REGTECH FOR REGULATORS ACCELERATOR

# Adopting SupTech for Anti-Money Laundering: A Diagnostic Toolkit

Rochelle Momberg and Louis de Koker

January 2020

Website **www.bfablogal.com/R2A**

Twitter **@BFAGlobal** | **@R2Accelerator**

# The RegTech for Regulators Accelerator (R$^2$A)

The RegTech for Regulators Accelerator (R$^2$A) partners with leading financial sector authorities to pioneer the next generation of tools and techniques for regulation, market supervision, and policy analysis. Accessing new datasets and analyzing available data more effectively allows financial authorities to establish a body of knowledge and evidence to drive smart policy reforms that promote financial inclusion and ensure financial stability, integrity, and consumer protection. R$^2$A accelerates these advances by helping authorities re-imagine how they collect and manage data, and by developing new solutions that strengthen their capabilities. Through R$^2$A, partner financial authorities seek to harness technology to improve the speed, quality, and comprehensiveness of information in support of targeted, risk-based decision-making. R$^2$A also engages closely with technology innovators to create structured opportunities for them to propose solutions and collaborate with financial authorities in the design and testing of promising ideas. Launched in October 2016, R$^2$A has already partnered with Bangko Sentral ng Pilipinas (BSP) and the Mexican Comisión Nacional Bancaria y de Valores (CNBV) to develop and test next-generation SupTech solutions. The project objective is to contribute to the creation of a global SupTech marketplace where robust solutions are available and demonstrated for key use cases. BFA Global is the managing partner and lead implementer project.

In August 2018 R$^2$A launched an AML working group with the aim to provide AML supervisors with a platform to discuss "*Solutions to Strengthen Financial Integrity and Combat Financial Crime*." The heterogeneous group of participants includes financial authorities from Uganda, Mexico, Peru, the Philippines, Australia, the United Kingdom, Singapore and the United States responsible for AML/CFT regulation and supervision from a variety of high-income, middle-income, and low-income countries.

R$^2$A is an initiative sponsored by Flourish and implemented by BFA Global.

# Acknowledgement

The authors received valuable support and contributions from the BFA Global R$^2$A team and from other stakeholders and partners in drafting this paper. We acknowledge in particular the overall leadership and the technical input provided by Simone di Castri (BFA Global and R$^2$A Director), Tania Churchill (Director, Enterprise Analytics, Australian Transaction Reports and Analysis Centre – AUSTRAC), and Arend Kulenkampff (BFA Global), who collaborated closely in the development of this toolkit.

We also acknowledge and thank our reviewers, including members of the R$^2$A Working Group and of La Trobe University, Australia (LTU), in alphabetical order: Pompeu Casanovas (La Trobe LawTech, La Trobe Law School, LTU); Cid Contreras Cesar (Mexican Comisión Nacional Bancaria y de Valores); Daswin da Silva (La Trobe Business School, LTU); Farida Saraid Paredes Falconi (Peruvian Superintendencia de Banca y Seguros); and Robin Newnham (Alliance for Financial Inclusion).

No endorsement of this paper was sought from any financial authority, nor should any be inferred from participation in its development or review. The authors and BFA Global are solely responsible for its content.

# Table of Contents

# Acronyms

| | |
|---|---|
| AI | Artificial Intelligence |
| AML | Anti-Money Laundering |
| API | Application Programming Interface |
| BI | Business Intelligence |
| CD | Compact Disc |
| CFT | Counter Financing Terrorism |
| CSV | Comma Separated Value |
| CTR | Cash Transaction Report |
| DB | Database |
| DBMS | Document-based management system |
| DNFP | Designated Non-Financial Businesses and Professions |
| DIS | Data Integration Strategy |
| EDW | Electronic Data Warehouse |
| ETL | Extract-Transform-Load |
| ELT | Extract-Load-Transform |
| ESB | Enterprise Service Bus |
| FATF | Financial Action Task Force |
| FI | Financial Institutions |
| FIU | Financial Intelligence Unit |
| FTP | File Transfer Protocol |
| GDBS | Graph database management system |
| GIS | Geographic Information System |
| HDFS | Hadoop Distributed File System |
| IT | Information Technology |
| JSON | JavaScript Object Notation |
| LEA | Law Enforcement Agency |
| ML | Machine Learning |
| MoU | Memorandum of Understanding |
| NLP | Natural Language Processing |
| NoSQL | Not only SQL |
| QA | Quality Assurance |
| RDMS | Relational database management system |
| RE | Reporting Entities |
| RPA | Robotic Process Automation |
| SFTP | Secure File Transfer Protocol |
| STR | Suspicious Transaction Report |
| SQL | Structured Query Language |
| VPN | Virtual Private Network |
| XBRL | Extensible Business Reporting Language |
| XML | Extensible Markup Language |

# Organization acronyms

| | |
|---|---|
| BoI | Bank of Italy |
| BSP | Bangko Sentral ng Pilipinas |
| CNBV | Mexican National Banking and Securities Commission |
| FCA | Financial Conduct Authority |
| DAG | Data Analytics Group |
| MAS | Monetary Authority of Singapore |
| NBR | National Bank of Rwanda |
| OeNB | Central Bank of the Republic of Austria |
| CBN | Central Bank of Nigeria |

# Introduction

The success of anti-money laundering and combating of financing of terrorism (AML/CFT) measures is deeply dependent on the capacity of financial intelligence units (FIUs) and AML/CFT supervisory authorities to collect and analyze comprehensive data accurately and speedily. The ability to do so, especially without creating undue cost burdens on the government, regulated entities, and customers, has been a challenge since the adoption and implementation of global AML/CFT standards. Technology now holds the promise to revolutionize the way in which FIUs and supervisory agencies can manage and analyze data, increasing the effectiveness and efficiency of AML/CFT while also strengthening data governance and data security.

The framework presented in this paper assists AML/CFT supervisory authorities in the analysis of their current data technologies and the identification of potential upgrades. It is presented as a first version to be improved and expanded, informed by user experiences.

The paper starts with a conceptual overview of supervisory technologies (SupTech) and related concepts, briefly setting out the techno logies available for implementation to realize a SupTech-supported financial authority. The question-based toolkit then enables a financial authority to consider its current state by examining the supervisory and intelligence lifecycles, enabling financial authorities to work towards their SupTech vision.

# Understanding SupTech

Several "tech" terms are often used in relation to financial and regulatory technologies. While these terms lack clear and consistent definitions, they are useful to support the discussion of these technologies:

"**RegTech**" is often used to refer to regulatory technologies that regulated entities and supervisors can use to respond to regulatory requirements and to create for example digital native regulation. Regulated entities and supervisors however require different, though complementary, RegTech solutions: Regulated entities would benefit from compliance technologies ("**Comptech**") while supervisors require technologies to support the collection and analysis of data relating to regulated entities to support their supervisory functions, including their enforcement functions. "**SupTech**" is used to label the latter. It refers to "the use of new technologies for internal supervisory purposes" as well as "the use of technologically enabled innovation by supervisory authorities",[1] and is often associated with the aim of assisting supervisory agencies in digitizing reporting and regulatory processes and help them comply with their mandate.

SupTech offers the potential to either radically improve existing supervisory tools or develop considerably better ones.[2] In addition, it holds the promise of enabling FIUs and supervisors to respond appropriately to the rising tide of data that underpins, drives and is generated by **Fintech** – the use of new technologies to provide financial services.[3]

[1] World Bank Group (2018), Discussion note: From Spreadsheets to Supervision; Basel Committee on Banking Supervision (BCBS) (2018), Sound Practices: implications of fintech developments for banks and bank supervisors; and Financial Stability Board (2017), Artificial intelligence and machine learning in financial services: Market developments and financial stability implications

[2] Dirk Broeders and Jermy Prenio (2018), Innovative technology in financial supervision (SupTech) - the experience of early users, Financial Stability Institute Insights No 9, Bank for International Settlements.

[3] Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018a), Financial Authorities in the Era of Data Abundance: RegTech for Regulators and SupTech Solutions.

While the Comptech and SupTech elements of RegTech can be distinguished, there are important touchpoints: Ideally Comptech solutions would not only support compliance management by regulated entities but would also support SupTech employed by FIUs and AML/CFT supervisors, for example by enabling faster and standardized compliance data collection by the government agencies. Given the sensitivity of the data concerned, they also share the need for appropriate data protection and data governance measures and for appropriate protection of individual rights, especially in relation to privacy.

## Technology stack

FIUs are increasingly using technology to facilitate the collection and analysis of AML/CFT-related financial intelligence. AML/CFT supervisors are similarly implementing innovative technologies to support effective and efficient risk-based supervision[4]. This, combined with the industry digitization trend, could enhance the risk-based approach to intelligence collection and supervision[5] in a number of ways, for example:

- Automated data collection solutions could ease compliance burdens of regulated institutions while increasing the speed of supervisory data collection;
- Centralized data warehouses for intelligence and supervisory reports can facilitate data mining; and
- Dynamic risk indicator dashboards and early warning systems can support intelligence and supervisory analysis and inform early and effective interventions.

When considering how technology may assist a specific agency it is helpful to understand key technologies and how they map to the core functions of the agency and may relate to each other. It is also important to understand the capability and limitation of each technology in its area of application.

Innovative technologies that can be considered within the Suptech AML context can be divided into two layers:

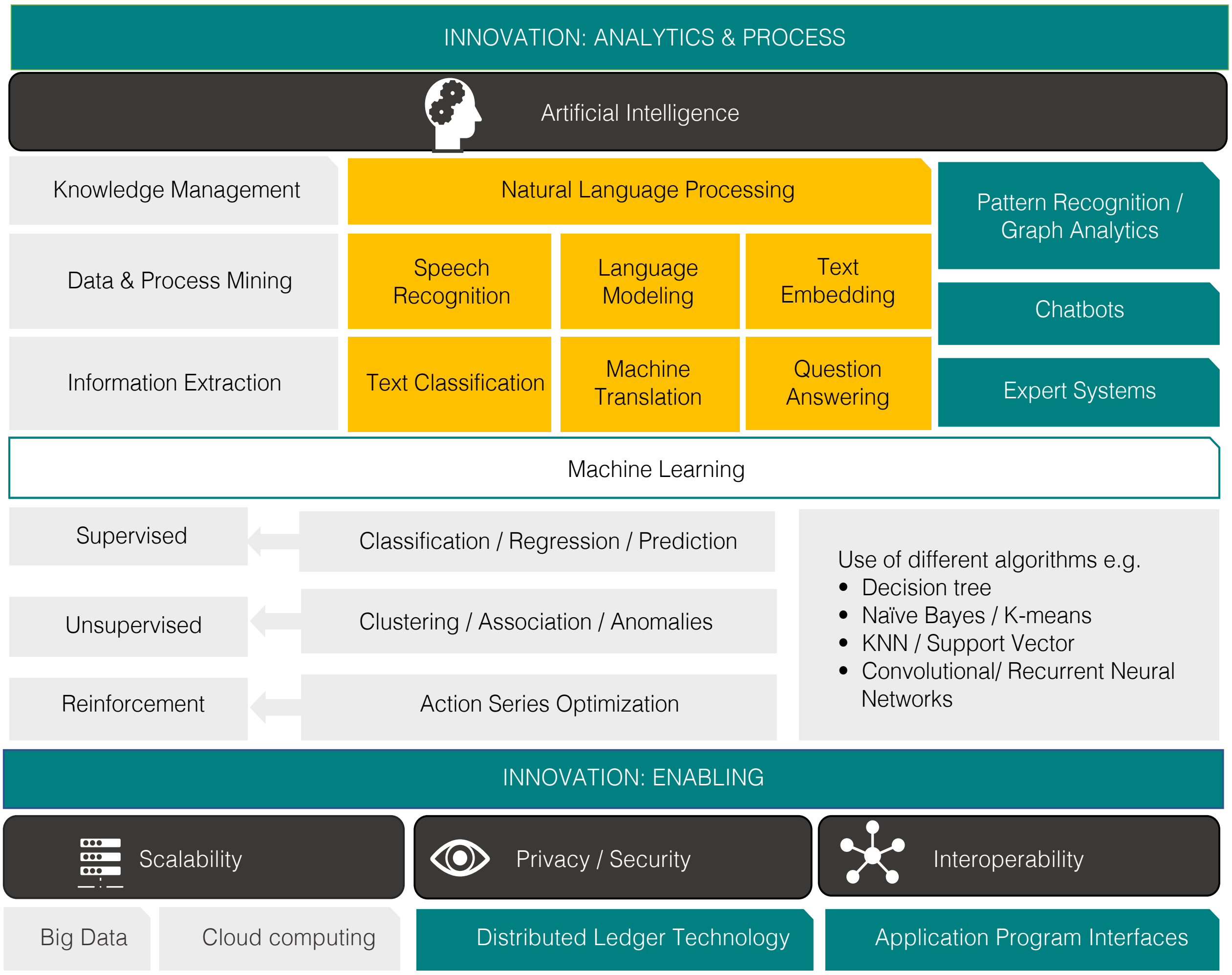
- Innovation relating to analytics and data management, and
- Innovation within the enabling environment.

The diagram below provides a visual representation of these two layers.

[4] Simone di Castri, Stefan Hohl, Arend Kulenkampff, and Jermy Prenio 2019, The suptech generations, Financial Stability Institute Insights on policy implementation No 19, Bank for International Settlements.

[5] Oliver Wyman (2018), Supervising Tomorrow.

*Figure 1: Emerging technologies for SupTech AML*



In the field of data science many of these new technologies are often clustered under the broad umbrella of "Artificial Intelligence" or "AI"[6], which refers broadly to technologies that enable machines, especially computer systems, to simulate human **intelligence** processes. These technologies in turn are powered by and associated with other technologies, for example those related to Big Data and enhanced data analytics.

[6] The term AI was coined in 1956 by Dr. John McCarthy at the Dartmouth Conference, however the groundwork for it was done by mathematicians such as al-Khwārizmī (who developed algebra and whose name informed the term "algorithm"), philosophers such as Leibniz, Hobbes and René Descartes and also computer scientists such as Alan Turing, who pioneering work on how and when computers could exhibit human-like or equivalent intelligence in 1950. See Hackernoon (2017), Robotic Process Automation and Artificial Intelligence.

