



---

# How AI and Big Data Can Help Banks Adapt to a New Accounting Standard

**Scott Liao**

Professor of Accounting  
Rotman School of Management  
University of Toronto

## **CPA Ontario thought leadership and research**

As a thought leader, CPA Ontario examines those issues and trends impacting the accounting profession and Ontario's business landscape. We fuel thoughtful discussion about the challenges and opportunities on the horizon by supporting research from universities and colleges across the Province with the aim of advancing the profession.

This whitepaper was funded by CPA Ontario in support of academic freedom and the views expressed herein are not necessarily the views of CPA Ontario.

To learn more about CPA Ontario thought leadership and research, visit our Insights page  
[www.cpaontario.ca/insights/research](http://www.cpaontario.ca/insights/research)

# How AI and Big Data Can Help Banks Adapt to a New Accounting Standard

---

## **Abstract**

In the immediate aftermath of the 2008 financial crisis, accounting standards for banks needed to change. The old backward-looking for measuring loan losses had contributed to the crisis and would be replaced with a new forward-looking system. In Canada and the United States, among other countries, new standards have since been adopted to safeguard the system. But the new models present challenges for banks attempting to adopt them. Many banks do not have sufficient data collection and analysis capabilities in place nor the tools needed to predict forward-looking estimation. That's where Big Data and Machine Learning come in. Banks can benefit from innovative Artificial Intelligence and analytics approaches as they seek to more accurately estimate lifetime expected losses.

---

\* I would like to acknowledge the financial support of the CPA Ontario Centre for Accounting Innovation Research at the Rotman School of Management. This white paper is partly based on the chapter "A Brave New World: The Use of Non Traditional Information in Capital Markets" in the book Economic Information to Facilitate Decision Making: Big Data, Blockchain and Relevance, ed. Kashi Balachandran, World Scientific Publishing (forthcoming).



**1.**  
**In response to  
the financial crisis:  
new loan loss  
accounting models**



---

The incurred loan loss recognition model used by banks and financial institutions prior to 2008 was at least partially to blame for the 2008 global financial crisis. That model required banks to recognize loan losses after the credit loss becomes “probable,” which tends to be backward looking. In response to the shortcomings of the incurred loan loss recognition model, new expected loss accounting standards have since been introduced. In 2018 Canada and other IFRS countries implemented IFRS 9, a model that requires financial institutions to recognize either 12-month expected credit loss (ECL) or lifetime ECL, depending on the change in credit risk. In 2020, the United States introduced its current expected credit loss (CECL) model, which requires “life of loan” estimates of losses.

Implementation of the new models posed a problem – many financial institutions did not have existing data infrastructure and statistical modelling systems to estimate expected losses (KPMG, 2018). According to Moody’s (2015), Standard & Poor’s (2017), and Deloitte (2017), banks found implementation of the IFRS 9 model more challenging than expected. The challenges experienced included determination of appropriate data needed for analyses, the collection of data, forward-looking cashflow and default risk estimation, macroeconomic indicator prediction, etc.

---

#### CHALLENGES FACING BANKS IMPLEMENTING IFRS 9

- determination of appropriate data needed for analyses
- the collection of data
- forward-looking cashflow
- default risk estimation
- macroeconomic indicator prediction

#### AN ADDITIONAL CHALLENGE: COVID-19

These challenges have been amplified by the COVID-19 pandemic. It is uncertain how many businesses will default or how the unemployment rate will be affected by the pandemic and related lockdowns, and how long this uncertainty will last.

Despite the challenges faced by banks in IFRS countries, implementation of expected loss accounting standards is imperative. Accurate loan loss recognition can have significant consequences to banks, real sectors and the economy (see Beatty and Liao, 2014).

The background of the slide is a dark blue gradient with various financial data visualizations. On the left, there is a green bar chart with four bars of increasing height. In the center and right, there are several line graphs and candlestick charts in white and light blue, overlaid on a grid. Some of the text in the background is blurred, including the word 'NA' and the number '0'.

## **2. Understanding the difference: incurred loss vs. expected loss**

Governments and market participants have criticized incurred loan loss recognition methods for being backward-looking. The US Government Accountability Office (2013), for example, noted that “existing [incurred loss] accounting rules made it difficult for examiners to require banks to make provisions to increase their loan loss allowances when it became clear the credit troubles were on the horizon”. The Financial Crisis Advisory Group (2009) argued the model’s probability threshold delayed the recognition of loan losses, thereby contributing to the financial crisis. Due to the challenges, the Financial Accounting Standards Board (FASB) (2010) noted that “elimination of the current probability threshold for recognition of impairment was widely supported”.

