Big Data and its Epistemology

Martin Frické, SIRLS, The University of Arizona, 1515 E. First St., Tucson, AZ85719 mfricke@u.arizona.edu

Abstract

The paper considers whether Big Data, in the form of data-driven science, will enable the discovery, or appraisal, of universal scientific theories, instrumentalist tools, or inductive inferences. It points out, initially, that such aspirations are similar to the now discredited inductivist approach to science. On the positive side, Big Data may permit larger sample sizes, cheaper and more extensive testing of theories, and the continuous assessment of theories. On the negative, data-driven science encourages passive data collection, as opposed to experimentation and testing, and hornswoggling ('unsound statistical fiddling'). The role of theory and data in inductive algorithms, statistical modeling, and scientific discoveries is analyzed, and it is argued that theory is needed at every turn. Data-driven science is a chimera.

Introduction

There are quantities of recorded data as never before, largely thanks to computers, networks, sensors, and how cheap and fast the processing and storage now is. In 2013, according to the EMC/IDC Digital Universe Studies, there is about 4 zettabytes of stored information, and this amount is doubling every two years (Gantz & Reinsel, 2011). A zettabyte is (10 power 20) bytes and the total amounts to something like several hundred CD-ROMs for every man, woman, and child on Earth. The sources of this data are many and varied. There are data intensive sciences such as astronomy, genomics, experimental particle physics, and ocean sciences. For example, in astronomy there are indefinitely many celestial bodies; they can be probed across a wide range of the electromagnetic spectrum; and there are instruments to collect and record this abundance of data. (This can be seen in the Sloan Digital Sky Survey (Sloan Digital Sky Survey, 2013).) Then there are sensor networks that can produce a sometimes continuous data about geographical locations, the ocean, the weather, or atmospheric conditions (for example, the Oceans Observatories Initiative (Oceans Observatories Initiative, 2013)). Then there is much data about us created by our behavior in conjunction with computers, smart phones, electronic or digital transactions, location services and the like (do not be alarmed, but there are rumors that the National Security Agency NSA Data Center in Utah has a memory of, or can process, 5 zettabytes (Skeptics Stack Exchange, 2013)). In sum, there is Big Data.

What does this mean for epistemology? Some commentators are giddy. Hans Rosling tells us

> The data deluge… is leading us to an ever greater understanding of life on Earth and the Universe beyond….
> …[it may] transform the process of scientific discovery
> The more data there is the more discoveries can be made. (Rosling, 2010)

Luciano Floridi suggest that scientists be alert to the opportunity of detecting the ‘patterns’ (and ‘absence of patterns’) in the data as a way to advance knowledge (Floridi, 2012). He writes

> … the pressure … on … genetics or medicine, experimental physics or neuroscience, is to be able to spot where the new patterns with real added value lie in their immense databases and how they can best be exploited for … the advancement of knowledge. (Floridi, 2012)

Others point to the ‘fourth paradigm’ for science—new algorithmic, computational and analytical tools to produce gold for us from this data resource (Bell, Hey, & Szalay, 2009; Hey, Tansley, & Tolle, 2009) In prospect, then, some see the advent of data-driven science.

Let us be careful here. Of course, computers are vital today and moving into the future of science. They can create, run, and test thousands of hypotheses, models, and simulations in the blink of an eye. Shotgun sequencing in genomics, to provide an example, definitely needs raw computing power. But most of this would be what might be called computer enhanced hypothetico-deductivism—it still would be guess-and-test, but guess-and-test on computer steroids. However, ‘data-driven’ science is not intended to be this, instead

> Data-intensive science consists of three basic activities: capture, curation, and analysis (Hey et al., 2009).

Somewhat similarly, Michael Schmidt and Hod Lipson’s well known paper is entitled ‘Distilling free-form natural laws from experimental data’ (Schmidt & Lipson, 2009) (and see also (Hillar & Sommer, 2012) and (Waltz & Buchanan, 2009)). It would not be uncommon at this point to cite Chris Anderson’s ‘The end of theory: the data deluge makes the scientific method obsolete’ (Anderson, 2008). (It should be mentioned, though, that apparently Anderson never believed or advocated the theses of his own paper but wrote it to provoke response (see (Norvig, 2008)), he was merely being a journalist shouting ‘Fire’ in an academic theatre.) Even so, there are many commentators who think that it is data that is now going to drive much of science.

What is a preliminary view?

The very fact that we have more data means that we know more—at the very least we know that extra data. And we likely know more besides. Apparently if a young man, of a certain kind, comes by a lost or stolen credit card, he fills the tank of his own car with gas, fills his friends' cars with gas, then buys himself a pair of sports sneakers. Real time data acquisition and monitoring of credit card transactions would identify and reveal the fraud in an instant. Some valuable learning has taken place. But that, in a manner of speaking, is uninteresting—it is hardly giving us a greater understanding of life on Earth and the Universe beyond. It is hardly even a low level nomic or statistical generalization about certain young men and lost or stolen credit cards—the behavior may be entirely different tomorrow or next week or next year. What about some of the more scientific examples? As its title indicates, the Sloan Digital Sky Survey is a survey—that is, it is a list of celestial bodies, some newly discovered, and their properties (Sloan Digital Sky Survey, 2013). The Ocean Observatories Initiative is looking to make observations (Oceans Observatories Initiative, 2013). Surveys are fine, and so too are observations, but what they are doing methodologically is gathering more data. They are not in themselves offering any explanations or theories or solving scientific problems, or aiming to do anything of that nature. The same point is true of several other examples of Big Data ventures, such as a good portion of biodiversity science or genomics science. They aim to provide lists, catalogs, and classifications.

What would be interesting is whether we could make scientific discoveries; whether we would or could or might learn something new that went beyond the new data and its surface deductive consequences: that new theories, explanations and predictions would become available. The standard view in the philosophy of science, since Sir Karl Popper and earlier commentators (possibly even back to the Scottish philosopher David Hume), is that while there is a 'logic of scientific justification' (i.e. methods of testing scientific theories once they are available and presented for consideration) there is no 'logic of scientific discovery' (i.e. no routine and semi-mechanical way of producing valuable scientific theories in the first instance)

> The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it.(Popper, 1959)

Is Big Data going to produce a logic of scientific discovery?