Today technology is enabling a wide range of different AML/CFT analytics and data management applications, such as:

- VPN channels with strong encryption to securely send and receive data
- Robotic process automation for data validation and cleaning
- API-based data input and pull approaches for extracting regulatory data and reports directly from supervised entities
- Modern data warehouse & data lakes enabling advanced storage mechanisms and data analytics
- Cloud computing for scalable storage and flexible computing power
- Aggregating micro data, combining structured and unstructured data to deepen and broaden analytics
- Real-time data streaming enabling real-time monitoring
- Algorithms for quality and integrity rules built into a data model for automated alerts
- Chatbots[7] for enabling automated communication to provide advice (adequate suggestions) on simple regulation interpretation questions
- Automated alert engines built into a workflow process
- Natural Language Processing (NLP) to implement machine-readable regulations combined with chatbots to support regulatory compliance questions
- NLP and machine learning models applied to unstructured data to perform sentiment analysis
- Machine learning with application of deep learning techniques for classification of suspicious transaction reports
- Machine learning models detecting anomalies within real-time data feeds or identifying different levels of risks and instances of suspicious activity and underlying issues such as fraud
- Geographic information systems (GIS) to map data flows and cross-check geo-tagged data

## ENABLING TECHNOLOGIES

### Big Data

Refers to complex data sets that can be generated, analyzed and stored using a variety of digital tools, data elaboration techniques and information systems. "Big Data" is generally associated with three "V" characteristics namely: Volume, Velocity and Variety. Some experts add Value and Veracity to the list of characteristics. AI and machine learning rely on Big Data to refine decision-making processes and increase predictive power.

### Cloud Computing

Refers to the use of an online network ("cloud") of hosting processors to increase the scale and flexibility of computing capacity.

### Distributed Ledger Technology

These technologies enable nodes in a network to securely propose, validate and record state changes (or updates) to a synchronized ledger that distribute across the network's nodes. Use-cases include compliance management, digital identity solutions and cybersecurity.

### Application Program Interface (API)

The means by which a piece of computer software communicates with another.

*This toolbox provides definitions of key technologies in the Glossary but assumes that users will ensure that necessary technology knowledge is available to teams implementing this tool.*

[7] Deloitte (2017). Conversational Chatbots.

The application of such technologies to the AML/CFT use case requires change management and organizational change to embed the technology appropriately. Ideally agencies will also secure an appropriate in-house skillset across different roles such as data scientists, data analysts, data engineers and system developers to name a few.

**While implementation of comprehensive AI solutions still lies far in the future for many agencies, the use of some of the AI technologies that are available, may radically improve current AML/CFT processes.**

AI, in itself, is a developing field. Principles and standards guiding its development and to foster trust and promote ethics and responsible stewardship are still emerging. In 2019 the OECD contributed to this debate by setting out the following five complementary values-based principles for the responsible stewardship of trustworthy AI:[8]

- Inclusive growth, sustainable development and well-being;
- Human-centred values and fairness;
- Transparency and explainability;
- Robustness, security and safety; and
- Accountability

**Establishing an ethical framework for AI in society requires a comprehensive understanding of the opportunities and risks that the design and use of the technology presents. Ethical frameworks should be accompanied by appropriate governance frameworks that enhance the opportunities presented by technology while efficiently mitigating the risks[9]. Legal protection must be available for individuals and institutions whose rights may be infringed by such technologies, for example as result of biased algorithmic analysis, and appropriate remedies should be available where infringements occur. These frameworks should ideally not be developed by the state and simply enforced from the top or just left to the stakeholders to develop as a form of self-regulation. They are best designed combining elements of both these approaches in a so-called "middle-out" approach[10].**

[8] OECD Recommendation (2019).

[9] AI4People (2018), Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.

[10] AI4People (2019), On Good AI Governance: 14 Priority Actions, a S.M.A.R.T. Model of Governance, and a Regulatory Toolbox 's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, 16.

# AML/CFT supervisory authorities and SupTech

The AML/CFT framework is data intensive. Regulated entities are required to file reports on suspicious transactions and other reportable transactions (e.g. those involving more than a set amount of cash or cross-border transfers) to their national FIU. The FIUs in turn must process increasing volumes of data and are challenged to identify and respond to patterns in the data that point to money laundering or terrorist financing in order to make useful analysis and data available to their law enforcement and intelligence end-user agencies. Supervisory authorities in turn collect due diligence and compliance-related data and must analyze it to identity potential integrity and compliance vulnerabilities in regulated entities.

While AML/CFT intelligence and supervisory functions can be differentiated they are sometimes performed by one agency. In Australia, for example, AUSTRAC is both the national FIU and Australia's anti-money laundering and counter-terrorism financing regulator. Even in such cases, however, the intelligence and supervisory functions tend to operate separately as they have different objectives and requires different processes and skills. For purposes of this paper, the two functions are therefore distinguished and treated as if they are performed by different bodies. It should further be noted that the intelligence functions of FIUs are fairly homogenous globally as they all operate within the FATF standards framework. AML/CFT supervisory bodies on the other hand are more diverse as they supervise different institutions and industries and may have a range of additional non-AML/CFT functions. This paper, however, focuses on the generic AML/CFT functions of an AML/CFT supervisory body.

## FIUS AND DATA ANALYSIS

The international AML/CFT regulatory standards set by the Financial Action Task Force (FATF) require an FIU to add value by analyzing the information received and held by it. While all the information should be considered, an analysis may focus on each single report it receives or, depending on the type and volume received, on appropriately selected information.[11]

The standards explicitly encourage FIUs to use analytical software to process information more efficiently and assist in establishing relevant links, while supporting and retaining the human judgement element of analysis. Analytics and machine learning approaches look for anomalies or high-risk indicators, enabling analysts to more effectively target regulatory and supervisory risks. The following types of analysis should be conducted:

- **Operational analysis:** Using available information as well as other information that can be obtained (for example information to identify specific targets such as persons, assets, criminal networks and associations), to follow the trail of particular activities or transactions, and to determine links between targets and possible proceeds of crime, money laundering, predicate offences or terrorist financing.
- **Strategic analysis**: Using available and obtainable information, including data that may be provided by other competent agencies, to identify money laundering and terrorist financing-related trends and patterns. Such analysis can assist agencies to determine money laundering and terrorist financing-related threats and vulnerabilities. This in turn would support policymaking and risk-based supervisory decisions.

[11] FATF Recommendations INR 29 par 3

Importantly FIUs should have the ability to disseminate, spontaneously and upon request, information and analytical results to government agencies and counterparts[12]. Given the sensitivity of the data and information, all information received, processed, held or disseminated by the FIU must be securely protected, exchanged and used only in accordance with agreed procedures, policies and applicable laws and regulations.[13]

The technological capabilities of FIUs differ. Some developed country FIUs, for example, have in-house technical expertise to develop and maintain their technological capabilities. A number of FIUs use goAML solutions offered by UNODC. goAML is an integrated database and intelligent analysis system offering solutions data collection, analysis and data exchange solutions.[14] The technological capacity of many other FIUs, especially smaller FIUs, is very basic.

The Egmont Group of FIUs developed the FIU Information System Maturity Model (FISMM), to enable FIUs of different sizes to assess the maturity level of their processes and IT systems.[15] This framework is largely focused on financial intelligence rather than supervisory processes and functions.

## SUPERVISORY BODIES AND DATA ANALYSIS

Supervisory agencies require access to information and the ability to analyzes it effectively for a range of AML/CFT functions, including risk-based supervisory functions. In terms of the FATF standards, some of the key functions include:

- Licensing or registration of regulated entities.[16]
- Preventing criminals or their associates from holding, or being the beneficial owner of, a significant or controlling interest, or holding a management function in, a financial institution.[17]
- Assessing the exposure of regulated institutions and the industry to money laundering and terrorist financing risks.
- Issuing AML/CFT regulations that respond to national and industry-related money laundering and terrorist financing risks.
- Issuing guidelines regarding identified money laundering and terrorist financing risks and appropriate risk mitigation measures.
- Supervising or monitoring, and ensuring compliance by, regulated entities with AML/CFT rules.
- Identifying regulatory breaches and applying remedial actions.

To perform such functions they need to form a clear understanding of the money laundering and terrorist financing risks present in a country.[18] They also require on-site and off-site access to all relevant information on the specific domestic and international risks associated with customers, products and services of the regulated entities, including the quality of the compliance function of the institutions.[19] These functions can all benefit from appropriate technology support.

[12] FATF Recommendations INR 29 par 3

[13] FATF Recommendations INR 29 par 7

[14] United Nations, About goAML.

[15] UNODC (2017), UNODC's Software for Financial Intelligence Unit.

[16] FATF Recommendations INR 26.

[17] FATF Recommendations INR 26.

[18] FATF Recommendations INR 29 par 3.

[19] FATF Recommendations INR 29 par 3.

## DATA PIPELINE CHALLENGES

FIUs and supervisory agencies face several data pipelines challenges.

FIUs can be thought of as a data broker, connecting the creators of the data (regulated institutions) with the end users of the data (law enforcement and partner agencies) while adding value to the data for its end users.

While the international standards and expectations of FIUs and AML/CFT supervisors are clear, many regulated entities are challenged to perform key tasks effectively and efficiently. Current data management challenges of these entities result in time-consuming and expensive processes that fail to meet key objectives, often resulting in over reporting. Some FIUs may therefore be drowning in data. It has been estimated that typically, 80-90% of suspicious transaction reports filed with FIUs are not providing operational value to active law enforcement investigations.[20] This is often due to a combination of factors, including the inability to link risk, money flows and persons of interest across the growing data holding of regulated entities as well as FIUs. Costly systems are being maintained by regulated entities and FIUs to file, receive and manage these reports and this investment is wasted if the system does not provide appropriate support to law enforcement and national intelligence agencies. More importantly though, the protection that these measures may afford society against crime and terrorism is diminished.

Supervisory agencies face several challenges across the data life cycle such as data requirement clarity, infrequent data collection, ineffective data validation and not having adequate tools to manage large datasets. Supervisory data is often collected through multiple channels and methods which may not all be equally secure. These challenges become compounded by delayed submission of data and incomplete data, which results in a lack of quality data that has a knock-on effect on the quality of data analysis. Different stages of mature technologies[21] within institutions and agencies affect their ability to adapt to changes in data requirements. These challenges often result in missed opportunities in monitoring key risk metrics and identifying early signs of misconduct.

Specific challenges also exist in terms of data storage. Data lakes or warehouse can be created that can facilitate raw and unstructured data to optimally integrate with modeled and structured data for effective analysis. Enhanced security measures are required to protect the valuable pool of data from unauthorized access and attacks. Alternatively, data may remain in the different databases but mined using federated data mining solutions that enable interrogation of the data in different silos. Solely utilizing in-house storage systems without consideration of cloud storage or hybrid systems can limit the ability to manage the increasing amount data collected, especially where the data is combined with unstructured data such as CCTV camera footage.

**When data is stored in a cloud environment, it may be subject to the laws of other countries.[22] Cloud storage therefore raises data sovereignty questions.**

[20] Nick Maxwell and David Artingstall (2017), The Role of Financial Information-Sharing Partnerships in the Disruption of Crime, Royal United Services Institute for Defence and Security Studies. London. 5.

[21] Defined in Glossary.

[22] AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.

The questions arising in relation to data management by both FIUs and supervisory agencies emphasize the need for multi-disciplinary technology teams with appropriate skills and capacity to support agencies to meet their needs in the most effective and efficient manner. Current teams often reflect skills that may be limited to basic Structured Query Language (SQL), data modelling and business intelligence tools and their task is often complicated by their limited size of the team.

Data sharing similarly reflect current channels and methods which are not deemed secure, such as external sharing using email or unencrypted USB memory sticks. The implementation of systems like FIU.net, used within the European Union for information sharing between FIUs, should be more widely considered. And expanding the system's ability to include sharing with external parties across a secure environment needs to be developed.

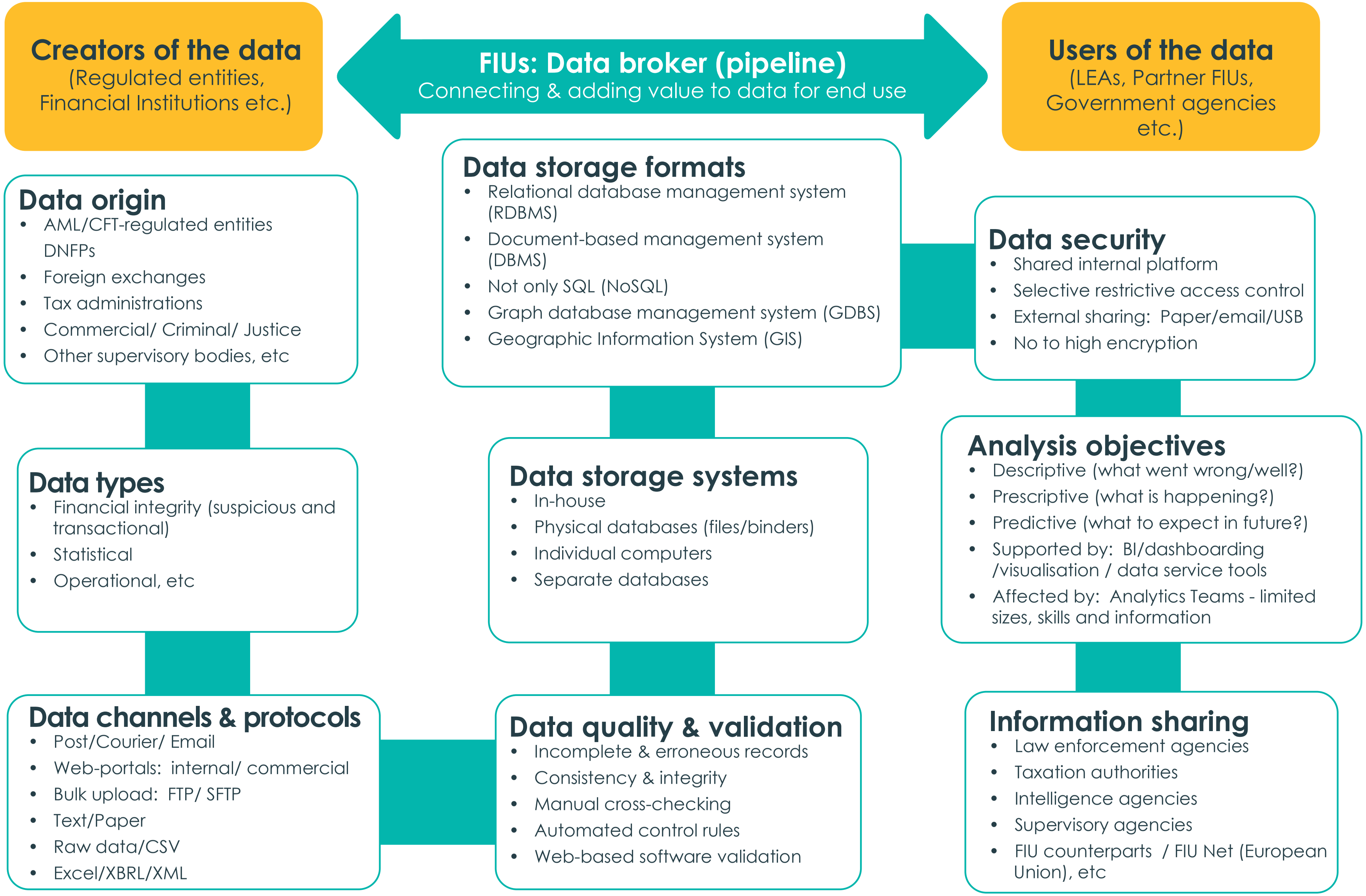
Data pipelines do not exist in isolation and should be viewed in their organizational and socio-political contexts. Given the sensitivity both of the data and the objectives of the channels they should be operate within appropriate ethical and governance frameworks that enhance their usefulness while mitigating attendant risks, for example relating to data protection and privacy.[23] Legal protection must be available for individuals and institutions whose rights may be infringed by such technologies and appropriate remedies should be available where infringements occur.

