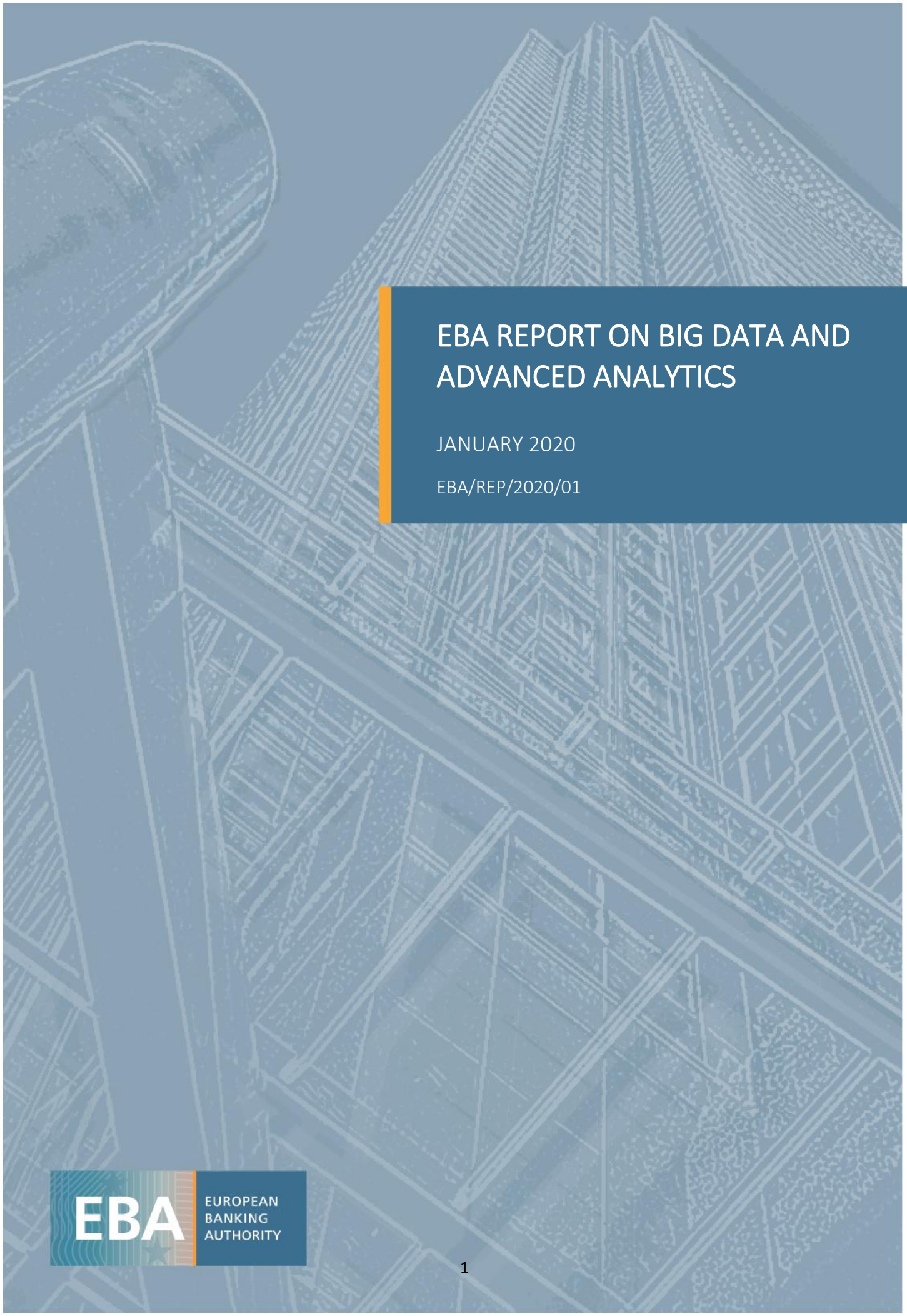


EBA REPORT ON BIG DATA AND ADVANCED ANALYTICS

JANUARY 2020

EBA/REP/2020/01

EBA EUROPEAN BANKING AUTHORITY

Contents

Abbreviations 3

Executive summary 4

Background 8

1. Introduction 11

1.1 Key terms 12

1.2 Types of advanced analytics 14

1.3 Machine-learning modes 15

2. Current landscape 16

2.1 Current observations 16

2.2 Current application areas of BD&AA 19

3. Key pillars 25

3.1 Data management 25

3.2 Technological infrastructure 27

3.3 Organisation and governance 28

3.4 Analytics methodology 29

4. Elements of trust in BD&AA 35

4.1 Ethics 35

4.2 Explainability and interpretability 35

4.3 Fairness and avoidance of bias 37

4.4 Traceability and auditability (including versioning) 39

4.5 Data protection and quality 40

4.6 Security 41

4.7 Consumer protection 42

5. Key observations, risks and opportunities 43

5.1 Key observations 43

5.2 Key opportunities 43

5.3 Key risks and proposed guidance 44

6. Conclusions 47

Annex I 49

Annex II 53

Annex III 58

Abbreviations

AI	Artificial Intelligence
AML/CFT	Anti-Money Laundering/Countering the financing of terrorism
API	Application Programming Interface
BD&AA	Big Data and Advanced Analytics
CCTV	Closed-Circuit television
EBA	European Banking Authority
ECB	European Central Bank
ESAs	European Supervisory Authorities
EU	European Union
FinTech	Financial Technology
GDPR	General Data Protection Regulation
GPS	Global Positioning System
ICT	Information and Communication Technology
ML	Machine Learning
NIST	US National Institute of Standards and Technology
NLP	Natural Language Processing
RegTech	Regulatory Technology
SupTech	Supervisory Technology

Executive summary

A data-driven approach is emerging across the banking sector, affecting banks' business strategies, risks, technology and operations. Corresponding changes in mindset and culture are still in progress. Following the cross-sectoral report by the Joint Committee of the European Supervisory Authorities (ESAs) on the use of big data by financial institutions[1], and in the context of the EBA FinTech Roadmap, the EBA decided to pursue a 'deep dive' review on the use of big data and Advanced Analytics (BD&AA) in the banking sector. The aim of this report is to share knowledge among stakeholders on the current use of BD&AA by providing useful background on this area, along with key observations, and presenting the key pillars and elements of trust that could accompany their use.

The report focuses on BD&AA techniques and tools, such as machine learning (ML) (a subset of Artificial Intelligence (AI)), that go beyond traditional business intelligence to gain deeper insights, make predictions or generate recommendations using various types of data from various sources. ML is certainly one of the most prominent AI technologies at the moment, often used in advanced analytics due to its ability to deliver enhanced predictive capabilities.

BD&AA are driving fundamental change in institutions' business models and processes. Currently, BD&AA are part of most institutions' digital transformation programmes, along with the growing use of cloud services, which is perceived in some instances to facilitate the use of BD&AA. Core banking data are currently the main flow feeding data analytics, rather than other data sources such as external social media data, due to institutions' concerns about the reliability and accuracy of external data. A key constraint for institutions is the integration of BD&AA into existing business processes, as they recognise the need to develop relevant knowledge, skills and expertise in this area. Institutions appear to be at an early stage of ML use, with a focus on predictive analytics that rely mostly on simple models; more complex models can bring better accuracy and performance but give rise to explainability and interpretability issues. Other issues such as accountability, ethical aspects and data quality need to be addressed to ensure responsible use of BD&AA. At this stage, institutions leverage BD&AA mainly for customer engagement and process optimisation purposes (including RegTech), with a growing interest in the area of risk management.

Key pillars of BD&AA

This report identifies four key pillars for the development, implementation and adoption of BD&AA, which interact with each other and are thus not mutually exclusive. These pillars require review by institutions to ensure they can support the roll-out of advanced analytics.

[1] https://eba.europa.eu/documents/10180/2157971/Joint+Committee+Final+Report+on+Big+Data+%28JC-2018-04+%29.pdf

The four pillars are listed below.

(i) **Data management**

Data management enables the control and security of data for enterprise purposes taking into account data types and data sources, data protection and data quality. A successful data management approach, which builds trust and meets legal requirements, could lead to improved decision-making, operational efficiency, understanding of data and regulatory compliance.

(ii) **Technological infrastructure**

Technological infrastructure entails processing, data platforms and infrastructure that provide the necessary support to process and run BD&AA.

(iii) **Organisation and governance**

Appropriate internal governance structures and organisational measures, along with the development of sufficient skills and knowledge, support the responsible use of BD&AA across institutions and ensure robust oversight of their use.

(iv) **Analytics methodology**

A methodology needs to be in place to facilitate the development, implementation and adoption of advanced analytics solutions. The development of an ML project follows a lifecycle with specific stages (e.g. data preparation, modelling, monitoring) that differs from the approach adopted for standard business software.

The elements of trust

The report finds that the roll-out of BD&AA specifically affects issues around trustworthiness and notes a number of fundamental trust elements that need to be properly and sufficiently addressed and which cut across the four key pillars. Efforts to ensure that AI/ML solutions built by institutions respect these trust elements could have implications for all the key pillars. The trust elements are:

- **Ethics:** in line with the *Ethics guidelines for trustworthy AI* from the European Commission's High-Level Expert Group on AI[2], the development, deployment and use of any AI solution should adhere to some fundamental ethical principles, which can be embedded from the start in any AI

[2] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

project, in a sort of 'ethical by design' approach that can influence considerations about governance structures.

- **Explainability and interpretability:** a model is explainable when its internal behaviour can be directly understood by humans (interpretability) or when explanations (justifications) can be provided for the main factors that led to its output. The significance of explainability is greater whenever decisions have a direct impact on customers/humans and depends on the particular context and the level of automation involved. Lack of explainability could represent a risk in the case of models developed by external third parties and then sold as 'black box' (opaque) packages.

  Explainability is just one element of **transparency**. Transparency consists in making data, features, algorithms and training methods available for external inspection and constitutes a basis for building trustworthy models.

- **Fairness and avoidance of bias:** fairness requires that the model ensure the protection of groups against (direct or indirect) discrimination[3]. Discrimination can be a consequence of bias in the data, when the data are not representative of the population in question. To ensure fairness, the model should be free from bias. Note, however, that bias can be introduced in many ways. Techniques for preventing or detecting bias exist and continue to evolve (a current research field).

- **Traceability and auditability:** the use of traceable solutions assists in tracking all the steps, criteria and choices throughout the process, which enables the repetition of the processes resulting in the decisions made by the model and helps to ensure the auditability of the system.

- **Data protection:** data should be adequately protected with a trustworthy BD&AA system that complies with current data protection regulation.

- **Data quality:** the issue of data quality needs to be taken into account throughout the BD&AA lifecycle, as considering its fundamental elements can help to gain trust in the data processed.

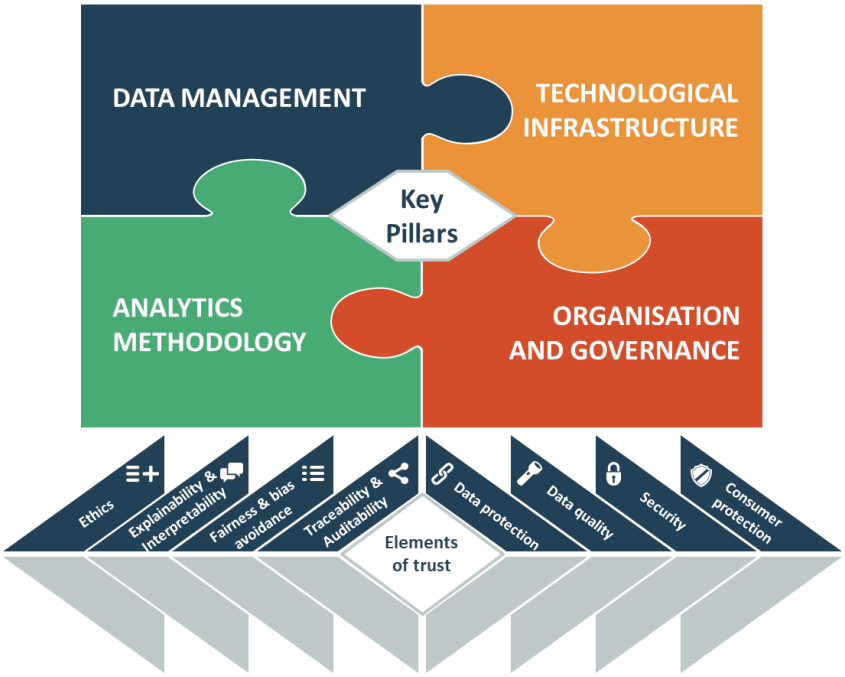
- **Security:** new technology trends also bring new attack techniques exploiting security vulnerabilities. It is important to maintain a technical watch on the latest security attacks and related defence techniques and ensure that governance, oversight and the technical infrastructure are in place for effective ICT risk management.

- **Consumer protection:** a trustworthy BD&AA system should respect consumers' rights and protect their interests. Consumers are entitled to file a complaint and receive a response in plain language that can be clearly understood[4]. Explainability is key to addressing this obligation.

[3] Discrimination (intentional or unintentional) occurs when a group of people (with particular shared characteristics) is more adversely affected by a decision (e.g. an output of an AI/ML model) than another group, in an inappropriate way.

[4] https://eba.europa.eu/documents/10180/732334/JC+2014+43+-+Joint+Committee+-+Final+report+complaints-handling+guidelines.pdf/312b02a6-3346-4dff-a3c4-41c987484e75

Figure 0.1: Key pillars and elements of trust in BD&AA



It was observed that, within institutions, the specific implementation of the key pillars may change over time. For example, from a regulatory perspective, the EBA's *Guidelines on internal governance*[5], on outsourcing arrangements[6] and on ICT and security risk management[7] set the baseline for a sound internal governance and resilient risk management framework. Nevertheless, technological infrastructure remains an ongoing challenge for most institutions as they deal with related legacy issues. In addition, the use of new, often diverse, sources of data and increased recognition of citizens' rights over that data creates specific challenges for data management inside institutions, which require attention and possibly targeted action.

Moreover, the need to build the trust elements into the development of advanced analytics applications, for example to ensure the explainability and ethical design of such solutions, will require ongoing work.

Going forward, the EBA will continue to observe (taking into account also other work being done by the ESAs and work being done in other international fora) and consider the pace of evolution of BD&AA in financial services (in line with its FinTech Roadmap), and, where appropriate, it will accompany this work with opinions and/or proposals for guidelines to achieve a coordinated approach to the regulatory and supervisory treatment of AI and BD&AA activities.

