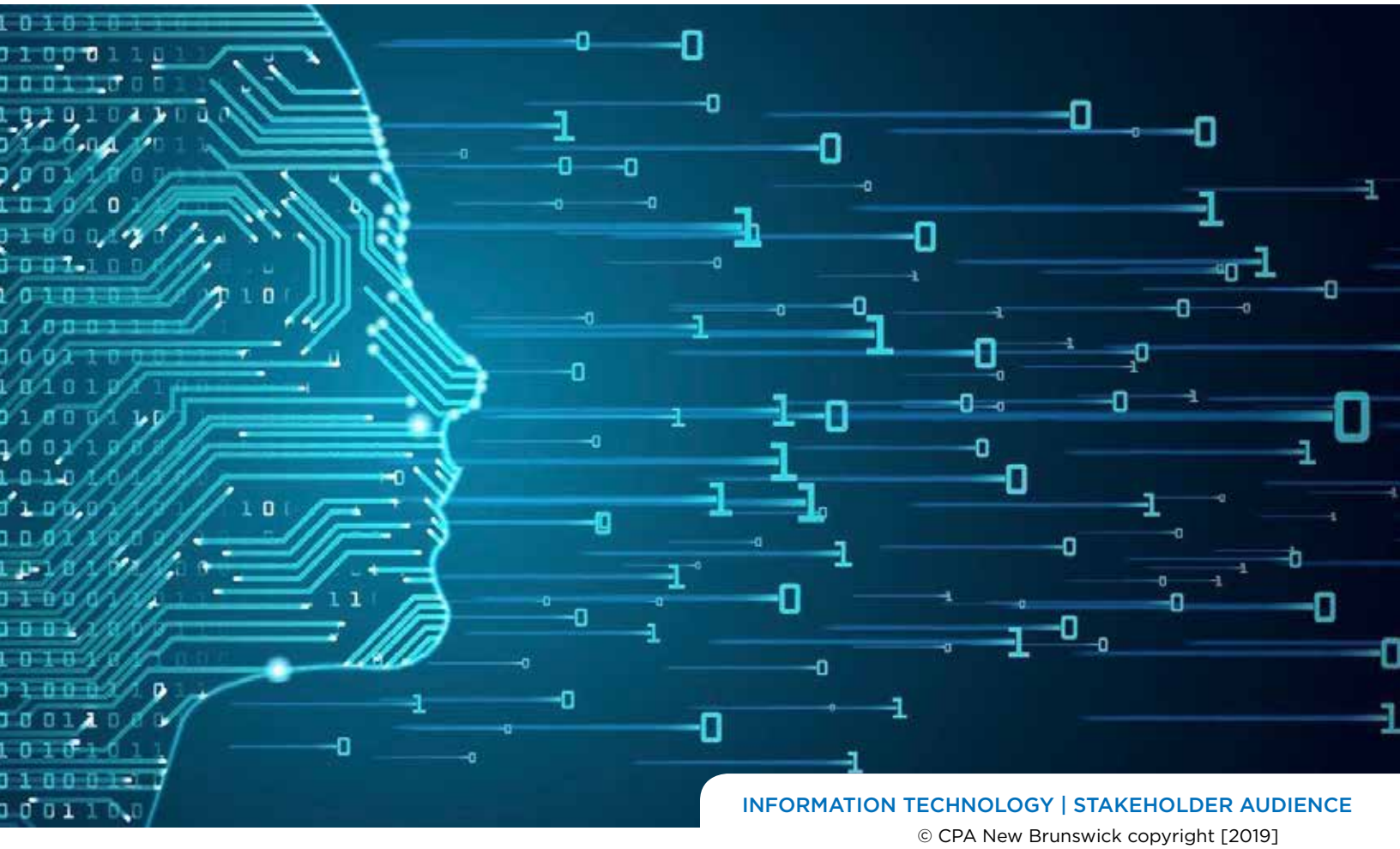
CHARTERED PROFESSIONAL ACCOUNTANTS NEW BRUNSWICK

# Intro to Machine Learning

## A CPA NEW BRUNSWICK MONOGRAPH

Kamalesh Gosalia, Ph.D., CFA, CPA, CGA • Rock Lefebvre, MBA, FCIS, FCPA, FCGA



INFORMATION TECHNOLOGY | STAKEHOLDER AUDIENCE

# Table of Contents

# Foreword

***"Software is eating the world."***

Marc Andreessen, Entrepreneur, Investor, Author and Software Engineer

The above quote, incidentally borrowed from co-founder of Netscape, reflects today's new normal – the new reality in which software drives our economy and propels our lives. The world's largest retailer, Amazon, has no physical stores, the largest video service by number of subscribers, Netflix, has no physical outlets or even cable networks, the world's largest recruiting interface, LinkedIn, does not employ recruiters. These are all software companies, just like other similar disrupters such as Google, EBay, PayPal, Uber, and Airbnb to name but a few.

Accountancy is not immune to the disruptive emerging technologies embraced by the world; those reshaping the very characterization of prosperity and striking at the very foundation of the profession and its value proposition. One such technology generally categorized as Machine Learning has particular relevance to CPAs and their clients – generating extensive insights from large data sets.

Accounting and other financial information can be viewed as a large set of structured data – hence highly amenable to machine learning, correspondingly responsive to current accounting and auditing processes, and decidedly impactful on the work that CPAs perform.

In such a scenario, the orientation and skills required are different and typically not found in conventional accounting education. Simply stated, the needed skills set is skewed more towards computer science and away from core accounting knowledge. At the risk of appearing alarmist, the imminent change has implications on the relevance of the profession and the profession's business models – a professional curriculum constrained by custom or tradition rather than one empowered by innovation. In the future, accounting firms will face and compete with formidable challenges set off by giant technology powerhouses and data aggregators such as Google for example.

It is entirely conceivable that future accounting firms will be predominantly staffed by data scientists with accounting and finance as their domain expertise.

As Bill Gates once said, we always overestimate the change that will occur in short term and underestimate the change that will occur in long term.

As Amy Webb, an NYU professor, explained at the World Economic Forum's 2019 annual conference in Davos, Switzerland, "The key issue, not just in a regulatory issue conversation framework, but just in general, is to bear in mind that there are just nine companies that control the future of Artificial Intelligence (AI)". AI represents the next era of computing and we ought to be paying attention to these companies. Those nine companies are the American giants Google, Microsoft, Amazon, Facebook, IBM, and Apple (commonly referred to as the G-MAFIA), as well as Chinese internet leaders Baidu, Alibaba, and Tencent (commonly referred to as the BAT companies).

It is against this backdrop that the Canadians should be heartened to note that two of Canada's notable pioneers in AI have been named co-winners of the 2018 Association for Computing Machinery (ACM) A.M. Turing Award for their work in deep neural networks.

The two Canadian winners are Geoffrey Hinton, a professor at the University of Toronto (UofT) and Yoshua Bengio, a professor at the University of Montreal. They share the prize with Yann LeCun, a professor at New York University. This award is considered equivalent to a Nobel Prize in Computer Science. Thus, their work, along with that of other Canadian researchers, has positioned Canada as the AI capital of the world.

To be clear, CPA New Brunswick and its authors are not computer scientists, nor computing experts, but rather CPAs who simply wish to table this monograph (not a research paper) with its members – presenting a concise, coherent and curated account of machine learning and its limitless application. It does not present any new or original ideas on the subject and is based on countless information sources in public domain. The monograph does not assume any knowledge of programming languages or statistics and the intended audience is constituted as professional accountants – aiming to furnish them a high level, non-technical introduction to the subject; perhaps even a call to action which will in future enable CPAs to better respond to future opportunity.

The subject of machine learning is as colossal and obscure as an iceberg and we have no illusion that this monograph covers even the tip of the proverbial iceberg.

It will however hope to serve as a compass – helping to steer the drift of the accounting profession arousing further interest and deeper conversation about redefining the profession.

# Executive Summary

Worth accentuating:

1. A plethora of emerging technologies are striking at the very foundation of the accountancy profession. It is difficult to make long term predictions about the future. It is however increasingly apparent that the profession is about to undergo a monumental makeover.

2. These emerging disruptive technologies are: Cloud Computing, Blockchain, Robotic Process Automation, Big Data Analytics, Artificial Intelligence including Machine Learning and Natural Language Processing.

3. The profession must rediscover and transform itself to remain relevant and provide value in the new normal where "software is eating the world".

4. CPAs need be entrepreneurial, proactively preparing themselves for the imminent challenges and opportunities of the continually evolving modern world.

