Why machine learning is crucial to effective utilization of big data

Managing and gaining the most value from big data

July, 2018

Arthur D Little

Cutter Consortium
Access to the Experts

Content

Authors:

Greg Smith
Partner, Digital Problem Solving,
London
smith.greg@adlittle.com

Andreas Macek
Consultant, Digital Problem Solving,
London
macek.andreas@adlittle.com

Michael Papadopoulos
Chief archictect, Digital Problem Solving,
London
papadopoulos.michael@adlittle.com

Andrea Solda
Business Analyst, Digital Problem Solving,
London
solda.andrea@adlittle.com

This article originally appeared in *Cutter Business Technology Journal*, Vol. 30, No. 10/11, 2017
https://www.cutter.com/article/why-machine-learning-crucial-effective-utilization-big-data-497821. Republished with permission.

Executive summary

This article explores the challenge of managing and gaining the most value from big data. We highlight the increased opportunity associated with larger data sets, while illustrating the limitations of current methods and human intellect across the 4 Vs of big data (volume, velocity, variety, and veracity), ultimately resulting in lost value — the fifth V. We further show how organizations can use machine learning (ML) to address these limitations and realize the full value from big data. Finally, we highlight how cutting-edge companies employ ML to obtain greater value.

1. Challenges in big data applications

Big data is now a commonplace term that represents exponentially growing data. New terms, such as "data exhaust" for all the data being generated by our daily activities and "data lake" for new ways of storing data in its natural format, have also entered our lexicon. In fact, an entire field, data science, has been reinvented and rediscovered just to try to handle this explosive flow of information.

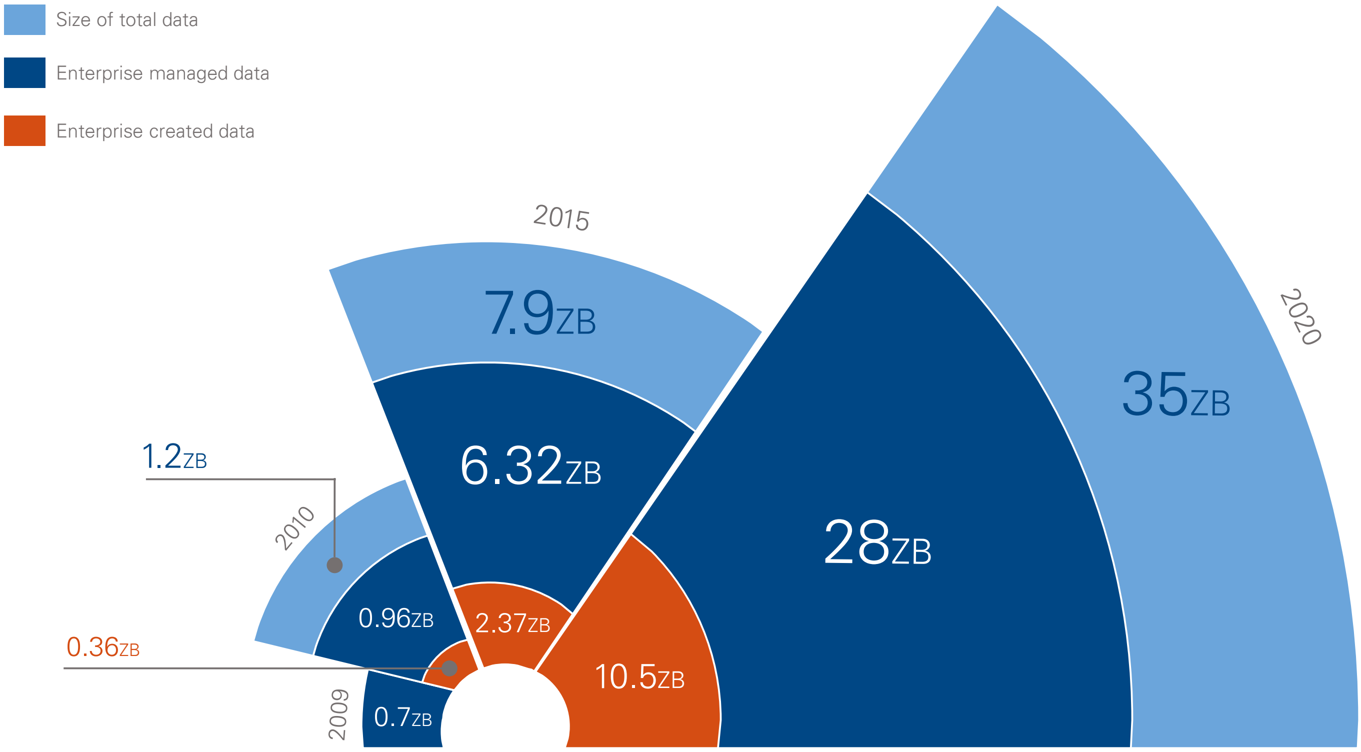
Nevertheless, big data has huge untapped value. The amount of data created in the world is only growing (see Figure 1); our ability to extract value and insight from it should naturally grow with it. Its hidden patterns and clusters will allow us to analyze data sets on a possible cure for tuberculosis or crime prevention, as well as derive more mundane yet valuable insights, such as predicting the value of apartments in urban suburbia or what customers may want to buy and how much they are willing to pay.

