The Role of Big Data, Machine Learning, and AI in Assessing Risks: a Regulatory Perspective

Scott W. Bauguess
Acting Director and Acting Chief Economist, Division of Economic and Risk Analysis

New York, New York

SEC Keynote Address:
OpRisk North America 2017

Thank you, Alexander [Campbell] for the introduction.

Thanks also to Genevieve Furtado and the other conference organizers for the invitation to speak here today, at the 19th Annual Operational Risk North America Conference. I understand that this is the Champagne Keynote address. Given that title, I feel obligated as an economist to share with you the reported last words of John Maynard Keynes – the father of modern macroeconomics: "I should have drunk more champagne." I hope my words here today do not inspire a similar sentiment. And finally, I must remind you that the views that I express today are my own and do not necessarily reflect the views of the Commission or its staff.[1]

My remarks this afternoon will center on a technology topic that is encroaching on many aspects of our lives and increasingly so within financial markets: Artificial Intelligence. Perhaps better known by its two-letter acronym "AI," artificial intelligence has been the fodder of science fiction writing for decades. But the technology underlying AI research has recently found applications in the financial sector – in a movement that falls under the banner of "Fintech." And the same underlying technology [machine learning and AI] is fueling the spinoff field of "Regtech," to make compliance and regulatory-related activities easier, faster, and more efficient.

This is the first time that I have addressed the emergence of AI in one of my talks. But I have spoken previously on the two core elements that are allowing the world to wonder about its future: big data and machine learning.[2] Like many of your institutions, the Commission has made recent and rapid advancements with analytic programs that harness the power of big data. They are driving our surveillance programs and allowing innovations in our market risk assessment initiatives. And the thoughts I'm about to share reflect my view on the promises – and also the limitations – of machine learning, big data, and AI in market regulation.

Perhaps a good place to begin is with a brief summary of where we were, at the Commission, 2 years ago. I remember well, because it was then that I was invited to give a talk at Columbia University on the role of machine learning at the SEC. I accepted the invitation with perhaps less forethought than I should have had. I say this because I soon found myself googling the definition of machine learning. And the answers that Google returned—and I say answers in plural, because there seem to be many ways to define it—became the first slide of that presentation.[3]

The Science of Machine Learning and the Rise of Artificial Intelligence

Most definitions of machine learning begin with the premise that machines can somehow learn. And the central tenets of machine learning, and the artificial intelligence it implies, have been around for more than a half a century. Perhaps the best known, early application was in 1959, when Arthur Samuel, an IBM scientist, published a

Electronic copy available at: https://ssrn.com/abstract=3226514

solution to the game of checkers. For the first time, a computer could play checkers against a human and win.[4] This is now also possible with the board game "Go," which has been around for 2,500 years and is purported to be more complicated and strategic than Chess. Twenty years ago, it was widely believed that a computer could never defeat a human in a game of "Go." This belief was shattered in 2016, when AlphaGo, a computer program, took down an 18-time world champion in a best-of-seven match.[5] The score: 4 to 1.

Other recent advancements in the area of language translation are equally, if not more, impressive. Today, if the best response to my question on the definition of machine learning is in Japanese, Google can translate the answer to English with an amazing degree of clarity and accuracy. Pull out your smart phone and try it. Translate machine learning into Japanese. Copy and paste the result into your browser search function. Copy and paste the lead paragraph of the first Japanese language result back into Google Translate. The English language translation will blow your mind. What would otherwise take a lifetime of learning to accomplish comes back in just a few seconds.

The underlying science is both remarkable and beyond the scope of this talk.[6] (Not to mention my ability to fully explain it.) But it is not too difficult to understand that the recent advancements in machine learning are shaping how AI is evolving. Early AI attempts used computers to mimic human behavior through rules-based methods, which applied logic-based algorithms that tell a computer to "do this if you observe that." Today, logic-based machine learning is being replaced with a data-up approach. And by data-up, I mean programming a computer to learn directly from the data it ingests. Using this approach, answers to problems are achieved through recognition of patterns and common associations in the data. And they don't rely on a programmer to understand why they exist. Inference, a prerequisite to a rule, is not required. Instead, tiny little voting machines, powered by neural networks, survey past quantifiable behaviors and compete on the best possible responses to new situations.

If you want a tangible example of this, think no further than your most recent online shopping experience. Upon the purchase of party hats, your preferred retailer is likely to inform you that other shoppers also purchased birthday candles. Perhaps you need them too? Behind this recommendation is a computer algorithm that analyzes the historical purchasing patterns from you and other shoppers. From this, it then predicts future purchasing-pair decisions. The algorithm doesn't care why the associations exist. It doesn't matter if the predictions don't make intuitive sense. The algorithm just cares about the accuracy of the prediction. And the algorithm is continually updating the predictions as new data arrives and new associations emerge.