INCURRED LOAN LOSS RECOGNITION MODEL	IFRS 9 EXPECTED LOSS (ECL) MODEL	ASC 326 CURRENT EXPECTED CREDIT LOSS (CECL) MODEL
<p><b>In effect globally from 1975-2018/20</b></p> <ul style="list-style-type: none"> <li>→ Incurred loss accounting requires “objective evidence of impairment as a result of one or more events that occurred after the initial recognition of the asset.” Therefore, impairment is recognized when a loss is probable based on past events and conditions at the financial statement date.</li> <li>→ This approach is criticized for delayed loss recognition that contributed to procyclical lending and excessive risk taking.</li> </ul>	<p><b>Implemented in Canada and IFRS countries in 2018</b></p> <ul style="list-style-type: none"> <li>→ ECL no longer requires loss to be recognized based on the “probability” threshold and past events and conditions giving rise to defaults.</li> <li>→ Instead, ECL requires banks to be forward looking and recognize expected credit losses in the period when loans are initiated based on a probability weighted expected loss estimate.</li> <li>→ Depending on whether the credit risk has increased significantly since initial recognition, banks are required to recognize either 12-month expected credit loss as default or lifetime ECL if credit risk increases.</li> </ul>	<p><b>Implemented in United States in 2020</b></p> <ul style="list-style-type: none"> <li>→ Similar to ECL, CECL no longer requires the probability threshold or loss incurrence.</li> <li>→ Different from ECL, CECL requires “life of loan” outright, which does not depend on whether credit risk has increased significantly since initial recognitions.</li> <li>→ Both ECL and CECL are expected to result in more timely loss recognition, thereby mitigating procyclical lending and improving transparency of financial reporting.</li> </ul>



---

The major difference between the incurred loss model and both versions of expected loss models is that incurred loss accounting requires “objective evidence of impairment as a result of one or more events that occurred after the initial recognition of the asset”. Impairment is therefore recognized when a loss is probable based on past events and conditions at the financial statement date. In contrast, CECL or ECL models do not require the loss to be incurred and impose no probability threshold for loss recognition. Instead, CECL or ECL models require a probability weighted expected loss estimate regardless of the loss probability.

Compared to incurred loss models, when banks adopt CECL or ECL, they need to collect more credit related information and use forward-looking information to estimate life-time expected losses. This involves understanding and forecasting general economic and market conditions (e.g., expected increase unemployment rates, interest rates, etc.), operating results or financial position of the borrower, expected or potential breaches of covenants, and expected delay or default in payments. The estimation also requires banks implementing statistical models that can accurately forecast expected losses. Based

on Deloitte (2017), “given the combination of a principles-based standard with a complex end-to-end production process, the implementation of CECL will be significantly more complex than that of other accounting standards.” In the Deloitte US CECL survey polling senior executives at US institutions, development of statistical CECL models is the most challenging implementation task, while obtaining data necessary for credit modelling and loss estimation, and defining data requirements to support model development are also named as other top challenges faced by banks.

When choosing statistical models, banks need to consider multiple factors including understandability and tractability. The model needs to be easily understandable because the board of directors, audit committees and management all need to be able to understand the model inputs and outputs and the process of the loan loss estimation to explain to external stakeholders including auditors, regulators and shareholders. In addition, the bank needs to consider the timeliness of financial reporting. For example, regulatory reports are due within weeks after the quarterly end, so being able to collect, process and analyze information in a timely fashion including the execution of relevant internal controls and model validation is critical.

### **The difficulty in applying forward-looking factors as the COVID-19 pandemic evolves**

The use of forward-looking information to estimate credit loss is even more challenging given the uncertainty caused by COVID-19. Based on DeNiese and Cigna (2020), information used to assess the change in credit risk and to measure CECL or ECL is changing rapidly as the pandemic evolves. As such, financial institutions may find it difficult to apply forward-looking factors to their expected loss models. As the pandemic evolves, it is increasingly difficult to predict macroeconomic indicators such as unemployment rate, inflation, and borrower-specific default risk. To assist banks to implement the requirements of the standard, both FASB and International Accounting Standards Board (IASB) have provided guidance on how expected loss accounting should be applied. For example, deferral or payment holidays on loans should not result automatically in loans being considered a significant change in credit risk. The IASB recognizes that to incorporate COVID-19 and government support measures into expected loss estimations can be difficult, and has suggested banks exercise management judgments and discretions.