The "data-insight-action" chain has suffered from a distinct lack of validation, especially in the context of measurable value

The reason big data has failed to deliver on these promises in recent years is quite simple: enterprises apply the traditional data analytics techniques they have long used for small data, but now do so with the help of expensive tools with impressive-sounding names like Hadoop and Spark. While enterprises can now analyze data at scale and produce impressive and better-looking interactive dashboards and graphs, they are essentially deriving the same insights as they did a decade ago based on the same type of data and techniques used then. They are effectively unable to consume in a valuable way the plethora of new data available.

Traditional analytics techniques tend to drive analytics functions to be effectively classic reporting functions, regardless of the

Figure 1: Data growth rate is exponential, with data growth in 2017 exceeding the total amount of data produced throughout human history



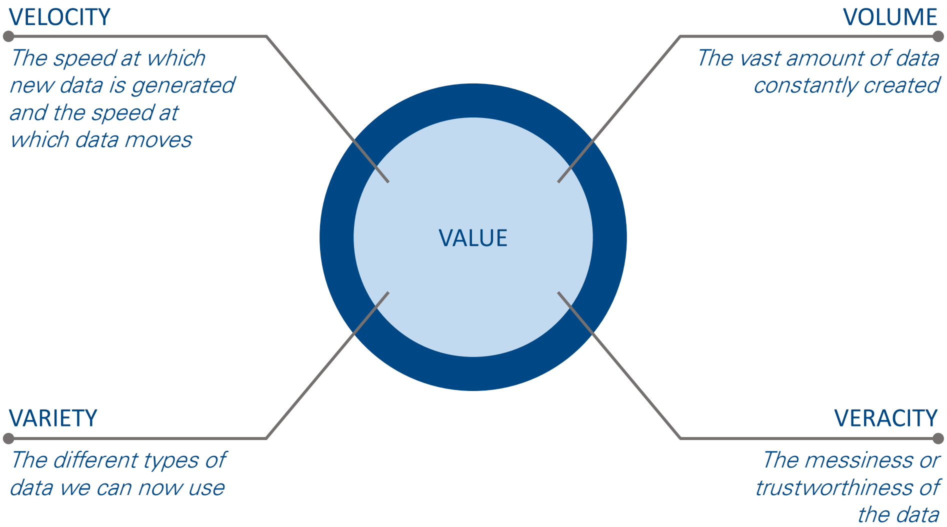
Source: Marr.

However, in recent years the promises of big data have been frustrating. Enterprises wanting to extract hidden value from petabytes of raw data have not managed to do so successfully.

volume of data or the tools used. Thus, instead of analyzing business intelligence to produce actionable insights, the enterprise produces simplified reports containing sums, counts, averages, and medians, with the occasional SQL query added to the mix.

