

# MANAGEMENT OF AI AND MACHINE LEARNING IN THE RESIDENTIAL MORTGAGE INDUSTRY



# CONTENTS

“Any sufficiently advanced technology is indistinguishable from magic.”

– Arthur C. Clarke, author of Profiles of the Future; An Inquiry Into the Limits of the Possible

Introduction . . . . .	3
AI/ML Defined . . . . .	4
AI/ML vs. Traditional Software . . . . .	5
Measuring Model Performance . . . . .	8
Quality Assurance and ML Ops . . . . .	9
Model Drift . . . . .	11
Combating Drift: Model Tuning . . . . .	13
Architecture, Data Governance and Information Security . . . . .	14
Stakeholders . . . . .	15
AI/ML and Risk Management . . . . .	16
Explainability, Explained . . . . .	17
The Regulatory Landscape . . . . .	20
Summary. . . . .	22
Addendum / Discussion of F1 . . . . .	23
Suggested Reading . . . . .	24

## Introduction

In the residential mortgage industry, as in popular culture, artificial intelligence (AI) is portrayed as near magic. Adherents tout many abilities that are fast exceeding our comprehension. In the colloquial imagination, AIs can drive cars autonomously; read and interpret everything; translate languages on the fly; write scientific papers about themselves; testify in court as expert witnesses; cure diseases and – if that wasn't enough – explore the universe from quantum to cosmic scales, faster than the human mind could possibly absorb.

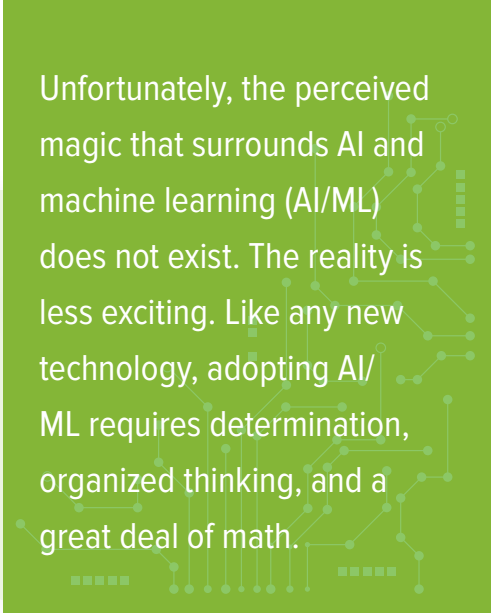
Unfortunately, the perceived magic that surrounds AI and machine learning (AI/ML) does not exist. The reality is less exciting. Like any new technology, adopting AI/ML requires determination, organized thinking, and a great deal of math.

There is no doubt that AI/ML technologies are transforming our world, but their capabilities are much more nuanced than portrayed in films, media and advertising hype. For many reasons, some of which we will explore in this document, actual solutions using AI/ML are attacking seemingly mundane problems like finding important data elements in documents. Solutions that eliminate costly human work, e.g. making subjective judgments about human behavior as it pertains to mortgage lending, are likely to be years away from any significant adoption.

Traditional software is deterministic – it produces the same output from a given initial state. AI/ML is non-deterministic. The same input into an AI/ML system doesn't always yield the same result. AI/ML systems exhibit characteristics of human intelligence, including the ability to learn and solve problems in non-deterministic ways – starting with chaotic initial states and providing results that rely on statistics and probability.

AI/ML will not necessarily produce a consistent output from a given initial state, but it can provide useful answers to questions it is designed to answer. For example, every human's voice is slightly different. The sounds we each make when saying a given word vary in pitch, inflection, pace, and in other ways. Voice recognition "listens" to enough people talking to be able to determine that a sound you make might be a specific word or pattern of words. The word pattern with the highest probability wins and is assigned to your utterance. In essence, AI/ML is often highly sophisticated educated guessing.

Since AI/ML is increasingly being leveraged to solve practical problems in the mortgage industry and is a new technology domain for financial institutions, lenders' AI/ML strategy must comprehend and take into account the different risk management and regulatory compliance challenges this technology presents. Existing financial



Unfortunately, the perceived magic that surrounds AI and machine learning (AI/ML) does not exist. The reality is less exciting. Like any new technology, adopting AI/ML requires determination, organized thinking, and a great deal of math.



**Key Point:** *This paper is intended to assist mortgage industry managers and executives as they consider how to safely benefit from AI/ML. AI/ML is different enough when compared to traditional software that lenders will need to develop new management disciplines to meet this challenge.*

accounting rules, government regulations, business process standards, and auditing methods are based on testing known inputs versus expected outcomes. Such testing is not viable for AI/ML solutions, and no single standard yet exists for alternative testing approaches.

In addition to the lack of clear AI/ML testing and audit standards, AI/ML has recently been subject to heightened concern by regulators. While regulatory guidance is not definitive as of this writing, model **explainability** and **transparency** are focus areas for regulators and other stakeholders.

The principal regulatory concern centers on fair lending: proving that no consumer is harmed by bias in an application of AI/ML in the lending process. In this document and **Dark Matter** discourse, the regulatory risk associated with the possible introduction of bias that can result in unfair treatment of a consumer is called “Regulatory Bias Risk”.

## AI/ML Defined

There are many sources of information about AI/ML. In fact, for most businesspeople, there is too much information. Internet searches for either term will yield thousands of articles, treatises, research papers and software product offerings. This article will use the term AI/ML to refer to software solutions that use one or more of the methods under the general AI/ML umbrella. Definitions in this domain are often fluid so below are a few simplified terms to help clarify:

- **Artificial Intelligence**, also called machine intelligence, is a general term or patterns of thought. AI is characterized as using an “intelligent agent” to perform a given task.
- **Machine Learning** involves the creation of algorithms and methods that can “learn” or get better when provided with more data. Machine learning can take several forms, but this paper will refer most often to supervised learning systems. The recommendations and conclusions stated will apply equally to other machine learning approaches.
- **Natural Language Processing (NLP)** allows machines to read and use human languages. Document classification based on the words identified on a page is an example of NLP.
- **Deep Learning** is a type of machine learning algorithm that uses multiple layers of model elements to derive a solution.
- **Predictive Analytics** is a broad category for application of techniques including data mining, modeling, and machine learning in business to seek patterns in potentially large sets of data to make predictions about future or unknown events.

**Key Point:** AI and Machine Learning is a general category that includes many disciplines. The category has been a focus of academic research and technology investment since the dawn of the computer age. AI/ML capabilities have seen some adoption in the financial services industry, but currently available solutions are just scratching the surface. Mortgage lenders need to be sufficiently conversant in the language of AI/ML so that they can identify opportunities and manage risks appropriately.

## AI/ML vs. Traditional Software

Traditional software technologies, including business rules management (BRM) systems, such as Black Knight’s decisioning engine, are deterministic: inputs are directly traceable to precise and expected outputs. See Figure 1.