5. This monograph provides a high level and non-technical introduction to one of the important emerging technologies: Machine Learning, also known as statistical learning because of its theoretical foundation in statistics.

6. Machine Learning is a subset of Artificial Intelligence (AI), which is itself a subset of Data Science. It provides systems the ability to automatically learn and improve from gathering data and experience as well as algorithms, without being explicitly programmed.

7. Machine Learning is a multi-disciplinary fusion of computer science, mathematics including statistics, and domain knowledge. It is concerned with predictive and prescriptive analytics.

8. An important assumption in machine learning is that the training data set is representative of the entire population of data sets. It means that the training data set, and test data set are similarly distributed. Also, the quality of data is important, and correlation should not be understood to imply misidentified causation.

There should be a priori logical anticipation of relationships between dependent and independent variables.

9. Most recent advances in Artificial Intelligence have been achieved by applying Machine Learning to very large data sets. Machine Learning algorithms detect patterns and learn how to make predictions and recommendations by processing data and experiences, rather than by receiving explicit programming instructions.

The algorithms also adapt in response to new data and experiences to improve efficacy over time.

10. Machine Learning applications can be expected to grow exponentially due to rapid advances in hardware and software development. However, Machine Learning may not be able to solve all problems because of insufficient high-quality data, wrong selection of tasks, algorithms tools, and lack of resources.

11. Machine Learning can be categorized on the basis of the algorithms deployed for deriving insights from the raw data input. An algorithm is a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

12. The six main categories of Machine Learning are Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Reinforcement Leaning, Deep Learning and Ensemble Learning.

13. Supervised Learning methods train a model using a labeled dataset.

14. Unsupervised Learning methods train a model to find patterns in an unlabeled dataset.

15. Semi-Supervised Learning methods are a class of Machine Learning tasks and techniques that also make use of unlabeled data for training (typically, a small amount of labeled data with a large amount of unlabeled data).

16. Reinforcement Learning refers to goal-oriented algorithms, which learn how to attain a complex objective (goal) or maximize along a particular dimension over many steps.

17. Deep Learning methods mimic the human brain and train models with several levels of abstractions from the raw data input to output.

18. Ensemble Learning methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

19. Selecting the right Machine Learning algorithm depends on several factors, including, but not limited to data size, quality and diversity, as well as what answers businesses want to derive from that data.

20. There are countless algorithms available for Machine Learning. The most important algorithms are:

- Linear Regression (Supervised Learning – Regression)
- Logistic Regression (Supervised Learning – Classification)
- Linear Discriminant Analysis (Supervised Learning – Classification)
- Naïve Bayes (Supervised Learning – Classification)
- Support Vector Machine (Supervised Learning – Classification),
- Decision Trees (Supervised Learning – Classification/Regression)
- Random Forests (Supervised Learning – Classification/Regression)
- K-Nearest Neighbors (Supervised Learning – Classification)
- AdaBoost (Supervised Learning – Classification)
- K-Means Clustering Algorithm (Unsupervised Learning – Clustering)
- Artificial Neural Networks (Deep Learning/Reinforcement Learning)

21. There are many open-source and proprietary tools available for developing and implementing Machine Learning applications. Some of these tools are user friendly, require no coding, and use Graphical User Interface (GUI).

22. Machine Learning has applications in such diverse fields as Manufacturing, Business, Health Care, Education and Law Enforcement.

There are many potential areas for application of Machine Learning in accounting and finance such as general accounting, fraud detection, credit rating, deriving market insights, content creation, algorithmic trading, process automation and anti-money laundering oversight.

23. There are many free and inexpensive resources including MOOCs (Massive Open Online Courses) available for interested learners who intend to gain a better understanding of Data Science, Artificial Intelligence, and Machine Learning.

# Accounting Profession And Emerging Technologies

***"It is dangerous to make forecasts, especially about the future."***

Yogi Berra, professional baseball catcher, manager, and coach, 18-time All-Star, and 10-time World Series championship player

Amongst all the learned professions, accountancy has been the most resilient to change – until recently. Accountants have earned the reputation of being the most conservative among all the professionals. The basic structure, methods and processes of the profession has essentially remained the same since the beginning of the industrial revolution. The advances in science and technology so far only made these methods and processes more efficient but have not altered the basic structure of the profession.

A plethora of emerging technologies under development in this century has, for the first time, shown potential to transform the profession beyond recognition.

As quoted above, it is dangerous to make forecasts, especially about the future. However, it is reasonable to conclude that more likely than not, the accounting profession is about to undergo a monumental makeover. Financial Reporting is already digitized with the advent of eXtensible Business Reporting Language (XBRL). XBRL shifts focus from data extraction to data access and data analysis.

XBRL ensures that the data that already exists is searchable, comparable and useable. XBRL makes access to data subordinate to the presentation of data.

It is difficult to list an inventory of emerging technologies that will be a game changer for the profession. The following paragraphs describe those that have the highest potential to impact and revolutionize the profession.

## Cloud Computing

Cloud computing is the technology that employs a cluster of remote servers hosted in different geographic locations. It is enabled via internet storing, managing, and

processing of data, instead of a local server or a personal computer. It is available on demand and on pay-as-you-go basis. The leading providers of cloud computing are Amazon, Microsoft, IBM, Salesforce and Alibaba.

Cloud computing can be divided into three distinct types:

### Software as a Service (SaaS)

This refers to software that is hosted on a third-party infrastructure but delivered to a client organization's end users as a service and often accessed through a specific web portal. For CPAs, this is the most relevant type of cloud computing.

When SaaS offers accounting applications, they are commonly referred to as Cloud Accounting. Cloud Accounting is rapidly gaining currency and will change the extant nature of basic accounting functions of retrieving, storing and processing of financial data.

### Platform as a Service (PaaS)

This offers platforms upon which apps and services can be built. It is primarily geared towards developers and operations professionals and not particularly relevant to business entities.

### Infrastructure as a Service (IaaS)

This offers the storage, networking, and computational resources needed to run a business.

Cloud computing has many forms and variations such as public, private and hybrid cloud. While adopting this technology for finance purposes, CPAs must carefully evaluate suitability, direct costs, indirect costs and benefits.

The main advantages of using cloud computing are:

- Access to computing resources anytime and anywhere;
- No need to maintain and update physical infrastructure for computing needs; and,
- Flexibility of scaling computing resources in response to business requirements.

However, the adoption of cloud computing limits the scope of computing only to the applications on offer by the provider of cloud computing.

The adopters of cloud computing need to adapt their business models and processes to suit these applications, rather than running things the other way around. CPAs also need to consider implications for security and privacy of data.

While building a business case for adopting cloud computing, all relevant factors should be given due consideration.

## Robotic Process Automation (RPA)

Robotic process automation (RPA) is the application of software and algorithms to perform routine repetitive tasks and operations without any external human intervention. Contrary to popular belief, RPA does not require physical robots, but is rather characterized by a mesh of software and algorithms.

RPA frees knowledge workers from mundane tasks and enables them to focus on tasks of higher value. Any RPA tool should have the following core functionalities:

- A bot should be able to interact with various other systems either through screen scraping or Application Programming Interface (API) integrations;
- A bot should be able to make decisions and determine its actions based on inputs gathered from other systems; and,
- A bot should have an interface to program itself.

The following three vendors are leading providers of RPA tools:

- UiPath (https://www.uipath.com/)
- Blue Prism (https://www.blueprism.com/)
- Automation Anywhere (https://www.automationanywhere.com/)

The evolution of Robotic Process Automation can be described in four distinct phases:

### Assisted RPA

Assisted RPA or RPA 1.0 means the RPA software automates various activities and applications running on the user's desktop.

The simplest example of this could be a simple 'cut-and-paste' of information from one screen to another in a word processor or a spreadsheet application. Assisted RPA does require real-time human-system interaction.

Assisted RPA does save time and costs and serves to simplify complex tasks while improving user experience.

### Unassisted RPA

Unassisted RPA or RPA 2.0 means the RPA software is deployed on several machines to run without requiring any external intervention. The robots work around the clock and can be monitored through dashboards representing a significant improvement over assisted RPA.

### Autonomous RPA

Autonomous RPA or RPA 3.0 incorporates and leverages other powerful technologies such as Artificial Intelligence, Machine Learning and Computer Vision. It represents the current and latest state of most potent available technologies.

### Cognitive RPA

Cognitive RPA is still under development and aims at making use of structured as well as unstructured data. It attempts to incorporate distinct algorithms and technology approaches such as natural language processing, data mining, semantic technology and text analytics.
Cognitive RPA, when developed, will automate predictive analytics and decision-making processes.