[5] https://eba.europa.eu/regulation-and-policy/internal-governance/guidelines-on-internal-governance-revised-

[6] https://eba.europa.eu/regulation-and-policy/internal-governance/guidelines-on-outsourcing-arrangements

[7] https://eba.europa.eu/eba-publishes-guidelines-ict-and-security-risk-management

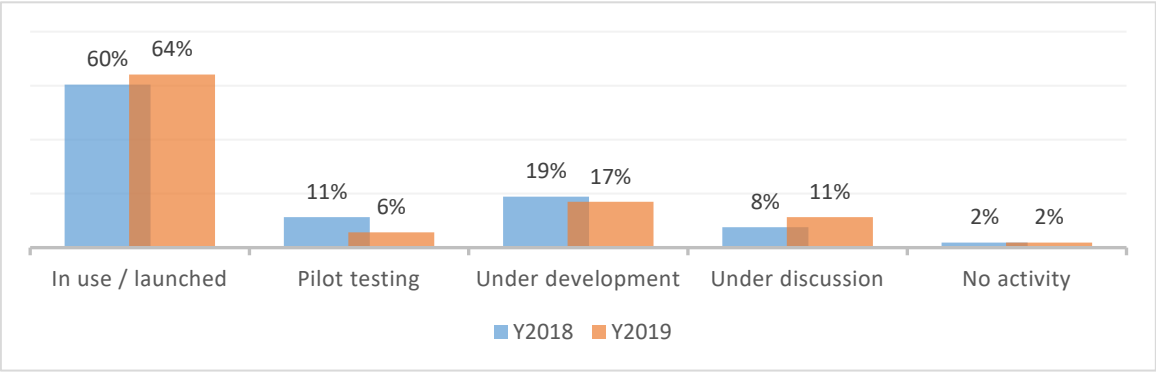
Background

Article 1(5) of the Regulation establishing the EBA (Regulation (EU) No 1093/2010) requires the EBA to contribute to promoting a sound, effective and consistent level of regulation and supervision, ensuring the integrity, transparency, efficiency and orderly functioning of financial markets, preventing regulatory arbitrage and promoting equal competition. In addition, Article 9(2) requires the EBA to monitor new and existing financial activities.

These mandates are key motivations underpinning the EBA's interest in financial innovation in general and more specifically in FinTech. The EBA decided to take forward work in relation to FinTech by publishing its FinTech Roadmap setting out its priorities for 2018/2019. One of the priorities set out in the EBA FinTech Roadmap is the analysis of the prudential risks and opportunities for institutions arising from FinTech, including with regard to the development and sharing of knowledge among regulators and supervisors. This thematic report, a step towards this priority, follows the EBA's *Report on the prudential risks and opportunities for institutions arising from FinTech*[8] as well as the ESAs' *Joint Committee final report on big data*[9].

In the context of its ongoing monitoring, the EBA has observed a growing interest in the use of Big Data Analytics (as noted in the EBA risk assessment questionnaires); institutions see potential in the use of advanced analytics techniques, such as ML, on very large, diverse datasets from different sources and of different sizes. Figure 0.2 shows that institutions are using BD&AA to a significant extent in their operations, with 64% of institutions reporting having already launched BD&AA solutions, while within 1 year around 5% of institutions moved from a pilot testing and/or development phase to deployment. In general, almost all institutions are exploring the use of BD&AA.

Figure 0.2: Use of Big Data Analytics across EU institutions



Source: EBA risk assessment questionnaires (autumn 2018 and autumn 2019)

8 https://eba.europa.eu/documents/10180/2270909/Report+on+prudential+risks+and+opportunities+arising+for+institutions+from+FinTech.pdf

9 https://www.esma.europa.eu/sites/default/files/library/jc-2018-04_joint_committee_final_report_on_big_data.pdf

This report provides background information on BD&AA, along with an educational perspective, and describes the current landscape as regards their use in the banking sector, without making policy recommendations or setting supervisory expectations in this regard. It aims to share knowledge about and enhance understanding of the practical use of BD&AA, noting the risks and challenges currently arising from the implementation of such solutions, such as the integration and coordination of institutions' legacy infrastructures with new big data technologies.

For the purposes of this report, the EBA has engaged with a number of stakeholders (e.g. credit institutions, technology providers, academics and data protection supervisors) to better understand the current developments and approaches as well as to exchange views on this area. The report was also enriched by input from the EBA risk assessment questionnaires (conducted on a semi-annual basis among banks), discussions with competent authorities and input from their subject matter experts, a literature review and desk research.

During its interactions with the industry, the EBA noted that the development of BD&AA applications in the banking sector is at an early stage (in terms of sophistication and scope), with growing investments and potential opportunities. Therefore, it is important for the regulatory and supervisory community to understand and closely follow these developments to ensure any potential risks posed by BD&AA are properly managed going forward.

This report adheres to the EBA's overall approach to FinTech with regard to technological neutrality and future-proofing, as it does not intend to pre-empt and prescribe the use of BD&AA across the banking sector. The structure of the report can be summarised as follows.

Section 1 – Introduction: this provides an overall introduction to the report and basic background information, for example on key terms, types of advanced analytics and ML modes.

Section 2 – Current landscape: this describes the current landscape, including high-level observations on the use of advanced analytics in banking (mainly based on industry interactions and experience to date). This section also includes a general description of current applications of BD&AA, supported by data from the EBA risk assessment questionnaire (autumn 2019).

Section 3 – Key pillars: this section illustrates the four key pillars for the development, implementation and adoption of BD&AA. These pillars (data management, technological infrastructure, organisation and governance, and analytics methodology) are described in more detail, including the important steps in the ML process, such as data preparation and modelling.

Section 4 – Elements of trust in BD&AA: the overarching elements of trust to be respected throughout the development, implementation and adoption of BD&AA are discussed in this section, such as ethics, explainability, interpretability, traceability and auditability.

Section 5 – Key messages: in this section, all the key messages presented in Sections 1-4 are summed up, in an effort to clearly convey the key messages of this report.

Section 6 – Conclusions: final remarks and views are presented in this section, including proposals and thoughts on the way forward in this area.

A number of relevant publications (e.g. the European Commission's High-Level Expert Group on AI's *Ethics guidelines for trustworthy AI*, the Basel Committee on Banking Supervision's *Sound practices on the implications of FinTech developments for banks and bank supervisors*, the Financial Stability Board's *Artificial intelligence and machine learning in financial services* and recent publications on AI from a number of competent authorities), have been taken into account for the purposes of this report.

1. Introduction

Technological change is leading to increasing amounts of data being collected, processed, shared and used in digital form at lower cost and on a larger scale. Managing data is not new but the ability to store huge amounts of data in any format and analyse it at speed is. The growing volume and increased analysis of data have led to the emergence of **Big Data**. There are many definitions of Big Data but for the purposes of this report we have used the ESAs' tentative definition[10], according to which Big Data refers to large volumes of different types of data, produced at high speed from many and varied sources (e.g. the internet of things, sensors, social media and financial market data collection), which are processed, often in real time, by IT tools (powerful processors, software and algorithms).

Big Data innovations in the leisure and retail sectors have created financial services customers who increasingly expect a more personalised service. For example, platform-based business models have been designed to significantly increase the number of providers a consumer can access, as well as the number of providers in any market. End users are also changing the way in which they pay for goods and services, with an increasing reliance on the use of non-cash payment services, which generates a digital data footprint that can be monetised. These developments are actively changing the role of the intermediary in financial services, and industry incumbents are adapting to meet changing consumer demands.

To analyse Big Data, institutions are increasingly using **advanced analytics**. Advanced analytics include predictive and prescriptive analytical techniques, often using AI and ML in particular, and are used to understand and recommend actions based on the analysis of high volumes of data from multiple sources, internal or external to the institution. Typical use cases include customer onboarding, fraud detection and back office process automation.

Cloud computing has been an enabler of advanced analytics, as the cloud provides a space to easily store and analyse large quantities of data in a scalable way, including through easy connectivity to mobile applications used by consumers. Tools available for data science purposes for use on site and/or in cloud-based environments also appear to be increasing.

As digital transformation continues, BD&AA may be used in an effort to influence and direct consumer behaviour, and as open banking and BD&AA evolve they will raise compelling questions. For example, the use of AI in financial services will raise questions about whether it is socially beneficial, whether it creates or reinforces bias, and whether the technology used is built and tested to prevent harm. Other questions – relating, for example, to accountability and how advanced analytics technology is controlled by humans, whether privacy safeguards are appropriate and transparent, and what scientific rigour and integrity sits behind the design of the

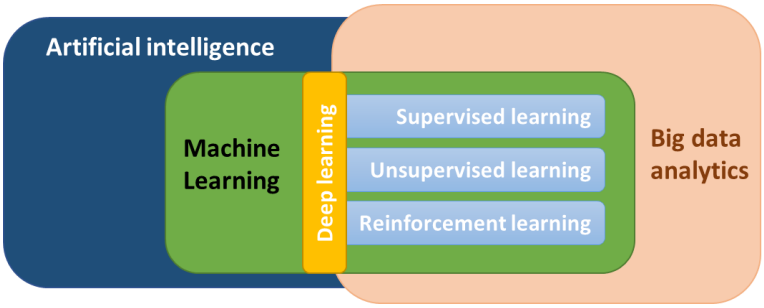
[10] https://eba.europa.eu/documents/10180/2157971/Joint+Committee+Final+Report+on+Big+Data+%28JC-2018-04+%29.pdf

technology – will need to be considered, taking into account ethical issues regarding privacy, bias, fairness and explainability.

As BD&AA become more prevalent, supervisors are also juggling an environment where customers and institutions can maximise the opportunities of competition, innovation and Big Data while seeking to ensure that customers do not suffer harm because of innovation.

This report considers the use of and interaction between BD&AA and their potential for use in financial services. In considering such interaction, ML was identified as a particular branch of AI that institutions are currently using. Key features of ML are set out in Section 4 to help readers understand the potential risks and benefits that may come with its use.

Figure 1.1: Pivotal role of ML in AI and Big Data analytics



Source: Financial Stability Board, 'Artificial intelligence and machine learning in financial services' (November 2017)

1.1 Key terms

For the purposes of this report, it is important to create a common understanding of the key terms to allow the reader to fully appreciate the substantive sections. In this regard, standard/pre-existing definitions from international bodies have been used whenever they were available.

Big Data

Big Data generally refers to technological developments related to data collection, storage, analysis and applications. It is often characterised by the increased volume, velocity and variety of data being produced (the three Vs) and typically refers (but is not limited) to data from the internet. In addition, increased variability with respect to consistency of data over time, veracity with respect to accuracy and data quality, and complexity in terms of how to link multiple datasets are characteristics of Big Data[11]. However, as noted in the ESAs' *Joint Committee final report on big data*,[12] any definition of a fast-evolving phenomenon such as Big Data should remain flexible to accommodate the inevitable need for future adjustments.

[11] Mario Callegaro and Yongwei Yang (2017), 'The role of surveys in the era of Big Data'.

[12] https://eba.europa.eu/documents/10180/2157971/Joint+Committee+Final+Report+on+Big+Data+%28JC-2018-04+%29.pdf

Big Data come from a variety of sources and include social media data and website metadata. The internet of things contributes to Big Data, including behavioural location data from smartphones and fitness-tracking devices. In addition, transaction data from the business world form part of Big Data, providing information on payments and administrative functions. The increased availability of data has led to improved technologies for analysing and using data, for example in the area of ML and AI.

Advanced analytics

Advanced analytics can be defined as 'the autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools, typically beyond those of traditional business intelligence'; it is often based on ML, 'to discover deeper insights, make predictions, or generate recommendations. Advanced analytics techniques include those such as data/text mining, machine learning, pattern matching, forecasting, visualization, semantic analysis, sentiment analysis, network and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, neural networks'[13].

Data science

Data science is an interdisciplinary field involving extracting information and insights from data available in both structured and unstructured forms, similar to data mining. However, unlike data mining, data science includes all steps associated with the cleaning, preparation and analysis of the data. Data science combines a large set of methods and techniques encompassing programming, mathematics, statistics, data mining and ML. Advanced analytics is a form of data science often using ML.

Artificial intelligence

The independent High-Level Expert Group on AI set up by the European Commission has recently proposed the following updated definition of AI[14], which has been adopted for the purposes of this report: 'Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from these data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimisation), and robotics (which includes

[13] https://www.gartner.com/it-glossary/advanced-analytics/

[14] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)'.

Currently, many AI applications, particularly in the financial sector, are 'augmented intelligence' solutions, i.e. solutions focusing on a limited number of intelligent tasks and used to support humans in the decision-making process.

Machine learning

The standard on IT governance ISO/IEC 38505-1:2017 defines ML as a 'process using algorithms rather than procedural coding that enables learning from existing data in order to predict future outcomes'.

ML is one of the most prominent AI technologies at the moment, often used in advanced analytics due to its ability to deliver enhanced predictive capabilities. ML comes in several modes, and the main ones are described in Section 1.3.

1.2 Types of advanced analytics

Advanced analytics techniques extend beyond basic descriptive techniques and can be categorised under four headings:

- **Diagnostic analytics:** this is a sophisticated form of backward-looking data analytics that seeks to understand not just what happened but why it happened. This technique uses advanced data analytics to identify anomalies based on descriptive analytics. It drills into the data to discover the cause of the anomaly using inferential statistics combined with other data sources to identify hidden associations and causal relationships.

- **Predictive analytics:** this forward-looking technique aims to support the business in predicting *what could happen* by analysing backward-looking data. This involves the use of advanced data mining and statistical techniques such as ML. The goal is to improve the accuracy of predicting a future event by analysing backward-looking data.

- **Prescriptive analytics:** this technique combines both backward- and forward-looking analytical techniques to *suggest an optimal solution based on the data available at a given point in time*. Prescriptive analytics uses complex statistical and AI techniques to allow flexibility to model different business outcomes based on future risks and scenarios, so that the impact of the decision on the business can be optimised.

- **Autonomous and adaptive analytics:** this technique is the most complex and uses *forward-looking predictive analytics models that automatically learn from transactions and update results in real time* using ML. This includes the ability to self-generate new algorithmic models with suggested insights for future tasks, based on correlations and patterns in the data that the system has identified and on growing volumes of Big Data.

1.3 Machine-learning modes

As mentioned in Section 1.1, ML is a subcategory of AI that uses algorithms able to recognise patterns in large amounts of data via a learning process in order to make predictions based on similar data. For this reason, ML is very often used in predictive analytics solutions.

The learning is done by means of suitable **algorithms**, which are used to create predictive **models**, representing what the algorithm has learnt from the data in order to solve the particular problem. Their performance improves as more data are available to learn from (to train the model).

ML algorithms can be grouped based on the learning mode.

- In **supervised learning**, the algorithm learns from a set of training data (observations) that have labels (e.g. a dataset composed of past transactions with a label indicating whether the transaction is fraudulent or not). The algorithm will learn a general rule for the classification (the model), which will then be used to predict the labels when new data are analysed (e.g. data on new transactions).

- **Unsupervised learning** refers to algorithms that will learn from a dataset that does not have any labels. In this case, the algorithm will detect patterns in the data by identifying clusters of similar observations (data points with common features). Important problems addressed using unsupervised learning algorithms are clustering, anomaly detection and association.

- In **reinforcement learning**, rather than learning from a training dataset, the algorithm learns by interacting with the environment. In this case, the algorithm chooses an action starting from each data point (in most cases the data points are collected via sensors analysing the environment) and receives feedback indicating whether the action was good or bad. The algorithm is therefore trained by receiving rewards and 'punishments'; it adapts its strategy to maximise the rewards.

  Furthermore, regardless of the mode adopted, some complex ML solutions can use a deep-learning approach.

- **Deep learning** means learning using deep neural networks. **Neural networks** are a particular type of ML algorithms that generate models inspired by the structure of the brain. The model is composed of several layers, with each layer being composed of units (called **neurons**) interconnected with each other. Deep-learning algorithms are neural networks that have many hidden layers (the number of layers can vary from tens to thousands), which can make their structure very complicated, so much so that they can easily become black boxes.

2. Current landscape

2.1 Current observations

In the context of its ongoing monitoring of financial innovation, and through its interactions with the competent authorities and stakeholders, the EBA has made a number of observations in the area of BD&AA that are relevant to the financial sector, including the following.