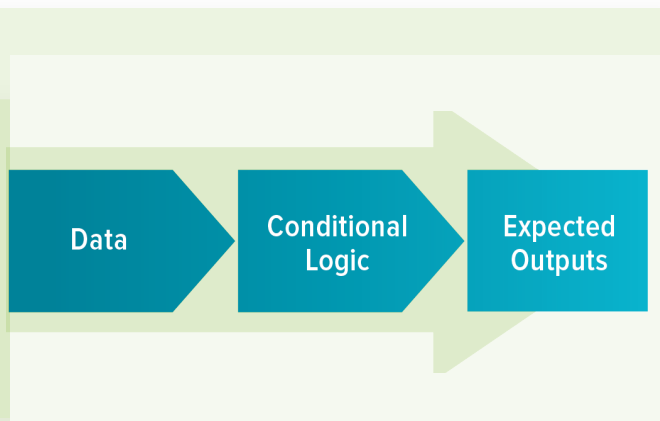


Figure 1 – Deterministic Software

Traditional deterministic software remains appropriate in meeting most business requirements. No matter how complex it might become, traditional software’s conditional logic can always be tested, even if the number of test permutations is potentially vast. The software development life cycle relies on this premise: software testing is definitive, and a set of properly constructed tests will confirm the validity of delivered code to its intended purpose.

AI/ML technology is different. It can address business problems that do not yield deterministic results. When inputs are chaotic and there are a range of possible outcomes, a probabilistic approach is required. AI/ML algorithms are applied to often-large sets of data, and the outcomes are evaluated for their fitness in solving the problem. Algorithms are organized to be deployed for a specific purpose that creates a “model” for the intended solution. Models always answer a specific question based on the inputs they are given. Data scientists need to be adept at defining the question to be answered before beginning model development.

Take the application of AI/ML to classification of loan documents as an example (Figure 2). In this example, the question assigned to the model is “Given a document and a list of types, what type should be assigned to the document?”

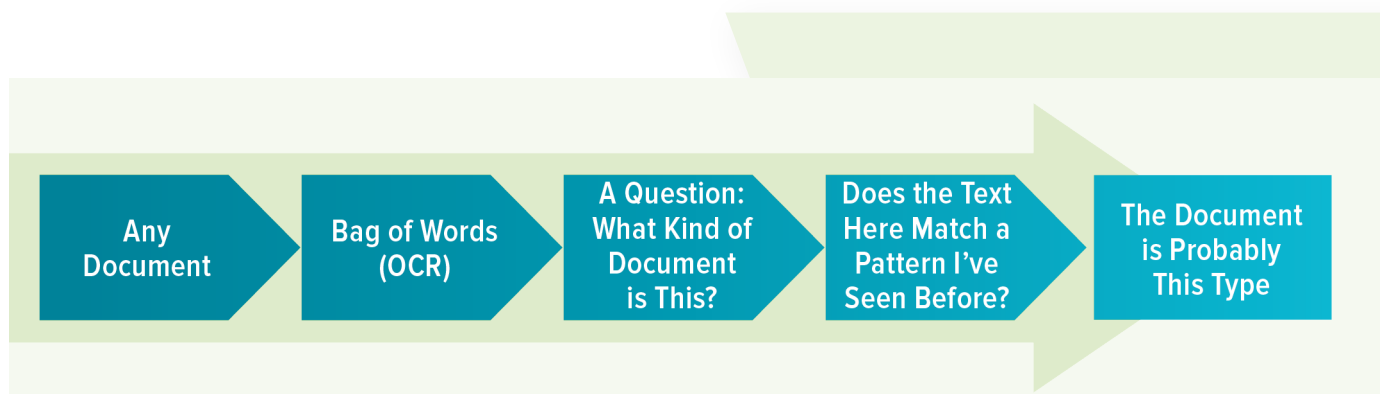


Figure 2 – AI/ML problem solving



In Figure 2, Black Knight’s originations-based AI platform, [AIVA](#)<sup>®</sup>, receives a package containing uncategorized documents. The goal is to assign a document type to each page received – classifying the documents in the package.

After using optical character recognition (OCR) to determine the text that each document contains (a bag of words), the system looks at the contents of each page to try to find patterns that can be used to classify it. The problem is that document content is highly variable even within a given type, and there are hundreds of document types in the mortgage business. Additionally, there could be custom or non-standard documents contained within the initial set of uncategorized documents.

To perform the classification task, a natural language processing model can be used to predict a document’s type based on patterns in the text. Cues like the presence of common words, the order of those words in the document and word adjacency (combinations of words) are used to make the prediction.

The result, which is a predicted document type, is not certain. The model might get it right or it might get it wrong, just as a group of humans looking at a large population of documents would. It is this uncertainty, the probability that the model’s predictions are correct, that creates a software development conundrum: how do you manage probabilistic systems when traditional software quality measurement focuses on producing the correct answer every time? The answer is that you cannot manage them that way. You can only measure how often the model gets the right answer when applied to real-world data. This raises two further implications:

1. You need a way to gather a large enough set of the correct answers to the question that the model intends to solve, this is also known as ground truth.
2. You need a way to measure how well a model worked by comparing its output to ground truth, and by inspecting production results for “rightness”.

Gathering and maintaining ground truth is a substantial barrier to model management, as it usually requires human effort: model trainers putting eyes on screens to do the work the model is intended to automate. This work is repeated to create a “large enough” set of high-quality examples of correct answers to feed the model training process. “Large enough” is a “one size fits none” description; training data requirements vary widely from one use case to another, since the data must represent the variability expected in production. This presents a cost and resource challenge that AI/ML managers must answer with careful resource planning.

**Key Point:** *Using ground truth to train and then measure model rightness is one AI/ML analog for traditional testing, with the significant difference that you do not expect the software to be right all the time.*

When a model has a specific objective – for instance, classifying a document – measuring rightness is paramount. The model training process relies on ground truth to determine model performance. If you have a large enough data set that includes both inputs and answers, you can measure how well a model performs through simple comparison. Humans can count how often output matches expectation to measure accuracy. If done in a well-controlled way, the human comparison process yields both high-quality ground truth and performance statistics.

AI/ML teams use ground truth for both training and measuring model performance. A model is trained using the ground truth, and then the model is fed the question part of the ground truth. When the model result matches the human-generated answers, the model is right. Running many of these tests yields the statistics needed to judge the model's rightness.

The performance of a model as a solution depends not only on how often it gets the right answer, but also on its consistency, measurability and mutual exclusivity among other things. Consistency is a model attribute that indicates how well a model improves given more training data. Measurability expresses how meaningful performance statistics are in the context of the problem being addressed. Mutual exclusivity means that the model yields only one result for a given input.

These three attributes are largely defined when setting up the problem to be solved. For example, if a classifier is required to identify a document as one thing in one circumstance, and another in a different circumstance, the model will be inconsistent, have problems with measurability and, by design, not be mutually exclusive. Such a model will fail. In such a case, the problem needs to be reworked to solve the consistency, measurability and mutual exclusivity problems.



