Last Update: 18 October 2021

Introduction of Big Data and AI

Yoshimasa Satoh, CFA
https://www.linkedin.com/in/yoshimasa-satoh-cfa-84b6b92b/

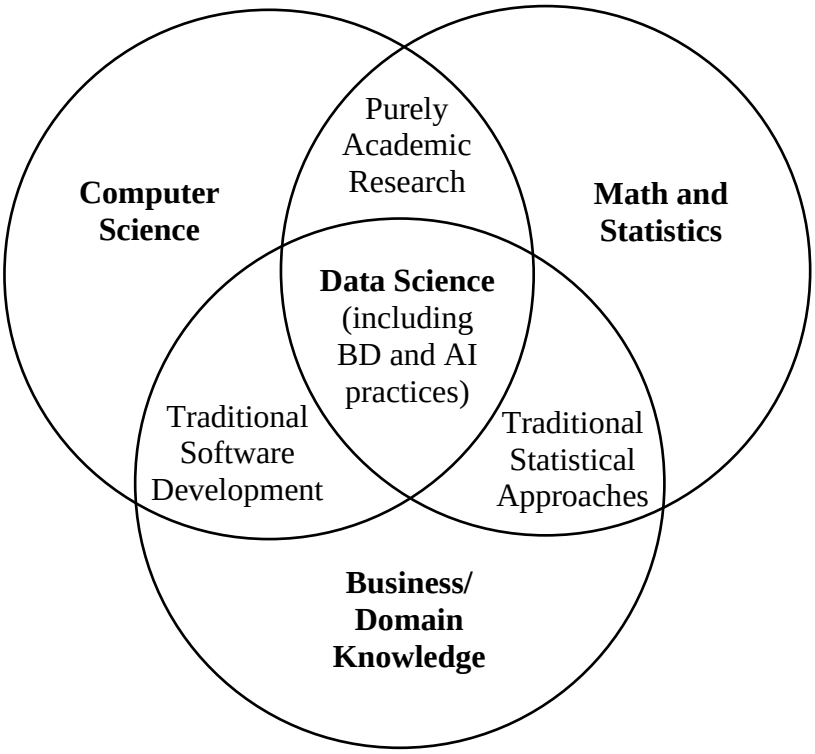
This paper is for informational and educational purposes only.

Last Update: 18 October 2021

1. Data Science

Generally speaking, data science is an inter-disciplinary field that overlaps with math and statistics, computer science, and business/domain knowledge as in Figure 1; data science can be used to extract insights from a wide variety of data with the help of mathematical/statistical techniques, computer algorithms, and clear goals of business/domain. Big Data (BD) and Artificial Intelligence (AI) practices are key areas of Data Science.

Figure 1. Venn Diagram for Data Science



2. Big Data (BD) and Artificial Intelligence (AI) as an Important Part of the Fourth Industrial Revolution

The steam engine in the first industrial revolution, electricity in the second, and internet technology in the third, fundamentally changed human history. The fourth industrial "digital" revolution has been occurring and is characterized by a fusion of technologies that is blurring the lines between the physical, digital, and biological spheres.[1] The fourth industrial revolution has the similar transformative potential as the precedent three industrial revolutions; actually, some experts say it's much higher than the three industrial revolutions in the past. As a huge part of the fourth industrial digital revolution, BD and AI are of extraordinary importance.

A technological advance can impact — even transform — the global economy let alone a particular sector. Financial services industry is no exception.

When it comes to BD and AI in the financial services industry, some people anticipate a rosy future with various new sources of productivity growth, cost containment, and ultimately higher corporate profits and/or personal compensations. Others worry about the jobs that might be lost to machines mainly through process automation and even replacing highly intellectual activities.

The digital revolution raises a series of critical questions. How financial, technology, legal/compliance, and corporate management professionals address them will go a long way in determining who will successfully adapt and who might be rendered obsolete.[2]

3. Big Data (BD)

BD has four Vs as in Table 1.

Table 1. Big Data (BD) and four Vs

#	V	Description
1	Volume	- Data volumes are growing to far larger sizes as more data are being generated, captured, processed, stored, and utilized.
2	Variety	- BD is structured/semi-structured/unstructured data from traditional/alternative sources.
3	Velocity	- BD is high frequency data on a real-time or near-real-time basis.
4	Veracity	- BD needs to be accurate to be utilized.

