand that or believe that, just attend some of their webinars where they make it painfully obvious that marketing FMCG products or grocery retailing is not their strong suit (cf. Retailwire 2012, 2013 and IBM 2012). These examples are filled with 1980s technology and a wonderful sense of naiveté about what really goes on in the real world.

It's not clear that marketers actually know what they are doing with this information either, as this cartoon from Dilbert suggests.

FIGURE 1



THE NEW PARADIGM

I told you in the title that Big Data is actually a small idea and you shouldn't worry so much. Here's why you shouldn't worry – because only one of two things will happen:

1. We will come to realize that the intractability of the problems I've discussed is true, or mostly true, and we'll walk away from this concept. Of course, this will come after an incredible amount of money will have been spent trying to build data bases and thousands of man-hours will have been wasted trying to create models for this data.
2. Someone will actually figure out how to do Big Data, and it won't be you. It won't be me either, if that helps you feel any better. Even then, they will have spent millions and millions of dollars and countless man hours to figure it out and they will never recoup that investment.

Either way, you have already won, so you can relax. What the Big Data discussion has taught us is that asking questions might be a good thing – we might find out useful things. Big Data wants you to let the computer do the asking, but that's not because the computer is smarter; it's because the proponents think they can build a business out of it. We in marketing research are supposed to be the experts in asking and answering marketing questions. All we have to do is do that job and do it well. That job starts with a well-grounded but basic knowledge of your own business. When I start a project on a business new to me, I usually ask two questions. I want to know why shoppers buy my client's product and I want to know why category buyers don't buy my client's products. You all might be amazed at how few clients actually know the answer

to that question. Those two answers actually provide the platform for everything the brand does and should raise the questions that help guide sales, marketing, new product development, shopper marketing, and so forth.

Big Data should have made us comfortable with asking questions again, but the shotgun approach to correlating everything and seeing what pops up has yet to show itself as a useful paradigm. Do not be seduced by the 3 "V"s – you don't need the volume of data, much of the variety is irrelevant, and you can't react at the velocity of data so don't bother trying. Instead, focus on knowing more about your brand and what makes it tick. When you don't know, get help, and get quality help. Last year we listened to Angry MR Client moan about how bad her suppliers were. To this (and to others who expressed the same complaints) I had a simple answer – hire better suppliers. We have lots of bright people doing lots of bright research in our industry in almost any area of inquiry. Hire them to figure out the answers to your problems when you need outside help.

Finally, be skeptical with respect to marketing research. We have seen a proliferation of new marketing platforms and new research technologies and we have, in my opinion, been overly eager to pounce on these as a solution to some problem or the opening of a new opportunity. We don't know whether social media means anything yet. We don't know if you can do a decent job of asking questions via a mobile device in geo-appropriate real time. We don't know if two questions on a Google survey is sufficient, and if so, when or for what problems. We don't know whether facial recognition or biometric measurement is actually of any value. All show promise; none show that they can make us better marketers or researchers. So relax, and keep on asking questions, hopefully better and better ones.

REFERENCES

Court, David, Elzinga, Dave, Mulder, Susan, and Velvik, Ole Jergen (2009). The Consumer Decision Journey. McKinsey Quarterly, June 2009.

Duhigg, Charles (2012). How Companies Learn Your Secrets. New York Times, 16 February 2012.

IBM (2012). Science Meets Assortment Optimization. IBM Corporation, 2012.

Laney, Douglas (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Meta Group. Stamford, Connecticut.

Needel, Stephen P. (2008). Where Has All the Science Gone? ESOMAR Congress. Montreal, 2008.

Needel, Stephen P. (2012). Is Eye Tracking Making Us Blind and Other Research Maladies. Marketing Research in a Mobile World North American Conference. Cincinnati, 2012.

Needel, Stephen P. (2013). An Ecclestiastical Perspective on Big Data. Presented at the MRIA 2013 National Conference, Niagara Falls, Canada.

Murphy, Lenny. (2012). ESOMAR Congress Was a Pivotal Event for MR. Greenbook Blog, 18 September 2012.

Passingham, Judith, McCaughan, David, Murphy, Lenny, Koornstra, Sjeord, Cooke, Michael, and Baker, Reg. (2012). White Hat vs. Black Hat: Ensuring the Future Growth of Market Research. ESOMAR Congress, Atlanta. Also available at http://vimeo.com/53492501

Retailwire, (2013). "Think You Know Assortment? Chances Are You Don't." Webinar presented by Retailwire.com, 1 January 2013.

Retailwire (2012). "Big Data's Big Impact On Retail." Webinar presented by Retailwire.com, 30 October 2012.

http://www.greenbookblog.org/2013/01/11/everyone-is-a-research-agency-under-the-radar-presents-the-snap-together-solution/

THE AUTHOR

Stephen Needel is the Managing Partne, Advanced Simulations, LLC., United States.