There are cases, such as with some deep learning models, where the goal is to identify patterns that are not otherwise perceivable among large amounts of data. In such cases, rightness may be unimportant early in the model development process. Early experimentation leads to interesting insights, which can be explored more deeply in subsequent experimentation. In these cases, ground truth is somewhat replaced by well-curated data. In the long run, however, a measure of rightness will be needed to certify model suitability to a business purpose.

The focus on math and statistics is another difference between the practice of traditional software development and AI/ML. Where traditional software is engineered, AI/ML is largely the province of data science. This language difference is not a nuance of practitioners, it is a reasonable way to distinguish the craft. Just as with the physical sciences, data scientists use experimentation to create a model of a thing, and then to apply the model to determine its veracity. The data scientist job title may be misleading in some organizations – it is sometimes erroneously applied to traditional database management and visualization roles.

**Key Point:** *AI/ML IS different – it requires different approaches, systems, skills, and controls versus managing traditional software development.*

In an AI/ML practice, a data scientist's role focuses on deep statistical analysis and math skills, experience with a range of software tools, as well as knowledge of the business problem being attacked. They must understand big data sets, be adept at data mining and advanced storage technologies, and be able to explain what the data is telling them about the problem. Finally, they need to have a scientific mindset – learning through experimentation; problem solving through collaboration; willingness to abandon work that does not bear fruit, and an inventive approach to applying data insights to address challenges through software. This exploration of problems with goal-driven experimentation requires an artist's touch – blending technical skill, curiosity and creativity.

## Measuring Model Performance

Because of its probabilistic nature, the performance of an AI/ML system can be explained, but it cannot be tested in the same way as traditional software. The output of an AI/ML model is a “probably,” not a consistently predictable result. No matter how good the model is, it will occasionally be wrong, so any single test is not dispositive even if it has a clear result. The next test using the same inputs could have a different result.

The conundrum: if testing is not dispositive, how does a businessperson gain enough confidence in a model to find it useful? In discussion of AI/ML models, businesspeople naturally seek to understand the “accuracy” of a model – does it get the objectively right answer, and if so, how often?

In the document classification example, a model predicts the most likely document type for a given page or pages. Humans can review the model's predictions and score each result as correct or incorrect. If the model got it wrong, the human could supply the right answer. If these interactions are tracked, the resulting statistics can then be aggregated to determine how well the model performs.

Simply counting how often a model is correct is insufficient. Data scientists need to know how well a model makes predictions and how well those predictions perform. The  $F_1$  statistic (the F-measure or balanced F-score) is helpful since it covers these extra dimensions of rightness, especially for models where direct measurement of predictions is possible. The  $F_1$  statistic is a number between zero and 1; higher numbers indicate better performance. See section 13.1 below for a brief discussion of how  $F_1$  is calculated.



Applying  $F_1$  as a tool for determining model performance is nuanced. In some business cases a minimum  $F_1$  may be stipulated as a requirement for a model to be acceptable. Managers should be aware that this absolutist approach can be misleading and can also needlessly limit the return on AI/ML. This is especially true when very high  $F_1$  is considered a requirement by business teams.



**Key Point:** *Models are not tested – their performance is measured. Statistics about model results must be well understood by managers and stakeholders. High performance is not necessarily required for a model to be useful.*

Managers need to correctly weigh the quest for rightness as they set performance goals for a model. This has always been true of human processes: cost-effective quality management programs are central to managing any business, and such programs rarely expect to get it right 100% of the time. The same is true of AI/ML models. Business value can accrue from the presence of a prediction, even if the prediction has a relatively low  $F_1$ .

Consider a case where a model is used to eliminate stare-and-compare work by extracting data from a document and comparing it with expected data – to check that a W2 belongs to a given borrower, for example. If a model can identify the subject person for the W2 and that name matches expected data, the data can be considered verified, and no human work needs to be done. If a match is not made, then the normal stare-and-compare process can proceed. In a case like this, a relatively low  $F_1$  can significantly reduce the amount of work, with every correct prediction providing business value. The key is in devising systems that use the correct predictions and ignore the incorrect ones.

When it is not possible to know in advance if a prediction is correct – for example when a model is used to replace human data entry – it then becomes important to understand how  $F_1$  impacts the business outcome.

For business managers, understanding the statistics is crucial in deciding when a model is good enough. In some applications – such as when a given output directly impacts a business outcome – teams may require very high  $F_1$ . When the objective embodies lower risk and provides business efficiency, determining a lower acceptable  $F_1$  target should be considered carefully. With all the costs of an AI/ML program, managers should always focus on extracting the maximum business value.

This performance statistic-based approach works in applications of AI/ML where there is a correct answer to the question a model is designed to answer. However, that is not always the case.

## Quality Assurance and ML Ops

Once a model performs appropriately, it is time to ready the model for production deployment. In another departure from traditional software development, quality assurance (QA) testing of the model itself does not add material value, as the data science team has proven that the performance of the model meets business needs. Nevertheless, depending on the complexity of the systems that surround the model, QA has a role in proving that the model operates properly when its data sources and downstream systems are in place. Traditional QA testing is appropriate to this task, just as would be true if a new or upgraded component were being readied for deployment.

While QA's mission is focused on integration testing in its traditional sense, test plans need to consider edge cases and systemic failure modes that are not able to be defined by specific input data test cases. For example, a model may behave erratically under a very improbable set of data. Model output in such a case may be poorly formed, or the model may fail to process at all. Downstream systems need to be programmed to recognize and handle these error states, even if they are difficult to reproduce. The dilemma for QA is that the data set that triggers an error mode may or may not trigger the same error every time. QA and release management teams need to plan for this risk by focusing appropriately on the tails of the input data distribution.

Technical management of model change is another area of divergence between AI/ML and traditional software. To accommodate the differences inherent to model building, training, retraining and deployment, the practice of ML Ops has evolved. ML Ops is a set of practices and tools that adapts DevOps' (development operations) continuous improvement practices for AI/ML model deployment.

Models require some elements of traditional software development. Business goals are translated into software, software needs to be stored and made available for testing, and tested software needs to be merged with existing software and deployed to production. Among other things, DevOps practitioners ensure that documentation and control systems for all these processes meet business, governance, and security standards. Sound AI/ML management adds several responsibilities:

- **Management and tracing of training data** – governance of and change control over ground truth and question data.
- **Experimentation management** – support for data scientists to conduct experiments in a secure, controlled manner.
- **Methods to manage feature engineering** – building a feature of a model, then iterative training and evaluation of the performance of a model with each new feature.
- **Model history** – a solid archive of models for an extended period.
- **Model packaging** – containerization and methods to manage the AI/ML workflow.
- **Model deployment** – since model updates and retraining are frequent, a CI/CD (continuous integration/continuous delivery) approach to development and deployment is typical. Even if surrounding software does not follow a frequent delivery model like CI/CD, the change rhythm of the AI/ML pipeline needs to be managed for rapid change.
- **Model monitoring** – tools and methods to monitor models for failures and model drift.
- **Model risk assessment** – tools, methods and data strategies to identify model bias and other risks.