A diagrammatic view of a data intelligence pipeline is provided below, providing an overview of current situational analysis across supervisory institutions drawn from interviews and the BFA Global survey: The State of RegTech.[24]

[23] AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.

[24] BFA Global (2018), The State of RegTech: The Rising Demand for "Superpowers", BFA Global RegTech for Regulators Accelerator (R$^2$A) survey.

*Figure 2: Intelligence data pipeline*



**Creators of the data**
(Regulated entities, Financial Institutions etc.)

**FIUs: Data broker (pipeline)**
Connecting & adding value to data for end use

**Users of the data**
(LEAs, Partner FIUs, Government agencies etc.)

**Data origin**
- AML/CFT-regulated entities
  DNFPs
- Foreign exchanges
- Tax administrations
- Commercial/ Criminal/ Justice
- Other supervisory bodies, etc

**Data types**
- Financial integrity (suspicious and transactional)
- Statistical
- Operational, etc

**Data channels & protocols**
- Post/Courier/ Email
- Web-portals: internal/ commercial
- Bulk upload: FTP/ SFTP
- Text/Paper
- Raw data/CSV
- Excel/XBRL/XML

**Data quality & validation**
- Incomplete & erroneous records
- Consistency & integrity
- Manual cross-checking
- Automated control rules
- Web-based software validation

**Data storage systems**
- In-house
- Physical databases (files/binders)
- Individual computers
- Separate databases

**Data storage formats**
- Relational database management system (RDBMS)
- Document-based management system (DBMS)
- Not only SQL (NoSQL)
- Graph database management system (GDBS)
- Geographic Information System (GIS)

**Data security**
- Shared internal platform
- Selective restrictive access control
- External sharing: Paper/email/USB
- No to high encryption

**Analysis objectives**
- Descriptive (what went wrong/well?)
- Prescriptive (what is happening?)
- Predictive (what to expect in future?)
- Supported by: BI/dashboarding /visualisation / data service tools
- Affected by: Analytics Teams - limited sizes, skills and information

**Information sharing**
- Law enforcement agencies
- Taxation authorities
- Intelligence agencies
- Supervisory agencies
- FIU counterparts / FIU Net (European Union), etc

## Towards an AML/CFT SupTech vision

New data technologies hold great promise to improve every element of the current AML/CFT information management life cycle. The pace of adoption by a specific supervisory authority will however depend on a range of factors for example:

- The technology maturity level of the supervisory authority and its staff together with the commitment levels of senior management will influence decisions, such as whether to purchase commercial off-the-shelf solutions (COTS) versus develop in-house solutions, adopt in-house versus outsourced hosting, support and maintenance, amongst others.
- The adoption of automated data collection solutions will depend on the technology maturity levels of regulated institutions while information-sharing technologies will need to be considered in conjunction with the other government agencies (e.g. law enforcement, intelligence and supervisory agencies) that receive information from the agency.
- Employment of enhanced data analytics by FIUs will need to comply with the international AML/CFT standards set by FATF that encourage FIUs to use analytical software while cautioning that such tools cannot fully replace the human judgement element of analysis.[25] This may be revisited as FATF's trust in enhanced analytics grow but for now FIUs are required by FATF standards to ensure human oversight of the analytical process.

Where the implementation of new data technologies requires standardization of data, questions about the reach of the innovation are also relevant. Will it only affect new data being collected or will it also require changes to existing data holdings? If the latter, how extensive are the changes and how large or small is the current data holding?

It is important for supervisory authorities to adopt a comprehensive SupTech vision and ensure they enjoy full government and industry support for planned innovation. A key factor to take into account is that the basic premise of SupTech is Regtech, and their developmental processes are often dependent on or interacting with each other. Understanding the International Regtech Association (IRTA) principles for Regtech firms[26] and knowing which core principles such institutions may follow when implementing Regtech within a financial sector environment should be considered when developing a vision.

- **RegTech/SupTech[27] solutions can increase the evidence base for policy development and rulemaking. Richer datasets allow for more sophisticated modeling and impact analysis when defining prudential requirements, for example. Adding new channels of data collection can improve the feedback loop between financial authorities, FSPs, and consumers. For instance, new means of collecting and analyzing customer feedback (e.g., chatbots) and understanding their behaviors (e.g., AI) can give consumers a louder voice in shaping regulation. Such inputs could induce more user-friendly and consumer-oriented regulation.**

[25] FATF Recommendations INR 29 par 3.

[26] IRTA (2018), Principles for Regtech firms.

[27] Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018a), cit.

- **RegTech/SupTech solutions can streamline licensing and compliance procedures of firms seeking to enter new markets or lines of business, thereby further lowering the regulatory barriers to entry.**
- **RegTech/SupTech can help to craft smarter policies to drive financial inclusion and close the gender gap. Access to richer supply-side data adds context and color to demand-side surveys on financial inclusion, thus painting a more holistic and nuanced picture of the degree of financial exclusion in a given market. Such data can then serve as quantifiable benchmarks against which to measure progress towards financial inclusion goals. For instance, disaggregating high-quality supply-side data by sex (e.g., how many men and women are reached by a given channel or product) can expose gender gaps in financial service delivery, which can then be triangulated with data from customer complaints to understand and address barriers to usage. Since women are excluded from the financial system at a higher rate than men, targeting gender-based measures could yield disproportionate returns to investment in financial inclusion programs.**[28]
- **RegTech/SupTech can facilitate efficient financial intelligence collection to enhance law enforcement and national security.**

The vision should include the development of appropriate ethical and legal frameworks for data governance to ensure that opportunities to further national security, law enforcement and supervisory objectives are enhanced while risks to individuals and their rights are minimized and practical and fair remedies are available when such risks eventuate.

The implementation plan to embed new technologies in its operations should ideally take a staged approach using the Agile Methodology,[29] an iterative and collaborative approach to developing software that has spread to broader project management, to assist agencies to navigate their context and environment and adopt solutions to address their most pressing data pain points. For authorities who want to explore specific SupTech tools first before committing substantial resources, there are helpful institutionalized or one-off methodologies such as innovation labs, accelerators or tech sprints.[30]

The Toolkit below is designed to enable AML/CFT agencies to perform a diagnostic to determine key pain points across the intelligence cycle and identify potential SupTech solutions that may address them. It provides guiding principle discussions and diagnostic questions.

[28] Global Banking Alliance for Women (2015), Measuring Women's Financial Inclusion; The Value of Sex-Disaggregated Data.

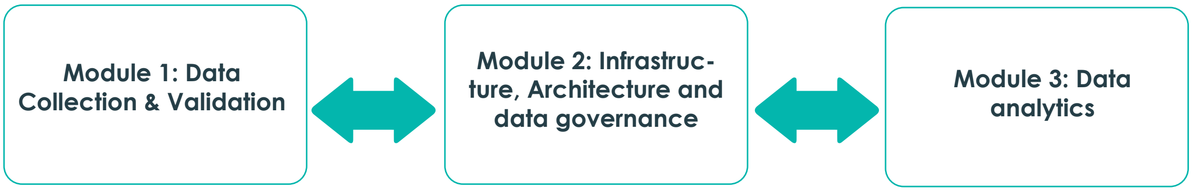
[29] For the applicable principles, see the Manifesto for Agile Software Development. See also Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018b), The RegTech for Regulators Accelerator (R²A) Process Giving Financial Authorities Superpowers, BFA Global RegTech for Regulators Accelerator (R$^2$A) white paper.

[30] Simone di Castri, Stefan Hohl, Arend Kulenkampff, and Jermy Prenio (2019), cit.

# Toolkit: A SupTech diagnostic

In performing a diagnostic to assess your organization's innovation potential, requirements and readiness profile, background preparation research is necessary. This will identify the processes and systems that exist and provide insight into what opportunities can be explored. To identify opportunities for innovation the questions at each juncture can guide you.

The diagnostic process is consolidated into 3 phases, combining different aspects of the data pipeline. These phases influence each other on different levels. The modules are the same for both the intelligence and supervision cycles, however, the focus and as such the applicable choice of innovation and how it is implemented could differ.



**Module 1: Data Collection & Validation**

- Data Origin
- Supervisory Processes
- Data Pain Points
- Collection & Validation processes & flows

**Module 2: Infrastructure, Architecture and data governance**

- Data Pipeline Understanding
- Data Management Pain Points
- Supervisory Processes

**Module 3: Data analytics**

- Data Analytics Landscape
- Tooling Pain Points
- Supervisory Processes
- Data sharing & dissemination
- Organizational & Human Resource Pain Points

There are 3 types of diagnostics questions in the Toolkit:

- Strategic open-ended questions that depends on individual organizational context.
- Situational analysis and pain point questions with guiding options set out to assist in determining a set of options that can be combined in different manners to develop a future strategy. The options are set out in the Diagnostic Questionnaires.
- Situational analysis and pain point questions with guiding options providing an element of linearity that will assist in developing a step-based innovation roadmap. These options are set out in the Diagnostic Questionnaires and marked with an ↑.

Not all the questions will align with each individual organization. Therefore, in some cases, a category "other" is specified as part of the options provided which can then defined according to a best fit for individual context.

When answering the diagnostic questions, remember to look at the end use of the data, i.e., what is the intelligence or supervisory purpose served by the process? For example, given a specific intelligence purpose, are there other types of data that would improve the end use product? If so, this could flow into a change of data fields collected by regulated institutions or result in purchasing external data sources.

## MODULE 1: DATA COLLECTION & VALIDATION

**This module incorporates the following aspects of the data pipeline: Origin, types, channels, protocols, quality and validation**

International guidelines and local laws and regulation to a large extent dictate from whom (regulated entities) and what type of data (reporting sources) should be collected. A thorough understanding of the existing data environment with a clear definition of data requirements is key to fulfilling the supervision mandate. A deeper appreciation of the capabilities of these regulated entities is also important to ensure the development of a pragmatic innovation path. At the same time, streamlining the collection process to ensure consistent, relevant and accurate data can unlock the value of big data that will enable a whole new level of analysis.

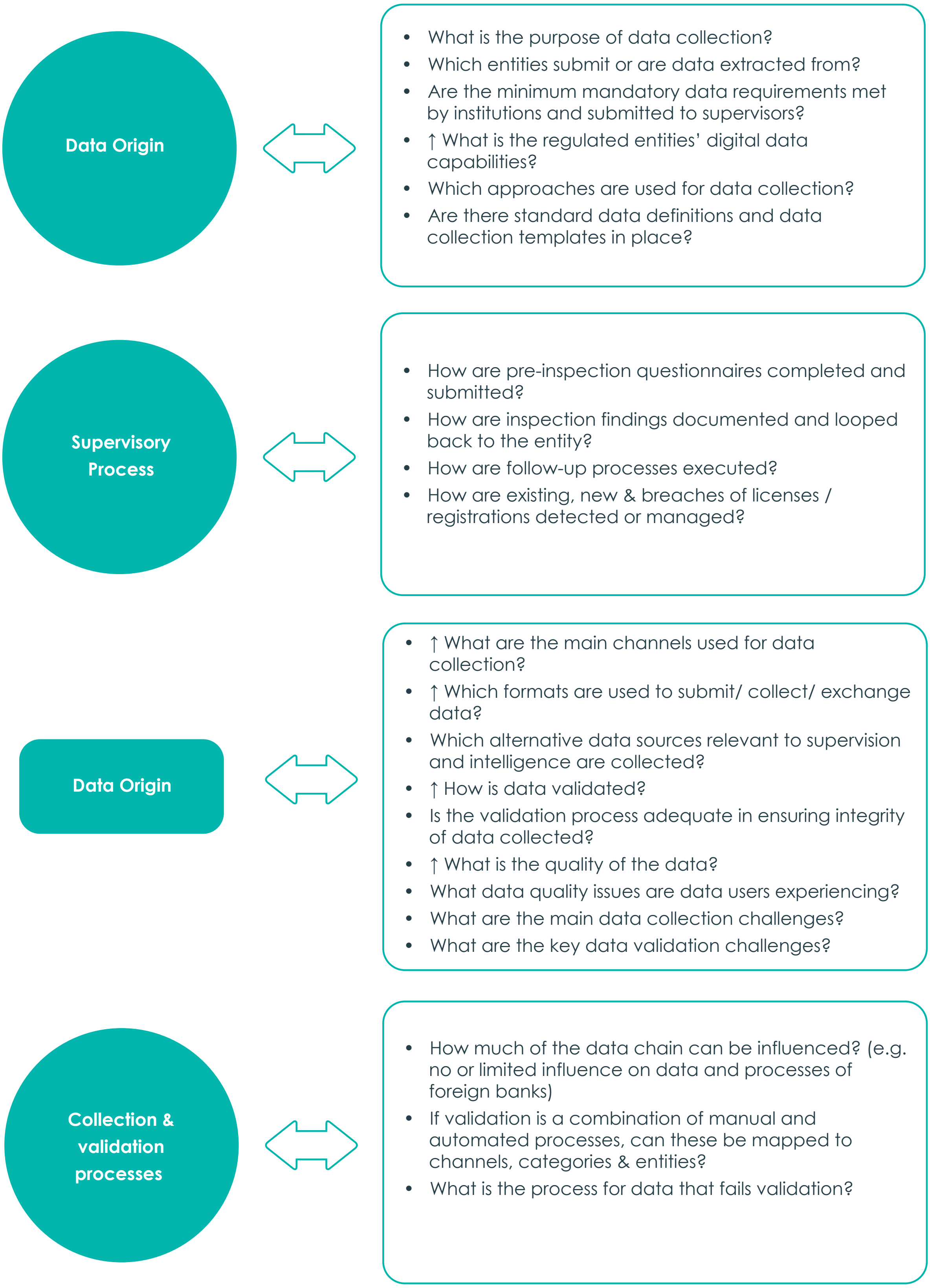
**Big Data analytics[31] can provide opportunities in key areas of interest to central banks, namely (i) the production of statistical information; (ii) macroeconomic analysis and forecasting; (iii) financial market monitoring; and (iv) financial risk assessment.**

**As a note of caution, feedback from BFA Global central bank pilot projects consistently highlights the complex privacy implications of dealing with Big Data, and the associated reputational risks.**

Given the continuous evolution of data, issues with data quality will persist and therefore AI systems and machine learning algorithms must be sufficiently robust to respond to inherent data quality issues. Whilst data quality is important, a key point to bear in mind is that much depends on context. It is often more important to ensure that the data is fit for purpose than fully accurate and up to date. For example, while a customer address may turn out to be invalid or outdated, it may still be useful for entity resolution. Diagnostic questions to consider in determining where and which innovation or SupTech applications can be implemented is set out below. Guiding options for the questions below are set out in the Diagnostic Questionnaire.

