

Martinello, Alessandro; Jensen, Thais Lærkholm; Grenestam, Erik; Meyer, Bjørn Bjørnsson

Research Report

AI and machine learning in the financial sector: Five focus points

Economic Memo, No. 3

Provided in Cooperation with:

Danmarks Nationalbank, Copenhagen

Suggested Citation: Martinello, Alessandro; Jensen, Thais Lærkholm; Grenestam, Erik; Meyer, Bjørn Bjørnsson (2022) : AI and machine learning in the financial sector: Five focus points, Economic Memo, No. 3, Danmarks Nationalbank, Copenhagen

This Version is available at:

<https://hdl.handle.net/10419/261829>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

DANMARKS NATIONALBANK

1 APRIL 2022 — NO. 3

AI and machine learning in the financial sector: Five focus points

Alessandro T. Martinello
Principal Data Scientist
FINANCIAL STABILITY
alem@nationalbanken.dk

Bjørn B. Meyer
Data Scientist
FINANCIAL STABILITY
bbm@nationalbanken.dk

Thais L. Jensen
Head of Analyses and Models
FINANCIAL STABILITY
tij@nationalbanken.dk

Erik A. Grenestam
Senior Data Scientist
FINANCIAL STABILITY
egr@nationalbanken.dk

The viewpoints and conclusions stated are the responsibility of the individual contributors, and do not necessarily reflect the views of Danmarks Nationalbank. The authors thank Per Andersen, Rastin Matin, Rikke R. Nissen, Thomas Krause for valuable sparring, input, and discussions.

AI and machine learning in the financial sector: Five focus points

Abstract

The financial sector and its regulators have an obligation to explore the use of Artificial Intelligence (AI) and machine learning. These technologies hold the potential of sharpening business processes and improving the resiliency of both individual financial institutions and the financial sector as a whole. Yet, users of these tools carry the responsibility to continuously balance their potential benefits against the risks which these technologies can amplify. This paper highlights the need of a model governance structure, and suggests five focus points which financial institutions should consider when moving from simple, static statistical models to complex, self-learning AI systems and machine learning algorithms.

The financial sector's interest in artificial intelligence (AI) and machine learning is growing at an accelerating pace (Refinitiv, 2020; Babina, Fedyk, He, & Hodson, 2021).¹ This growth is fuelled not only by the increased availability of large datasets and the popularisation of machine learning algorithms through open-source technology, but also by the competitive pressure that fintech start-ups are placing on established financial institutions.

AI systems have a wide range of applications. They can improve the quantity and quality of services offered by the financial sector through e.g. data-driven product customisation, innovative financial solutions and trading algorithms. They can also sharpen or aid decisions normally taken by humans and automate business processes (e.g. customer and email routing or financial advice through chatbots).

AI also holds the potential of benefitting society as a whole. By allowing financial institutions to obtain more accurate estimates of their financial exposure, the likelihood of loan defaults and the riskiness of their portfolio, they can increase an institution's resiliency and thereby the structural robustness of the financial sector. In turn, more robust credit portfolios allow banks to increase their credit supply and provide liquidity to the most productive assets in an economy.

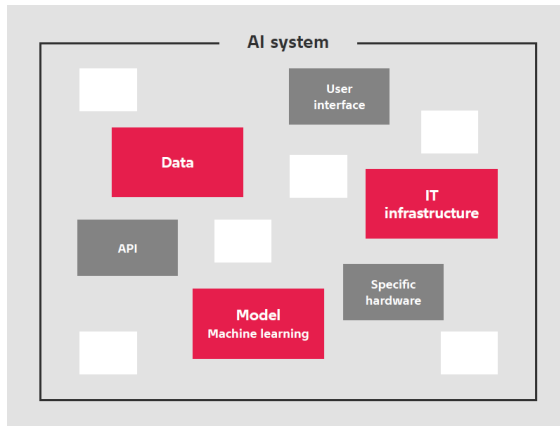
Yet, these technologies also raise ethical, operational, and regulatory dilemmas. Delegating decisions to a machine weakens the link between corporate actions and accountability. It can reinforce biases, introduce algorithmic discrimination, and reduce the transparency of decisions that affect customers.

¹ The exact boundaries between machine learning and simpler statistical models can be blurry. Even a relatively simple model such as a logistic regression can be viewed as a machine learning approach, especially if a system is in place to automatically re-train and calibrate model

parameters as new data becomes available. Box 1 provides a definition of these terms for the purposes of this paper, and provides examples of how AI and machine learning are used in the financial sector.

Artificial intelligence and machine learning

Box 1



Artificial intelligence and machine learning are often used interchangeably – sometimes as buzzwords – with unclear established definitions. In this paper, we follow the definitions used in the European Commission proposal for a regulation laying down harmonised rules on artificial intelligence ("Artificial Intelligence Act") (COM/2021/206 final) and refer to AI as any form of programmatic system able to produce content, recommendations, decisions, allocations, or any other form of output that can influence a human user or the surrounding environment. Artificial intelligence systems therefore include chatbots, recommenders, and constrained optimisers from Operation Research (e.g. in portfolio selection or financial planning systems).

Such a system consists of several components. These typically include at least data, an IT architecture, and a model – often a machine learning model. In this paper, we therefore refer to an AI system as the self-contained combination of components that together generate inputs and recommendations to a user, and to machine learning as specific types of models that can be a component of the AI system.

Machine learning consists of statistical methods characterised by their flexibility in adapting to the data at hand (OECD, 2021). Particularly relevant for the financial sector are supervised machine learning approaches, i.e. methods able to recognise patterns across a set of explanatory variables (features) that are associated with a specific outcome and can therefore then be used for prediction purposes. Supervised machine learning is used in many AI applications, from natural language processing (chatbots, translation engines, named entity recognition) to prediction purposes (churn analysis, time-series forecasting, distress predictions). Examples of supervised machine learning include random forest, gradient boosted trees, and deep learning (Ozbayoglu, Gudelek, & Sezer, 2020).

This paper presents an overview of the challenges that a financial institution may face when deploying machine learning technologies, from both a regulatory and an operational perspective. A solid model governance structure is necessary to mitigate risks associated with the deployment of machine learning models. This paper suggests five focus points to consider when building such a governance structure.