**Key Point:** AI/ML and related systems require new methods and processes to manage record keeping, software delivery, and risk associated with the delivery of new models, retraining updates and other model changes. Managers need to understand the impact of this new technology on QA and technology delivery operations.

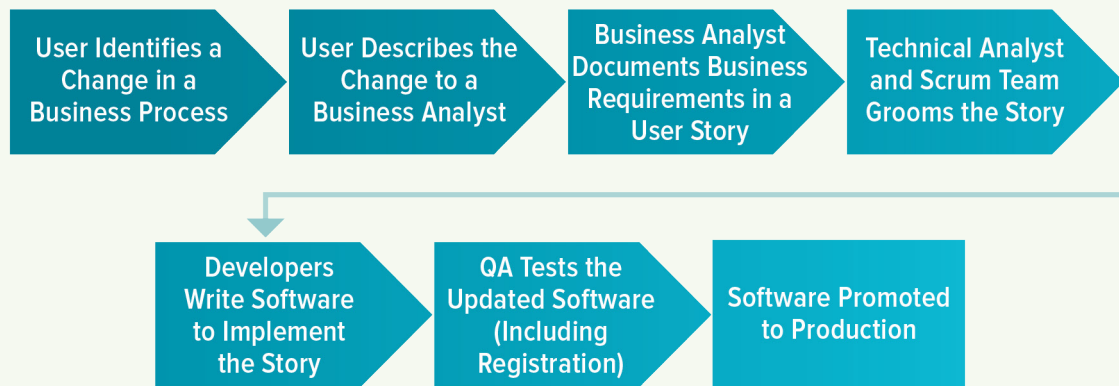


Figure 3 – Traditional software change management

## Model Drift

In managing change in traditional software, the change process is often linear and predicated on the steady state of the system and the ability of the system’s users to identify the need and nature of required change.

Model performance can “drift” over time. When the data stream, presented to a production model changes, a model can fail to adapt to the changes, resulting in more errors. Since a model is expected to be wrong some of the time, end users may see new errors but may not notice the reduced overall accuracy. Model drift is usually only identifiable by examining statistics about the frequency of errors (such as F<sub>1</sub>). Even then, the reasons for poor model performance are not always (or even usually) clear.

One contributor to poor performance could be the introduction of a new model or addition of new data into the data stream presented to a model. For example, when **Dark Matter** added models to classify drivers’ licenses, which in part searched for photographs, we found that existing models for other documents that contained photos were adversely impacted. As one of our team members reported on an early iteration, the AIVA solution thought everything with a photo was a drivers’ license, including several pages of property appraisals, which had previously been classified with high accuracy by AIVA.

A model can drift into poor performance because the problem being solved was not well understood (the basic premise of the model is wrong), or for other less-obvious reasons. Sometimes, to make progress a model needs to be abandoned and an entirely new approach applied to the problem.

**Key Point:** AI/ML managers must have processes in place to monitor model performance. In addition to monitoring  $F_1$  and other statistics related to the model, they should conduct regular reviews of the data used for training to ensure that it is sufficiently representative of production data. Managers need to have processes to quickly evaluate the performance of new models in production and ways to take corrective action if needed.

While it is important to remeasure model performance regularly, managers also need to consider that a model might change how it “perceives” the world over time. This means maintaining awareness in these areas:

- When models are trained, the data used is usually derived from production data, with sensitive information removed or anonymized. Retraining with more data over time is an important measure to prevent model drift or biasing that may be inherent to the original training data. That new training data may change outcomes. Using a vendor with wide industry coverage allows for a variety of input patterns that can keep the models fresh and increasing in accuracy and speed.
- Business conditions could yield changes to the patterns the model is “seeing.” Such changes include shifts that could adversely impact performance, or even invalidate the model or one of its features. In handling documents, classifiers and extraction models can react poorly to the introduction of new document types, changes in document standards, or large shifts in the distribution of documents versus the training data. Natural language processors can be sensitive to regional differences, so managers should be sensitive to mergers or acquisitions in a community where language elements vary. Organizations can change a vendor who provides data, resulting in subtle changes in model performance.
- Testing of new model versions cannot be comprehensive. New models may perform well during training, but drift significantly when launched in production. Managers need to have ways to evaluate the performance of production models to quickly identify if a new model performs poorly. A poorly performing model may need to be removed and replaced by an earlier model.
- Regular contact and trouble-reporting from end users is crucial. If a model begins to drift, its end users will likely be the first to detect the problem. Always investigate reports of higher-than-expected error rates or unexpected behavior. This may seem obvious, but since models are expected to have some errors, it is tempting to dismiss error reports as being within the expected range. Humans are very good at observing that something seems wrong. Listen to your clients.
- Some types of models have humans in the loop to review and validate the “rightness” of answers. In such cases, validation drift related to the performance of teams of reviewers can appear. Managing the training, expectations, consistency, and reliability of what is validated and how human validation is accomplished is also critical. Human behavior is nuanced, and human idiosyncrasies can impact ground truth gathering and validation processes.



## Combating Drift: Model Tuning

Once models begin to drift, they need to be retrained, or tuned. Ideally, the factors that could have caused drift would be understood, but this is not usually a reasonable expectation. As a result, retraining is not just a repeat of a previous training exercise.

The main vectors for model drift are discussed above, but here is a concise breakdown:

1. **Upstream business environment changes.** Some aspect of the business processes that create the production data used by the model has changed, resulting in production data that is structurally different enough from the original training ground truth to cause drift.
2. **Production data distribution changes.** In any business process, the distribution of objects can shift over time. When this happens, former outliers can become significant, and dominant features can reduce in frequency, resulting in lower overall  $F_1$  as well-performing aspects of a model are overwhelmed by features that were previously less significant.
3. **Upstream system changes.** Some part of the systems that feed data to the production model has been changed, introducing data variability that is not expected by the model.

The retraining process begins with assessments and planning.

- Examination of the latest measurement statistics
- Analysis of original ground truth versus samples from current production
- Identification of retraining focus items

Retraining benefits from production use of the model, in that the model continuously produces more ground truth – but only if a valid sample of output is reviewed for quality. Managing the quality of training data is crucial, so the steps taken to identify and validate ground truth when the model was initially trained should be not only maintained, but continuously improved.

Additionally, ground truth is not a static concept. Consider if you were designing models for reading and interpreting social media – the data, structure, and information provided in a MySpace® blog feed in 2000 is different from a Facebook® post in 2010 and again different from a tweet or snap in 2022. Likewise, in the mortgage industry, we have had multiple iterations over the years of GSE forms and required reporting data sets.



**Key Point:** Unlike traditional software, AI/ML is not a set-and-forget technology. A regular drift detection review of model feature performance and end-user feedback is important. Following up on model drift when identified may be critical to business objectives, so early detection of drift is very valuable. Successful organizations will establish a rigorous retraining routine, including a well-curated ground truth repository and clear assignment of responsibility for model performance.