The problem with traditional analytics and reporting functions — even when using shiny new tech — is that they rely on a human to specify and direct the questions asked of data. Traditionally, a human user has specified the questions to be asked to run known calculations. This human specification of a problem is a limitation in the big data space due to the very nature of big data. Indeed, the 4 Vs (see Figure 2) represent that the relevant questions may be unknown. The human intellect struggles to see the correlations hidden in big data that help to specify a problem and the questions to be asked, and is even less capable of creating the intellectual model necessary to extract the fifth, and most important, V: value.

Figure 2: The 4 Vs of big data (volume, velocity, variety, and veracity) produce the most important V: value.



Source: Arthur D. Little

Arthur D Little

2. The machine learning advantage

Machine learning is a specialized section in the field of artificial intelligence (AI) that allows computers to learn on their own. Computers are provided a set of rules, rather than explicit programming instructions on what to do, and can then self-train and develop a solution by themselves. With ML techniques, computers iteratively learn from data and also find potential hidden insights in data.

Table 1 contrasts the limitations of the human intellect with machine learning, which is entirely data-driven (so no bias) and runs at machine scale. ML does not rely on human intelligence and direction to model big and varied data. Unlike humans, machine learn¬ing improves when working with growing data sets. The more data is fed into an ML system, the more the system will learn, resulting in the production of higher-quality insights. (A caveat must be included here about the existence of examples of applying good models to bad data, resulting in bad outputs. Appropriate analysis of the data is required to ensure its quality.)

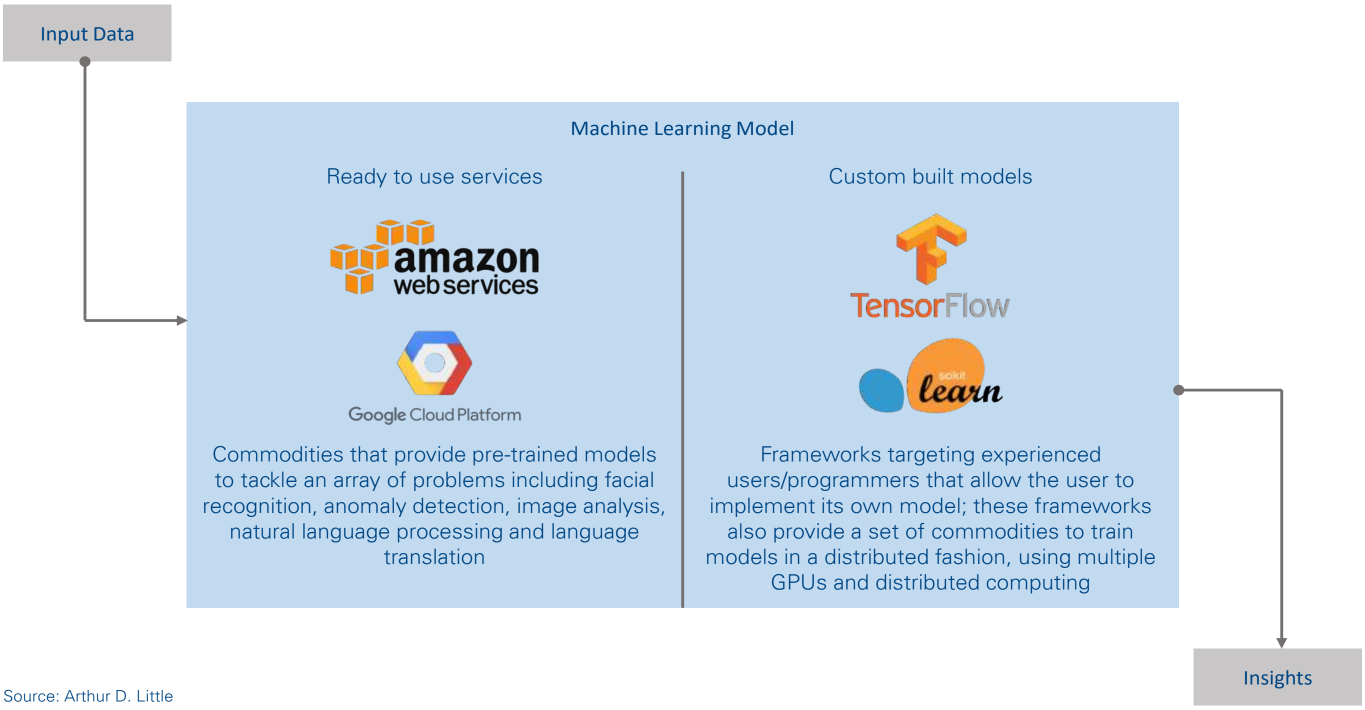
Machine learning is thus ideal for extracting hidden value from big data. ML techniques are commoditized today and, as a result, they are easily accessible with a low cost of entry (see Figure 3). All the big cloud providers (i.e., Amazon Web Services, Google, Azure) have their own ML platforms and software as a service offerings of mature open source AI frameworks (e.g., MXNet and TensorFlow).