While not a new paradigm, AI and machine learning amplify existing risks

Box 1 defines an AI system as any form of programmatic system producing content able to affect its surrounding environment – a system that often incorporates machine learning models as one of its primary components. While AI systems and machine learning models are gaining momentum driven by the surge in the amount of available data, financial institutions have relied for decades on micro-level datasets and statistical models for informing their decision-making processes. Machine learning as a method in itself does not constitute a new paradigm for the financial sector: The degree of complexity of a model does not fundamentally change the ethical responsibilities of a financial institution or the types of risks to which it is exposed.

Even a simple linear regression can be an obscure black box for those without statistical training. Even a simple logistic regression has the potential to unfairly discriminate across race or gender without proper supervision. Likewise, the decision-making processes of a human decision maker can be all but transparent, with existing implicit biases and prejudices being hard to detect.

While machine learning carries the promise of making decision-making processes more precise, impartial and objective, blindly trusting these models with minimal or no supervision and oversight risks achieving the exact opposite. By increasing model complexity, by automating the update of model

parameters, and by potentially removing human layers between data, model, and output, machine learning models and AI systems can substantially amplify the inherent risks of data-driven decision-making.

Machine learning and AI amplify risks of algorithmic discrimination

Algorithmic discrimination has drawn substantial attention in recent years, as cases of unfair treatment across race and gender by AI systems using machine learning models emerged.² Machine learning models are designed to be able to achieve better performance than simpler, static statistical approaches.

From a business perspective, higher performance is desirable, as it allows for increased efficiency in e.g. pricing, customisation, and recommendations. Yet by attempting to extract as much information as possible from available data to maximise predictive performance, machine learning models are more prone than simpler models to unintentionally discriminate (Kleinberg J. , Ludwig, Mullainathan, & Rambachan, 2018).

Using available information to customise e.g. access to credit to different individuals is part of the core business of a financial institution and does not per se constitute discrimination. For example, we expect a credit institution to price loans differently according to the probability of default of the customer. Discrimination occurs when the estimated default probability of equally risky borrowers differs nonetheless systematically across race, gender, or any other protected attribute.

As the potential of discrimination dramatically increase reputational and legal risks, we argue that financial institutions ought to adopt a strategy of awareness by monitoring algorithmic bias across protected groups (see box 2). In fact, Article 10, paragraph 5 of the European Commission proposal

for an Artificial Intelligence Act explicitly allows—under appropriate security measures—the processing of available personal data and otherwise protected characteristics for the exclusive purpose of bias monitoring, detection, and correction (COM/2021/206 final, 2021).

The sources of algorithmic bias are multiple and can occur at any stage of model development, from data collection (e.g. an image dataset featuring primarily Caucasian people, or a loan application data where female applicants have been historically unfairly rejected) to model development and deployment. As a consequence, there is no single recipe or guidebook to detect and correct such biases. While the literature provides competing bias metrics, no unifying measure of algorithmic bias exists, as examples that might be unbiased according to one metric can be biased according to another.

The choice of bias metrics, relevant protected groups, and approaches to ensure the absence of algorithmic bias should therefore be grounded in context-specific best practices. Ultimately, the goal and responsibility of financial institutions is to hold their automated systems, including AI systems and machine learning models, accountable to the same ethical and legal standards as their human driven decisions. In fact, algorithms have the potential of being more transparent in their recommendations than human decision-making. With the right safeguards and appropriate governance, algorithms can be a powerful force for equity (Kleinberg J. , Ludwig, Mullainathan, & Sunstein, 2018; Kleinberg J. , Ludwig, Mullainathan, & Sunstein, 2020).

² A prominent example is the recent debate about criminal risk scores used in several U.S. states (Angwin, Larson, Mattu, & Kirchner, 2016).

Monitoring algorithmic bias in machine learning models

Box 2

In a legal framework, information about customers is typically divided in protected attributes, e.g. gender and race, and permissible attributes that are allowed as input to business decisions, e.g. income and assets. The concern is that an algorithm is unfair/biased in the allocation of resources, like access to credit, in a way that disproportionately disadvantage a protected group of individuals.

To eliminate algorithmic bias, simply excluding protected attributes from the data (*fairness through unawareness*) is not only insufficient, but also counterproductive. The risk of algorithmic discrimination through triangulation, yielding discriminatory outcomes in e.g. mortgage allocation, remains even when excluding protected indicators (Fuster, Goldsmith-Pinkham, & Ramadorai, 2022). Statistical models extract as much information as possible from the data that can help them optimise their performance, and machine learning models excel at this task: Deep learning can predict race even from unmarked medical imaging (Banerjee, et al., 2021). Instead, protected attributes can be used to actively monitor, identify and correct algorithmic bias. This strategy (*fairness through awareness*) also allows the use of machine learning models to actively detect bias in existing decision processes, automated or not (Martinello, Mønsted, Matin, Steffensen, & Laursen, 2021).

We introduce three intuitive appealing fairness definitions that can be tested on a statistical model for key protected attributes. While the metrics generalise to most cases where individuals are categorised, we focus on the example of credit underwriting. Machine learning and various permissible variables are used to assess the risk of default. Customers assessed with a probability of default lower than a given threshold are considered unlikely to default and granted a loan are labelled $L=1$. Conversely, customers with high risk assessments and no loan offers are labelled $L=0$. Whether a customer in the data actually repays (defaults) is labelled $R=1$ ($R=0$). $C=1$ denotes membership of the group of customers sharing the protected attribute, e.g. women or an ethnic minority.

The simplest metric to monitor is the raw difference in granted loans across groups. *Demographic Parity* refers to both groups having the same probability of getting a loan.

Demographic Parity definition: $P(L = 1|C = 1) = P(L = 1|C = 0)$

Because differential treatment in many cases can be justified by permissible attributes, this measure can only serve as an initial benchmark. More complex metrics are necessary to inform on algorithmic discrimination. Unfortunately, not only are there many competing metrics in relation to classification errors across groups building on different intuitions of fairness, they are also often impossible to satisfy simultaneously (Kleinberg, Mullainathan, & Raghavan, 2018). Hence, the choice of the appropriate metric depends on how the model will be used, established practices in specific business areas and the model risks that are most important to minimise.

