Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **3**, Number **1**, January **2018**
doi:10.3934/bdia.2018001
pp. **22–42**

UNDERSTANDING AI IN A WORLD OF BIG DATA

Richard Boire

Environics Analytics
33 Bloor St. East, Toronto, Ont. M4W3H1, Canada

(Communicated by Xiaomin Zhu)

Abstract. Big Data and AI are now very popular concepts within the public lexicon. Yet, much confusion exists as to what these concepts actually mean and more importantly why they are significant forces within the world today. New tools and technologies now allow better access as well as facilitating the analysis of this data for better decision-making. But the discipline of data science with its four-step process in conducting any analysis is the key towards success in both non-advanced and advanced analytics which would, of course, include the use of AI. This paper attempts to demystify these concepts from a data science perspective. In attempting to understand Big Data and AI, we look at the history of data science and how these more recent concepts have helped to optimize solutions within this 4 step process.

1. Introduction.

Data Science, now the more glamorous term as opposed to data mining, and its practitioners have been executing projects with extremely large amounts of data and have deployed techniques using advanced machine learning techniques for many years. The ever growing recognition of both the social and economic benefits of data analytics and data science have resulted in increasing demand for these disciplines. The appetite for more data and more advancements in machine learning have been one outcome for this increased demand. The use of Big Data and AI are direct concepts and terms that are being deployed as enablers in this brave new world.

The increasing importance of data analytics and data science applications has been manifested in a number of business and social applications. Let's look at the social applications. Rudy Giuliani, when he was mayor of New York City, was a pioneer in the use of data and analytics in how to better allocate police resources. The success of data analytics in improving law enforcement in New York City has now been adopted in many cities throughout North America. Besides its impact on law enforcement, its impact has been extended to other areas that are under the purview of governments be it municipal, state/provincial or federal. Examples include the use of data and analytics for better transit and transportation, management of health care as well as the use of more advanced techniques in identifying the specific courses of actions in preventing negative health care outcomes. Government policies and strategies, which in the past, were based on very broad macro/aggregate type analysis can now be based on a much finer and more granular type of analysis.

2010 *Mathematics Subject Classification.* 91C20, 62F07.

Key words and phrases. Artificial intelligence, data science, big data, machine learning, predictive analytics.

This allows a much higher degree of analysis in terms of generating insights as to why a specific government policy or strategy led to a certain outcome.

Some of the other newer areas that are leveraging big data and advanced analytics are sports and human resources. In some sense, both human resources and sports are linked in that sport organizations, given finite salary cap numbers in how much they can spend on player personnel, need to use analytics on how to best align their money with the available players on the market. The pioneering book on sports analytics was Michael Lewis' book called Moneyball which looked at baseball and how a team like the Oakland A's need to use analytics and big data in order to compete against very large payroll teams such as the Boston Red Sox and New York Yankees. The use of data to explore new metrics that contributed to team success were now considered as key components in assessing player value. The Oakland A's were able to hire so-called "undervalued" players who were otherwise not considered premium players. Yet according to the A's, they demonstrated those key "winning" components. The end result was a playoff berth which had previously eluded them for 8 years.

Most other sports are now embracing analytics and data particularly where there are rules regarding salary caps per team. This stringent regulation necessitates the use of data and information and the result is a growth in general managers and/or assistant general managers who are versed in the use of data and analytics but also have very strong domain knowledge regarding the sport. One needs to know the intricacies and mechanics of the sport besides just how the game is played. This requires technical knowledge of the mechanics of the sport and what teams and players have to do in order to be successful.

In the non-sports world, human resource professionals are leveraging data on applicants and their likelihood to be long-term "good to excellent" employees. The discipline of human resources analytics is becoming a common business process in many organizations today. This increased demand for analytics in a world awash of big data has seen tremendous growth within academia as attested by the emergence of data science and analytics disciplines within many post secondary institutions. The emergence of online education has just accelerated this trend. Beyond the academic world, software tools and technologies have also exploded in an attempt to satisfy this growing need. The tools are increasing the empowerment of analytics to larger groups of people that were previously considered less technical. At the same time, many of these new tools increase the operationalization and automation of our solutions, thereby increasing the speed of delivery of these solutions.

Yet, as we will state throughout this chapter, the human practitioner is the key to success. Big Data and artificial intelligence are enablers but optimization of solutions cannot occur without a deep understanding of the data science process. It is the hope of this chapter to at least provide some insight on some of the key considerations within this data science process.

1.1. **Evolution of Big Data.** Disruption and transformation seem to be the prevailing hallmark trends of our world today. In fact, one might argue that they are the only constants within our increasingly frenetic environment. For those of us who are baby boomers, the changes over the last 25 years have been nothing less than startling. Technology has been the real engine behind these changes which has significantly enhanced productivity levels but often at the expense of jobs. But let's look at these changes as it relates to Big Data and AI. Neither of these concepts or terms are new. In big data, the 5 V's are often presented as the key factors in

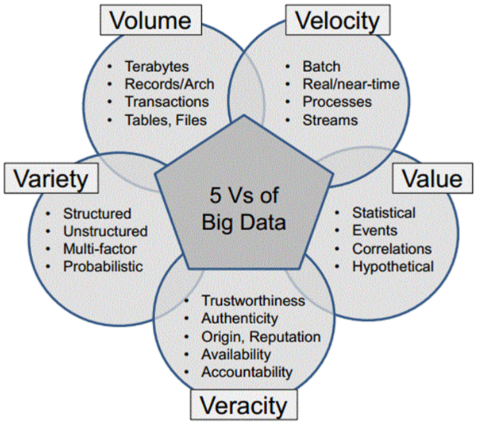


Figure 1. [1] The 5 V's of Big Data

operating within this kind of environment. Figure 1 is a chart which depicts these 5 V's.

Back in the early eighties when I began my data science career, big data was not a new phenomenon at least in terms of volume which is one of the 5 V's of big data. Direct marketers such as Reader's Digest were conducting CRM or one to one marketing campaigns to each of their 6 million customers. In order to process all this customer data alongside the tens of millions of transaction records, a mainframe computer was built and housed in a space comprising two meeting rooms. The primary competitive advantage of direct marketers was the ability to target individual-level customers. Direct marketers such as Reader's Digest employed machine learning techniques (regression and decision trees) as a means to target customers. In developing these models, the actual processing of the data with these techniques was an overnight process as in those days data scientists used samples as the source data to build models. Sample sizes for building models were determined by the fact that the sample would generate at least 500 responders (250 for model development and 250 for model validation). Numbers below 500 responders would be considered too sparse to generate a robust model. For example, a model to be developed from an initiative that typically generates a 2% response rate would need 25000 names in order to provide a minimum of 500 responders (500/.02). But the real big data component was the deployment of these models against the 6mm customers as well as updating these 6mm customers alongside the tens of millions of records associated with their campaign and transaction history. Multiple model scores would be produced for each business initiative whereby each model score would represent the customer's likelihood to respond to some Reader's Digest initiative. Due to the demands on this system which included other functions such as bill processing, the processing of individual customer data was executed once a month and on weekends. As one can observe, we were dealing with big data but in a very inefficient manner.