Equally interesting would be to find out that some of our most cherished theories and explanations were mistaken, or some of our favored prediction techniques are unreliable, i.e. to find out refutations. In practice, as we know from Pierre Duhem, Imre Lakatos, and others, there is more to practical refutation than the plain production of a counter-example (Duhem, 1914; Lakatos, 1970, 1974a, 1974b). Nevertheless, we do have an epistemological interest in actual and potential counter-examples.

The present paper now has a problem, or question, to consider: might Big Data enable the discovery, appraisal or validation, of universal scientific theories, instrumentalist tools, or inductive inferences. And, secondarily, might Big Data facilitate refutation?

Data

What is data? We can characterize data as follows. Data is anything recordable in a relational database in a semantically and pragmatically sound way. The semantics require that the recordings be understood as true or false statements. The pragmatics suggest that we favor recording what seem to be concrete facts, i.e. singular and relatively weak statements, and that interpreted recordings be true statements (Frické, 2009). Of course, our knowledge of data, or what constitutes data, is fallible. Data is therefore conjectural.

Data is also theory-laden in the following sense. Data is typically gathered using instruments, measuring instruments, and sensors. Here the notion of instrument is understood to be a very wide one. And those instruments are constructed or adopted in the light of what we know, or theories we have, about what we are measuring and what the instruments are telling us about some kind of reality beyond themselves and their apparent surface indications. (Bogen, 2013)

Big Data is both fallible and tainted by theory.

Inductivism

An immediate concern about the present enchantment with data-driven Big Data is just that it might be inductivism, the hoary punching bag from the philosophy of science. The philosopher Karl Popper has told us that it is utterly pointless to collect observations and to hope that similarities in the observations will somehow appear and allow good theories to emerge.

> … the belief that we can start with pure observations alone, without anything in the nature of a theory, is absurd; as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society to be used as inductive evidence. This story should show us that though beetles may profitably be collected, observations may not. (Popper, 1963)

And the novelist Laurence Sterne teased inductivists with his character Tristram Shandy who took a year to write up each day in his diary (of course, a modern Tristram Shandy, using Big Data techniques, might be able to make some kind of attempt at real-time recording)(Sterne, 1759).

That inductivism is a mistaken philosophy of science is not controversial—it is received wisdom.

Passive Observation versus Active Experimentation

Science made a great leap forward with the advent of the experimental method (in the modern era, roughly from Robert Hooke onwards, about 1660). What is so special about this? We are looking for lawlike or nomological connections, to connect causes with effects. But what we observe can be misleading due to confounds. Confounds are other conditions or variables which correlate either with the causes, or with the effects, to mask what is really happening at a causal level. The experimental method is in part a technique to deliberately manipulate Nature so that known possible confounds are controlled for. With the experimental method we ask Nature questions which are deliberately framed. With a typical scientific problem and a tentative hypothesis offered to solve it, there are known confounds and maybe also unknown confounds. The known confounds are controlled for. There is a view on experimental technique, the standard view from Sir Ronald Fisher, that the problem of unknown confounds can be addressed by randomization, for example using Randomized Controlled Trials (RCTs). Actually, it seems that randomization via Randomized Controlled Trials (RCTs), is not necessary (Urbach, 1985; Worrall, 2007); however attention certainly does need to be paid to potential confounds, both known and unknown. Doing this typically requires active intervention to produce certain kinds of data. Testing theories again typically requires experimentation. Some theories can be tested against plain observation; and sometimes Nature performs ‘natural experiments’ to produce suitable results without active intervention; but mainly testing invites experimentation.

To use an illustration from David Freedman of the growth of scientific knowledge (Freedman, 1999 (revised 2002), 2008): More than a few smokers have or get lung cancer; more than a few smokers have or get cirrhosis of the liver. The reasons are:- smoking causes lung cancer, but smoking does not cause cirrhosis of the liver, rather drinking causes cirrhosis of the liver (and many smokers just happen also to drink). How we know all this is a complex story. Much of it is careful and specific observation in the light of other scientific and medical theories which we hold in high regard; some of it is experimentation, also conducted against a backdrop of accepted medicine and epidemiology. One point is for sure, though, it would be very hard, if not practically impossible, to read knowledge like this directly from 'patterns' in observable data. To give slightly more detail to part of it: Richard Doll, and Austin Bradford Hill, had the idea or theory or guess that smoking caused lung cancer. They then sought and obtained extensive data about which doctors smoked, and which did not, and what happened to them (and which doctors gave up smoking and what happened to them). The data would not have been sought and aggregated without the theory. The data would have meant nothing without the theory. Of course, this is just one example. But once again it was Karl Popper who made the relevant point years ago: everything is similar to everything else in certain respects, and everything is different to everything else in certain respects, so the mere looking for similarities does not take you very far (Popper, 1963).

Big Data is not itself incompatible with experimentation. But it is the friend of passive observation—it encourages passive observation. Conducting surveys and making observations amount to passive observation.

The Curve-fitting Problem, Machine Learning and Statistical Modeling

It is convenient to remind ourselves of some features of inductive inference and statistics using the classical 'curve-fitting' problem (Forster, 1995; Freedman, 1999 (revised 2002)).

Suppose there is an x-y plane, a two-dimensional graphical coordinate system, with some data points in it. These points are, or represent, the known data. And it is assumed, as a truth of the matter about the world, that there is some nomic or causal connection between the points by virtue of their x, and y values. The task for science is to draw a curve through those points which will successfully anticipate the location of future or unknown data points. This requires an inductive inference. The data is singular, the curve is universal, and the inference from one to the other is deductively invalid. Indefinitely many curves might be drawn that include the known data, but a requirement is that just one curve be chosen.

At the start of the analysis, it is not known what is cause and what effect (is the x value producing the y value, or the other way around? does a predisposition to own stocks make you rich, or does being rich give you a predisposition to own stocks?). But it is assumed that there is a law-like connection. This assumption is often reasonable to make. In science, physical science, say, there are many laws; we know of many. In social science the assumption might not be so reasonable; laws, even robust statistical regularities, have proved elusive (Meehl, 1978). In free form data analysis, the assumption would itself require critical scrutiny.

Then if there is another variable, not in the original analysis or model thus far, for example, z, that affects x, then the whole analysis can be totally and completely wrong (this is the so-called 'omitted variable' bias in regression).

