Preface: Big Data Is Not About The Data![1]

Gary King[2]
Institute for Quantitative Social Science
Harvard University

A few years ago, explaining what you did for a living to Dad, Aunt Rose, or your friend from high school was pretty complicated. Answering that you develop statistical estimators, work on numerical optimization, or, even better, are working on a great new Markov Chain Monte Carlo implementation of a Bayesian model with heteroskedastic errors for automated text analysis is pretty much the definition of conversation stopper.

Then the media noticed the revolution we're all apart of, and they glued a label to it. Now "Big Data" is what you and I do. As trivial as this change sounds, we should be grateful for it, as the name seems to resonate with the public and so it helps convey the importance of our field to others better than we had managed to do ourselves. Yet, now that we have everyone's attention, we need to start clarifying for others -- and ourselves -- what the revolution means. This is much of what this book is about.

Throughout, we need to remember that for the most part, Big Data is not about the data. Data is easily obtainable and cheap, and more so every day. The analytics that turn piles of numbers into actionable insights is difficult, and more sophisticated every day. The advances in making data cheap have been extremely valuable but mostly automatic results of other events in society; the advances in the statistical algorithms to process the data have been spectacular, and hard fought. Keeping the two straight is crucial for understanding the Big Data revolution and for continuing the progress we can make as a result of it.

Let's start with the data, the big data, and nothing but the data. For one, the massive increase in data production we see all across the economy is mostly a free byproduct of other developments underway for other purposes. If the HR team in your company or university installed new software this year, they will likely discover a little spigot which spews out data that, it will turn out, can be used for other purposes. Same for your payroll, IT infrastructure, heating and cooling, transportation, logistics, and most other systems. Even if you put no effort into increasing the data your institution produces, you will likely have a lot more a year from now

[1] In R. Michael Alvarez, ed., In press, *Computational Social Science: Discovery and Prediction*. Cambridge University Press.
[2] Albert J Weatherhead III University Professor, and Director, Institute for Quantitative Social Science, Harvard University (IQSS, 1737 Cambridge Street, Cambridge, MA 02138); GaryKing.org; King@Harvard.edu; 617-500-7570.

than you do today. In many areas where you need to purchase data, you'll find its prices dropping as it becomes commoditized and ever more automatically produced. And if you add to these trends a bit of effort, or a bit of money, you will see vast increases in the billions of bits of data spewing forth.

Although the increase in the quantity and diversity of data is breathtaking, data alone does not a Big Data revolution make. The progress in analytics making data actionable over the last few decades is also essential.

So Big Data is not mostly about the data. But it is also not about the 'big' since the vast majority of data analyses involve relatively small data sets, or small subsamples of larger data sets. And even many truly immense data sets do not require large scale data analyses: if you wanted to know the average age in the US population, and you had a census of 300 million ages, a random sample of a few thousand would yield accurate answers with far less effort.

Of course, the goal for most purposes is rarely the data set itself. Creating larger and larger quantities of data that is not used can even be downright harmful -- more expensive, more time consuming, and more distracting -- without any concomitant increase in insight about the problem at hand. Take the following data sources, each with massive increases in data pouring in, even more massive increases in the analytics challenge posed, and little progress possible without new developments in analytics.

Consider social media. The world now produces a billion publicly available social media posts every other day, the largest increase in the expressive capacity of humanity in the history of the world. Any one person can now write a post which has at least the potential to be read by billions of others. But yet no one person, without assistance from methods of automated text analysis, has the ability to understand what billions of others are saying. That is, when we think of social media data as data, it is nearly useless without some type of analytic capacity.

Or consider research into exercise. Until recently, the best data collection method was to ask survey questions, such as "did you exercise yesterday?". Suppose your survey respondent has an answer, is willing to tell you, intends to give a genuine response, and that response happens to be accurate. (Not likely, but at least possible.) Today, instead, we can collect nearly continuous time measurements on hundreds of thousands of people carrying cell phones with accelerometers and GPS or Fitbit-style wearables. In principle, the new data is tremendously more informative, but in practice what do you do with hundreds of millions of such unusual measurements? How do you use these data to distinguish an all out sprint on a stationary bike from an all in sit by a couch potato? How do you map accelerometer readings into heart beat or fitness measures? What is the right way to process huge numbers of traces on a map from GPS monitoring, all at

different speeds and in different physical locations and conditions? Without analytics, and likely innovative analytics tuned to the task at hand, we're stuck paying to store a very nice pile of numbers without any insights in return.

Or consider measuring friendship networks. At one time, researchers would ask a small random sample of survey respondents to list for us their best friends, perhaps asking for their first names to reduce measurement error. Now, with appropriate permissions, we now have the ability to collect from many more people a continuously updated list of phone calls, emails, text messages, social media connections, address books, or bluetooth connections. But how do you combine, match, disambiguate, remove duplicates, and extract insights from these large and diverse sources of information?

Or consider measurements of economic development or public health in developing countries. Much academic work still assumes the veracity of officially reported governmental statistics, which is of dubious value in large parts of the world and just plain made up in others. Today, we can skip governments and mine satellite images of human-generated light at night, road networks, and other physical infrastructure. Internet penetration and use provides other sources of information. But how are you supposed to squeeze satellite images into a standard regression analysis expecting a rectangular data set? These data too require innovative analytics, which fortunately is improving fast.

Moore's Law is the historically accurate prediction that the speed and power of computers will double every 18 months, and the result of this repeated doubling has benefited most parts of society. However, compared to advances in analytics, Moore's law is awfully slow. I've lost track of how many times a graduate student working with me has sped up an algorithm by a factor of 100 or 1000, by working on a problem for more like 18 hours rather than 18 months.

Not long ago, a colleague came to the institute I direct and asked for help from our computing staff. The statistical program he was running every month started crashing because the volume of his data had increased and it overwhelmed his computer's capacity. He asked them to spec out what a new computer would cost so he could figure out how big a grant he would need to seek. The answer: $2 million. A graduate student noticed the answer, and she and I worked for an afternoon to improve his algorithm -- which now runs on his off-the-shelf laptop in about 20 minutes. This is the magic of modern data analytics. As terrific as the developments summarized by Moore's Law, they don't come close to modern data science.

Whether you call what we all do by one of the longstanding names -- such as statisticians, political methodologists, econometricians, sociological methodologists, machine learning specialists, cliometricians, etc. -- or some of the newer emerging names -- such as big data

analysts, data scientists, or computational social scientists -- the current and likely future impact of these areas on the world is undeniable. From an institutional perspective, we see considerable power coming from the unprecedented and increasing connections and collaborations, and even to some extent tentative unification, across all these fields.

Throughout the history of each of these areas, the biggest impact has increasingly emerged from the tripartite combination of innovative statistical methods, novel computer science, and original theories in a field of substantive application. I hope this book will clarify for us all the distinctive perspective and high impact that researchers in these areas have had. The benefits for the rest of academia, commerce, industry, government, and many other areas depend on it.