Nonetheless, two widely used metrics are particularly useful – not only because they are simple and intuitive, but also because they describe how classification errors can work differently for different groups. Both measures build on the idea that the accuracy of the model should work in the same way for all groups. While final metric choices should depend on specific models and business areas, choosing to ignore substantial unbalances in either of these metrics should be justified.

Equal Opportunity checks whether false negative error rates are balanced across groups. This definition implies that groups should not differ in their risk of being denied a loan when it would actually be repaid.

Equal Opportunity definition: $P(L = 0|R = 1, C = 1) = P(L = 0|R = 1, C = 0)$

Equalised Odds is more restrictive and requires both equal true positive rates and equal false positive rates. This definition implies that qualified customers are equally likely to get a loan, and if they are not qualified, they are equally likely to get rejected.

Equalised Odds definition: $P(L = 1|R = 1, C = 1) = P(L = 1|R = 1, C = 0)$ and $P(L = 0|R = 1, C = 1) = P(L = 0|R = 1, C = 0)$

Machine learning models can be less explainable than simpler statistical approaches

The more interpretable and explainable a model is, the easier it is to document it and to detect unintended model behaviour, e.g. leading to algorithmic discrimination. Explainability of a model is therefore a value in itself, complementary to performance.

While a more complex model can sometimes boost performance, complexity does not have to result in a “black-box” model where the inner workings are incomprehensible to a human. Steps can and should be taken both before and after model development to improve its interpretability (see box 3). Examples include the computation of feature influences, and imposing constraints during model training to incorporate domain-specific knowledge.

Explainable AI

Box 3

In the machine learning literature, the terms explainability and interpretability are often used interchangeably. They represent the details and reasons a model provides that make its functioning and predictions easy to understand.

In many fields, including financial services, explainability is often viewed as equalling performance when ranking desirable model characteristics. There are three main reasons for the importance of model explainability (Adadi & Berrada, 2018):

The need to justify: In AI/ML applications, developers and users need to know why a model reached a particular conclusion given a set of inputs. For example, a bank who are looking to implement a new internal ratings-based model must be able to explain to a regulator why the model has assigned a particular default probability to an asset.

The need to control and improve: For model developers and maintainers, explainability is one of the most important tools to control model behaviour and improve performance. Explaining model decisions can uncover flaws and errors to be corrected. As an explainable model is more informative to the user, explainability can facilitate model adoption and oversight and mitigates the risk of unexpected model behaviour.

The need to discover: Explainability can uncover relations between model features to better understand underlying processes and answer questions about causality.

Techniques that facilitate model explainability can be divided into two categories, ex-ante and ex-post. As the name suggests, ex-ante explainability refers to choices by the model developer prior to actually implementing the model. This includes the simple approach of choosing a model that is inherently interpretable. An interpretable model is defined by several characteristics (Sudjianto & Zhang, 2021). Examples include additivity (can outcomes be represented as a weighted sum of inputs?), linearity (is the impact of a feature proportional to the value of the feature?) and visualisability (can the impact of model features easily be presented in a graph?).

Explicitly incorporating domain knowledge into a model (e.g. knowledge about a bank’s lending process is relevant domain knowledge when predicting default probability) is encouraged (Roscher, Bohn, Duarte, & Garcke, 2020). Domain knowledge can drive explainability by e.g. shaping the set of training data, selecting relevant model features and restricting features and algorithms to reflect real-world relationships between features (e.g. income should not have a negative impact on default probability). As ex-ante explainability approaches place restrictions in either the range of available models or in how freely the model can adapt to the data, these approaches typically result in performance losses for predictive tasks.

The other explainability paradigm, ex-post explainability, refers to techniques that can be used to understand model outcomes after implementation and training. This toolkit is often model agnostic and can be used across a range of models. The perhaps most common ex-post explainability tool consists in feature importance measures such as SHAP (Lundberg & Lee, 2017), which quantify the average impact that a particular feature has on the outcome.

The advantage of ex-post explainability is that it places minimal restriction on model choice and training. However, ex-post explainability tools entail a simplification of model behaviour that may not provide an accurate representation of model behaviour and therefore cause difficulty for conceptual soundness evaluation for model risk management (Molnar, et al., 2021).

As in the case of algorithmic bias, there is no definitive cookbook on how to interpret machine learning models, with strategies depending both on the type of model being developed and on the relevant stakeholders. Different stakeholders have different explainability requirements. A domain expert using the model needs insight into how specific features influence model predictions in order to trust the model. Developers and maintainers need transparency on model training and performance. Regulatory entities need transparency on the process of model development, and explanations on how the model performs across protected attributes to certify compliance with existing regulations (Arrieta, et al., 2020).

As model complexity increases, stakeholder-specific explainability requirements become harder to address through a single approach. While for simple statistical models the same documentation might satisfy the requirements of model maintainers and regulatory entities alike, complex machine learning models require specific efforts to ensure model explainability across competences and needs for oversight.

Regulation on AI is growing

Automated systems exploiting machine learning models have attracted the attention of both national and international regulators, see box 4. Specifically, the recent European Commission proposal for an Artificial Intelligence Act represents the most comprehensive, detailed and regulatory effort to date in a European context (COM/2021/206 final, 2021).

The proposal is subject to further amendments, and a deadline for implementation in national law has not been established yet. Nonetheless, in its current version administrative fines for non-compliance are set to up to 30 million euros, or, for a company, up to 6 per cent of the previous year's worldwide revenue, whichever the highest.

The goal of this regulation is not to supersede existing requirements, e.g. in terms of non-discrimination and fair treatment, to which any system or process must comply regardless of its degree of automation. Rather, it proposes an additional set of checks and balances specific to AI systems.

The proposal by the European Commission first defines high-risk AI systems. Notably for the financial sector, Annex III specifically refers to systems used to evaluate the creditworthiness of natural persons or establish their credit score as high-risk AI systems. Second, it proceeds in laying down a range of governance requirements to which high-risk systems must be subject. These requirements include the registration of high-risk AI systems in a centralised database, the establishment of a dedicated risk management system and data governance, adherence to principles of transparency and explainability, and the involvement of human oversight.