That there are indefinitely many curves consistent with the data means that the actual choice of curve must be based on considerations other than those offered by the data itself. For example, it might be thought that the curve should be a straight line, or be quadratic, or be 'simple', or… The functional form of the connection between the data is not known and is not determined by the data that is under consideration. Those extra considerations are *inductive bias*. Then, with the inductive bias that is adopted, the data can be suitable or unsuitable in various ways. For example, suppose that the inductive bias is that the curve be a straight line; then one data point will not determine a single line; the set of hypotheses (straight line curves) would *overfit* the data; on the other hand, were there to be three data points, not in a straight line, the set of hypotheses would *underfit* the data (because the inductive bias was not rich enough to generate even a single hypothesis that would fit the data).

There might also be the need for an *error term*. It would be unusual to assume that any numerical data to hand is exactly correct and accurate, instead the standard practice would be to imagine it composed of a true value summed with an error term. This error term feeds into confidence about results of the analysis as a whole. How large the error is matters, and so does the distribution of the errors (for example, whether they have a normal distribution). Once again, information about size and distribution of errors does not itself come from the data to hand.

The curve-fitting as described is the counterpart of statistical regression, but it can easily be adapted to statistical classification. For this, the y values could be categorical or integral. The task is to say whether a particular x value does or does not have property Y (and that might be indicated by a y-value of 0 or a 1) or which property any individual x has among the properties U,W,Y ...(say values 0,1, 2, 3 ...). It may be a slight stretch to call this ‘curve fitting’, because the curves may well be a little unusual (for example, being discontinuous). However, the process still is that of finding a function of unknown form, it is inductive, it relies on an inductive bias to prune the indefinitely many possible functions, it requires avoiding omitted variable bias, and it may require care with errors and their distribution.

So, the form of the equation sought is

$$y=F(x) + \varepsilon$$

or, in the multivariate case where y might depend on several variables x, z, w, ...

$$y=F(x,z,w,\ldots) + \varepsilon$$

Given all this, it is near miraculous that curve fitting ever works either for discovery or for justification. But, as a rational reconstruction of some scientific discovery or some scientific justification, it does work, at least some of the time, and the successes of statistics in physical science are a testimony to that. It cannot, and should not, work all the time simply because for arbitrary x and y, y does not have to be connected with x. In fact, most of the time y will not be connected with x (and a successful curve fitting analysis should either fail to reveal a connection or directly suggest that there is not one).

The central point is: the data itself does not speak. What is required is a huge amount of background knowledge, or assumptions, or prior research of one kind or another.

There is an acid test or touchstone with curve fitting. It is that of further testing. If, somehow or other, curve fitting produces a candidate curve, that candidate curve should be tried on new data. Then the new data might fit the curve or not. The curve should be tested; attempts should be made to 'falsify' it; then passing the tests (i.e. failing to falsify the curve) often amounts to a certain kind of corroboration (or replication).

Inductive algorithms are a central plank of the Big Data venture. In certain circumstances, 'machines', that is, computers and algorithms, seemingly can learn from data. There is a research area addressing this at the intersection of machine learning, artificial intelligence and data processing (see, for example, (Dietterich, 2003)).

At the core there are three problems here: supervised regression, supervised classification, and unsupervised analysis. The adjective 'supervised' in this context means that the supplied pre-analysis data is conceived of as being governed by a function from inputs to outputs, then the task is to find that function or to find some way of predicting what outputs will be produced by new, hitherto unseen, inputs of the same kind as those initially observed.

‘Unsupervised’ does not make this assumption, and so the input data is just raw data that is to be made sense of. Supervision means that there are rewards, the right answers are known in at least some cases—they are known with the supplied data. Unsupervised learning takes place without rewards.

In essence, supervised regression is the curve-fitting problem, or possibly statistical modeling, attacked by algorithm. In supervised learning the observed data is the *training set*, and the training set together with the *inductive bias, lack of omitted variable bias*, *a theory of errors*, etc., are sufficient for inductive algorithms and machine learning.

Unsupervised learning is much more open than supervised learning. Of course, there is the data, and that data has statistical properties (often means, standard deviations, clustering, distributions etc.). And the data can be guessed to be a sample from a wider set of data which itself has a distribution. These are, in a manner of speaking, standard statistical inferences and techniques, but in ordinary statistics usually more is known about the sample (for example, whether it is random) and more can be known about the underlying population (for example, its likely distribution). In this paper, we will focus on the supervised form of machine learning.

There certainly are some successes with machine learning, such as text parsing, image and hand-writing recognition, spam filters, and credit card fraud detection. And the apparent inductive nature of these has certainly given rise to debate (an example is provided in (Allen, 2001a, 2001b; Gillies, 2001; Kell & Oliver, 2004; Kelley & Scott, 2001)).

However, whether there are practically or theoretically successful inductive algorithms need not occupy us. Our interest is with some different questions. Would these inductive algorithms, or the machine learning approach in general, work using only data? Are they driven by data? Would they benefit from more data? Would they be any the better with Big Data? Is something special going to happen now that we have Big Data?

Supervised algorithms depend mainly on a training set and an inductive bias. Once the inductive bias is set, the algorithm designers know exactly how large they would like the training set to be. There is a tradeoff between possible underfit and possible overfit. But the theory and the design dictate the data requirements. For example, in the curve-fitting illustration above, two different data points will fix the straight line. Being provided with a thousand and two data points would not improve this. Of course, there is fallibility, the hypothesized relationship might not be a straight line and a third observation, or a one thousand and second observation, might be the one to kill the straight-line hypothesis. But there are diminishing returns to repeat observations or tests of the same kind (Howson & Urbach, 2006). The designers know exactly what they want in the training set.

In slightly more elaborate examples, which are not different in principle, the training set needs to be representative of the underlying data in the sense of being a suitable sample of the kinds or types that might be there. In text processing, for example, there may be the desire for the training set to be larger than the one that is available, say from the extant documents that we actually have in our hands, but the desire here is usually not for more of the same stuff but more of different stuff should there be different stuff. The point remains though: with inductive algorithms there is insight as which data is required and the heuristic is not: bigger is better, without limit.