The background of the slide is a night-time aerial view of a city, likely Dubai, with numerous skyscrapers illuminated. Overlaid on this cityscape is a complex network of glowing blue and purple lines that represent digital data or connectivity. These lines are vertical and horizontal, with some curving, and many end in small, bright circular nodes. The overall aesthetic is futuristic and technological.

### **3. Loan loss provision channels and their impact on the economy**

---

There are mainly two channels through which loan loss provisioning may affect bank behaviors that have real impacts on the economy (Beatty and Liao, 2014): the first is regulatory capital, the second is market discipline.

Under the regulatory capital channel, when the economy is booming, some banks under-recognize their loan loss provisions. These banks will have to then recognize losses in bursts to correctly reflect the defaults. During an economic downturn, this additional loss recognition can prompt these banks to cut their lending to avoid violating the regulatory capital minimums, because bank lending is assigned the highest weight in risk-weighted asset calculations as the denominator in the regulatory capital ratio. This behavior can destabilize the economy. While the expected loss models are designed to address this behavior, if banks' expected loss recognition is not accurate or timely, bank lending can still be procyclical and therefore defeat the purpose of the expected loss standards.

In the market discipline channel, timely provisions assist market participants to monitor bank risk-taking. Bushman and Williams (2015) found that banks with more timely loss recognition are less likely to take excessive risk and contribute less to systemic failures. This is because transparent provisioning facilitates external stakeholder monitoring. Bank system stability can be jeopardized, as we witness in the 2008 financial crisis, if bank provisions are not correctly recognized to reflect lending risk and bank risk-taking is not properly monitored. While expected loss models have a potential to make provisioning more transparent and thereby improve stakeholder monitoring, if it is not implemented correctly, stakeholders may be still left in the dark. Therefore, to correctly implement the expected loss models to facilitate monitoring is critical to the banking system.



**3.**

**Big data: three approaches to predict expected credit losses**

While there is no prescribed method for predicting expected credit losses, there are a number of analytics approaches to estimating expected losses that banks, companies, and auditors can consider when implementing their information system and statistical models.

## 1. The cross-sectional model

Harris et al. (2018) developed a structural model of expected credit losses (ExpectedRCL) using bank-specific periodic disclosures including historical net charge-offs, allowance for loan losses, loan loss provision, and fair value of loans. To construct ExpectedRCL, banks can use these bank credit loss indicators and time-varying coefficients based on cross-sectional regressions in the structural model. The authors argue that because these known predictors partially explain expected credit loss individually, combining all of the expected loss from these variables can outperform each of these variables separately in predicting expected credit loss. The authors of this study found that ExpectedRCL substantially outperforms in predicting one-year-ahead

realized charge-offs and bank failures beyond these known credit loss predictors over the next year. Because this study used data collected from bank regulatory consolidated financial statements (Y-9C) from 1996 to 2015, which is easy to acquire to construct this ExpectedRCL, the model is easy to implement for banks.

The cross-sectional model is useful only when estimating the 12-month expected credit loss but not the lifetime credit losses. Whether this model is equally predictive of life-time credit losses is unknown. In addition, the data used in this study are collected from the incurred loss accounting regime and whether the findings of this study can extend to the post-CECL or post-ECL data is uncertain because banks are likely to have changed their lending or financial reporting after the regime change.

### The cross-section model

After collecting data from bank regulatory consolidated financial statements (Y-9C) from 1996 to 2015, Harris et al. acquired the time-varying coefficients from running the following regression by quarter.

$$\frac{NCO_{i,t}}{AveLoans_{i,t}} = \alpha_{0,t} * + \alpha_{1,t} \frac{NCO_{i,t-1}}{AveLoans_{i,t-1}} + \alpha_{1,t} \hat{y}_t \frac{\Delta NPL_{i,t-1}^{unexp}}{Loans_{i,t-1}} + \alpha_2 \frac{NPL_{i,t-1}}{Loans_{i,t-1}} + \alpha_{3,t} LoansYield_{i,t-1} + \alpha_{4,t} FloatLoanRatio_{i,t-1} + \alpha_{5,t} \frac{RELoans_{i,t-1}}{Loans_{i,t-1}} + \alpha_{6,t} \frac{ConsLoans_{i,t-1}}{Loans_{i,t-1}} + \varepsilon_{i,t-1}$$