**Reminder: Remember to look at the end use of the data, i.e., what is the intelligence or supervisory purpose served by the data and processes?**

[31] BIS (2019), The use of big data analytics and artificial intelligence in central banking, IFC Bulletin No 50.



Data Origin

- What is the purpose of data collection?
- Which entities submit or are data extracted from?
- Are the minimum mandatory data requirements met by institutions and submitted to supervisors?
- ↑ What is the regulated entities' digital data capabilities?
- Which approaches are used for data collection?
- Are there standard data definitions and data collection templates in place?

Supervisory Process

- How are pre-inspection questionnaires completed and submitted?
- How are inspection findings documented and looped back to the entity?
- How are follow-up processes executed?
- How are existing, new & breaches of licenses / registrations detected or managed?

Data Origin

- ↑ What are the main channels used for data collection?
- ↑ Which formats are used to submit/ collect/ exchange data?
- Which alternative data sources relevant to supervision and intelligence are collected?
- ↑ How is data validated?
- Is the validation process adequate in ensuring integrity of data collected?
- ↑ What is the quality of the data?
- What data quality issues are data users experiencing?
- What are the main data collection challenges?
- What are the key data validation challenges?

Collection & validation processes

- How much of the data chain can be influenced? (e.g. no or limited influence on data and processes of foreign banks)
- If validation is a combination of manual and automated processes, can these be mapped to channels, categories & entities?
- What is the process for data that fails validation?

Before we look at innovations that can enable and result in improvement, let's first look at how to deepen an understanding of data pain points.

**A good starting point:** Performing a mapping exercise. This will allow a visual overview of strengths and weaknesses that may exists across the different data sets collected and identify specific pain points.

The exercise can be done from two perspectives:

- Every specific data source
- Every entity/organization where data originates

The example below shows an example of a manual color-coded mapping of each specific data source for a fictional agency.

| Data type | Submission format | Submission channel | Data received quality | Data validation proces |
|---|---|---|---|---|
| STR/ SAR/ CTRs | XML/ XBRL | Internal Secure web based portal | Incomplete | Moderately inadequate |
| Cross border/ EFTs | XML/ XBRL | Commercial Secure web based portal | Complete | Slightly inadequate |
| Sanction information | Paper/Word/ PDF | Fax/ Email attachment | Tidy | Fully inadequate |
| Legal persions registry | Database | Commercial Secure web based portal | Incomplete | Slightly inadequate |
| Biometrics | Database | Other | Poor | Completely inadequate |
| Criminal investigations | Paper/Word/ PDF | Fax/ Email attachment | Incomplete | Slightly inadequate |
| Identify verifiers (Government issued IDs) | Database | Bulk upload (FTP/ SFTP system) | Poor | Moderately inadequate |
| Population demographics | Text/ CSV | CDs/ External Harddrive | Complete | Adequate |
| Asset data | XML/ XBRL | Internal secure web ased portal | Incomplete | Moderately inadequate |
| Audio/ voice/ speech files | Database | Other | Poor | Completely inadequate |
| Credit records | Database | API-Based system | Actively maintained | Adequate |
| Red flags | Text/ CSV | Fax/ Email attachment | Tidy | Fully inadequate |
| Customs data | Text/ CSV | API-Based system | Actively maintained | Moderately inadequate |

| | | | | |
|---|---|---|---|---|
| No digitization | Limited/ inadequate digitization | Initial digitization innovation efforts | Adequate digitization | Advanced digitization |

**Essentially the mapping specification will analyze, on a field-by-field basis, the level of digitization of the data pipeline category. The digitization levels need to be defined from the lowest current functioning level to the highest desired level within the organization.**

**The next step**: Gaining some insights into what might be coming down the line, for instance, what new data sources are coming online and in what format may this present itself, or what new services are being implemented or which new risk types or other factors may affect data dictionaries.[32]

[32] Generally defined as a centralized repository of information about data with information such as its origin, its relationships to other data, usage, and format.

**And the final step**: Looking at the validation processes and flows by considering tests, checks, rule definitions, approvals and other relevant processes and flows. These are the first steps in ensuring an effective data integration strategy (DIS), i.e. a strategy to combine data residing in different sources to provide users with a unified view of the data. Hand in hand with a DIS goes the implementation of a data pipeline. Modern data pipelines are designed for two major tasks[33], the first of which is to define what, where, and how data is collected. The second task relates to the automation of processes to transform and store data for use in further analysis and visualizations.

It is important to decide whether innovation will be implemented narrowly focused on specific functions or specific data sets only or on a broader or even a whole-of-enterprise basis. As mentioned earlier a staged approach, using the Agile Methodology[34] - an iterative and collaborative approach - may often be better than a large-scale Waterfall approach that follows a sequential, linear process. Consideration must also be given to the next levels of DIS and the data pipeline, including interfaces, automated processing, storage and analysis. Diagnostic questions for these are discussed in the modules that follow.

Having an end-to-end vision of the entire ecosystem and designing an effective and supportive data pipeline should enable operations on different levels:

- What business value, working around the current system and process limitations, can be provided now?
- How can the pain points and limitations identified be best addressed?
- Which people and appropriate skills can be integrated into the agency to solve the challenges?

Suptech applications deals with moving towards automated and real-time data collection and validation techniques. Useful techniques include examples such as using NLP to auto-extract narratives from quarterly and annual reports by regulated entities or inspection reports by the agency; implementing machine learning algorithms to recognize patterns and concerns in the available data.

| Innovation Solutions | |
|---|---|
| Data origin: | Machine Learning algorithms (e.g. Automated licensing process; e.g. determining missing data) (Goes hand in hand with data analytics) |
| | Also refer Module 3: Determining missing data through reporting volumes and sector analysis |
| Data channels: | API-based systems;<br>Chatbots (e.g. Automated whistle blowing or citizen tips guidance) |
| Data collection: | Pull and/or Push approaches<br>Require: Harmonized definitions; Standardized validation rules<br>(Goes hand in hand with data storage solutions) |
| | Also refer to Module 2: Older solutions with limitations: Electronic Data Warehouse (EDW) & Data dictionaries; Newer solutions: Data lakes or pipelines |
| Data Validation: | Encrypted VPN Channels (secure communication methods);<br>IT Validation Software (data integration);<br>Natural Language Processing (NLP) (e.g. text-analysis)<br>Machine Learning algorithms (e.g. building in automated rules, rejections & alerts) |

[33] Garrett Alley (2018), Data integration vs Data pipeline.

[34] For the applicable principles, see the Manifesto for Agile Software Development. See also Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018b), cit.

**Staying on top of innovative solutions and emerging technologies within the broader AI field can be achieved by reviewing reports such as the Gartner Hype Cycle,[35] which represents the maturity and adoption of technologies and application and assist with interpreting technology hype.**

Understanding where the blockages are will ensure innovations are applied in such a manner that gaps are tackled with the greatest degree of benefit. Innovation approaches can be implemented step-by-step or in an integrated approach.[35]

**The more integrated approach can have a knock-on effect on the entities submitting data, in terms of designing data dictionaries for instance, which may have cost and time implications.**

## INTELLIGENCE AND DATA GATHERING

As previously stated, a deeper appreciation of the capabilities of FIs/REs are essential. Generating financial intelligence relies heavily on which and what type of information is gathered by the regulatory entities, how these are processed and consolidated into alerts for FIUs. Challenges experienced by FIs and REs can span a range of issues that will affect an FIUs ability to collect the relevant data, such as dealing with legacy systems that do not talk to each other but house data disparately and presents false alerts. Other challenges can relate to, for example missing data, such as location data on an alert. An FI may lack location detail for a suspicious transaction as a result of IP addresses being spoofed presenting a challenge in determining accurate location and the secondary challenge of obtaining the relevant data from service providers given data privacy, sharing and other security issues.

With this in mind, let's look at a scenario presenting a challenge and innovation solution adopted from an intelligence generating viewpoint:

A key pain point: multiple disparate data channels not allowing for easy integration of new data source types.

Step-by-step SupTech solution:

- Enhancing data standardization across the source types
- Establishing a web-based portal and bulk upload through older API-systems, such as File Transfer Protocol/ Secure File Transfer Protocol (FTP/ SFTP);
- Establishing encrypted data submission channels, allowing only entities that lack the capacity to report via the portal to continue sending reports in alternative ways;
- Utilizing a tool for converting different data schemas into a single Extensible Markup Language/ JavaScript Object Notation (XML/JSON) format, such as Apache Daffodil or implementing an open source data tool for integrating data sources such as Apache Nifi;
- Exploring data mining technologies that support data mining across different data source types to allow for more effective data mining of older data types.

[35] Gartner (2019), Top Trends on the Gartner Hype Cycle for Artificial Intelligence.

Alternative integrated SupTech solution: designing an end-to-end architecture solution by moving towards a data collection pull approach (extracting data rather than waiting for it to be submitted) with implementation of a data warehouse that enables the collection of both structured and raw data in a more real-time manner.

**A word of caution: Plotting an architecture means testing it with (real) use cases in which you may have "semi-structured data" as well.**

The National Bank of Rwanda is one of the first agencies to use a data pull approach. The Rwanda example also illustrates **how validation methodology and processes can be innovated** with data integrity check mechanisms, which can include checks for receipt of data, data completeness, consistency, and correctness. By studying the datasets, variables and their relations, and assessing the quality requirements, automated validation rules and integrity checks can be defined to check against data collected and build in automated rejections and alerts.

**A word of caution: Reaching to implement the latest and most applicable data mining solutions without having a good foundation will generally not lead to effective solutions.**

### TEXT BOX 1: NATIONAL BANK OF RWANDA (NBR)

**Problem statement:** Supervised entities manually compile reports from their data systems to submit to NBR resulting in delayed submission, inconsistent reporting and often inaccurate or missing data. This created complex validation processes and complicated integration with BNRs internal system data.

**SupTech Innovation Solution:** Implementing an e**lectronic data warehouse (EDW) system**, which included: A **direct access and automated raw data-pull** for NBR from the supervised entities IT systems. The development of a **data dictionary** with supervised entities writing a data script to map the data dictionary to information in its own systems. The extraction of supervised entity mapped data into a "staging area" for BNR access and data-pull. An **encrypted VPN channel** with data integrity check mechanisms. Data frequency and consistency improved with some data daily and others in near real-time collection taking place. Quality and integrity rules delivered automated data rejection and email alerts to entities. The EDW enables quick and flexible large data analysis abilities.

**Benefits, challenges and risks – Supervisor:** An improvement in data consistency and accuracy to the Supervisor, together with consistent and standardization of reports were achieved by adapting supervisory processes and methodologies to fully leverage collected data. Not all manual reporting could be discontinued due to entity data gaps. Historical data require cleaning to align to the new data dictionaries and analytical processes. Internal business processes require streamlining.

**Benefits, challenges and risks – Supervised Entities:** There was a once-off cost/investment for supervised entities as a result of the data mapping, but this also led to improving their own data quality which benefits their internal risk management processes. New concerns and challenges regarding data privacy and operational and reputational risks became a factor requiring the development of mitigation strategies.

## SUPERVISION AND DATA GATHERING

The supervisory processes rely heavily on template-based reporting across its off-site and on-site inspection and remedial action processes and data collection. Available data is generally well-structured but is dependent on the case management processes and systems in place by FIs and REs. Data quality and the speed of data collection can therefore pose challenges. Further challenges may lie in making sense of the data and developing the capacity and human resources to interpret increasing amounts of data in real time and making informing appropriate supervisory decisions. Differences in regulated entities can also complicate industry-wide standardization and affect the ability to interpret changes in regulations and proper application thereof.

Now let's look at a scenario presenting a challenge and innovation solution adopted from a supervisory perspective: **Key pain point:** Supervised entities manually populating multiple spreadsheet-based report templates submitted over insecure channels, such as email or more secure web-portals. Further pain points resulting from this process were delayed or late submission affecting the quality and processing of data which ultimately presented challenges in the level of data available for effective data analysis. An **integrated SupTech solution** were applied to ensure the best degree of benefits.

