NYU STERN

NEW YORK UNIVERSITY · LEONARD N. STERN SCHOOL OF BUSINESS

Working paper CeDER-12-01

May 2012

http:// http://hdl.handle.net/2451/31553

Data Science and Prediction

Vasant Dhar

Professor, Stern School of Business

Director, Center for Digital Economy Research

March 29, 2012

Abstract

The use of the term “Data Science” is becoming increasingly common along with “Big Data.” What does Data Science mean? Is there something unique about it? What skills should a “data scientist” possess to be productive in the emerging digital age characterized by a deluge of data? What are the implications for business and for scientific inquiry? In this brief monograph I address these questions from a predictive modeling perspective.

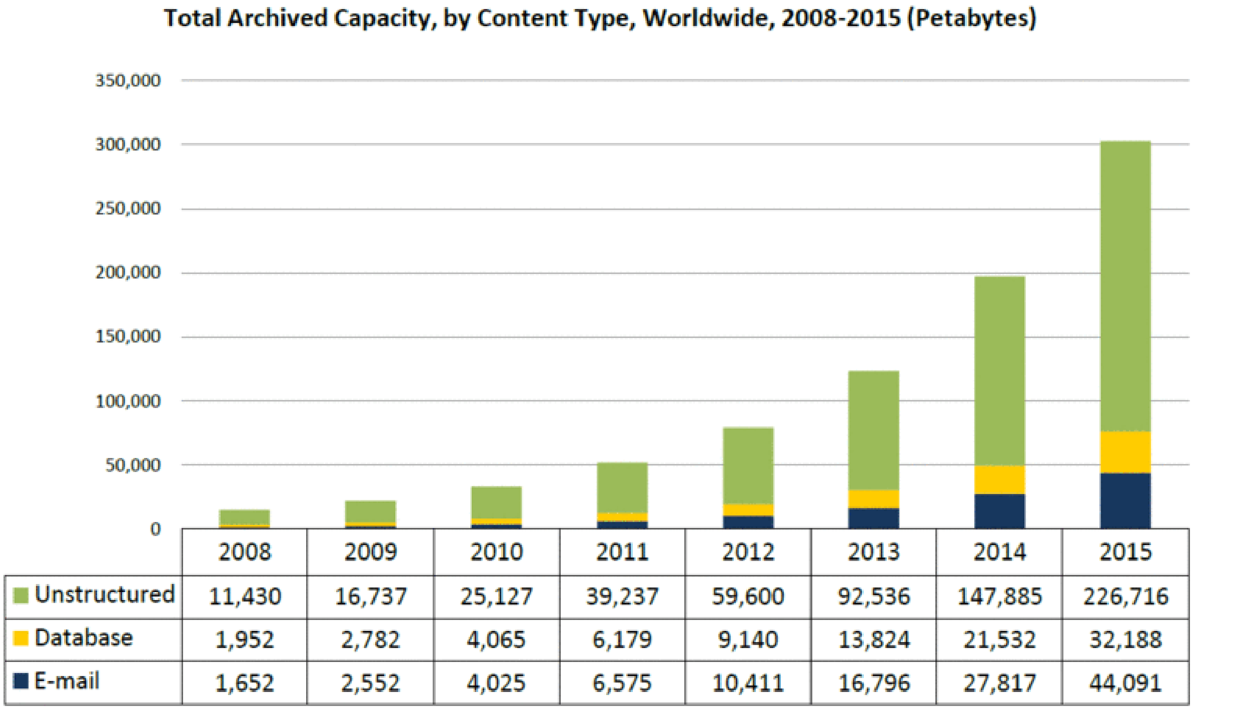
1. Introduction

The use of the term “Data Science” is becoming increasingly common along with “Big Data.” What does Data Science mean? Is there something unique about it? What skills should a “data scientist” possess to be productive in the emerging digital age characterized by a deluge of data? What are the implications for scientific inquiry?

The term “Science” implies knowledge gained by systematic study. According to one definition**,** it is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.[1] *Data* Science might therefore imply a focus around data and by extension, Statistics, which is a systematic study about the organization, properties, and analysis of data and their role in inference, including our confidence in such inference. Why then do we need a new term, when Statistics has been around for centuries? The fact that we now have huge amounts of data should not in and of itself justify the need for a new term.