Level of BD&AA adoption

Institutions are currently developing or implementing **digital transformation** programmes, which include the growing use of advanced analytics across business functions. Institutions are at different stages in the use of AI and other techniques (along with the related governance), including the extent to which they incorporate digital and data aspects into their core strategies, which can act as a strong lever to reshape the entire business model. The time required for institutions to move to an advanced analytics solution, from the early stages to production stage, ranges from 2 to 12 months.

In general, a number of factors appear to affect the adoption of new technologies by institutions. While **risk controls** remain a strong concern for institutions, the following points have also been observed in relation to the use of advanced analytics.

- Institutions appear to be at an early stage of ML use, with a focus on predictive analytics that rely on simple models, prioritising explainability and interpretability over accuracy and performance.

- Material processes seem to remain on premises or be shifted incrementally to private or hybrid clouds, rather than an all or nothing moves to the public cloud taking place.

- The integration of BD&AA solutions into existing legacy systems is a challenge for many institutions and currently acts as a key constraint on the adoption and deployment of ML solutions.

- Core banking data (i.e. proprietary and internal structured data) are currently the main flow feeding data analytics, rather than other data sources such as external social media data, due to concerns about and issues with the reliability and accuracy of external data.

- Institutions recognise the need to develop knowledge, skills and expertise in AI. For example, institutions may not fully understand the different forms of advanced analytics or may not be aware of the potential existence of bias in the data or in the algorithmic model itself.

Many institutions engage external technology providers for BD&AA-related services (e.g. tools for model development) and, from the technology providers' perspective, **the responsibility for the**

models' use, performance and supervision is transferred to their customers (i.e. the institutions), with institutions being accountable when deploying and operating products and services.

Use of cloud services

Institutions are increasingly **relying on cloud service providers** to overcome issues with legacy systems. Heightened competitive pressure, changing customer behaviour and the speed of new technological releases force institutions to move faster, resulting in an increasing interest in the use of cloud outsourcing solutions in the banking industry.

Greater use of BD&AA is perceived to be facilitated by the use of cloud services, which promises high levels of availability, scalability and security.

Leveraging data sources

Institutions hold a significant number of data**,** generated internally based on customers' behaviour or acquired from external data providers or new data collection devices, with potential for mining new types of data (e.g. unstructured enterprise data).

To support BD&AA applications, some institutions are exploring the use of algorithms and ML models available from open-source libraries. The quality of the data to be used to feed the models is important: the slogan 'Garbage in, garbage out' means that applications are only as reliable as the quality and quantity of data used.

Human involvement and explainability

At this stage, the involvement of humans is required to make decisions based on advanced analytics-related techniques. In this new paradigm, solutions are no longer pure IT; new skills in data science are required and a gap has appeared between business and IT experts.

Institutions appear to recognise the importance of explainability and its possible consumer protection implications, and they seem to be working towards addressing these issues. Although no simple and unique approach appears to exist at this stage (academic research is ongoing), institutions seem to prefer the implementation of relatively **simple algorithms**, which tend to be more explainable, in an effort to avoid black box issues (e.g. a preference for decision trees and random forests rather than deep-learning techniques). The **modelling process** may be rather **iterative** to ensure a balance between explainability and accuracy.

Data protection and data sharing

Today, more than ever before, personal data protection brings new concerns to be addressed, from regulatory, institutional and customer perspectives. Both the General Data Protection Regulation (GDPR)[15] and the 'Principles for effective risk data aggregation and risk reporting' of the Basel

[15] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504&from=EN

Committee on Banking Supervision[16] have resulted in an increased focus on proper data governance frameworks and strategies to be put in place by institutions.

In particular, the GDPR principle of accountability (Article 5(2)) requires that institutions be able to ensure and demonstrate compliance with the principles relating to the processing of personal data. The GDPR can therefore be regarded as an opportunity to strengthen customers' trust in how institutions process their data.

Taking into account the international dimension, there is an increasing convergence on the adoption of data protection rules and principles closely mirroring the GDPR model. This trend towards convergence on privacy and data protection could attenuate the possible impact of the different legal frameworks of some non-EU countries.

The European Data Protection Supervisor and the European Data Protection Board are actively working, notably issuing opinions and guidelines, to ensure that institutions can optimise their activities relying on their data sources without undue impact on the interests and fundamental rights and freedoms of the persons concerned by the data processing.

In relation to the sharing of personal data with external parties, a mixed picture can be observed, as some institutions share their customer data, anonymised beforehand, with technology providers for model-training purposes. Some other institutions share their customer data with universities and public institutions only, and not actively with other commercial enterprises.

Also in this regard, the GDPR contains rules that allow the sharing of personal data, including transfer to institutions in third countries, subject to appropriate safeguards (Chapter V of the GDPR).

Bias detection

Bias is a strong concern that can hamper the accuracy and fairness of models. Some institutions address this issue at the model development stage by removing specific variables (i.e. sensitive attributes) and paying attention to the dataset used for training the model. Various statistical techniques are explored to help in detecting bias, while an iterative approach may help to gradually strengthen models against bias.

Software tools

Institutions frequently use **open-source frameworks** to implement BD&AA solutions. This covers programming languages, code versioning, and big data storage and management; the background seems more diverse for data analytics tools and data visualisation tools, where no dedicated tools appears to prevail and in-house solutions are used combined with ad hoc tools as needed.[17]

[16] 'Principles for effective risk data aggregation and risk reporting', January 2013, BCBS (https://www.bis.org/publ/bcbs239.pdf).

[17] A non-exhaustive list of tools mentioned by banks responding to the EBA's questionnaire is as follows: Python, R and Scala for programming language; R, Scikit-Learns, Pandas, Tensor flow and Keras functions for data science libraries; Git,

Moreover, it appears that it is not always the case that the aforementioned tools support the entire data science process that leads to a specific output in a **reproducible** way, as in some institutions only the source code is recoverable while in other institutions all relevant events are reproducible.

2.2 Current application areas of BD&AA

BD&AA in financial services may have multiple applications, reflecting data pervasiveness and advanced analytics adaptability. They may be used to improve existing services from efficiency, productivity and cost savings perspectives or even to create new services or products as business opportunities. Therefore, all functions across an institution may benefit from such applications. A number of general areas where BD&AA applications are already in use or being developed are presented below, along with an overview of selected BD&AA use cases.

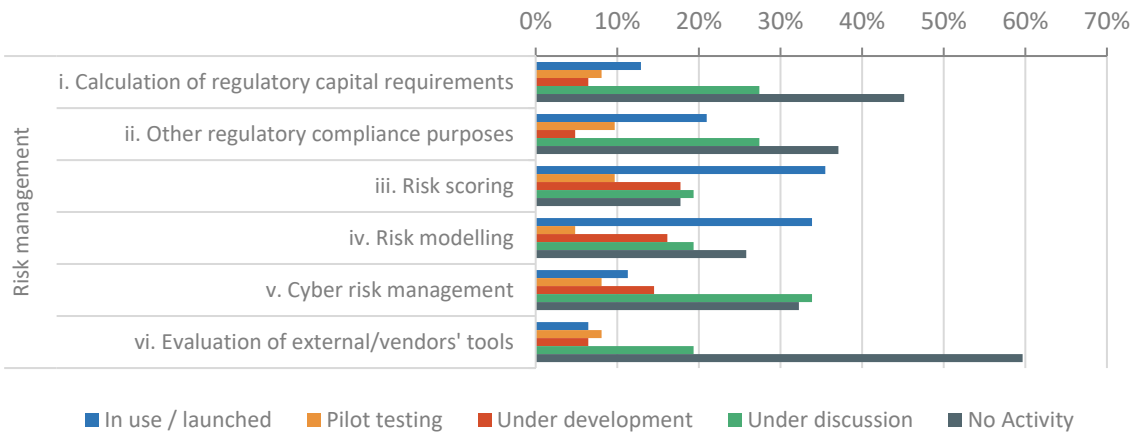
Risk mitigation

The use of BD&AA tools in the area of **risk mitigation** appears to be increasing. In particular, growing use was observed for risk-scoring and risk-modelling purposes (Figure 2.1). For example, credit scoring for primary customers may benefit from the use of advanced analytics models fed by the vast data held by institutions (sometimes combined with external data), while non-bank customers can be assessed taking advantage of application programming interface (API) access to payment data, bringing new services such as instant lending to non-bank customers or pre-approved loans for primary customers.

Institutions acknowledge the growing focus on **operational risks**, such as cyber-risk, and fraud and anti-money laundering/countering the financing of terrorism (AML/CFT) issues. Through the use of BD&AA techniques, institutions are exploring more efficient ways to save costs and ensure compliance. Figure 2.1 shows that the use of BD&AA for regulatory compliance purposes and cyber-risk management seems to have potential in the financial services sector. For example, such techniques are being used to detect fraud on payment transactions (especially in real time) and to detect high risk customers but also to streamline the whole **fraud detection** process. This relies on institutions' vast quantities of backward-looking data combined with external datasets and pattern detection techniques (examining customer behaviour) provided by ML algorithms.

Spider, PyCharm and R Studio for code versioning; Spark and Hadoop for big data storage and management; KNIME, H20 and Elastic/Kibana for data analytics; and R Shiny and JavaScript for data visualisation.

Figure 2.1: Current use of Big Data Analytics for risk management purposes



Source: EBA risk assessment questionnaire (spring 2019)

Moreover, similar processes, such as '**know your customer**' processes, can involve leveraging BD&AA techniques based on document and image processing backed with facial recognition, streamlining the whole onboarding processing and other document processing techniques.

Practical use case: automated credit scoring

Automated credit scoring is a tool that is typically used in the decision-making process when accepting or rejecting a loan application. Credit-scoring tools that use ML are designed to speed up lending decisions while potentially limiting incremental risk. Lenders have typically relied on credit scores to make decisions on lending to firms and retail clients. Data on transaction and payment histories from institutions historically served as the foundation for most credit-scoring models[18].

The most common ML methods used to assess credit applications are regression, decision trees and statistical analysis to generate a credit score using limited amounts of structured data. However, due to the increased availability of data, institutions are increasingly turning to additional data sources, unstructured and semi-structured, including on social media activity, mobile phone use and text message activity, to capture a more accurate view of creditworthiness.

BD&AA are also being applied to analyse large volumes of data in a short period of time. For example, analysis of the timely payment of utility bills enables access to new variables to calculate a credit score even for individuals who do not have enough credit history.

While there are benefits arising from the use of BD&AA to assess institutions' and individuals' creditworthiness, like any program or product, it is not without its risks. For example, a large retail institution could use powerful statistical models to model a borrower's repayment

[18] https://www.fsb.org/wp-content/uploads/P011117.pdf

behaviour. However, the institution's sales staff could game the system and coach uncreditworthy customers on how to be granted a loan. As a consequence, a substantial share of the institution's credit decisions could be based on dubious data. This could result in many borrowers falling into arrears, while the institution's data team struggle to work out how much of its information is reliable and suitable for informing future lending decisions.

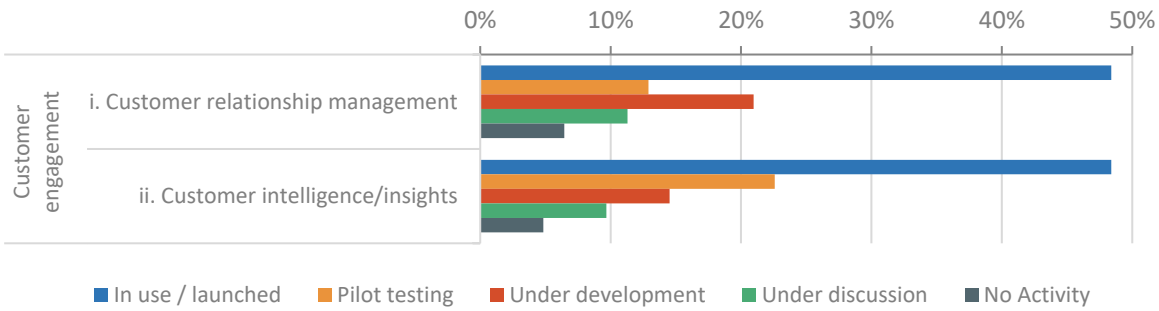
Another notable use case that has some potential is reliance on advanced analytics for the calculation of regulatory capital requirements (Figure 2.1). Possible outcomes might be optimal portfolio segmentation for model building or better performance and improved quantitative and qualitative parts of the models. From a prudential framework perspective, it is premature to consider ML an appropriate tool for determining capital requirements, taking into account the current limitations (e.g. 'black-box' issues). Sufficient testing would be needed on the efficiency of such models under various conditions, especially under changing economic conditions (e.g. in a downturn), to avoid overreliance on historical, and especially on most recent, data.

Customer interaction

Considerable use of BD&AA is observed in the area of customer engagement, with a focus on customer relationship management as well as improving customer intelligence and gaining better customer insights (Figure 2.2).

A customer's voice can be converted into text (through natural language processing (NLP)) for automated analytics: for example, phone calls transcribed and processed could be classified in terms of estimated customer satisfaction, allowing ad hoc relationship management and enabling an automatic assessment of the potential for upselling or cross-selling. Similar management could be applied to chat, email and chatbot channels.

Figure 2.2: Current use of Big Data Analytics for customer engagement purposes



Source: EBA risk assessment questionnaire (spring 2019)

For example, chatbots could assist institutions in handling increasing volumes of customer demands (including peaks), saving employee resources that could then be shifted from administrative tasks to added-value customer interaction tasks (including personal advice or support).

Processing inputs from users is a major source of data, embracing both structured and unstructured formats, from digital forms to paper documents to be processed for subscription and biometric data for facial recognition provided during digital onboarding.

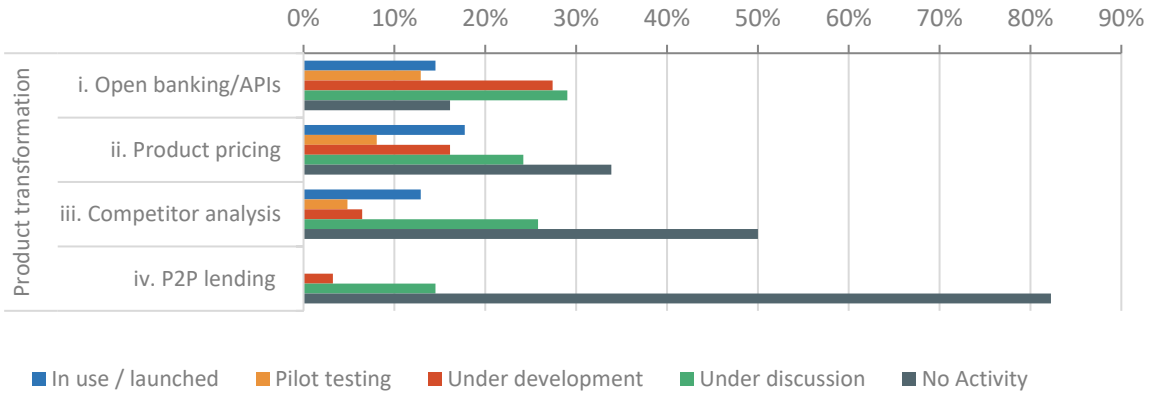
Market analysis

Customer insight is one of the cornerstones of institutions' marketing, used to develop new business or maintain existing business. It aims to improve **customer understanding** through better customer segmentation analysis backed with ad hoc models (e.g. an affordability model) that allow the use of advanced analytics, while **customer churn** or **customer behaviour** is reflected through dedicated analytics (e.g. propensity modelling[19]) fed by customer interaction data. This can help institutions to propose relevant and tailored financial services to their customers in a timely manner.