In general, it's often based on behavior and interactions of humans and computer algorithms. To establish causal connections between an explained variable and multiple explanatory variables, which have repetitive patterns in cross-sectional data and/or time-series data, you need sophisticated knowledge and experience in raw data processing, AI theories and techniques, implementation of AI algorithms, computer utilization, and analyses of the results, e.g., predicted outcomes. Time-series data can decomposed into seasonality (i.e., cyclical patterns), structural trends, structural patterns after removing seasonality and trend, and other unexplained noises (residual).

Sources of BD in general society include, but not limited to, social media, web-scraped data, crowd-sourced information such as surveys, depersonalized credit card and point-of-sale details, and collated web search trends. Earnings conference call recordings and transcripts, and satellite images (to count the number of cars in a parking lot of a store for sales estimation) are also famous and popular examples with the press, but any (selected, collected, and pre-processed) BD has to be proved as a source of insights, not a pile of garbage.

BD had been hard and/or costly to collect, clean, store, and analyze; however, by decreased costs in cloud computing and harnessing advances in AI, it has been looking increasingly likely that we can find new and more accurate relationships within BD, which often have high dimensionality and non-linearity. Namely, a explanatory variable interacts with other many explanatory variables (i.e., high dimensionality) and its relationships (between an explained variable and an explanatory variable, or explanatory variable vs another explanatory variable) are non-linear.

It should be noted that AI, "machines," are only as intelligent as the data it learns from, and humans involved with that intellectual activities. Raw data itself, data cleaning, and knowledge/experience about how data is generated, collected, processed, stored, and analyzed, do matter. Additionally, understanding of an objective of the analysis and even subjective expert human judgement based on knowledge and experience are critical as well.

Nothing works off the shelf. It is a common misperception that by just putting a lot of raw and dirty data, fancy computers, and smart people together, you'll be able to extract useful signals instantly. The cold reality is that it does not work that easy. Although BD might provide new opportunities in the long run, BD often does not have enough data history so far; many people still do not have sufficient knowledge and experience either. Starting earlier and pursuing sustained efforts in BD utilization for focus areas with the help of AI techniques are keys. It's never too late to launch a BD and AI initiative and keep it to a manageable level.

4.1. Artificial Intelligence (AI)

Simply put, AI is automated machines that can [1] learn from large amount of data (often regarded as BD), [2] determine underlying structures/relationships in the data, [3] predict results for new unseen data, and then [4] act for themselves basically with no or a little human intervention. AI can make their own decisions when it faces with a new situation, in the same (or similar) way that humans can.

In supervised learning of AI, we provide data and "expected results" to AI and then AI provides "rules" (or a structural relationship amongst data elements) back to us. In the case of unsupervised learning of AI, data

only is provided to a machine and the machine finds plausible results on their own in addition to a structure of data elements. These rules derived by AI tend to be something humans cannot easily and intuitively understand.

Both traditional statistical techniques and AI techniques analyze observations to reveal some underlying rules; however, they usually diverge in their assumptions, techniques, and even terminologies.

If we look at traditional statistical approaches, it relies on foundational assumptions and explicitly structured models. That is, observed samples are assumed to be drawn from a specified underlying probability distribution which can be explicitly and descriptively modelled and relatively straightforward for humans to intuitively understand. We traditionally provided observation data and explicit rules to old-fashioned machines so that we could achieve results. Differently put, a dependent/explained variable (y) is considered to be explained by a pre-defined descriptive equation with independent/explanatory variables (X's) and related coefficients. This basically deductive, a priori restrictive assumptions can fail to describe a reality due to a strictly and incorrectly pre-specified data structure model.

On the contrary, AI techniques are used to extract insights from large amounts of data with no such restrictions. One of the most important goals of AI techniques is to streamline decision-making processes by generalizing (i.e., "learning") from known training data to determine an underlying structure in the data, that is, a relationship between a target (y) and features (X's), in a more flexible manner. The emphasis is on the ability of the algorithm to build the data structure model and make predictions ($\hat{y}$) based on the model with no or limited help from humans. After explicitly selecting an AI technique, which can be either deductive (supervised learning) or inductive (unsupervised learning), the selected AI technique itself finds patterns from training data, build a model, apply the model to unseen validation data/test data for hyperparameter[3] setting/target forecasting, respectively, and then the loop goes on until it reveals an optimal structure under a given set of conditions. One of the weaknesses built into AI techniques is a black-box problem; cause and effect links between a target and features tend to be complicated and implicit. Thus, there are some cases that are almost impossible to explain and visualize in a manner that humans can intuitively understand. More troublingly, mere correlations are most commonly mixed up with causations; what you do really want to know are cause-and-effect relationships, not merely calculated nominal correlations.

