The top 14 challenges for today's model risk managers: Has the time come to think about going beyond SR11-7?

Received (in revised form) 11th January, 2019

Jon R. Hill
leads the New York chapter of the Model Risk Managers International Association (MRMIA). With over 20 years of experience in diverse areas of quantitative finance, Jon is recognised as a subject matter expert in model risk management, governance and validation, and is the author of numerous publications on these topics. Jon holds a PhD in biophysics from the University of Utah. He is a frequent speaker and chairperson at model risk conferences throughout the USA and Europe.

200 East 63rd St., Apt. 4A, New York, NY 10065, USA
E-mail: jonhill@optonline.net

Abstract Model risk management (MRM) is, in the broadest sense, a governance framework that has an impact on every phase of the model life cycle, including model validation. MRM is a relatively immature discipline compared to validation, a practice most banks, and more recently insurance companies, have had in place for 10 to 15 years. Because it is a recently evolved discipline, model risk managers are confronted with a set of challenges that traditional model validation has not had to contend with. Addressing these expanded challenges in model risk management requires creativity, resourcefulness and support from a firm's senior management, as well as additional staffing with a variety of specialised skills, both quantitative and qualitative. This paper attempts to identify and discuss industry standard solutions and potential alternative approaches to the 14 greatest challenges that model risk managers face today and in doing so raises the question: Has the time come to think seriously about going beyond the recommendations of SR11-7/OCC2011-12, the Federal Reserve Bank (FRB)/Office of the Comptroller of the Currency (OCC) joint landmark guidance that set the standard for model risk management at US financial firms in 2011?

Keywords: *model risk management, model governance, model validation, model identification, risk-tiering, aggregate model risk, model risk appetite, model inventory, model usage, SR11-7, OCC2011-12, SR15-18, CCAR, CECL, FRTB, challenger models, model interdependency, machine learning, big data*

INTRODUCTION

What is model risk management?

Beginning with the financial crisis of 2008 (and thrust into public awareness by the 2012 London Whale debacle at J. P. Morgan), model risk has become a topic of strategic importance that is attracting attention at the board level (both the board risk committee (BRC) and the board of directors (BOD)) of many leading financial firms. While the BRC is by far the most frequently mentioned recipient of reports on model risk, it is not exclusively a matter of internal interest as model risk is also increasingly attracting regulatory attention. The model related risk types most commonly analysed and reported are:

(1) Data: incomplete, corrupt, erroneous or missing data.
(2) Implementation: model errors introduced by incorrect or incomplete implementation.

(3) Statistical: uncertainties introduced by the chosen methodology, such as convergence of a Monte Carlo simulation or the standard error of a regression.
(4) Parameters: limitations or uncertainties introduced by invalid or incomplete underlying assumptions.
(5) Calibration: errors resulting from incorrect or inaccurate fitting of model parameters to data.
(6) Misuse: use of valid models for inappropriate applications.
(7) Interpretation: incorrect interpretation of model results.
(8) Dependence: use of data produced by upstream models based on assumptions that are inconsistent with downstream (recipient) models.
(9) Inventory: risks associated with incomplete or inaccurate model inventories, use of unvalidated models or models that have been retired or failed validation, opacity of model usage, orphan models in inventory (ie models without clear lines of ownership).

To address, remediate and manage these types of model risks, many financial firms have evolved to a significantly expanded form of model discipline known generically as model risk management (MRM) in the industry. MRM is charged with a much broader set of responsibilities than model validation, the traditional approach to mitigating model risk. MRM serves as an independent function responsible for overseeing all aspects of model risk over the entire model life cycle and ensuring compliance with both internal policies and external regulatory guidelines, such as SR11-7[1] and SR15-18.[2] MRM therefore has an impact on all six phases of the life cycle of financial models, model validation being one key central phase (see Figure 1).

This paper will describe 14 of the most pressing challenges today's model risk managers are confronted with in order to carry out the responsibilities described above. Industry standard practices as well as potential alternative approaches for addressing these challenges will be put forward in this paper. These are:

(1) Establishing a comprehensive and holistic framework for model risk governance.
(2) Positioning model risk as the fourth leg of a firm's enterprise risk management.
(3) Incorporating big data and machine learning into model risk management.
(4) The challenge of challenger models for comprehensive capital analysis and review (CCAR).
(5) Model identification: is it a tool or a model?
(6) What to do about validating vendor models?
(7) Creating and managing a complete and accurate comprehensive model inventory.
(8) Developing a dynamic model inventory.
(9) Identifying and representing complex model interdependencies.
(10) Validation of expert judgment models.
(11) Verification of model input data quality.
(12) Tiering models by risk rating.
(13) Assessing model risk in the aggregate.
(14) Improving the quality of model and validation documentation.

This list of challenges suggests the level of knowledge and expertise in diverse areas to which model risk managers must aspire in order to be effective. Doing so may lead MRMs at some financial institutions to embrace solutions that go well beyond the bar set for model risk management by the Federal Reserve Bank (FRB) and the Office of the Comptroller of the Currency (OCC) in 2011 with the joint release of bulletins SR11-7 and OCC2011-12.[1,3] To this point, sections 3, 8, 9 and 14 of this paper describe fertile opportunities for forward-looking firms to elevate their model discipline above the baseline established by SR11-7/OCC2011-12.

The 14 topics addressed in this paper are solely the author's opinions about the major challenges confronting MRM today, based on 20 years of extensive experience in the banking industry. They are not intended to represent the opinions of any particular financial firm or those of any of the author's former employers. Further, in this paper the term 'financial firm' will be used in the broadest sense to include top tier investment and retail banks, hedge funds, credit unions and insurance companies—or any entity engaged in an area of finance that relies on the use of quantitative or qualitative models to support its business activities.

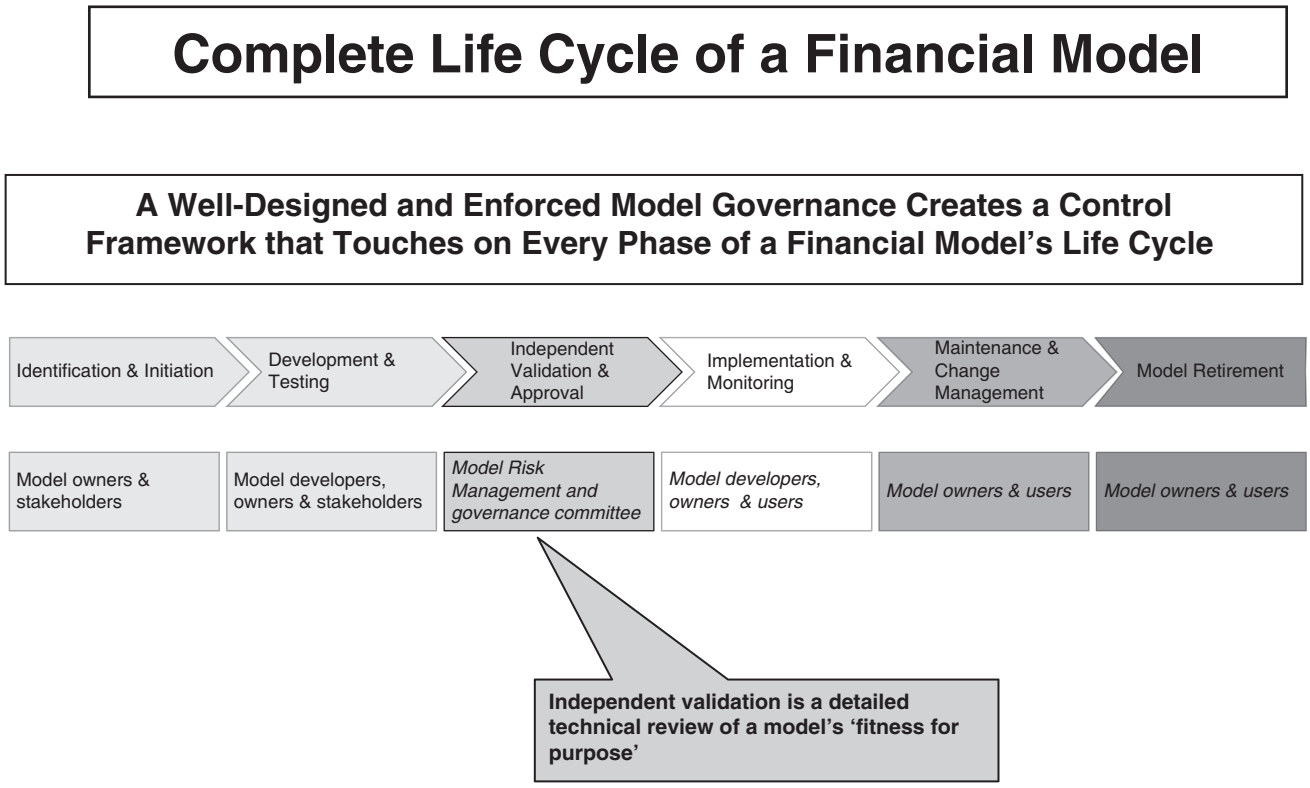

Figure 1: The six phases of the model life cycle, all of which are affected to some degree by model governance

Each of these 14 challenges will be discussed in a medium level of detail in the following sections of this paper.