Combined with sales analysis (e.g. automated monitoring of sales or cross-selling modelling), product analysis (e.g. consumer loan pricing or cross-product analysis) and network marketing analysis (e.g. corporates interacting in a common business flow), customer insight can support institutions' market understanding.

Figure 2.3 shows that there is limited use of BD&AA for product transformation purposes, with some interest in competitor analysis and in the area of open banking.

Figure 2.3: Current use of Big Data Analytics for product transformation purposes



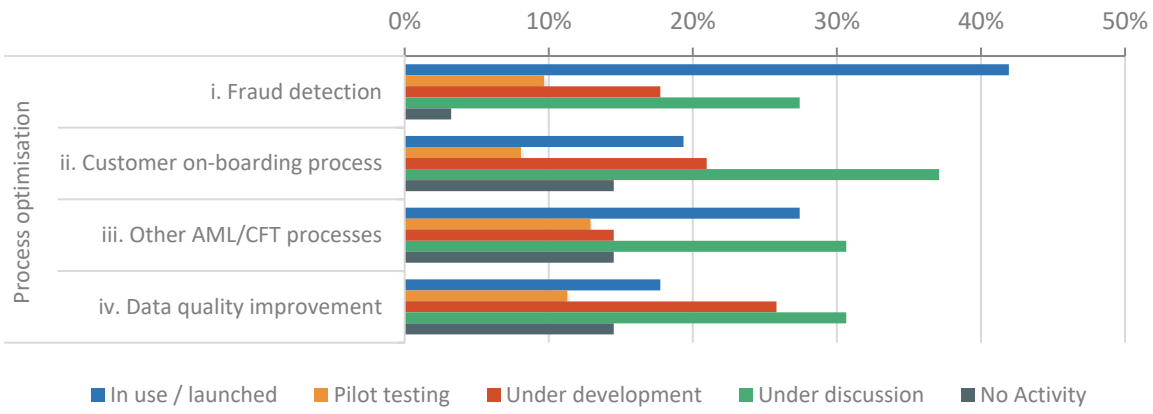
Source: EBA risk assessment questionnaire (spring 2019)

Back office automation

General automation of tasks performed across the functions of an institution leverages BD&AA solutions in an effort to **save costs or maintain staff resourcing levels during business growth**. robotic process automation techniques combined with AI allow the automation of such tasks.

[19] A propensity is an inclination or natural tendency to behave in a particular way.

Figure 2.4: Current use of Big Data Analytics for process optimisation purposes



Source: EBA risk assessment questionnaire (spring 2019)

These tasks could include generation of client conventions, incoming email classification and routing, control solutions, internal chatbots to help staff answer queries or cleaning of customer inputs to redact data.

BD&AA solutions are mostly used in optimising the process of fraud detection, as well as other AML/CFT processes. Institutions are also exploring the use of BD&AA to automate customer onboarding processes and improve data quality (Figure 2.4).

Practical use case: fraud detection

Fraud detection use cases vary according to the type of fraud targeted (internal fraud, payment fraud, identity fraud, loan fraud, etc.); however, the rationale is broadly the same.

The institution relies on a predictive model previously trained with backward-looking data on customers' behaviour cross-referenced with supplementary data, such as transactional data, for greater accuracy. Some extra features can be set up to enrich the model, such as rules that would highlight an obvious fraud pattern (e.g. a speed feature combining for one given credit card the timestamp and retailer location for successive payment transactions: the higher the value of the speed feature, the more likely it is that fraudulent copied credit cards are in use).

Predictive models may rely on supervised ML algorithms (fed by training data labelled as fraudulent or not) that can learn the fraudulent patterns based on past frauds and consequently detect a potential fraud case. Unsupervised machine algorithms, aiming to detect anomalies in behaviour (reflecting rare or unusual patterns), can also be used, in combination with predictive models, to ensure sufficient predictive capability and accuracy in the fraud detection process.

In operational processes, when it comes to detecting fraud, predictive models can be applied in real time with the purpose of preventing fraudulent transactions. As part of the business process,

the model receives as input the flow of business data to be checked and gives as a result a score assessing the potential for fraud for each entry in the flow.

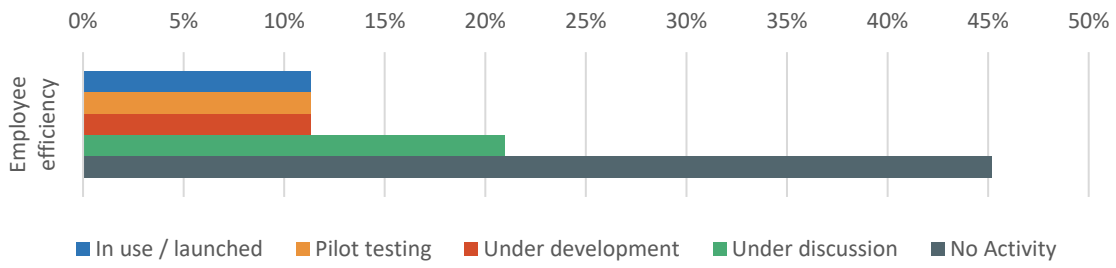
When the score given by the model for a particular entry reaches a predefined threshold, the entry is considered suspicious, i.e. potentially fraudulent.

An alert is then triggered and the entry (i.e. financial transaction) is quarantined until a compliance officer can manually check it. If the model is accurate, the compliance officer should have fewer cases to check and consequently be able to perform a more efficient assessment of the cases flagged as potentially fraudulent. The compliance officer makes a decision based on the explainable output provided by the predictive model and on the ad hoc investigation that he or she carries out.

To further improve the efficiency of the model, new recognised patterns resulting from the fraud detection process can be collected to retrain the model on a regular basis (a feedback loop).

However, the use of BD&AA for employee efficiency purposes is overall at a preliminary stage (Figure 2.5).

Figure 2.5: Current use of Big Data Analytics for employee efficiency purposes



Source: EBA risk assessment questionnaire (spring 2019)

3. Key pillars

This section introduces and illustrates four key pillars for the development, implementation and adoption of BD&AA, namely:

1. data management
2. technological infrastructure
3. organisation and governance
4. analytics methodology.

These pillars interact with each other, are not mutually exclusive and form the preconditions for the advanced analytics process described in Section 3.4.1.

3.1 Data management

Data management enables an institution to control and secure data used for enterprise purposes. To be able to manage data, one needs to know where the data are located, from where they are collected, the type and content of the data and who has access to them. In this section, the main aspects of data management including data types and data sources, data security and data protection, and data quality are introduced.

3.1.1 Data types and data sources

BD&AA applications rely on analysing large sets of different types of data. As more data have become available and the ability to store and analyse these data has increased, the need to consider the types of data being stored and analysed has become increasingly relevant. In this context, many different forms of data exist, including the following types.

Structured data: this refers to data that exists in a format that has been sorted or organised into standard fields and categories to give it a structure, for example data such as files and records in databases and spreadsheets that can be sorted and interrogated based on certain attributes.

Unstructured data: this data type has not been sorted or organised in a predetermined way and consists of a wide variety of data that are inherently difficult to search and make sense of. Gaining insights from unstructured data requires advanced analytics, skills and competence. Volumes of unstructured data significantly exceed those of structured data and these data are increasing rapidly, being generated from many disparate sources such as sensors, audio media, video media, GPS, CCTV and social media.

Semi-structured data: a type of data that contains semantic tags but does not conform to the structure associated with typical relational databases. Such data have some defining or consistent characteristics but may have different attributes. Examples include email and XML and other markup languages.

There is also a need to consider the data sources and the legal rights and obligations that derive from the data in question.

Whereas data types describe the inherent characteristics of the data and their basic formats, data sources specify the origins of the data used for BD&AA, which can be either internal data derived from the institution itself or external data collected or acquired from external entities.

Institutions predominantly collect and use internal data for their BD&AA models. The most commonly used internal data include customer transaction data, data on the use of other banking products (e.g. credit cards) and data on loan repayment behaviour. Although the use of external data by institutions is currently limited, it is worth noting that external data used by institutions include financial data (with prior customer consent), sociodemographic information about the customer, credit bureau data and public data on, for example, property values, negative news data affecting the institution and its clients, and the economic situation (e.g. the unemployment rate). In this context, it is important to validate external data before using it (see Section 3.1.3 and Section 4).

3.1.2 Data security and data protection

Information security is defined as the protection of information and information systems from unauthorised access, use, disclosure, disruption, modification or destruction in order to provide confidentiality, integrity and availability[20]. In the context of BD&AA, data security and model security are of particular importance as both are aspects of information security that are essential for the proper functioning of BD&AA algorithms. While data security focuses on protecting the confidentiality, integrity and availability of data, model security addresses attacks and corresponding protection measures specific to ML models (see Section 4 for more details).

To ensure data security, the protection needs of the data used for BD&AA first have to be identified and classified. Following this, appropriate safeguards for data security need to be defined and implemented. These safeguards must include appropriate **technical and organisational measures** to ensure a level of security appropriate to the risk[21]. These measures could be addressed as part of an overall information security management strategy or, alternatively, by establishing dedicated roles and a security management framework specifically for data security in relation to BD&AA, enabling the management body to develop a strategy and procedures to monitor, rapidly detect and respond to data security incidents relating to BD&AA systems, their data sources and third-party technology providers.

The importance of data protection in the context of BD&AA needs also to be reflected appropriately at the organisational and management levels of institutions. In particular, institutions need to comply with the GDPR throughout the entire lifecycle of a BD&AA application (e.g. the development

[20] https://csrc.nist.gov/glossary/term/information-security

[21] Article 32 of the GDPR ('Security of processing') provides that 'the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk'.

and production processes) when using personal data for training models or for other purposes during the steps in the BD&AA process.

3.1.3 Data quality

Data are one of the key strategic assets in banking because the decisions made depend on the data available and their veracity. Erroneous data poses a risk to an institution; therefore, data quality risks need to be identified and incorporated into an overall risk management framework.

The concept of data quality is overarching and needs to be considered at each step shown in the advanced analytics methodology presented in Figure 3.1. Like data security, data quality need to be considered throughout the whole BD&AA lifecycle. This is especially true with regard to the first two activities, data collection and data preparation. During data collection, cleaning and conversion, data quality aspects such as redundancy, inconsistency and incompleteness need to be addressed. Data that are doubtful or derived from unknown sources and ingested into analytics may result in erroneous outputs and consequently lead to wrong decisions.

The consideration of fundamental elements of data quality during the BD&AA process can help to gain trust in the data processed. A high level of governance is necessary to achieve data quality objectives. Although there are different approaches to categorising aspects of data quality, some of the most common categories (accuracy and integrity, timeliness, consistency and completeness) are presented in Annex I.

3.2 Technological infrastructure

The second pillar refers to the technology foundation in place for developing, implementing and adopting BD&AA. According to the US National Institute of Standards and Technology (NIST) Big Data reference architecture, the technology of BD&AA is based on three components, which are **infrastructure, data platform and processing**[22] (more details in Annex I).

The **infrastructure** component includes networking resources to transmit Big Data either into or out of the data centre, computing resources (e.g. physical processors and memory) for executing the software stack required for BD&AA and storage resources (e.g. storage area network and/or network-attached storage) to ensure the persistence of the data.

The **data platform** component manages all the data used by an advanced analytics system and provides the necessary API to enable data to be accessed. Finally, the **processing** component provides the necessary software to support the implementation of advanced analytics applications.

The processing component enables the processing of the data according to its volume and/or velocity (e.g. in batch or streaming mode), in support of advanced analytics applications (see also Section 3.4).

[22] NIST, *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*, June 2018, p. 16 (https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-6r1.pdf).

3.3 Organisation and governance

Another key pillar for the development, implementation and adoption of BD&AA is the establishment of appropriate internal governance structures and measures, as well as the development of sufficient skills and knowledge.

3.3.1 Internal governance structures and measures

Adaptable existing internal governance structures and/or the possibility of implementing new structures if necessary can support the development of BD&AA across the organisation as well as ensuring robust oversight of their use. This can be further supported as follows.

Governance structure, strategy and risk management: these require clear roles and responsibilities within the governance structure and an adequate understanding by the board of directors of the adoption and use of BD&AA, taking accountability for the related risks. The adoption and use of BD&AA need to be integrated into a risk management framework, alongside appropriate controls to mitigate risks and measures to ensure the responsible use and auditability of BD&AA applications. Moreover, in case of malfunctioning advanced analytics systems, fallback plans should be developed for core business processes to ensure continuity of operations and regulatory compliance.

Transparency: this means adherence to the fundamentals of explainability and interpretability (please refer to Section 4) to enable adequate risk management and internal audit, as well as effective supervision, supported by systematic documentation, sufficient justification and communication of important elements of BD&AA applications (e.g. with regard to material decisions, limitations of models and datasets adopted, circumstances of discontinuation, and model choices and decisions). The evaluation of model outputs can support understanding of the models and form part of a continuous effort to improve traceability and ensure that performance remains aligned with set objectives. Furthermore, a risk-based approach can be adopted in terms of the level of explainability, as requirements can be made more or less stringent depending on the impact of BD&AA applications (e.g. the potential impact on business continuity and/or potential harm to customers).

External development and outsourcing: the need to adhere to the EBA's *Guidelines on outsourcing arrangements* (EBA/GL/2019/02) applies to the use of externally developed and/or sourced BD&AA applications; institutions cannot outsource responsibility to external providers and thus they remain accountable for any decisions made. Moreover, adequate scrutiny of and due diligence on data obtained from external sources, in terms of quality, bias and ethical aspects, could be included in the risk management framework.

3.3.2 Skills and knowledge

BD&AA can be complex and difficult to understand, while their applications may not always function as intended and can result in risks for the institution, its customers and/or other relevant stakeholders. Staff across an institution may increasingly come to rely on BD&AA applications to

support them in their work, suggesting the need for sufficient training in the correct use of such applications to minimise errors and enhance opportunities. Furthermore, all staff and board members could have sufficient understanding of the strengths and limitations of BD&AA-enabled systems.

Level of understanding of management body and senior management: relevant and up-to-date competence and expertise on the part of the management body can help to ensure sufficient understanding of the risks associated with BD&AA applications for core business processes. Sufficient training for the managers responsible for the use of BD&AA applications could increase understanding of specific challenges, issues and risks. Furthermore, the management body is in a position to stimulate and facilitate the sharing of knowledge and experiences regarding the use of BD&AA.

Level of understanding of the second and third line: similarly, relevant and up-to-date knowledge and experience could be gained by second and third line of defence employees to ensure that they sufficiently understand the risks to core business processes and address the right challenges.

Level of understanding of developers of advanced analytics-enabled systems: data scientists could be trained to understand the impact of the input parameters used in BD&AA applications, their interconnectedness and other potentially relevant parameters (that are not included in the model). There is a need for business knowledge, given that they are trying to identify solutions to related problems. On the other hand, senior management need to be able to understand the explanations provided. As data cleansing, data visualisation and AI algorithm development require in-depth knowledge, it may be more effective to hire experts specialising in a specific field rather than staff having a basic knowledge of all aspects. Moreover, the establishment of a 'facilitator' (with sufficient understanding of both business and technology) to connect business and technology teams and the creation of a data science function to develop models using business knowledge could be other supportive measures.

Level of understanding of staff working day to day with advanced analytics-enabled systems: appropriate training (e.g. development of educational programmes) could be provided to staff working day to day with BD&AA applications to improve awareness regarding the responsible use of BD&AA applications and ensure sufficient understanding of their strengths and limitations.