Having worked with a variety of financial institutions in the late 80's and early 90's, these above issues were only accelerated if one considers that we now have hundreds of millions of transactions as opposed to tens of millions of records. Besides increasing volumes, the second V of big data was the increasing variety of data as there were simply many more data fields to be used by the data scientist in building

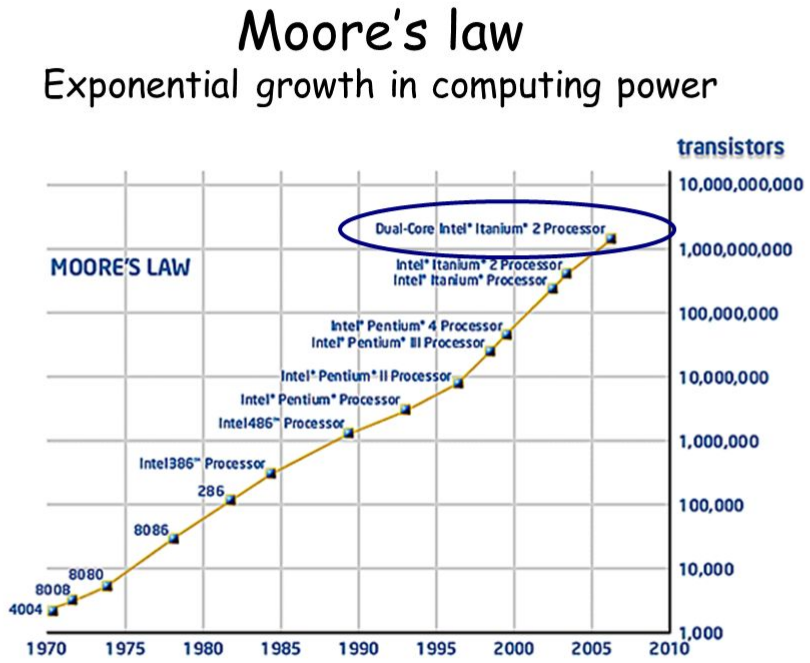


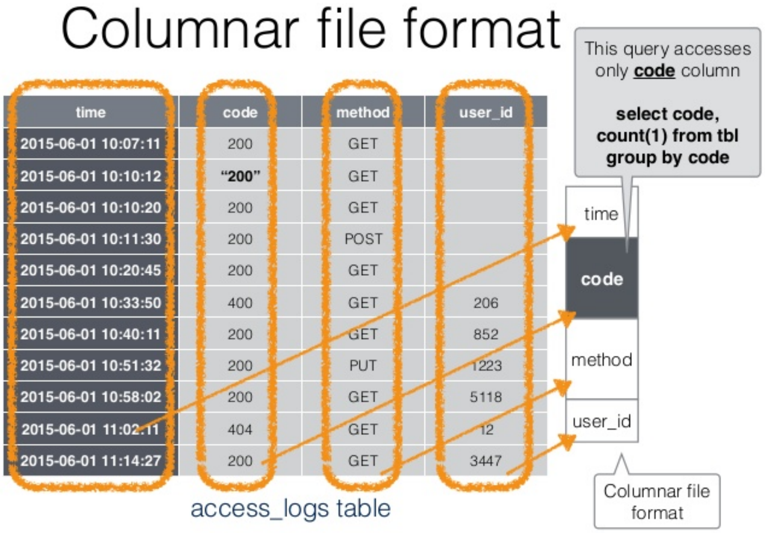
FIGURE 2. [2] Moore's Law

a solution. Once again, machine-learning techniques were utilized to develop many different kinds of predictive models for the many products and services that are delivered by financial institutions.

Advances in technology in the 90's allowed data scientists to develop models directly on their PC. Increased computing power allowed the user to generate larger sample sizes. In fact, sample sizes in excess of 100000 names were not unusual. The processing of model development no longer needed to occur in a batch overnight run but rather could be conducted in either seconds or minutes depending on the CPU power of the computer. A quick look at the chart (Figure 2) looks at the growth in computing power over the last several decades. Note how that line becomes increasingly steeper after 1995 which would be the time that the development of machine-learning tools were being conducted more frequently on a PC rather than on a mainframe system.

Deployment of these models, though, against millions of names reverted back to the batch process of scoring the names either overnight or on a weekend. Certainly financial institutions were not unfamiliar with the notions of Big Data. Having discussed volume and variety of data as common factors to be dealt within that environment, the remaining factors of velocity, veracity, and value also needed to be addressed. The issues of value and veracity are also not new as both these issues deal with the quality of the data and how to extract value from it. This will be discussed in more detail when we explore the four step process of data science.

Velocity of data implied how often does information need to be updated for analytics and machine learning purposes. Database technology in the mid 90's relied



Columnar file format
This query accesses only code column
select code, count(1) from tbl group by code
time
code
method
user_id
Columnar file format
access_logs table

time	code	method	user_id
2015-06-01 10:07:11	200	GET	
2015-06-01 10:10:12	"200"	GET	
2015-06-01 10:10:20	200	GET	
2015-06-01 10:11:30	200	POST	
2015-06-01 10:20:45	200	GET	
2015-06-01 10:33:50	400	GET	206
2015-06-01 10:40:11	200	GET	852
2015-06-01 10:51:32	200	PUT	1223
2015-06-01 10:58:02	200	GET	5118
2015-06-01 11:02:11	404	GET	12
2015-06-01 11:14:27	200	GET	3447

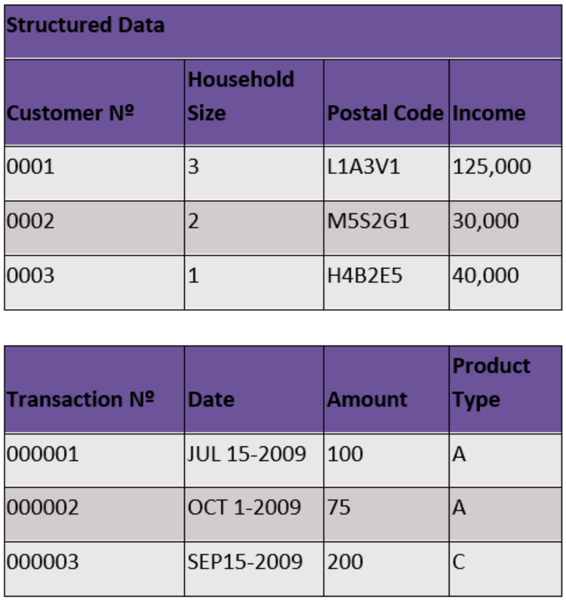
FIGURE 3. [3] Columnar File Format

primarily on relational database technology where match keys or links between files were indexed such that the join between two files could occur at a much faster rate than if the match key or link were unindexed. This capability was seemingly acceptable in a batch frame type environment. But as data volumes grew, answers to the demand for analytics continued to accelerate and as a result this situation became no longer acceptable. Hence the need for a new type of database technology which has often been referred to as columnar or inverted flat file technology. The basis of this technology is that instead of one field, typically the link field between files being indexed, all fields are now indexed. See Figure 3.

The obvious question, here, is once the concepts of relational database technology were understood and became the norm, why would this same technology of indexing not be applied to all fields within a database. The underlying constraint was disk size. If a given field is indexed, its disk size is greatly expanded. This theory is greatly magnified when all fields are indexed. The technology was unworkable in that the huge disk size of the database prevented any function or operation from effective completion due to the limited availability of memory. In effect, the I/O ability capabilities of these operations were altered in the sense that the read operations occurred in a much faster timeframe while the write operations were significantly slower due to the increased disk size of the database.

So what happened to fundamentally change this paradigm. The game changer was advances in compression technology which could mitigate the impact of disk size thereby leveraging the already existing efficiencies of the read operation while now allowing for quicker write functions due to smaller database disk sizes and increased memory [4].

Columnar or inverted flat file technology became a more accepted reality where analytics and the need for quicker analyses against hundreds of millions of records became the norm. This was manifested by the increased number of vendors offering services in the analytics software that would provide these type of capabilities.



Structured Data			
Customer Nº	Household Size	Postal Code	Income
0001	3	L1A3V1	125,000
0002	2	M5S2G1	30,000
0003	1	H4B2E5	40,000

Transaction Nº	Date	Amount	Product Type
000001	JUL 15-2009	100	A
000002	OCT 1-2009	75	A
000003	SEP15-2009	200	C

Figure 4. Example of Structured Data

Virtually all of the major DB vendors offered this type of capability by the early 2000's.