Due to the difficulty of establishing universal standards for concepts such as algorithmic bias and explainability, the European Commission proposal take a principle-based approach. A principle-based approach establishes general principles (such as explainability and non-discrimination) to be respected but avoids providing specific metrics on which to evaluate such principles. This approach has the crucial advantage of being flexible, easily adapting to the emergence of new technologies and shifting of collective understanding of principles.

However, such an approach puts responsibility on practitioners to develop and maintain best practices and governance structures for adhering to general principles. For example, the current proposal never addresses fairness directly and only specifies that any metric used to measure eventual discriminatory impacts are to be included in the technical documentation of the system required in Article 11 [Annex IV]. Yet, no specific metric is mentioned, with choices on appropriate metrics (see box 2 for a selected sample of popular fairness metrics) and

monitoring practices being left to the system developer.

Regulatory framework

Box 4

The attention of regulators to the use of machine learning and artificial intelligence is growing both nationally and internationally.

Nationally, the Danish FSA has published a guide on good practices when using supervised machine learning (Finanstilsynet, 2019). Other international regulators are also increasing their awareness and attention on how machine learning is used in the financial sector (BaFin, 2021). The Bank of England and the Financial Conduct Authority have launched in 2020 an AI Public-Private Forum, and recently released a report highlighting challenges, risks, and best practices in adopting AI in financial services, in line with the conclusions of this paper (AIPPF, 2022).

At the European level, the European Banking Authority has first published a report on big data and advanced analytics (EBA, 2020), followed by a discussion paper on machine learning for IRB models (EBA, 2021). The European Commission has presented a proposal for an Artificial Intelligence Act (COM/2021/206 final), proposing large fines of up to 30 million euros, or 6 per cent of a company's worldwide annual revenue in case of non-compliance.

In addition to these regulations and guidelines, any automated system employing machine learning is also subject to all other existing legal frameworks and specifically those ensuring equal treatment and preventing discrimination. At the European level, these frameworks are constituted by several directives, including those on racial, gender, and employment equality (2000/43/EC; 2000/78/EC; 2004/113/EC; 2006/54/EC).

Building a common language: From regulatory principles to rules and minimum requirements

To establish an effective governance structure, abstract regulatory principles must be translated into concrete rules and quantifiable minimum requirements. This translation requires a common language to be established across an organisation. Different stakeholders in a governance structure,

from legal experts to developers and data scientists, must be able to seamlessly refer to the same concepts.

Establishing a common language across different expertise areas can be a challenge even for the most mature organisations. Referring to common standards can therefore help the process of translating regulatory principles into concrete rules and requirements. Efforts for establishing standards on machine learning and AI governance are currently ongoing within the ISO/IEC JTC 1/SC 42 domain (ISO, 2022).³ These standards are likely to increasingly become a natural element of machine learning and AI model governance.

Credit rating and capital requirements

Capital requirements stand out as an application area in the financial sector as, despite the area's relevance for credit institutions, machine learning models have so far been very sparsely used in internal ratings-based (IRB) approaches to calculate regulatory capital (EBA, 2021). IRB models are subject to thorough regulatory supervision, and financial institutions have been cautious in adopting complex models for determining capital requirements. Moreover, Basel III capital floors limit potential gains for banks in this area.

Regulations on capital requirements (575/2013; 2019/876) nonetheless include requirements and suggestions for the use of machine learning.

Challenges in the adoption of machine learning models for capital requirement purposes are due not only to an increased complexity and opacity of machine learning models, but also to the frequent or automated recalibration of model parameters, requiring stringent governance requirements for documenting and auditing model updates. These governance requirements make investments in machine learning technology for IRB modelling costlier than in other areas.

³ Particularly relevant for the purposes of this paper are the standards on Transparency taxonomy of AI systems (ISO/IEC AWI 12792), Artificial intelligence – risk management (ISO/IEC DIS 23894), Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality

model for AI systems (ISO/IEC CD 25059), and Governance implications of the use of artificial intelligence by organizations (ISO/IEC FDIS 38507).