There is a slightly different question concerning more data. In the above curve-fitting example, the problem was very constrained. There was a variable, x, and a variable, y, whatever they might be or designate, and then the inductive algorithm was going to look for a relationship between them. But more data might provide values for other variables, say, z, u, v, a, b, c… etc. and then the inductive algorithms could look for relationships between any of those hitherto unscrutinized variables and some chosen output variable, perhaps y. So, for example, with email spam, an inductive algorithm might tell of a relation between the presence or absence of the word 'lottery' in the subject line and email being or not being spam. Now, the designers of the spam filter may have no interest in further data about 'lottery' in the subject line. But Big Data could provide them with different data, for example, the presence or absence of the words 'Nigeria' and 'inherited' in the body of the email (SCAMwatch, 2013); and the inductive algorithms could work on possible relationships between the new variable or variables and an email being spam.

Fine, but two points. The algorithms would not want to look at all of the other possible variables and relationships, for example, a possible relationship between Yorkshire's cricket scores in 1948 and spam in today's email, for that approach would be subject to the strictures of Popper and others against inductivism. A simple consideration from rapid growth shows that considering all possible relationships is out. Suppose the algorithms considered a thousand variables, which is surely not a large number in Big Data terms; there are (2 exp 1000) subsets of those variables; this number of subsets, which is a proxy for number of possible functional input variable combinations, is considerably larger than the number of particles in the Universe (which has been estimated as approximately 10 exp 80). So, anything like a global search through possible correlations will never be a practical computing possibility. And, second, if numerical values for variables are conceived of as consisting of a real value plus a systematic or random error, then, as we will see later with multiple comparisons, considering relationships between pairs of variables from a large possible variable set is pretty well guaranteed to find some apparent connections, 'false positives', even if there are no real connections. So, again, the designers of the inductive algorithms need to have prior ideas on possible connections, and once they have those and a chosen inductive bias, they know exactly what data they want (and they do not want to be drenched by a fire-hose of data).

Connectionism, neural nets, random forests etc. in machine learning, are at heart a black-box instrumentalist version of inductive algorithms (usually without the curve fitting visualization support) (see, for example, (Breiman & Cutler, 2013)). There are connected nodes and links, some means of permeating inputs to outputs, and then 'weights' or parameters that could be adjusted to get the specified inputs to produce the desired or observed output. The weights might be Bayesian conditional probabilities relating firings of one node to firings of others, or parameters in some kind of generalized Markov chain process. With supervised learning, once again there is a training set of input data, each with known corresponding output data, and the parameters are tuned suitably. Then the black box is released on new input data and hopefully acceptable output data will be predicted. And often it is. So, for example, with 'Random Forests' there are a number of yes/no decision trees working on different random subsets of components of the input data, and these are tuned to produce collectively the requisite output of the training data (Breiman & Cutler, 2013). There is no real explanation of why or how the input connects with the output; however, predictions are made, and often these predictions are pretty good. In an instrumentalist sense, the right answers are forthcoming.

With supervised black-box models, the data requirements for the training data set is much the same as it is in the plain inductive algorithm case. Representative training data is required, but beyond that there is no special interest in more data.

Machine learning is neither theory free, nor in need of fecund data.

One part of statistics is to make predictions, or to say how output variables are dependent on input variables. Each of these collections of input and output variables can be coalesced together to form a vector, so then the prediction task is produce a function that relates the output vector to the input vector. What happens in between is a 'black box'. One typical approach at this point is to make a model, which is some statistical conjecture as to what is happening within the black box (and thus remove the blackness from the box). The model starts life as a story as what is connected with what and how the variables might be related. And then there is the actual statistics, which will usually be a regression analysis (i.e. a curve fitting problem) of some kind (what exactly it will likely be depends on the number of independent variables, i.e. the input vector components, the nature of the values of the output components (for example, whether they are continuous or categorical), and so forth…) The upshot hopefully is a statistical connection between input and output and an 'explanation' or rationale as to what is happening in the box.

Intellectually the process as described is not a whole lot different from machine learning, apart from the fact that it is not algorithms that are doing the heavy lifting; rather it is intelligent human statisticians (who can use algorithms if they wish, but can also use any other smarts and insights that they might have). (Of course, historically, it is the statistical modeling that comes first, then the machine learning, which is basically an algorithmic implementation of some of the statistical practices.)

And the role of data is much the same in statistical modeling and machine learning. Statistical modelers want data. But they do not want unlimited data. Sometimes, maybe even often, larger sample sizes are better, but usually with limits. They have theories or tentative models, and from these they know exactly what data is required. And at least some of the time, they know that parts of their model, or some of their assumptions, are false. For example, they might want to assume that some error factor has a normal distribution when they know that it does not. More data in these cases is no help at all. More data is not going to convert a distribution that is not normal into a distribution that is normal.

There are also pure black-box statistical modelers, just like black-box machine learners (Leo Breiman might be an example of both (Breiman, 2001)). These want to connect input with output, without any story in the box in between. And exactly the same points about data can be made. The input vectors and the output vectors, or rather the components of the input vectors and the components of the output vectors, cannot be any variables whatsoever. Too much simply is a non-starter. The problem has to be tightly circumscribed by outside theories. Then once it is, the data requirements are known and not usually extensive. There is no need for Big Data.

Independently, and separately from considerations of Big Data, there is an obvious objection to black-box modelers and black-box machine learners and it is that in science not only do we want theories in the black-box to explain the connection between input and output but often there is the desire to penetrate deeper and to explain the theories themselves. For example, Boyle's Law connects pressure and volume (for a fixed amount of gas at a fixed temperature), it can help predict what the pressure of a gas will be if its volume changes. No doubt a simple black box can connect pressure and volume and thus apparently replace Boyle's Law for the purposes of prediction. But in science, Boyle's law is itself a problem to be explained, and, indeed, it has been explained by the Kinetic Theory of Gases. Random Forests, to mention one black-box modeling example, would simply not open the way to this kind of scientific knowledge.

Evidence, Refutation, and Continuous Assessment

Christine Borgman draws attention to Michael Buckland's pithy phrasing that data is

> "alleged evidence," (Borgman, 2012; Buckland, 1991)

Some unpacking might be done here. One aspect of 'alleged' is just that the evidence or data is fallible. We have that noted already. Another is that someone, or something, or some logical relations, want to use the recordings as 'evidence'.