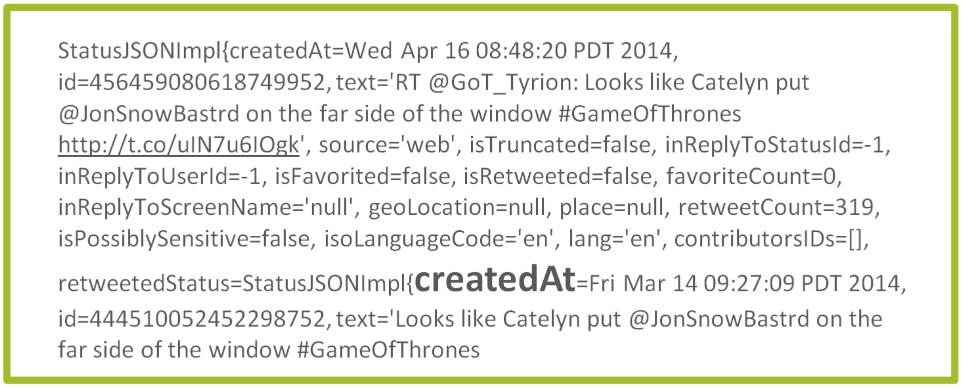
Certainly, organizations now understood the potential of conducting quick analytics on hundreds of millions of records. However, the analytics paradigm was largely confined to structured data with rows and columns where rows represent records and columns represent fields. See Figure 4.

With the recognition that analysis of data through the granular approach of data science was now yielding significant business gains, the need for alternative sources of data became an area of exploration. The advent of social media with Facebook in 2003 alongside the emergence of other providers such as Linked-In, Twitter, etc. provided new platforms where data was being captured on individuals. But virtually all of this data is either semi-structured or unstructured. In other words, the source of the data is not arising as structured rows and columns. Instead, the data is arriving as objects rather than as a column field within a record. See Figure 5 for below example of Twitter data.

In the above schematic, I have highlighted the term "createdAt" as one object (related to date of when tweet occurred). This type of data is referred to as semi-structured data. Meanwhile, unstructured data represents the actual post of what someone would write within the specific social media platform. See below example.

"I really like the RRSP product and will continue to invest every year. However, the level of service is sub-standard and I will be looking at other companies. But it will be difficult since I have so many products with this institution."

The application of data science techniques to unstructured data actually represents a discipline called text mining. In text mining, we are trying to explore all the



StatusJSONImpl{createdAt=Wed Apr 16 08:48:20 PDT 2014,
id=456459080618749952, text='RT @GoT_Tyrion: Looks like Catelyn put
@JonSnowBastrd on the far side of the window #GameOfThrones
http://t.co/uIN7u6IOgk', source='web', isTruncated=false, inReplyToStatusId=-1,
inReplyToUserId=-1, isFavorited=false, isRetweeted=false, favoriteCount=0,
inReplyToScreenName='null', geoLocation=null, place=null, retweetCount=319,
isPossiblySensitive=false, isoLanguageCode='en', lang='en', contributorsIDs=[],
retweetedStatus=StatusJSONImpl{**createdAt**=Fri Mar 14 09:27:09 PDT 2014,
id=444510052452298752, text='Looks like Catelyn put @JonSnowBastrd on the
far side of the window #GameOfThrones

Figure 5. Example of Twitter Data

text (corpus) data with the intention of trying to uncover insights from the data. Unlike search engine technology where we know what we are looking for within the text, text mining involves the analysis of text data without knowing what we are looking for. The objective in building text mining solutions is the identification of major themes or topics from the text.

But the ability to use and analyze this social media data in effect required a new quantum leap of database technology due to the exponential growth in the volume of data. The typical approaches of employing sequential database technology as a means to process data was no longer acceptable. So what was the breakthrough? Let's think about this in terms of search engine technology. In the late 90's, submitting a google search engine request was essentially attempting to match our request versus all content on the web. Think of the vast amount of data in the worldwide web and it is easy to understand that traditional database technologies would no longer suffice. The technology employed by these early Web tech pioneers would scrape the entire web and then deliver results based on the content of my search engine query. Despite my delight at how quickly the response was generated and given my understanding of databases, I often wondered how this was achieved. Enter Doug Cutting, the founder of Hadoop technology which represented the technical capability of data being processed in parallel. The creation of parallel nodes where components of the data could be partitioned for processing as opposed to all the data being processed at once represented the great breakthrough in processing speed. This parallel type of processing is often referred to as HDFS or Hadoop File Distributed System. Figure 6 is a schematic of what this means when looking at sequential vs. parallel processing.

In the HDFS type technology, the data is first mapped to the different nodes and is then reduced in terms of the output that we need which then produces the overall result. Hence from the above diagram, we can see why the term Map-Reduce is the underlying technology behind Hadoop. But you may wonder why the name Hadoop. It originates with Doug Cutting who coined the name Hadoop after his young son's toy elephant.

With access to this data now available as a result of this technology, organizations began to explore the analytics possibilities. This has led to the development of NOSQL databases such as MongoDB and Riak which are just a couple of the type of platforms that can process structured, semi-structured, and unstructured data.

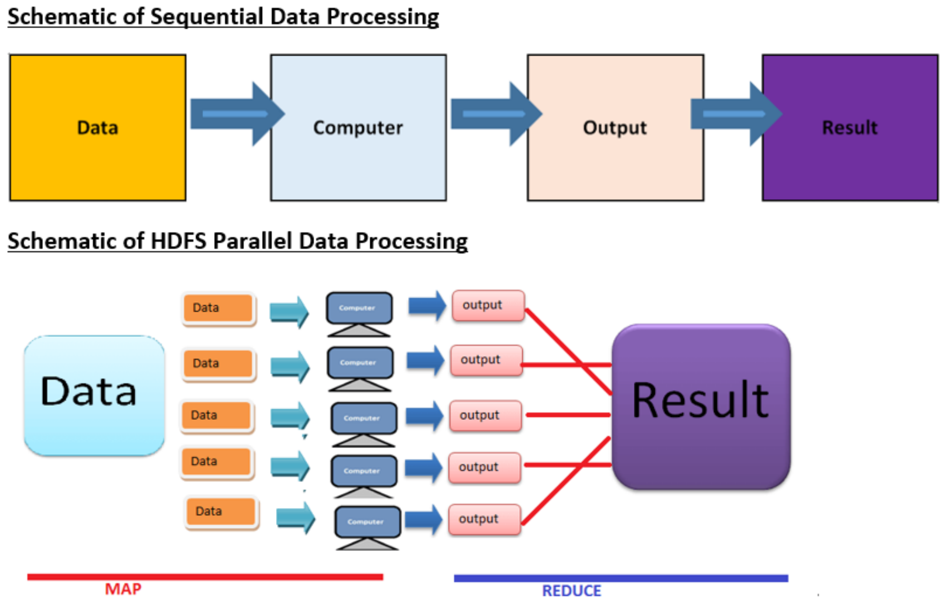


FIGURE 6. Sequential vs. Parallel Data Processing

The main advantage of such a system is there is no rigid structure or protocols unlike relational databases which have very strict protocols that must be adhered to when designing a database. For example, a relational database would have a series of tables with designated fields in each table. In a NOSQL system, we may actually have relational database tables within this system as well as data that does not conform to a rigid database structure such as social media data. But all this data in a NOSQL type system can be analyzed through SQL type tools.

The challenge, though, with these original tools was that data could only be processed in batch as opposed to processing data in real-time or streaming type data. Apache Spark solved this dilemma through its in-memory processing technology which utilized the Map Reduce technology whereby data can be processed in-memory within each node of the computer. Waiting times for analytics projects can be significantly reduced from 10 times to 1000 times or more [6].

Meanwhile, the development of GPU's (graphical processing units) actually represent the hardware to really leverage this parallel processing architecture. Each GPU has its own ability for in-memory processing which is the key to fast processing.