3.4 Analytics methodology

The fourth pillar refers to the analytics methodology that is in place for the development, implementation and adoption of BD&AA. The process of building advanced analytics solutions generally follows the high-level methodology illustrated in Figure 3.1, which includes four main phases, namely data collection, data preparation, analytics and operations[23].

[23] Various frameworks exist that describe in more detail the phases of the analytics process or, more generally, the data science process. Relevant examples include the CRISP-DM and the NIST big data reference architecture.

Figure 3.1: Advanced analytics methodology

During the **data collection** phase, data are collected from internal data sources (e.g. databases, data warehouses, data lakes) and/or external sources (e.g. external data providers, the internet). This phase is followed by the **data preparation** phase, in which raw data are transformed to make them ready for analysis. Examples of data preparation tasks include data validation (e.g. checksums, format checks), cleansing (e.g. eliminating or fixing bad records or data fields from the dataset), scaling[24] and aggregation. During this phase, having good data quality and good data governance in place are clear success factors.

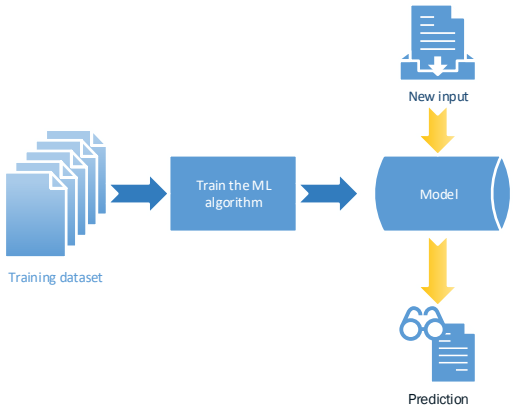
Subsequently, the **analytics** phase uses techniques, such as ML, to develop models that extract knowledge from data. More information about the use of ML in the advanced analytics process is provided in Section 3.4.1. Finally, the **operations** phase enables the end user or another system to request and access the results and insights gained by the model. Furthermore, this phase also includes monitoring and maintenance of the advanced analytics solution to ensure that results remain accurate over time.

3.4.1 Advanced analytics and machine learning

Advanced analytics often uses ML to gain deeper insights, make predictions or generate recommendations for business purposes. This is done by means of suitable ML **algorithms** able to recognise common patterns in large volumes of data via a **learning (or 'training')** process. The result of the learning process is a **model**, which represents what the algorithm has learnt from the training data and which can be used to make predictions based on new input data, as illustrated in Figure 3.2.

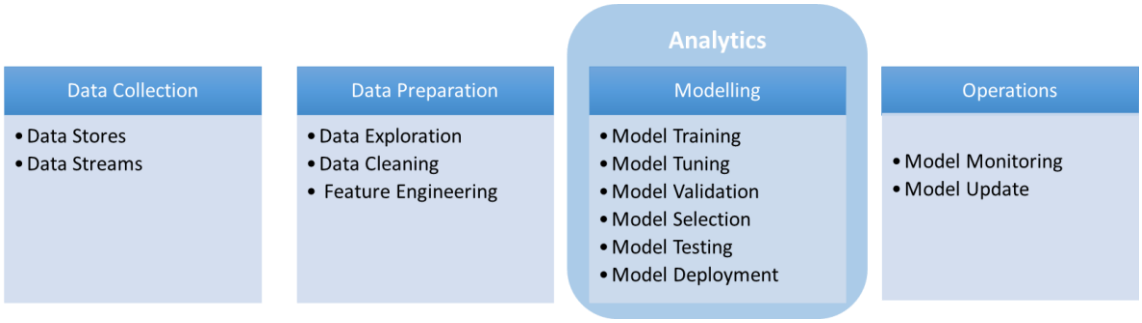
[24] Scaling consists in transforming the data so that they fit within specific scales, so that they can be compared (e.g. data in different currencies converted into data in one currency for comparison).

Figure 3.2: A model is a representation of what the algorithm has learnt from the training data and is used to make predictions based on new input data



The development of a ML model follows the generic analytics methodology described in Section 3.4, although some specificities apply, especially in the data preparation phase (which also includes feature engineering tasks), analytics (which basically consists in building the ML predictive model via modelling activities) and operations (which comprises the specific activities related to the monitoring and updating of the ML model), as illustrated in Figure 3.3. The specificities of the ML development process are presented in following subsections.

Figure 3.3: Advanced analytics process using machine learning



Data preparation

Feature engineering

In ML, a feature is an input variable that will be used by the model to make predictions. Feature engineering is the process consisting of transforming the input data into features that best represent the problem being predicted. Features are created via a sequence of data transformation steps (e.g. rescaling, discretisation, normalisation, data mapping, aggregation, ratios, etc.), usually involving some coding. Furthermore, dimensionality reduction techniques (e.g. principal component analysis[25]) can be applied to reduce the number of input variables.

[25] Principal component analysis is a statistical procedure that can be used as a dimension-reduction tool to reduce a large set of variables to a smaller set that still contains most of the information included in the larger set.

Given that the features can be the result of several data transformation steps, the link with the original raw data can be difficult to reconstruct. For this reason, carefully documenting the steps applied to generate the features is a success factor (this can be facilitated by the use of a data science development platform[26]).

Features can heavily affect (positively or negatively) the performance of the model. For example, a feature that contributes predominantly to the predictions made by the model may not be desirable, since the final prediction will strongly depend on the value of that variable only, instead of being linked to several features. This could lead to inaccurate or discriminatory results.

The importance of each feature as contributor to the prediction of the model can be measured via feature importance analysis, which may be useful also in helping to spot some cases of target leakage[27].

Analytics (modelling)

In the modelling phase, selected algorithms are trained in order to generate several candidate models and then the best for the particular business problem under analysis is selected.

A common technique used in ML modelling is to split the available data into three groups: **training data**, **validation data** and **test data**. The first dataset will be used to train the model, the second dataset will be used in the next step to validate the predictive capacity of the trained model and to tune it and, finally, the third dataset will be used in the testing phase for the final evaluation of the trained, fit and tuned model.

Model training

Model training (also called 'learning') consists in feeding the training dataset to the algorithm to build the model. The challenge of this phase is to build a model that fits the given dataset with sufficient accuracy and has a good generalisation capability on unseen data, i.e. a model that is **fit**. An ML model generalises well when its predictions on unseen observations are of a similar quality (accuracy) to those made on test data.

An **over fit** model is a model that has learnt too many details of the training data and therefore performs poorly when generalising and making predictions on new samples. This problem often arises when too many features have been selected as inputs for the model or, in the case of models based on decision trees, when the tree is too large and complex (in this case, overfitting may be reduced by removing those branches that are not relevant for the predictive performance of the model).

[26] Data science development platforms are integrated platforms covering all the steps of a data science/ML development project. Examples are DataRobot, DataIKU, Microsoft Azure Machine Learning Studio, Google Cloud Machine Learning, Amazon AWS SageMaker, etc.

[27] Target leakage occurs when the training dataset contains some information related to the variable being predicted (output/target variable) that normally is not available at the time of the prediction. Models with target leakage tend to be very accurate in development but perform poorly in production.

On the other hand, **under fitting** occurs when the model has not captured the underlying patterns in the data and is thus too generic for good predictions. This often happens when the model does not have enough relevant features.

It is difficult to achieve the right fit.

Model tuning

After the training phase, models are calibrated/tuned by adjusting their hyper-parameters. Examples of hyper-parameters are the depth of the tree in a decision tree algorithm, the number of trees in a random forest algorithm, the number of clusters *k* in a *k*-means algorithm, the number of layers in a neural network, etc. Selection of incorrect hyper-parameters can result in the failure of the model.

Model validation and selection

After the models are tuned, they are evaluated (or validated) against validation datasets, to check their prediction accuracy on data that is different from the dataset used for training.

The simplest model validation techniques use only one validation dataset. However, to build more robust models, a *k*-fold cross validation technique[28] can be used.

The model selection phase is tightly coupled with the model validation phase, since both activities are based on comparing various measures of performance (e.g. prediction accuracy). At the end of this process, the 'best' model is selected, taking into account other criteria such as computational performance, simplicity and explainability.

Model testing

In a separate testing phase, the chosen model can be checked again with new data not yet used (the testing dataset) for a final evaluation. In this phase, together with the business users, some final settings can be applied to tune the model (e.g. the set-up of the cut-off threshold in a classification problem, which defines the probability of an item falling into one class or the other and therefore the trade-off between false positives and false negatives).

Model deployment

The deployment of the model into production should be integrated into the standard change management process of the institution. As part of a standard change management process, integration tests are performed to validate the interaction of the model with the other parts of the system, such as application interfaces or databases, verifying that the end-to-end process will work appropriately in production.

[28] This technique consists of dividing the dataset into *k* subsets, each of which is used as the validation set while the other, *k* – 1 subsets are combined to form the training set. The results of the *k* validation tests are compared to select the highest performing and most robust model.

A model is integrated into a business process either in a fully automated way or with a 'human in the loop' involved in critical decisions. Sometimes, the process may be automated and the decision is delegated to a domain expert only if the decision is not clear enough (e.g. the model gives a credit score between *x* and *y*).

Optionally, the model may be deployed in production only after a parallel testing period during which the new candidate model (also called the '**challenger model**') runs in parallel with the old system to enable a comparison of the results.

The decision to implement the model in production is made after considering the overall business benefits and also based on other criteria such as explainability and fairness (see Section 4 for more details).

A model may run on premises, in the cloud or even on end users' devices, such as mobile phones.

Operations

Once in production, end users or other systems access the outputs of the model (e.g. via APIs or user interfaces). Keeping the model monitored and updated are success factors during this phase.

Model monitoring

After the selected model has passed the validation and test phases and all the components are integrated and tested, it is important to consider the overall business benefits and risks that the implementation of the model in production brings.

Having the model continuously monitored will help to promptly detect whether its performance is worsening or deviating from the expected behaviour (e.g. unintentional discrimination), thus making it possible to take appropriate remediation measures, such as selecting new features or retraining the model.

The model monitoring can be automated, via the implementation of 'early signals' or threshold warnings (e.g. accuracy under a specific value), or performed by domain experts who periodically analyse the outcomes of the system. Performance can also be monitored by comparing the outcomes of the ML model with those of the previous system running in parallel.

Model update

When monitoring identifies deviations from the expected performance or behaviour, it may be necessary to review the features used by the model. In addition, the model may be retrained periodically using new datasets to keep it up to date. The frequency of retraining will depend on the particular case. For example, the model should be retrained frequently when incoming data change rapidly. Details on software maintenance can be found in Annex I.

4. Elements of trust in BD&AA

4.1 Ethics

As mentioned in the *Ethics guidelines for trustworthy AI* from the European Commission's High-Level Expert Group on AI[29], the development, deployment and use of any AI solution should adhere to some fundamental ethical principles such as respect for human autonomy, prevention of harm, fairness and explainability.

These principles can be embedded from the start in any AI project, in a sort of 'ethical by design' approach. This also means that the business case made at the beginning of an AI project can include a high-level analysis of conformity to ethical principles and refuse unethical solutions. Having an ethical policy in place enforcing the above principles and setting the boundaries for acceptable and unacceptable use cases is recommended. Such a policy can apply also when the AI solution (or part of it, for example external data[30]) is purchased from a third-party vendor.

Furthermore, setting up an ethics committee (to validate new AI use cases, periodically review fairness metrics from live models, etc.) or integrating it into an existing similar committee is also recommended, especially when AI technology is widely used in an institution.

More details about how to embed ethical principles within the overall AI lifecycle are provided in the document cited above, *Ethics guidelines for trustworthy AI*, to which we refer the reader for more details.

Nevertheless, some key ethical aspects are described in the following sections, to provide all the information required to support our conclusions.

4.2 Explainability and interpretability

ML models can quickly become "black boxes", opaque systems for which the internal behaviour cannot be easily understood, and for which therefore it is not easy to understand (and verify) *how* a model has reached a certain conclusion or prediction. The opaqueness of a ML solution may vary depending on the complexity of the underlying model and learning mode. For example, neural networks tend to be more opaque, due to the intrinsic complexity of the underlying algorithm, than decision trees, the internal functioning of which can be more easily understood by humans. This **technical opaqueness** is directly linked to the opposing concept of **explainability**.

A model is **explainable** when it is possible to generate explanations that allow humans to understand (i) how a result is reached or (ii) on what grounds the result is based (similar to a justification).

[29] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[30] Particular attention should be paid when acquiring external data to ensure that it is of good quality and appropriate for the target context.

In the first case (i), the model is **interpretable**, since the internal behaviour (representing how the result is reached) can be directly understood by a human. To be directly understandable, the algorithm should therefore have a low level of complexity and the model should be relatively simple.

In the second case (ii), techniques exist to provide explanations (justifications) for the main factors that led to the output. For example, one of the simplest explanations consists in identifying the importance of each input variable (feature) in contributing to the result, which can be done via a feature ranking report (feature importance analysis).

There are other explainability techniques (this is a current research field that is constantly evolving), for example the following.

- (Global) surrogates, i.e. generating simplified models more interpretable than the black box model: an example of this technique are tree interpreters/decision tree approximations, which consist in approximating the model to a simpler tree highlighting the main decision steps (i.e. summarising the reasoning process).

- The LIME[31] technique (local interpretable model-agnostic explanations) generates simplified (interpretable) models that represent **local** explanations, i.e. that are valid only for the individual prediction. Such models cannot be used to provide **global** explanations, i.e. explanations that are generally valid for all predictions.

- Contrastive explanations[32] explain the behaviour of the model in the vicinity of the data point whose explanation is being generated, by generating two contrasting explanations, for example in the form of two opposite feature-ranking diagrams, showing the features that contributed most to that result (with their importance) and the features that least contributed to that result. Furthermore, several small modifications are applied to the value of each input feature to identify the threshold after which the prediction result would have changed. This is then depicted in a final diagram in the form of a 'confidence level' or in the form of a range of values (see Figure 5.1 below for an example). A common method of generating contrastive explanations is using Shapley values[33].

Finally, to improve the explainability of the entire advanced analytics process, institutions can develop specific tools to help simulate the output given certain input data (with a view on all intermediary steps).

The need for explainability is higher whenever decisions have a direct impact on customers/humans and depends on the particular context and the level of automation involved.

[31] Marco Tulio Ribeiro, 'Why should I trust you? Explaining the predictions of any classifier', 2016.

[32] https://arxiv.org/pdf/1802.07623.pdf; https://arxiv.org/pdf/1811.03163.pdf

[33] https://arxiv.org/pdf/1906.09293.pdf

Similarly, the need for explainability is also strong when a human is involved after the AI/ML model, is required to take the final decision based on the results produced by the model and therefore needs to understand why a particular result was generated. That explanation will be even more important when the decision impacting a consumer is taken fully automatically by the machine: in that case, regulations such as the GDPR have introduced the right of the data subject (i.e. the person who is the subject of the data being processed) to receive 'meaningful information about the logic involved' whenever there is 'automated decision-making, including profiling'[34].

In this context, it is important to note that the GDPR has also introduced, whenever there is an automatic decision, the right to human intervention in the process ('the right not to be subject to a decision based solely on automated processing'[35]).

Lack of explainability could represent an important risk in the case of AI/ML models developed by external third parties and then sold as opaque black box packages. The institution acquiring the package needs to have enough means, including explanations, to validate the results produced by the package without being strongly dependent on the external provider. In addition, black box products are more difficult to maintain and to integrate with other systems.