(1) ESTABLISHING A COMPREHENSIVE AND HOLISTIC FRAMEWORK FOR MODEL RISK GOVERNANCE

'*Model validation is a necessary but not sufficient condition for effective model risk management*'—Dennis Bennett, founder of the Model Risk Management International Association (MRMIA)[4]

As recently as 2015, model risk management at nearly all firms consisted almost entirely of performing model validations, prioritised by risk, to the level of rigour mandated by SR11-7.[1] More recent developments, however, partly inspired by the challenges of assessing firm-wide model risk in the aggregate, developing a complete firm-wide model inventory and identifying interdependencies between models, has increased awareness of model risks that do not arise from any one model, but of those that exist between and external to models within what has been described most effectively as *the model ecosystem*. This phrase captures an awareness that most models as used in financial firms today do not function independently of other models within the ecosystem. The contrast is that while traditional model validation addresses the risks within models, it does not adequately address those risks in the ecosystem that exist outside of and between models.

A description and discussion of many types of model ecosystem risks may be found in a seminal paper by Martin Goldberg.[5] (Some examples identified in this paper include the following: input data source risk, governance risk, model suitability and multiple version risk, jurisdictional regulatory risk, model misclassification risk, assumption and

overlay risk and design/calibration risk.). Model risk management, as described herein, has evolved since 2012 as an umbrella function in recognition of the need to address those model ecosystem risks not addressed by traditional validation. (Although not described in the Goldberg paper,[5] inventory risk, which is discussed in section 8, also clearly arises from the model ecosystem.)

The list below highlights eight concepts that are key to model governance and risk management:

(1) Model governance has an impact on all six phases of the model life cycle: (i) model identification and initiation; (ii) development and testing; (iii) validation and approval; (iv) implementation and ongoing monitoring; (v) maintenance and change management; and (vi) retirement (Figure 1).
(2) Model validation is the central component the life cycle, serving as the gatekeeper to production by addressing the risks within a quantitative model. New models should not as a rule be approved for production without first passing validation. The rigour and sophistication of the validation should be comparable to the complexity and materiality of the models and the firm's overall size and complexity of model use.
(3) Model validation should be performed by staff with education and skills comparable to those of the model developers and should be accorded a level of compensation and authority that supports their ability to pose 'effective challenge'. (As described in SR11-7, the term 'effective challenge' means 'critical review by objective, informed parties who have the proper incentives, competence, and influence to challenge the model and its results.'[1])
(4) One of the key responsibilities of model risk managers (as opposed to model validators) is to identify and mitigate risks arising from the model ecosystem that are *outside and between models* and therefore not addressed by traditional model validation.
(5) A robust model governance framework will provide complete coverage for the following: policies and procedures, roles and responsibilities for ownership, control and compliance, model inventory, model risk assessment and all model-related documentation.
(6) Differentiating between tools and models is a current area of intense focus at most financial firms.
(7) The biggest challenge for model inventory is ensuring completeness and accuracy. At most firms this is accomplished through an attestation process: obtaining confirmation from model stakeholders that the inventory of models for their area of responsibility is complete to the best of their knowledge.
(8) Numerical measures of model risk are very diverse in nature or in some cases not feasible at all, for example, risk models. A qualitative score card type of approach that combines quantitative metrics and expert judgment can provide a common measure across all model types and form the basis for aggregation of model risk, a key requirement of SR11-7.

To be effective, model risk managers must develop competence in each of these eight key areas of model discipline and should be compensated commensurate with model developers.

Figure 1 presents a standard phase diagram of the complete life cycle of models that are used in financial firms, from conception and implementation to eventual retirement. All firms committed to practising effective model risk discipline should develop policies and procedures that will lead to full compliance with the FRB guidelines for model risk management over all six life cycle phases as set out in SR11-7.[1] Typical core responsibilities of MRM at SR11-7-compliant firms should include, but are not limited to:

(1) Creating, maintaining and enforcing an MRM framework and set of governing policies.
(2) Identifying and classifying models by type and risk tier (typically high, medium and low).
(3) Performing independent validations and approvals of new models or of significant changes to existing models.
(4) Performing annual reviews and periodic (with frequency based on risk tier) revalidations of existing models.

(5) Creating and maintaining a firm wide model inventory.
(6) Performing ongoing model performance monitoring.
(7) Overseeing and approving compensating controls (ie overlays, exposure limits, valuation adjustments) applied to mitigate model weaknesses identified by validation.
(8) Reviewing and managing validation findings, including monitoring of follow up actions and escalating unresolved issues to the appropriate model risk oversight committees when necessary.
(9) Maintaining SR11-7 compliant validation documentation, including assessments of developers' documentation for SR11-7 compliance, especially inclusion of developmental evidence, specification of model inputs and all upstream model dependencies.
(10) Socialising validation schedules and scope to model stakeholders and relevant committees.
(11) Submitting aggregate model risk reports that highlight concentrations of model risk to the model oversight committee (MOC), BRC, BOD and any other relevant model risk review groups on a regular periodic (monthly or quarterly) basis.
(12) Approving or disapproving exceptions to MRM policies and escalating any approved exceptions to the model oversight committee hierarchy, up to and including the BRC.

Model oversight structures and model risk committees

Model governance policies should describe the structure and relationships of model risk oversight committees, beginning at the lowest levels with working groups populated from model risk and validation staff to address specific issues. The model risk working groups should report and escalate issues into an MOC composed of middle and senior level managers tasked with monitoring model risk and performing review and challenge sessions on new model development and validation outcomes. MOCs should escalate ultimately to a board risk committee composed of the most senior risk managers. Global firms that have regional subsidiaries in the form of independent holding companies (IHCs) will be required to tailor model risk policies and procedures that are compliant with the regional regulatory requirements, but should also be compatible with the firm's global policies and procedures. The chief model risk officer should serve as chair of the MOC with provisions for escalation to enterprise risk management and ultimately to the board risk committee.

(2) POSITIONING MODEL RISK AS THE FOURTH LEG OF A FIRM'S ENTERPRISE RISK MANAGEMENT

Market, credit and operational risks are the traditional three legs of enterprise wide risk management at all financial firms. Gaining equal footing for MRM so that it may serve as a fourth leg of a firm's enterprise risk management framework requires increasing the visibility, and thereby the awareness of, model risk to senior management. In order to accomplish this, the MRM framework should include a hierarchy of committees that oversee and approve MRM policies and procedures, review validation outcomes through review and challenge sessions and provide escalation paths for MRM issues that require review and approval at more senior committee levels, such as a model oversight, risk process and board risk committee. A central responsibility of the MRM framework is for a model risk appetite to be established, typically at the board risk committee level, and for regular (monthly or quarterly) model risk reports, including model risk (red-amber-green) RAG heatmaps, to be submitted up the committee hierarchy with high visibility to senior management. The model risk reports should characterise the current level of model risk in the aggregate, highlighting any areas of concern, such as concentrations of red model risk status identified by the RAG heatmaps.

Firm-wide model risk in the aggregate is discussed in further detail in section 11.

(3) INCORPORATING BIG DATA AND MACHINE LEARNING INTO MODEL RISK MANAGEMENT

There is extensive academic literature in the field of machine learning (ML)/artificial intelligence (AI)

describing applications for financial services. It is well established that ML algorithms tend to perform better on data-rich applications, so the convergence of big data (BD) with ML is especially promising for model risk management and validation. Richard Bellman, the father of dynamic programming theory and optimal control, argued in his classic 1957 work[6] that high dimensionality of data (ie many elements that can be incorporated into a model) is a fundamental hurdle to many modelling applications, especially in the context of pattern recognition types of applications for which learning complexity grows at a much faster pace than the degree of dimensionality of the data. Bellman lamented this as 'the curse of dimensionality'.[6] ML, on the other hand, thrives on high dimensionality of data and so Bellman's curse becomes an enchantment when ML is combined with BD in the context of model risk management.

ML and BD are topics far too broad to treat in any detail here. This section will present instead the author's views about how these rapidly evolving fields will have a significant impact on the way in which model validations are performed over the next five- to ten-year horizon.

The procedures employed by banks and other financial institutions for performing full SR11-7 compliant model validations have been essentially stagnant since the release of SR11-7 in 2011.[1] Although there have been incremental improvements in documentation, testing and tools that reduce time and effort for replication (such as MatLab and R) it is essentially the same process that was first codified by OCC2000-16[7] in the year 2000. The bottom line is that significant improvements in validation methodology that will result in streamlined replication, testing and documentation, expanded uses of data and a substantial overall reduction of manual effort are long overdue but will soon arrive.

These overdue improvements will very likely be driven by a major disruption introduced by the convergence of model validation with ML and BD over the next 5 to 10 years. The word 'disruption' is not used here in a negative connotation, but in a very positive sense of replacing some of the most inefficient and tedious parts of the validation process with new technologies that free the quantitative validators to focus in the more demanding cognitive aspects of model validation such as judging the model's fitness for purpose, the validity of its underlying assumptions, appropriateness of the input data sources chosen by the developers, the quality of the model documentation, including developmental evidence, and assessment of model's behaviour under stressed conditions.