After the regression is run by quarter, the estimated coefficients are then applied to the respective variables to calculate the expected credit loss as the following equation:

$$\widehat{ExpectedRCL}_{i,t} = \hat{\alpha}_{0,t} + \hat{\alpha}_{1,t} \frac{NCO_{i,t}}{AveLoans_{i,t}} + \hat{\alpha}_{1,t} \hat{y}_t \frac{\Delta NPL_{i,t}^{unexp}}{Loans_{i,t}} + \hat{\alpha}_2 \frac{NPL_{i,t}}{Loans_{i,t}} + \hat{\alpha}_{3,t} LoansYield_{i,t} + \hat{\alpha}_{4,t} FloatLoanRatio_{i,t} + \hat{\alpha}_{5,t} \frac{RELoans_{i,t}}{Loans_{i,t}} + \hat{\alpha}_{6,t} \frac{ConsLoans_{i,t}}{Loans_{i,t}}$$

The variables used in the model are all publicly disclosed information, so they are easy to extract and construct.

In the regression,  $NCO_{i,t}$  is measured as the net charge offs of firm  $i$  at time  $t$ ,  $AveLoans_{i,t}$  is the average balance of loans held by firm  $i$  during period  $t$ ,  $Loans_{i,t-1}$  is the total of loans held for investment of firm  $i$  at time  $t-1$ , and  $NPL_{i,t-1}$  is nonperforming loans of firm  $i$  at time  $t-1$ .  $NPL$  is defined as the total of non-accruing loans, restructured loans, and accruing loans that are at least 90 days delinquent.  $LoansYield_{i,t-1}$  is measured as the firm  $i$ 's ratio of tax-equivalent interest income on loans to the average balance of loans over period  $t-1$ ,  $FloatLoanRatio_{i,t-1}$  is calculated as the proportion of loans of firm  $i$  at time  $t-1$  that reprice or mature within one year.  $RELoans_{i,t-1}$  and  $ConsLoans_{i,t-1}$  are measured as the total real estate and consumer loans of firm  $i$  at time  $t-1$ , respectively. Finally, unexpected change in NPL ( $\Delta NPL_{i,t}^{unexp}$ ) at time  $t$  is measured as the actual NPL minus total loans at time  $t$  multiplied by the ratio of NPL to total loans at time  $t-1$ .

---

## 2. The vintage analysis approach

To address the shortcomings of the cross-sectional model to predict 12-month credit loss, Wheeler (2019) introduced a “vintage analysis” to estimate bank-specific lifetime expected credit losses at each balance sheet date. This approach “requires analysis of the performance of a static pool of financial instruments over time to determine marginal loss rates each period after the vintage is formed and cumulative loss rates over the life of the instruments. These loss rates are then applied to vintages of similar financial instruments outstanding at the balance sheet date to estimate remaining life-of-loan losses for a portfolio”.

The benefit of this vintage analysis approach is that it predicts expected losses over the next five years, compared to the cross-sectional model (e.g., Harris et al., 2018) which does not predict loan losses beyond one year. So, while the recognized allowance under the cross-sectional model is associated with one period ahead net-charge-offs only, Wheeler’s approach is associated with second through sixth year ahead charge-offs. In addition, this vintage analysis acknowledges that the probability of future losses depends both on the type of loan and the time that the loan has been outstanding. This vintage analysis also assumes that past loss rates are the best predictor of future loss rates, which may not be valid. And, like the cross-sectional model of Harris et al., the data used in this vintage analytics approach (from 1990-2016) also predate the CECL regime, so the analysis may not necessarily translate to the new accounting standard.

Wheeler (2019) disaggregates loans into four types: single-family residential real estate loans, non-single family real estate loans, consumer loans (including credit card and automobile loans), and other non-real estate loans (including commercial and industrial loans). For each of four loan types, Wheeler regresses current charge-offs on past loan originations and uses the estimated coefficient for period  $n$  originations as the marginal loss rate  $n$  periods after origination as in the following regression, where  $i$ ,  $j$ , and  $t$  represent firm, loan type, and quarter, respectively:

$$CO_{ijt} = \beta_{ij1}LO_{ijt-1} + \dots + \beta_{ijN}LO_{ijt-N} + \varepsilon_{iq}$$

CO denotes gross charge-offs and LO denotes loan originations. Wheeler then sums the estimated coefficients over  $n$  periods, i.e.,  $\sum_{m=1}^n \beta_{ijm}$  to estimate lifetime cumulative loss rate for each loan origination vintage and then multiplies this lifetime loss rate by the loans initiated in the vintage  $N$ . Finally, he adds up estimated losses of loans initiated in all vintages as the measure for lifetime expected credit losses. After this step, Wheeler makes further adjustments to the expected credit losses for macroeconomic factors (for more detail see Wheeler, 2019).