The short answer is that it is different in several ways. First, the raw material, the “data” part of Data Science, is increasingly heterogeneous and unstructured – text, images, and video, often emanating from networks with complex relationships among its entities. Figure 1 shows the relative expected volumes of unstructured and structured data between 2008 and 2015, projecting a difference of almost 200 pedabytes in 2015 compared to a difference of 50 pedabytes in 2012. Analysis, including the combination of the two types of data with requires integration, interpretation, and sense making, increasingly based on tools from linguistics, sociology, and other disciplines. Secondly, the proliferation of markup languages, tags, etc. are designed to let computers interpret data automatically, making them *active agents* in the process of sense making. In contrast to early markup languages such as HTML that were about displaying information for human consumption, the majority of the data now being generated by computers is for consumption by other computers. In other words, computers are increasingly doing the background work for each other. This allows decision making to *scale:* it is becoming increasingly common for the computer to be the decision maker, unaided by humans. The shift from humans towards computers as decision makers raises a multitude of issues ranging from the costs of incorrect decisions to ethical and privacy issues. These fall into the domains of business, law, ethics, among others.

[1] *The Oxford Companion to the History of Modern Science* New York: Oxford University Press, 2003.



	2008	2009	2010	2011	2012	2013	2014	2015
🟩 Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
🟨 Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
🟦 E-mail	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

Source: Enterprise Strategy Group, 2010.

Figure 1: Projected Growth in Unstructured and Structured Data

From an epistemological perspective, the data explosion makes it productive to visit the age old philosophical debate on the limits of induction as a scientific method for knowledge discovery. Specifically, it positions the computer as a credible generator and tester of hypotheses by ameliorating some of the known errors associated with statistical induction. Machine learning, which is characterized by statistical induction aimed at generating robust predictive models, becomes central to Data Science.

From an engineering standpoint, it turns out that scale matters in that it has rendered the traditional database models somewhat inadequate for knowledge discovery. Traditional database methods are not suited for knowledge discovery because they are optimized for fast access and summarization of data *given what the user wants to ask* (i.e. a query), not *discovery* of patterns in massive swaths of data when the user does not have a well formulated query. Unlike database querying which asks "what data satisfy this pattern (query)," discovery is about asking "what patterns satisfy this data?" Specifically, our concern is in finding *interesting* and *robust* patterns that satisfy the data, where interesting is usually something unexpected and actionable, and robust means a pattern that is expected to occur in the future.

What makes an insight actionable? Other than domain-specific reasons, it is prediction. Specifically, what makes an insight actionable is its predictive power in that the return distribution associated with an action can be reliably estimated from past data and therefore acted upon with a high degree of confidence.

The emphasis on prediction is particularly strong in the machine learning and "KDD" communities. Unless a learned model is predictive, it is generally regarded with skepticism. This position on prediction mirrors the view expressed strongly by the philosopher Karl Popper as being a primary criterion for evaluating a theory and for scientific progress in general.[2] Popper argued that theories that sought only to explain a phenomenon were weak whereas those that make "bold predictions" that stand the test of time and are not falsifiable should be taken more seriously. In his well-known treatise on this subject, *Conjectures and Refutations,* Popper presented Einstein's theory of relativity as a "good" one since it made bold predictions that could be easily falsified. All attempts at falsification on this theory have indeed failed on this date. In contrast, Popper argued that a theory like Freud's, which could be "bent" to accommodate almost any scenario is weak in that it was virtually unfalsifiable.

The emphasis on predictive accuracy implicitly favors "simple" theories over more complex ones, a point we return to shortly. Data Science is increasingly about prediction on observations that will occur in the future. This requires a unique mindset, one that has heretofore seen little representation in typically academic curricula, in social science literature, and in commerce.

In the remainder of this short monograph I will discuss the implications of Data science from a business and research standpoint. I first talk about the implications for skills – what will people in industry need to know about and why? How should educators thinking about designing programs that deliver the skills most efficiently and enjoyably? Finally, what kinds of decision-making skills will be required in the era of big data and how will these be different from the past?

The second part of my answer is aimed at research. How can scientists use the abundance of data and massive computational power to their advantage in scientific inquiry? How does this new line of thinking complement our traditional methods of scientific inquiry? How can it augment the way we think about discovery and innovation?

2. Implications for Business

According to recent report by McKinsey and Company[3], the volume of data is growing at a rate of roughly 50% per year. This translates into a roughly 40-fold increase in ten years. Hundreds of billions of messages are transmitted on social media daily and millions of videos are uploaded daily across the Internet. As storage becomes virtually costless, most of this information is stored because businesses generally associate a positive option value with data. In other words, since it *may* turn to be useful in ways not yet foreseen, why not just keep it? (An indicator of how cheap storage has become is the fact that it is possible to store the world's current stock of music on a $500 device!)