Machine learning is mature enough today to be successfully applied to assess the conceptual soundness of a model's underlying methodology, to analyse data quality and perform basic remediations such as replacing missing or obviously incorrect data, truncating outliers and transforming non-stationary time series to stationary processes. ML may also be employed to generate test suites designed to reveal known weaknesses of certain classes of models, including stress tests, and to assess validation tests results. ML has been successfully applied in some instances to solving stochastic differential equations. ML will eventually assume the majority responsibility for the construction and testing of challenger/benchmark models, construction of complete test suites and managing their execution on both the primary and challenger/benchmark models, and complete outcomes analysis, including exchanging challenger/benchmark models with champions as a result of the analysis.

ML may assume the labour-intensive tasks of assessing of the quality of the input data as well as performing basic data cleansing activities as described in the previous paragraph. BD may provide a wealth of additional extensive data resources with higher dimensionality than currently available, such as econometric and demographic data. Machine language models for probability of default (PD) used for credit risk applications could significantly improve performance by accessing the troves of demographic data available from federal agencies, such as census data that describe age and geographic distributions of populations combined with income statistics that can refine predictions of PD.

For further reading on this topic, the references at the end of this paper include selected texts for introductory and advanced topics in ML/AI.[8–12]

Validation of machine learning models

The use of ML methodologies as benchmarks in model validation does not require validation of the ML benchmarks themselves, but if ML models are employed by the model developers as champion models, they are subject to model validation just as any other first line of defence against model risk (FLOD) model would be. This introduces some additional complexities because ML models, especially those that employ deep learning, often lack transparency. For a model to be validated, its conceptual soundness must be assessed, its documentation must explain how the model produces output and the output must be explainable. Validation of ML models presents special challenges compared to traditional model validation in the following areas:

(1) Conceptual soundness: ML models are far less familiar and less well understood by validation practitioners. This will make it more difficult for practitioners to assess the fitness-for-purpose and demonstrate the suitability of an ML model for an intended application.[13]
(2) Model documentation: Per SR11-7, model documentation should be self-contained and sufficiently detailed to allow third-party reviewers to replicate the model without recourse to developers or the model's code. This is often a difficult enough task for traditional validation and is made much more so for ML models; documentation and explanation of the ML model's selection of a champion model for example will become even more difficult owing to its relative opacity and higher dimensionality.
(3) Outcomes analysis and model testing: Firms that use ML models will have to rethink procedures for sensitivity and back-testing and may have to apply more computationally intensive techniques to test the accuracy and stability of ML models. Outputs and sensitivities may be quite different from those of traditional models due to the less organised and higher dimensionality of ML models.
(4) Use of vendor ML models: There is a much higher prevalence of vendor models that employ ML techniques than there are that employ traditional modelling. SR11-7 requires the same level of validation and model risk control rigour for vendor as for internally developed models. This is a daunting enough prospect for traditional models due to the lack of vendor transparency into its intellectual property (see section 6), requiring banks to relax their validation rigour somewhat and rely more on benchmarking, sensitivity and stability testing and outcomes analysis. In the case of vendor ML models, validation will be even more challenging. As described in item (3) above, there are additional complications associated with these standard techniques; banks in the position of having to validate vendor-supplied ML models will have to devise and rely on softer validation methodologies that are bespoke to ML, such as more frequent periodic performance monitoring, assessment of developmental evidence and justification of the applicability of the ML model's fitness for purpose.[14]

(4) THE CHALLENGE OF CHALLENGER MODELS FOR CCAR

A requirement for firms undergoing the FRB's comprehensive capital analysis and review (CCAR) stress tests, as described in an FRB guidance letter from December 2015, SR15-18,[2] is the creation of challenger models to benchmark and sanity-check the performance of the champion models developed for CCAR:

> *'A firm should use benchmark or challenger models to assess the performance of its primary models for all material portfolios or to supplement, where appropriate, the primary models'*—from SR11-7[1]

Although challenger models are a form of benchmark, is there a difference between a benchmark model in the traditional business-as-usual (BAU) sense and CCAR challenger models? SR15-18 uses the terms challenger and benchmark interchangeably, but in reality most firms make a practical distinction that has become industry practice. Challengers are a form of benchmark model that are distinguished by a singular requirement: challenger models must be developed by the first line of defence, whereas the broader class of

BAU benchmark models may be developed by either the first or second LOD (FLOD and SLOD respectively). Model validators (within the SLOD) often develop benchmark models to compare outputs and sensitivities against the models submitted for validation by the FLOD. SLOD benchmark models do not have to be validated so long as they are used as intended for the validation process[15] and not by the FLOD as primary models.

SR15-18 recommends that the most important (or primary) CCAR models should have one or more challenger/benchmark models, but is not prescriptive about how the challenger/benchmarks are to be developed or how they may differ from BAU benchmark models traditionally used by developers and validators. In practice, however, the most effective challenger/benchmark models are those that implement a different methodology from that of the champion. This orthogonality of methodology is particularly important for those classes of models for which the 'right' answer is not known, such as operational risk, incremental risk charge (IRC) and comprehensive risk measure (CRM) estimators that are not viable candidates for back-testing due to their high confidence levels.

Benchmark models used solely for validation do not need to be validated themselves but SR15-18 states that in such cases firms must still 'assess the rigor of all benchmark models and benchmark data used to ensure they provide reasonable comparisons'.[2] Unfortunately, SR15-18 does not provide guidance explaining what is meant by 'assess the rigor' or how that differs from a validation—this distinction is left to the judgment of the firm's model risk managers.

The main purpose of challenger/benchmark models is to serve as a comparative testbed for the CCAR primary (or champion) models. Significant differences between primary and challenger/benchmark models should be addressed by model developers by modifying the primary model, applying model overlays (adjustments to the model output) or replacing the primary model with a different estimation methodology. One such possibility is to replace the primary model with a better performing challenger/benchmark. Because of this possibility, if time and resources allow, it is a useful practice for the SLOD to perform full validations on both primary and challenger/benchmark models. This conservative practice would also help to inform the FLOD about the comparative performances of their primary and challenger/benchmark models and offer additional support for deciding whether to replace the primary model with a challenger/benchmark. Additionally, if a challenger/benchmark model is used to calibrate or add an overlay to a primary model or to contribute in any other way to the firm's estimates (such as by averaging primary and challenger/benchmark results together) then SR15-18 makes clear that the challenger/benchmark model should receive an appropriate level of rigorous review. For CCAR, this should be interpreted as a recommendation to perform a full validation equivalent to that for primary models.

SR15-18 is also not prescriptive regarding organisational structure within the FLOD for development of the challenger/benchmark models. At some firms the challengers may be developed by the same team that develops the primary champion models; at others separate independent teams may be tasked with development of challenger/benchmark and primary models. There are trade-offs between these two approaches to development. If a CCAR development team has responsibility for both primary and challenger/benchmark models, the two types of models may exhibit functional similarities (ie the apple does not fall far from the tree). For example, a model whose only difference from the primary is that it uses different input data can be considered a legitimate benchmark for the purposes of testing by both FLOD and SLOD. But if there are serious flaws in the primary model's methodology, these flaws may be replicated within the benchmark model as well.

On the other hand, if the challenger/benchmark is developed by a separate and independent FLOD team whose developers have less in-depth knowledge and experience with the methodology and products the primary model is designed for, the result may be challenger/benchmark models that are not very challenging at all.

Federal bank examiners have been known to ask during CCAR model exams how many challenger/benchmark models were selected to

replace primary models during the CCAR process. 'None' is not a recommended answer as it would suggest that the challenger/benchmark models were not sufficiently challenging to the primary models they were compared against. At the other extreme, '50 per cent' would also not be a good answer as it would indicate that the development process for the primary champion models was flawed and lacked the rigour necessary for CCAR stress testing and capital planning. An acceptable answer should fall somewhere between these two extremes.

One question remains: Who gets to decide if the primary model should be replaced by a better performing challenger/benchmark? In this situation the answer should be obvious: the model stakeholders in the FLOD, who have ownership of and responsibility for using the primary model for the CCAR stress tests, should also have sole responsibility for selecting the most suitable among the primary model candidates. This not a decision that should be made or influenced by the SLOD validation team although the FLOD can decide to choose a benchmark developed by the SLOD as the FLOD champion, with restrictions on validation as noted.[15]

(5) MODEL IDENTIFICATION: IS IT A TOOL OR A MODEL?

One area of increasing industry and regulatory focus is the classification of quantitative methods that are implemented in software into one of three buckets that could be loosely described as calculators or tools, near models and true models. The historical trend in the US financial industry has been to keep the set of true models used by a firm to a minimum owing to the amount of effort and resources required for development, testing, validation and ongoing monitoring to be compliant with SR11-7.

A case in point would be anti-money laundering (AML) applications. Historically, few if any banks classified AML tools as models because their function is to monitor financial transactions to identify questionable activities that might be instances of money laundering. Most AML models operate by applying relatively straightforward threshold filters to flag suspicious transactions. Since these filters do not involve algorithms with underlying assumptions or complex mathematics they were historically treated as tools rather than models subject to validation, even though the types of activities they were designed to detect carried potential for great reputational and financial risk to banks.