This data-driven approach is far easier to apply and is proving in many cases to be more accurate than the previous logic-based approaches to machine learning. But how does it help a market regulator to know that purchasers of protein powder may also need running shoes?

The simple, and perhaps obvious, answer is that regulators can benefit from understanding the likely outcomes of investor behaviors. The harder truth is that applying machine learning methods is not always simple. Outcomes are often unobservable. Fraud, for example, is what social scientists call a latent variable. You don't see it until it's found. So, it is more challenging for machine learning algorithms to make accurate predictions of possible fraud than shopping decisions, where retailers have access to full transaction histories—that is, complete outcomes for each action. The same is true for translating languages; there is an extremely large corpus of language-pair translations for an algorithm to study and mimic.

Two years ago, tackling these types of issues at the Commission was still on the horizon. But a lot of progress has been made since then, and machine learning is now integrated into several risk assessment programs—sometimes in ways we didn't then envision. I'm about to share with you some of these experiences. But let me preview now, that while the human brain will continue to lose ground to machines, I don't believe it will ever be decommissioned with respect to the regulation of our financial markets.

The Rise of Machine Learning at the Commission

Electronic copy available at: https://ssrn.com/abstract=3226514

Let me start by giving you some background on staff's initial foray into the fringes of machine learning, which began shortly after the onset of the financial crisis. That is when we first experimented with simple text analytic methods. This included the use of simple word counts and something called regular expressions, which is a way to machine-identify structured phrases in text-based documents. In one of our first tests, we examined corporate issuer filings to determine whether we could have foreseen some of the risks posed by the rise and use of credit default swaps [CDS] contracts leading up to the financial crisis. We did this by using text analytic methods to machine-measure the frequency with which these contracts were mentioned in filings by corporate issuers. We then examined the trends across time and across corporate issuers to learn whether any signal of impending risk emerged that could have been used as an early warning.

This was a rather crude proof-of-concept. And it didn't work exactly as intended. But it did demonstrate that text analytic methods could be readily applied to SEC filings. Our analysis showed that the first mention of CDS contracts in a Form 10-K was by three banks in 1998. By 2004, more than 100 corporate issuers had mentioned their use. But the big increase in CDS disclosures came in 2009. This was, of course, after the crisis was in full swing. And identification of those issues by the press wasn't much earlier. We analyzed headlines, lead paragraphs, and the full text of articles in major news outlets over the years leading up to the financial crisis and found that robust discussions of CDS topics did not occur until 2008. During that year, we found a ten-fold increase in CDS articles relative to the prior year.

Use of Natural Language Processing

Even if the rise in CDS disclosure trends had predated the crisis, we still would have needed to know to look for it. You can't run an analysis on an emerging risk unless you know that it is emerging. So this limitation provided motivation for the next phase of our natural language processing efforts. This is when we began applying topic modeling methods, such as latent dirichlet allocation to registrant disclosures and other types of text documents. LDA, as the method is also known, measures the probability of words within documents and across documents, in order to define the unique topics that they represent.[7] This is what the data scientist community calls "unsupervised learning." You don't have to know anything about the content of the documents. No subject matter expertise is needed. LDA extracts insights from the documents, themselves using the data-up approach to define common themes – these are the topics – and report on where, and to what extent, they appear in each document.

One of our early topic modeling experiments analyzed the information in the tips, complaints, and referrals (also referred to as TCRs) received by the SEC. The goal was to learn whether we could classify themes directly from the data itself and in a way that would enable more efficient triaging of TCRs. In another experiment, DERA – the Division of Economic and Risk Analysis – research staff examined whether machine learning could digitally identify abnormal disclosures by corporate issuers charged with wrongdoing. DERA research staff found that when firms were the subject of financial reporting-related enforcement actions, they made less use of an LDA-identified topic related to performance discussion. This result is consistent with issuers charged with misconduct playing down real risks and concerns in their financial disclosure.[8]

These machine learning methods are now widely applied across the Commission. Topic modeling and other cluster analysis techniques are producing groups of "like" documents and disclosures that identify both common and outlier behaviors among market participants. These analyses can quickly and easily identify latent trends in large amounts of unstructured financial information, some of which may warrant further scrutiny by our enforcement or examination staff.