The viability of using large amounts of data for decision making became practical in the 80s. The field of "Data Mining" started to burgeon in the early 90s as relational database technology matured and

[2] Popper, Karl. Conjectures and Refutations, 1968

[3] Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, June 2011.

business processes became increasingly automated. Early books on Data Mining[4] [5] [6] from the 90s described how various methods from Machine Learning could be applied to a variety of business problems. There was a corresponding explosion in the available software tools geared towards leveraging transactional and behavioral data for purposes of explanation and prediction.

An important lesson learned during the 90s was that machine learning "works" in the sense that these methods detect subtle structure in data relatively easily without having to make strong assumptions about linearity, monotonicity, or parameters of distributions. The downside is that these methods is that they also pick up the noise in data[7] and often have no way of distinguishing between what is signal and what is noise, a point we return to shortly.

Despite the drawbacks, there is a lot to be said for methods that don't force us to make assumptions about the nature of the relationship between variables before we begin our inquiry. This is not a trivial issue. Most of us are trained to believe that theory must originate in the human mind based on prior theory and data is then gathered to demonstrate the validity of the theory. Machine Learning turns this process around. Given a large trove of data, the computer taunts us by saying "If only you knew what question to ask me, I would give you some very interesting answers based on the data!"

The reality is that we often don't know what question to ask. For example, consider a healthcare database of individuals who have been using the healthcare system for many years, where a group has been diagnosed with Type 2 diabetes and some subset of this group has developed complications. We would like to know whether there is any pattern to complications and whether the probability of complication can be predicted and therefore acted upon.

What could the data from the healthcare system look like? Essentially, it would consist of "transactions," that is, points of contact over time of a patient with the healthcare system. The system records service rendered by a healthcare provider or medication dispensed on a particular date. Notes and observations could be part of such a record. Figure 2 shows what the raw data would look like for 5 individuals, where the data are separated into a "clean period" which captures history prior to diagnosis, the red bar which represents the "diagnosis" and the "outcome period" which consists of costs and other outcomes such as the occurrence of complications. Each colored bar in the clean period represents a medication, showing that the first individual was on three different medications prior to diagnosis, the second individual was on two, and the last three were on a single medication. The last two individuals were the costliest to treat and had complications represented by the downward pointing red arrows whereas the first three individuals had no complications.

[4] Gregory Piatetsky-Shapiro and William Frawley, eds., *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991

[5] Dhar, V and Roger Stein, *Seven Methods for Transforming Corporate data Into Business Intelligence*", Prentice-Hall, 1997

[6] Linoff, G and Michael Berry., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Wiley, 1997.

[7] Tukey, J W. *Exploratory Data Analysis*. Addison-Wesley 1977, referred to this phenomenon as "fitting to the noise."