The out-sized risks associated with what appeared to be simple non-models along with regulatory encouragement resulted in banks re-classifying their AML tools as models subject to validation. A recent trend at most financial firms is for the diffuse boundary between models and non-models to be pushed progressively towards the non-model side of the divide despite resistance from the non-model stakeholders. This is one reason the model count at most firms is progressively increasing.

A second factor influencing model count at many banks is the number of new models that banks are required to develop to satisfy regulatory requirements for CCAR, current expected credit loss (CECL) and the Fundamental Review of the Trading Book (FRTB). In this new environment of increased regulatory scrutiny and demands for new quantitative applications, affected firms need to develop a sound framework for distinguishing tools from models. To avoid uncertainties that inevitably arise during model identification, this framework should be clearly defined in a firm's model risk governance policy and procedures.

As an illustration of the differences between non-model and model is provided in Table 1. It develops a calculator into a model in four steps by starting with a simple deterministic function that is clearly a non-model and adding additional functionality and complexity until the application crosses through an ill-defined grey area into becoming a true model.

The darker shaded third row of Table 1 describes where the battle lines are typically drawn at most firms between the first and second lines of defence. Model owners (developers and supervisors) in the FLOD understandably prefer to have the smallest possible inventory of true models that are subject to full validation since it adds significant overhead to their workload. Validators in the SLOD on the other hand need to be confident that the firm's validation procedures are compliant with the broad definition of models and fulfil the requirements for validation as articulated in SR11-7.

There is thus a natural tension between the FLOD and SLOD around those software implementations

Table 1: From tool to model in three easy steps

Application function	Tool or a model?
• Anna creates a spreadsheet that aggregates the groups trading positions for reporting.	• Definitely a tool, not a model.
• Stephanie expands the spreadsheet with risk numbers for each position from the daily risk report.	• Completely deterministic, so still a tool, not a model.
• James enhances the spreadsheet with 'what-if' calculations for potential buys & sells.	• This could be a model if the calculations involve assumptions. This is the grey area between tool and model, so let's call it a 'toodel'.
• Svetlana codes a probabilistic risk calculation into the spreadsheet so the 'what-if' scenarios can be run intra-day	• This application has definitively crossed over into the domain of becoming a true model.

that fall into the grey netherworld resulting in many animated discussions over exactly how models are distinguished from tools. Within the darker grey area of Table 1, validators understandably tend to lean towards being conservative and strive to push the demarcation between models and tools downward, in direction of true models, as they must defend their work to their senior model management, internal audit and in regulatory exams on model risk management.

Should quantitative non-models be subject to validation, review or just a basic verification?

Non-models, tools or calculators that produce a quantitative output, even something as simple as a spreadsheet that adds a column of profit and loss data, still need to be independently reviewed for correctness of implementation. This process should be thought of as a simple verification, or sanity check, of the accuracy, appropriateness and correctness of implementation of the calculator rather than a light-touch validation or review. Whatever procedures a firm decides on for treatment of non-models should be clearly defined in the firm's global model governance policy and procedures documents.

Pushback from model owners: Arguments often made for why it is not a model

Model owners (developers, users and supervisors) have often proven adept and creative in articulating descriptions of their quantitative software implementations that are designed to avoid the dreaded classification as 'true models' that must be submitted to the SLOD for validation, annual review and ongoing monitoring processes. Ten of the most ridiculous but sometimes amusing arguments that have been presented to model risk managers at various firms in order to justify why their quantitative software should not be considered models are listed in Appendix A ('It's not a model because …').

(6) WHAT TO DO ABOUT VALIDATING VENDOR MODELS?

Vendor, or third party, models are frequently used by firms that lack either the expertise or wherewithal to develop their own models inhouse. It can therefore be more economically feasible to license or purchase outright models developed by external, or third party, firms. SR11-7 is quite clear on the issue of validation: vendor models must be validated to the same level of rigour as internal models. The crucial question then is, who is responsible for validating a vendor's model?

Should the vendor be required to provide evidence of independent validation of its models to SR11-7 standards? To quote SR11-7, 'Vendors should provide appropriate testing results that show their product works as expected.'[1] This suggests that the onus should be on the vendor to provide evidence that the vendor's model has been validated to SR11-7 standards, which is the optimal solution from the client's perspective. SR11-7 states, however,

that clients are also responsible for validating their particular use of the model as well as any customisations performed by the vendor for the client.

Vendors have a substantial investment of resources in developing complex proprietary models and tend to be very reluctant to provide the type of detailed documentation required for validation in order to protect their intellectual property. The optimal solution from the client's perspective is for the vendor to validate its own model to SR11-7 standards, or to engage a third party to perform the validation. The vendor would understandably be reluctant to release the validation report itself as this would expose the proprietary details of the model's design, but the vendor should be able to provide an attestation (perhaps by a third party) that its model performs as intended and has been validated to SR11-7 standards.

If the vendor is not willing or unable to provide proof that its models perform as intended and have been validated to SR11-7 standards, the onus falls upon the client to perform the validation. Depending on how forthcoming the vendor may or may not, be, this can result in a 'black box dilemma', ensuring that the model performs as intended without really knowing the details about what is going on inside the black box. This can result in unresolvable stalemate if the validation relies on replication or benchmarks that produce results that are not in agreement with the vendor's model output.[16]

In this situation, one option might be to modify the level of rigour by testing the vendor model under a variety of scenarios and/or stresses and assessing whether the model output is as expected. For pricing models this would be a straightforward approach if market prices for the products that model is intended for are available. For risk models this approach would not be practical as there are no comparable public data available for risk calculations. In such cases the use of expert judgment regarding the model's performance might be the preferred recourse for validation.

Given the difficulties clients may encounter in validating vendor models, it is highly advisable for clients make proof of validation part the contract negotiations with vendors. A client might ask vendors if they consider their models to be SR11-7 compliant. If the vendor responds in the affirmative, then the client can request confirmation that the model have been validated to SR11-7 standards. If the vendor is unable or unwilling to provide such confirmation, it may be advisable for the client to consider other vendors. Even so, clients are still required to validate their use of the vendor model.

There remains the question of where responsibility for the reliability of upstream input data into a vendor model should reside? If the input data is acquired or provided by the vendor, then verification of the quality and reliability of the data should naturally be the vendor's responsibility. If, however, the input data is provided by the client, this responsibility would fall either to the client or to the provider of the client data.

(7) CREATING AND MANAGING A COMPLETE AND ACCURATE COMPREHENSIVE MODEL INVENTORY

SR11-7 requires banks to 'maintain a comprehensive … firm-wide inventory of all models'.

Internal audit is responsible for assessing the 'accuracy and completeness' of the firm's model inventory.[1] The guidance further recommends that inventories describe the purpose and products the model is designed for, its *actual or expected usage*, any restrictions on usage, model sub-components (which may be other models), the type and source of inputs to the model, model outputs and the intended downstream applications of the outputs. It is recommended that inventory should also identify by name the model owners (developers and model supervisors) and validators, validation dates (completed and planned), date of last model update, dates and results of ongoing monitoring, any exceptions granted to the model and expected model lifetime.

Although not specifically stated in SR11-7, it should be obvious that in addition to the above, a complete model inventory should include the unique model ID assigned to the model, the model name, the model's status (ie active, inactive or retired) and

the validation status (validated or unvalidated). A comprehensive inventory would also include the model documentation submitted by developers for validation, the completed validation document(s) and the validation outcome (eg passed, passed with conditions or failed). A complete inventory would also list the number of findings resulting from the validation (including the findings' risk level, eg high, medium or low risk), dates by which the findings are to be remediated and the number of findings that remain outstanding (ie not yet remediated and approved by validation). The number of findings resulting from the model validation along with the risk levels can be a useful contribution to an estimate of model risk.

SR11-7 does not offer guidance regarding the implementation of model inventory platforms, such as spreadsheets versus SQL or other types of online databases. Spreadsheet-based inventories may be adequate for relatively small numbers models (on the order of hundreds or less), but larger firms may have several thousand or more models with unique IDs. For firms that strive to comply with the SR11-7 recommendation for a single, firm-wide inventory a spreadsheet approach may quickly become unwieldy. Since the first decade of the 21st century, most firms have invested in either building out in-house inventory databases or buying third party systems. Both approaches embody trade-offs between cost and ownership but in the long run a firm is probably better off building their own inventory system in-house.

Regardless of the type of platform selected by a firm for their model inventory (eg spreadsheet or online database, in-house or vendor), SR11-7 requires confirmation that the model inventory is complete and accurate to the best of the firm's knowledge. At virtually every firm with a model inventory, this is accomplished through a manual attestation process overseen by the firm's model risk managers. In a typical scenario, a model risk manager will obtain a list of models assigned to a model supervisor from the inventory and submit this list (usually by e-mail) to the supervisor who is then asked to attest that the list is a complete and accurate description of the complete set of models that the supervisor is responsible for. This exercise is typically repeated annually for all model supervisors and all models in inventory. In practice, this process can be less straightforward and error-prone than the description implies, as some of the models in inventory may have become 'orphans' owing to staff turnover, or some models may have become inactive or retired without the inventory being updated accordingly.