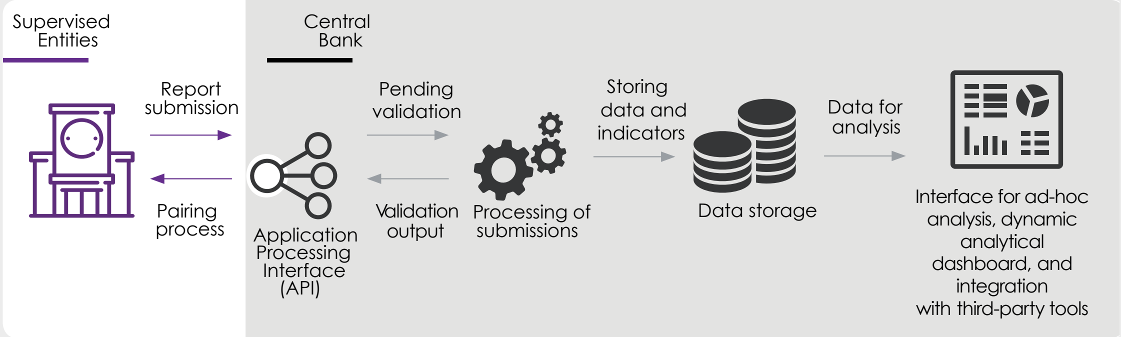
### TEXT BOX 2: BANGKO SENTRAL NG PILIPINAS (BSP)

**Problem statement:** Insecure transmission of manually populated template reports, a process prone to human error and requiring multiple validation iterations.

**SupTech Innovation Solution:** An **API-based data input approach** extracting regulatory reports directly from banks databases, converted into a single encrypted XM-based file & pushed to BSPs processing queue. BSPs new processing engine runs near real-time validation tests with automated responses flagging abnormal values. Other solutions included a **centralized secure access-controlled database** and a**nalytical tools** including a web-based pivot table tool and charts depicting risk indicators.

**Benefits, challenges and risks – Supervisor:** The data input solution enable better and more timely analytics and a more flexible reporting format leading to more efficient maintenance and management.

**Benefits, challenges and risks – Supervised Entities**: Automated process, decreasing human error and less corrective actions required resulting in time savings, cost savings and increased efficiency.


Supervised Entities
Central Bank
Report submission
Pairing process
Application Processing Interface (API)
Pending validation
Validation output
Processing of submissions
Storing data and indicators
Data storage
Data for analysis
Interface for ad-hoc analysis, dynamic analytical dashboard, and integration with third-party tools

Source: Simone di Castri, Matt Grasser, and Arend Kulenkampf (2018),An API-based Prudential Reporting System for the Bangko Sentral ng Pilipinas (BSP) R2 A Project Retrospective and Lessons Learned, BFA Global Case Study, https://static1.squarespace.com/static/583ddaade4fcb5082fec58f4/t/5beb3a3b352f53a18862a9e1/1542142560572/R$^2$A+API+Case+Study.pdf

**In creating a synergy and harmonized approach or model, a key factor is cooperation from all stakeholders and a collaborative transformation process that are co-defined by the authorities and regulated entities alike.**

**Diagnostic Questionnaire:** Guiding options for the respective questions are provided to assist in strategic development of an innovation roadmap. Some options are guides for a situational analysis of pain points whilst other provide innovations that can be considered for implementation. These do not always provide a linear progression but rather a set of options that can be combined in different manners to provide the best possible solution applicable for the organization. Guiding options with an element of linearity are marked with an arrow (↑). Not all the questions will align with each individual organization. In some cases, a category "other" are provided as part of the options where answers can be defined according to a best fit for individual context.

## Data Collection and Validation

| | | | | | |
|---|---|---|---|---|---|
| For what purposes are the data collected? | Prudential | Statistical | Financial Integrity | Operational | Financial Inclusion |
| | Capital and risk requirements | Growth in accounts, regions, service type | Risk identification and analysis, STRs, CTRs, cross-border transactions, etc. | Number of entities and persons, balance sheet compositions, performance indicators & sector information (size & importance) | Indicators |
| What is the intended use or generated intelligence of the collected data? | Sanction/ Watch Lists | Relationship plotting | Typologies | Prosecution & Recovery | Other |
| | Background checks, lead generation and suspect profiling | Identification of motives, associations and underlying criminal activity | Risk patterns and trends identification | Identification of property/ assets for seizing | Define |
| Which entities submit data or are data extracted from? | Regulated entities | Government departments/ agencies/ supervisors | Commercial businesses/ bureau's | Fiscal administrations | National security and law enforcement |
| | Financial and non-financial institutions, lawyers, casinos etc | National identification authority, companies' registry, supervisory agencies | Credit bureaus, newsfeeds | Tax authorities, exchange control authorities | Police, intelligence agencies, courts, etc. |
| ↑ What are the general digital data capabilities of the entities from which data are collected? | Manual | Multiple Incompatible systems | Semi-Automated | Fully Automated | Integrated |
| | Records are manually entered and kept in paper ledgers, scorecards, application forms, etc. | Raw unstructured data housed in legacy systems across different departments with limited data standardization and aggregation | Processed structured data in ETL/ELT architecture able to transform data into regulatory required formats for submission | The system is optimized to allow direct controlled access for supervisor pull approach of raw unstructured and processed structured data collection | Records are kept in excel, database or other formats across different modules (product/ customer/ core banking/ mobile etc.) and imported/ exported or calculated |
| Which approaches are used for collecting data | Separate files | Input | Pull | Other | |
| | Different channels utilized (text/ excel/ pdf/ memory sticks etc.) | Sets of standardized granular data automatically uploaded by institution into supervisor database | Raw granular data extracted by supervisor from institutions IT system | Define | |
| Are there standard data definitions and data collection templates in place? | None | Only for FIs | Across FIs and some non-FIs | Only for government agencies | Across all entities |
| | | | | | |

| Data Collection and Validation | | | | | |
|---|---|---|---|---|---|
| ↑ What are the main channels used for collecting data? | Post/ Courier/ Fax/ SWIFT messaging | Email/ CD's/ External Hard drive | Internal/ Commercial secure web-based portal forms | Bulk upload (FTP/ SFTP system) | API-based systems |
| | | | A specially designed website to bring information from diverse sources, like emails, online forums and search engines, together in a uniform way | Server-client communication protocol with varying degrees of security/ encryption for upload of multiple sources | |
| ↑ Which formats or protocols are used to collect and exchange data? | Paper/ Text Files / Comma-Separated Values (CSV) | Spreadsheets | Flat files | Vector Files | Data exchange formats |
| | CSV output format is a text file with each record from the crawl per line, with the columns separated by commas | e.g. Excel, Extensible Business Reporting Language (XBRL), Google Sheets | dbs data exchange if files in Document-based management systems (DBMS) or relational database management systems (RDBMS) | .gbd file-based in Geographic Information Systems (GIS) | Extensible Markup Language (XML), JavaScript Object Notation (JSON) |
| Which alternative data sources relevant to supervision and intelligence are collected? | Survey & questionnaire data | Virtual asset service provider data and other cryptocurrency data | Social Media/ Website/ internet records | Call detail records/ mobile application data | GIS/ Satellite/ Other Government agency data |
| | | (if not yet covered under regulated entities above) | | | If other, define (e.g. company data) |
| ↑ How is data validated? | Manual cross-checking of different reports | Automated data validation rules | Implementation of IT data validation tools | Other | |
| | | | Specify | Define | |
| How adequate is the data validation process in ensuring integrity of data collected? | Completely inadequate | Slightly inadequate | Moderately inadequate | Adequate | Fully adequate |
| | | | | | |
| ↑ What is the data quality? | Poor | Incomplete | Complete | Tidy | Actively maintained |
| | Missing rows (people/address level entities missing in the data) | Missing columns (variables missing) | Minimal missing data but errors in data collection such as typos | Minimal missing data and no errors in data collection | QA and review processes are in place, and a feedback mechanism throughout the data chain has been put into practice |

## Data Collection and Validation

| | | | | | |
|---|---|---|---|---|---|
| What are the main data collection challenges encountered? | Delayed submission | Inconsistency | Incompleteness | Incorrect interpretation of requirements | Low quality of data |
| | Due date report receipts | Same report indicators/ previous period frequency/ cross-institutional report elements | Missing data fields/ incompatible data/ fragmented information | Reporting format difficulties/ definitional issues/ system differences/ contradictory inputs | Gaps and errors in aggregation and standardization of raw data |
| What are the key validation challenges experienced? | Human errors | Definitional errors | System errors | Calculation errors | Other |
| | | | | | Define |

## MODULE 2: INFRASTRUCTURE, ARCHITECTURE AND DATA GOVERNANCE

**This module incorporates the following aspects of the data pipeline: Storage formats & repositories, hosting, data management and policies.**

Achieving effective data integration require a robust data pipeline, which refers to the movement of data through a series of processes with the transformation of the data as it moves along the chain, for instance producing the right data for data scientists to enable machine learning to develop algorithms that would deliver relevant results for law enforcement end users.

In automating the data pipeline there are many factors to be addressed requiring appropriate expertise, capacity and good planning processes. Various tools can be combined to connect systems whether they are on-premises, in the cloud or hybrid systems. Cloud computing allows for greater and more flexible storage and computing power and may decrease costs. But storing regulatory data in the cloud generally require stronger oversight[36] and appropriate governance.

**Governance of data is undoubtedly a board level issue, with significant implications for strategy, business model, IT architecture, and capital investment, as well as assurance, reporting, and management structures.**[37]

Another factor to consider is whether any tools require appropriate hardware or specialist set-up and whether that is feasible for the agency. Key decisions are required, such as whether to outsource, whether to buy out-of-the-box solutions and the related vendor choices and service level agreement management, which additional software and hardware are necessary and how to scale. These can all present challenges along the way of developing and innovating a data pipeline. Software technologies and protocols can be customized to automate the management, transformation and movement of data as well as data access security. Data intensive environments requiring large storage abilities and high computing power demands specialized workload-optimized hardware platforms that provides the right central processing unit (CPU), graphics processing unit (GPU) and Application-Specific Integrated Circuit (ASIC), to name a few. Finding the sweet spot between hardware, software and business facilitates successful design. When acquiring resources, it is important to obtain outsourced or develop in-house technology that will be eventually be able to systematically execute on multiple AI projects.

**Reiteration: Implementing a data pipeline approach necessitate thinking about the pipeline itself, it different parts and their functions and how best to support experimentation and continuous improvement.**

Data must be copied, moved between storage repositories, reformatted for each system, and/or integrated with other data sources.[38] A thorough understanding of different storage repositories for optimal implementation is necessary as a comprehensive design will result in lessening and streamlining analytical work. Data repositories all have the same core function of storing data, but differ in relation to their purpose; the types of data stored; where the data originates from; who has access to the data; what type of data quality it requires and what type of analytics can be performed from it. It is important to focus on setting up a modern data architecture

[36] See Dias and Staschen (2017).

[37] Ernst & Young (2018), Data governance: securing the future of financial services.

[38] SnapLogic (2019), Data pipeline architecture.

incorporating relational databases, data warehouses, data lakes and data marts with the goal of using each one for what they were designed to do.

**An end-to-end or modern architecture does not necessarily mean that legacy systems and existing IT infrastructure is completely removed but could instead apply the two-speed principle,[39] where traditional data warehouses are scaled down and integrated with high-speed transactional architecture.**

By enabling a Big Data approach, it brings a shift in how data is integrated and processed. The more traditional way of batch processing moves towards real-time integration and processing. Choosing a batch processing technology depends on several factors, but both models are valuable, and each can be used to address different use cases. While the batch processing model requires a set of data collected over time, streaming processing requires data to be fed into an analytics tool, often in micro-batches, and in real-time.[40] An example of micro-batch processing model is set out in the diagram below. Note that semantic interoperability, incorporating aspects such as ontology building, merging and aligning, is required for the model to be effective.

**"What is Semantic Interoperability? (IGI)[41]**

- **Denotes the ability of different applications and business partners to understand exchanged data in a similar way, implying a precise and unambiguous meaning of the exchanged information.**
- **The ability of two or more computer systems to exchange information and have the meaning of that information accurately and automatically interpreted by the receiving system.**
- **Encompasses the meaning of data elements and the relationship between them. It includes developing vocabulary to describe data exchanges and ensures that data elements are understood in the same way by communicating parties."[42]**

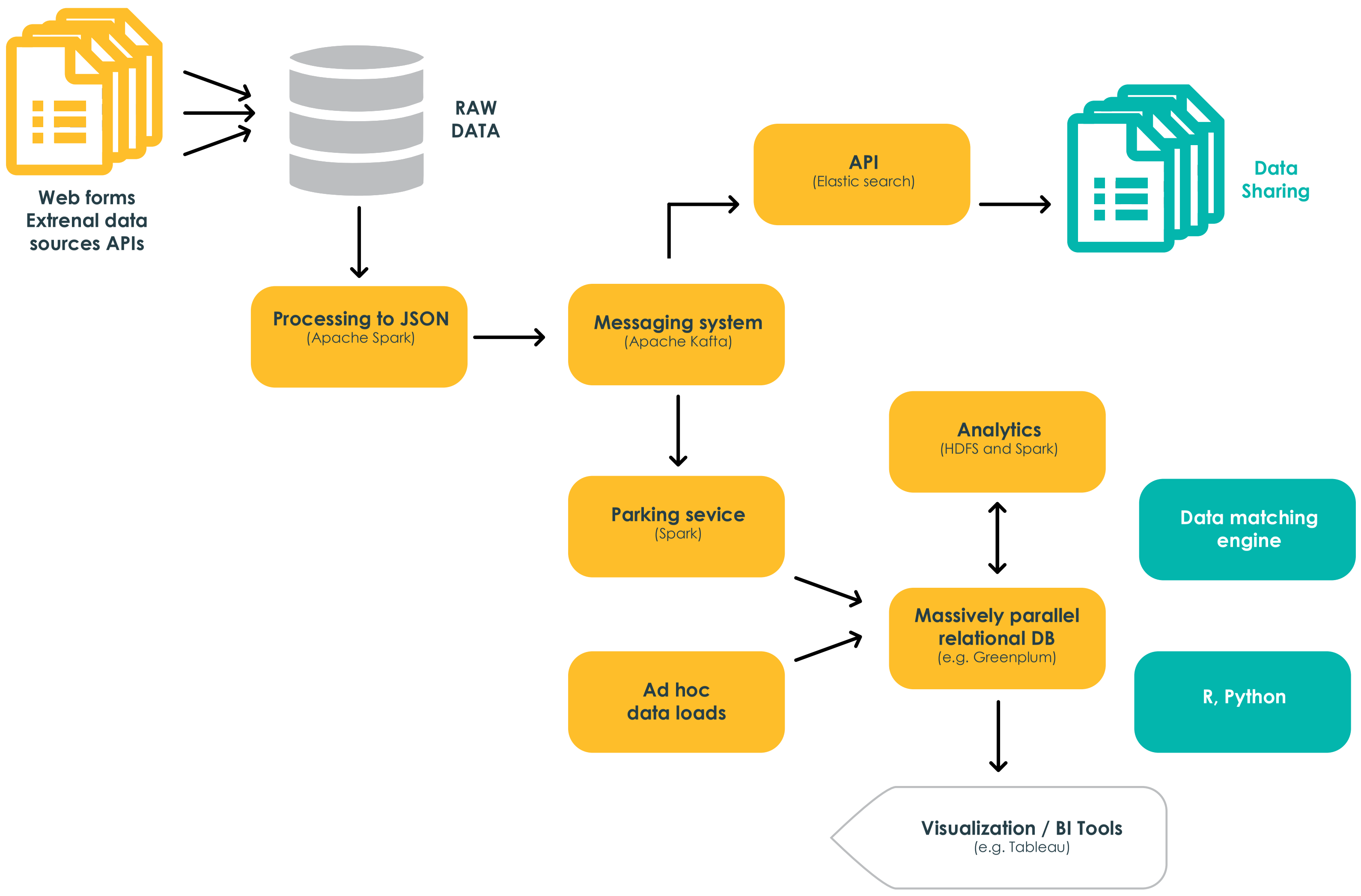
[39] McKinsey (2017), Why you need a digital data architecture to build a sustainable digital business.