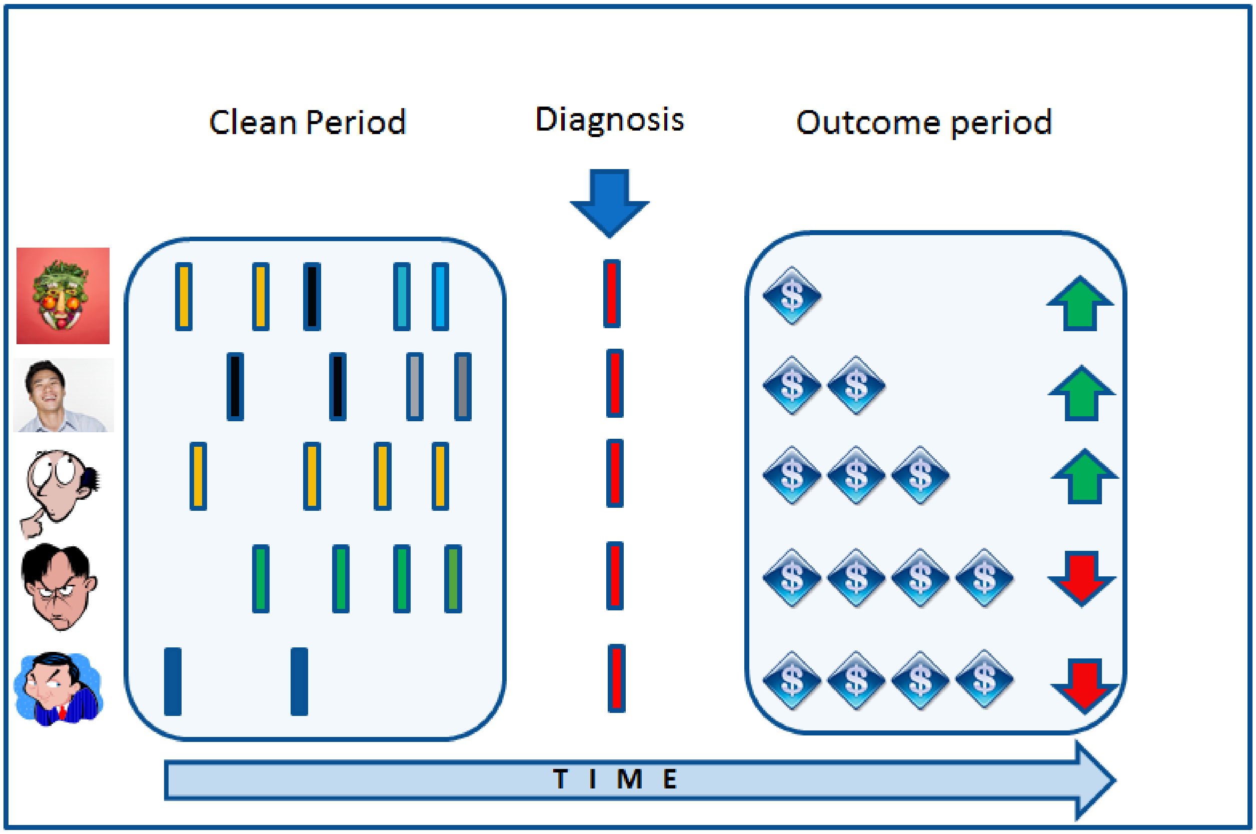
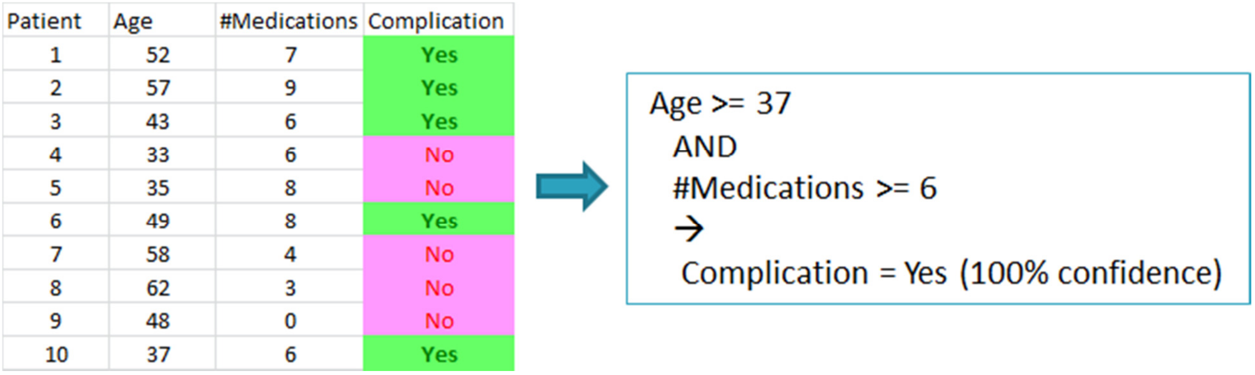


Figure 2: Healthcare Use Database Snippet

It is non-trivial to extract the interesting patterns from a large temporal database of the type above. For starters, the raw data across individuals typically needs to be aggregated into some sort of canonical form before useful patterns can be discovered. For example, suppose we count the number of prescriptions an individual is on at every point in time without regard to the specifics of each prescription as one approximation of the "health" of an individual prior to diagnosis. Another identifier might be the specific medications involved, where green and blue might be "severe" medications.

From the above data, a "complications database" might be synthesized from the raw data. This might include demographic information such as a patient's age and their medical history including a list of current medications aggregated into a count in which case we get a summary table of the type below. A learning algorithm, designated by the right facing blue arrow in Figure 3, could then be applied to discover the pattern shown on the right of the table. The pattern represents an abstraction of the data. Essentially, this is the type of question we *should* ask the database, if only we knew what to ask!



Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes

Age >= 37
AND
#Medications >= 6
→
Complication = Yes (100% confidence)

Figure 3: Extracting Interesting Patterns of Health Outcomes From Healthcare System Use

Why is the pattern on the right side interesting? To appreciate this, suppose the overall complication rate in the population is 5%. In other words, a random sample of the database would on average, contain 5% complications. Under this scenario, the snippet on the right hand side could be very interesting since its complication rate is many times greater than the average. The critical question here is whether this is a pattern that is robust and hence *predictive*, that is, likely to hold up on unseen cases in the future. The issue of determining robustness has been addressed extensively in the machine learning literature.[8]

If the above table is representative of the larger database, the box on the right *tells us the interesting question to ask our database*, namely, “What is the incidence of complications in Type 2 diabetes for people over 36 who are on more than five medications?” In terms of actionability, such a pattern would suggest being extra vigilant about people with this profile who do not currently have a complication because of their high susceptibility to complications.

The general point is that when the data are large and multidimensional, it is practically impossible for us to know a priori that a query such as the one above is a good one, that is, one that provides a potentially interesting insight. Suitably designed machine learning help find such patterns for us. Equally importantly, these patterns must be predictive. Typically, the emphasis on predictability favors Occam’s razor since simpler models generally have a higher chance of holding up on future observations than more complex models, all else being equal.[9] For example, consider the diabetes complication pattern above:

Age > 50 and #Medication > 6 → Complication_rate=100%

A simpler competing model might ignore age altogether, stating that people over 50 develop complications. The goodness of such a model would become more apparent when applied to future data. Does simplicity lead to higher future predictive accuracy in terms of lower false positives and false

[8] Perlich, C, Provost, F., and Simonoff, J., Tree Induction vs. Logistic Regression: A Learning-Curve Analysis, *Journal of Machine Learning Research,* 4, 2000.

[9] Dhar, V., Prediction in Financial Markets: The Case for Small Disjuncts, *ACM Transactions on Intelligent Systems and Technologies*, volume 2, Number 3, April 2011.

negatives? If so, it is favored. The practice of “out of sample” and “out of time” testing is used to assess the robustness of patterns from a predictive standpoint.

When predictive accuracy becomes a primary objective, the computer tends to play a significant role in model building and decision making. It builds predictive models through an intelligent “generate and test” process, the end result of which is an assembled model that is the decision maker. In other words, it takes automates Popper’s criterion of predictive accuracy for evaluating models at a scale that has not been feasible before. It is notable that the powerhouse organizations of the Internet era which include Google, and Amazon, and most of the emerging Web 2.0 companies have business models that hinge on predictive models based on machine learning. Indeed the first machine that could arguably be considered to pass the Turing test, namely, IBM’s Watson, could not have done so without extensive use of machine learning in how it interpreted questions. In a game like jeopardy where understanding the question itself is often a nontrivial task, it is not practical to tackle this problem through an extensive enumeration of possibilities. Rather, the solution is to “train” a computer to interpret questions correctly based on large numbers of examples.

Machine learning skills are fast becoming a necessary skill set in the marketplace as companies reel under the data deluge and try to build automated decision systems that hinge on future predictive accuracy. A basic course in machine learning is an absolute necessity in today’s marketplace. In addition, knowledge of text processing or “text mining” is becoming essential in light of the explosion of text and other unstructured data in healthcare systems, social networks, and other forums. Knowledge about markup languages such as XML and its derivatives is also essential as more and more content becomes tagged and hence capable of being interpreted automatically by computers.

Knowledge about machine learning must build on more basic skills which fall into three broad classes. The first is Statistics. This requires a working knowledge of probability, distributions, hypothesis testing and multivariate analysis. This knowledge can be acquired in a two or three course sequence. The last of these topics, multivariate analysis, often overlaps with the subject of econometrics which is concerned with fitting robust statistical models to economic data. Unlike machine learning methods which make no or few assumptions about the functional form of relationships among variables, multivariate analysis and econometrics by and large focus on estimating parameters of linear models where the relationship between the dependent and independent variables is expressed as a linear equality.

The second set of skills for a data scientist comes from Computer Science and pertains to how data are internally represented and manipulated by computers. This is a sequence of courses on data structures, algorithms, and database systems. The well-known textbook “Data Structures + Algorithms = Programs” expresses the fact that a program is a procedure that operates on data. Database systems are specialized programs optimized to access, store, and manipulate data. Together with scripting languages such as Python and Perl, database systems provide fundamental skills required for dealing with reasonably sized datasets. For handling very large datasets, however, standard database systems built on the relational data model has severe limitations. The recent move towards Hadoop for dealing with enormous datasets signals a new set of required skills for data scientists.