1.2. **Enter the Discipline of AI.** Certainly, the ability to analyze more data and in a more efficient manner has now yielded the potential to create better solutions. But it is the synergy between Big Data and AI that are essentially transforming the economy. The discipline of AI or artificial intelligence research is not a new topic as research was conducted many decades ago at institutions such as MIT and Stanford. Although serious academic research was devoted to this area, the social and public reaction to notions concerning AI was that these were useful topics to consider within the genre of science fiction. The notion of robots and self-autonomous vehicles were common themes within this genre. Yet, there was this implicit recognition

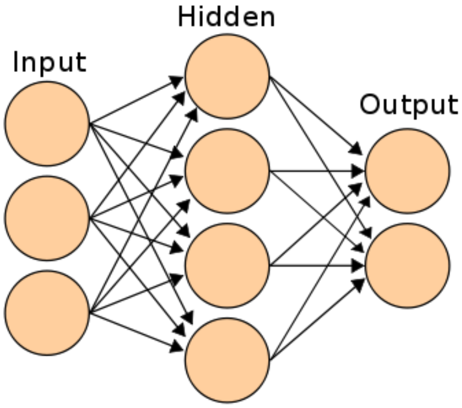


Figure 7. Schematic of Simple Neural Net-One Hidden layer

that these so-called "sci-fi" scenarios would indeed happen at some future period. For many of us, this seemed to exemplify a period that was beyond our lifetime. And in fact, given the work and research that had been done in artificial intelligence up until the last five years, minimal success was being achieved within this area. To give you some perspective, academic articles were published in the early 90's on using AI for image recognition. In those days, the success rate was in the neighborhood of 40%-50% which was clearly not an outcome that would advocate the widespread usage of AI. Yet, it is the last 5-10 years that have witnessed transformational results with success rates on image recognition now reaching in excess of 95%. Not only are we seeing its successful use in image recognition but we are also observing its widespread usage in the area of text recognition. So what are the underlying reasons for its huge success in today's environment? However in order to truly understand the magnitude of this change in success, one needs some basic level understanding of how AI works. First, the underlying mathematics behind AI is the concept of the neural net. Much as the name sounds, the mathematics attempts to replicate the processes that underpin the human neurological system. See Figure 7 a simple one hidden layer neural net system where there are three layers (input layer, hidden layer, and output layer).

Alongside each of these layers are nodes. In the example above, we have three nodes in the input layer, four nodes in the hidden layer and two nodes in the output cell. In neurological terms, we might think of the nodes as actual nerve cells or neurons with the arrows representing the synapses which transmit messages between neurons. Through the mathematics that is used within the neural nets which involved feedbacks loops, weights are assigned to these nodes. These weights are adjusted through these feedback loops in its attempt to optimize a given mathematical procedure. Let's take a look at these weights schematically - see Figure 8.

Within neural nets, a number of different algorithms can be employed, but the commonality amongst them is the ability to identify a point within a given data distribution where incremental change within the data is not occurring or in fact is being minimized. These algorithms will differ in many cases due to the activation functions which are attempting to identify that optimal point where incremental

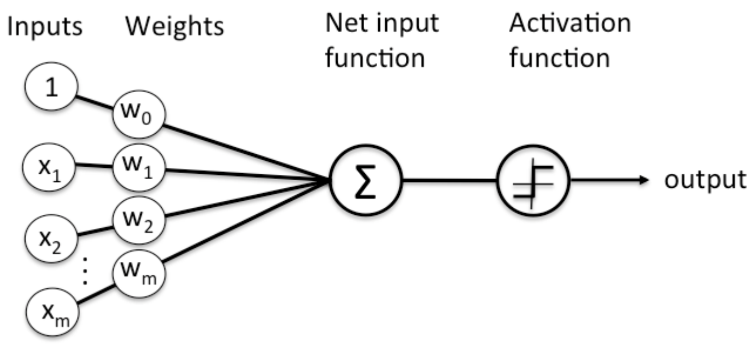


FIGURE 8. [7] Schematic of Weights within Neural Net Structure

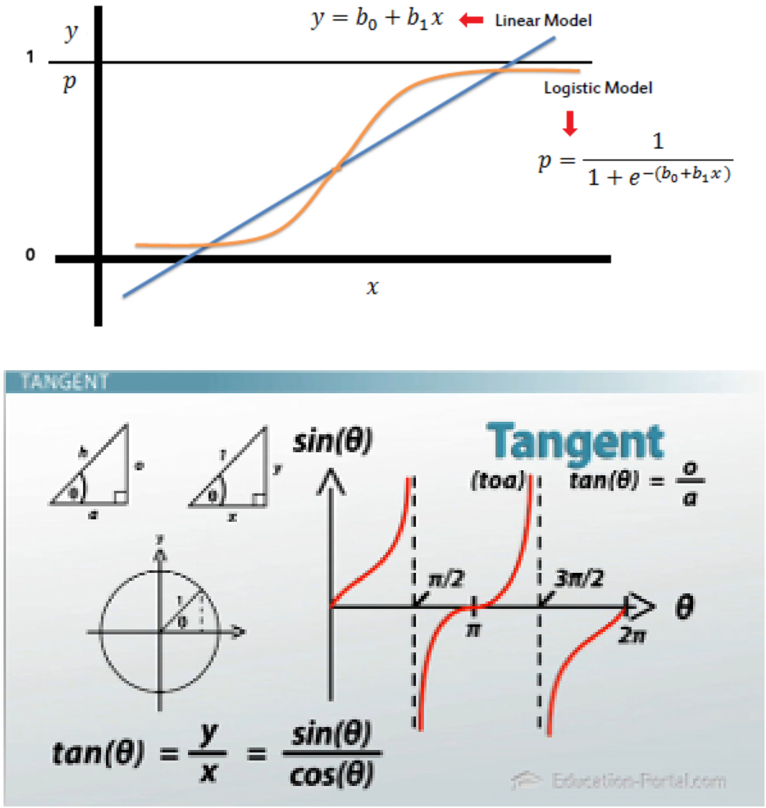


FIGURE 9. [8] Examples of some Optimization Algorithms

change is being minimized. A number of different activation functions can be employed with some examples being linear, logistic, tan, etc.

There are a number of other activation functions that can be used to optimize a given outcome - see Figure 9. But as will be discussed below, they are not the

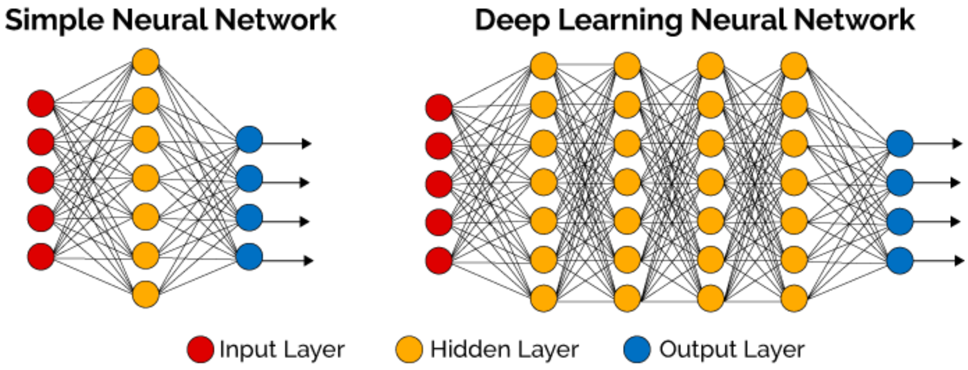


FIGURE 10. Examples of Neural Nets

only mechanisms that one can use in order to optimize outcomes within a neural net. The mathematical functions discussed above are not new and have existed for decades. However, the real game changer has been the ability to utilize more hidden layers as well as increasing the number of nodes. This ability to utilize multiple hidden layers and more nodes has allowed AI data scientists to more fully leverage the mathematics where large volumes of data or Big Data are critical to its success. This is ultimately the essence of what we call deep learning. See Figure 10 - Examples of Neural Nets.

In other words, Big Data and the ability to process it represented the missing link in obtaining success in the early days. It is the distributed parallel processing nature of Big Data technologies which has allowed analysis of these large data volumes to occur. Within AI, its most common oft-touted success has been in the area of image recognition. As stated earlier, dramatic improvements in performance have resulted in many different business applications with image and text and voice recognition being the most common. In image recognition, convolutional neural nets represent the deep learning approach while the deep learning approach for text and voice is recurrent neural nets.

In layman terms, convolutional neural nets utilizes the neural net structure by commencing its analysis on the edges around an image and moves deeper and deeper into the image until the object itself is identified. Meanwhile recurrent neural nets utilizes its multiple hidden layer structure by focussing its analysis in a more sequential or time series approach toward any text or voice data. Once again, the use of multiple hidden layers complemented with the use of many nodes is the key to increased model accuracy and performance. But this cannot happen without data and under these scenarios large amounts of data or Big Data is a fundamental requirement.