Explainability is just one element of **transparency**. Transparent systems could provide visibility with regard to not only the model (via explanations) but also the entire process used to build the model and the process in which the AI model is embedded when in production. *Transparency consists therefore in making data, features, algorithms and training methods available for external inspection and constitutes a basis for building trustworthy models*.

4.3 Fairness and avoidance of bias

Another important aspect of a trustworthy model is its fairness. ***Fairness*** *requires that the model ensures the protection of groups against (direct or indirect) discrimination*[36]. **Discrimination** can affect in particular smaller populations and vulnerable groups (e.g. discrimination based on age, disability, gender reassignment, marriage or civil partnership, pregnancy or maternity, race, religion or belief, sex, sexual orientation, etc.). To ensure fairness (non-discrimination), the model should be free from bias.

[34] Article 12, recital 58, Articles 13 and 14, and Article 22(1) of the GDPR. In this regard, please also see the European Data Protection Board's *Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679* (https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053).

See also Council of Europe, *The protection of individuals with regard to automatic processing of personal data in the context of profiling*, Recommendation CM/Rec(2010)13, and explanatory memorandum, 23 November 2010 (https://rm.coe.int/16807096c3).

In the context of the Organisation for Economic Co-operation and Development, see 'Recommendation of the Council on artificial intelligence', 22 May 2019 (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449).

[35] Article 22 of the GDPR.

[36] Discrimination (intentional or unintentional) occurs when a group of people (with particular shared characteristics) is more adversely affected by a decision (e.g. an output of an AI/ML model) than another group, in an inappropriate way.

Bias is 'an inclination of prejudice towards or against a person, object, or position'[37]. Bias can be introduced in many ways, including the following.

- It can be present in the (training/validation/test) input dataset. For instance, a common form of bias is human bias (when data are labelled according to a person's own view and therefore reflect the bias of that person).
- It can be introduced via the online learning process, when new, biased data are fed to the model in real time.
- Bias may also occur when ML models make non-linear connections between disparate data sources, despite those sources being validated individually for certain characteristics/variables.
- It can be introduced into the model during the development phase through inadvertent coding of biased rules, for example (algorithmic bias).

Most commonly, data contain bias when they are not representative of the population in question. This can lead to **discrimination**, for example when a class of people less represented in the training dataset receives less or more favourable outcomes simply because the system has learnt from only a few examples and is not able to generalise correctly. However, discrimination can exist without bias or direct discrimination; it can result from sensitive attributes serving as input variables, regardless of bias.

Techniques exist to **prevent** or **detect** bias (active or passive de-biasing). For example, controls can be implemented during the data preparation and feature engineering phases to prevent or detect bias and discrimination. Furthermore, statistical analysis (e.g. data skewness analysis) can be applied to the training dataset to verify that the different classes of the target population are equally represented (under-represented classes can be incremented by oversampling or over-represented classes can be reduced in size). In addition, techniques (and libraries) exist to test models against discriminatory behaviour (e.g. using crafted test datasets that could lead to discrimination).

Finally, a model running in a production environment can be regularly monitored to ensure that it has not deviated into discriminatory behaviour. In addition, it should be noted that having a diverse workforce (i.e. composed of a good balance of men and women and people from different backgrounds and with complementary skills) can also help to ensure the early detection of bias/discrimination issues and represents, therefore, a competitive advantage in building fair solutions.

It should be noted that bias detection and prevention techniques are a current field of research that is continuously evolving.

[37] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

4.4 Traceability and auditability (including versioning)

Traceability

All the steps and choices made throughout the entire data analytics process need to be clear, transparent and traceable to enable its oversight. This includes, inter alia, model changes, data traceability and decisions made by the model.

In addition, it is important to track and document carefully the criteria followed when using the model in a way that is easily understood (e.g. including a clear indication of when a model should be retired), the alternatives (e.g. model choices) and all the relevant information on each step throughout the process.

Moreover, institutions could keep a register of the evolution of the models. Having all the versions of a model registered enables an institution to compare different models or perform a roll-back if necessary.

The steps involved in a decision made by a model can be tracked from data gathering (including from third-party data sources) to the moment when the decision is made and even beyond, as, when a model is retired, institutions could still be able to explain how its results were produced and why its decisions were made.

To enable the repetition of the process by which a decision was made, the correct version of the model and data could be used. Sometimes, the model and data will need be recovered from repositories with previous versions of models and data.

Some institutions leverage an integrated platform to ensure the traceability of all the phases of a data science development process. These platforms usually include versioning features to keep track of the evolution of the model.

In summary, with traceable solutions, institutions are able to track all the steps, criteria and choices throughout the process, which enables the repetition of the processes resulting in the decisions made by the model and helps to ensure the auditability of the system.

Auditability

A traceable solution, for which there are detailed audit logs for all phases of the process that can be used to identify ‘who did what, when and why’, facilitates oversight of the system, as it makes it possible to follow the whole process and gain better insights.

Furthermore, when establishing an audit programme, institutions should focus not only on the model but also on governance aspects (including data governance).

4.5 Data protection and quality

Data are value, and as such should be adequately protected. This is particularly relevant when it comes to personal data.

Further to the general security measures described in Section 3.1.2, when managing personal data a trustworthy BD&AA system needs to comply with the current regulation on data protection (the GDPR). According to the GDPR, institutions should have a lawful basis for processing personal data,[38] an example being the customer's consent to processing the data for a specific purpose.

To re-use personal data for purposes other than those for which they were initially collected, the new purpose should be compatible with the original one;[39] otherwise, other grounds for lawfulness of the processing are required. Pursuant to the 'compatibility test'[40], *training a model with existing personal data may not be considered a compatible purpose and the data may need to be properly anonymised*.

Similarly, personal data may be shared with a third party only when there is a lawful basis (e.g. in case of outsourcing supported by customer-specific consent) and in compliance with other relevant GDPR requirements.

As a general principle, the customer should be informed about any data processing performed on his or her personal data[41] (the right to 'transparent information, communication and modalities for the exercise of the rights of the data subject').

In addition, customers have the right to demand human intervention and not be subject to a decision based solely on automated processes, including profiling, when the decision produces legal effects concerning him or her or significantly affects him or her[42]. Customers have the right to obtain an explanation of the automated decision and to challenge it.

When buying personal data from third parties, institutions should carry out due diligence to assess the quality of the data and that the data were collected in compliance with the relevant GDPR requirements. The institution (or 'data controller') is **accountable**[43] for, among other things, the **lawful, fair and transparent** processing of personal data and ensuring that personal data were collected for specified, explicit and legitimate purposes and not further processed in a manner incompatible with those purposes[44]. These principles also apply to personal data collected from publicly available social media, in relation to which it may be challenging to demonstrate

[38] Article 6 of the GDPR.

[39] Article 6(4) of the GDPR.

[40] Based on the criteria under Article 6(4) of the GDPR.

[41] Article 12 of the GDPR.

[42] Article 22 of the GDPR.

[43] Responsible for and able to demonstrate compliance with Article 5(2) of the GDPR.

[44] Article 5 of the GDPR.

compliance with the GDPR, in particular with regard to the need for a legal basis for processing and compliance with transparency obligations.

4.6 Security

Whenever there is a new technology trend, there are also new attack techniques exploiting security vulnerabilities, and AI does not escape this universal rule.

Some of the main types of attack affecting in particular ML include the following:

- model stealing
- poisoning attacks
- adversarial attacks (including evasion attacks).

Model stealing/extraction attacks are used to 'steal' models by replicating their internal functioning. This is done by simply probing the targeted model with a high number of prediction queries and using the response received (the prediction) to train another model. The cloned model can reach a high level of accuracy, even above 99.9%.

In **poisoning** attacks, attackers deliberately influence the training data to manipulate the results of a predictive model. This type of attack is especially valid if the model is exposed to the internet in online mode, i.e. the model is continuously updated by learning from new data.

An **adversarial** attack consists in providing a sample of input data that has been slightly perturbed to cause the model (in this case a classifier) to misclassify it. In most cases, such a small perturbation (basically representing a 'noise') can be so subtle that a human does not even notice it (e.g. when the problem concerns an image classification and the input image has been altered by a noise not noticeable to the human eye).

A particular example of an adversarial attack is an **evasion** attack, which consists in crafting the input to cause the model to avoid detection of a particular object/element.

Researchers have recently developed some new defence techniques to defeat these attacks, for example by adding a de-noiser before the input stage. Furthermore, open-source libraries are being developed that offer techniques to test the robustness of models against this kind of attack.

Currently, this kind of attack is not yet often seen in practice, partly because, for the most part, the models implemented by institutions are not directly connected to internet, reducing exposure to such attacks.

It is important to maintain a technical watch and be regularly updated about progress on security attacks and related defence techniques.

4.7 Consumer protection

A trustworthy BD&AA system should respect customers' rights and protect their interests. Customers sometimes accept the conditions for a service without reading them carefully, giving consent to abusive terms.

In addition to the personal data protection aspects discussed above, models may give rise to other consumer protection issues: they may make it possible to exploit customers' data patterns to maximise profit without considering customers' interests, leading to misconduct issues.

The use of alternative sources of data, such as behavioural data, can result in financial inclusion when customers gain access to financial services that they could not access before (e.g. due to a lack of financial information). However, some customers may be unfairly excluded from financial services if they do not share the data required or do not have that data at all (e.g. non-digital customers). In addition, those customers in higher risk categories may be excluded, since the principle of spreading risk among a larger set of users is no longer respected.

Some customers may be favoured over others if they learn how to behave so that the model makes a particular decision. In addition, some consumers may have concerns that their day-to-day behaviour limits the services they can access, since they feel monitored.

If consumers encounter problems due to an unsatisfactory service, they are entitled to file a complaint and receive a response in plain language that can be clearly understood[45]. Explainability, again, is key to addressing this obligation.

[45] https://eba.europa.eu/documents/10180/732334/JC+2014+43+-+Joint+Committee+-+Final+report+complaints-handling+guidelines.pdf/312b02a6-3346-4dff-a3c4-41c987484e75

5. Key observations, risks and opportunities

This section summarises the key findings from Sections 1-4, in an effort to clearly convey the key messages of this report.

5.1 Key observations

- Institutions are at different stages of BD&AA development, depending on their maturity in these technologies and the extent to which they integrate digital and data issues into their core strategies. Typical use cases are fraud detection, customer relationship management and back office process automation.

- Institutions mostly use internal data; the use of external data, including social media data (for customer insight purposes), is currently limited. For BD&AA development, some institutions use open-source solutions. At this stage, limited use of complex/sophisticated models by institutions was observed. The focus appears to be on the development of simpler algorithms, such as regression models, that can be traced and explained rather than deep-learning algorithms.

- In terms of internal organisation, some institutions were observed to explore the use of advanced analytics with small unstructured datasets within the ICT department, while others had clear business goals and action plans set by the management body, to be implemented by dedicated data science departments.

- To overcome issues with legacy systems, many institutions are increasingly relying on technology companies for the provision of both infrastructure and cloud services for BD&AA purposes.

5.2 Key opportunities

- The impact of BD&AA on the leisure and retail sectors has resulted in financial services customers increasingly expecting a more personalised service, with an increasing reliance on the use of non-cash payment services. These developments are actively changing the role of institutions and creating opportunities for the financial services sector. In this context, when it comes to the processing of personal data, compliance with the GDPR is not only a legal obligation but also a driver for increasing (or at least not losing) customers' trust in financial services.

- The opportunities of BD&AA lie in increased customer satisfaction, better customer insights to improve the offering for customers and reduce customer churn, optimisation of

processes leading to reduced costs, and new lines of business. In addition, BD&AA can also assist in the fine-tuning of risk mitigation and fraud reduction.

- Interpretable models may have many possible uses and present many opportunities in the short term, with documentation and responsible use of ML algorithms (algorithmic responsibility, system assurance, validation and verification) being possible support measures.

5.3 Key risks and proposed guidance

- BD&AA can be complex and difficult to understand. Their applications may not have deterministic behaviour and their outputs are correct according to a probabilistic measure; this may harm the institution itself, its customers and/or other relevant stakeholders.

- The implementation of a **governance and methodological framework** on BD&AA could promote its responsible use; this should include appropriate **documentation and sufficient justification** of material decisions based on BD&AA applications. Other explainability techniques also exist (as discussed in Section 4), as, depending on the specific algorithm used, the intrinsic level of explainability can vary. In addition, the limitations of models and datasets adopted should be documented and communicated, as should the circumstances under which the use of a particular model or dataset is to be discontinued. A model is explainable when it is possible to generate explanations that allow humans to understand (i) how a result is reached or (ii) on what grounds the result is based (similar to a justification).

- Explainability requirements can be applied following a **risk-based approach**, becoming more stringent as the significance of the model increases (e.g. the potential impact on business continuity and/or potential harm to customers).

- The integration of ML solutions with existing legacy systems may raise ICT risks, such as risks to data security and protection, data quality, change management, and business continuity and resilience.

- To support the development and functioning of models, a traceable solution (including model versioning) that assists in tracking all steps, criteria and choices through the entire process should be used. In this way, it should be possible to repeat and verify the decisions made by the model, helping to ensure the **auditability** and **traceability** of the system.

- A **human in the loop** (where necessary) should be involved in the decisions taken by a model periodically to assess whether the model is performing correctly, depending on the criticality of the decision and the possible impact on the consumer.

- To ensure that a model's outputs remain accurate over time and that the model is not deviating from its expected behaviour, model performance could be regularly monitored

(e.g. via the set-up of automatic alerts or via periodical expert reviews) and the model could be periodically updated.

- As most BD&AA solutions perform human tasks more efficiently, employees and management are likely to increasingly rely on BD&AA in the long term to support their work. Like other tools, BD&AA tools can result in potential misconduct when the organisation does not properly **train its employees** in all three lines of defence to use and understand BD&AA applications. From the front line to the boardroom, an institution runs a risk if employees do not have sufficient understanding of the strengths and limitations of the BD&AA-enabled systems they work with. There is currently a shortage of skills, as it is not easy to find human resources with all the necessary knowledge and experience (e.g. in data science, business, IT, statistics).

- Risks related to the use of BD&AA can arise from **reliance on third parties**. These include a lack of third-party knowledge and control, vendor lock-in, concentration risk, model maintenance, etc. **Accountability** for outsourced BD&AA systems remains with the institution and cannot be delegated (in line with the EBA's *Guidelines on outsourcing arrangements* (EBA/GL/2019/02)). Therefore, it can be a success factor if institutions that rely on technology providers and other third parties when using BD&AA include these risks in their risk management strategy. For example, lack of explainability can be a risk in the case of models developed by external third parties and then sold as opaque black box packages.

- The data analytics platforms of some institutions are based on **open-source frameworks**. Therefore, open-source frameworks are also crucial for some institutions in their provision of advanced analytics infrastructure, data platforms and processing software. A success factor for institutions could be being aware of the risks resulting from the use of open-source solutions and having measures in place to proactively mitigate them.

- As BD&AA applications take on tasks that previously required human intelligence, it is a success factor to ensure that the outputs of these systems do not violate institutions' **ethical standards** (e.g. ensuring that models are free from bias and model outputs are not discriminatory). This moral obligation could go above and beyond the fulfilment of applicable legal requirements. Institutions could ensure, for example by using a code of ethics, an ethics policy and/or an ethics committee, that their customers, as well as other stakeholders, can trust that they will not be mistreated or harmed – directly or indirectly – because of the firm's deployment of BD&AA.