See Figure 4 (below) for a simplified view of the ground truth and retraining cycle for document classification.

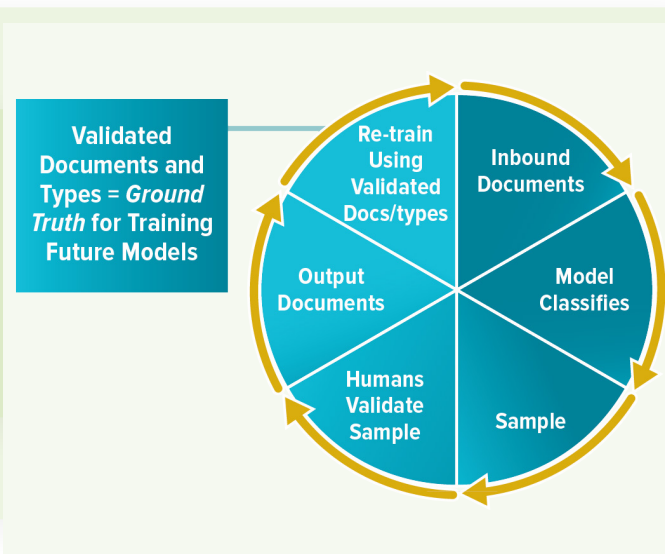


Figure 4 – Model retraining cycle, document classification

When the focus for the retraining has been identified, experimentation begins. The core work is much like original training, except that instead of a theoretical target for the model’s performance, the data science team needs to improve on the now-drifted model’s performance. The team goal is likely to be to achieve results somewhat better than the original model’s best performance.

Once a model achieves the desired level of performance, the ML Ops process deploys the new model.

Any AI/ML solution must include responsibility for ground truth management, including a data management plan and robust validation that may include humans-in-the-loop or human audit/validation processes.

## Architecture, Data Governance and Information Security

Since data is the currency of AI/ML, data governance and processes to gather ground truth and effective storage strategies are crucial. Large amounts of data about all the objects used to create models can result in complex storage requirements. In most cases, original artifacts, such as images for document processing or voice and video files, are needed to solve problems like speech recognition or to identify features in “real world” recordings. Since data scientists are never sure at the outset what data will be needed, when asked what data they need, any good data scientist will reply “all of it”.

Data management methods and costs are significant considerations for any AI/ML initiative. For data science professionals, clean, comprehensive and high-quality data sources are critical. Stakeholders in information security, legal, and compliance demand governance and controls engineered to protect, anonymize and appropriately use client and consumer data. Data architecture must be engineered for performance and be optimized to be financially sustainable. Regulators and end-users demand explainability, which will require storage and reporting of data that provides evidence of how the model arrived at each result. Failing any one of these needs, especially if the failure results in excessive cost or risk, could overwhelm the AI/ML value proposition.

This kind of bad outcome can be avoided with solid planning, careful financial forecasting, and good decisions about architecture, tools, data governance and information security. Each of these disciplines plays a role in successful use of AI/ML technology.

**Key Point:** AI/ML is emerging technology, and it does not stand alone. The core infrastructure and human capital that makes your traditional software work effectively is also needed for AI/ML. The secret is to gain an understanding of how the traditional and the disruptive differ, and how they work together.

## Stakeholders

Managers should also invest time providing stakeholders across the organization with educational materials and regular briefings. Using AI/ML will impact executive decision-makers, compliance, legal, human resources (HR) and audit teams and a successful program will provide them with knowledge of AI/ML-specific risks and issues.

- **Executives, decision-makers** – Executive management teams and boards of directors will likely be part of the decision-making process for any application of AI/ML in lending, especially given the concern expressed by regulators recently. Whether building your own or purchasing a vendor's AI/ML solution, the investment in and risk acceptance for this new technology is likely to require senior management or even board of directors approval. Providing regular briefings to decision-makers will improve their understanding of the risks, benefits, and policy/process changes needed to support AI/ML. Better understanding will lead to better outcomes when approvals are needed or if challenges arise.
- **Compliance and legal** – These teams are your allies in understanding the regulatory, contractual, design, and counterparty risks you need to manage in your AI/ML program. The sooner they know the details of AI/ML plans, the better able they will be to give meaningful input and adjust governance programs to emerging needs. Clear contractual language is needed to make sure data scientists have the rights to the data they need.
- **Privacy** – With existing and emerging data privacy regulations, an experienced privacy partner is essential to make sure information is accessed, stored and used in a compliant manner.
- **Human Resources** – From recruiting to career development and performance management, your HR team will need to understand entirely new job categories. Job descriptions for data science, ML Ops and AI/ML engineering will be needed if in-house AI/ML is planned. Even if third-party solutions include AI/ML, managers and staff will need new skills and knowledge. HR can help managers determine if new job descriptions are needed, or if current employees have critical skills or if they can meet needs with targeted training. HR engagement is usually important in determining if consultants or contractors are the best choice to fill a given need. Sharing knowledge early with HR partners will accelerate delivery when resource demand is realized.
- **Audit** – Internal controls and oversight programs will need to be developed and implemented. Partnering early with internal and external auditors will avoid bad audit outcomes due to misunderstandings, missed control requirements and late identification of risks. Auditors should be aware of new AI/ML programs and have a solid knowledge base far in advance of any audit.

***Key Point:** Managers should develop a communication and education plan for stakeholders to avoid obstacles and to facilitate informed decision-making for AI/ML success.*

## AI/ML and Risk Management

Because many characteristics of AI/ML software differ significantly from traditional software, managers need new tools to evaluate and manage the risk of their application in business. There are several risk categories that are especially important to consider.

- 1. Financial Risk** – The cost of developing, deploying and maintaining traditional software is well understood. The variability of the time invested in design, development and deployment work is managed through the software development life cycle and project management methods. In AI/ML the situation is more complex. All the traditional cost elements still apply, plus the added costs for ground truth gathering, model experimentation, training and retraining. Those costs are often difficult to understand in advance and can have a large impact if not properly managed.
- 2. Operational Risk** – AI/ML systems can be computationally intensive and require large amounts of storage. Model training is particularly challenging to manage – training runs can take extended periods of time and consume computing resources quickly. Models also require integration with other systems and may include ancillary capabilities to perform sample testing and gathering of ground truth. Change control risks are magnified by the need for regular model drift-related changes due to model decay and the limited ability of traditional quality assurance testing to detect problems in the AI/ML world. Finally, sample testing and ground truth gathering are usually human endeavors, adding a set of operational and information security risks.
- 3. Business Expectations/Reputational Risk** – A lack of understanding as to the complexity and capability of current-state AI/ML, combined with popular culture depictions of the technology and vendor sales hype, all point to a magic-like perception of the technology. As a result, expectations are very high and climbing. For vendors, unmet client expectations can result in dissatisfied customers. For lenders, the value delivered by a specific AI/ML solution may not meet the assumptions that were used to justify acquisition and maintenance costs. Beyond an additional layer of financial risk, there is also significant personal capital risk to the solution's advocates.
- 4. Information Security Risk** – Data scientists need access to the data used to evaluate and train models. There are cases – such as in document classification and extraction – where the information in the documents includes non-public personal information (NPI). Data science teams must work closely with their information security partners to be sure all data is protected, and a rigorous ongoing information security risk management stance is strongly recommended.