Another key driver of AI success is the signal to noise ratio. In other words, is the pattern or trend we are trying to capture in the data relatively strong when comparing it to the data that is truly random? My article on “Is predictive analytics for marketers really that accurate?” published in the Journal of Marketing Analytics goes into much greater detail on this topic [9].

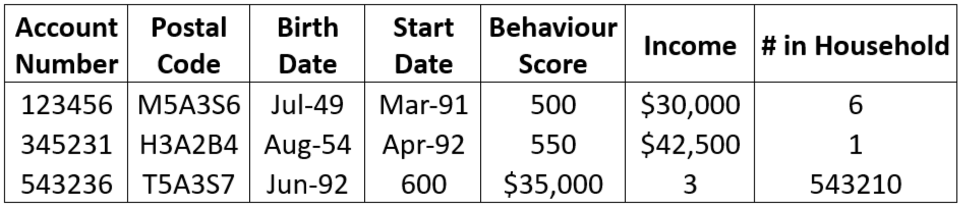
1.3. **The Data Science Process-Problem Identification.** The popularity of AI has created its own persona in that many organizations with lots of data seem to accept the notion of “I can apply AI and obtain my solution”. But this couldn’t be further from the truth." AI represents just one technique within an entire process of developing a solution from raw data. This process, the "data science" process is a four step process involving these four steps of which AI can represent one key component. The book [10] "Data Mining for Managers: How to use data (big and small) to solve business problems" which was authored by me and published by Palgrave Macmillan goes into much greater detail on this subject. Listed below are the 4 steps in the data science process

- Problem Identification
- Creation of the Analytical File
- Application of the Right Data Mining Tools in Developing the Solution
- Deployment and Measurement of Solution.

Let's explore each of these steps in more detail. Problem-identification is arguably the most important component or stage in the process. The best data analytical techniques amount to nothing if they are not directed towards solving the right problem. Many data scientists will indicate that this stage really has nothing to do with data science and more with the business given their objectives and strategies. The data scientist will argue that the business needs to define the business challenge or problem with the data scientist reacting to the needs of the business. This is flawed thinking as the data scientist alongside the business stakeholder needs to collaborate in identifying what are the real challenges or issues of the organization.

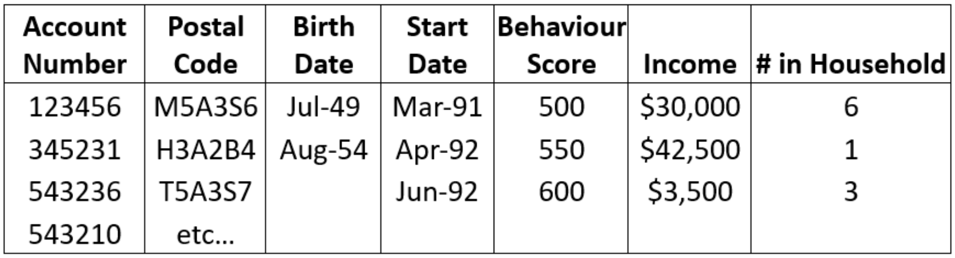
This first stage would involve the sharing of information from the business stakeholder to the data scientist. This might involve reports and analyses that have been conducted on various key business initiatives. Presentations and documents related to business strategies and goals would also be shared with the data scientist. At the same time, interviews and meetings would be set up between the business stakeholder and the data scientist in order to identify other key challenges and issues that may not emerge from the sharing of reports, analyses, and presentations. The underlying objective of all this collaboration is for both the data scientist and the business stakeholder to take ownership over this step. With this approach, the data scientist can become more proactive in identifying business problems rather than being reactive to the business stakeholder with a "you tell me what to do" type attitude.

There are many examples of how lack of collaboration have yielded sub-optimal results. Perhaps the best example is in the area of customer retention. Most organizations will without a doubt express the need to retain customers. In data science, one of the best approaches to this problem is to build a model to predict those customers who are most likely to defect or to become non-engaged. Marketing is then able to deploy more resources towards this high risk group in an attempt to save more customers. The data scientist upon hearing this need would then go ahead and build a retention model to target high risk defectors. But is this the real problem? In any customer file of any organization, there are going to be groups of customers who already exhibit low activity or engagement. But do we want to really consider these type of customers from this group as being our high risk target group? Instead, the right approach would have been to build a predictive model that targets high risk defectors who are high value or highly engaged. A more collaborative approach as described above would have identified the better approach thereby not wasting time and resources in pursuit of a model that targets the wrong group of customers. Note that no AI routine would have identified this issue.



Account Number	Postal Code	Birth Date	Start Date	Behaviour Score	Income	# in Household
123456	M5A3S6	Jul-49	Mar-91	500	$30,000	6
345231	H3A2B4	Aug-54	Apr-92	550	$42,500	1
543236	T5A3S7	Jun-92	600	$35,000	3	543210

Figure 11. Sample of 3 records



Account Number	Postal Code	Birth Date	Start Date	Behaviour Score	Income	# in Household
123456	M5A3S6	Jul-49	Mar-91	500	$30,000	6
345231	H3A2B4	Aug-54	Apr-92	550	$42,500	1
543236	T5A3S7		Jun-92	600	$3,500	3
543210	etc...					

Figure 12. Sample of 3 records-Fixed

1.4. **The Data Science Process-Creation of the Analytical File.** In many cases, the second stage of creation of the analytical file is the most time consuming component of the data science process. The objective here is to transform raw data into meaningful inputs which can be used in whatever solution we are trying to build such as a predictive model, reports, etc. The first step is for data to be extracted and more importantly the right data which is necessary to solve the appropriate business problem. Presuming that the right data has been extracted, the data audit process is conducted on all this extracted data. This process comprises three steps:

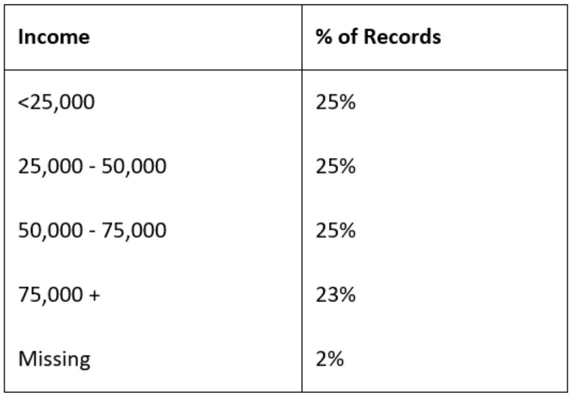
- Random sample of records
- Frequency Distributions
- Data Diagnostics

Once the data is extracted, these three steps are conducted on each file of data. At our organization, we have automated this process. The first report is a random sample of 100 records which actually depicts what the data looks like where the record is a row and the fields or source variables are columns. See Figure 11 for example.

From the above, you can observe that some of the fields in record 3 have gone awry. This is due to the fact that there are missing values in the birthdate field which were not picked up correctly by the software. Once this is corrected, this table (Figure 12) would appear as follows:

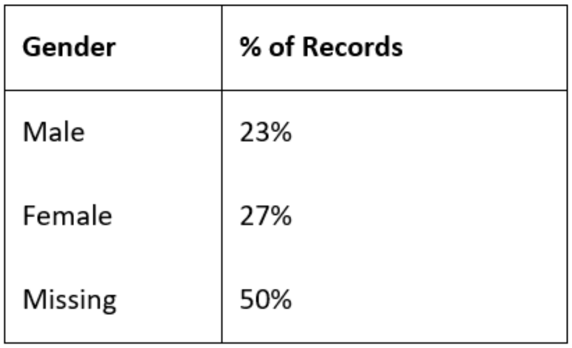
The second report represent frequency distributions which depict how the values of a given field or variable are distributed amongst its records. Both numeric as well as character type data are explored in these frequency distribution reports. See examples of these reports (Figure 13 and Figure 14).