The final skill set is the most non-standard and elusive, but probably what differentiates effective data scientists. This is the ability to *formulate* problems in a way that results in effective solutions. Herbert Simon, the famous economist and "father of Artificial Intelligence" argued that many seemingly different problems are often "isomorphic" in that they have the identical underlying structure. Simon demonstrated that many recursive problems, for example, could be expressed as the standard Towers of Hanoi problem, that is, with identical initial and goal states and operators. Simon observed these differently stated problems took very different amounts of time to solve, representing different levels of difficulty even though they had the identical underlying structure. Simon's larger point was that is easy to solve seemingly difficult problems if represented creatively.[10]

In a broader sense, formulation expertise involves the ability to see commonalities across very different problems. For example, many problems of interest have "unbalanced target classes" usually denoting that the dependent variable is interesting only a small minority of the time. As an example, very few people commit fraud in population, very few people develop diabetes, and very few people respond to marketing offers or promotions. Yet, these are the cases of interest that we would like to predict. Such problems pose challenges for models which have to go out on a limb to make such predictions which are very likely to be wrong unless the model is very good at discriminating among the classes. Experienced data miners are very familiar with such problems and at knowing how to formulate problems in a way that give a system a chance of making correct predictions under conditions where the priors are stacked heavily against it.

The above represent "core skills" for data scientists over the next decade. The term "computational thinking" coined by Seymour Papert[11] and elaborated by Wing[12] is similar to the core skills we describe, but also encompasses abstract thinking about the kinds of problems computers are better at than humans and vice versa, and its implications. There is a scramble at universities to train students in the core skills, and electives that are more suited to specific disciplines. The McKinsey study mentioned earlier projects are roughly 200 thousand additional "deep analytical" positions and 1.5 to 2 million "data manages" over the next five years.

The projection of almost two million managers is not just about managing data scientists, but about a fundamental shift in how managerial decisions are being driven by data. The famous Ed Demming's quote has come to characterize the new orientation from intuition-based decision making to fact-based decision making: "in God we trust, everyone else please bring data." This isn't going to be an easy transition, requiring organizations to focus on change management. Imagine informing an expert clinician that the success rate based on his prescribed regimen is 40% or that the costs associated with the treatment are double the national average with an inferior outcome. Or a chief economist that the data support a hypotheses at odds with her theory? Or a veteran marketer that his proven methods no

[10] Simon, Herbert A.; Hayes, John R. The understanding process: Problem isomorphs. Cognitive Psychology, Vol 8(2), Apr 1976, 165-190.

[11] http://www.papert.org/articles/AnExplorationintheSpaceofMathematicsEducations.html

[12] Wing, J., Computational Thinking, Communications of the ACM, March 2006

longer work? Or the sports manager of a professional basketball team that his strategy is flawed against teams that have a majority of left-handed sluggers. The list goes on. Fundamentally managers will have to adapt their information gathering and decision making strategy in this new world.

More generally, we are moving into an era of big data where for many types of problems, computers are inherently better decision makers than humans, where “better” could be defined in terms of cost and accuracy. This shift has already happened in the world of data-intensive finance where computers now make the majority of investment decisions often in fractions of seconds as new information becomes available. The same is true in areas of online advertising where millions of auctions are conducted in milliseconds every day, air traffic control, routing, and many types of planning tasks that require scale, speed, and accuracy simultaneously. This trend is likely to accelerate in the near future.

3. Knowledge Discovery

In his provocative article titled “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete[13],” Chris Anderson drew on the famous quote by George Box that “All models are wrong, but some are useful,” arguing that with the huge amounts of data now available, we don’t need to settle for wrong models or any models for that matter. Anderson pointed out that prediction is of paramount importance to businesses, and that data can be used to let such models emerge using machine learning algorithms, largely unaided by humans. Anderson points to companies such as Google as symbolizing the triumph of machine learning over top-down theory development. Google’s language translator doesn’t “understand” language, nor do its algorithms know the contents on webpages. Nor does IBM’s Watson “understand” the question it is asked. There are dozens of lesser known companies that likewise are able to predict the odds of someone responding to a display ad, etc., without any solid theory, but rather, based on gobs of data about behaviors of individuals and the similarities and differences among these behaviors.

Anderson’s article set off a vigorous debate in academic circles. How can one have science and predictive models without first articulating a theory?