### RPA in Accounting and Finance

RPA has the potential to make accounting and finance functions more efficient, effective, and timely. It has already transformed accounting and finance functions such as processing of receivables, payables, payroll, reconciliations, month-end routines and preparation of custom month-end management reports. Tax Accounting is also highly amenable to RPA.

Blackline (https://www.blackline.com/) is one of the leading providers of cloud software that automates and controls the entire financial close process.

RPA also has applications in testing of internal controls, analytical procedures, as well as performing and documenting standard low risk audit procedures which do not require subjective professional judgement.

## Blockchain

The origin of Blockchain is synonymous with its first practical application, Bitcoin, which was first conceptualized in a white paper published in 2008 under the pseudonym Satoshi Nakamoto. (https://bitcoin.org/bitcoin.pdf)

Blockchain is a shared, immutable ledger that facilitates the process of recording transactions and tracking assets in a business network. An asset can either be tangible or intangible. Virtually anything of value can be tracked and traded on a blockchain network.

Blockchain eliminates risks and costs of recording as well as transacting. It is also extremely fast compared to the conventional methods of recording and transacting.

In essence, a Blockchain is a growing list of records, called blocks, which are linked and secured by cryptography. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data. The transactions are recorded in the ledger of two contracting parties and also on a shared, secured and distributed ledger.

Thus, Blockchain offers triple entry bookkeeping. The new transactions originate with one user, but propagate to a network of identical ledgers, without control from a central authority. Many blockchains are programmable and hence facilitate the automation of new transactions and controls via ‘smart-contracts’.

The leading developers of this technology are IBM, Hitachi and Daimler AG. This particular technology is still under development; requiring proper regulatory framework and safeguarding privacy.

A detailed technical description of how Blockchain works is naturally beyond the scope of this monograph.

## Big Data and Analytics

The Merriam Webster Dictionary defines Big Data as an accumulation of data that is too large and complex for processing by traditional database management tools.

An article by Forbes states that data is growing faster than ever before, and by the year 2020, about 1.7 megabytes of new information will be created every second for every human being on the planet. (https://tinyurl.com/yxacvca6)

Big Data is characterized by the following attributes:

### Volume

Big Data originates from diverse sources, including business transactions, social media, information from sensors and machine-to-machine data.

### Velocity

Big Data streams are generated at high velocity and in real-time aided by radio-frequency identification (RFID) tags, sensors and smart metering.

### Variety

Big Data may be structured such as numeric data in traditional databases, or it may be unstructured such as text documents, email, video and audio.

### Variability

In addition to the increasing volumes, velocities and varieties of Big Data, its flows can be highly inconsistent with periodic peaks and dips.

### Complexity

Big Data may originate from multiple sources, which makes it difficult to link, match, cleanse and transform across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages before further analysis can be undertaken.

When Big Data is combined with powerful analytics, it can generate valuable business insights that may enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

Big Data finds applications in all sectors of economy including financial services.

## Accounting is Big Data

Financial Accounting systems generate structured Big Data that can be leveraged to generate insights by combining the same with proper analytics.

Different data sources such as text, video and audio are being integrated into accounting information systems. CPAs need to enhance their data analytic skills to deal with large volumes of available data, including automatically mined data.

CPAs in businesses must change from decision-making supporters to business partners, create value for businesses and cultivate an evidence-based decision-making culture rather than one based on management opinions.

They should also be able to evaluate, manage and mitigate risks related to the privacy and security of data on premises as well as in the cloud.

With Big Data, auditors can analyse both structured and unstructured data to identify potential transactional anomalies (e.g. unauthorized disbursements), patterns of behaviour (e.g. split payments to bypass transaction limit), and trends (e.g. increased fraudulent transactions before a big holiday).

As a consequence of using automatic data collection and rule-based analysis techniques to identify errors, auditors may shift responsibilities from detecting errors in data to judging which errors are worthy of further investigation.

In 2015, CPA Canada, the American Institute of CPAs (AICPA), and Rutgers Business School partnered to create the Rutgers AICPA Data Analytics Research Initiative, which hopes to help integrate data analytics into the audit process to enhance audit quality.

The following four examples explain how the initiative will affect auditing:

- Population-level tests would be feasible over traditional sampling because of the digitization of transaction data and reduced costs of data analysis.
- The role of auditors with the emergence of Big Data will move from statement-level assurance to data-level assurance.
- Auditors will need to use text analytics to manage unstructured data, such as text in the management discussion and analysis sections of financial reports.
- Auditors will face a less challenging task in asserting the existence of fixed assets if each of the asset's records are complemented with pertinent audio, video, and textual information.

It can be concluded that Big Data and analytics will have increasingly important implications for accounting, even as new types of data become accessible.

The video, audio, and textual information made available via Big Data can provide for improved financial accounting, managerial accounting, and financial reporting practices.

In financial accounting, Big Data will improve the quality and relevance of accounting information, thereby enhancing transparency and stakeholder decision making. In managerial accounting, Big Data will contribute to the development and evolution of effective management control systems and budgeting processes.

In financial reporting, Big Data can assist with the creation and refinement of accounting standards, helping to ensure that the accounting profession will continue to provide useful information as the dynamic, real-time, global economy evolves.

## Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) can be defined as the ability of a machine to perform cognitive functions normally associated with human minds, such as perceiving, reasoning, learning, and problem solving.

AI is not a single technology with specific application, but a collection of technologies which have general purpose applications across the domains. Examples of technologies that enable AI to solve business problems are Robotics, Computer Vision, Natural Language Processing (NLP), and Machine Learning (ML).

Machine Learning is the subject matter of this monograph. It is also known as statistical learning because of its theoretical foundation in statistics.

ML is a subset of artificial intelligence; in fact, it is simply a technique for realizing AI. It is a method of training algorithms such that they can learn how to make decisions and/or predictions without explicit programming instructions.

Training in Machine Learning entails giving a lot of data to the algorithm and allowing it to learn more about the processed information.

ML is highly relevant for CPAs as it can enable them to automate data analysis of complex data sets and make predictions based on such analysis. The subsequent sections of this monograph provide a high-level introduction and overview of ML tools, algorithms and examples.

The monograph provides a non-technical exposure to ML and no prior knowledge of computer science concepts, programming languages or statistics is assumed.

## Conclusion

Accountancy has historically concerned itself with retrieving, processing and analyzing financial information to provide insights useful for decision making.

Perhaps for the first time in its history, the profession faces serious challenges to its relevance.

The profession cannot remain in a state of denial and must rediscover itself to add value in a new normal wherein the financial information and transactions will be:

- captured digitally,
- recorded on a Blockchain that guarantees immutability,
- retrieved by Structured Query Language (SQL),
- integrated with non-financial and unstructured data,
- processed by RPA,
- presented in XBRL format,
- analyzed by ML and NLP enabled AI to generate insights and predictions, and
- accessed in the cloud at any given time and place.

The new normal will require new sets of skills and a tailored acumen within the new business models and technologies of the 21st century.

# Machine Learning: A Primer

***"We are drowning in information and starving for knowledge."***

Rutherford D. Roger, famous for his passion for knowledge, great quotations, and affirmations

A generally accepted formal definition of Machine Learning is as follows:

***"A computer program is said to learn from experience E with respect to some class of task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."***

Tom Mitchell (Machine Learning 1997)

Machine Learning is such an important component of Artificial Intelligence that both terms are frequently used interchangeably. In reality, Machine Learning is a subset of Artificial Intelligence. Artificial Intelligence itself comprises all technologies including Machine Learning that enable the machine to mimic human intelligence.

Artificial Intelligence and Machine Learning are actually different but related concepts. One way to think about the relationship of AI and ML is that the former is a problem while the latter is one of the possible solutions. If the end goal is that a computer can solve a problem with the cognitive abilities of (human) intelligence, the processing of algorithms through data to apply to new and larger situations is one method of getting there.

Machine Learning teaches computers to do what comes naturally to humans and animals: learn from experience. Machine Learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. The machine learns directly from training data (examples) and is not directed by explicit programming instructions. Hence the more the data, the better the outcome.

An important assumption in Machine Learning is that the training data set is representative of the entire population of data sets. It means that the training data set, and test data set are similarly distributed. Also, the quality of data is important and the principle of Garbage In, Garbage Out applies.

Also, correlation should not be understood to imply misidentified causation. There should be a priori logical anticipations of relationships between dependent and independent variables.

Machine Learning algorithms can figure out how to perform important tasks by generalizing from examples. This is often feasible and cost-effective where manual programming is not. As more data becomes available, more ambitious problems can be tackled. As a result, Machine Learning is widely used in computer science and other fields.