4.2. Machine Learning

Machine Learning (ML) is a subset of AI, but many people often equate ML with AI nowadays. ML algorithms use statistics to find (especially non-linear) patterns in massive amounts and various types of data. They then can use those patterns to provide predictions on a target ($\hat{y}$) based on features (X's), and then make decisions. ML techniques can be more effective than linear regressions, a famous category of traditional statistical methods, in the presence of multi-collinearity where features are correlated each other.

Within ML, supervised learning infers patterns between a set of inputs/features (X's) and the actual output/target (y) within training data. The inferred pattern is then used to map a given input set (X's) into a predicted output ($\hat{y}$). Supervised learning requires a labeled data set, one that contains matched sets of observed inputs (X's) and the associated output (y).

On the contrary, unsupervised learning does not make use of a labeled data; there are inputs (X's) that are used for analysis without any target (y) being supplied. Algorithms seek to discover a structure within the data themselves and predict a target ($\hat{y}$) accordingly.

Target (y) variables can be continuous for regression (i.e., numerical prediction) or discrete for classification (including clustering). The former is basically supervised learning only while the latter can be supervised or unsupervised (e.g., clustering).

4.3. Deep Learning

Deep Learning (DL) is a subset of ML. DL is based on artificial neural networks, which is a type of learning modeled by reference to human brains. DL is the basis of many major breakthroughs, including natural language processing (NLP), speech recognition, voice synthesis, image classification, face recognition, hyper-realistic photo generations, and AlphaGo by DeepMind, which plays the board game Go. It should be noted that this is just a tiny fraction of what AI could be in the long term.

4.4. Other Miscellaneous Topics of AI

If we look back into history, the first AI boom is in 1950s-1960s. It is the age of reasoning as search. The second boom is in 1980s and called the age of knowledge representation. Appearance of expert systems that are capable of reproducing simple rule-based human decision making was symbolic. We are in the third boom after University of Toronto developed and announced their DL technique back in 2006.

Furthermore, there is a grand idea to develop something resembling (or even enhancing) human intelligence, which is often referred to as artificial general intelligence, or AGI. Some experts believe that ML (especially DL) will eventually get us to AGI with enough data, computing power, and theoretical advance in AI techniques. However, most would agree there are many and big missing pieces and it's still a long way off. AI may have mastered Go, but in other ways it is still much dumber than a toddler. Most importantly, there are many, not only technical, but also legal, ethical, and philosophical issues for AI-based decision making processes and responsibility for the results.

In that sense, AI as a whole is still aspirational, and its definition is constantly evolving. What would have been considered AI in the past may not be considered AI today. Because of this, the boundaries of AI can get really confusing, and the term often gets mangled to include any kind of computer programs.

Even so, if we narrow down and clearly define an objective, and make the best use of properly prepared data and an appropriate AI technique while understanding limitation, then it is likely that we can acquire insights from data for the benefit of our business/domain. Most importantly, it is highly recommended to begin with setting a clear business goal at the very beginning so that we can avoid fruitless efforts and making big mistakes. If you confound means with the end, it is unlikely to succeed; typical mistakes are as follows.

"Is there anything we can do with existing data in our own databases and spreadsheets as they are? It is kind of messy data, but I do not want to clean up as pre-processing is burdensome."
"We need to buy or implement AI tools first because we should start using it asap (no matter what a business goal is.) We all are in the throes of the fourth industrial, digital revolution backed by AI, aren't we? We must hurry ourselves."
"I studied a new and innovative AI theory. It looks cool. I want to give it a try as soon as possible."
"Look at those great AI-backed analysis and visualization tools!"

5. Challenges and Pros/Cons of Artificial Intelligence (AI) Utilization

There are many challenges of AI in general as described in Table 2.