It is commonly, and correctly, observed that there is no such thing as evidence *simpliciter*, rather something is or is not evidence only in relation to a hypothesis (or several hypotheses). Then Bayesianism would tell us that a factual statement is evidence for a hypothesis in so far it is to be expected conditional on the hypothesis and not expected conditional on the negation of the hypothesis (Howson & Urbach, 2006).

A hypothesis or some hypotheses lead the way. A hypothesis is needed first. Then the hypothesis illuminates evidential requirements. Much potential data can play the role of evidence with a chosen hypothesis, in the sense of increasing or decreasing the *posterior* probability of that hypothesis. Practical considerations enter at this point. The 'data' may be already available; it may require passive observation to collect; it may even require deliberate experiment to produce. And getting the data may cost money, it may involve risk, risk of damage or injury, it may involve ethical considerations, and so on. There are practical decision-theoretic cost-benefit analyses to be made regarding the acquisition of evidence. There is a delicate interplay between epistemic and non-epistemic factors.

Big Data has an important role here. Likely collecting data will become progressively cheaper. In at least some cases, the techniques will lower risk. On the other hand, in some cases, ethical considerations will or may become more pronounced (privacy might be an example of this). But, all in all, Big Data should be able to give us better evidence for our theories.

Not all evidence is positive evidence. There are refutations. Initially, when appraising a theory, the theory will be subject to whatever testing seems necessary. Then the theory will be used (to make predictions, constitute explanations, and the like). From an epistemological point of view there are diminishing returns from repeating the same tests (Howson & Urbach, 2006). We might distinguish here test types and test tokens. There are diminishing returns from more tokens of the same type. What typically would be sought are tests of a different type, when the theory is used in new areas or under new conditions. However, going back to the repetition of the same kind of test. There is almost always some epistemological virtue, however small, from more data of the same kind. It is just that in the past the time, effort, and expense of collecting this routine data suggested that it was not worthwhile bothering. However, Big Data techniques change the trade-offs. If the data was essentially close to being free and getting it does not require any time or effort, then, of course, it would be good to have it. What could happen here is a type of ‘continuous assessment’. The instruments, or data collection techniques, could collect data all the time and alert us to any ‘exceptions’ to any of our theories. There is no need under this scenario to record and store all the data—recording the exceptions, and giving notification of them, would probably be enough. And, certainly, some Big Data examples pretty well do this. The Large Hadron Collider (LHC), for instance, detects many, really many, events, but it does not record all of them, nor anything like that, instead it uses ‘triggers’ etc. to record just ‘interesting’ events (and theory informs as to what is ‘interesting’).

Post Hoc Hornswoggling

Statistics, by its very nature, tends to be *post hoc*. The numbers are provided in advance from elsewhere, then analysis is done on them. This gives rise to special difficulties in the case of the statistics of science and hypothesis testing, especially if the numbers are both generating the hypotheses and then testing those same hypotheses. The relevant objection or concern is well known, David Freedman and Michael Babyak phrase it thus:

> Generally, replication and prediction of new results provide a harsher and more useful validation regime than statistical testing of many models on one data set. Fewer assumptions are needed, there is less chance of artifact, more kinds of variation can be explored, and alternative explanations can be ruled out. (Freedman, 1991)

> If you use a sample to construct a model, or to choose a hypothesis to test, you cannot make a rigorous scientific test of the model or the hypothesis using that same sample data.(Babyak, 2004)

There are many known possible errors to make in statistical analysis, and many books and articles on them (for example, (Faraway, 1992; Huff, 1954; Mills, 1993)). The kinds of errors will be familiar to real statisticians and they should be known to working researchers. But they are systematically ignored in current research and publication. This is bad enough, but *post hoc* analysis, in the form of data-driven science, has the potential to make it much worse.

Here are a few common errors: null hypothesis significance testing, stepwise regression, multiple comparisons, subsetting, overfitting, univariate screening, dichotomizing continuous variables, etc. (Bretz & Hsu, 2007; Cohen, 1994; Johansson, 2011; Lykken, 1991; Maxwell & Delaney, 1993; Meehl, 1978).

Exhaustive analysis of these is not a practical possibility here (and it would be beyond the competence of the present author). But it is useful to mention two of them.

The familiar null hypothesis significance testing (NHST), with its 'the null hypothesis was rejected with a p=0.05', has its origins as a mashup of Fisher's views on trying to refute a Null Hypothesis and Neyman-Pearson's decision theoretic suggestions on choosing between a Null Hypothesis and a Hypothesis Under Test. Its use in social and behavioral science research journals has been widespread and continues to be widespread. It has always been controversial, but the arguments about it have only recently erupted with full force, say from about 1995 (Cohen, 1994; Johansson, 2011; Lykken, 1991; Meehl, 1978; Rodgers, 2010; Wagenmakers, 2007). However, this ongoing dispute has yet to be felt in publication practices. How might this work in with Big Data? One way might be this. First of all, everything is correlated with everything else (Lykken, 1991;

Meehl, 1978). This means that any Null Hypothesis (that there is no correlation) is ‘quasi-false’ (it actually is false, but ‘quasi-false’ is more evocative) (Lykken, 1991; Meehl, 1978). (Cohen, 1994; Maxwell & Delaney, 1993). In turn, most samples related to the Hypothesis Under Test will show some correlation. Now there is only the p to worry about. But p gets smaller as sample size gets bigger. And amplifying sample size is exactly what Big Data is capable of. So it would be trivial for Big Data to produce arbitrarily many hypotheses significant at a p=0.05 level.