---

### 3. The refined cross-sectional model

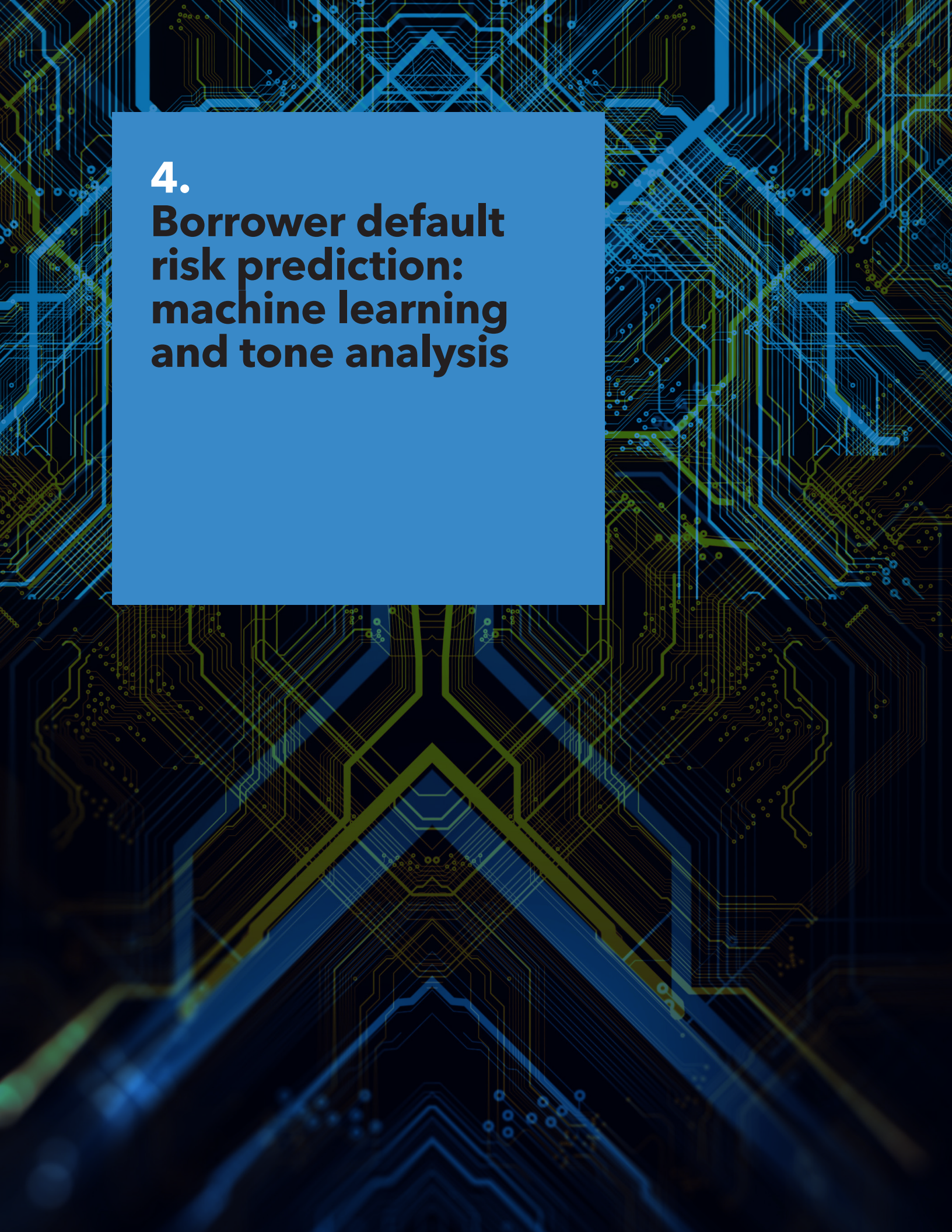
Building on the the 12-month expected credit loss of the cross-sectional model, Lu and Nikolaev (2019) refined the model to extend to lifetime expected loss predictions. This refined model breaks down expected loss into two components: a sector-wide (aggregate) component and a non-systematic, bank-level component. Lu and Nikolaev contend that while the cross-sectional model may be good at predicting the bank-level component, it is poor at predicting aggregate losses. Their solution augments cross-sectional regression with a dynamic latent factor forecasting methodology to extract the factors most effective at predicting aggregate losses. This model outperforms the cross-sectional model but is less straightforward and requires more econometric and technical training to implement.