Table 1: Human intellect vs. ML models.

4Vs	Human Intellect	Machine Learning
Volume	According to a recent study by the Massachusetts Institute of Technology (MIT),[1] the average processing speed in humans is around 45 bits per second, with skilled individuals going up to 60 bits per second. That means that if a person were to spend 24 hours a day just reading (without ever stopping, not even to eat or drink or sleep), that person would process slightly above 5 MB of data. It would take such a person around 536 days to go through a 1TB data set – and at the end of the process, the person would very likely not remember the entire data set.	ML has no issue with volume; in fact, the more data you have, the higher-quality results you expect. In terms of capability to process volume, the machine is constrained only by the size of the machine (a modern laptop can likely read 1TB data in a few seconds).
Velocity	Data changes too quickly for a human to understand the consequences and react in a positive manner. For example, the New York Stock Exchange typically produces 1 TB of trade information during each trading session. In simplistic terms, 1 TB is equal to a couple million decent-sized books. Humans simply cannot absorb the data fast enough and, even more important, are unable to sift through the data to find the valuable actionable insight	ML operates at machine speeds: a successfully trained model can react to input data at speeds impossible for a human. Recent advances in neural nets allow for immediate reaction to inputs "close to edge"– where the event happens. Reaction times of an efficient model can be measured in tens of milliseconds.
Variety	Most enterprises now have data from multiple sources (e.g., databases, real-time social media streams, IoT reporting) and in various formats (e.g., structured, unstructured). The average human intellect will almost certainly be unable to construct and hold in memory a data model the considers every single possible correlation and relationship among data from multiple sources and in multiple formats. The variety of data needs to be simplified for human consumption either via description, aggregation, or summarization, but simplification loses the potential of big data and makes finding hidden insights unlikely	Variety is not a problem for machines, as extremely complex models can be built and held in machine memory. ML operates at machine scale with modern machines capable of holding and making use of terabytes of information in milliseconds
Veracity	We cannot be confident that the data we have is clean, usable, and of high quality – and even worse, humans are prone to errors of judgement such as confirmation bias and will tend to attribute more worth to data that confirms biases – even if it means losing out financially	ML has no confirmation (or other) bias when analysing data. Proper algorithms and models will also trend to recognize "bad" data. Significant advances in automated algorithms exist that allow the machine to effectively ignore noise and outliers.

1 "New Measure of Human Brain Processing Speed." *MIT Technology Review*, 25 August 2009 (https://www.technologyreview.com/s/415041/new-measure-of-human-brain-processing-speed/).
Source: Arthur D. Little

Figure 3: Commodity ML platforms.



Source: Arthur D. Little

3. Value from applying ML to big data

A small number of cutting-edge companies have effectively exploited big data's potential. These companies have a strong focus on data gathering, recognize the potential value locked in their data, and effectively use machine learning to convert potential value into real, measurable business value. In this section, we outline examples of three such companies and the way in which they utilize ML in practice.