**Key Point:** *In the short run, demand for explainability in the financial services industry is likely to stem from regulatory concerns and potentially soon-to-be accepted standards.*

- Regulatory Risk** – AI/ML solutions must be documented and sufficiently explainable to support audits and client model validation and must be designed to allow clients to control regulatory bias risk.

Lenders have their own risk management programs that must adapt to the opportunities offered by AI/ML. Black Knight’s AI/ML strategy accounts for large-lender model validation requirements related to the Basel Accords and US Regulators’ safety and soundness standards.

## Explainability, Explained

AI/ML systems are technologically complex, and are often driven by advanced math and sophisticated algorithms. They are generally considered to be black boxes whose inner workings can be difficult to grasp without study by experts. Regulators, auditors and business managers alike are all trying to understand if and how AI/ML solutions can be relied upon to operate properly. This is challenging, to say the least, when they cannot see how the thing inside the box really works. The idealized goal of explainability is to therefore use technology and documentation standards to turn AI/ML into a glass box.

While the glass box is an alluring idea, understanding of a given model is audience specific. The kind of explanation that is meaningful to a regulator will likely differ from the explanation that is sensible to an operations manager, or a data scientist, or an end user. This means that managers developing an explainability framework for a given model need to carefully consider the needs of the explanation’s audiences.

In attempting to meet this demand, managers should work on explainability concurrently with initial development and deployment of a model. Without accounting for business explainability, it may be difficult to determine if the AI/ML solution delivers intended value. If carefully considered, design artifacts and business requirements support creating an explainable model before the first experiment is undertaken.

Current approaches to explainability center around the concept of “causality”, where one tries to measure the relative influence of specific features in arriving at a model output. Causality is a challenging concept; just because something appears to have an effect does not mean it actually does. Spurious correlations are chief offenders in inferring causality in wildly complex, real-world scenarios.



A model for explainable AI called XAI has been discussed by the US Government’s Defense Advanced Research Project Agency (DARPA) and others. From the end user’s perspective, explainability means that each result of a model makes sense. The user understands why a model made a given prediction. They have a sense of the limits of the model so that they understand when to trust its results, and when to discount its output. See Figure 5 (below).

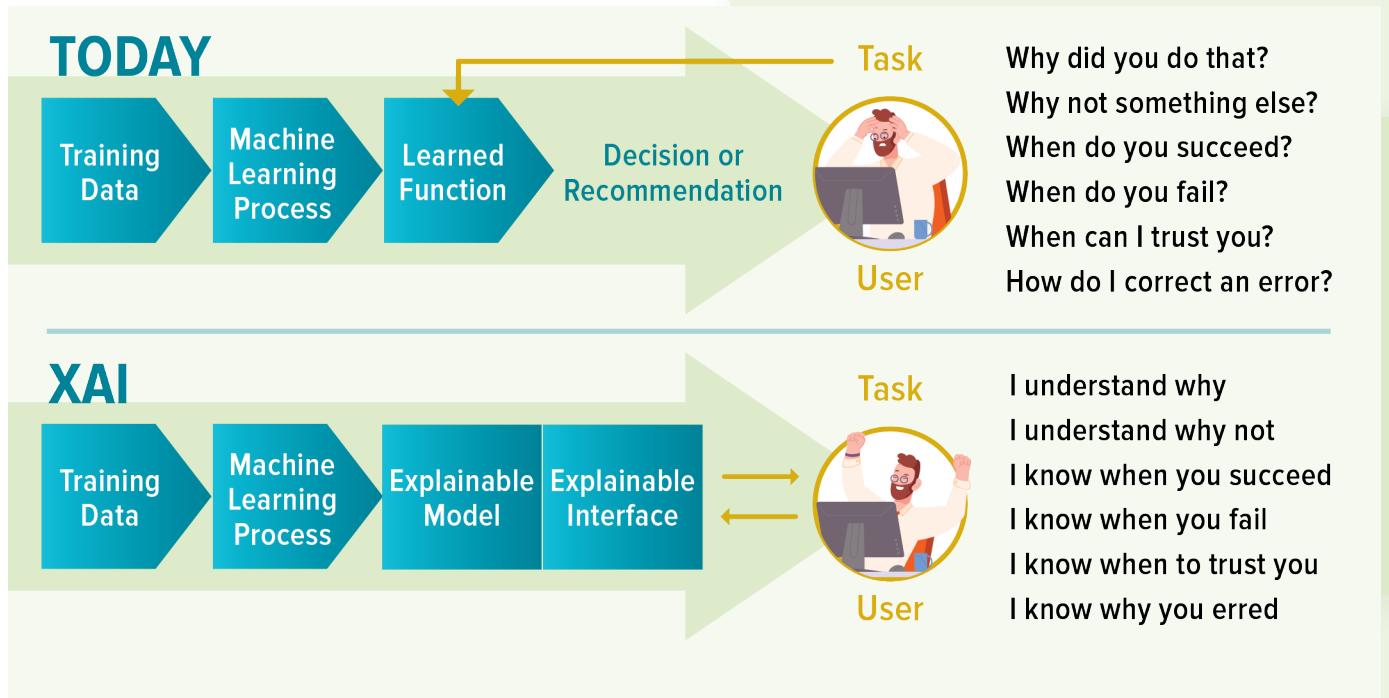


Figure 5 – XAI Concept.<sup>1</sup>

Among other things, the previously mentioned DARPA XAI model includes analytics, Q&A and model interaction elements for users to obtain the transparency desired. This model works well for interactive applications and expresses many foundational elements for explainability ideals quite well. It relies on feature attribution as a core element of explainability and sets high expectations for practitioners to democratize the understanding of the data that drives the model through visualization and model interaction. See Figure 6 (next page).

This XAI model may not be satisfying for all audiences, especially regulators. From a regulatory standpoint, explainability and transparency have a goal – detecting and eliminating bias that could harm consumers. The level of interactivity implied by this type of XAI is also likely to be uncomfortable for lenders.

To apply XAI to a given AI/ML solution, managers need to prioritize the search for both technical and regulatory bias, along with the search for a solution to a given problem. Careful design is important in managing training data, as well as in the experiment/build process with the goal being an explainable answer to the question the model is intended to answer. Some explanations are straightforward – in the case of document processing, the user can look at a document to determine if the model classified it properly. They can also see why a blurry document was not properly classified. A document classifier can rely on OCR text- or image-based information to identify a document, so explaining that it looks for specific text or pixel patterns to classify a document type is generally sufficient.