Table 2. Challenges of AI in general

Interpretability, Accountability, Auditability	- It's a "black box[4]" and also affected by mutual interactions amongst data elements. Both outputs (i.e., predicted target values) and forecasting processes inside of a machine tend to be hard for humans to intuitively understand. "Use the simplest tool that does solve a problem" like the Occam's razor principle might be a quick countermeasure to overcome it. - More dispersion between growing computing power and theories lagging behind it are also an issue from an accountability perspective.
Overfitting and Generalization	- We need to strike an appropriate balance between overfitting and underfitting. - Overfitted models incorporate noises, random fluctuations, and/or spurious correlations without cause-and-effect links of the training dataset into its learned relationship in the trained model; in contrast, the relationship does not necessarily

	apply to validation and test dataset, which are used for hyper-parameter setting and final testing, respectively. It might have perfect hindsight for training dataset because of its too much complexity, but not generalized foresight for validation/test dataset. This could be an issue not only in AI utilization, but also from raw data collection and processing prior to AI method applications.
Data acquisition and computational costs	- BD and AI often require huge computational, time, and monetary costs.
Security	- Machines and data can be misused by malicious manipulation, e.g., data falsification.
Confidentiality	- Confidentiality, especially privacy protection in relation to personal data on the retail business side, has to be provided. A trade secret in institutional business has to be secured as well.
Ethics, Judgement	- If machines learn some humans' unfair/unethical/illegal behavior, then the machines might think it's normal and acceptable. Who decides what's right or wrong, and how?

Note: This table is not necessarily mutually exclusive, collectively exhaustive.

All human jobs are not necessarily replaced by machines at least partially due to these reasons. Actually, humans equipped with machines are expected to replace others, by using a combination of AI and human intelligence (HI). It is likely AI can transform our business, but it is unlikely the mass extinction event for many humans might fear. Rather, those human teams that successfully adapt to the evolving landscape will persevere. Those that don't are likely to render themselves obsolete.

If we take a closer look at AI, there are pros and cons as in Table 3.

Table 3. Pros and Cons for AI

Pros	- AI can be used even for non-continuous data (outliers), which usually reduce the accuracy of traditional statistical methods. - More precautionary measures can be implemented with AI algorithms for unusual data. For instance, some trading activities at banks are very difficult to define and capture by clearly written surveillance rules. Especially, Deep Neutral Network (DNN) can express non-smooth function and it could have better performance than other basic methods. - A combination of traditional interpretable/accountable statistical methods and sophisticated AI techniques can be one countermeasure for a black-box problem of AI. - Ultimately, market crashes and/or a widespread financial crisis might be mitigated (or prevented before it happens) if appropriate precautionary measure are in place across organizations in a comprehensive manner. It might sound too naive though.
Cons	- It still tends to be a black box by its nature; that means there is no clear intuitively-understandable relationship between inputs (explanatory variables) and outputs (explained variable). It's problematic if you are not sure about an AI method used and why a certain result is presented. Achieving deep knowledge and experience in AI matters. - AI can be misused to hide hostile attacks; say, a rogue trader might be able to deceive trade surveillance machines by outsmarting explicit rules/implicit principles and implemented algorithms. - Algorithms, objective facts, work as programmed and do not tell a lie; however, rogue traders misuse algorithms to make their own trading activities almost invisible. - Actual data of market abuse practices is small or "tiny" data in the first place. It makes both [1] splitting data into training, validation, and test datasets and [2] statistically significant tests tough. Because of this and dynamically changing unstable environment in financial markets, an S/N (Signal-to-Noise) ratio tends to be lower; namely, there could be many and big noises and false positives/negatives. Suspicious order and/or trade data within a very short time horizon is even more difficult to detect. Generating appropriate artificial datasets is a key. - One of the most difficult aspects of applying AI, including ML and DL, in financial markets, is intractableness of taking trial and error approaches. It is not like voice recognition, natural language processing, automated driving, and so forth that can be tested over and over again in the real world.[5]

Note: This table is not necessarily mutually exclusive, collectively exhaustive.

Conclusion

After clearly setting a business objective, a target (distinct numbers/words for classification or continuous numbers for a regression problem) and features within raw data, an AI method, and hyperparameters, should be carefully chosen, given, and trained/validated/tested.[6] Those data science-boosted practices can serve for sophisticated, proactive, and precautionary business practices.