Like Wheeler, Lu and Nikolaev validated their expected loan loss allowance estimates by comparing the association of their estimates with the recognized allowance under the incurred loss model with future charge-offs. Where Lu and Nikolaev's model differs is its use of the sum of net charge-offs in the next three years rather than individual year's losses to capture future losses. They found that allowances reported under the incurred loss model are equal to 60 percent of estimated lifetime losses and that both their measure of expected losses and the recognized allowance are incrementally associated with the sum of net charge-offs in the next three years. Relying on a set of simplified assumptions to estimate and validate their expected loss estimates, Lu and Nikolaev used current information to predict future losses for the existing portfolio of loans. While they do not assume that past loss rates are the best predictor of future loss rates, they do assume that the association between current information and future loss rates does not change over time and is stable for the period from 1991 to 2017. This model does not distinguish by type of loan or by loan vintage.

For the sector-wide component of expected losses, Lu and Nikolaev first estimate factors based on hundreds of macroeconomic indicators. The model is implemented independently for the loan categories of real estate, commercial, and other. For the bank-specific component of expected losses, the authors modify the cross-sectional model to improve forecasting losses over long-term horizons by extending the training set to all past observations. They also use logit link function to replace the linear regression that sometimes generates unreasonable forecasts.

When considering long-term prediction, Lu and Nikolaev assume that all loans mature in five years regardless of loan type and length of time outstanding. In contrast to the use of past credit losses to predict future losses in the vintage analytics approach, Lu and Nikolaev use coefficients from models of one- to five-year ahead future charge-offs on measures of current loan portfolio and macroeconomic conditions to estimate the future loss rates for the existing loan portfolios.





**4.**  
**Borrower default  
risk prediction:  
machine learning  
and tone analysis**



---

To further their ability to predict borrower default risk – which is an important input to forecast expected losses for consumer loans – banks can incorporate Artificial Intelligence technology, in the form of machine learning. For estimating expected credit losses for large commercial and industrial loans with varying borrower idiosyncratic credit risk, banks may want to estimate the expected loss as a product of the probability of default multiplied by the loss given default. While the traditional approaches to estimating the probability of default, based on publicly available accounting and market information including Altman Z-Score, Ohlson O-Score and expected default frequency (EDF) are still widely used, the new technology and methodologies such as machine learning and textual analysis can significantly improve the predictability for borrower bankruptcy.

Donovan et al. (2019) found that textual information explains significant variation in credit default swap (CDS) spreads beyond the traditional credit risk measures. They used textual information from conference call transcripts and management discussion and analysis (MD&A) disclosures from 2002 to 2016 as inputs into three machine learning approaches: Support Vector Regression, Supervised Latent Dirichlet Allocation, and Random Forest Regression Trees. In addition to these machine learning methods, tone analysis can also be used to predict credit risk even though it is not considered a machine learning approach. Each of the four approaches is discussed below.

### **Tone Analysis**

In the accounting literature, Li (2008) was the first to extract forward-looking information from firm disclosure by examining the association between firm future performance and annual report readability. Using the Fog Index to estimate the level of education needed to understand documents, he noted that firms whose annual reports require higher education to appreciate tended to report lower earnings. While this approach is intuitive, the most

frequently occurring “complex” words that lead to a high Fog index in business documents are words that are readily understood by investors such as “financial”, “company”, “operations”, “management”, “employees” and “customers”. Therefore, the prediction based on this simple approach can be flawed. To overcome this issue, data analysts can introduce pre-determined word lists to measure a document’s tone. Tone is usually defined as the ratio of positive words over the number of negative words. Research suggests firms whose conference calls have a positive tone tend to have higher market returns during the conference call and higher future performance. The quality of tone analysis depends on the dictionaries used.

#### **tone analysis dictionaries**

Data analysts can use pre-determined dictionaries to measure positive versus negative tone in documents. The quality of any analysis will depend on the dictionary used.

#### **Harvard GI word list**

One popular dictionary is the Harvard GI word list. However, Loughran and McDonald (2011) found that 75 percent of Harvard GI negative words do not actually have negative meanings in the financial context. For example, “crude”, “cancer”, and “mine” do not necessarily have negative meanings in the oil, pharmaceutical, and mining industries. Words like “no”, “not”, “without”, and “gross” also do not have negative meaning in typical accounting documents. And frequently occurring positive words like “respect”, “necessary”, “power”, and “trust” do not always have positive meanings when describing future or current operations.