It is not an uncommon MRM experience for the model supervisor's list to contain inconsistencies or outright errors, such as models that are missing from the list due to omissions or errors in the inventory or models in the list that the supervisor does not recognise. Resolving these issues may require multiple iterations of inventory review and the attestation process, a manual and time-consuming process.

(8) DEVELOPING A DYNAMIC MODEL INVENTORY

The model inventory described in SR11-7 and required by US regulators is a static repository of indicative data. Even after fully populating such a model inventory there will remain some persistent and glaring blind spots in almost every firm's model risk discipline in the form of *residual inventory risk*, a topic rarely described or discussed in formal papers or at MRM conferences. Most, if not all, financial firms share a common set of often over-looked shortcomings in understanding and tracking actual model usage, a topic that is only obliquely alluded to in SR11-7: 'The guidance further recommends that inventories describe … [the model's] *actual or expected usage*.'[1]

Is it possible to go beyond the basic requirements for the type of static inventory described in SR11-7? In the current context, the phrase 'actual usage' might be interpreted more broadly than 'the purpose and products the model is designed for' language from SR11-7. Today's static model inventories are characterised by an absence of data that captures the dynamics of model usage: accurate historical data describing the *how, when and where* of model execution.

A lack of transparency in fully understanding the dynamics of model usage appears to be endemic in the financial industry. Such opacity is revealed by a systematic inability to answer any of the following

types of usage questions with confidence and quantitative accuracy supported by data:

(1) What is the exact number of different models that have been used over the last year?
(2) How often has each model been executed, by day, by month, by year? Can you identify the most frequently and least frequently executed models?
(3) Where are the firm's models being used? By business unit, legal entity, geographic regions?
(4) Can you provide a complete list of the models used by each of the above entities over the last year?
(5) Are there any models in your inventory with an active status that were not executed during the last year?
(6) Are there any models that were executed on any of your firm's computers that do not appear in inventory? Please provide a full listing.
(7) Are you able to provide a full list of the IDs of models that exhibit significant seasonality? If so, what are the peak and troughs of seasonal model usage?
(8) Were there any instances of a retired model still being executed during the last year?

A recent paper by this author appearing in the *Journal of Structured Finance*,[17] has proposed a solution for addressing these areas of opacity in model dynamics. The proposed solution relies on the introduction of a model-embedded identity token operating in tandem with an embedded transponder tracking function. The paper demonstrates in detail how all of the eight questions cited above could be answered by accurately and comprehensively tracking just three critical pieces of model-related information for each execution event: model ID, time stamp and MAC or IP address.[18] This approach may also provide a platform for automating the manual (and error-prone) process of attestation to the completeness and accuracy of model inventories that all financial firms rely on today to satisfy a requirement of SR11-7.[1]

What can be accomplished by collecting just these three pieces of information for model execution events over a significant period? The answer is 'quite a lot', in combination with a little bit of imagination. The graphical dashboard reproduced in Figure 2 demonstrates various ways of displaying usage tracking information for a synthetic portfolio of 100 models collected for 100,000 randomised execution events at various geographic locations.

The uppermost plot tracks usage frequency for any model over the simulation window; the histogram in the lower right quadrant plots the distribution of execution frequencies, from most to least used models. The global map displayed the lower left quadrant illustrates the distribution of geographic locations based on MAC or IP addresses of the executing processors. This map can be magnified for detailed identification of individual models or animated to illustrate the dynamic changing global patterns of usage for all or any subset of models in the portfolio.

(9) IDENTIFYING AND REPRESENTING COMPLEX MODEL INTERDEPENDENCIES

SR11-7[1] briefly alludes to the importance of understanding model dependencies as part of the challenge of aggregating model risk: 'Aggregate model risk is affected by interaction and dependencies among models.'.

FRB bank examiners have recently begun placing greater emphasis on understanding upstream and downstream dependencies among models, for example, as part of the CCAR inventory requirement. The current practice at nearly every financial firm is to require model owners to identify all model inputs as part of the model documentation. If any model inputs are in turn produced by upstream models, model owners are expected to track each input back to its source and identify any models that contribute to the input in the required model documentation. If this manual process is performed thoroughly all upstream model dependencies should be identified, including multiple levels of upstream dependencies (ie models that are upstream to an upstream model). Nevertheless, this is where the manual process of identification by attestation often falls

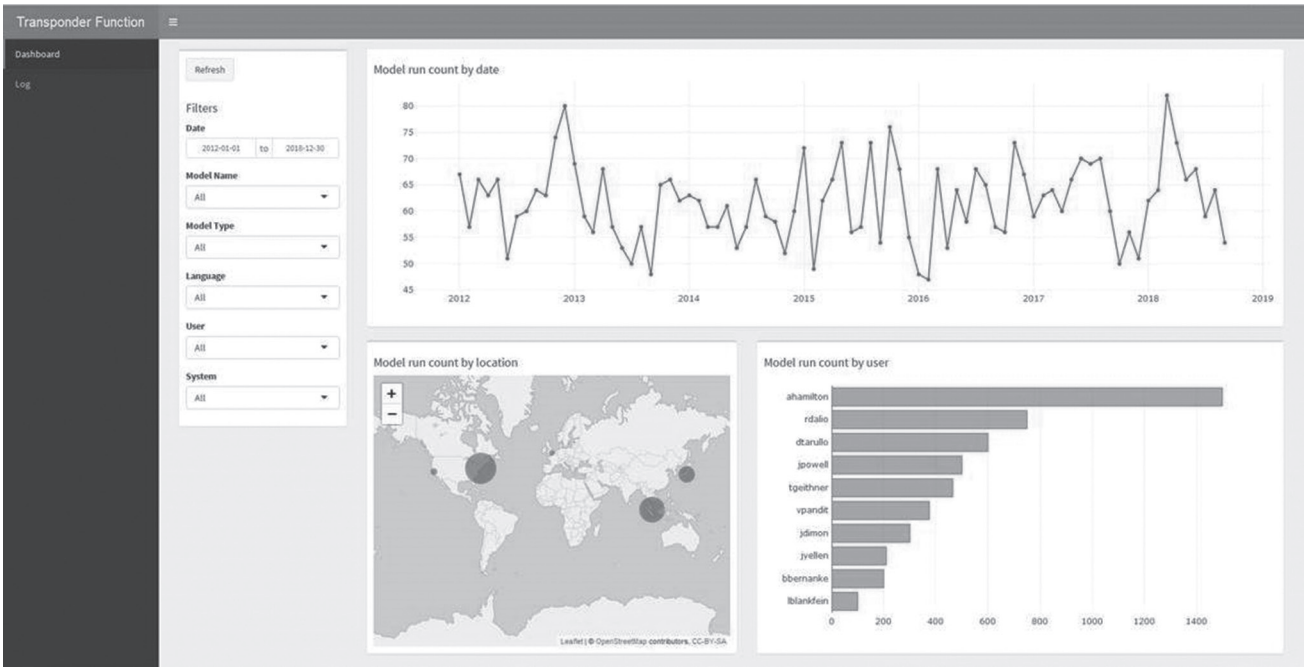

Figure 2: A prototype transponder function and dashboard display used for this simulation were developed in collaboration with the author by David Leonard at FI Consulting, Arlington, VA. The usage plots were produced by collecting three crucial data points for each simulated execution event: ID, timestamp and MAC/IP address (model name is optional). The graphical dashboard was implemented on an Amazon Web Services (AWS) cloud platform

short because there is typically no single line of ownership for upstream and downstream models. Model owners may trace their input data back to its first generation of upstream dependency, but often will neglect to trace any inputs to the upstream models to determine if there are second or third generations (levels) of upstream model dependencies. An alternative methodology that will automate the capture of all upstream dependencies by attaching an identity token to the output produced by each model has recently been proposed the author.[17] This paper presents an argument for replacing the manual attestation process that all firms currently rely on for identifying model interdependencies with a tracking scheme that is based on actual model execution sequences that are captured by passing embedded identity tokens from upstream to downstream models.

A related challenge is how to represent the multiple generations of model interdependencies in ways that can be easily understood. Dependency maps can be generated in a variety of forms, the simplest being a list of all the pairwise relationships that have been identified by model owners and validators (or by an identity token-passing scheme such as one proposed by the author[17]).

As the number of models and complexity of interdependencies increases, a linear list of each of the pairwise dependencies will rapidly become unwieldy and difficult to visualise by simple inspection. Hierarchies can be a useful way to diagram one-to-many dependency relationships, but also do not lend themselves readily to many-to-many interdependency relationships as they quickly become cumbersome and convoluted.