A second example is multiple comparisons. In one setting, multiple comparisons can be connected with data correlations and false positives involving hypotheses. The result can elevate some of these correlations to conjectured causal or quasi-causal connections. So, for example, if, in the data, A is observed to be somewhat correlated with B; and there is a hypotheses H to the effect that A brings about B; A might indeed do this, or the observed correlation might be the result of chance i.e. a pure accident; accepting the hypothesis H if the connection is a pure accident would be a Type-I error, it would be a ‘false positive’. As a particular example, suppose a researcher gathers much data about mothers and fathers and birth defects in their children (Mills, 1993), and starts with the hypothesis that whether the mother has a job is related to birth defects—that is: there is a ‘test’ group of mothers with jobs and a ‘control’ group of mothers without jobs and the Null Hypothesis is that the number or rate of birth defects of their children are the same. Suppose that the researcher carries out the data analysis and statistics correctly and plans to reject the Correlation Hypothesis if it has a p value greater than 0.05. If the p value is less than this the hypothesis will be accepted, which means that once in a while (in fact 0.05 of the time for a true Null Hypothesis) the Correlation Hypothesis will be accepted completely by chance even though the alleged connection might be spurious. [Care is needed with the phrasing here. A p value of 0.05 means

Probability(Null Rejected given that Null True)=0.05

not

> Probability(Null True given that Null Rejected)=0.05

so the number of false positives cannot be estimated from the number of rejections—a hundred rejections is entirely consistent with no false positives and with one hundred false positives. However, if the Null Hypothesis is true in each of a hundred independent tests there will be (roughly) 5 false positive rejections of it.] To continue, with the first hypothesis there is a small chance of a false positive. Suppose the initial hypothesis is rejected and the researcher decides to go on to a second comparison, perhaps with whether birth defects are related to whether the father has a job. Again there is the 95% 5% or 0.95 0.05 significance level for this second hypothesis considered on its own. And again, suppose the Null Hypothesis is true. But the chances for at least one of the two false hypothesis succeeding by chance is now $(1-(0.95)^2)$ which is about 0.1. So now, getting at least one false positive among the two attempts is more likely. Do this for thirteen to fourteen different comparisons (age, weight, education, smoking, drinking, etc.) and the odds go past 0.5 and change in the researchers favor. Chance is starting to guarantee a 'successful' hypothesis; the researcher is likely to find a quasi-causal hypothesized correlation whether or not there is one. The real problem arises with the publication of the research. If the fifteenth hypothesis shows that, say, smoking is apparently connected with birth defects, the researcher will likely conveniently forget the other fourteen hypotheses and claim a p of 0.05 for the smoking correlation hypothesis. This is just plain wrong as statistics and as sound research. (Peter Austin and Meredith Goldwasser give a good example of something similar, with dichotomizing, in 'showing' that having the Pisces star sign is linked to heart failure (Austin & Goldwasser, 2008)).

Donald Berry writes about multiple comparisons

> Most scientists are oblivious to the problems of multiplicities. Yet they are everywhere. In one or more of its forms, multiplicities are present in every statistical application. They may be out in the open or hidden. And even if they are out in the open, recognizing them is but the first step in a difficult process of inference.

> Problems of multiplicities are the most difficult that we statisticians face. They threaten the validity of every statistical conclusion. (Berry, 2007)

In the same paper, Berry relates an autobiographical anecdote of how, in the first grade, he concluded that redheads were intelligent from having observed two redheaded classmates that were the brightest in the class, he continues

> The two brightest kids in any class are necessarily similar in some other way – perhaps in several other ways. Perhaps they are both girls, both boys, both tall, both short, extreme in height (one may be very tall and the other very short), both of the same nationality or religion or ethnic group, both overweight, both underweight, have similar hair color, have buck teeth, have freckles, speak with a lisp, can run fast, cannot run fast, are handsome, are not handsome, etc. So I was doomed. I was bound to learn something that was wrong! (Berry, 2007)

Big Data offers an open invitation to the problems of multiplicities.

There seem to be three ways out of some of this fiddling. Educate the researchers of the need to tell of unpublished failures as well as of published successes. It is doubtful they would do this under their own initiative, but nevertheless the advice to do so is good. And it can be imposed by policy. For example, there are Clinical Trial Registries, which require the describing of the experiment before it is done. The International Committee of Medical Research Editors write

> … the ICMJE requires, and recommends that all medical journal editors require, registration of clinical trials in a public trials registry at or before the time of first patient enrollment as a condition of consideration for publication….

> The purpose of clinical trial registration is to prevent selective publication and selective reporting of research outcomes [and more purposes are listed] …
> (International Committee of Medical Research Editors, 2013)

Second, consider, in the light of other independent knowledge, whether the published correlation hypothesis might be plausible or sound. For example, given what we know, smoking might indeed cause birth defects (while our present knowledge would not be so accommodating of the suggestion that the father's astrological sign causes birth defects). And third, seek replication: ask the authors, or others, to repeat the data gathering with new data and to show that the conjectured correlation still holds.

Multiple comparisons are a problem for ordinary researchers today. But they will be a bigger problem for attempted 'data-driven science'. In the example, data-driven science will presumably have much data about birth defects and properties of the parents, and others, and so be able to run many multiple comparisons. Maybe the computational research process will be transparent, maybe it will not. Likely, the human researchers will not know in detail what the computer algorithms have done. Data-driven science temperamentally wishes to ignore outside knowledge and wants to let data patterns speak for themselves. This is a mistake. Any analysis would be better with priors, Bayesian prior probabilities, or similar, for any hypotheses. And, again in spirit, it is not set up to gather very specific new data that would amount to attempted falsification leading to replication.

Quite what constitutes replication is an open question (Borgman, 2012; Jasny, Chin, Chong, & Vignieri, 2011). Big Data may have past data about two different settings, e.g. time periods or geographical regions, and, in certain circumstances, one of them could be regarded as a replication of the other. But the best kind of replication is when the new data is genuinely new. Gathering replications could be automatic, or perhaps the human researchers could be prompted of the needs. Christine Borgman has an insightful discussion of replication and Big Data, in particular between such notions as reproducibility, repeatability, validation, and verification (Borgman, 2012).