The third report, which we refer to as the data diagnostics report, explores in depth the content of each field or variable. It displays the number of missing values, the number of unique values, the format of each variable alongside some



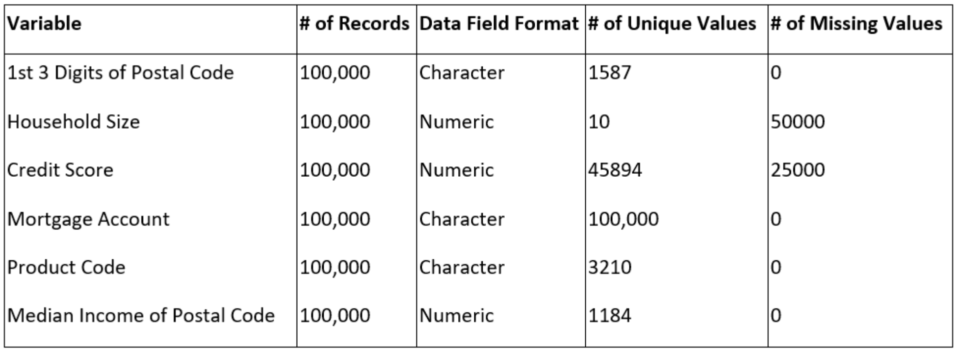
Income	% of Records
<25,000	25%
25,000 - 50,000	25%
50,000 - 75,000	25%
75,000 +	23%
Missing	2%

Figure 13. Frequency Distribution of Numeric Variable



Gender	% of Records
Male	23%
Female	27%
Missing	50%

Figure 14. Frequency Distribution of Character Variable



Variable	# of Records	Data Field Format	# of Unique Values	# of Missing Values
1st 3 Digits of Postal Code	100,000	Character	1587	0
Household Size	100,000	Numeric	10	50000
Credit Score	100,000	Numeric	45894	25000
Mortgage Account	100,000	Character	100,000	0
Product Code	100,000	Character	3210	0
Median Income of Postal Code	100,000	Numeric	1184	0

Figure 15. Example of Data Diagnostics

basic statistics of each field such as the mean, median, and standard deviation for variables that are numeric. See Figure 15 as an example of such a report but for sake of brevity, we did not include columns related to the basic statistics.

The purpose of these data audit reports is to convey some initial insight concerning the raw data. This insight would be used to determine what fields would be

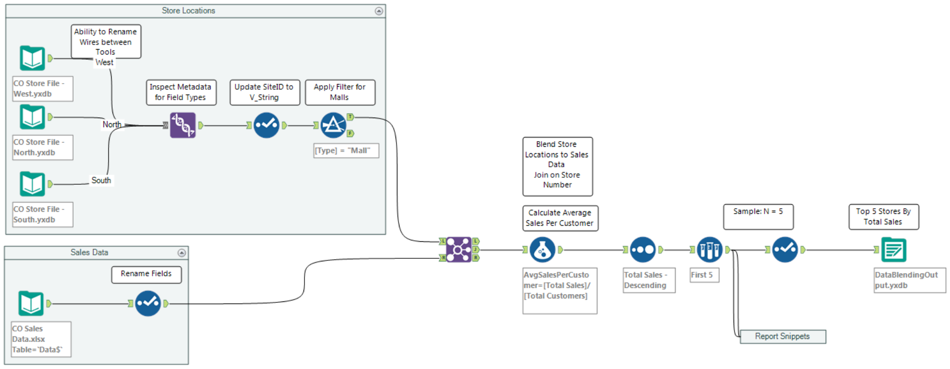

Store Locations
Ability to Rename Wires between Tools
West
CO Store File - West.yxdb
North
CO Store File - North.yxdb
South
CO Store File - South.yxdb
Inspect Metadata for Field Types
Update SiteID to V_String
Apply Filter for Malls
T
F
[Type] = "Mall"
Sales Data
Rename Fields
CO Sales Data.xlsx Table=`Data$`
L
R
L
J
R
Blend Store Locations to Sales Data Join on Store Number
Calculate Average Sales Per Customer
AvgSalesPerCustomer=[Total Sales]/[Total Customers]
Total Sales - Descending
First 5
Sample: N = 5
Report Snippets
Top 5 Stores By Total Sales
DataBlendingOutput.yxdb

Figure 16. Example of Alteryx Software

useful vs. not useful for future analysis. You will remember that these automated data audit reports are conducted on each file separately.

Another component to this second step is how to link or join all the files together in creating one analytical file. This process of joining files together is not just the simple linkage between files but also the derivation of new variables. In fact, variable derivation is arguably the most important component of the data scientist's work. In a data science exercise to build a predictive model, it is not uncommon for well over 90% of the variables to be comprised of derived variables as opposed to raw source variables. Much of this work was historically done by practitioners who knew how to program or code. In the early days, either one coded in SPSS or the more common data science software platform of SAS. More lately, coding languages such as R or Python are gaining great traction due to their open source nature of their platforms. However, the ability to conduct more advanced analytics was still largely confined to practitioners who knew how to code. But this has changed with the advent of new software which empowers more practitioners but does not require programming or coding skills. Instead, a deep understanding of how to work data to solve a business problem is required. But rather than code for this, the practitioner uses drop-down GUI icons to invoke the functions or tasks that needs to be used to create the analytical file. See Figure 16 as one example of one such software program from Alteryx.

In the above schematic, we see a flow chart which depicts how raw source data is transformed to yield the top 5 stores for this retailer as well as a report. From the above, icons, instead of programming code are used to work the data in order to deliver the solution. The end result with software tools such as Alteryx as well as other similar tools is now a broadening of the community to non-programmers who are now empowered to create the analytical file.

Much of the work done here either through automated tools or through coding is done to create variables or what is referred to as feature engineering. Some components of AI research have been devoted to this area of feature engineering. Using the concept of auto encoders, AI can be used to generate the optimum features or variables which would comprise the inputs of a model. This research is not new and has been ongoing for at least the last decade. This research has been primarily focussed in the area of image and text recognition where the raw data is more proscribed. For example, AI's ability to detect an edge from the image pixel is a form of feature engineering. Recognizing that feature engineering represents a very significant portion of the data scientist's work, research in this area continues as to expand in an attempt to provide software that facilitates this very labor-intensive process. The challenge is that other business applications do not lend themselves to the more automated feature engineering routines. Within the area of consumer behaviour, can I take raw transaction behaviour and use automated feature engineering to generate key meaningful inputs. For example, in trying to predict a given consumer behaviour, one key approach is the ability to derive variables or features from the longitudinal nature of the data. Longitudinal information refers to the historical information of the customer and the ability to create variables on how customer information has changed such as change in spend between the current 6 months and the previous 6 months. One can easily create many of these type of variables based on different time periods. For data scientists, the question to answer is: "Can this type of feature engineering be truly automated through software?" With all the tools that are now available in the analytics space, data scientists need to recognize the benefits of these tools while at the same time recognizing their limitations particularly in the area of feature engineering where the end objective is a more effective approach in creating the analytical file.

1.5. **The Data Science Process-Application of the Right Data Mining Tools in Developing the Solution.** The third step involves the application of the right data mining tools in developing the solution. This step is where we determine whether we are conducting analysis or building reports or in fact building a model. If we are building reports or conducting analysis, how do we lay out the analytical file such that it can be meaningfully used in one of the more visual type platforms in order to derive key insights? If we are building models, a variety of machine-learning techniques can be utilized. Artificial Intelligence or deep learning is just one option that can be considered. In much of my world of trying to predict customer behaviour, we do not necessarily look at error rates but rather how well AI delivers better rank ordering of names in terms of the desired modelled behaviour. This perspective explores the use of a model which has been developed off a certain group of names but is then deployed against a holdout or validation group. Within the validation group, names are scored and placed into decile groups with decile 1 representing the top scored names and decile 10 the lowest scored names. See Figure 17 which demonstrates the results of a response model.

In this example above, the better performing model is the logistic regression model. If we are going to employ AI techniques such as deep learning, does deep learning providing a steeper curve. For many of us in the customer analytics sector, we often observe that traditional techniques perform just as well if not better than deep learning neural net techniques. In other words, AI does not improve the slope of the curve.