- When using BD&AA, institutions can encounter **data-related risks** such as poor or unavailable data sources (e.g. where does the data come from?), data security (e.g. can the data be tampered with or leaked?), data protection (e.g. can the institution use personal data?), data quality (e.g. how accurate, timely, consistent and complete is the data?). When the data in the model are not of sufficient quality or are incorrect, there is a risk of outputs being biased or not being sufficient (noting, however, that bias is not only related to data

quality). Maintaining data quality is the basis for the responsible use of advanced analytics. Moreover, the application's behaviour depends significantly on the data used for training, testing and validation (as described in Section 3.4). A success factor for institutions is having in place a data management framework in which data quality is a key priority.

- For **data protection,** institutions must comply with the GDPR throughout the entire lifecycle of a BD&AA application (e.g. the development and production processes). This means that, when using personal data for training models or for other purposes during steps in the BD&AA process, institutions must adhere to the current regulation on personal data (the GDPR).

- **Data security** and **model security** will become increasingly important. A success factor will be that data security and model security for BD&AA are addressed appropriately at the organisational and management levels of institutions, for example within a dedicated unit or as a part of overall information security management within the institution. Appropriate safeguards for data security and model security need to be defined and implemented, and data security could be addressed throughout the entire lifecycle of a BD&AA application.

6. Conclusions

This report provides an overview of the use of BD&AA in the banking sector and has been prepared with the aim of sharing knowledge and increasing understanding of the development, implementation and adoption of BD&AA, noting the risks and challenges that could arise.

BD&AA provide the opportunity to repurpose data from their original intended use and to provide insights and find correlations between datasets that might not otherwise have been envisaged. New channels for collecting data resulting from technological developments mean that sources of data and methods of obtaining data (e.g. collected, derived, inferred and provided) are constantly evolving. This may lead to changes in the provision of some financial services, as well as posing risks to institutions, hence the need to pay attention to the **risk assessment** of new tools/solutions.

Currently, in the view of stakeholders, the development of BD&AA solutions is still at an early stage, with further adoption expected in the future. This may be due to institutions' approach to new technological solutions, affected by issues relating to their legacy systems, and to adequacy of skills, expertise and knowledge, as well as by data and security concerns.

Trust in BD&AA solutions is essential to allow the utmost benefit from the potential opportunities and at the same time ensure the proper, secure and responsible use of such solutions. A framework for responsible use and trustworthiness in BD&AA could be underpinned by a set of fundamental elements (listed in Section 4), namely ethics, explainability and interpretability, fairness and bias prevention/detection, transparency and auditability, data protection and data quality, customer aspects and security.

The effectiveness of human involvement depends on the level of understanding of the outputs of the models, making **explainability** a key feature for model accuracy and representativeness. Explainability is not a standard feature to be delivered, as the degree of explainability for ML models needs to be linked to the type of related function being performed. Ongoing research and development of tools and techniques may assist in addressing current issues with explainability and interpretability, as well as bias detection and prevention, which could possibly facilitate the responsible use of more sophisticated advanced analytics solutions.

Proper and effective **internal governance** frameworks and appropriate **organisational measures** need to be in place during the entire model development process, and investment in technical skills and knowledge are important to leverage the potential opportunities of data analytics. Understanding the quality of data available for use in advanced analytics is key, as the maintenance of **data quality** is the basis for the responsible use of advanced analytics and a key priority in an appropriate data governance framework. It is also important to have control of BD&AA tools provided by external parties and to have clarity on the responsibility of each participant (at all relevant levels of the institution). As data security and model security are crucial for the proper

functioning of BD&AA algorithms, adequate technical and organisational measures need to be considered.

The need for necessary **competence** will become increasingly important when the use of AI/ML techniques becomes more widespread in the financial services industry, raising an important challenge for institutions, supervisors and regulators. Training and development, as well as closer engagement between the relevant stakeholders, could be an appropriate starting point for addressing this challenge.

Adherence to the above elements of trust is expected, with the possibility of a **risk-based approach** towards certain aspects, such as explainability and interpretability, depending on the impact of each BD&AA application. For example, stricter requirements may apply when there is a potential impact on business continuity or potential harm to the customer.

The current trend and pace of change may soon raise the question of the need to develop AI/ML policies or regulatory frameworks for the application of AI/ML in an effort to facilitate its proper development, implementation and adoption within institutions. The EBA will continue monitoring these developments as part of its mandate on innovation monitoring.

Having in mind the four key pillars for the development, implementation and adoption of BD&AA, the existing regulatory framework is deemed sufficient (at this stage) in the areas of ICT, security and governance. Possible steps could focus in particular on **data management and ethical aspects**, as these appear to be the prevailing areas with a potential need for direction.

Annex I

Data quality (Section 3.1)

Data quality categories	Description
Accuracy and integrity	**The accuracy and integrity** of the data need to be inspected closely to detect errors, in particular when data are from external or less trusted sources but also when using internal data. The collection of relevant and high-quality data from the beginning could result in greater accuracy than that found in data that needs extensive cleaning. However, to gain a satisfactory level of quality, data usually needs to be run through at least some cleaning procedures.
Timeliness	Data collected can lose their validity over time. This is especially true for real-time data or data in high transactional environments. **Timeliness** is therefore a dimension that needs to be considered when data are used in models.
Consistency	Issues concerning the **consistency** of data can result from the use of heterogeneous data sources and legacy systems. The main challenge is to channel the data into one consistent data source, which serves as a single point of truth. Harmonisation and consolidation of the data involves combining different data sources so that they become comparable for the defined use cases.
Completeness	**Completeness**, from a technical perspective, means that a data field is filled with the data expected or defined by rules. For example, an empty field where one would expect a date of birth to be could lead to the conclusion that the field was either not set as being mandatory or that there is a general upstream issue in the data collection process that needs attention. Another issue is the use of incomplete datasets that are not useful for a specific use case. The outcome can have very limited value due to the limited dataset used.

Source: 'Data quality for data science, predictive analytics, and big data in supply chain management' Hazen, Benjamin T. & Boone, Christopher A. & Ezell, Jeremy D. & Jones-Farmer, L. Allison, 2014 and Basel Committee on Banking Supervision, *Principles for effective risk data aggregation and risk reporting*, Principles 3-6 (https://www.bis.org/publ/bcbs239.pdf)

Working towards meeting these quality requirements is not an easy task. In Big Data environments, where data are collected from multiple data sources on multiple hardware platforms and where each data source comes with its own data quality problems, this is even more challenging.

Multi-platform data environments are becoming reality in many organisations with the addition of on premises or cloud-based data lakes[46] and enterprise data hubs to organisational architectures that already include data warehouses and analytics platform servers. In a traditional data warehouse, the data ingested into the warehouse are supposed to be well prepared and quality assured. This is quite the opposite in data lakes, where the institution wants to retain the raw data in its native format, including structured, semi-structured and unstructured data. Based on the raw unfiltered nature of the data, it is possible to discover unexpected attributes and data relationships. When data are ingested into a data lake, one of the most common challenges for an institution is the ability to find, understand and trust the data actually needed. This is usually because the data are not understandable and can even be conflicting. Without establishing context for the data ingested, the institution could find its data lake becoming a data swamp.

The 'fit for purpose' principle could be the basis for the time required for the establishment of a data structure for the data ingested into a data lake. The typical step is to establish only the data structures needed based on the requirements of the advanced analytics project in question. In this way, how much work is invested in describing the data ingested into the data lake is restricted. The metadata for the data structures are fed into a data catalogue, with data items such as what information is available, why it is there, what it means, who owns it and who is using it.

Thorough cleaning and preparation of the data are required before it can be used as an input to an advanced analytics model. A number of tools and procedures are available to address various data quality problems (e.g. matching, profiling and enrichment).

Technological infrastructure (Section 3.2)

According to the NIST Big Data reference architecture, the underlying technology of advanced analytics is based on three components, which are **infrastructure, data platform and processing**[47].

Of the three, the processing component provides the technology on the application level to enable the advanced analytics methodology described in Section 3.4. Different types of processing can be applied to the three distinct processing phases of Big Data, namely data collection, analytics and access. In particular, the data collection phase can be supported by stream processing, as data may enter an advanced analytics system at high velocity. On the other hand, the analytics phase can be a batch process performed at a specified time and thus handled by batch processing, whereas interactive processing can be applied to retrieve data, which constitutes the access phase[48].

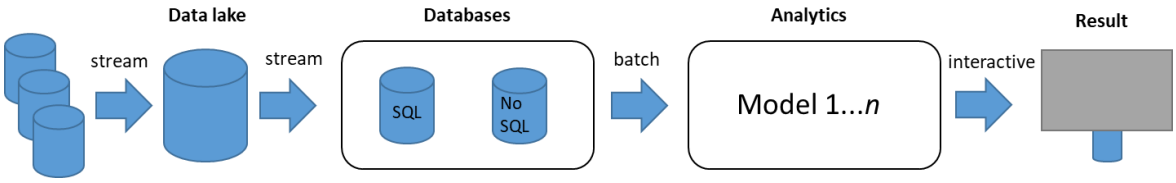
In Figure 8.1, a simplified illustration of an example of the technological infrastructure for BD&AA is outlined.

[46] A data lake is a storage repository that holds a vast amount of raw or refined data until it is accessed.

[47] NIST, *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*, June 2018, p. 16 (https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-6r1.pdf).

[48] NIST, *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*, June 2018, p. 44 (https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-6r1.pdf).

Figure 8.1: Technological infrastructure



In practice, institutions first ingest raw data into data lakes after collecting it from data sources. A data lake is a storage repository that holds a vast amount of raw or refined data until it is accessed[49].

The collection and ingestion of data can be supported by stream processing in real time, whereas the analytics step can be supported by batch processing.

It was observed that the Hadoop open-source platform is widely used to manage data lakes, while data warehouse platforms, proprietary or open source, are more often used to collect data from different sources (e.g. from subsidiaries). In addition to the example of technological infrastructure illustrated in Figure 8.1, it is also possible to bypass the data lake step. It was observed that some institutions do not use data lakes to store raw data for advanced analytics; instead, they ingest the collected data directly into databases, where they are then analysed by advanced analytics models.

Following this, data from the databases are analysed using an ML model. Finally, the user retrieves the results from the model, for example through interactive processing.

Software maintenance (Section 3.4)

The development of an ML model may lead to the creation of complex software systems that can be difficult to maintain. To build reliable and maintainable systems, a sound development approach needs to be adopted. ML developers should be aware of the specific issues that can affect the software development lifecycle in the context of ML, including the following.

- The many different ways that software platforms allow mixing of different data sources to feed ML models can impact system deployment, operations (especially when models are updated in real time in an online learning mode) and maintenance.
- Hidden dependencies between labels of a supervised learning solution can influence the predictions of the model in an underhand way and increase the cost of reviewing the system when these labels are tightly coupled in the code.
- Models designed as slight variations of existing models and implemented by using the existing model and simply learning a small variation generate cascade systems with strong dependencies, for which it is difficult to maintain the code (including bug fixing).

[49] NIST, *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*, June 2018, p. 44 (https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-6r1.pdf).

- Coding made up of many quick additions of experimental code paths in order to rapidly test the effect of the changes during model execution (due to the very iterative and experimental nature of the modelling phase) and then not removed when they are no longer needed, or 'glue code' written to customise general-purpose packages (often found in the market or in open-source libraries), can rapidly lead to 'spaghetti code' issues and increase the cost of documenting and correcting the software.

- The programming/development skills of the members of staff composing the data science team working on the ML project may vary, as may the quality of the code produced. This is because the skills required of data scientists encompass different fields (statistics, mathematics, IT) and is difficult to find experts in all domains.

Moreover, an ML system should be designed taking into account its non-deterministic nature and the need for a computational environment able to scale its performance. One key consideration in this context is capacity planning for the processing power required to efficiently execute the ML model during development and also when the model is deployed into production. For instance, when the model is deployed in an online learning mode (i.e. the learning is done continuously and the model is continuously updated with new data feeds), the model changes constantly, as does the computational power that it requires. Therefore, the allocation of capacity should be sufficiently elastic and scalable to quickly accommodate new needs in terms of the resource required, as in cloud or distributed computing, for example.

Annex II

Machine learning modes and data analytics solutions

As described in Section 1, each problem may be better addressed by a specific learning mode and some specific algorithms (examples of some widely used algorithms are presented in Table 9.1; note that a given algorithm may solve different problems).

Table 9.1: Links between ML modes, problems and algorithms

Learning mode	Problem	Algorithm (examples)
Supervised learning	Classification	Logistic regression
		Decision tree
		Naive Bayes classifier
		K-nearest neighbours
		Support vector machine
		Neural network
		Deep learning
		Hidden Markov model[50]
	Regression	Linear regression
		Non-linear regression
Unsupervised learning	Clustering	*K*-means
		DBSCAN
		Principal component analysis
		Hidden Markov model
	Anomaly detection	*K*-nearest neighbours
		Bayesian belief network
		Decision tree
		Support vector machine
	Association	Bayesian belief network
		Decision tree
		Neural networks
Reinforcement learning	Optimal action selection	*Q*-learning[51]
		SARSA[52] (state–action–reward–state–action)
		DQN[53] (double *Q*-learning)
		DDPG[54] (deep deterministic policy gradient)

[50] https://en.wikipedia.org/wiki/Hidden_Markov_model

[51] https://en.wikipedia.org/wiki/Q-learning

[52] https://en.wikipedia.org/wiki/State%E2%80%93action%E2%80%93reward%E2%80%93state%E2%80%93action

[53] https://en.wikipedia.org/wiki/Q-learning

[54] https://en.wikipedia.org/wiki/Reinforcement_learning

		PPO[55] (proximal policy optimisation)

In **supervised learning** the chosen algorithm 'learns' a classification (or regression) model on the basis of a large set of training data and, after the training, the model can be used to predict the class of new observations. Each class is identified by a label that can be either a nominal feature (e.g. a code) or a numerical feature (e.g. a continuous value). In the first case, the model selection concerns a **classification** problem and in the second a **regression** problem, which could be linear or non-linear.

Supervised learning allows the identification of general rules that exploit the knowledge learnt to process new input data. Known classification algorithms include logistic regression, decision tree, naive Bayes classifier, *k*-nearest neighbours, support vector machine, neural network and deep neural network.

Among data analytics solutions, **text analytics** encompasses a set of techniques and approaches able to extract specific information from textual contents[56] that are valuable data for predictions. For example, text analytics can be used for sentiment analysis, which is the prediction of the negative, neutral or positive sentiment of a text according to an accuracy measure. **Naive Bayes** is one of the most popular algorithms for text analytics. The naive Bayes algorithm uses a slight modification of the Bayes formula to determine the probability that certain words belong to text of a specific type. The 'naive' part of naive Bayes come from the fact the algorithm treats each word independently and a text is considered simply as a set of words. The wider context of the words is then lost using naive Bayes.

In ML, the selection of the right model depends also on the checks on the model assumptions. For instance, a linear regression can consider and predict only numerical values, while a decision tree will learn the best possible rules from well-labelled training data. For any classification problem, the prediction performance can be evaluated by a statistical measure, such as the ROC-AUC (receiver operating characteristic area under the curve) method[57], the confusion matrix or F1 score[58].

For example, a confusion matrix represents the performance of a predictive model that classifies cases into two categories. Actual cases are compared with predictions and different metrics can be calculated, as shown in Figure 9.1.

[55] https://en.wikipedia.org/wiki/Reinforcement_learning

[56] Text analysis is often referred as a partial NLP, because it intends not to fully understand the text's meaning but, rather, to retrieve high-quality data from it.