For epistemology, it is testing not ‘replication’ that is the core demand. As a perhaps apocryphal example, Galileo dropped two balls of different mass off the Leaning Tower of Pisa and refuted the Aristotelian view that heavier bodies fall faster and corroborated his own theory that all freely falling bodies, with negligible air resistance, fall with the same acceleration, that

> …a bird-shot falls as swiftly as a cannon ball (Galilei, 1638)

There are two tests that might be done here. A seventeenth century scientist might seek to test or replicate Galileo’s ‘experiment’, or ‘study’, or ‘clinical trial’, or ‘observations’, or ‘data’, and that might involve protocols with leaning towers, two small masses, and the like. Or a seventeenth century scientist might go directly for an attempted refutation of Galileo’s theory and devise perhaps totally new ways of checking how freely falling bodies fall. Let us imagine further for a moment here. If we grant Galileo’s Law of Freely Falling Bodies its formulation in terms of constant acceleration, a smart enough seventeenth or eighteenth century scientist could reason that the Moon was falling towards the Earth also with the *same* constant acceleration as the dropped cannon ball and thus the centripetal acceleration of the Moon (available from its period and radial distance) should equal the acceleration of the cannon ball (available from the height of the tower and the time of fall). [Newton did this reasoning, and more. In fact, the Moon’s centripetal acceleration is thousands of times smaller than it should be under Galileo’s law (and thus the Moon’s acceleration refutes Galileo’s law).] Of course, the latter testing, testing Galileo’s theory, is better than the former testing, testing Galileo’s data, but to do it you need Galileo’s theory, you do not need his data. This is theory-driven science not data-driven science, theory-driven testing leading, not really to replication, but instead to corroboration. In passing, notice how raised-eyebrow unlikely it is to suppose that there is some ‘pattern’ in the data concerning the height of the Leaning Tower of Pisa, the time of fall of a cannon ball, the period of the Moon, and the radius of the Moon's orbit.

Post hoc analysis does not have to be unsound (and research statistics has plenty of theoretical and practical answers to multiple comparison problems and similar (see, for example, (Benjamini, 2010; Motulsky, 2013)). But, if the object of research is to find law-like connections, actually doing some experiments or gathering completely new data is a good idea.

At first glance, Big Data might excel at heuristics: at finding or suggesting or discovering possible candidate correlations; the problem with that is the numbers—every twenty or so comparisons of variables is going to produce a false positive, there would be a deluge of false positives. An interesting Big Data example is that of genetic association studies which in essence look at whether particular genes are associated with specific traits (height, weight, aggression, susceptibility to various medical conditions and so on); it seems fair to say that there have been problems in that discipline with studies not being able to replicate the studies of others. For example, Joel Hirschhorn et al. [2002] point out that of 600 genetic associations reported in studies only 6 were able to be consistently replicated (i.e just 1 in 100 was replicable) (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002). There may be more similar problems with data-driven science for

> …virtually any data-driven decision about modeling will lead to an overly optimistic model. (Babyak, 2004)

and

> A regression analysis usually consists of several stages such as variable selection, transformation and residual diagnosis. Inference is often made from the selected model without regard to the model selection methods that preceeded it. This can result in overoptimistic and biased inferences. (Faraway, 1992)

John Ioannidis has a paper explaining why most research findings are false (Ioannidis, 2005). Most research findings in psychology, medicine, information science, behavioral science, and social science may indeed be false, or, at least, the evidence offered for them may be unsound or inconclusive (Begley, 2013; Naik, 2011; Yong, 2012). There is a good portion of such modern science, or modern research in those fields as it is practiced and published, that is in a parlous state (Begley, 2013; Begley & Ellis, 2012; Ioannidis, 2005; Ioannidis & Khoury, 2011; Nosek, Spies, & Motyl, 2012; Yong, 2012). To amplify

> Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. (Ioannidis, 2005)

and the further argument is that typically the studies are irreproducible, the data is irreproducible, the data is unreliable, there is a lack of positive and negative controls, there is the inappropriate use of statistics (often leading to results that the investigator 'likes'), there is the investigator's ignoring of negative results, there is a pro-positive-result publication bias, and more… (Begley, 2013; Begley & Ellis, 2012; Ioannidis, 2005; Ioannidis & Khoury, 2011; Nosek et al., 2012; Yong, 2012). [This is worth a pause. The present journal is one of the finest in information science, yet it is likely that most of what it publishes, if there is any statistics present in the articles in question, is just wrong. This conclusion, which is certainly distressing, and perhaps also surprising, follows from the arguments, and evidence, provided in the articles cited above.]

Any research findings produced by data-driven science have the potential to be as bad if not worse. For example, data-driven science might envisage separate research teams working independently on the same shared data, but, if carried out naively, this is very similar to using multiple comparisons (there is some chance of the first team producing a false positive, there is an even better chance of either the first team or the second team producing a false positive, and so on, then, typically, only positive results are published). And some protections, for example, Registration, are inimical to data-driven science (if the plan is to collect, curate, analyze why should you want to register, collect, curate, register again, analyze?).

In sum, data-driven science is too *post hoc*. There are reasons, some suggested here, to suppose that data-driven science will or would find many spurious connections. Data-driven science could easily lead to apophenia and a wild outbreak of hornswoggling.

Conclusion

The ability to gather large amounts of data both cheaply and easily does have advantages: sample sizes can be larger, testing of theories can be better, there can be continuous assessment, etc. But data-driven science, the ‘fourth paradigm’, is a chimera. Science needs problems, thoughts, theories, and designed experiments. If anything, science needs more theories and less data.

References

Allen, J. F. (2001a). Bioinformatics and discovery: induction beckons again. *BioEssays: news and reviews in molecular, cellular and developmental biology, 23*(1), 104–107.

Allen, J. F. (2001b). Hypothesis, induction and background knowledge. Data do not speak for themselves. Replies to Donald A. Gillies, Lawrence A. Kelley and Michael Scott. *BioEssays: news and reviews in molecular, cellular and developmental biology, 23*(9), 861–862.

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired, 16*(07).

Austin, P. C., & Goldwasser, M. A. (2008). Pisces did not have increased heart failure: data-driven comparisons of binary proportions between levels of a categorical variable can result in incorrect statistical significance levels. *Journal of Clinical Epidemiology, 61*(3), 295-300.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*, 411–421.

Begley, C. G. (2013). Reproducibility: six red flags for suspect work. *Nature, 497*, 433–434.

Begley, C. G., & Ellis, L. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science, 423*, 1297-1298.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal, 52*(6), 708–721.

Berry, D. A. (2007). The difficult and ubiquitous problems of multiplicities. *Pharmaceutical Statistics, 6*(3), 155-160.

Bogen, J. (2013). Theory and observation in science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2013 Edition)*.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059–1078.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science, 16*(3), 199–231.

Breiman, L., & Cutler, A. (2013). *Random Forests*. Retrieved 10/25/2013, from http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Bretz, F., & Hsu, J. C. e. (2007). Special issue on multiplicity in Pharmaceutical Statistics, *Pharmaceutical Statistics* (Vol. 6, pp. 151-250).