Five focus points for model governance

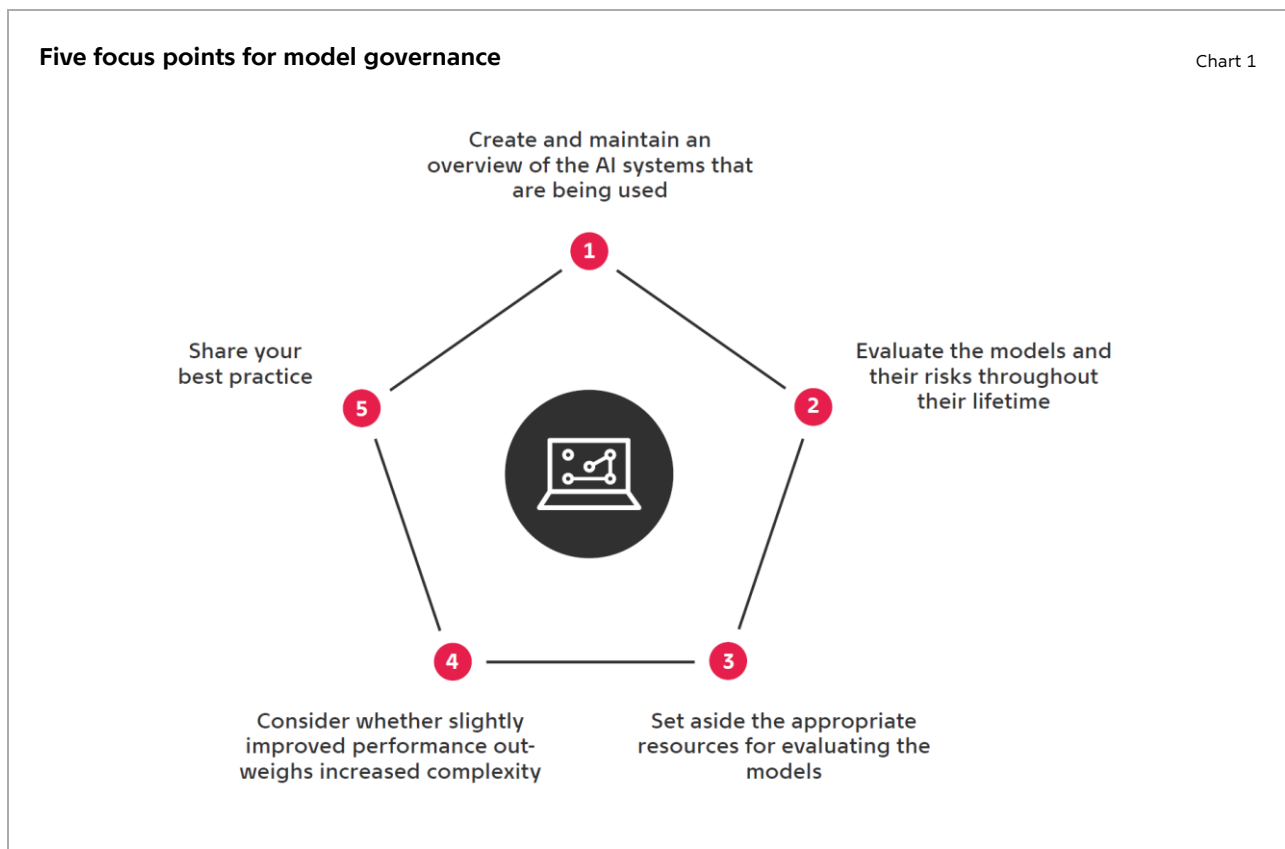
Any deployed automated system or statistical model requires some form of governance for regulating its maintenance, usage and accountability. Because without proper oversight AI and machine learning amplify inherent risks of using models in decision-making processes, and due to the increasingly stringent regulatory framework, AI systems require an even stronger governance (McKinsey & Company, 2021).

A rigid governance structure cannot match the needs and practices of all specific institutions and might not be appropriate for every specific AI system. Financial institutions will have to develop model governance approaches according to their needs. This section suggests five focus points to consider when moving from static, simple auxiliary statistical models to more complex, self-learning AI systems.

1: Create and maintain an overview of deployed AI systems

Many institutions have already developed several machine learning models, at least for internal use or as auxiliary input in the decision-making process. Other institutions might currently be using AI systems purchased by external vendors. Nonetheless, a governance structure for each of these systems is required.

Often the existence of these systems and how they are used are known only to direct stakeholders, with no centralised catalogue specifying how machine learning models have been constructed, which data the AI system exploits, and how these systems are used in the business practice of an institution. Building a model catalogue gives an overview of the AI systems that an institution has deployed across all potential stakeholders, and can be used for governing both the development and the procurement of AI systems.



2: Evaluate models and their risks through a lifecycle perspective

A solid governance structure is necessary once a model is deployed. However, too stringent requirements in terms of paperwork, documentation, and risk assessment can stifle the agile development of new solutions based on dynamic and changing business needs. Protecting and fostering creative innovation is key to unlock the potential of machine learning in the financial sector.⁴

Incrementally applying a model governance in line with the lifecycle of a machine learning model balances the need for documentation, structure and accountability with the need to innovate and experiment. Chart 1 shows a simplified stage-gate process for the development of a lifecycle approach to model governance.

At the stage of project prioritisation and resource allocation, a lean model impact risk assessment can ensure an agile governance approach appropriate for the AI system in question. At this stage, the behaviour of the final model is unknown, and the quality of available data is often uncertain.

The goal of the preliminary model impact risk assessment is then to determine as early as possible the appropriate model governance and shape the controls and constraints applicable to the model. Therefore, while high risk models require more stringent controls and higher standards, low-risk AI systems can be developed under more agile standards.⁵

The High-Level Expert Group on Artificial Intelligence set up by the European Commission has published a series of ethical guidelines for trustworthy AI (AI-HLEG, 2019), which includes an early assessment checklist. This list is meant to be comprehensive, and therefore many questions would be irrelevant for any

specific model. However, with proper logical rules in place allowing to focus only on the questions relevant for a specific AI system, this checklist can be a powerful tool on which to structure a preliminary model impact risk assessment.

As model development progresses and data is procured, identifying risks becomes easier (see box 6). The appropriate metrics for measuring e.g. algorithmic fairness can be selected according to the specific model, context, and use cases. Supporting the application of a governance process in parallel becomes easier, and ensures that time and resources are invested only in the necessary documentation, tests, and controls. While proof-of-concepts and preliminary investigations do not necessarily require an extensive, structured model governance, they should be exploited to inform an organic development of such systems for model deployment.

Well-defined governance, defined through broad stakeholder involvement (e.g. from business, risk management, and legal areas) and complemented by a pre-deployment risk assessment, should nonetheless be in place by the time of model deployment.

Once a model is deployed, its lifecycle is far from over. The performance of a machine learning model faced with a changing world can easily deteriorate over time. Regular re-assessments of its performance and adherence to pre-established minimum requirements is therefore necessary not only whenever a model is retrained, but at regular intervals, even if the model parameters remain unchanged.