The type of graphic that can most effectively capture the complexities of multiple generations of many-to-many interdependencies employs a network graph theoretic approach. Figure 3 provides one example of complex network representations using hypothetical data and relationships. It was created with Gephi, an open source network visualisation package written in Java. Because this type of graphic scales readily with size and complexity, thousands of

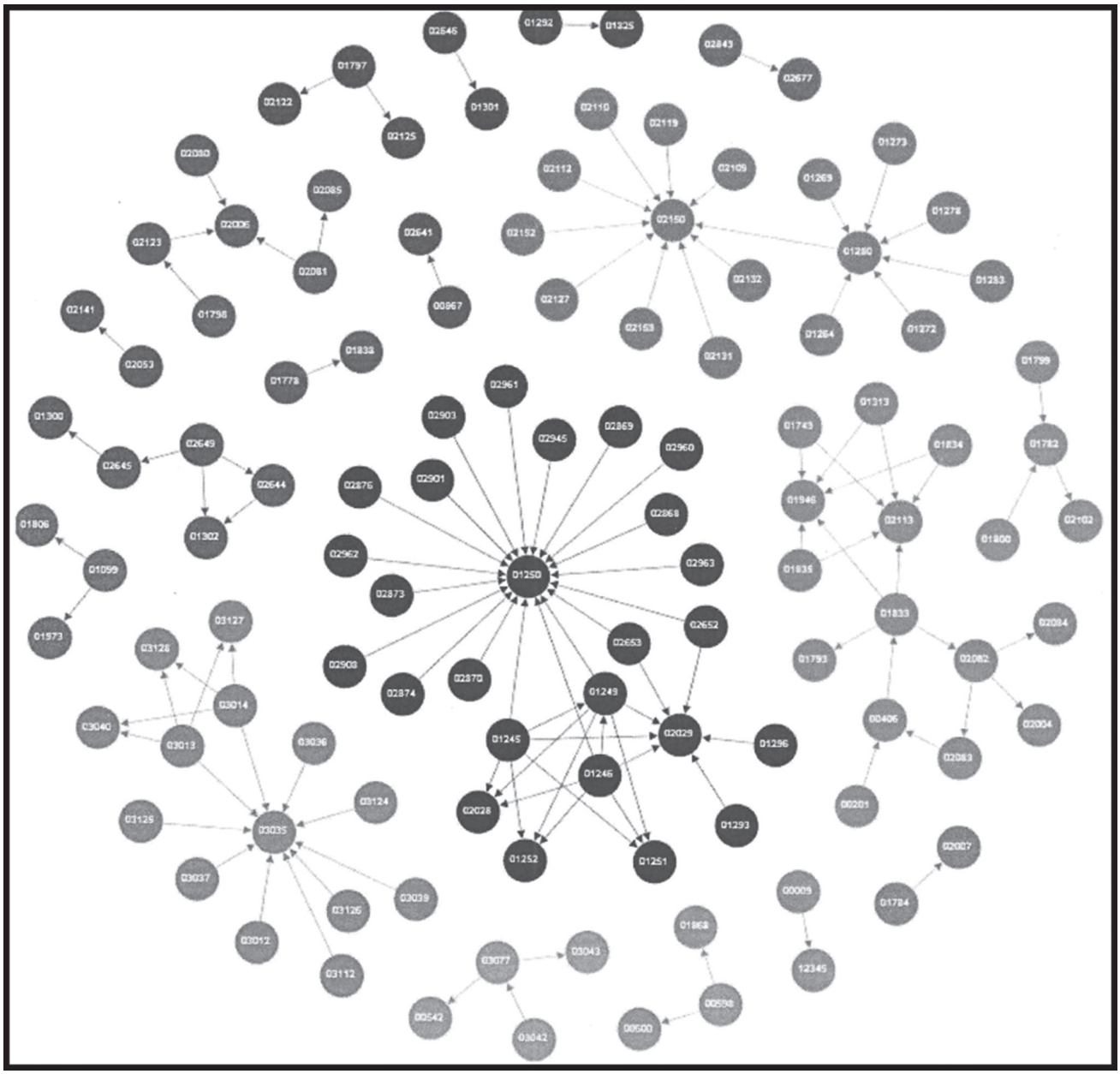

Figure 3: An example of a hypothetical network graphic of complex model interdependencies. Each node represents a single model and can be sized and colour-coded according to complexity, materiality, risk-tier or other indicative model metrics. The node labels are pseudo model IDs created for this example. This graphic was produced with Gephi, an open source network visualisation package written in Java

models with hundreds of complex interdependencies can be represented in a single plot that can be scrolled and zoomed to exam minute details. Each node in the network graph represents a single model; each arrow represents a single pairwise dependency relationship with the arrowhead attached to the downstream member of the pair. Nodes can be sized and/or colour-coded according to complexity, materiality, risk tier or other useful model indicatives.

To create such a network graphic it is necessary only for MRM to compile a list of pairwise

upstream/downstream model dependencies in the form of vectors (eg a→b, b→c, a→c, b→d, etc).[19] Gephi combines all of the pairwise dependencies into the type of network graphic reproduced in Figure 3, readily displaying multiple generations of many-to-many interdependencies. This example uses a globular galaxy configuration that places the model nodes with the greatest number of dependencies in the central regions and those with the fewest in the peripheral arms of the galaxy. Many other types of display configurations are possible.[20]

(10) VALIDATION OF EXPERT JUDGMENT MODELS

Models that rely on expert judgment are also subject to the SR11-7 mandate for validation:

> The definition of model also covers quantitative approaches whose inputs are partially or wholly qualitative or based on expert judgment, provided that the output is quantitative in nature.[1]

Expert judgment, as the name implies, embodies the opinion of an individual who is highly qualified to make quantitative estimates that can be used as an input to a model that produces a quantitative output. An example of how expert judgments are formulated could be the scenario analysis workshops that are used to estimate various types of losses owing to operational risk (OR).

Typically, a group of five to ten senior managers who are familiar with the firm's history of operational risk losses are appointed to participate in a firm's scenario analysis workshop, which usually takes place over several days. The workshop participants are provided with information available from anonymous external industry resources, such as ORX[21] or ABA[22] consortium data, about the size of losses realised by other firms of comparable size in seven event type categories of operational losses defined by Basel II.[23] After becoming familiarised with the firm's internal history of OR losses in each of the seven event types and informed about losses experienced by peer firms provided from external loss data resources, the workshop attendees may be tasked with estimating quantitative values for 1-in-5-year, 1-in-10-year, 1-in-30-year and 1-in-100-year losses for each of the seven event types. The results of the scenario workshops are then used in combination with the firm's own internal loss history as inputs to the firm's operational risk model(s) to produce a quantitative estimate of the firm's operational risks.

A scenario workshop for carrying out an operational risk exercise is a classic example of the use of expert judgment as input to a quantitative model and it presents validators with unique challenges, since part of the input to the operational risk model is formed by expert judgment. The dilemma for validators is simply put: how can the opinions of 'experts' be validated? A facetious response would be to create a second panel of comparable 'experts' to use as a benchmark comparison against the first panel, the equivalent of independent replication of a computer model implemented in software. Unfortunately, it is highly unlikely that two different panels of experts would arrive at comparable estimates of operational risk in all seven categories with a degree of numerical accuracy that would be required for validating quantitative models.

How, then, to validate model inputs based on expert judgment? This is one of the major challenges faced by model risk managers today. One approach that has been adopted by a few firms is to verify the process by which the expert opinions are produced. This approach, which is more correctly referred to as process verification rather than validation, would involve certifying the selection process and qualifications of the workshop participants, assessing the validity of the external data used to inform the experts about losses experienced by other firms, and broadly reviewing the process by which the workshop participants arrive at the various 1-in-*x*-year losses for each of the seven event types. Process verification should also confirm how the resulting scenario data is used as input to the firm's operational risk model(s).

Since process verification requires a different set of skills from those required for traditional validation of quantitative models, firms that pursue this approach to verification of qualitative model components should consider creating a team of

process verification specialists that is distinct from the quantitative validation teams.

(11) VERIFICATION OF MODEL INPUT DATA QUALITY

Data quality is a topic that is often overlooked by validators focused on quantitative modelling methodologies; however, it should be obvious that the performance of any model that requires some quantitative input is heavily dependent on the quality of the data provided to the model at runtime. Time series data, for example, often have missing values or outliers that result from erroneous data and need to be assessed and 'scrubbed' to replace missing values and assess three sigma or greater outliers for veracity. Given the large amount of data required by many complex models, assessment of the quality of input data should be part of any rigorous model validation process. Data quality assessment and scrubbing, however, require a different set of skills from those typically possessed by the quant-mathematicians who develop and validate today's complex models.

Rather than requiring FLOD and SLOD quants to invest resources in verifying data quality, it is probably a more effective solution for firms to hire skilled data specialists and to form data verification and scrubbing teams separate from the model development and model risk management organisations. Under this scenario, model and validation documents for each model should attest that the quality of all data inputs to the model have been assessed, remediated whenever necessary and verified for appropriateness for use by the receiving model or models.

It is possible that if the data verification and model teams reside in highly separate organisations, neither side will have a sufficient understanding of the other side's requirements; that is, data specialists may have little or no understanding of the assumptions underlying the downstream models that receive their data, while model developers and validators may likewise have little or no understanding of the assumptions underlying the construction of the inputs their models receive. This not an uncommon situation and can be addressed by mutual education of model developers and data verifiers. This type of mutual awareness education is most effectively overseen by the firm's MRM.