Machine Learning uses a variety of algorithms that iteratively learn from data to improve it, describe it, and predict outcomes. As the algorithms ingest training data, it is possible to produce more precise models based on that data.

Machine Learning algorithms find natural patterns in data that generate insights and help make better decisions and predictions. They are used every day to make critical decisions in medical diagnosis, stock trading, energy load forecasting, and more.

Media sites rely on Machine Learning to sift through millions of options to give you song or movie recommendations. Retailers use it to gain insights into their customers' purchasing behaviour.

Most recent advances in AI have been achieved by applying Machine Learning to very large data sets. Machine Learning algorithms detect patterns and learn how to make predictions and recommendations by processing data and experiences, rather than by receiving explicit programming instructions. The algorithms also adapt in response to new data and experiences to improve efficacy over time.

In recent times, commercial Machine Learning applications have experienced exponential growth. This has become possible by the synergy generated by the enablers described below:

- Computing has become extremely powerful and cheap;
- Data storage has similarly become highly affordable and innovative;
- Distributed computing has greatly improved the ability to analyze complex data in record time;
- There are more commercial data sets available across domains such as health care and finance. Many of them are available as cloud services Application Programming Interfaces (APIs);
- Development of Machine Learning algorithms is highly facilitated because of the wide availability of open source frameworks and libraries;
- Visualization of data has become more user-friendly and accessible to non-technical users.

The most critical part of the training process for Machine Learning is to acquire sufficient high-quality data for testing the model. Training for Machine Learning is an iterative and continuous process, at each step, providing more data and/or refining the model. The entire Machine Learning cycle can be broadly summarized as follows:

### Data acquisition

The first step is to source relevant data for the application under development. The data should be high-quality and extensive.

### Data preparation

This step is also called data cleaning or data shaping or data wrangling. The data should be accurate, clean and secured.

### Algorithm selection

The most appropriate algorithm for the application under development should be identified.

### Model training

A Machine Learning model is a mathematical representation of a real-world process. The algorithm selected needs to be trained on the data to create the model. The training process may be supervised, unsupervised, or reinforcement learning. The detailed description of this process is provided in other sections of this monograph.

### Evaluation

The model should be evaluated to ensure that the algorithm selected is the most appropriate.

### Deployment

Decisions should be made if the model should be deployed in the cloud or on premises.

### Testing

The model should be tested with new and previously unseen data to make predictions.

### Assessment

The validity of predictions made by the model should be assessed and the refinement of data, model and algorithm should be implemented as deemed appropriate.

## Machine Learning in Practice

A Machine Learning project requires collaboration between business analysts, executives and data scientists. The following are the typical steps for implementing such a project:

• Identify a problem that has a business outcome and for which sufficient high-quality data is readily available.

• Implement a pilot project which is a subset of the main project. A successful or even a failed pilot project can generate very useful insights.

• Evaluate the pilot project and, on the basis of knowledge gained, modify the data sources, algorithms and assumptions as required.

• Scale up the pilot project to a full-fledged project and identify other areas of the business where Machine Learning can add value.

## The Future of Machine Learning

Data is the new gold, and Machine Learning is the technique to mine it. Machine Learning is emerging as one of the most important developments in the software industry.

While this advanced technology has been around for decades, it is only now becoming commercially feasible.

Machine Learning techniques are essential tools to create value for businesses that want to understand the hidden value of their data.

The following developments can be visualized for the future of Machine Learning:

### ML Embedded Applications

ML will be seamlessly embedded in the most common applications such as e-commerce websites.

### Dynamic Models

The offline models are static and based on old data, and hence become less accurate when the business environment changes. The models deployed in cloud would be dynamic and will be able to ingest new data that reflects new realities and thus will be able to retain the accuracy of predictions.

### Data as a Service (DaaS)

Sourcing, training and labelling of data is time-consuming and expensive. This will result in new business opportunities such as supplying pre-trained data suitable for a specific application.

### Machine Learning as a Service (MLaaS)

MLaaS vendors will offer cloud-based tools such as image recognition, voice recognition, data visualization, and deep learning. A user will be able to upload data to a vendor's cloud, and then the data will be processed in the cloud to generate the desired output.

### NLP Interface

With the evolution of Natural Language Processing technology, the human-machine interface will be predominantly in spoken and written words.

### Automated ML

Routine and tedious tasks in ML workflows such as data cleaning and data visualization will be automated, and ML platforms with Graphic User Interface (GUI) will evolve. Hence ML will become more accessible to non-technical users. Similarly, optimum algorithm selection will be automated, greatly facilitating the work of developers.

### Hardware Innovation

New hardware technology such as Graphical Processing Units (GPUs) will greatly speed up processing and hence enable implementation of complex algorithms and massive data.

In the end, it should be remembered that ML is not magic, nor the panacea for instinctively solving all problems. ML often fails to deliver expected results. This may be because of insufficient high-quality data, wrong selection of tasks, algorithms tools, and lack of resources.

# Machine Learning Types

***"Any sufficiently advanced technology is indistinguishable from magic."***

Sir Arthur C. Clarke, science writer and futurist, inventor, undersea explorer, and television series host – co-writer of the screenplay for the film 2001: A Space Odyssey

Machine Learning is a subset of Artificial Intelligence, which itself is a subset of Data Science. Data Science is concerned with descriptive, diagnostic, predictive and prescriptive analytics. Descriptive analytics describes what happened in the past; diagnostic analytics answers why it happened; predictive analytics forecasts what is most likely to happen in the future; and prescriptive analytics prescribes the most logical course of action for achieving the desired outcome.

Machine Learning is focused on predictive and prescriptive analytics, depending upon the nature of analytics and algorithms employed. This section provides an overview of the most popular types of Machine Learning. A non-technical and high-level synopsis of the algorithms for Machine Learning is presented in a separate section.

## Supervised Learning

Supervised Learning algorithms make predictions based on a set of examples, e.g. historical sales can be used to estimate future prices. With supervised learning, you have an input variable that consists of labeled training data and a desired output variable. You use an algorithm to analyze the training data to learn the function that maps the input to the output. This inferred function maps new, unknown examples by generalizing from the training data to anticipate results in unseen situations.

Supervised Learning is simply a formalization of the idea of learning from examples. In Supervised Learning, the learner (typically, a computer program) is provided with two sets of data, a training set and a test set. The idea is for the training set to "learn" from a set of labeled examples in the training set so that it can identify unlabeled examples in the test set with the highest possible accuracy.

The goal of the learner is to develop a rule, a program, or a procedure that classifies new examples (in the test set) by analyzing examples it has been given that already have a class label.

For example, a training set might comprise of images of different types of fruits, where the identity of the fruit in each image is given to the learner. The test set would then consist of more unidentified pieces of fruit, but from the same class. The goal is for the learner to develop a rule that can identify the elements in the test set.

There are many different approaches that attempt to build the best possible methods of classifying examples of the test set by using the data given in the training set. In Supervised Learning, the training set consists of n ordered pairs (x1, y1), (x2, y2), ..., (xn, yn), where each xi is some measurement or set of measurements of a single example data point, and yi is the label for that data point.

For example, a xi might be a group (sometimes called a vector1) of five measurements for a patient in a hospital, including height, weight, temperature, blood sugar level, and blood pressure.

The corresponding yi might be a classification of the patient as “healthy” or “not healthy”. The test data in Supervised Learning is another set of measurements without labels: (xn+1, xn+2, ..., xn+m).

As described above, the goal is to make educated deductions about the labels for the test set (such as “healthy” or “not healthy”) by drawing inferences from the training set. Supervised Learning problems may be further sub-classified as follows:

### Classification

When the data are being used to predict a categorical variable, Supervised Learning is also called classification. This is the case when one is assigning a label or indicator, e.g. dog or cat, to an image. When there are only two labels, this is called binary classification. When there are more than two categories, the problems are called multi-class classification.

### Regression

When predicting continuous values, the problem becomes a regression problem.

### Forecasting

This is the process of making predictions about the future based on the past and present data. It is most commonly used to analyze trends. A common example might be the estimation of next year’s sales based on the sales of the current year and previous years.

## Semi-Supervised Learning

The challenge with Supervised Learning is that labeling data can be expensive and time consuming. If labels are limited, you can use unlabeled examples to enhance Supervised Learning. Because the machine is not fully supervised in this case, we say the machine is semi-supervised.

With Semi-supervised Learning, you use unlabeled examples with a small amount of labeled data to improve the learning accuracy.