Acknowledgements

The author is immensely grateful to all the contributors for their comments on an earlier version of the paper. However, any potential errors are the author's own, and should not tarnish the reputations of these esteemed professionals.

Notes

The material presented is for informational and educational purposes only. The views expressed in this paper are the views solely of the author and are subject to change; moreover, the views do not necessarily represent the official views of the author's employer.

1. See World Economic Forum (2016).

2. There is the concept ABCDS of financial technology.
   A: Artificial Intelligence
   B: Blockchain
   C: Cloud Computing
   D: Big Data
   S: Information Security
   ABCDS may very well transform the financial services industry. Blockchain, cloud computing (e.g., AWS, which stands for Amazon Web Services), and information security, tend to be closer to the technology infrastructure layer, rather than the application layer, and are beyond the scope of this paper.

3. A hyperparameter is a special parameter whose value must be set by a human, such as, data scientist, AI engineer, or researcher before getting learning started, and then used to control the learning process. It requires discretionary human expert judgement based on knowledge, experience, and constraint conditions in business processes, data, AI methods, the amount of time spent, and the level of precision desired. Two of representative hyperparameters in Machine Learning (ML), which is a subset of AI, are learning rate and mini-batch size. The former controls how quickly the model is optimized to training data. The latter is the number of training examples utilized in one iteration of learning.
   By contrast, the values of other non-hyper parameters (typically weights of features in a model, i.e., coefficients) are derived via training of AI in an objective manner, not by human's subjective judgement.

4. Concerning a black box problem, why don't humans understand what AI (especially ML including DL techniques) implies? There are two major reasons.
   First, there are too many features as inputs and arbitrarily set hyperparameters for humans to understand them intuitively. It's also very difficult to narrow down features since interactions amongst them are complicated. Simply put, it's a quantitative challenge.
   Second, AI tends to be complicated (e.g., deeper in DL) and there are so many data transformations; that makes outputs non-intuitive for humans while only machines themselves can interpret it. Simply speaking, it's a qualitative challenge.

5. If we can develop real-world-like virtual markets in the future, then trial and error approaches could be employed. If that's the case, then Reinforcement Learning (RL) is one of options. RL models learn from interacting with themselves (and/or data generated by the same algorithms.) It also uses neural networks in DL, which can be either supervised or unsupervised.

RL made headlines in 2017 when DeepMind's AlphaGo program beat the reigning world champion at the ancient game of Go. The RL algorithm involves an agent that should perform actions that will maximize its rewards over time, taking into consideration the constraints of its environment.
RL with unsupervised learning has neither direct labeled data for each observation nor instantaneous feedback. With RL, the algorithm needs to observe its environment, learn by testing new actions (some of which may not be immediately optimal), and reuse its previous experiences. The learning subsequently occurs through millions of trials and errors. Academics and practitioners are applying RL in a similar way in investment strategies where the agent could be a virtual trader who follows certain trading rules (actions) in a specific market (an environment) to maximize its profits (rewards). However, the success of RL in dealing with the complexities of financial markets is still an open question.

6. Even if selected carefully, each single model will have a certain error rate and will make noisy predictions. However, if you use ensemble learning (EL) to combine the predictions from various models and take the average result of many predictions, it leads to a reduction in noise and thus more accurate predictions.
EL typically produces more accurate and more stable predictions than the best single model. In fact, in many prestigious ML competitions, an ensemble method is often the winning solution so far.
EL can be achieved by an aggregation of either heterogeneous learners—different types of algorithms combined with a voting classifier—or homogenous learners—a combination of the same algorithm but using different training data based on the bootstrap aggregating (i.e., bagging) technique.

References

World Economic Forum. 2016. "The Fourth Industrial Revolution: what it means, how to respond." https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Robert Kissell, PhD, and Barbara J. Mack. 2019. "Fintech in Investment Management." *CFA Institute Refresher Reading* 2022 CFA Program Level I Reading 55.
https://www.cfainstitute.org/membership/professional-development/refresher-readings/fintech-investment-management
Note: This reading is available to CFA Institute members only.

Kathleen DeRose, CFA, and Christophe Le Lannou. 2020. "Machine Learning." *CFA Institute Refresher Reading* 2022 CFA Program Level II Reading 4.
https://www.cfainstitute.org/membership/professional-development/refresher-readings/machine-learning
Note: This reading is available to CFA Institute members only.