Domino's Pizza

Domino's Pizza has been able to successfully switch from traditional reporting techniques to machine learning to leverage the value of its data. Data on customer preferences, time orders are placed, and the types of food and beverages ordered are used to tailor the company's promotions. Domino's predictive model is so good it can predict when you are likely thinking about pizza and breadsticks and, in a clear example of data-insight-action, can push out targeted offers at that exact moment. The ML system in place is constantly improving its performance as it analyzes the new data available, helping the company achieve faster growth than any of its competitors. In fact, Domino's growth rate is comparable to that of digital or Web-only companies.

EasyJet

At EasyJet, AI — specifically, ML — has been a valuable tool in the company's kit bag. In the last few years, EasyJet has been using an ML system to optimize food inventory on its planes by predicting what kind of food (and how much) will be needed on a flight. The system looks at a series of features, such as weather, time of year, type of customer, and so on, to make its predictions.

Insight into optimal food inventory is extremely valuable to the company and creates measurable cost savings by avoiding unnecessary expense for food that would be wasted, and by reducing the plane's weight, resulting in less fuel consumption and translating into less money spent on fuel.

PayPal

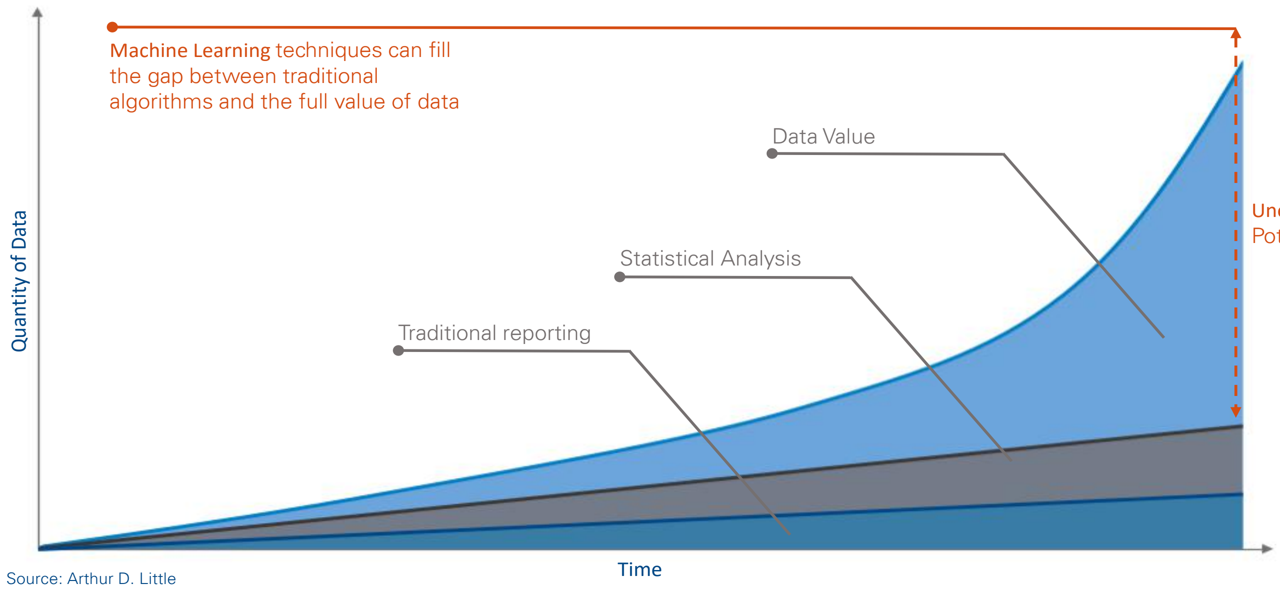
PayPal's fraud detection system uses machine learning to keep the fraud rate at around 0.32% — far better than the 1.32% average that other merchants see. The algorithms mine data from the customer's purchasing history — in addition to reviewing stored patterns of likely fraud — to determine whether a suspect transaction was, for example, the innocent action of someone traveling for work or possibly a fraud attempt. The amount of data that PayPal controls is staggering, with US $354 billion in payments in 2016 from over 4 billion transactions by its more than 170 million unique customers. Given the sensitivity of this data, PayPal has put itself at the forefront of innovation when it comes to security, managing to improve its ML system over time and scaling the system's complexity to match the pace of increase in the company's data. With more data and the power to process it, PayPal's system can get even better.