Moreover, working with our enforcement and examination colleagues, DERA staff is able to leverage knowledge from these collaborations to train the machine learning algorithms. This is referred to as "supervised" machine learning. These algorithms incorporate human direction and judgement to help interpret machine learning outputs. For example, human findings from registrant examinations can be used to "train" an algorithm to understand what pattern, trend, or language in the underlying examination data may indicate possible fraud or misconduct. More broadly, we use unsupervised algorithms to detect patterns and anomalies in the data, using

Electronic copy available at: https://ssrn.com/abstract=3226514

nothing but the data, and then use supervised learning algorithms that allow us to inject our knowledge into the process; that is, supervised learning "maps" the found patterns to specific, user-defined labels. From a fraud detection perspective, these successive algorithms can be applied to new data as it is generated, for example from new SEC filings. When new data arrives, the trained "machine" predicts the current likelihood of possible fraud on the basis of what it learned constituted possible fraud from past data.

An Example of Machine Learning To Detect Potential Investment Adviser Misconduct

Let me give you a concrete example in the context of the investment adviser space. DERA staff currently ingests a large corpus of structured and unstructured data from regulatory filings of investment advisers into a Hadoop computational cluster. This is one of the big data computing environments we use at the Commission, which allows for the distributed processing of very large data files. Then DERA's modeling staff takes over with a two-stage approach. In the first, they apply unsupervised learning algorithms to identify unique or outlier reporting behaviors. This includes both topic modeling and tonality analysis. Topic modeling lets the data define the themes of each filing. Tonality analysis gauges the negativity of a filing by counting the appearance of certain financial terms that have negative connotations.[9] The output from the first stage is then combined with past examination outcomes and fed into a second stage [machine learning] algorithm to predict the presence of idiosyncratic risks at each investment adviser.

The results are impressive. Back-testing analyses show that the algorithms are five times better than random at identifying language in investment adviser regulatory filings that could merit a referral to enforcement. But the results can also generate false positives or, more colloquially, false alarms. In particular, identification of a heightened risk of misconduct or SEC rule violation often can be explained by non-nefarious actions and intent. Because we are aware of this possibility, expert staff knows to critically examine and evaluate the output of these models. But given the demonstrated ability of these machine learning algorithms to guide staff to high risk areas, they are becoming an increasingly important factor in the prioritization of examinations. This enables the deployment of limited resources to areas of the market that are most susceptible to possible violative conduct.

The Role of Big Data

It is important to note that all of these remarkable advancements in machine learning are made possible by, and otherwise depend on, the emergence of big data. The ability of a computer algorithm to generate useful solutions from the data relies on the existence of a lot of data. More data means more opportunity for a computer algorithm to find associations. And as more associations are found, the greater the accuracy of predictions. Just like with humans, the more experience a computer has, the better the results will be.

This trial-and-error approach to computer learning requires an immense amount of computer processing power. It also requires specialized processing power, designed specifically to enhance the performance of machine learning algorithms. The SEC staff is currently using these computing environments and is also planning to scale them up to accommodate future applications that will be on a massive scale. For instance, market exchanges will begin reporting all of their transactions through the Consolidated Audit Trail system, also known as CAT, starting in November of this year.[10] Broker-dealers will follow with their orders and transactions over the subsequent 2 years. This will result in data about market transactions on an unprecedented scale. And, making use of this data will require the analytic methods we are currently developing to reduce the enormous datasets into usable patterns of results, all aimed to help regulators improve market monitoring and surveillance.

We already have some experience with processing big transaction data. Using, again, our big data technologies, such as Hadoop computational clusters that are both on premises and available through cloud services, we currently process massive datasets. One example is the Option Pricing Reporting Authority data, or OPRA data. To help you grasp the size of the OPRA dataset, one day's worth of OPRA data is roughly two terabytes. To illustrate the size of just one terabyte, think of 250 million, double-sided, single-spaced, printed pages. Hence, in this one

Electronic copy available at: https://ssrn.com/abstract=3226514

dataset, we currently process the equivalent of 500 million documents each and every day. And we reduce this information into more usable pieces of information, including market quality and pricing statistics.

However, with respect to big data, it is important to note that *good* data is better than *more* data. There are limits to what a clever machine learning algorithm can do with unstructured or poor-quality data. And there is no substitute for collecting information correctly at the outset. This is on the minds of many of our quant staff. And it marks a fundamental shift in the way the Commission has historically thought about the information it collects. For example, when I started at the Commission almost a decade ago, physical paper documents and filings dominated our securities reporting systems. Much of it came in by mail, and some [documents] still come to us in paper or unstructured format. But this is changing quickly, as we are continuing to modernize the collection and dissemination of timely, machine-readable, structured data to investors.[11]

The staff is also cognizant of the need to continually improve how we collect information from registrants and other market participants, whether it is information on security-based swaps, equity market transactions, corporate issuer financial disclosures, or investment company holdings. We consider many factors, such as the optimal reporting format, frequency of reporting, the most important data elements to include, and whether metadata should be collected by applying a taxonomy of definitions to the data. We consider these factors each and every time the staff makes a recommendation to the Commission for new rules, or amendments to existing rules, that require market participant or SEC-registrant reporting and disclosures.