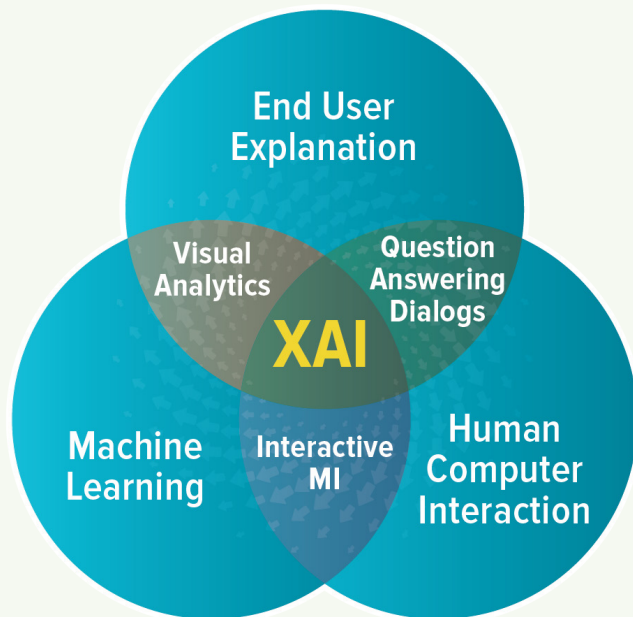


Figure 6 – candidate XAI Model – XAI Emphasis (from DARPA DARPA-BAA-16-53)<sup>2</sup>

Most AI/ML models require deeper explanation. Consider facial recognition. Assume an AI/ML model is tasked with determining if a person in a picture is smiling or not. Such a model could include a feature to examine eyes. An algorithm first identifies eyes, classifies them as left and right, then looks for squinting or lines in the outside corners of each. A second feature identifies the person's mouth and assesses the geometry of the lips to determine if the corners are uplifted. If the person is squinting, evidenced by lines in the corners of the eyes AND the lip corners are uplifted, then the person is predicted to be smiling. The model is explained by two features – one for the eyes and one for the mouth.

Now consider the weaknesses of the model for people with beards or eyeglasses. Data scientists would need to engineer features that determine the presence of facial hair and find alternatives if eyes are partially or fully covered. What else about the training data could also cause bias? Were a diverse set of ethnicities included in the training so that eye shape and facial hair impacts are correctly handled? More model features are needed, and careful selection of training data is a must.

To meet the explainability challenge, data scientists need a strong partnership with business management and legal/compliance teams. These stakeholders should be enlisted to question the output of each model, asking the data science team to answer their questions about its transparency and explainability. Those questions will often lead to software requirements, reporting on outcomes, visualizations, technical diagrams and narrative explanations.

One additional concern of note, especially to technology providers: any reader of this document has learned one thing – developing quality AI/ML is not cheap. Investment in the full set of capabilities needed to manage AI/ML solutions is considerable. The pursuit of transparency cannot overwhelm the protection of intellectual property. Explainability demand cannot overcome a software owner's right to protect their investment by requiring disclosure of trade secrets.

Lenders, technology providers and their regulators are embarking on a journey in pursuit of a solid explainability paradigm. The additional research, documentation and software engineering needed will be a non-trivial cost of delivering AI/ML solutions.

**Key Point:** *Explainability requires more than a description of a model's features. Among other things, data scientists need to carefully curate and describe the source and composition of training data, provide transparency about how models operate, and be able to document and report on measures taken to detect and correct for model bias.*

## The Regulatory Landscape

Financial services industry regulators and other regulatory bodies have been increasingly vocal about their concerns regarding the application of AI/ML to consumer financial transactions. Among them:

- CFPB statements reinforce the need for deliberation about the application of AI/ML because such systems could introduce bias that can “result in credit determinations that are unfair to consumers.”<sup>3</sup>
- AI guidance issued by the Federal Trade Commission emphasizes that AI algorithms should be transparent, explainable, fair, empirically sound and foster accountability.<sup>4</sup>
- In February 2022, a fair housing advocacy group proposed a framework for auditing AI/ML solutions that is focused on detecting bias that can adversely impact consumers.<sup>5</sup>
- Other organizations, including the OCC and National Institute of Standards and Technology, are also working on AI/ML standards aimed at developing and using AI/ML in a trustworthy and responsible manner.<sup>6</sup>
- In July 2022, National Mortgage News reported that the FHFA was starting its own Office of Financial Technology, which includes a reference to compliance activities and assuring use of technology in “a responsible and equitable manner.”<sup>7</sup>
- On May 26, 2022, CFPB issued Circular 2022-03 . This circular warns consumers and lenders against “...credit decisions based on certain complex algorithms, sometimes referred to as uninterpretable or “black-box” models, that make it difficult – if not impossible – to accurately identify the specific reasons for denying credit or taking other adverse actions.” In a footnote following this statement, CFPB further states: “While some creditors may rely upon various post-hoc explanation methods, such explanations approximate models and creditors must still be able to validate the accuracy of those approximations, which may not be possible with less interpretable models.”<sup>8</sup>

International trends are similar, if not more aggressive. Two examples:

- In April 2021, The European Parliament began consideration of The Artificial Intelligence Act.<sup>9</sup>
- In 2022, the Canadian Parliament considered sweeping regulation of AI and Machine Learning under the proposed Artificial Intelligence and Data Act.<sup>10</sup>

Despite this considerable activity, as of this writing no regulators have proposed rules or provided detailed guidance to the industry about regulatory risk management of AI/ML models.

When it comes to understanding regulatory bias risk, lenders should consider adopting their own standards while the regulators deliberate. **Dark Matter** has developed an interim model for regulatory bias risk that classifies an

AI/ML solution based on increasing risk tiers:

- **Tier 1** – Identifying/Learning: an AI/ML tool that simulates manual work and data validation
  - a. Expected outcomes are testable by direct observation
  - b. Auditable and verifiable through sampling
  - c. Minor to no risk
- **Tier 2** – Presenting/Predicting: an AI/ML tool that produces analytics and interpretation of large data sets resulting in predictions or insights
  - a. Such use exceeds the combination of a Tier 1 AI/ML solution deterministic rules engine
  - b. Unknown outcomes are supported by underlying data, therefore explainable through analysis of the data
  - c. Data insights require human review and interpretation to be operationalized outside of the AI/ML product
  - d. Minor to no risk for some applications, and moderate risk due to reliance on predictive results for other applications
- **Tier 3** – Decisioning/Suggesting: an AI/ML tool that creates operational decisions or limits the human operator’s decisions based upon analytics or interpretations of data
  - a. Unknown outcomes are supported by underlying data, therefore explainable through analysis of the data but individual predictions are applied immediately
  - b. Operationalized models that can directly modify process outcomes. Example: risk models that are applied to underwriting decisions.
  - c. Higher risk due to immediate application of predictions to operational processes that impact business and consumer outcomes



**Key Point:** Financial services regulators have significant concern about the risk of bias in models, as well as concern that models could result in consumer harm. Pending definitive guidance, managers should adapt their risk management programs to incorporate these perceived regulatory risks as factors in assessing investment in AI/ML solutions. As first steps, managers should:

- Work with their internal risk management partners to assess XAI and similar explainability approaches
- Actively discuss explainability with internal organizations and vendors who provide AI/ML services.