## Unsupervised Learning

When performing Unsupervised Learning, the machine is presented with totally unlabeled data. It is asked to discover the intrinsic patterns that underlies the data, such as a clustering structure, a low-dimensional manifold, or a sparse tree and graph.

### Clustering

Grouping a set of data examples so that examples in one group (or one cluster) are more similar (according to some criteria) than those in other groups. This is often used to segment the whole dataset into several groups. Analysis can be performed in each group to help users find the intrinsic patterns.

### Dimension reduction

Reducing the number of variables under consideration. In many applications, the raw data have very high dimensional features and some features are redundant or irrelevant to the task. Reducing the dimensionality helps to find the true, latent relationship.

Unsupervised Learning is best suited for when the problem requires a massive amount of unlabeled data. For example, social media applications have large amounts of unlabeled data.

Understanding the meaning behind this data requires algorithms that are able to classify the data based on the patterns or clusters it finds. Therefore, the Unsupervised Learning conducts an iterative process of analyzing data without human intervention.

Unsupervised Learning is used for email spam-detecting technology. There are far too many variables in legitimate and spam emails for an analyst to flag unsolicited bulk email. Instead, Machine Learning classifiers based on clustering and association are applied in order to identify unwanted email.

Unsupervised Learning algorithms segment data into groups of examples (clusters) or groups of features. The unlabeled data creates the parameter values and classification of the data. In essence, this process adds labels to the data so that it becomes supervised.

Unsupervised Learning can determine the outcome when there is a massive amount of data. In this case, the developer doesn't know the context of the data being analyzed, so labeling isn't possible at this stage. Therefore, Unsupervised Learning can be used as the first step before passing the data to a Supervised Learning process.

Unsupervised Learning algorithms can help businesses understand large volumes of new, unlabeled data. Similar to Supervised Learning, these algorithms look for patterns in the data; however, the difference is that the data is not yet understood.

For example, in healthcare, collecting huge amounts of data about a specific disease can help practitioners gain insights into the patterns of symptoms and relate those to outcomes from patients.

It would take too much time to label all the data sources associated with a disease such as diabetes. Therefore, an Unsupervised Learning approach can help determine outcomes more quickly than a Supervised Learning approach.

## Reinforcement Learning

Reinforcement Learning analyzes and optimizes the behavior of an agent based on the feedback from the environment. Machines try different scenarios to discover which actions yield the greatest reward, rather than being told which actions to take. Trial-and-error and delayed reward distinguish Reinforcement Learning from other techniques.

Reinforcement Learning is a behavioral learning model. The algorithm receives feedback from the data analysis, so the user is guided to the best outcome. Reinforcement Learning differs from other types of Supervised Learning because the system isn't trained with the sample data set.

Rather, the system learns through trial and error. Therefore, a sequence of successful decisions will result in the process being "reinforced," because it best solves the problem at hand.

Reinforcement Learning is used for self-driving cars. In many ways, training a self-driving car is incredibly complex because there are so many potential obstacles. If all the cars on the road were autonomous, trial-and-error would be easier to overcome.

However, in the real world, human drivers can often be unpredictable. Even with this complex scenario, the algorithm can be optimized over time to find ways to adapt to the state where actions are rewarded.

One of the easiest ways to think about Reinforcement Learning is the way an animal is trained to take actions based on rewards. If the dog gets a treat every time he sits on command, he will take this action each time.

## Deep Learning (Neural Networks)

Deep Learning is a specific method of Machine Learning that incorporates neural networks in successive layers in order to learn from data in an iterative manner.

Deep Learning is especially useful when trying to learn patterns from unstructured data.

Deep Learning methods (complex neural networks) are designed to emulate how the human brain works so that computers can be trained to deal with abstractions and problems that are poorly defined. Neural Networks and Deep Learning are often used in image recognition, speech, and computer vision applications.

A Neural Network consists of three or more layers: an input layer, one or many hidden layers, and an output layer. Data is ingested through the input layer. Then, the data is modified in the hidden layer and the output layers based on the weights applied to these nodes. The typical Neural Network may consist of thousands or even millions of simple processing nodes that are densely interconnected.

The term "Deep Learning" is used when there are multiple hidden layers within a neural network. Using an iterative approach, a neural network continuously adjusts and makes inferences until a specific stopping point is reached. Neural networks are often used for image recognition and computer vision applications. Deep Learning is a machine learning technique that uses hierarchical neural networks to learn from a combination of unsupervised and supervised algorithms.

Deep learning is often called a sub-discipline of Machine Learning. Typically, machine learns from unlabeled and unstructured data. While deep learning is very similar to a traditional neural network, it will have many more hidden layers. The more complex the problem, the more hidden layers there will be in the model.

There are many areas where Deep Learning will have an impact on businesses. For example, voice recognition will have applications in everything: from automobiles to customer management. In the Internet of Things (IoT) and manufacturing applications, deep learning can be used to predict when a machine will malfunction.

## ML Algorithm Selection

When choosing an algorithm, three aspects should be taken into account: accuracy, training time, and ease of use. Many users put the accuracy first, while beginners tend to focus on algorithms they know best.

Beginners tend to choose algorithms that are easy to implement and that generate results quickly. This works fine, as long as it is just the first step in the process, but the next step should be to use more sophisticated algorithms to strengthen understanding of the data, and further improve the results.

# Machine Learning Algorithms

***"An algorithm is a methodical set of steps that can be used to make calculations, resolve problems and reach decisions. An algorithm isn't a particular calculation, but the method followed when making the calculation."***

Yuval Noah Harari, historian, author, and professor

Selecting the right machine learning algorithm depends on several factors, including, but not limited to data size, quality and diversity, as well as what answers businesses want to derive from that data.

Additional considerations include accuracy, training time, parameters, data points and much more.

Therefore, choosing the right algorithm is both a combination of business needs, specifications, experimentation and time available.

It is difficult to judge which algorithm will perform the best before experimenting with others.

The following is a high level non-technical account of the most commonly employed algorithms for Machine Learning tasks.

### Linear Regression (Supervised Learning – Regression)

Linear regression is the most basic type of regression. Simple linear regression allows us to understand the relationships between two continuous variables.

### Logistic Regression (Supervised Learning – Classification)

Logistic regression focuses on estimating the probability of an event occurring based on the previous data provided. It is used to cover a binary dependent variable – that is where only two values, 0 and 1, represent outcomes.

## Linear Discriminant Analysis (Supervised Learning – Classification)

Logistic regression is a classification algorithm traditionally limited only to two-class classification problems. For more than two classes, Linear Discriminant Analysis (LDA) algorithm is the preferred linear classification technique.

The representation of LDA is pretty straight forward. It consists of statistical properties of your data, calculated for each class. For a single input variable, this includes:

- The mean value for each class, and
- The variance calculated across all classes.

Predictions are made by calculating a discriminate value for each class and making a prediction for the class with the largest value. The technique assumes that the data has a Gaussian distribution (bell curve), so it is a good idea to remove outliers from the data beforehand. It's a simple and powerful method for classification predictive modeling problems.

## Naïve Bayes (Supervised Learning – Classification)

The Naïve Bayes classifier is based on Bayes' theorem and classifies every value as independent of any other value. It allows us to predict a class/category, based on a given set of features, using probability. Despite its simplicity, the classifier does surprisingly well and is often used due to the fact it outperforms more sophisticated classification methods.

## Support Vector Machine (Supervised Learning – Classification)

Support Vector Machine or SVM can be used for both regression and classification tasks. However, it is widely used in classification objectives. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N= the number of features) that distinctly classifies the data points.

## Decision Trees (Supervised Learning – Classification/Regression)

A Decision Tree is a flow-chart-like tree structure that uses a branching method to illustrate every possible outcome of a decision. Each node within the tree represents a test on a specific variable – and each branch is the outcome of that test.

## Random Forests (Ensemble Supervised Learning – Classification/ Regression)

Random Forests or 'Random Decision Forests' is an ensemble learning method, combining multiple algorithms to generate better results for classification, regression and other tasks. Each individual classifier is weak, but when combined with others, can produce excellent results.

The algorithm starts with a 'decision tree' (a tree-like graph or model of decisions) and an input is entered at the top. It then travels down the tree, with the data being segmented into smaller and smaller sets, based on specific variables.

### K-Nearest Neighbours (Supervised Learning – Classification)

The K-Nearest-Neighbour algorithm estimates how likely a data point is to be a member of one group or another. It essentially looks at the data points around a single data point to determine what group it is actually in. For example, if one point is on a grid and the algorithm is trying to determine what group that data point is in (Group A or Group B, for example) it would look at the data points near it to see what group the majority of the points are in.