The observation that “patterns emerge before reasons for them become apparent” [14] tends to resonate universally among people, particularly in financial markets, marketing, and even healthcare. If this is true, Box’s observation becomes very relevant: if a problem is non-stationary and a model will be approximate, why not build the best predictive model based on data available until that time and just update it periodically? Why bother developing a detailed causal model if it is poor at prediction and more importantly, likely to get worse over time due to “concept drift?”

Anderson’s point has particular relevance in the health, social, and earth science in this era of big data since these areas have generally been characterized by the lack of solid theory but where we are now

[13] Wired Magazine, June 23, 2008

[14] Dhar, V., and D Chou., A Comparison of Nonlinear Models for Financial Prediction, IEEE Transactions on Neural Networks, June 2001.

seeing huge amounts of data that can serve as grist for theory building. For illustration, contrast the areas of physics and social sciences which lie at opposite ends of the spectrum in terms of the predictive power of their theories. In physics, a theory is expected to be “complete” in the sense that a relationship among certain variables is intended to explain the phenomenon *completely*, with no exceptions. Such a model is expected to make perfect predictions (subject to measurement error, but not error due to omitted variables or unintended consequences). In such domains, the explanatory and predictive models are synonymous. The behavior of a space shuttle, for example, is completely explained by the causal model that describes the forces acting on it. This model can also be used to predict what will happen if any of the inputs change. It is not sufficient to have a model which is 95% sure of outcomes, and leave the rest to chance. Engineering follows science.

In contrast, the social sciences are generally characterized by incomplete models that are intended to be partial approximations of reality, often based on assumptions of human behavior known to be simplistic. A model that is correct 95% of the time in this world would be considered very good. Ironically, however, the emphasis in social science theory development is on proposing theories that embody causality without serious consideration of their predictive power. When such a theory claims that “A causes B” data are gathered to confirm whether the relationship is causal. But its predictive accuracy could be poor because the theory is incomplete. Indeed, it is not uncommon for two experts in the social sciences to propose opposite relationships among the variables and to offer diametrically opposite predictions based on the same sets of facts. Economists, for example, routinely disagree on both theory and prediction, and are often wrong in their forecasts.

How could big data put these domains on firmer ground?

Hastie et al enumerate that errors in prediction come from three sources[15]. The first type is from misspecification of a model. For example, a linear model that attempts to fit a nonlinear phenomenon will generate an error simply because the linear model imposes an inappropriate bias on the problem. The second source of error is from the use of samples for estimating parameters. The third is due to randomness, even when the model is perfectly specified.

[15] Hastie, T, Tibsharani,R, Friedman, J., *The Elements of Statistical Learning:* Data Mining, Inference, and Prediction, Springer 2009.