[40] Mark Balkenende (2018), TheNewStack: The Big data debate – Batch versus Stream Processing.

[41] IGI Global, Disseminator of Knowledge.

[42] Refer to text box 1: National Bank of Rwanda (NBR).

*Figure 3: Data Pipeline Example – Micro batch processing*



Diagnostic questions to consider in determining where and which innovation or SupTech applications can be implemented are set out below. Guiding options to some of the below questions are set out in the Diagnostic Questionnaire.

**Data pipeline Understanding**

- Where is the data hosted?
- How do you store your data?
- Which storage repositories are being used and combined to enable optimal analysis?
- How does the data get processed?
- How well are data input and storage mechanisms documented?
- Which software tools are being used and combined for optimization of specific areas?
- What are the user tool preferences?
- What are the security requirements?
- Which aspect / risks in cloud computing exists or is most applicable for consideration of service?

**Data management pain points**

- How is historical data managed?
- How accessible is the data?
- Is there a comprehensive data governance policy in place?
- Who is responsible and accountable for breaches of data governance policies and any applicable laws?
- What data privacy policies do you have in place?
- Are there security policies in place for each of the data sources?
- How is data ownership and sharing facilitated and managed?
- What are the key limitations in data access and sharing?

**Supervisory processes**

- What are the key current inspection and monitoring processes?
- How well are inspection and monitoring processes documented?
- How will data be sensitized, pseudonymized or fully anonymized?
- Is legacy system interoperability required?

SupTech applications deals with moving towards enhanced data consolidation, integration and processing with greater and more flexible storage that supports an array of tools and data analytics and build internal capacity. Integration and interoperability between systems can be achieved through APIs to increase efficiency and provide platforms for innovation.

**The fastest growing category of APIs[43] over the past five years is for the sharing and analyzing of data across various applications – an area where the characteristics of the API determine the value of the application or deem it untenable for use in the real world.**

| Innovation Solutions | |
|---|---|
| Data hosting: | Cloud / Hybrid systems |
| Data storage: | A modern storage system comprising of RDMS, EDWs, data lakes & data marts |
| Data accessibility: | Machine readable restricted APIs; open APIs |
| Data sharing: | Encrypted software technology (FIU.net); data lakes |
| Data processing: | ETLs; APIs; Enterprise Service Bus (ESB); Edge Machine Learning |
| Reporting processes: | Auditing Algorithms; Chatbots |

The UK Financial Conduct Authority (FCA) make use of cloud solutions for collecting, storing and processing market data involving "billions of data elements" daily, which is flexibly addressed by the auto-scaling cloud facilities.

## INTELLIGENCE AND SUPERVISION

Whilst FIUs and Supervisory bodies may have different objectives or ultimate end purpose of the data use, building an end-to-end architecture often follow the same paths and innovation implementations.

The important factor is to ensure the architecture supports the objective.

**A good starting point:** Similar to performing a data mapping exercise for data collection, the development of a data flow map to assist in identifying where data resides, how it is processed and where or how it will be accessed and shared, will assist in identifying pain points and where innovations can or should be implemented.

Let's look at a scenario presenting a challenge and innovation solution adopted from an intelligence perspective:

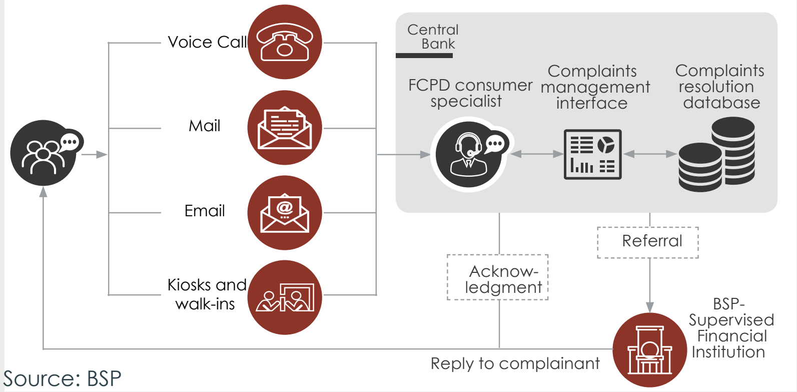
The Bangko Sentral ng Philipinas (BSP) faced the **pain points** of inconsistent, inaccurate and limited data resulting from outdated technology, inability to digitally capture and integrate data due to existing channel methodology in place. They implemented an **integrated SupTech solutions** including APIs, NLP, chatbots, machine learning algorithms, a modern database system and a software analysis tool to provide an **end-to-end architecture overhaul**.

[43] Adam Hughes (2018), What is an API? Everything You Need to Know.

## TEXT BOX 3: BANGKO SENTRAL NG PILIPINAS (BSP)

### Problem Statement:

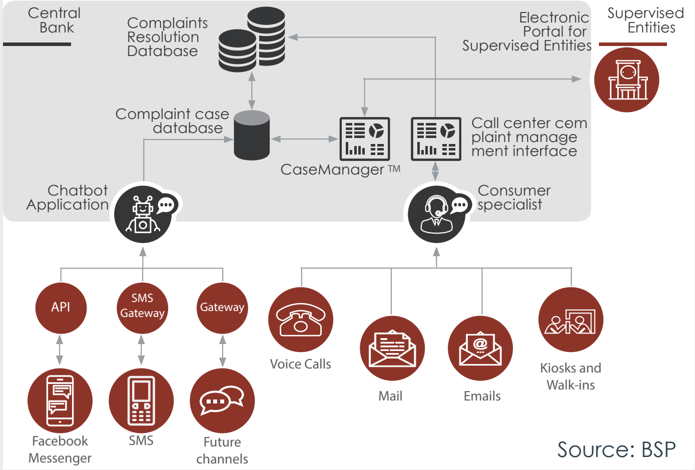
Inefficiencies in complaints data architecture and process on the levels of capturing, classification, storage and analysis of consumer complaints. Administrative burden from highly manual processes and **outdated technologies** with face-to-face interaction requirements and a high reliance on **channel infrastructure** create incomplete, inconsistent and inaccurate information, as well as a customer bias resulting in limited insights for supervision or policy making. An inability in providing timely remediation and proper resolutions with limited visibility of customer experiences resulting **low usage** of system and subsequent **data limitations** raising the possible failure of flagging potential consumer or market conduct risks.



Source: BSP

### SupTech Innovation Solution:

A Big **Data architecture** and AI-driven analytics tool leveraging mobile handsets, which included: A multiple communication channel customer complaints platform compatible with smart and feature phones utilizing chatbots and APIs for **message transmission** and cross-lingual **natural language processing (NLP) technology for message interpretation** following pre-defined conversation flow. A **supervised machine learning model** continuously teaches the chatbot to correctly interpret end-user "intents" and classify complaints into one of ten categories. A **revamped database** allows for auto-updates by the chatbots, manual updates by administrative staff and importing of historical data enables deeper machine learning. CaseManagerTM, a **new interface** allows for analytical views, internal logic configuration and handling and tracking information.



Source: BSP

Now let's look at a scenario presenting a challenge and innovation solution adopted from a supervisory perspective:

The **pain points** experienced by the Central Bank of the Republic of Austria (OeNB) included several disparate reporting, related risks and non-standardized processes and issues related to evidence of regulations followed.

An integrated innovation approach was developed by implementing a data push collection approach and a multidimensional data cube system on a joint platform with a central interface,[44] providing an **end-to-end architecture overhaul.**

**Standardized data definitions applied at the creation of financial reporting should reflect the complete lifecycle of data from collection through dissemination.**

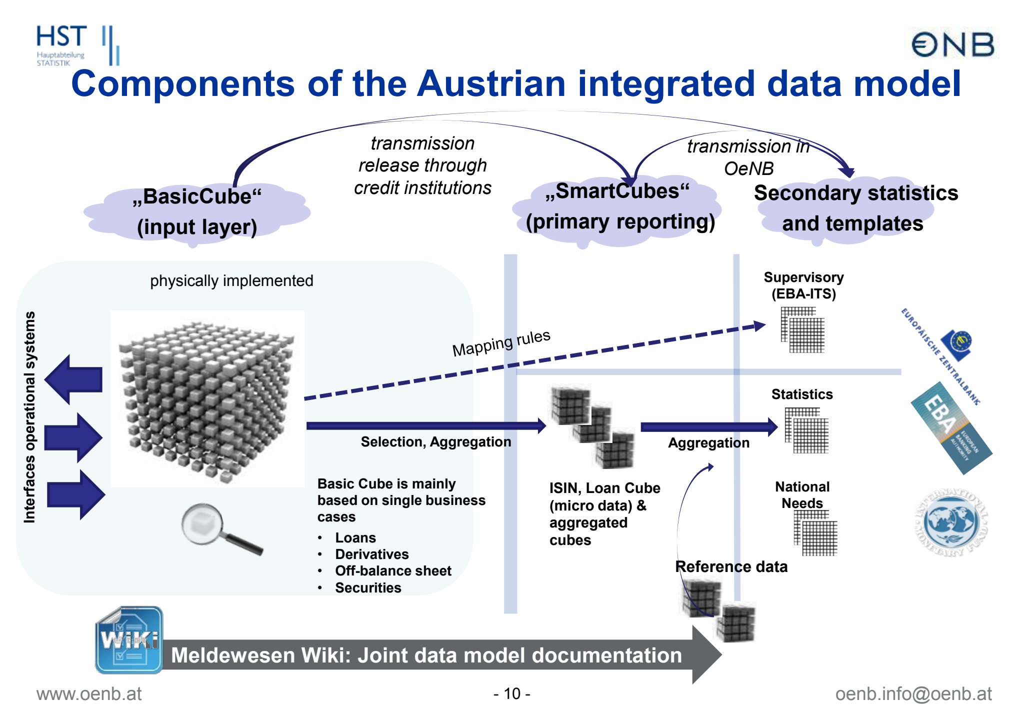
## TEXT BOX 4: CENTRAL BANK OF THE REPUBLIC OF AUSTRIA (OENB)

**Problem statement:** Separate disclosure processes and different reporting obligations create risks of misinterpretations with high instances of queries to the OeNB and inefficient, inconsistent and often redundant reporting from the entities resulting in complexities and high costs for all stakeholders.

**SupTech Innovation Solution:** Integration of content through a harmonized data model based on micro data to define reports with a shared IT-solution. The data model minimizes room for interpretation with definitions and validation rules operating in a standardized language. Central enrichment, aggregation, quality assessment and correction procedures.

**Benefits, challenges and risks – Supervisor:** Legal challenges regarding anonymization in terms of test data were faced during the development stage. The system also created a high level of dependent processes and required a complete adjustment of the organizational structure from vertical to horizontal departments.

**Benefits, challenges and risks – Supervised Entities:** Cost savings, better data quality and reduced queries for correction in data submission. An integrative data model innovative solution jointly developed fostering two-way understanding and transparency.



HST Hauptabteilung STATISTIK

ƏNB

**Components of the Austrian integrated data model**

„BasicCube“ (input layer)
*transmission release through credit institutions*
„SmartCubes“ (primary reporting)
*transmission in OeNB*
Secondary statistics and templates
physically implemented
Interfaces operational systems
Mapping rules
Selection, Aggregation
Aggregation
Supervisory (EBA-ITS)
Statistics
National Needs
Basic Cube is mainly based on single business cases
• Loans
• Derivatives
• Off-balance sheet
• Securities
ISIN, Loan Cube (micro data) & aggregated cubes
Reference data
Wiki
Meldewesen Wiki: Joint data model documentation
EUROPÄISCHE ZENTRALBANK
EBA
INTERNATIONAL MONETARY FUND

www.oenb.at - 10 - oenb.info@oenb.at

Source: EIFR - OENB

[44] BearingPoint, Regulatory reporting platform for Austrian Banks.

**In creating a synergy and harmonized approach or model, a key factor is cooperation from all stakeholders and a collaborative transformation process that are co-defined by the authorities and regulated entities alike.**

The OeNB model also includes data validation based on unsupervised learning, a machine learning style, we discuss in the next module: data analytics.

**Diagnostic Questionnaire:** Guiding options for the respective questions are provided to assist in strategic development of an innovation roadmap. Some options are guides for a situational analysis of pain points whilst other provide innovations that can be considered for implementation. These do not always provide a linear progression but rather a set of options that can be combined in different manners to provide the best possible solution applicable for the organization. Guiding options with an element of linearity are marked with an arrow (↑). Not all the questions will align with each individual organization. In some cases, a category "other" created, where options or answers can be defined according to a best fit for individual context.