[57] ROC-AUC is a performance measurement for classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability.

[58] F1 score is a measure adopted for binary classification that represents the cut-off between precision and recall.

Figure 9.1: Confusion matrix example

Confusion matrix		Actual case	
		Class 1	Class 2
AI/ML prediction	Class 1	TP (Number of True Positives)	FP (Number of False Positives)
	Class 2	FN (Number of False Negatives)	TN (Number of True Negatives)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (Rate\ of\ \textbf{correct}\ predictions)$$

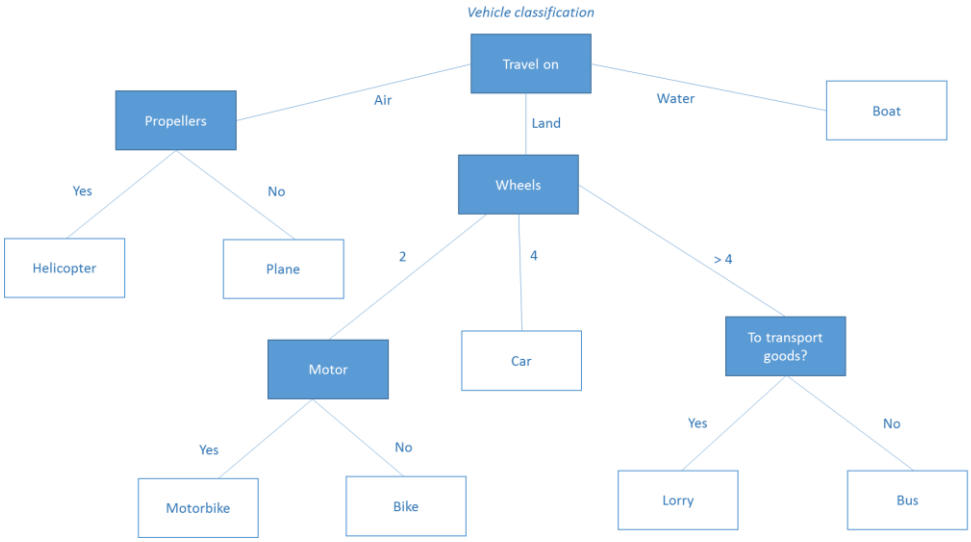
$$Precision = \frac{TP}{TP + FP} \quad (Rate\ of\ \textbf{positive predictions}\ that\ are\ \textbf{correct})$$

$$Sensitivity(recall) = \frac{TP}{TP + FN} \quad (Rate\ of\ \textbf{positive cases}\ that\ were\ \textbf{predicted}\ as\ \textbf{positive})$$

$$Specificity = \frac{TN}{TN + FP} \quad (Rate\ of\ \textbf{negative cases}\ that\ were\ \textbf{predicted}\ as\ \textbf{negative})$$

A **decision tree** algorithm splits observations along their dimensions (feature space) into smaller clusters decreasing the input entropy with regard to a given prediction function. Each cluster is then treated independently and is further split until a decision can be taken with an adequate level of entropy (not necessarily the optimal level). To maximise the efficiency of this process, the decision tree algorithm tests all features to find the most important contributor to the reduction in entropy by the largest amount in the class after the split. The entropy measure allows the selection of a specific decision tree, pruning the space of possible branches. Decision trees can also be used to explore which are the most important features for a particular dataset, information that can then be used in other algorithms. The decision tree allows the regularisation of input variables, making possible the selection of other ML algorithms that perform best on the reduced set of features.

Figure 9.2: Decision tree example



Random forests is a technique used to boost the efficiency of decision tree models by creating randomly a set (a forest) of slightly varying decision trees that model the same target. A cross-validation method makes it possible to select a target decision tree with reduced sensitivity to possible errors and noise.

Unsupervised learning refers to algorithms that implement models able to detect autonomously patterns in the data by identifying clusters of similar observations. Important problems addressed by unsupervised learning algorithms are clustering, anomaly detection and association.

- **Clustering** aims to group observations into clusters of similar data, i.e. two data in the same cluster would be as similar as possible to each other and as dissimilar as possible to data in other clusters. Relevant clustering algorithms are *k*-means, DBSCAN,[59] principal component analysis and hidden Markov model. Cluster identification is often used to discover unseen possible classifications of observations.

  K-means is a widely used clustering algorithm. In *k*-means, the algorithm starts assigning *k* cluster seeds randomly within the dataset. Data points are assigned to the nearest cluster seed, which is then moved to the average position of the data points. This process is carried out iteratively until stable clustering is achieved.

  The same methods mentioned for clustering can be used for model selection, in particular as cross-validation methods. Important algorithms are *k*-nearest neighbours, Bayesian belief network, decision tree and support vector machine.

- **Anomaly detection** (or outlier detection) is intended to detect data points that are not similar to most of the other observations in the dataset. Anomaly detection is related to clustering because observations that cannot be assigned to any one cluster are anomalies.

In **reinforcement learning**, rather than learning from a training dataset, the algorithm learns by interacting with the environment or from another algorithm that sends a feedback on the model output for a specific input.

Model selection approaches for all learning modes can be grouped into four categories: cross-validation, complexity criteria, regularisation and network pruning/growing.

- In **cross-validation** methods, many ML models are built using a different data distribution between training and test and measured on an independent validation set. The procedure is computationally demanding and sometimes requires additional data withheld from the entire dataset.

- In **complexity criterion-based** methods, many models are built using many training datasets and hence these methods are computationally demanding; although a validation set is not required, an information criterion to select the model according to a trade-off between complexity and accuracy is required (e.g. Akaike's prediction error criterion[60]).

[59] Density-based spatial clustering of applications with noise has been one of the most used clustering algorithms since the end of 1990. It groups data points on the basis of a density function that depends on the problem to be solved and the feature space.

[60] The Akaike information criterion can be expressed as a function with two components, one for measuring training error and the other for penalising complexity.

- **Regularisation** methods are more efficient than cross-validation techniques, but the results may be suboptimal.

- **Pruning/growing methods** can be considered to fall into the regularisation class; these methods often involve making restrictive assumptions, resulting in models that are suboptimal.

The aforementioned model selection methods can be used, with appropriate fine-tuning.

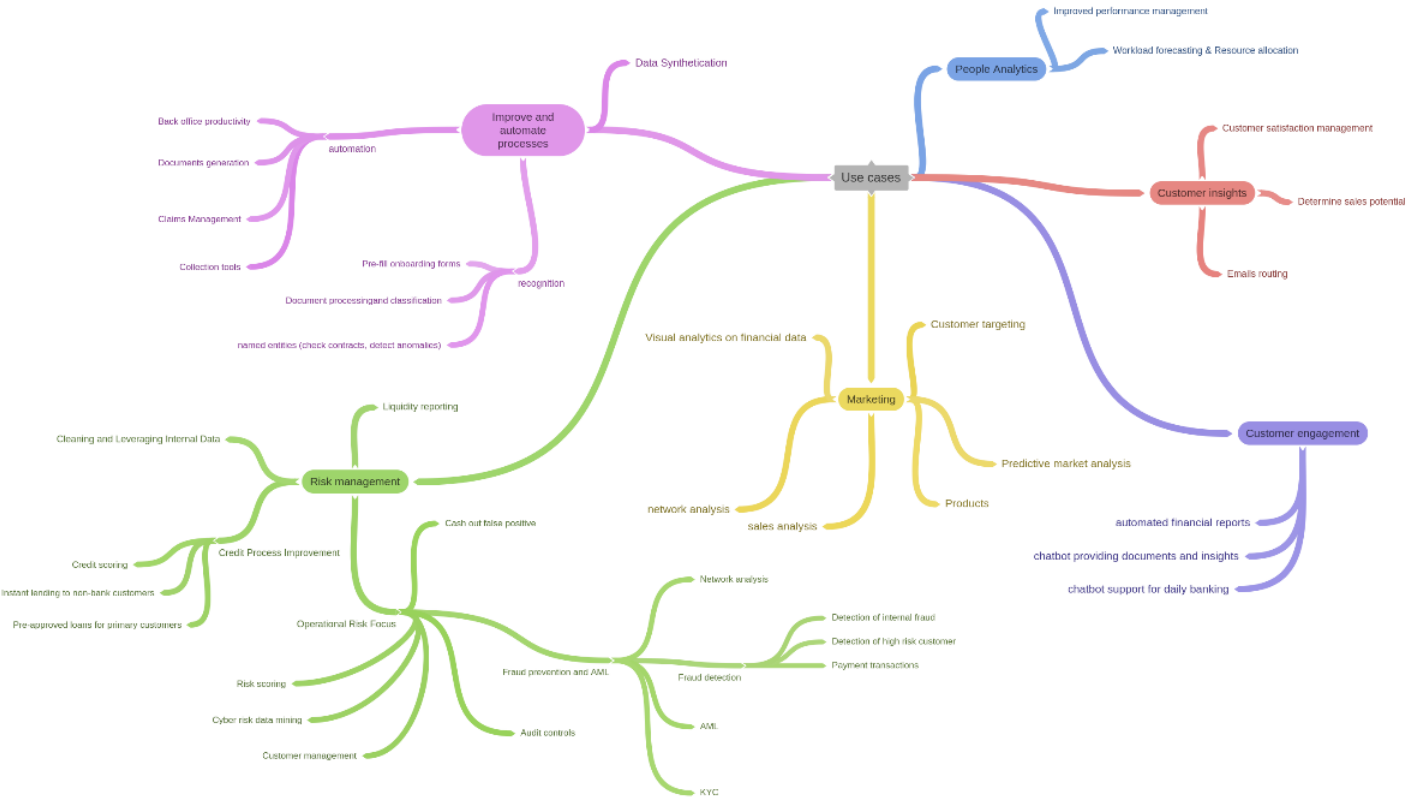
The generalisation error of an ML model can be estimated via either cross-validation or bootstrapping. **Cross-validation** and **bootstrapping** are both **resampling** methods. Resampling varies the training set numerous times based on one set of available data.

Cross-validation differs from bootstrapping in that bootstrapping resamples the available data at random with replacement, whereas cross-validation resamples the available data at random without replacement. Cross-validation methods do not evaluate the trained model on observations that appear in the training dataset, whereas bootstrapping methods typically do that. Cross-validation methods split the dataset such that a sample does not appear in more than one validation set.

Annex III

Current BD&AA applications heuristic map

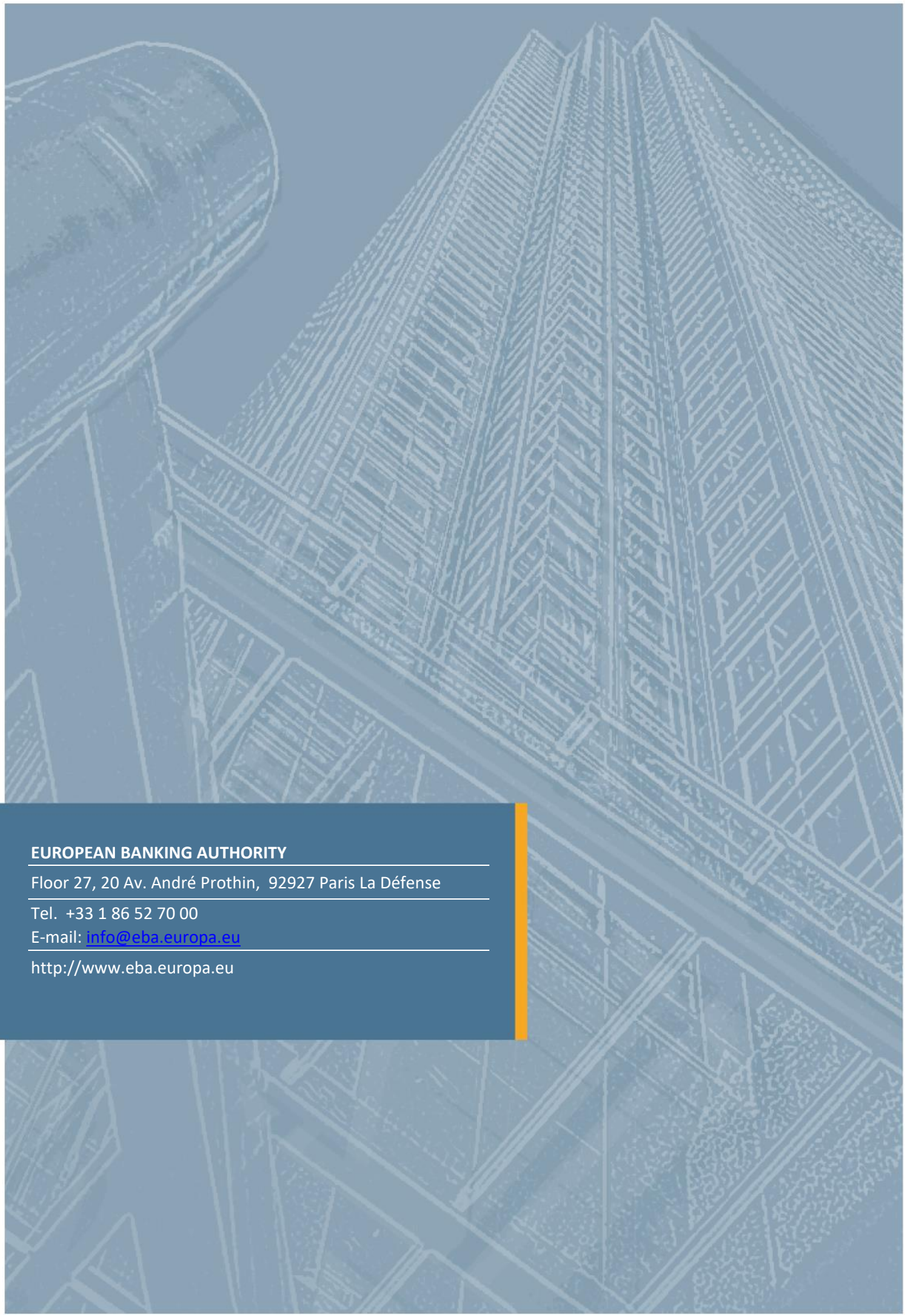
The following figure depicts a classification of the current main applications of BD&AA in the banking industry.



CONTRIBUTORS TO THE REPORT

The following contributed to the drafting of this report.

Co-Chair of the Task Force on IT Supervision	European Central Bank Banking Supervision – Single Supervisory Mechanism
Denmark	*Lars Brander Ilsøe Hougaard* Danish Financial Supervisory Authority
France	*Laurent Camus* Autorité de contrôle prudentiel et de résolution (Banque de France)
Germany	*Xu Zhu* *Jan Kiefer* Federal Financial Supervisory Authority (BaFin)
Greece	*Charalampos Paganos* Bank of Greece
Italy	*Giovanni Rumolo* Banca d'Italia
Luxembourg	*Anna Curridori* Commission de Surveillance du Secteur Financier
Netherlands	*Noor Witteveen* *Joost van der Burgt* De Nederlandsche Bank
Norway	*Jarleif Lødøen* Financial Supervisory Authority of Norway
Spain	*Carolina Toloba* Banco de España
United Kingdom	*Steven McWhirter* Financial Conduct Authority
Workstream lead/EBA Secretariat	*Andreas Papaetis* European Banking Authority



EUROPEAN BANKING AUTHORITY

Floor 27, 20 Av. André Prothin, 92927 Paris La Défense

Tel. +33 1 86 52 70 00

E-mail: info@eba.europa.eu

http://www.eba.europa.eu