### AdaBoost (Ensemble Supervised Learning – Classification)

Boosting is an ensemble technique that attempts to create a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. Models are added until the training set is predicted perfectly, or a maximum number of models are added.

AdaBoost was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting. Modern boosting methods build on AdaBoost, most notably stochastic gradient boosting machines.

### K-Means Clustering Algorithm (Unsupervised Learning – Clustering)

The K-Means Clustering algorithm is a type of unsupervised learning, which is used to categorise unlabelled data, i.e. data without defined categories or groups.

The algorithm works by finding groups within the data, with the number of groups represented by the variable K. It then works iteratively to assign each data point to one of K groups based on the features provided.

### Artificial Neural Networks (Deep Learning/Reinforcement Learning)

An artificial neural network (ANN) comprises 'units' arranged in a series of layers, each of which connects to layers on either side. ANNs are inspired by biological systems, such as the brain, and how they process information. ANNs are essentially a large number of interconnected processing elements, working in unison to solve specific problems.

ANNs also learn by example and through experience, and they are extremely useful for modelling non-linear relationships in high-dimensional data or where the relationship amongst the input variables is difficult to understand.

# Machine Learning Tools

***"We become what we behold. We shape our tools and then our tools shape us."***

Marshall McLuhan, Canadian philosopher

Machine learning tools can be classified either as Platforms or Libraries. A platform provides everything to implement a project, whereas a library only provides specific capabilities for what is required to implement a project. However, some machine learning platforms are also libraries.

ML tools can also be distinguished by their interface. Some tools provide Graphical User Interface, some provide Command Line Interface and some others provide Application Programming Interface (API).

One more way of distinguishing ML tools is on the basis of whether a tool is installed on a local machine on-premises or accessed remotely on a third-party server in the cloud.

The following is a non-exhaustive list of important ML tools which can be differentiated as follows:

## Free and Open-Source Tools

Python, (https://www.python.org) is a high-level programming language. It is one of the most popular programming languages and particularly so for Data Science and Machine Learning.

It has very rich libraries for data shaping, analysis and visualizations. It is the language of choice for developing industrial-strength Machine Learning Applications.

Scikit-learn, (https://scikit-learn.org/stable/) is a free software Machine Learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random

forests, gradient boosting, and k-means. It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

R, (https://cran.r-project.org/) is a derivative language of another programming language S. It is most suitable for statistical computing and graphics. It also has all the necessary and capable packages (executable codes) for implementing advanced Machine Learning algorithms.

Apache Hadoop, (https://hadoop.apache.org) is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. It is at the center of a growing ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and Machine Learning applications.

Hadoop can handle various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analyzing data than relational databases and data warehouses provide.

Apache Spark, (https://spark.apache.org) is a general-purpose distributed data processing engine that is suitable for use in a wide range of circumstances. On top of the Spark core data processing engine, there are libraries for SQL, Machine Learning, graph computation, and stream processing, which can be used together in an application.

Cognitive Microsoft Toolkit, (https://www.microsoft.com/en-us/cognitive-toolkit/) previously known as CNTK and sometimes styled as The Microsoft Cognitive Toolkit, is a deep learning framework developed by Microsoft Research. Microsoft Cognitive Toolkit describes neural networks as a series of computational steps via a directed graph.

H2O, (https://www.h2o.ai/) is an open-source software for big-data analysis. It is produced by the company H2O.ai. H2O allows users to fit thousands of potential models as part of discovering patterns in data.

Orange, (https://orange.biolab.si/) is an open-source data visualization, Machine Learning and data mining toolkit. It features a visual programming front-end for explorative data analysis and interactive data visualization. It can also be used as a Python library.

Waikato Environment for Knowledge Analysis (WEKA), (https://www.cs.waikato.ac.nz/~ml/weka/) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is a free software licensed under the GNU General Public License.

Massive Online Analysis (MOA), (https://moa.cms.waikato.ac.nz/) is a free open-source software project specific for data stream mining with a concept drift. It is written in Java and developed at the University of Waikato, New Zealand.

## Proprietary Tools with Free and Open-Source Community Editions

KNIME, (Konstanz Information Miner), (https://www.knime.com/), is free and open-source data analytics, reporting and integration platform. KNIME integrates various components for Machine Learning and data mining through its modular data pipelining concept.

A graphical user interface and use of Java Database Connectivity (JDBC) allows assembly of nodes blending different data sources, including pre-processing (ETL: Extraction, Transformation, Loading), for modeling, data analysis and visualization without, or with only minimal, programming.

RapidMiner, (https://rapidminer.com/) is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics.

It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development. It supports all steps of the Machine Learning process including data preparation, results visualization, model validation and optimization.

## Exclusive Proprietary Tools

Amazon Web Services (AWS), (https://aws.amazon.com/) is a subsidiary of Amazon that provides on-demand cloud computing platforms to individuals, companies and governments, on a paid subscription basis. The technology allows subscribers to have at their disposal a virtual cluster of computers, available all the time, through the Internet.

In 2017, AWS comprised more than 90 services spanning a wide range including computing, storage, networking, database, analytics, application services, deployment, management, mobile, developer tools, and tools for the Internet of Things (IoT).

IBM SPSS Modeler, (https://www.ibm.com/products/spss-modeler) is a data mining and text analytics software application from IBM. It is used to build predictive models and conduct other analytic tasks. It has a visual interface which allows users to leverage statistical and data mining algorithms without programming.

Wolfram Mathematica, (https://www.wolfram.com/mathematica/) is a modern technical computing system spanning most areas of technical computing — including neural networks, machine learning, image processing, geometry, data science, visualizations, and others.

MATLAB (Matrix Laboratory), (https://www.mathworks.com/products/matlab.html) is a multi-paradigm numerical computing environment and proprietary programming language developed by MathWorks.

MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other programming languages. Although MATLAB is intended primarily for numerical computing, optional toolboxes are available for the development of ML applications.

Microsoft Azure, (https://azure.microsoft.com/en-ca/) is a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through Microsoft-managed data centers. It provides software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS). It supports many different programming languages, tools and frameworks, including both Microsoft-specific and third-party software and systems.

SAS ("Statistical Analysis System"), (https://www.sas.com/en_ca/home.html) is a software suite developed by SAS Institute for advanced analytics, multivariate analysis, business intelligence, data management, machine learning and predictive analytics.

Statistica, (http://www.statistica.com/) is an advanced analytics software package originally developed by StatSoft. It provides data analysis, data management, statistics, data mining, machine learning, text analytics and data visualization procedures.

There are very large number of tools available for ML tasks other than those described above. They vary in level of sophistication and functionalities. For beginners in ML without advanced knowledge of programming and statistics, Orange and Weka offer an ideal balance between ease of use and functionalities.

A good option for Excel users with basic understanding of statistics is XL Miner, an Excel Add-In (https://www.solver.com/xlminer-data-mining). However, this option typically works best for a small data set, while a very large data set may crash Excel.

Microsoft has also enhanced its Business Intelligence (BI) tool, Power BI, and incorporated automated AI and ML without requiring any coding (https://powerbi.microsoft.com/en-us). Another BI tool Tableau (https://www.tableau.com/) has similar capabilities.

# Machine Learning Applications

## (Accounting & Finance)

***"Aim for simplicity in Data Science. Real creativity won't make things more complex. Instead, it will simplify them."***

Damian Duffy Mingle, cartoonist, scholar, writer, curator, lecturer, teacher, and a New York Times bestselling graphic novelist

Many organizations are bringing ML into their enterprise data analytics practices to help expose hidden insights and make smarter recommendations to make business decisions. This is especially helpful in big data analytics and handling increasingly large and complex data sets.

Machine Learning can also identify behavioral trends within an organization to make suggestions to users that appear similar to others such as which data sources to use for analysis, or which analytical content is the most relevant to help answer a particular question.

Other areas of continued development include advanced and predictive analytics. Machine Learning can help automate advanced statistical analyses and automatically apply models with the highest confidence, allowing less advanced users to take advantage of complex models.

More advanced users can explore and modify calculations, which not only addresses trust and transparency, but also allows testing of different what-if scenarios.

Machine learning is also being used in analytics to help users query their data with natural language. This essentially means learning to interpret human intent and semantics behind questions and translating requests into a structured query language.

With advances in natural language and other smart analytics capabilities powered by AI and ML, people without traditional data skills will be able to work with data in new and exciting ways to get new insights.

Machine learning has applications in diverse fields such as Manufacturing, Business, Health Care, Education and Law Enforcement. There are many potential areas for application of Machine Learning in accounting and finance. The following paragraphs describe some of these possible applications; some in more details than others.