**Infrastructure, Architecture and Data Governance**

| | | | | | |
|---|---|---|---|---|---|
| Where is the data currently hosted? | Physical databases, / Desktop Individual supervisors' computers | In-house storage, | Dedicated / private servers | Shared platforms | Cloud |
| | Such as files and binders and separate devices | separate databases for each internal department<br>OR<br>separate databases according to category of institution | | Such as credit registry system etc. | Private or public |
| ↑ How do you store your data? | Analog | Digital Files | Silo Databases | Consolidated Databases | Constellation of systems |
| | Paper records are currently kept in paper ledgers, scorecards, application forms, etc. | Records are kept in files, such as a set of Excel spreadsheets, in a static Google Sheets file, or in a file-based db like SQLite, for example. | Records are kept in a standalone SQL-based or NoSQL/Graph MIS, or a Google Sheet that is connected to an application with AppScripts, for example | Records are kept in a normalized, tidy database (each row is an observation, each column is an attribute), contains data that's been thoroughly validated by a web application against the schema defined in the documentation. | in addition to accounting for the principles of tidy databases, the system is highly optimized based on needs/locations of users, e.g. via sharing, edge caching, replication, serverless infrastructure, etc. |
| Which storage repositories are being used and combined to enable optimal analysis? | Relational Database Management System | Operational Data Store | Data Warehouse | Data Mart | Data Lake |
| | Structured, numerical data, text and dates organized in a relational model captured from a single source using normalized, static schemas to provide organized & consistent data | Transactional data from multiple sources, cleansed and compliant; ingest, integrate, store and prep data for operations or analytics to feed data warehouse or run fast queries on real-time data; denormalized | Upfront workload for data formatting (could result in bottlenecking); relational data from transactional systems, operational databases & application stored for BI, batch reporting and data visualization; curated centralized data captured from multiple sources, denormalized | Relational data subsets; curated data captured from a data warehouse or external sources for specific application or users; normalized & denormalized | Enable easy input of structured and unstructured data with the workload transferred to the user of the data to enable preparation and use in specific context |
| Which aspects / risks in cloud computing exists or is most applicable for consideration of service? | Suitability of cloud services | Costs associated with service | Security and privacy laws, standards and guidelines | Unpredictable electricity interruptions | Other |

## Infrastructure, Architecture and Data Governance

| | | | | | |
|---|---|---|---|---|---|
| | Is it necessary?<br>What is the impact and choice between private versus public cloud? | Service Level Agreement and vendor acquisition processes in place | Regulation is in place to protect service providers and citizens and deal with data breaches/ losses | Stability of connection<br>Probability of data losses | Define |
| Which software tools are being used and combined for optimization of specific areas? | Relational Databases | Massively Parallel Databases | Graph Databases | NoSQL | Pipeline |
| | E.g. PostgreSQ:<br>SQLServer<br>Oracle | E.g. GreenPlum<br>Terradata | E.g. JanusGraph<br>MongoDB<br>Neo4J | E.g. HDFS (Hadoop)<br>Cassandra | E.g. Apache Kafka<br>Apache Nifi |
| ↑ How does your data get processed? | Manual | Statically Automated | Pipeline | Other | |
| | Human manually integrating, humans moving files, humans entering data into system | Spreadsheet macros, robotic process automation (RPA) to collect/merge/ process files, basic scripts that are triggered by new data coming in | Application Programming Interfaces (APIs) are used for moving data place-to-place, synchronous extract-transform-load (ETL) processes are defined and microservices are in place, data generally processed in automated and asynchronous manner (e.g. DataStreams) | Define | |
| ↑ How well are the input and storage mechanisms documented? | None | Manual ERD & Data Dictionaries generated | Generated API Documentation | Platform Documentation | Other |
| | Documentation does not exist, and knowledge of the data storage mechanisms live strictly within the leadership team's mind. | An entity relationship diagram can be manually generated by the database system (for relational db storage) and a set of descriptions of all stored fields (and their data type!) exists and is kept up to date | The application is created in such a way that the APIs can generate docs on how fields enter the system | Considers developer experience, includes code samples and easy-to-embed libraries, interactive documentation includes sample API calls | Define |

## Infrastructure, Architecture and Data Governance

| | | | | | |
|---|---|---|---|---|---|
| ↑ How is historical data managed? | Poor | Incomplete | Complete | Tidy | Actively Maintained |
| | No history kept (key time series data is overwritten), data manually input without validation (e.g. freeform text inputs vs drop-downs/radio buttons) | Granular historical data is not stored but aggregated updates overwrite existing data | Historical data is stored, and new data gets appended with timestamp, preserving old values | All history is kept in line with legislation and/ or policies, and data migration processes and policies are in place to allow for older data usage | QA and review processes are in place, and a feedback mechanism from analysts to DBAs has been put into practice |
| ↑ How accessible is the data? | Completely Closed | Proprietary | File-Based but Standardized | Restricted APIs | Open APIs |
| | (No API or export capability) Only accessible within the application where it is collected | Can be accessible outside the application but proprietary format, requiring specialized analysis software | All machine readable in standard open format (CSV, JSON, XML, database) | All machine readable in standard open format and available through an API | given the proper credentials, others can integrate with the data set to augment/ create their own product solutions |
| Who is responsible and accountable for breaches of data governance policies and any applicable laws? | No One/ Unclear | Junior Staff members | Business department heads | IT department team | Senior Executive Team and Board |
| ↑ What privacy and data protection policies are in place? | None | Limited | Ad Hoc | Comprehensive | Integrated |
| | No policies exist around collection, use, transfer, sharing, storage and deletion of data | Organization has some general policies in place for the collection, use, transfer, storage and deletion of data. | Organization has some policies in place relating to the collection, use, transfer, sharing and deletion of some of its data. | Organization has comprehensive policies in place for the collection, use, transfer, sharing and deletion of data. | The organisation's comprehensive policies are supported by its ethical policy and appropriate training of for all relevant employees. |
| ↑ How is data ownership and sharing facilitated and managed? | Departmental & no sharing | Departmental & Ad-hoc access | Inter-disciplinary & Dashboards | Organization wide & Specialized | Entity wide & Controlled |
| | Only accessible within the department where it is collected. | Can be accessible outside the department but requires case-based approvals, access to shared drive folders, or other bureaucratic processes. | Insights are available to those who need it but require a request to the dashboard developer to incorporate new features | Access is available to granular data across departments but requires a special role-based classification or specialized skillset (e.g. ability to write SQL queries) to make use of this ability. | Access to granular data is generally made available across entities and is in compliance with well-defined data use and privacy policies (MoUs etc.). |

## MODULE 3: DATA ANALYTICS

**This module incorporates the following aspects of the data pipeline: *Analysis objectives, information sharing***

To modernize the supervisor's financial intelligence analytics, it is important to ensure that the data organization layer in the end-to-end architectures discussed in the previous module has been appropriately developed. Significant aspects that form the baseline for data science to work optimally, include:

- Access to all relevant data, not siloed sets of data or only a specific subset
- Relevant tools, skills and platforms to query and pre-process the data
- Relevant tools, skills and platforms to create machine learning models and implement the algorithms

***Different projects may require different tools. If vendor products are considered, the roadmap to new features may be limited to those offered by the vendor products.***

The diagnostic process should ideally assist you to develop a few AI projects that can be prioritized and executed within a sequence to provide cross-functional support and benefits. This may require partnering with external consultants with deep AI knowledge that can be combined with internal teams that have deep domain knowledge to build AI solutions that can show progress within a short timeframe.

**A good starting point:** Mapping the type of analysis (descriptive/ prescriptive) to the ***use case or problem to be solved*** (risk identification and sector mapping, criminal conduct and fraud detection, supervisory resources and alert management etc.) will assist in defining the required analytical results. It will also enable identification of the types of datasets (e.g., transaction and contextual factors) required for the analysis, their level of detail and type (granular, aggregate, un/structured) and the data storage repositories from which data will be sourced to perform the analytics. If there are gaps in any of these areas, it will identify an architecture pain point that can be further probed through the diagnostic questions highlighted in module 2.

**Search and database technology: AI-based search algorithms can better match users with what they are looking for making qualitative aspects easier to search and compare.**

**The next step:** By expanding the data mapping of collection, flows and analytics to incorporate additional governance and human resource aspects, such as 'the who' and 'the skill' related to data access, management and stakeholder engagement it can provide an analytical trail of personnel information that can assist in defining security processes and policy design as well as identify skills and workflow gaps and provide a valuable cross multi-disciplinary team analysis. Key skills span 5 distinct areas: Programming Fundamentals (across different languages); Mathematics and Advanced Statistics; Machine Learning Algorithms; Data Modelling and Software Engineering.

A key aspect of building an AI transformative strategy is to ensure broad AI training for employees in addition to hiring the correctly qualified staff. This will ensure an enterprise wide understanding and ability for divisions to identify their needs and build sustainable solutions. This training can include digital content from open source or online courses or alternatively hiring consultant to provide a mixture of in-person and digital content.

**The final step:** A path to take machine learning models from idea to prototype to development and finally production. So how do you think through such a path? One example is the AI Canvas,[45] that follows this path:

| | |
|---|---|
| Prediction: | • What is the objective, e.g., reduce uncertainty, predict new risks?<br>• What do you need to know to make the decision or improve the service? |
| Judgement: | • How do you value different outcomes and errors? |
| Action: | • What are you trying to achieve, e.g. generating automated data rejection and email alert |
| Outcome: | • What are the metrics for the task success? |
| Input: | • Which data is needed to run the predictive algorithm?<br>• Concurrent data fed to the machine learning algorithm to produce a prediction. |
| Training: | • What data is needed to train the algorithm?<br>• Using historical data to generate the machine learning algorithm. |
| Feedback: | • How can the outcomes be used to improve the algorithm? |

Given the different use cases, there are many potential models that can be applied. It is important from the outset to develop ethical frameworks and privacy and data protection policies as discussed and detailed in module 2. It is also advisable to review models to determine whether they meet their intended objective and require any adjustments.

The choice: Which machine learning (algorithms trained to recognize a specific pattern)) will best solve the problem defined and deliver the analytical results desired?

- The choice of algorithm (the hypothesis set) is dependent on different factors, such as the nature of the data, computational ability and time, the suitability of it for your problem and so forth.
- Algorithms need to be developed and adjusted to draw appropriate value from the data.
- There is an array of algorithms, many of them available from open source.
- Machine learning models will learn what it is taught according to the training data.
- Training the algorithm requires rigorous processes.

In determining the best choice, two approaches to categorizing algorithms can be considered, which are:

- By learning style (supervised/unsupervised/reinforced etc.)
- By similarity in form or function (regression/decision tree/Bayesian etc.)

There are only a few main learning styles or learning models that an algorithm can have and this approach is useful because it forces you to think about the roles of the input data and the model preparation process and select one that is the most appropriate for your problem in order to get the best result.[46]

[45] Harvard Business Review (2018), A Simple Tool to start making decision with the help of AI.

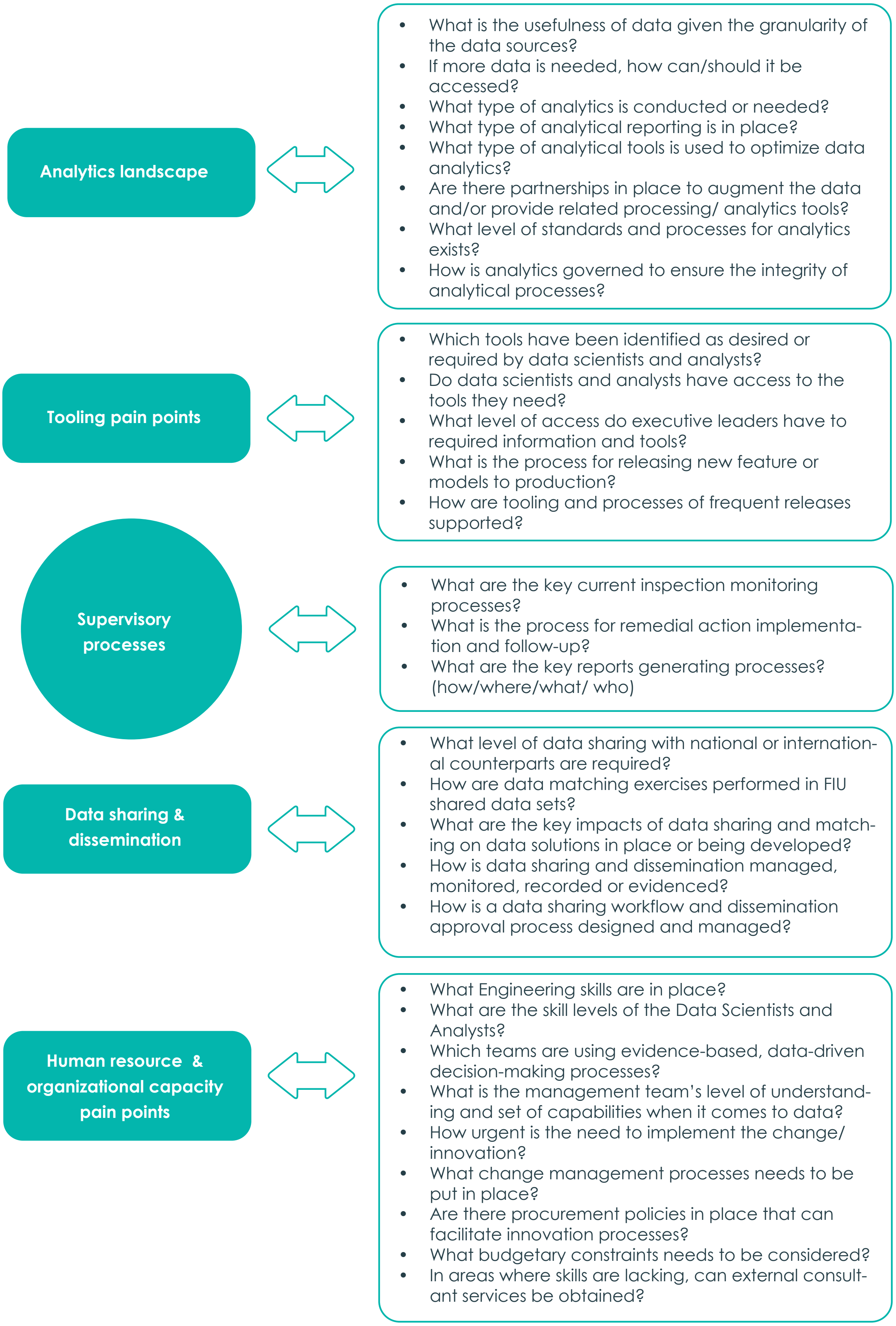
[46] Brownlee (2019), Machine learning mastery: A tour of ML algorithms.