#### **Financial based dictionary**

Based on their criticisms, Loughran and McDonald developed their own dictionary for the analysis of accounting and business documents by examining words used in a large sample of 10Ks from 1994-2008. In this dictionary, there are six different word lists: negative, positive, uncertainty, litigious, strong modal and weak modal. The dictionary is considered extensive including 354 positive and 2,329 negative words. Based on this dictionary, Donovan et al. (2019) measured tone as the difference between positive and negative word counts divided by the sum of positive and negative words. They found this tone measure to be predictive of firm future credit events such as credit rating downgrades, bankruptcies and stock price decline.



---

## Support Vector Regression (SVR)

SVR places weights on individual words and phrases to explain a dependent variable. Donovan et al. (2019) use CDS spreads as the dependent variable to capture credit risk. When estimating SVR, the weights on individual words or phrases are determined by algorithms that simultaneously minimize the coefficient vector magnitude and the estimation errors to reduce overfitting. A unique weight is placed for each one- and two-word phrase count that is included in the conference call transcript or MD&A. Using SVR, Donovan et al. found the predicted CDS spreads successfully predict actual CDS spreads and other credit events such as credit downgrades and bankruptcy risk beyond other known determinants of credit risk. In this model, many words and phrases with higher weights in predicting CDS spreads are often intuitive and can be linked to firms' default risk, like "growth", "slide", "free cash", "liquid", "earn", "cash flow", "matur", "ebitda", "credit", or "oper expens". However, for some other higher weight words, the relation between many other top words and default risk can be more difficult to appreciate, like "ga", "espn", "pulp", "pleas go" or "de".

## Supervised Latent Dirichlet Allocation (sLDA)

sLDA is similar to unsupervised LDA in that it categorizes words and phrases in a disclosure into a set of latent topics using an algorithm. The algorithm assumes all disclosures share the same set of topics, but allows the mix of each topic to vary by disclosure. An easy way to appreciate LDA is to think of it as a qualitative version of factor analysis, classifying words or phrases that often co-occur when discussing latent topics. As an example, if LDA returns a word list of "vaccine", "pandemic" and "payment holiday", we may have a good idea that the latent topic is the COVID-19 pandemic. sLDA adds another layer to the unsupervised LDA model. Because disclosures often discuss similar items or activities in different ways, sLDA groups words that discuss similar notions into topic groups to explain a dependent variable. According to

Donovan et al., most positively and negatively predictive topics for CDS spreads under sLDA appear to intuitively capture credit risk. For example, within one topic group, negative topics (topics associated with lower credit risk) included "share repurchases", "good", and "dividend", while positive topics (topics associated with higher credit risk) included "facil", "loss", and "reduct". Like SVR, however, some words or phrases lack economic intuition in relation to credit risk. Further, some words or phrases appear in both positive and negative topics, which increases the complexity of this approach.

## Random Forest Regression Trees (RF)

One drawback to both SVR and sLDA is that they treat each word or phrase independently. The relation, combination, or order of these words/phrases is not considered, even though the interrelation between the words/phrases often carries important context for predicting credit risk or other response variables.

A regression tree, on the other hand, uses an iterative partitioning process that creates a decision "tree" to predict the value of response variables.

### STANDARD REGRESSION TREES

The standard regression tree method, based on Frankel et al. (2016), uses an iterative partitioning process that creates a decision "tree" by recursively partitioning observations based on one- and two-word phrases to predict an outcome variable. Each partition is identified with a node, a binary classification of the data. At each node, the algorithm examines each of the remaining binary splits of the data using the remaining words/phrases and chooses the next phrase or word that minimizes the sum of squared errors within each partition. The algorithm continues to partition the data using nodes until the number of observations within each partition falls below a prespecified number or when the sum of the squared errors within the partition is equal to zero. When the process stops, the average value of the response variable at each final node represents the predicted value of the response variable.

---

The random forest regression tree is an application of the regression tree method that reduces overfitting and improves generalization. The random forest method constructs a predetermined number of regression trees, with each tree using a randomly selected subset of words or phrases. As a result, each tree uses a different word or phrase as a starting node, allowing the random forest method to incorporate nonlinearities in the estimation. The average predicted value generated by all "trees" is set to be the predicted value of the response. Like SVR and sLDA, some words or phrases using RF lack economic intuitions in relation to credit risk. So, while "net loss", "dividend increase", "ebitda", "growth" or "strong" can be easily associated with increases or decreases in credit risk, "mr", "lng", "rasm" or "patel" are not easily interpreted.