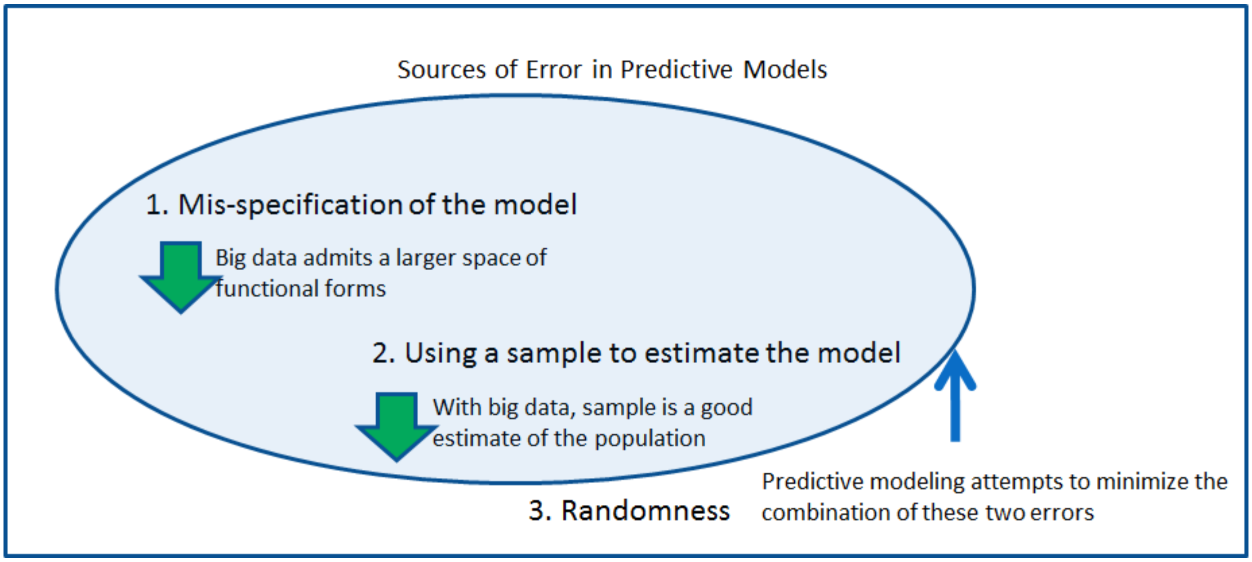

Sources of Error in Predictive Models
1. Mis-specification of the model
Big data admits a larger space of functional forms
2. Using a sample to estimate the model
With big data, sample is a good estimate of the population
Predictive modeling attempts to minimize the combination of these two errors
3. Randomness

Figure 4: Sources of Error in Predictive Models and Their Mitigation

As illustrated in Figure 4, big data allows us to significantly reduce the first two types of errors. Large amounts of data allow us to consider richer models than linear or logistic regressions simply because there is a lot more data to test such models and compute reliable error bounds. Big data also eliminate the second type of error as sample estimates become reasonable proxies for the population.

The theoretical limitation of observational data, regardless of how big it is, is that it is generally "passive," representing what actually happened in contrast to the multitude of things that *could have happened* had the circumstances been different. In the healthcare example, it is like having observed the use of the healthcare system passively, and now having the chance of understand it in retrospect and extract predictive patterns from it. The data do not tell us what could have happened if some other treatment had been administered to a specific patient or to an identical patient. In other words, it does not represent a clean controlled randomized experiment where the researcher is able to establish controls and measure the differential impact of treatments on matched pairs.

Interestingly however, we are now in an era where there is increasing possibilities of conducting large scale randomized experiments on behavior on the Internet and uncover interesting interactions that are not possible to observe in the laboratory or through observational data alone. In a recent controlled experiment on "influence versus homophily" conducted on Facebook via an "app," Aral and Walker uncovered the determinants of influence on online video games.[16] Their results include patterns such as "Older men are more influential than younger men," "people of the same age group have more influence on each other than from other age groups," etc. These results, which are undoubtedly peculiar to video games, make us wonder whether influences are different for different types of products, and more generally that influence is more complicated than we thought previously, not amenable to simple generalizations like Gladwell's[17] concept of "super influencers." The last of these assumptions has also

[16] Aral and Walker, Science, forthcoming.

[1717] Gladwell, M., The Tipping Point: How small things can make a big difference, Little Brown, 2000.

been questioned by Goel et.al[18] who observe in large scale studies that influence in networks is perhaps overrated.

More generally, however, social science theory building is likely to get a big boost from big data and machine learning. Never before have we been able to observe human behavior at a degree of granularity we are seeing now with increasing amounts of human interaction and economic activity being mediated by the Internet. While there are clearly limitations to the inductive method, the sheer volume of data being generated not only makes it feasible, but practically speaking, little us with little as an alternative. We do not mean to imply that the traditional scientific method is "dead" as claimed by Anderson. To the contrary, it continues to serve us well. However, we now have a new and powerful method at our disposal for theory development that was not previously practical due to the paucity of data. That era is largely over.

4. Concluding Remarks

There is no free lunch. While large amounts of observational data provide us with unprecedented opportunity to develop predictive models, they are limited when it comes to explanation[19]. Since it is impossible to run controlled experiments, except by design, we cannot know the consequences of things that did *not* transpire. A limitation of this is that we are limited in our ability to impact the future through intervention that is possible when the causal mechanisms are well understood.

The second limitation of predictive modeling with causation is that multiple models that appear different on the surface might represent the same underlying causal structure, but there is no way to know this. For example, in the diabetes example, there could be multiple uncorrelated robust patterns that predict complications. The good thing, however, is that If they are predictive, they are still useful in that they could suggest multiple observable conditions that lead to complications that should therefore be carefully monitored.

Despite the limitations of observational data, however, the sheer size of the data allows us to slice and dice the data in many ways without losing sample size, a limitation that has traditionally hindered our ability to examine conditional relationships in data even if they were real. The ability to interpret unstructured data and integrate it with numbers further increases our ability to extract useful knowledge in real-time and act on it. Incredibly, HAL isn't just a fantasy now but an imminent reality.

[18] Goel, S., Watts, D., and Goldstein., The Structure of Online Diffusion Networks, ACM 2012.

[19] Shmueli, G., To Explain or To Predict? Statistical Science, Volume 25, Number 3, 2010.