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science, 42*(5), 351-360.

Cohen, J. (1994). The Earth is round (p<.05). *American Psychologist*, 997-1003.

Dietterich, T. G. (2003). Machine Learning. In *Nature Encyclopedia of Cognitive Science*. London: Macmillan.

Duhem, P. (1914). *La théorie physique son objet et sa structure, 2nd ed.,* (T. A. a. S. o. P. T. English Translation Phillip Wiener, Princeton: Princeton University Press, 1954., Trans.). Paris: Chevalier et Rivière. .

Faraway, J. J. (1992). On the cost of data analysis. *Journal of Computational and Graphical Statistics, 1*, 213-229.

Floridi, L. (2012). Big Data and their epistemological challenge. *Philosophy of Technology, 25*(4), 435-437.

Forster, M. (1995). The Curve-Fitting Problem. In R. Audi (Ed.), *The Cambridge Dictionary of Philosophy,* . Cambridge: Cambridge University Press.

Freedman, D. A. (1991). Statistical models and shoe leather (with discussion). *Sociological Methodology, 21*, 291–313.

Freedman, D. A. (1999 (revised 2002)). From association to causation: some remarks on the history of statistics. *Statistical Science, 14*(3), 243-258.

Freedman, D. A. (2008). Oasis or mirage? *Chance, 21*(1), 59-61.

Fricke´, M. (2009). The Knowledge Pyramid: a critique of the DIKW hierarchy. *Journal of Information Science, 35*, 131-142.

Galilei, G. (1638). *Discourses and Mathematical Demonstrations Relating to Two New Sciences (Discorsi e dimostrazioni matematiche, intorno à due nuove scienze).*

Gantz, J., & Reinsel, D. (2011). *Extracting value from chaos.* Retrieved 10/25/2013, from http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Gillies, D. A. (2001). Popper and computer induction. *BioEssays: news and reviews in molecular, cellular and developmental biology, 23*(9), 859 - 860.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific DIscovery*. Redmond, Washington: Microsoft Research.

Hillar, C., & Sommer, F. T. (2012). Comment on the article "Distilling free-form natural laws from experimental data". *(arχiv:1210.7273).*

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine, 4*(2), 45-61.

Howson, C., & Urbach, P. (2006). *Scientific Reasoning : the Bayesian Approach* (3rd ed.). Chicago: Open Court.

Huff, D. (1954). *How to lie with statistics.* New York: W.W. Norton.

International Committee of Medical Research Editors. (2013). *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: Publishing and Editorial Issues Related to Publication in Medical Journals: Clinical Trial Registration.* Retrieved 10/25/2013, from http://www.icmje.org/publishing_j.html

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8: e124. doi:10.1371/journal.pmed.0020124).

Ioannidis, J. P. A., & Khoury, M. J. (2011). Improving validation practices in "omics" research. *Science, 334*(6060), 1230–1232.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again. *Science, 334*(6060), 1225.

Johansson, T. (2011). Hail the impossible: p-values, evidence, and likelihood. *Scandanavian Journal of Psychology, 52*(2), 113-125.

Kell, D. B., & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays: news and reviews in molecular, cellular and developmental biology, 26*(1), 99–105.

Kelley, L. A., & Scott, M. (2001). On Allen's critique of induction. *BioEssays: news and reviews in molecular, cellular and developmental biology, 23*(9), 860 – 861.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. E. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, England: Cambridge University Press.

Lakatos, I. (1974a). Lakatos, I. Popper on demarcation and induction. In P. A. Schilpp (Ed.), The philosophy of Karl Popper (Vol. 1). LaSalle, Ill.: Open Court, 1974.

Lakatos, I. (1974b). The role of crucial experiments in science. *Studies in History and Philosophy of Science, 4*, 309-325.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking Clearly about Psychology* (Vol. 1): University of Minnesota Press.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*, 181–190.

Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the slow progress of Soft Psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Mills, J. L. (1993). Data torturing. *New England Journal of Medicine, 329*, 1196-1199.

Motulsky, H. J. (2013). *Intuitive Biostatistics*. USA: Oxford University Press.

Naik, G. (2011). Scientists' elusive goal: reproducing study results. *Wall Street Journal, CCLVIII*((130), A1, A16.).

Norvig, P. (2008). *All we want are the facts, ma'am*. Retrieved 10/25/2013, from http://norvig.com/fact-check.html

Nosek, B. A., Spies, J., & Motyl, M. (2012). Scientific Utopia: II - restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631.

Oceans Observatories Initiative. (2013). *Oceans Observatories Initiative*. Retrieved 10/31/2013, from http://oceanobservatories.org

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Popper, K. R. (1963). *Conjectures and Refutations*. London: Routledge and Kegan Paul.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *American Psychologist, 65*(1), 1–12.

Rosling, H. (2010). *The joy of stats*. Retrieved 10/31/2013, from http://www.gapminder.org/videos/the-joy-of-stats/

SCAMwatch, A. C. C. C. (2013). *'Nigerian 419' scams*. Retrieved 12/10/2013, from http://www.scamwatch.gov.au/content/index.phtml/tag/nigerian419scams

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*(5923), 81-85.

Skeptics Stack Exchange. (2013). *Will NSA's Utah Data Center be able to handle and process five zettabytes of data?* Retrieved 10/31/2013, from http://skeptics.stackexchange.com/questions/16829/will-nsas-utah-data-center-be-able-to-handle-and-process-five-zettabytes-of-dat

Sloan Digital Sky Survey. (2013). *Sloan Digital Sky Survey*. Retrieved 10/31/2013, from http://www.sdss.org

Sterne, L. (1759). *The Life and Opinions of Tristram Shandy, Gentleman*. Britain: Ann Ward (vol. 1–2), Dodsley (vol. 3–4), Becket & DeHondt (5–9).

Urbach, P. (1985). Randomization and the design of experiments. *Philosophy of Science, 52*(2), 256-273.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

Waltz, D., & Buchanan, B. G. (2009). Automating science. *Science, 3*(April 2009), 43-44.

Worrall, J. (2007). Why there's no cause to randomize. *British Journal for the Philosophy of Science, 58*, 451-488.

Yong, E. (2012). Bad copy. *Nature, 485*, 298-300.