# 5. Summary



---

Given the potential adverse consequences of untimely loan loss recognition, it is important for banks to accurately estimate expected credit losses under the new expected loss regimes - ASC326 (CECL) in the US and IFRS 9 (ECL) in Canada and IFRS countries. However, this can be a challenge. Financial institutions may not have existing models, expertise, or information systems to collect the data needed to forecast forward-looking credit losses. Big Data analytics developed by academics to forecast bank expected credit losses can help. These approaches are relatively easy to implement and allow banks to estimate expected loan losses directly. Machine learning may also help. While no machine learning approach has yet been applied directly to estimate expected loan losses, several machine learning methods are used to predict borrower default risk - as well as others predicting macroeconomic indicators - and these are important inputs for estimating expected losses. Machine learning approaches, though they require more data and technology literacy, can be superior to traditional approaches thanks to their enhanced computational power and flexibility to capture underlying relations between outcome and explanatory variables. Banks can benefit from these newly developed innovative approaches to more accurately estimate lifetime expected losses.

# References

---

- Beatty, A., and S. Liao. 2014. Financial accounting in the banking industry: A review of the empirical literature. *Journal of Accounting and Economics* 58 (2-3), 339-383
- Bushman, R. and C. Williams. 2015. Delayed expected loss recognition and the risk profile of banks. *Journal of Accounting Research* 53, 511-553.
- Cook, T. R., and A. S. Hall. 2017. Macroeconomic indicator forecasting with deep neural networks. Federal Reserve Working Paper.
- Deloitte. 2017. Developing and implementing current expected credit loss (CECL) estimation models. <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-cecl-credit-modeling-POV.pdf>
- DeNiese, N., and V. Cigna. 2020. COVID-19 uncertainties and expected credit losses. <https://rsmcanada.com/our-insights/ifrs-resource-center/covid-19-uncertainties-and-expected-credit-losses-for-credit-unions-and-lending-entities-reporting-under-ifrs.html>
- Döpke, J., U. Fritsche, and C. Pierdzioch. 2017. Predicting recessions with boosted regression trees. *International Journal of Forecasting* 33, 745-759.
- FASB. 2010. Exposure Draft on Financial Instruments.
- Financial Crisis Advisory Group, 2009. Report of the financial stability forum on addressing procyclicality in the financial system. [http://www.financialstabilityboard.org/publications/r\\_0904a.pdf](http://www.financialstabilityboard.org/publications/r_0904a.pdf).
- Frankel, R., J. Jennings, J. Lee. 2016. Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics* 45(2), 209-227.
- GAO. 2013. Financial Institutions Causes and Consequences of Recent Bank Failures. <https://www.gao.gov/assets/660/651154.pdf>
- Harris, T.S., U. Khan, and D. Nissim. 2018. The expected rate of credit losses on banks' loan portfolios. *The Accounting Review* 93, 245-271.
- KPMG. 2018. Implementation of the expected credit loss model for receivables. <https://home.kpmg/de/en/home/insights/2018/06/expected-credit-loss-receivables.html>
- Li, F., 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45, 221-247.
- Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66(1), 35-65.
- Lu, Y. and V. Nikolaev. 2019. Expected loan loss provisioning: An empirical model. Chicago Booth Research Paper No. 19-11
- Moody's. 2015. Implementing the IFRS9's expected loss impairment model: Challenges and opportunities. <https://www.moodyanalytics.com/risk-perspectives-magazine/risk-data-management/regulatory-spotlight/implementing-the-ifrs-9-expected-loss-impairment-model>
- Nakamura, E. 2005. Inflation forecasting using a neural network. *Economic Letter* 86, 373-378.
- Sermpinis, G., C. Stakinakis, K. Theofilatos and A. Karathanasopoulos. 2014. Inflation and unemployment forecasting with genetic support vector regression. *Journal of Forecasting* 33, 471-487.
- S&P. 2017. IFRS 9 implementation top five concerns. <https://www.spglobal.com/marketintelligence/en/news-insights/blog/ifrs-9-implementation-top-five-concerns>
- Wheeler, P. B. 2019. Unrecognized expected credit losses and bank share prices. Tulane University Working Paper.