4. Insights for the executives

The quantity and potential value of data is continuously increasing, but traditional reporting techniques and statistical analysis can mine only so much value from that data due to the very nature of big data. Figure 4 illustrates big data's unexploited potential. Assuming that the value of the data is the full area beneath the graph, traditional reporting techniques and business intelligence — the SQL queries and interactive charts — can manage to extract the value contained only in the area in dark blue at the bottom. Although traditional reporting improves over time, as its capacity to handle data and the speed of executing reports increases, a huge amount of potential value is lost. Statistical analysis and automated machine-driven decision making can also help unlock the value of data, but again, those techniques are limited and lose much potential value as they do not derive insights. The promise of machine learning is what lies in the light blue area at the top of the graph — and in the fact that there is so much untapped, unmined potential in the data that we create. Only with ML algorithms can you step up to the challenge and make better sense of and extract substantially greater value from your data.

While we, as humans, can process and draw insight from data sets, due to the rapid and continuous growth of big data, the limitations of human intellect stop us short of realizing anything near the full value and potential of our data sets, especially as these increase in size. Thus, organizations must look to machine learning to capitalize on the full value of data to achieve competitive advantage or, over time, risk falling behind their rivals.

Figure 4: The unexploited potential of big data.



Source: Arthur D. Little

Contacts

If you would like more information or to arrange an informal discussion on the issues raised here and how they affect your business, please contact:

Austria
Karim Taga
taga.karim@adlittle.com

Belgium
Frederik van Oene
vanoene.frederik@adlittle.com

China
Russell Pell
pell.russell@adlittle.com

Czech Republic
Dean Brabec
brabec.dean@adlittle.com

France
Vincent Bamberger
bamberger.vincent@adlittle.com

Germany
Fabian Doemer
doemer.fabian@adlittle.com

India
Srini Srinivasan
srinivasan.srini@adlittle.com

Italy
Saverio Caldani
caldani.saverio@adlittle.com

Japan
Yusuke Harada
harada.yusuke@adlittle.com

Korea
Kevin Lee
lee.kevin@adlittle.com

Latam
Guillem Casahuga
casahuga.guillem@adlittle.com

Middle East
Thomas Kuruvilla
kuruvilla.thomas@adlittle.com

Netherlands
Michael Kolk
kolk.michael@adlittle.com

Norway
Lars Thurmann-Moe
thurmann-moe.lars@adlittle.com

Singapore
Yusuke Harada
harada.yusuke@adlittle.com

Spain
Salman Ali
ali.salman@adlittle.com

Sweden
Petter Kilefors
kilefors.petter@adlittle.com

Switzerland
Fabian Doemer
doemer.fabian@adlittle.com

Turkey
Coskun Baban
baban.coskun@adlittle.com

UK
Greg Smith
smith.greg@adlittle.com

USA
Carl Bate
bate.carl@adlittle.com

Why Machine Learning Is Crucial to Effective Utilization of Big Data – Managing and gaining the most value from big data

Arthur D. Little

Arthur D. Little has been at the forefront of innovation since 1886. We are an acknowledged thought leader in linking strategy, innovation and transformation in technology-intensive and converging industries. We navigate our clients through changing business ecosystems to uncover new growth opportunities. We enable our clients to build innovation capabilities and transform their organizations.

Our consultants have strong practical industry experience combined with excellent knowledge of key trends and dynamics. ADL is present in the most important business centers around the world. We are proud to serve most of the Fortune 1000 companies, in addition to other leading firms and public sector organizations.

For further information please visit **www.adlittle.com** or **www.adl.com**.

www.adl.com/MachineLearning