## General Accounting

Machine learning can automate the processing of receivables, payables and account reconciliations. ML is also well-suited for transaction matching and interpretation of complex and lengthy contracts. There are many established players and start-ups active in this area, e.g. Blackline (https://www.blackline.com/), Receipt Bank (https://www.receipt-bank.com/) and Hubdoc (https://www.hubdoc.com/).

In addition to the leading enterprise accounting software providers, the small business accounting software providers such as QuickBooks Online (http://tinyurl.com/y5a6cu99 /) and Xero (https://www.xero.com/ca/) are aggressively incorporating AI and ML in their products.

## Fraud Detection

Fraud detection is a challenging problem. The fact is that fraudulent transactions are relatively rare; representing a very small fraction of activity within an organization. The challenge is that a small percentage of activity can quickly turn into big losses without the right tools and systems in place. As traditional fraud schemes fail to pay off, fraudsters change their tactics. The good news is that with advances in Machine Learning, systems can learn, adapt and uncover emerging patterns for preventing fraud.

Most organizations still use rule-based systems as their primary tool to detect fraud. Rules can do an excellent job of uncovering known patterns; but rules alone aren't very effective at uncovering unknown schemes, adapting to new fraud patterns, or handling fraudsters' increasingly sophisticated techniques. This is where Machine Learning may become crucial to fraud detection.

Data sets are only growing larger, and as the volumes increase, so does the challenge of detecting fraud. In fact, data is key when it comes to building machine learning systems. The adage that more data equals better models is true when it comes to fraud detection. Practitioners need their machine learning platform to scale as data and complexity increase.

There is no single machine learning algorithm or method that works. Success comes from the ability to try lots of different ML-based methods, trying variations on them and testing them with a variety of data sets. The data scientist needs a toolkit with a variety of supervised and unsupervised methods as well as a variety of feature engineering techniques. It is the application of Machine Learning in new and novel ways, like combining a variety of supervised and unsupervised methods in one system, to be more effective than any single method alone.

Once a Machine Learning model is developed, it should be integrated with operations and properly documented to explain its methodology.

Ongoing monitoring of Machine Learning fraud detection systems is imperative for success. As populations and the underlying data shift, expected system inputs degrade and therefore have an impact on overall performance.

This is not unique to Machine Learning systems; rule-based systems have the same challenge. But newer Machine Learning methods can adapt to new and unidentified patterns as underlying changes occur. This eliminates some, but not all, of the Machine Learning retraining and evaluation steps.

A good monitoring program proactively looks at the data entering the system, evaluates the Machine Learning model's predictions and explanations, and alerts administrators to shifting data trends and statistics before dramatic changes affect operations and the bottom line. A successful Machine Learning program should have an element of ongoing experimentation.

## Credit Rating

Credit bureaus have now adopted the use of big data to develop credit score models before they determine how creditworthy a business is. This is a major advantage because the lenders now have a way to accurately assess businesses that ask for loans.

And the good news for any business is that they can take advantage of big data to build their credit-scoring model in the manner they desire. It is a delicate process that may involve data as well as financial expertise. This will happen as the business continuously strives to develop the best business model ever.

Most businesses have been buying generic scores to improve their models, and experts say that this is an acceptable move. But custom models are a better method that should be proposed by any financial consultant who appreciates the benefits of data.

According to the experts, a custom model works with more data from different sources. Custom models may use account data, supplier information or customer relationship data among many other types of data. Therefore, there will be more sources to increase the accuracy of the model and make it more effective.

By now, all companies and lending institutions usually benefit from big data from different credible sources. This has positively changed their reliance on credit bureau data. Although the credit bureau's credit history of a company is the primary source of data, companies should seek other sources of data, even if it means buying it.

According to data scientists, models that have the best algorithm can dig deep inside the databases to obtain untapped correlations that further make the credit score data more useful. After all, machine learning was created to handle vast

amounts of data. However, all these new developments come with some challenges. The outputs can sometimes be confusing, even to the experts, especially if there is a problem with the algorithms. Therefore, problems can also occur during this process. This has caused some organizations to give up on the use of big data and Machine Learning.

From the above insights it can be concluded that big data has made some advancements in how organizations, businesses, and lenders create scoring models. The best part is the provision of more data than the data in the usual credit bureau reports. While the latter cannot be overlooked, scoring models bring increased accuracy and provide an alternative point of view for businesses.

## Market Insights

Using Machine Learning, money managers can identify market changes earlier than possible with traditional models. The potential of Machine Learning technology to disrupt the investment advisory industry is taken seriously by major financial institutions. JPMorgan, Bank of America, and Morgan Stanley are developing automated investment advisors or Robo advisors, powered by ML technology.

## Algorithmic Trading

Algorithmic trading automates the trading process by executing trades according to predefined criteria set by the trader or fund manager. In its simplest form, an "algo" trade can automatically buy (or sell) a quantity of stock when the price reaches a specific level.

ML technology offers a new and diverse suite of tools to make algorithmic trading more than automatic. ML makes algo trading intelligent. ML algorithms are designed to analyze historical market behavior and determine an optimal market strategy to make trade predictions.

## Process Automation

As financial institutions transition from spreadsheets to cloud-based data storage, a tremendous opportunity emerges. Machine Learning automates back-office and client-facing processes.

Even though blockchains can automate many processes through smart contracts, they have limitations. Fintech companies that want to maximize their operational efficiency need to add a ML layer to their data processes. The predictive power of ML identifies issues that will need human attention. ML even performs real-time audits of the institution's processes, and thus automates regulatory compliance.

## Content Creation

Much of the written communications of financial institutions is repetitive. Advances in Natural Language Processing (NLP) and Machine Learning have made usable

machine-generated content a reality. A leading provider of such service is Narrative Science (https://narrativescience.com/).

## Money Laundering

ML offers a long-needed solution to the problem of money laundering. ML is capable of identifying patterns that are unique to money laundering. ML software results in greater detection rates, fewer false positives, fewer false negatives and easier regulatory compliance.

## Conclusion

The adoption of ML is resulting in an expanding list of use cases in finance. It should surprise no one that tech mammoths such as Google, Microsoft, Amazon, and IBM are ahead of the curve on ML. They all offer their own Machine Learning platforms, with plug-and-play solutions for many financial services.

However, as powerful as ML technology is, there is no universal financial ML solution that fits every need. Many financial engineering applications require a custom ML solution to be implemented.

# Machine Learning Path & Resources

***"I have no special talent. I am only passionately curious."***

Albert Einstein, theoretical physicist and developer of the theory of relativity

Machine Learning is a subset of Artificial Intelligence, which itself is a subset of Data Science. Data Science is a multidisciplinary field that encompasses mathematics including statistics, computer science and domain knowledge.

Non-programmers with only basic mathematics and statistics skills may at first find ML formidable. However, there are many examples of successful data scientists who have transitioned to their new careers from non-technical background. Their secret is perseverance and an inquisitive mind.

The learning path for ML could be highly subjective depending upon the person's extant skills set, aptitude and style. Fortunately, there are tons of free or inexpensive resources including MOOCs available to help along towards this exciting journey to master Machine Learning. (https://tinyurl.com/ydy8lho3)

The following paragraphs list some of these resources. This list is highly subjective, non-exhaustive and non-exclusive. The learner should complement and customise it by further research in accordance with the learner's background and aptitude.

## Programming and Data Wrangling

First, you'll need to know at least one scripting language well enough to wrangle datasets, prototype models, perform analyses and visualize results.

The leading contenders are Python or R, as they are both open-source (free), widely adopted, and supported by active communities.

Python is more common in software start-ups and large tech firms. Python tends to be more flexible because it is a general-purpose programming language. It is also better for Deep Learning and processing data.

R/RStudio is popular in research, finance, and analytics. R is a statistical programming language that has mature libraries for econometrics, statistics, and Machine Learning.

## Python Resources

### Learn Python the Hard Way (Online Book)

Recommended for beginners who want a complete course in programming with Python. (http://tinyurl.com/hcj244h)

### LearnPython.org (Interactive Tutorial)

A short, interactive tutorial for those who just need a quick way to pick up Python syntax. (http://tinyurl.com/3daozd6)

### How to Think Like a Computer Scientist (Interactive Book)

Interactive "CS 101" course taught in Python that really focuses on the art of problem solving. This goes beyond the bare minimum needed to get started. (http://tinyurl.com/psd8lnu)

## R/RStudio Resources

### R for Data Science (Online Book)

Recommended for beginners who want a complete course in data science with R. (http://tinyurl.com/yy3phusp)

### Swirl (Interactive R Package)

A nice R package that can be installed for learning the language directly from inside RStudio (the most common interface used to run R). (https://swirlstats.com/)

### Introduction to Data Science with R (Video Series)

For those who learn better by watching someone else walk through the steps. (http://tinyurl.com/y59d3d7v)

## Statistics and Probability

A strong statistics foundation helps in fully understanding Machine Learning, conditional probability, and many other core skills.