The other issue with using deep learning techniques in customer analytics and predictive modelling is explainability which is in effect a very significant barrier. Most organizations want to gain a better understanding of what is actually driving a given consumer behaviour such as customer response or credit risk or attrition risk. Data science practitioners have historically always been able to explain the key components within a consumer behaviour model and its relationship to the desired consumer behaviour. For example, traditional machine learning algorithms such as multiple regression or logistic regression can look at the statistical measures such as the partial R^2 of a given model input variable and compare it to the total R^2

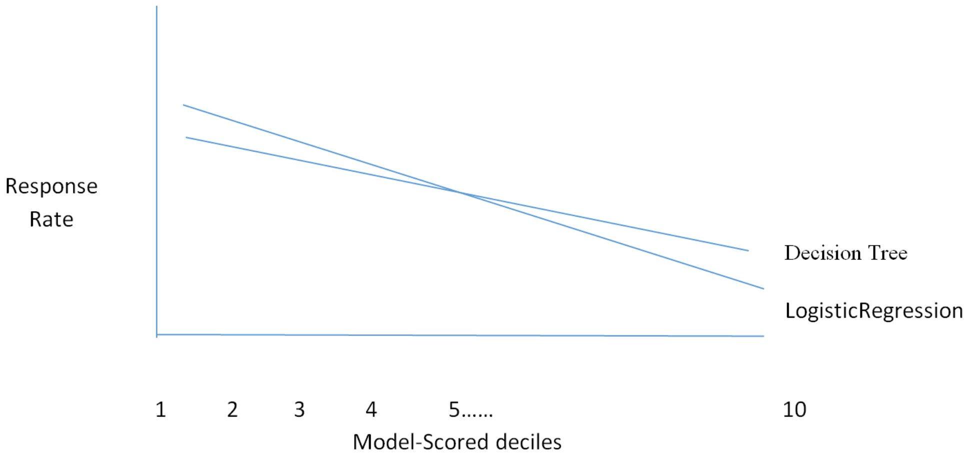
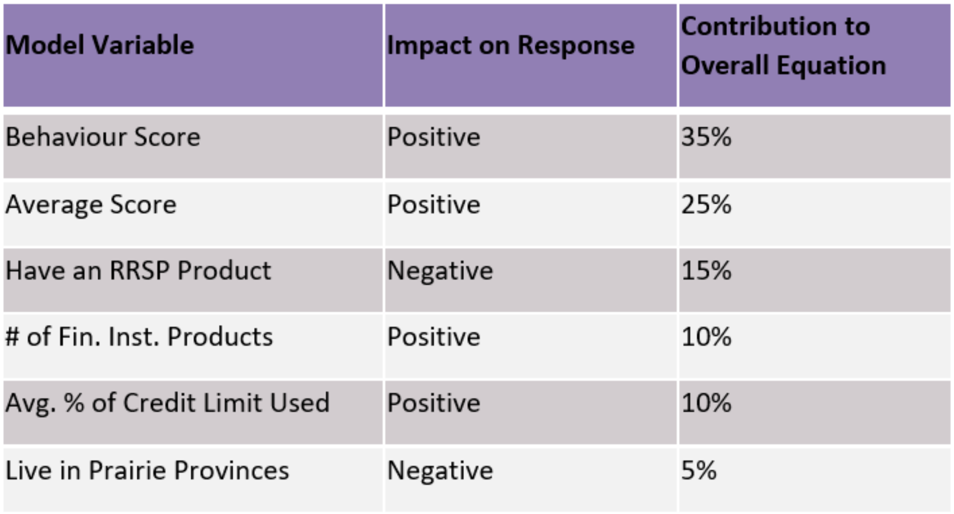


Figure 17. Example of Gains/Decile Table



Model Variable	Impact on Response	Contribution to Overall Equation
Behaviour Score	Positive	35%
Average Score	Positive	25%
Have an RRSP Product	Negative	15%
# of Fin. Inst. Products	Positive	10%
Avg. % of Credit Limit Used	Positive	10%
Live in Prairie Provinces	Negative	5%

Figure 18. Example of Final Model Variable Contribution Report

of the entire model in order to obtain the explainability of that variable. Figure 18 is a table of a predictive model which is attempting to maximize the likelihood of a bank customer responding to a credit card product. This table depicts the contribution of each model variable based on the partial R^2 relative to the total R^2. As well, it also indicates the impact (negative or positive) on the desired modelled behaviour of response.

Figure 18 clearly demonstrates clearly demonstrates which variables are the stronger vs the weaker variables within the overall model while at the same time exhibiting whether they exhibit a negative or positive response to the desired behaviour of response.

The use of a deep learning model poses explainability challenges due to the fact that neural nets seek to maximize a solution regardless of the distribution of the data(linear vs. non-linear). Much of the current research in using AI techniques, though, is in this area of explainability and there are some interesting approaches which can at least convey the importance or contribution but not necessarily the sign or impact. We are currently exploring a few options ourselves at this point in time.

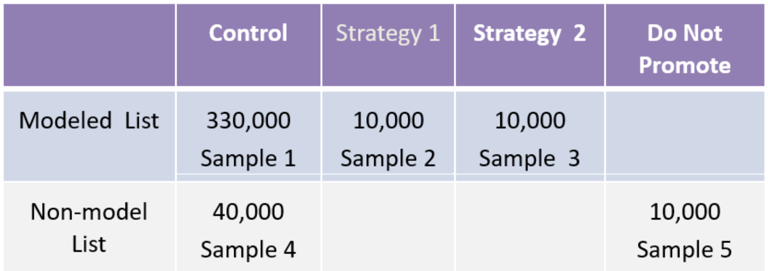
1.6. **The Data Science Process - Deployment and Measurement of Solution.** The last step in the data science process represents the implementation and measurement of the solution. A given data science exercise can yield an excellent solution but fail because it is not deployed in the right manner. Even it is deployed in the right manner, it may still be deemed to be unsuccessful because the measurement approach in evaluation of the solution is flawed. You will note that there is no AI influence in this process. In the deployment of a solution, an entire Q/C process needs to be established. For example, key information metrics would need to be assessed in order to see if there are major changes with the data upon the deployment of the solution. A good example of this is if the average response model score has declined by 50% between the current deployment and the last deployment of the solution. This would need to be examined in order to identify the reason for this dramatic change. A more practical example of how improper deployment can produce deleterious results was a property claims model that we developed for a client where we were trying to predict the likelihood of a claim. The model was developed and the performance of the model on the validation or holdout group was excellent. When the model was deployed, the initial results were abysmal. Upon further investigation, it came to our attention that the properties included both homeowners and renters when in fact the model was developed for homeowners only. Once apartments were excluded, the model validated quite well. Another key component is to determine how to effectively measure a given solution. Once again, collaboration is key between the business stakeholders and the data scientist. The two underlying questions are:

- What Kind of learning do we hope to attain
- What are the key business metrics that we hope to optimize

Depending on these two initiatives, the data scientist and the business stakeholder need to design a measurement matrix that addresses both these above questions. Figure 19 is one simple example of a measurement template that was designed for an RRSP marketing program to its existing customers where we were trying to evaluate the following:

- Did the overall campaign work?
- Did the model work
- Which strategy was best

The assessment of whether the campaign worked vs not worked would look at results between sample 4 and sample 5. A quick evaluation of whether the model worked vs. not worked would look at the comparison of results between sample 1 and sample 4. A deeper look at model performance would look at sample 4 and how modelled names ranked by decile actually perform within each decile. Performance by strategy would look at results of sample 1 vs. sample 2 vs. sample 3. As you can see, the matrix is designed not only to achieve the learning but also to maximize overall campaign results which is why most of the names overall are in sample 1 where the customer names have been targeted (modelled) and are in a known and



	Control	Strategy 1	Strategy 2	Do Not Promote
Modeled List	330,000 Sample 1	10,000 Sample 2	10,000 Sample 3	
Non-model List	40,000 Sample 4			10,000 Sample 5

FIGURE 19. Example of Final Model Variable Contribution Report

tried strategy (control). Throughout this process, there is no mention of AI as being utilized to help in this process.