(12) TIERING MODELS BY RISK RATING

SR11-7[1] calls for firms to develop quasi-quantitative measures of model risk:

> Model risk should be managed like other types of risk. Banks should identify the sources of risk and assess the magnitude. Model risk increases with greater model complexity, higher uncertainty about inputs and assumptions, broader use, and larger potential impact. (p. 4)

SR11-7[1] also allows for a risk-weighted approach to model validation:

> The range and rigor of validation activities conducted prior to first use of a model should be in line with the potential risk presented by use of the model. (p. 10)

Many model risk managers initially concluded, incorrectly, that the FRB/OCC guidance called for assigning a monetary value to model risk; however, FRB managers have clarified that the FRB does not expect model risk to be reported in terms of monetary impact, in the same way that market, credit and operational risk typically are. Assessing model risk remains one of the most difficult challenges confronting today's model risk managers because it is, at best, achieved by combining a majority of qualitative with a minority of quantitative measures in an ad hoc manner. This process will be described in more detail.

For the initial risk assessments of new models, banks have settled on a score card type of approach to estimating model risk in a form than can be more easily aggregated, such as using the results of a well-designed set of questions about models that are submitted to model owners and part of an annual review process. The questionnaires can be quite lengthy with anywhere from 10 to 100 questions, some of which are quantitative ('How many high-risk findings resulting from the initial validation?', 'How many validation or internal

audit findings are still open?', 'What is the model's anticipated materiality?', etc), but most of which require qualitative expert judgment, such as the model's complexity and degree of firm-wide reliance on the model. However they are posed, the questions should be designed to characterise four crucial sources of model risk:

(1) Complexity.
(2) Uncertainty in inputs and assumptions.
(3) Type and materiality of intended use.
(4) Number and complexity of upstream/downstream dependencies.

The resulting responses are typically weighted by relative importance and summed to produce a final scorecard risk rating number which necessarily has included a large measure of expert judgment.[24] The risk ratings for individual models can then be aggregated to a portfolio, business unit or firm level, as discussed in section 13 below.

For risk rating of models already in production, an annual review requirement, many firms rely on various forms of a simplified scorecard approach based on assessments of complexity and materiality to form a two-dimensional tiering matrix. A third dimension commonly used to improve granularity is sometimes defined as reliance or prevalence, a subjective measure of some mixture of how widely the model is used throughout the firm, how great the firm's reliance on the output of the model is or how many other models are dependent on the output. For example, one metric that can be used to estimate reliance is the number of upstream and downstream dependencies a model is known to have, which should be indicated in the inventory entry for each model, or more effectively, by the type of network interdependency graphic shown in Figure 3. Nevertheless, this may tend to underestimate reliance because while model owners should know all of the upstream dependencies, they often are not fully aware of all downstream consumers of the model's output.

Whatever methodology is adopted by a firm, the global model governance policy should clearly describe both the firm's model risk tiering procedures and how the validation protocols have been tailored for models in each tier. The protocols should define a firm's requirements for performing both annual reviews (required by SR11-7) and ongoing monitoring as well as the interval (in years) between full reviews (the term 'full review' is used here to indicate a level of rigour equivalent to an initial validation).

For example, firms may elect to assign a two to three year interval between full reviews for high risk models and three to five years or longer for medium and low risk models, in alignment with the risk-weighted approach allowed by SR11-7. Similarly, the level of rigour of initial validations should be aligned with the trade-offs between the assessed model risks and the available validation resources. It is reasonable to expect that all risk models such those employed for firmwide risk, CCAR and CECL would be classified as high risk and would be subject to a full SR11-7 compliant validation.[25]

A reduced level of rigour for medium risk models might be achieved by reducing requirements for the most time-consuming parts of a full validation, such as independent replication and testing by validators, instead relying on review and assessment of the testing performed by the model developers as a first line of defence (FLOD) against model risk, potentially requesting additional testing to be performed by the FLOD. For low risk models, a validation might only require that the model has documentation, a clearly defined line of ownership (developers and model supervisors), a minimal amount of testing by the FLOD to establish that the model performs as intended and quarterly ongoing monitoring.

Light touch annual reviews of low and medium risk models should at a minimum include a reassessment of whether the low/medium risk rating should be retained (re-using the questionnaires for the initial risk rating) or if the model should be moved to a higher risk tier.

Nevertheless, high, medium and low risk model validations are performed by a firm, the different levels of treatment should be clearly spelled out in a firm's model risk governance policy and procedures. Internal audit will also be likely to review the model risk policies and assess compliance, so it is important for governance to be enforced by model risk management.

(13) ASSESSING MODEL RISK IN THE AGGREGATE

SR11-7[1] explicitly recommends aggregation of model risk:

> Banks should consider risk from individual models and in the aggregate. Aggregate model risk is affected by interaction and dependencies among models; reliance on common assumptions, data, or methodologies; and any other factors that could adversely affect several models and their outputs at the same time. (p. 4)

The FRB/OCC guidance is not prescriptive in describing about how firms should go about defining metrics for model risk that can be aggregated across their various business units and legal entities. This has turned out to be a serious challenge for firms with large inventories of complex models as defining a model risk metric that can be aggregated across all types and classes of models is not a straightforward exercise. As discussed in section 12, no single number or complex metric can capture all aspects of risk across all model types, that in turn might be equated to a monetary risk value.

Unlike market, credit and operational risks, which are essentially numerical, model risk scorecard ratings may be represented by colours rather than numbers. A presentation format that is useful for presenting model risk status to senior management is based on assigning risk ratings into one of three colour buckets: red, amber and green, known as a RAG status. The RAG status for a firm's full complement of models can be summarised for the model oversight committees and board risk committee in the form of heatmaps, organised by business units, portfolios and class of model so that concentrations of high model risk, indicated by clusters of red colour, can be easily identified.

Recent efforts by FRB researchers have focused on innovative ways to supplement qualitative and scorecard assessments of model risk with quantitative components.[26] Nonetheless, current practice for assessing model risk still relies heavily on qualitative processes and expert judgment.

Model risk appetite is a set of thresholds applied to the model risk metrics by a senior model risk review committee and communicated to the board risk committee for review and approval. Thresholds may be specified in terms of a tolerance limit as percentages of models that may have red, amber and green RAG statuses. Heatmaps can be very useful in identifying concentrations of models that exceed RAG red tolerance thresholds, which may then be flagged for remediation.

(14) IMPROVING THE QUALITY OF MODEL AND VALIDATION DOCUMENTATION

Model documentation provided by developers is often found to be poorly written and with significant omissions (such as developmental evidence and stress testing) that render it inadequate to support a rigorous validation effort to the standards set by SR11-7. The most common underlying cause of insufficient model documentation is the unfortunate reality that model developers are rewarded for initiating models, implementing them in software, performing rudimentary testing for soundness and suitability for purpose and putting their models into production. What model developers are not typically rewarded for is producing publication quality documentation, a rare achievement in the 20-year experience of this author. Good documentation is not just that which is complete and detailed enough to support independent replication by validators without recourse to dialogue with the developers; good documentation should also be logically structured so that compliance with SR11-7 requirements[1] are satisfied (often accomplished by performing gap analysis between model documentation templates and SR11-7). In addition to the above, exceptional documentation would also be written with clarity and parsimony with attention to proper English grammar. For example, long, convoluted sentences should be broken up into shorter sentences; the use of personal pronouns should be avoided if possible (eg it may not be obvious to downstream readers who 'we' refers to); use of the passive voice ('It was found that the model failed testing …') should minimised in favour of active voicings ('Testing identified model weaknesses …).

Probably the most qualified judges of model documentation quality are the validation team

members, since they must rely on the FLOD documentation to perform an effective independent review. Validation documents typically include an assessment of the model documentation received from FLOD in the form of a check list ascertaining whether the documentation meets the content requirements of SR11-7[1] (such as developmental evidence, underlying model assumptions, detailed description of the methodology and so on). The quality of the writing is usually not included in such reviews, however. If validation teams rate the quality of documentation received from development teams and treat any serious deficiencies as findings requiring remediation, such practice will provide an incentive for developers to improve their documentation.

An even more innovative practice for improving documentation quality would be for the chief model officer to become a member of the salary and bonus review committees for the development groups. This practice is in place in at least one Tier 1 bank (Wells Fargo). If developers realise that their bonus treatment will be affected by the quality of the documentation they prepare, they will have a direct and measurable incentive to produce quality documentation.

Both model and validation documentation are areas that can have the potential to benefit enormously from language-parsing machine learning programmes. Machine learning applications in language processing have evolved to the point at which they can do substantially more than assess model and validation documents for completeness, clarity, structure and so on. More advanced machine learning applications are moving towards developing the ability to structure and compose draft documentation based on assimilation of supporting documentation along with input from modellers and validators.

The day may seem to be a long way off when a model document can be submitted to an ML programme that produces a validation document within an hour, complete with assessment of the model documentation, independent testing and detailed comparisons with machine learning-generated benchmarks. Nevertheless, progress in the field of machine learning is accelerating at a surprisingly rapid pace. Consider that ten years ago only a few visionaries believed they would live to see driverless cars operating on public roadways, yet in 2018 Google's Waymo autonomous vehicle was awarded the first permit to operate without a safety driver on California public roads.[27]

SUMMARY

Model risk management, when applied over the entire model life cycle, is a relatively immature discipline compared to model validation, a practice most finance firms have had in place for 10 to 15 years. Because it is a relatively new discipline, model risk managers are being confronted with a new set of challenges that traditional model validation has not had to contend with. Meeting these new model risk challenges often requires creativity, support from a firm's senior management and additional resources.