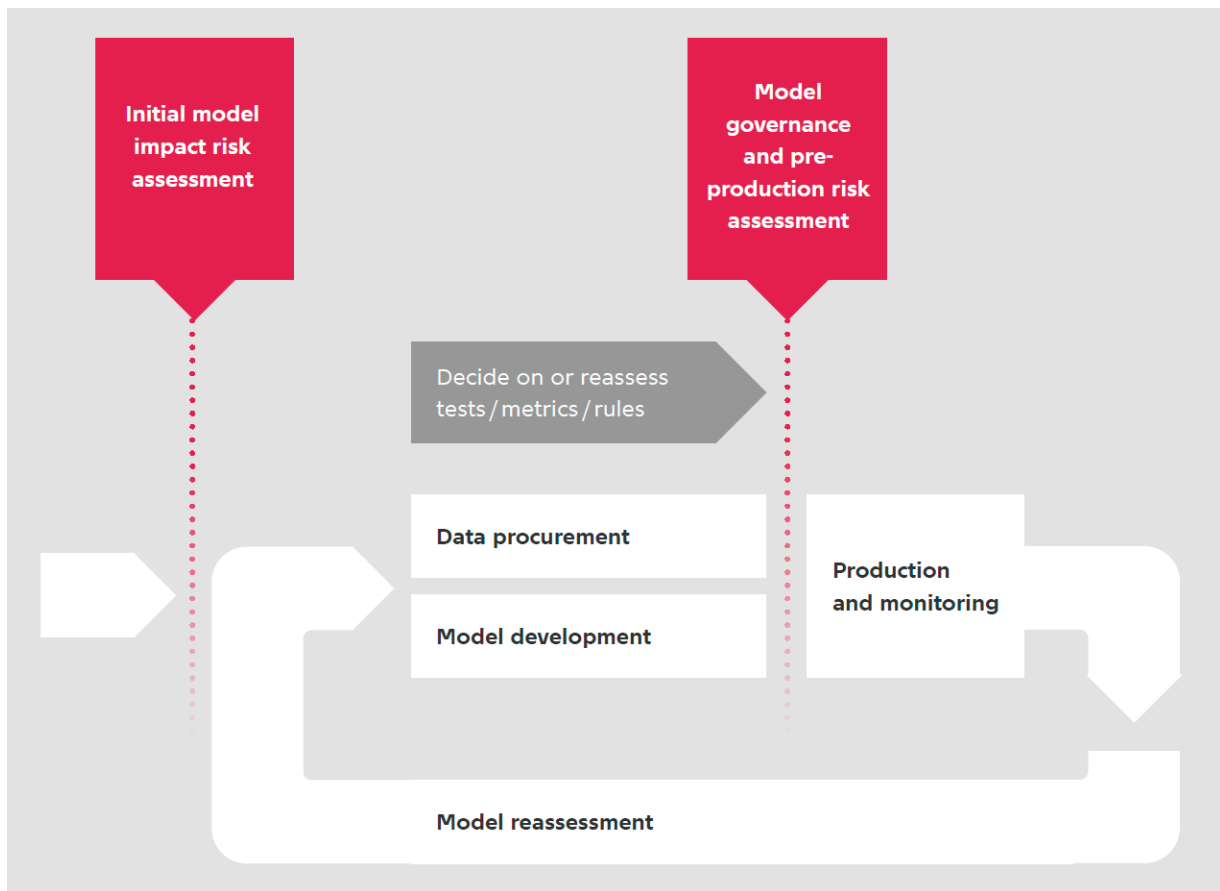
⁴ Regulators can also help the financial sector's innovation by encouraging experimentation in the adoption of AI. A noteworthy example of how US regulators have done this is in relation to the supervision over anti-money laundering (AML) systems. The Board of Governors of the Federal Reserve System, the Federal Deposit Insurance Corporation, the Financial Crimes Enforcement Network (FinCEN), the National Credit Union Administration and the Office of the Comptroller of the Currency released a joint statement on

innovative AML efforts, granting pilot AML programmes such as those exploiting AI systems grace from supervisory criticism, even if unsuccessful or if they expose gaps in existing AML compliance programmes. We are currently not aware of similar grace initiatives in Europe or Denmark.

⁵ Box 6 provides a taxonomy and overview of the risks associated with the development of a machine learning model.

An agile model governance approach following an AI system throughout its lifecycle

Chart 2



Note: The chart shows a simplified stage-gate process for aiding an agile development of a model governance structure. At project initiation, an initial model impact risk assessment can help determine the necessary checks, tests, metrics and rules that a model should be subject to, with high-risk models demanding stricter requirements. The appropriate governance structure, specific metrics, and monitoring strategies are then developed organically as data is procured and the AI system is developed. This concurring process helps ensuring that a proper governance structure is in place by the time of model deployment. After model deployment, periodical model reassessments take place, with their frequency depending on the pre-production risk assessment and governance structure. Each of these reassessments offer the chance to review not only models and data, but also the governance structure and KPIs, which can then evolve in line with the AI system itself.

The model governance should therefore define a system able to raise alarms in case algorithmic bias or performance metrics deviate from the acceptable range, with these ranges crucially depending on model risk. Automated systems, e.g. dashboards, are useful tools for these purposes.

Finally, when prioritising AI and machine learning projects, it is worth remembering that not only model governance, but also the maturity of an institution to AI systems need to grow organically over time. Establishing a common language requires diffused learning. Converging on a governance

structure satisfying the specific needs of an organisation takes efforts and most often involves a process of trial and error. For this reason, many organisations deliberately choose to begin developing and deploying low-risk models designed for internal use first and then progressively move towards use cases requiring stricter governance structures.

A taxonomy of risks associated with AI systems and machine learning models

Box 5

Deploying and maintaining a machine learning model carries multiple risks for an organisation, for instance:

- **Operational risks:** These include model risks, IT risks and all risks associated with the operational development and deployment of an AI system.
- **Legal risks:** These includes the risk of litigations, disputes and enforcements actions resulting from the application of the machine learning model, or non-compliance with existing regulations.
- **Reputational risks:** These include media exposure, issue management and reputational risks within the sector and among regulators following non-compliance or legal exposure.
- **Compliance risks:** These include the risk of non-compliance with existing regulations, not only specifically on AI systems, but also on fair treatment, capital requirements, privacy and data protection etc.

The relevance and severity of these risks might be very different both across a portfolio of machine learning models, and within a single model. A chatbot used for customer routing might carry low compliance risks, but substantive reputational risks in case of malfunctioning. A system used to predict corporate distress used internally in a financial institution might have low compliance risks, but high operational or financial risks in case bad model predictions are used uncritically to incorrectly price corporate loans and credit lines.

3: Evaluate models with appropriate skills and resources

Complying with regulation requires not only the establishment of a common language across business areas ranging from data science to legal and risk management, but also translating abstract regulatory principles into measurable minimum requirements. As models, data, and applications are context-specific, there can be no single recipe for ensuring regulatory compliance. Standardisations (e.g. ISO standards) can help ensure that minimum governance requirements are satisfied but cannot guide an institution on choosing the appropriate metrics for performance, fairness, and explainability.