## Summary

Technology and business managers alike should be familiarizing themselves with artificial intelligence and machine learning. Whether procured as part of a vendor solution or developed internally, AI/ML is different, and managers should be aware that those differences may place new demands on them, while also offering great benefits.

- Traditional software technologies, including business rule management (BRM) capabilities such as Decision Select, Black Knight's rules engine-based solution, are the appropriate technologies for meeting most mortgage origination business requirements since deterministic solutions are needed, i.e., inputs must be directly traceable to precise expected outputs.
- AI/ML technology, including the originations-based AI platform AIVA, is appropriate for addressing business problems that do not yield to deterministic approaches, i.e., where inputs are chaotic, where a range of possible outcomes exists, and therefore a probabilistic approach is required.
- All AI/ML solutions need new approaches to software development and deployment. They must be documented differently and be sufficiently explainable to support multiple audiences, including auditors, regulators and managers seeking to control regulatory bias risk.

Competitive lenders need to adopt this new technology, and given the risks involved, management has a responsibility to gain an understanding of the risks and risk management tools that are needed for success. This is true in the adoption of any new capability. While AI/ML is unique, with research, sound planning and the right investment, the journey can be rewarding.



## Addendum

### Discussion of $F_1$

To better understand measuring using  $F_1$ , feel free to consult your favorite statistics textbook. This section provides a business-friendly description.

For prediction scenarios like the examples in this document, there are four ways to evaluate the quality of a result. The model can provide a:

- **True Positive** – a prediction that is correct
- **True Negative** – a model correctly made no prediction
- **False Positive** – a prediction that is incorrect
- **False Negative** – no prediction when one should be made

The count of results in each of these categories is used to calculate two statistics:

- **Recall** – how often are relevant predictions made
  - Calculation: True Positives / (True Positives + False Negatives)
- **Precision** – how often are predictions relevant
  - Calculation: True Positives / (True Positives + False Positives)

Recall and precision each have a place in understanding model performance. Recall gives insight about model consistency; precision reflects how well it gets close to the right answer. The  $F_1$  statistic combines the two measures to provide a more balanced understanding. Mathematically,  $F_1$  is defined as the harmonic mean of precision and recall. Here is the formula:

$$F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

By the nature of the calculations, precision, recall and  $F_1$  all have a place in analyzing the viability of a given process in producing the right result.

## Suggested Additional Reading

Those seeking a more detailed understanding of AI/ML should consider additional study. Consider the following information sources:

- Google AI/ML course: <https://developers.google.com/machine-learning>
- AI/ML glossary from Science Magazine: <https://www.science.org/doi/full/10.1126/science.357.6346.19>
- The Brookings Institution: Glossary of AI and emerging technologies: <https://www.brookings.edu/blog/techtank/2020/07/13/the-brookings-glossary-of-ai-and-emerging-technologies/>
- DeepAI.org – AI/ML definitions: <https://deepai.org/definitions>
- International Journal of Artificial Intelligence & Applications: <https://airccse.org/journal/ijaia/ijaia.html>

## References

- 1 DARPA, “Broad Agency Announcement Explainable Artificial Intelligence (XAI)”, DARPA-BAA-16-53, <https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>, viewed March 28, 2022
- 2 Dr. Matt Turek, “Explainable Artificial Intelligence (XAI)” (Defense Advanced Research Projects Agency, <https://www.darpa.mil/program/explainable-artificial-intelligence><https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>), viewed May 10, 2022
- 3 Kate Berry, “CFPB warnings of bias in AI could spook lenders”, American Banker, January 31, 2022, <https://www.americanbanker.com/creditunions/news/cfpb-warnings-of-bias-in-ai-could-spook-lenders> (accessed February 11, 2022)
- 4 Andrew Smith, “Using Artificial Intelligence and Algorithms”, April 8, 2020, <https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-algorithms>
- 5 National Fair Housing Alliance, “The National Fair Housing Alliance Releases A New Framework for Auditing Algorithmic Systems: Purpose, Process, and Monitoring (PPM)”, February 17, 2022, <https://nationalfairhousing.org/the-national-fair-housing-alliance-releases-a-new-framework-for-auditing-algorithmic-systems-purpose-process-and-monitoring-ppm/>, (accessed February 22, 2022)
- 6 Office of the Comptroller of the Currency, Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation Bureau of Consumer Financial Protection, National Credit Union Administration, “Request for Information and Comment on Financial Institutions’ Use of Artificial Intelligence, including Machine Learning”, 86 FR 16837, March 31, 2021, <https://www.federalregister.gov/documents/2021/03/31/2021-06607/request-for-information-and-comment-on-financial-institutions-use-of-artificial-intelligence> (accessed February 21, 2022)
- 7 Brad Finlestein, “FHFA starts its own Office of Financial Technology”, National Mortgage News, July 18, 2022, <https://www.nationalmortgagenews.com/news/fhfa-starts-its-own-office-of-financial-technology> (accessed July 19, 2022)
- 8 CFPB, “Consumer Financial Protection Circular 2022-03”, May 26, 2022, <https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/> (accessed September 9, 2022)
- 9 European Commission, “Proposal For A Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts”, [https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC\\_1&format=PDF](https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF), (accessed July 18, 2022)
- 10 “Digital Charter Implementation Act, 2022”, <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading>, (viewed July 19, 2022)



## Statement of Confidentiality and Disclaimer

The information contained in or supplied with this document (the “Information”) is the property of Dark Matter Technologies LLC, or one of its affiliates or subsidiaries (collectively, “Dark Matter”) and contains the confidential, proprietary, and/or trade secret information of Dark Matter. The Information cannot be used, modified, extracted, copied, or distributed without the prior written consent of Dark Matter. With or without Dark Matter’s prior written consent, Dark Matter will not be liable for the disclosure or use of the Information. The intended recipient shall not use any part of the Information contained in, or supplied with this document, in any way to the competitive disadvantage of Dark Matter, and will take all steps designed to assure its compliance with this provision. Nothing herein shall be construed, nor is Dark Matter providing, any legal or compliance advice.

If this document contains proposal language, this proposal is neither an offer nor intended by Dark Matter, upon acceptance by the intended recipient, or otherwise, to create a binding agreement with Dark Matter. Such an agreement shall be reflected only by a definitive contract executed by both parties.

If this document contains feature or functionality specifications (the “Specifications”), Dark Matter does not guarantee the accuracy of the Specifications for any product, service, or functionality identified. The Specifications contained in this document may not be complete or up to date, and Dark Matter is not responsible for notifying the recipient of changes in the Specifications.

TM SM ® Trademark(s) of Dark Matter Technologies LLC, or an affiliate.  
© 2024 Dark Matter Technologies LLC. All Rights Reserved.