The Future of Artificial Intelligence at the Commission

So, where does this leave the Commission with respect to all of the buzz about artificial intelligence?

At this point in our risk assessment programs, the power of machine learning is clearly evident. We have utilized both machine learning and big data technologies to extract actionable insights from our massive datasets. But computers are not yet conducting compliance examinations on their own. Not even close. Machine learning algorithms may help our examiners by pointing them in the right direction in their identification of possible fraud or misconduct, but machine learning algorithms can't then prepare a referral to enforcement. And algorithms certainly cannot bring an enforcement action. The likelihood of possible fraud or misconduct identified based on a machine learning predication cannot – and should not – be the sole basis of an enforcement action. Corroborative evidence in the form of witness testimony or documentary evidence, for example, is still needed. Put more simply, human interaction is required at all stages of our risk assessment programs.

So while the major advances in machine learning have and will continue to improve our ability to monitor markets for possible misconduct, it is premature to think of AI as our next market regulator. The science is not yet there. The most advanced machine learning technologies used today can mimic human behavior in unprecedented ways, but higher-level reasoning by machines remains an elusive hope.

I don't mean for these remarks to be in any way disparaging of the significant advancements computer science has brought to market assessment activities, which have historically been the domain of the social sciences. And this does not mean that the staff won't continue to follow the groundbreaking efforts that are moving us closer to AI. To the contrary, I can see the evolving science of AI enabling us to develop systems capable of aggregating data, assessing whether certain Federal securities laws or regulations may have been violated, creating detailed reports with justifications supporting the identified market risk, and forwarding the report outlining that possible risk or possible violation to Enforcement or OCIE staff for further evaluation and corroboration.

It is not clear how long such a program will take to develop. But it will be sooner than I would have imagined 2 years ago. And regardless of when, I expect that human expertise and evaluations always will be required to make use of the information in the regulation of our capital markets. For it does not matter whether the technology detects possible fraud, or misconduct, or whether we train the machine to assess the effectiveness of our

Electronic copy available at: https://ssrn.com/abstract=3226514

regulations – it is SEC staff who uses the results of the technologies to inform our enforcement, compliance, and regulatory framework.

Thank you for your time today.

[1] The Securities and Exchange Commission, as a matter of policy, disclaims responsibility for any private publication or statement by any of its employees. The views expressed herein are those of the author and do not necessarily reflect the views of the Commission or of the author's colleagues on the staff of the Commission. I would like to thank Vanessa Countryman, Marco Enriquez, Christina McGlosson-Wilson, and James Reese for their extraordinary help and comments.

[2] SEC Speech, Has Big Data Made us Lazy?, Midwest Region Meeting of the American Accounting Association, October 2016. https://www.sec.gov/news/speech/bauguess-american-accounting-association-102116.html.

[3] http://cfe.columbia.edu/files/seasieor/center-financial-engineering/presentations/MachineLearningSECRiskAssessment030615public.pdf.

[4] Arthur Samuel, 1959, Some Studies in Machine Learning Using the Game of Checkers. IBM Journal 3, (3): 210-229.

[5] https://en.wikipedia.org/wiki/AlphaGo_versus_Lee Sedol .

[6] For an excellent layperson discussion on how machine learning is enabling all of this, *see*, *e.g.,* Gideon Lewis-Kraus, The New York Times, December 14, 2016, The Great A.I. Awakening.

[7] See http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf .

[8] See, G. Hoberg and C. Lewis, 2017, Do Fraudulent Firms Produce Abnormal Disclosure? Journal of Corporate Finance, Vol. 43, pp. 58-85.

[9] Loughran, Tim, and McDonald, Bill, 2011. When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. Journal of Finance 66: 35–65.

[10] *See, e.g.,* https://www.sec.gov/divisions/marketreg/rule613-info.htm.

[11] Securities and Exchange Commission Strategic Plan Fiscal years 2014-2018, https://www.sec.gov/about/sec-strategic-plan-2014-2018.pdf.

Electronic copy available at: https://ssrn.com/abstract=3226514