Thoroughly evaluating the robustness and performance of machine learning models therefore requires specific competences, and risk managers cannot lift this task alone. Data scientists, machine learning engineers, and developers should therefore not only be confined to model development and maintenance, but also allocated to model testing, vetting, and supervision. Separate roles ought to oversee the task of development and that of vetting and monitoring a model's performance, as role

separation enhances the quality of model oversight by limiting conflicts of interest.

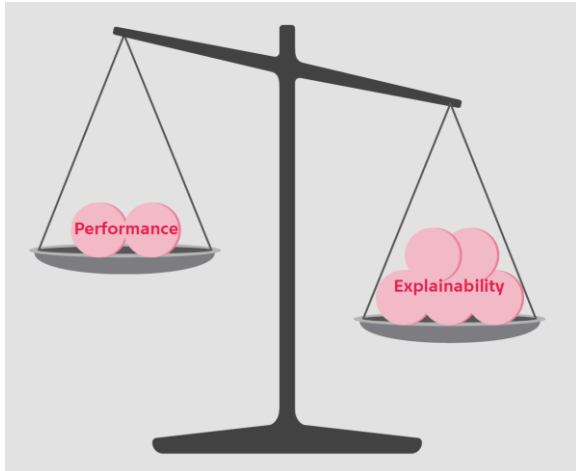
4: Consider whether marginal performance is worth added complexity

Minimising risks associated with machine learning models requires ensuring model explainability and adjusting the model if suspicions of model bias and unfairness arise. Taking these steps can have the practical consequence of decreasing a model's predictive performance (Hardt, Price, & Srebro, 2016).

Chart 3 illustrates this trade-off. Not every step taken to ensure model interpretability and algorithmic fairness implies decreased model performance. Graphical inspection of feature influences is a purely descriptive approach that does not change the structure of a model. Improving data collection to ensure that data used for training a model is unbiased can also increase model performance.

Machine learning models can imply a trade-off between their performance and their explainability

Chart 3



However, ensuring the consistency of how specific variables influence the model predictions with economic intuitions necessarily places constraints on a model designed to freely adjust its parameters to optimise performance. These steps should nonetheless be taken. Marginal increases in model performance are rarely worth increased compliance and operational risks.

Overall, risks should be justifiable. Increases in model and business performance resulting from switching from simpler models to complex machine learning systems and AI systems should be sufficient to justify undertaking additional risks.

5: Share best practices

Any single organisation is responsible for implementing best governance practices in their use of machine learning models. Yet, risks (reputational, financial etc.) of misuse and abuse of these models affect the entire financial sector.

Mitigating risks through appropriate governance improves the stability of not only individual financial institutions, but of the financial sector as a whole. To encourage a common strong adherence to regulatory principles, we encourage financial market

participants to periodically share implemented and planned best practice in terms of machine learning and AI system governance both internally and externally.

Internally, sharing best practices, strategies, and experiences across different business areas of an institution spreads knowledge within an organisation and fosters a common language across business areas, which facilitates new applications. Externally, it facilitates the development of common governance practices and of standards, and aids compliance efforts across the financial sector.

Summary

This paper argues that while AI and machine learning by themselves do not represent a paradigm shift for a financial institution, they amplify operational, compliance, legal, and reputational risks associated with model deployment. As a consequence, the regulatory framework is becoming increasingly stringent and severe. The recent proposal by the European Commission specifically represents a regulatory milestone, as it will impose strict governance requirements to, among others, models used for the risk assessment and distress probability of natural persons. Non-compliance with such regulation will result in fines as high as 6 per cent of a company's worldwide revenue.

Ensuring regulatory compliance and mitigating inherent risks requires institutions to develop appropriate model governance structures. This paper illustrates five focus points for financial institutions to consider when moving from simple, static statistical models to complex, self-learning machine learning algorithms. With proper supervision, oversight and model governance, machine learning and AI can deliver their promised potential and improve the resiliency of financial institutions and the robustness of the financial sector as a whole.

References

- 2000/43/EC. (2000, June 29). Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.
- 2000/78/EC. (2000, November 27). Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.
- 2004/113/EC. (2004, December 13). Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services.
- 2006/54/EC. (2006, July 05). Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast).
- 2019/876. (2019, 6 7). Regulation (EU) 2019/876 of the European Parliament and of the Council of 20 May 2019 amending Regulation (EU) No 575/2013 as regards the leverage ratio, the net stable funding ratio, requirements for own funds and eligible liabilities, counterparty credit.
- (2022, 03 10). Retrieved from Wikipedia/Fairness (machine learning): [https://en.wikipedia.org/wiki/Fairness_\(machine_learning\)#Metrics](https://en.wikipedia.org/wiki/Fairness_(machine_learning)#Metrics)
- 575/2013. (2013, 06 27). Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012 Text with EEA relevance.
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 52138-52160.
- AI-HLEG. (2019, April 8). Retrieved from <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.1.html>
- AIPPF. (2022). *The AI Public-Private Forum: Final report*. Bank of England.
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias*. Retrieved from ProPublica: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Arrieta, A., Diaz-Rodriguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 82-115.
- Babina, T., Fedyk, A., He, A., & Hodson, J. (2021). *Artificial Intelligence, Firm Growth, and Product Innovation*.
- BaFin. (2021). *Maschinelles Lernen in Risikomodellen - Charakteristika und aufsichtliche Schwerpunkte*. Bundesanstalt für Finanzdienstleistungsaufsicht.
- Banerjee, I., Bhimireddy, A., Burns, J., Celi, A. L., Chen, L.-C., Correa, R., . . . Gichoya, J. (2021). *Reading Race: AI Recognises Patient's Racial Identity In Medical Images*. Retrieved from <https://arxiv.org/abs/2107.10356>
- COM/2021/206 final. (2021, April 21). 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.
- EBA. (2020). *Report on Big Data and Advanced Analytics*. European Banking Authority (EBA).