This paper has attempted to identify and discuss approaches to the 14 greatest challenges facing model risk managers today. Of the 14, the 8 most demanding of time and resources are: (1) developing a firm-wide comprehensive and holistic model risk governance framework; (2) building a complete and accurate model inventory database; (3) incorporating machine learning and big data into model risk management; (4) developing procedures for reporting model risk in the aggregate to senior management; (5) risk tiering models; (6) accurately tracking model usage and mapping model interdependencies; (7) developing sound practices for distinguishing calculators, tools and near-models from true models; and (8) improving the completeness and overall quality of both model and validation documentation. These are the areas that firms should be dedicating most of their MRM efforts to today and in doing so consider that possibility that the time has come to think seriously about going beyond the baseline for MRM established by SR11-7/OCC2011-12 in 2011.[1]

Acknowledgments

The author is deeply indebted to David Leonard and the management of FI Consulting in Arlington, VA for developing the transponder prototype, simulator and graphical dashboard displayed in Figure 2.

The author is also grateful to Richard Harris for creating the synthetic dataset used for the Gephi model interdependence network graphic displayed in Figure 3.

References and notes

1 FRB/OCC (2011) 'Supervisory guidance on model risk management', 4th April, Supervisory and Regulatory Letter SR11-7, also distributed as OCC2011-12, Washington, DC., available at: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm (accessed 16th February, 2019).
2 FRB (2015) 'Federal Reserve supervisory assessment of capital planning and positions for LISCC firms and large and complex firms', SR15-18, Board of Governors of the Federal Reserve System, Washington, DC, available at: https://www.federalreserve.gov/supervisionreg/srletters/sr1518.htm (accessed 16th February, 2019).
3 Identifying, recruiting, training and retaining staff with the necessary skills and knowledge base to become competent MRMs also poses a particular challenge for human resources departments, one that is beyond the scope of this paper.
4 Bennett, D. (2018) 'Governance: Creating the professional framework for model risk management.' Stress Testing USA 2018, 6th November, Center for Financial Planning Conference Presentation, New York.
5 Goldberg, M. (2017) 'Much of model risk does not come from any model.' *Journal of Structured Finance*, Spring, pp. 32–37.
6 Bellman, R. E. (1957) *Dynamic Programming.* Princeton University Press, Princeton.
7 OCC (2000) 'Risk modeling: Model validation', 30th May, OCC 2000-16, Office of the Comptroller of the Currency, Washington DC, rescinded and replaced in 2011 by OCC2011-12, available at: https://ithandbook.ffiec.gov/media/resources/3676/occ-bl2000-16_risk_model_validation.pdf.
8 Khemani, D. (2017) 'A first course in artificial intelligence', McGraw Hill Education, Mumbai, 1st edn.
9 Mitchell, T. M. (2017) 'Machine learning', Mcgraw Hill, Manhattan, NY, 1st edn.
10 Nilsson, N. J. (1998) 'Artificial intelligence: A new synthesism', Morgan Kaufman, Burlington, MA.
11 Norvig, P. and Russell, S. J. (2015) 'Artificial intelligence: A modern approach', Pearson Education India, Carmel, Indiana, 3rd edn.
12 Thompson, W. (2017) 'Statistics and machine learning at scale: New technologies apply machine learning to big data'. SAS Institute, Cary, NC.
13 In a rare example of using a technology to solve a problem created by the same technology, some very recent R&D efforts by the Defense Advanced Research Projects Agency (DARPA) have been directed at developing machine learning validation models that are themselves implemented in machine learning methodologies. Explainable artificial intelligence (XAI) is in its nascent stage of development, but the concept is based on parsing through the code of another machine learning model in order to provide explanations for the outcomes the other model produces. XAI can then document the steps it performed to reach its conclusions. Eventually it may be able to perform and document complete validations of machine learning models.
14 A far better solution from the firm's point of view is to require the vendor to provide proof of the ML model's independent validation to SR11-7 standards and fitness for the client's purpose as a condition of sale. The client must still validate its internal use of the ML model, but the onus for the major part of the validation of the vendor's framework should correctly fall on the vendor.
15 It is possible for developers to adopt benchmark models developed by validation as their champion models, but this practice creates a dilemma over validation of those adopted models, which should not be performed by the validation team that developed them owing to a conflict of interest. One option is to engage external resources (ie consulting firms) to perform the validation of adopted benchmarks, at additional cost to the firm. Another option might be for a different team within the firm to perform to the validation, assuming a team with the appropriate skills and sufficient independence can be engaged.

16 Clients should consider placing the burden proof on the vendor to provide documentary evidence of the model's independent validation to SR11-7 standards and an attestation its fitness for the client's purpose as a condition of sale. This strategy can relieve the client firm of the major part of the onerous task of validating a 'black box' model. The client will still have to validate its specific application of the vendor's model, ie demonstrate the fitness of the vendor's model for the client's intended use, but the major part of the validation effort should correctly fall on the vendor.
17 Hill, J. R. (2018) 'Shouldn't a model "know" its own ID?' *The Journal of Structured Finance*, Fall, pp. 89–98.
18 A media access control (MAC) address is the hardware equivalent of an internet protocol (IP) address and uniquely identifies the processor executing the model code.
19 The upstream dependency information is typically available from model and validation documents, although these are often found to be incomplete, as model owners typically trace their required inputs only back to the first generation of upstream models and not to other models that may be upstream to the first generation. This is a common oversight at many firms that have no single line of ownership from upstream to downstream models. A more sophisticated approach based on passing identity tokens between models has been proposed by this author in a previous paper (see Hill, ref. 17 above).
20 For more information about Gephi and similar network visualisation tools, see: https://en.wikipedia.org/wiki/Gephi (accessed 16th February, 2019).
21 See https://managingrisktogether.orx.org/activities/loss-data (accessed 16th February, 2019).
22 Operational Risk Consortium (2015) 'Loss Data Collection & Reporting Guidelines, April, American Bankers Association, Washington, DC, available at: https://www.aba.com/Products/Surveys/Documents/ABA%20Loss%20Data%20Collection%20and%20Reporting%20Guidelines%20(approved%20April%202015).pdf (accessed 16th February, 2019).
23 The seven event types are: (1) internal fraud (IF); (2) external fraud (EF); (3) employment practices and workplace safety (EPWS); (4) clients, products and business practice (CPBP): (5) damage to physical assets (DPA); (6) business disruption and systems failures (BDSF); and (7) execution, delivery and process management (EDPM). For additional details, see the Wikipedia review at: https://en.wikipedia.org/wiki/Operational_risk (accessed 16th February, 2019).
24 It is critically important that the weighting schemes devised by the SLOD are not shared with model owners (developers, supervisors and users) lest they devise strategies to game their answers in order to obtain a lower risk rating, and therefore incur less overhead for validation engagements.
25 Fully compliant validations include complete assessment of any developmental evidence provided by developers (what modeling options were considered and the rationale for employing the chosen approach), assessment of underlying assumptions and suitability of the model for its intended applications. Compliant validations should also include benchmark comparisons, independent replication and testing, including stress testing, assessment of documentation and the posing of effective challenge by the validation team, and all known upstream and downstream dependencies. The validation should also state any restrictions on use of the model and a final appraisal of the model's fitness for purpose.
26 Brastow, R. and Brotke, L. (2018) 'Assessment of model risk in the aggregate: Contributions of quantification', *Journal of Risk Management in Financial Institutions*.
27 See: https://www.marketwatch.com/story/alphabets-waymo-first-to-test-driverless-cars-in-california-without-human-drivers-2018-10-30 (accessed 16th February, 2019).

APPENDIX A: 'IT'S NOT A MODEL BECAUSE …'

Ten classic evasions model stakeholders may use to avoid the validation process

On 12th November, 2018, Dennis Bennett, founder and CEO of the Model Risk Management International Association (MRMIA) professional association, sponsored a contest for the most ridiculous excuses attendees at a London MRM conference had heard from model owners/developers about why their quantitative software implementation should not be considered a model and therefore not subject to validation. The ten best are included in the list below, ranked (qualitatively) by increasing degree of absurdity. All of them are ridiculous — some are funny:

(1) It's not a model because it is just a calculation.
(2) It's not a model because we are still building it; this is just the first draft.
(3) It's not a model because it has a minimal impact.
(4) It's not a model because it is just a spreadsheet, and spreadsheets (end user computing — EUCs) are, by definition, not models!
(5) It's not a model because an (upstream) input into a (downstream) model cannot itself be a model.
(6) It's not a model because it's a vendor-supplied solution, and vendor solutions are the vendor's responsibility, not ours.
(7) It's not a model because it is only used for accounting, and is subject to a high level of qualitative adjustments.
(8) It's not a model because we only use it for reporting to regulators.
(9) It's not a model because we can't explain it to senior management.

Contest winner! The most ridiculous reason that it's not a model is:

(10) It's not a model because we don't have resources and we don't want to increase our model count and hence we really prefer not to consider this to be a model.