1.7. **Ongoing Business Applications.** Much of the discussion of this chapter has been devoted to the 4 step data science process because many people confuse AI as being the data science process which could not be further from the truth. There is no question that AI or deep learning techniques need to be considered but within the realm of the 4 step data science process and not a process unto itself. Certainly, AI or deep learning is now becoming mainstream in many business applications. In the area of insurance, the ability to better classify images of a claim into categories can be used to better predict the amount of what the claim will cost to the insurer. In health, classification of x-ray image data can be used as inputs to build better predictive diagnostic tools.

The use of chatbots in enhancing customer service has been a customer service application for the last decade. Yet, in the past, these chatbots were simply defined by business rules which represented a number of options and which were then applied dependant on what the customer had last said. With AI, the machine learns what the customer has said and then predicts what the best response should be rather than be limited to pre-defined business-rule options.

This concept of the machine learning from previous text can be applied to both health and law. Think of the time savings if a machine can learn from hundreds of previous case studies in order to present a number of concise options for the lawyer. This capability is not a replacement but rather a tool for the lawyer that provides more focussed areas of investigation. The same principle would apply in the area of health where machines could review thousands of case histories and once again provide the health practitioner with areas that he or she might want to review in more detail.

1.8. **Barriers to AI Deployment and Future Areas of Research.** In the area of consumer behaviour which is my area of domain expertise, predictive models have been utilized by certain organizations for decades particularly in the area of financial services. Both marketing behaviour and risk behaviour models are generated by teams of data scientists within these areas. Yet, in many cases, the traditional machine learning algorithms such as decision trees, logistic regression, SVM (service vector machines) are the choice rather than the deep learning algorithms of AI? Why? For these organizations, there is no lack of Big Data which as stated above is the key to successful model performance within the deep learning AI ecosystem. So then why the reluctance of using deep learning AI solutions within this area?

Two barriers exist with the first being what is often referred to as the signal to noise ratio while the second barrier is due to explainability. The first barrier of signal to noise ratio as described earlier essentially describes the ability of what data a predictive model can truly predict(explained variation or signal) versus what data is really random(unexplained variation or noise). In my many years of building models in the consumer behaviour space, this signal to noise ratio is quite small, hence the ability of the more traditional and less complex machine learning techniques to be as accurate if not more accurate than the deep learning AI techniques. Yet, we do not observe this condition within the ability to classify images or to predict text. In these non-human scenarios, the signal to noise ratio is large and when combined with a very large amount of data, deep learning techniques will surpass the more traditional machine learning type techniques.

Explainability as discussed earlier in the chapter is perhaps the bigger reason for their non-use in many business applications even outside the area of consumer behaviour. Businesses are uncomfortable with black box solutions and need to at least have a basic understanding of the key inputs that are driving the model algorithm. Current research has identified a number of different approaches which are still at an immature stage before being more fully embraced by the data science community.

Another interesting area of research is the use of AI in building more powerful solutions within a small data environment. Research to date has expounded upon AI's significant benefits in predicting outcomes but within a big data environment. Yet, many organizations operate within the typical "small" data environment. Small data can be a very subjective term but datasets under 1mm records and with under 200 columns have been in my experience labelled as small datasets. Certainly, new approaches to improving AI models in a small data environment would be a huge boon to data scientists.

2. **Conclusions.**

In a way, we can say that AI is the child of Big Data because without the Big Data disruptive changes in data processing, AI does not become a game changer. Both AI and Big Data are changing the landscape of the economy and society in general. Business magazines such as the Economist recognize that data is the new oil within the economy. It is still too early to observe what this overall impact will be but it will be transformative. Certain industries and jobs will become extinct while start-ups will emerge that will create new industries and jobs. These scenarios are already evident today. In certain major cities of Canada, Toronto in particular, innovation hubs are being created through both public and private financing. The thought process here is to explore and tinker with as many new ideas as possible recognizing that there will be many failures but it will be those few successes that will be the transformative engines of our new economy. Fail fast and learn quickly is the mantra of these new hubs. A new invigorated sense of entrepreneurialism is emerging today and it is this culture of potential opportunity that will be the foundation in building a strong economy which will be the genesis of new jobs and careers that otherwise do not exist today. Big Data and AI will be the focal point for much of this emergent activity.

REFERENCES

[1] Figure.1: The 5 V's of big data, Environics Analytics: Best Practices and Considerations in Big Data Analytics, June, 2018.

[2] Figure.2: Moore's Law, https://www.google.ca/search?hl=en&tbm=isch&source=hp&biw=1366&bih=651&ei=wd3pWuPdMqqPjwSUr4SQDQ&q=exponential+growth+in+computing+power&oq=growth+in+computing+power&gs_l=img.1.1.0j0i5i30k1.4574.14021.0.16389.28.27.0.1.0.0.182.2657.16j10.26.0....0...1ac.1.64.img..1.25.2467.0..0i24k1j0i8i30k1.0.eDlGB4j2AdI#imgrc=jhm-BdlhnmB2HM:.

[3] Figure.3: Columnar file formats, https://www.google.ca/search?hl=en&tbm=isch&source=hp&biw=1366&bih=651&ei=wd3pWuPdMqqPjwSUr4SQDQ&q=exponential+growth+in+computing+power&oq=growth+in+computing+power&gs_l=img.1.1.0j0i5i30k1.4574.14021.0.16389.28.27.0.1.0.0.182.2657.16j10.26.0....0...1ac.1.64.img..1.25.2467.0..0i24k1j0i8i30k1.0.eDlGB4j2AdI#imgrc=jhm-BdlhnmB2HM:.

[4] Index compression, https://nlp.stanford.edu/IR-book/html/htmledition/index-compression-1.html.

[5] Figure.6-Sequential vs. parallel data processing, https://www.google.ca/search?biw=1607&bih=678&tbm=isch&sa=1&ei=UVPwWu_uGoeYjwSkqrjwBA&q=sequential+db+processing&oq=sequential+db+processing&gs_l=img.3...0.0.0.123836.0.0.0.0.0.0.0.0.0..0.0....0...1c..64.img..0.0.0....0.jkNEKg1fCWO#imgdii=kH8ag2orN-LWNM:&imgrc=pBOBcUMsqlXNGM:&spf=1525699534175.

[6] Turn to in-memory processing when performance matters, https://searchdatacenter.techtarget.com/feature/Turn-to-in-memory-processing-when-performance-matters.

[7] Figure.8: Schematic of weights within neural net structure, https://www.google.ca/search?hl=en&tbm=isch&source=hp&biw=1366&bih=651&ei=bpvwWv2FM82O5wLqzLigCA&q=neural+net+simple+network&oq=neural+net+simple+network&gs_l=img.3...1065.21853.0.22452.38.24.0.14.14.0.120.1836.21j2.23.0....0...1ac.1.64.img..1.13.1052.0..0j0i24k1j0i10i24k1j0i10k1j0i7i30k1.0.nu7gREvNHkk#imgrc=13gO7BFbOGYZqM:.

[8] Figure. 9-Examples of some optimization algorithms, https://www.google.ca/search?hl=en&tbm=isch&q=logistic+function&chips=q:logistic+function,g_5:logistical&sa=X&ved=0ahUKEwjw-KD5oPTaAhWkpFkKHSxSDJwQ4lYIMCgA&biw=1366&bih=651&dpr=1#imgrc=oAHIGiD5uTjw2M: https://www.google.ca/search?hl=en&tbm=isch&q=tan+function+graph&chips=q:tan+function+graph,g_1:tangent,online_chips:cos+tan&sa=X&ved=0ahUKEwjK-IGYovTaAhVQwlkKHUBnC0cQ4lYIKygC&biw=1366&bih=651&dpr=1#imgrc=gWnErav-9CIbGM:.

[9] "Is predictive analytics for marketers really that accurate?", *Journal of Marketing Analytics*, May, 2013, https://link.springer.com/article/10.1057/jma.2013.8.

[10] "Data Mining for Managers: How to use data (big and small) to solve business problems", by Palgrave Macmillan, Oct, 2014.

E-mail address: richard.boire@environicsanalytics.com