- EBA. (2021). *Discussion paper on machine learning for IRB models*. European Banking Authority (EBA).
- Finanstilsynet. (2019). *God praksis ved brug af superviseret machine learning*.
- Fuster, A., Goldsmith-Pinkham, P., & Ramadorai, T. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance*, 5-47.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems* .
- ISO. (2022, 03 06). *Standards by ISO/IEC JTC 1/SC 42 - Artificial Intelligence*. Retrieved from iso.org:
<https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0>
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018, may). Algorithmic fairness. *Aea papers and proceedings*, pp. 22-27.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. (2018). Discrimination in the Age of Algorithms. 113-174.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. (2020). Algorithms as discrimination detectors. *PNAS*, 117(48).
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2018). Inherent trade-offs in the fair determination of risk scores. *ACM SIGMETRICS Performance Evaluation Review*.
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*.
- Martinello, A. T., Mønsted, B. M., Matin, R., Steffensen, S. A., & Laursen, K. H. (2021). *Female business owners pay higher interest rates on corporate loans*. Nationalbanken Working Paper Nr. 179.
- McKinsey & Company. (2021). *A strategic vision for model risk management*.
- Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Christian, S., . . . Bischl, B. (2021). *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*. Retrieved from <https://arxiv.org/abs/2007.04131>
- OECD. (2021). *Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers*. Retrieved from <https://www.oecd.org/finance/artificial-intelligence-machine-learning-big-data-in-finance.htm>
- Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications : A survey. *Applied Soft Computing*, 106384.
- Refinitiv. (2020). *The rise of the data scientist: Machine learning models for the future*. Retrieved from <https://www.refinitiv.com/perspectives/ai-digitalization/machine-learning-new-research-reveals-is-maturing-in-finance/>
- Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 42200-42216.
- Sudjianto, A., & Zhang, A. (2021). *Designing Inherently Interpretable Machine Learning Models*. arXiv preprint arXiv:2111.01743.

Data in new ways

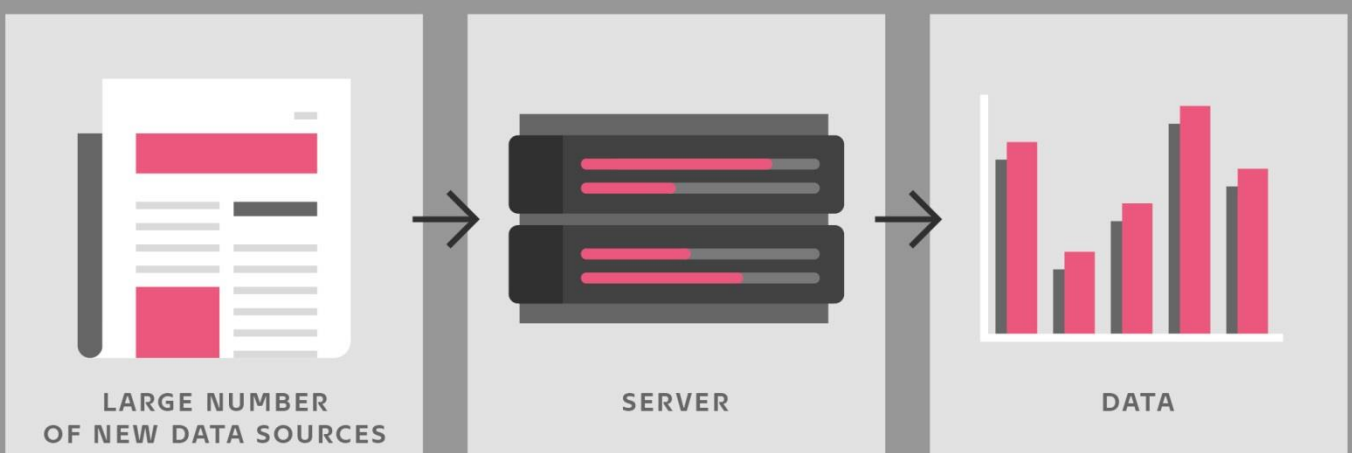
Data volumes have grown exponentially. By 2025, an estimated 450 exabytes of data will be created each day.

This is equivalent to hundreds of millions of personal computers being filled with data on a daily basis. The vast volumes of data are highly diverse, but new and sophisticated methods enable analysis of this data in new and more efficient ways.

New data types and new data collection methods may be used in various contexts in Danmarks Nationalbank's ongoing work.

In order to acquire more knowledge and a better basis for assessing the Danish economy, Danmarks Nationalbank focuses on new data types and methods in a series of publications of which this Economic Memo is one.

New data creates new knowledge



PUBLICATIONS



NEWS

News offers quick and accessible insight into an Analysis, an Economic Memo, a Working Paper or a Report from Danmarks Nationalbank. News is published continuously.



ANALYSIS

Analyses from Danmarks Nationalbank focus on economic and financial matters. Some Analyses are published at regular intervals, e.g. *Outlook for the Danish economy* and *Financial stability*. Other Analyses are published continuously.



REPORT

Reports comprise recurring reports and reviews of the functioning of Danmarks Nationalbank and include, for instance, the *Annual report* and the annual publication *Danish government borrowing and debt*.



ECONOMIC MEMO

An Economic Memo is a cross between an Analysis and a Working Paper and often shows the ongoing study of the authors. The publication series is primarily aimed at professionals. Economic Memos are published continuously.



WORKING PAPER

Working Papers present research projects by economists in Danmarks Nationalbank and their associates. The series is primarily targeted at professionals and people with an interest in academia. Working Papers are published continuously.

DANMARKS NATIONALBANK
LANGELINIE ALLÉ 47
DK-2100 COPENHAGEN Ø
WWW.NATIONALBANKEN.DK

Danmarks Nationalbank's Economic Memos are published at www.nationalbanken.dk. A free electronic subscription is also available at the website. The subscriber receives an e-mail notification whenever a new Economic Memo is published.

Text may be copied from this publication provided that the source is specifically stated. Changes to or misrepresentation of the content are not permitted.

Please direct any enquiries directly to the contributors or to Danmarks Nationalbank, Communications, Kommunikation@nationalbanken.dk.



**DANMARKS
NATIONALBANK**