Statistics and Probability (Khan Academy): Practical introduction to statistics and probability from Khan Academy. Recommended for getting up to speed quickly. (http://tinyurl.com/jomfddt)

## Data Collection

Everything hinges on the quality and quantity of data. There are four common ways to collect data:

### Internal Data

This is proprietary data that a company collects through its operations or through partnerships with other providers. This is usually the most relevant data.

### Online Database

Online datasets allow you to prototype before investing in proprietary data.

### Application Programming Interface (API)

APIs allow you to programmatically (and legally) access datasets that other companies collect. You can find anything from Twitter feeds to weather data to financial data.

### Web Scraping

Web crawling and scraping is a powerful tool that must be used responsibly and subject to the terms of services.

## API Resources

### Python: Requests QuickStart Guide (Tutorial)

How to use the requests library to request data from API's. (http://tinyurl.com/z8kmb33)

### R: httr QuickStart Guide (Tutorial)

How to use the R httr library to request data from API's. (http://tinyurl.com/y38awazw)

## Web Scraping Resources

### Scrapy

The Python framework for web scraping. (https://scrapy.org/)

### R: rvest (Tutorial)

Web scraping with the rvest library. (http://tinyurl.com/yya6bdua)

## Structured Query Language (SQL)

SQL is the lingua franca for database management and querying. Learning SQL also gives a better understanding of relational data in general (i.e. data in "table" format), which will improve data analysis in any programming language.

### Intro to SQL by Khan Academy (Course)

Comprehensive video series that covers every important SQL topic. (http://tinyurl.com/jw4zaed)

### sqlcourse.com (Interactive Tutorial)

Great to use as a review or a quick crash course. (http://tinyurl.com/y27kw7jq)

### SQL Fundamentals (Course)

Course that covers the basics of SQL. Includes quizzes along the way to test your understanding. (http://tinyurl.com/yyv5dokj)

## Data Visualization

Data visualization is important for exploratory analysis and for communicating your insights, and no list of data science resources would be complete without this topic.

Raw data can be difficult to interpret, so you'll need to investigate trends and distributions with plots and charts.

### Data Visualization in Python (Video Series)

Tutorial on using the matplotlib library in Python. (http://tinyurl.com/y5puteag)

### Data Visualization in R (Video Series)

Tutorial on using the ggplot library in R. (http://tinyurl.com/y4aslqj5)

## Applied Machine Learning

Machine Learning is a broad umbrella term that contains many sub-tasks. In a nutshell, it is about teaching computers how to learn patterns and models from data. For some people, Machine Learning is synonymous with data science, but in reality, it is a separate field that heavily overlaps with data science.

### Machine Learning by Andrew Ng (Video Series)

This is the most popular course when it comes to understanding the theory behind machine learning. (http://tinyurl.com/y3zp26ut)

### Elements of Statistical Learning

This is one of the classic textbooks of the industry, but it requires a solid math background. (http://tinyurl.com/y3vcedx8)

### An Introduction to Statistical Learning in R

Another classic textbook that has gentler math requirements. (http://tinyurl.com/y7eg3bw8)

## Miscellaneous

### Introduction to Business Analytics (Video)

Short introduction to how businesses use analytics. (http://tinyurl.com/z9conj2)

### Predict Titanic Survival (Kaggle Competition)

Kaggle is a site that hosts data science competitions, many of which are beginner-friendly. The Titanic Survival Prediction challenge is a classic, with detailed tutorials for both Python and R. (http://tinyurl.com/hyw2lgu)

# Glossary

### Algorithm

A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

### API

Application Programming Interface (API) is a set of subroutine definitions, communication protocols, and tools for building software. In general terms, it is a set of clearly defined methods of communication among various components.

### Artificial Intelligence

The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

### Artificial Neural Network

An Artificial Neural Network (ANN) is an information processing model that is inspired by the way biological nervous systems, such as the brain, process information.

### Blockchain

A public, permanent, append-only, distributed ledger.

### Central Processing Unit

A central processing unit (CPU), also called a central processor or main processor, is the electronic circuitry within a computer that carries out the instructions of a computer program by performing the basic arithmetic, logic, controlling, and input/output (I/O) operations specified by the instructions.

### Classification

Assignment of a given instance to one of a set of classes.

### Cloud Computing

The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.

### Data Science

The field of study that combines domain expertise, programming skills, and knowledge of mathematics including statistics to extract meaningful insights from structured and unstructured data.

### Database

An organized collection of data, generally stored and accessed electronically from a computer system.

### Deep Learning

Machine Learning methods that are used to train models with several levels of abstractions from the raw data input to output.

### Ensemble Learning

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

### Graphic Processing Unit

A graphics processing unit (GPU) is a computer chip that performs rapid mathematical calculations, primarily for the purpose of rendering images. A graphic processing unit is able to render images more quickly than a central processing unit because of its parallel processing architecture, which allows it to perform multiple calculations at the same time.

### Library

In the context of a programming language such as Python, a library is a collection of packages with related functionality.

### Machine Leaning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience and examples with data without being explicitly programmed.

### Model

A template formalizing the relationship between an input and an output. It has a fixed structure but modifiable parameters.

### Module

A module in Python programming language is a .py file that defines one or more function/classes which are intended to be reused in different codes of program.

### NLP

Natural Language Processing (NLP) refers to computer methods used to process human language and is also called computational linguistics.

### Package

A Python package refers to a directory (collection) of Python module(s). This feature comes in handy for organizing modules of one type at one place.

### Parallel Distributed Processing

A computational paradigm where the task is divided into small concurrent tasks, each of which can be run on a different processor and thus reducing computation time.

### Reinforcement Learning

It refers to goal-oriented algorithms, which learn how to attain a complex objective (goal) or maximize along a particular dimension over many steps.

### Semi-Supervised Learning

It is a class of Machine Learning tasks and techniques that also make use of unlabeled data for training (typically, a small amount of labeled data with a large amount of unlabeled data).

### SQL

Structured Query Language (SQL) is a standard language for storing, updating, manipulating and retrieving structured data in databases.

### Supervised Learning

Machine learning method used to train a model using a labeled dataset.

### Test Data

A set of observations used at the end of model training and validation to assess the predictive power of the model.

### Training Data

A set of observations used to generate Machine Learning models.

### Unsupervised Learning

Machine Learning method of training a model to find patterns in an unlabeled dataset (e.g. clustering).

### Web Scrapping

Action by software that automatically surfs the web and extracts information from web pages.

### XBRL

eXtensible Business Reporting Language (XBRL) is a type of XML (eXtensible Markup Language), which is a specification that is used for organizing and defining data. XBRL uses tags to identify each piece of financial data, which then allows it to be used programmatically by an XBRL-compatible program.

# About CPA New Brunswick

CPA New Brunswick is a professional organization representing more than 2,800 active and retired members and 300 future CPAs in New Brunswick.

Each provincial CPA organization is a member of the Chartered Professional Accountants of Canada (CPA Canada), which represents more than 200,000 professional accountants across Canada and Bermuda, making it one of the top five accounting designations in the world.

Under the Chartered Professional Accountants Act (Chapter 28), CPA New Brunswick is responsible for regulating the professional development of its members, and the protection of the public through its ethical standards and discipline process. CPA New Brunswick is also responsible for the training and certification of CPA Candidates.

CPAs work in every sector of New Brunswick. They are involved in a wide range of complex disciplines – from financial reporting, finance, and taxation; to strategy and governance, assurance, performance management, and information technology – and they volunteer their time and expertise for numerous community projects and charitable organizations.

Approximately 20% of CPAs work in public practice, but the balance work in diverse sectors of industry including government, education, and the not-for-profit sector. They offer a strong set of accounting and managerial skills required for today's complex and evolving environment.

These professional accountants are highly attractive to employers and recruiters for their solid training and expertise that contribute to improved efficiency and growth.

Throughout their professional careers, CPAs are subject to ongoing regulation to protect the public interest. Among other things, the Act provides for the regulation of members and firms, while the By-Laws assist to govern the operations of the Chartered Professional Accounting profession in New Brunswick.

CHARTERED PROFESSIONAL ACCOUNTANTS NEW BRUNSWICK

INFORMATION TECHNOLOGY | STAKEHOLDER AUDIENCE