**The uncertainty**:[47] How do we ensure that algorithm-driven decision-making can be trusted and held accountable? This uncertainty highlights the need for appropriate ethical and legal frameworks for data governance to support the use of AI technologies while mitigating the risks of negative impact on individual rights and providing appropriate, practical remedies when such risks eventuate.

Deploying a machine learning pipeline can be complex. Adding to that is the fact that often different parts of a machine learning pipeline are in different programming languages (R, Python, etc.) which may not work always well together. Some solutions that can assist with deployment include the serverless movement, containerization or microservices architecture.

Diagnostic questions to consider in determining where and which innovation or SupTech applications can be implemented are set out below. Guiding options to some of the below questions are set out in the Diagnostic Questionnaire.

[47] WEF (2019), The new physics of financial services.



**Analytics landscape**

- What is the usefulness of data given the granularity of the data sources?
- If more data is needed, how can/should it be accessed?
- What type of analytics is conducted or needed?
- What type of analytical reporting is in place?
- What type of analytical tools is used to optimize data analytics?
- Are there partnerships in place to augment the data and/or provide related processing/ analytics tools?
- What level of standards and processes for analytics exists?
- How is analytics governed to ensure the integrity of analytical processes?

**Tooling pain points**

- Which tools have been identified as desired or required by data scientists and analysts?
- Do data scientists and analysts have access to the tools they need?
- What level of access do executive leaders have to required information and tools?
- What is the process for releasing new feature or models to production?
- How are tooling and processes of frequent releases supported?

**Supervisory processes**

- What are the key current inspection monitoring processes?
- What is the process for remedial action implementation and follow-up?
- What are the key reports generating processes? (how/where/what/ who)

**Data sharing & dissemination**

- What level of data sharing with national or international counterparts are required?
- How are data matching exercises performed in FIU shared data sets?
- What are the key impacts of data sharing and matching on data solutions in place or being developed?
- How is data sharing and dissemination managed, monitored, recorded or evidenced?
- How is a data sharing workflow and dissemination approval process designed and managed?

**Human resource & organizational capacity pain points**

- What Engineering skills are in place?
- What are the skill levels of the Data Scientists and Analysts?
- Which teams are using evidence-based, data-driven decision-making processes?
- What is the management team's level of understanding and set of capabilities when it comes to data?
- How urgent is the need to implement the change/ innovation?
- What change management processes needs to be put in place?
- Are there procurement policies in place that can facilitate innovation processes?
- What budgetary constraints needs to be considered?
- In areas where skills are lacking, can external consultant services be obtained?

Defining the big-picture problem or specific goal for SupTech applications can originate from management, as policy questions from internal units, academics, etc. or as a need identified by supervisory and enforcement units. Exploring and developing solutions for these goals often necessitates the creation of dedicated units or by implementing an integrated multi-disciplinary team approach.

**Creating dedicated SupTech units: The Monetary Authority of Singapore (MAS) announced the formation of a new Data Analytics Group (DAG) which included a "Supervisory Technology Office (SupTech)"[48] that will conduct data analyses on supervisory and financial sector data in partnership with MAS departments.**

**Implementing a multidisciplinary team approach: The Bank of Italy (BoI)[49] created a team from its internal department that includes economists, statisticians and computer scientists to build a hardware and software infrastructure capable of dealing with big data that utilizes an array of tools such as Python, R, etc. and the open-source in-memory software layer, Spark. (refer to module 2: data pipeline example)**

Supervisory or enforcement units, such as an FIU, is in an ideal position to assess which data is available and what solutions can be generated from it.

| Innovation Solutions | |
|---|---|
| Descriptive analytics | Risk dashboards; Early warning systems |
| Augmented analytics | Pattern Recognition: Segments / anomalies |
| Prescriptive analytics | Expert systems (rule based) |
| Regulatory interpretation: | Natural Language Processing (NLP) (Machine readable regulation)<br>Chatbots for virtual assistance |
| Knowledge control | NLP; Sentiment Analysis; Authentication; Reasoning |
| Semi-structured | Graph & Link analytics; NLP: Text embedding; Taxonomy; Search |
| Predictive | ML/algorithms: Modelling policy simulation (clustering/ regression etc.) |

Solutions such as utilizing text-mining techniques can help understand what technologies or products can create disruption or which product customization are not adequately managed by the regulated institutions.[50]

**The U.S. Federal Reserve are reportedly using natural language processing (NLP) to help them identify financial stability risks.[50]**

[48] MAS media release (2017), MAS Sets up Data Analytics Group.

[49] FSI Insight No9 (2018).

[50] Giorgio Gasparri (2019), Frontiers in AI: Risks and Opportunities of RegTech and SupTech Developments. See also Quarles (2018), A Conversation on Machine Learning in Financial Regulation.

Other solutions include automating regulatory interpretation processes to support regulation and compliance[51] as well as creating an integrated and automated reporting template infrastructure. Converting regulatory text to a machine-readable format using NLP leads to greater consistency and improved compliance.

**The UK FCA is exploring the potential to implement machine-readable regulations. It is expected that this could assist supervisory agencies to reduce regulatory complexity and increase impact assessment.**[52]

**AUSTRAC is running a pilot program that aims to convert international funds transfer instruction (IFTI) rules to a decision tree format which can then be codified into software.**

## INTELLIGENCE AND DATA ANALYSIS

The use of data analytics to identify signals of emerging risks and macro-financial risks offer supervisors many opportunities, such as support in policy development and objectives. The CNBV example below, uses AML compliance data to produce customized reports for policy development purposes.

[51] Mustafa Hashmi, Pompeu Casanovas and Louis de Koker L. (2019), Legal Compliance Through Design: Preliminary Results.

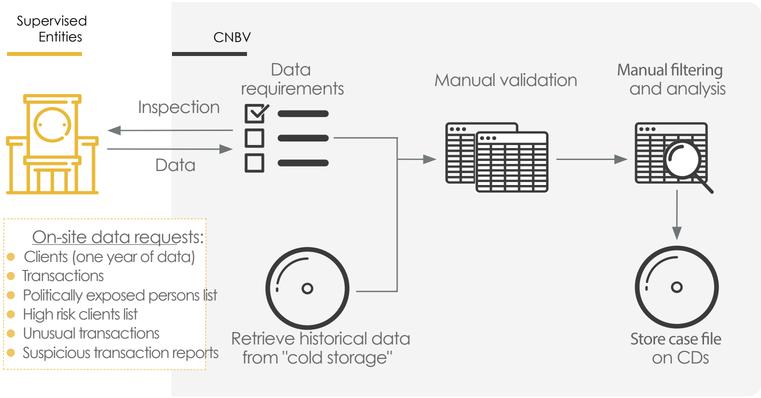
[52] Dirk Broeders and Jermy Prenio (2018), cit.

## TEXT BOX 5: MEXICAN NATIONAL BANKING AND SECURITIES COMMISSION (CNBV)

### Problem Statement:

Inefficiencies in AML data architecture combined with many FSPs categorized as high-medium risk results in inadequacies in drawing deep insights from data informing onsite visits or otherwise as well as delayed and unproductive auditing. These challenges arise from pain points identified across the AML supervision life cycle. In terms of **data extraction** this includes the submission of incomplete, erroneous, or improperly formatted reports. In terms of **data transmission**, it includes the data formats, file size limitations and security. In terms of **data storage**, it includes diverse data formats and a lack of a data warehouse system. And in terms of **data analysis** it includes reliance on manual loading of "cold storage" data loading and Excel analytics which prevents complex data mining and application of statistical models and data visualization.

The starting point:

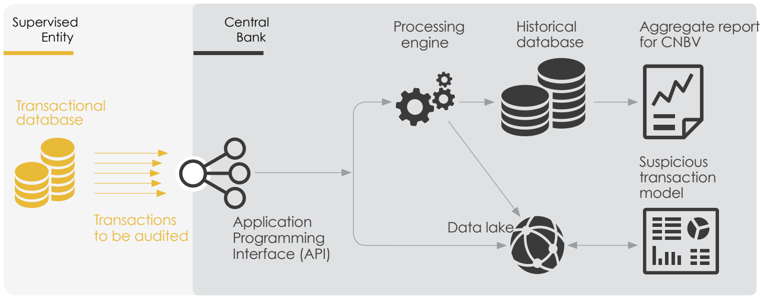


Source: CNBV

### SupTech Innovation Solution:

An API-based AML data architecture and **AI-driven analytics tool**, which included: Centralized platform to generate standardized, automated requests to supervised entities with raw data received through push or pull submission stored in a data lake. An API to establish secure, direct line of machine-to-machine data transmission feeding the data into a processing engine instantly running validations tests verifying quality, content and structure of reports and funneling processed data into the data lake creating a consolidated, single and access-controlled data architecture. AI-driven analytics that detects suspicious transactions using **predictive analysis** and ML techniques (clustering, neural networks, logistic regression, random forests) and **recommend AML** alerts using ML based on FIs underlying risks exposures. Dashboards and watchlist tracker provide a near real-time view of the AML risk landscape.

What success looks like:



Supervised Entity
Central Bank
Processing engine
Historical database
Aggregate report for CNBV
Transactional database
Transactions to be audited
Application Programming Interface (API)
Data lake
Suspicious transaction model

Source: CNBV

**Other innovations implemented includes:**

- **The experimental project of The Bank of Italy (BoI), in testing the use of machine learning and deep learning techniques to classify incoming suspicious transaction reports.[53] The US Federal Reserve[54] using bid data in its Comprehensive Capital Analysis and Review (CCAR) stress-testing process based on granular loan data used to project losses in each retail product.**

## SUPERVISION AND DATA ANALYSIS

The primary purpose of risk-based supervision is to ensure that AML/CFT efforts of FIs and REs are consistent with their risks and other non-risk-based compliance obligations. In achieving this purpose there are many ways in which innovations such as big data, machine learning and interactive insight systems can be developed, such as risk mapping and/or comparative reviews of FIs or REs to assess accuracy of their inherent risk reporting. adjust supervision priorities as and where necessary and ensure their suspicious alerts reported correspond with their risk levels.

With the increasing volume of data, mulling your way through it to draw inferences and make predictions with automated and manual systems has become a treacherous task. SupTech innovations can alleviate this burden. Priming data analytics, machine learning and NLP towards detecting outliers and anomalies are useful techniques. Machine learning techniques for outlier detection are usually grouped into model-based, proximity-based and angle-based.

**"Natural language processing (NLP), the technology that powers all the chatbots, voice assistants, predictive text, and other speech/text applications that permeate our lives, has evolved significantly in the last few years."[55]**

**There are a wide variety of open source NLP tools, with libraries mainly written in Python or Java programming languages. "Due to the fact that Python programming language is one of the best suited for Big Data processing, many tools and libraries are written for it. Solutions like Jupyter and other Big Data visualization tools are written in Python, and many other software instruments provide native Python functionality or support through APIs or various wrappers".[56]**

**NLP tool alternatives includes options such as NLTK, CoreNLP, Gensim, Spacy, OpenNLP (easy integration into Apache Spark tools) and IntelNLP Architect, to name but a few. To ensure you get the right insights from analyzing text, being aware of and understanding the different alternatives is key.**

An NLP tool should be chosen according to its use case.[57] One such use case is Named Entity Recognition (NER)[58] which can be implemented in customer support environments or other classification areas such as website news articles scanned with efficient search algorithms.The Financial Stability Board (FSB) mentions the potential of combining machine learning with NLP to identify patterns in the combination of trading data with behavioural data (e.g., communications among traders/employees).[59]

[53] FSI Insight No 9 (2018).

[54] Jagtiani et al (undated).

[55] Barker (2019), 12 open source tools for natural language processing.

[56] Fedak (2018), Towards Data Science: 5 Heroic Tools for Natural Language Processing.

[57] Bilyk (2019), The APP Solutions.

[58] Banerjee (2018), Introduction to Named Entity Recognition: A tool which invariably comes handy when we do Natural Language Processing tasks.

[59] FSB (2017), cit.

**The Monetary Authority of Singapore (MAS) is testing a tool to analyze high volumes of suspicious transactions reports and substitute labor-intensive and time-consuming procedures to filter the reports that demand further investigation (by humans) from false positives.**

Similarly, with the wealth of new forms of data available, the level of complex analysis escalates, but at the same time offer richer and more focused analytical abilities. This is achieved through combining the different data sources and types.[60]

SupTech applications can combine multiple data sources to support analytical work, especially by combining structured and unstructured data which provides a more holistic view of the data.

**The Bank of Italy (BoI), for instance, combines structured data (transaction reports) with unstructured data (twitter/press reviews) to enhance money laundering detection capabilities through sentiment analysis.**

**The US Consumer Financial Protection Bureau (CFPB)**[61] **uses both structured and unstructured complaints data to create company profiles, monitor activities, identify emerging risks and inform its risk-based methodology to prioritize FSPs. The system provides trend analysis, early warning tools, and scorecards. An algorithm-based tool named 'Spikes and Trends' flags short-, medium-, and long-term changes in complaints behavior.**

Moreover, by using SupTech it is possible to create data visualization with interactive dashboards and charts which can facilitate better insights and individual targeted reporting for data and information dissemination to stakeholders. The visualisation of outliers for further analysis however can be a challenging task given the increasingly high-dimensional nature of real-world datasets and the sparseness of outliers in terms of known patterns and trends. Latent spaces have been recently advanced as a solution to this complexity.

[60] See Simone di Castri, Matt Grasser, and Arend Kulenkampff (2020), The "Datastack": A Data and Tech Blueprint for Financial Supervision, Innovation, and the Data Commons.

[61] Toronto Centre (2018), SupTech: Leveraging Technology for Better Supervision.

The Central Bank of Nigeria revamped their systems to implement this type of analytical innovation.

## TEXT BOX 6: THE CENTRAL BANK OF NIGERIA (CBN)

**Problem Statement:**

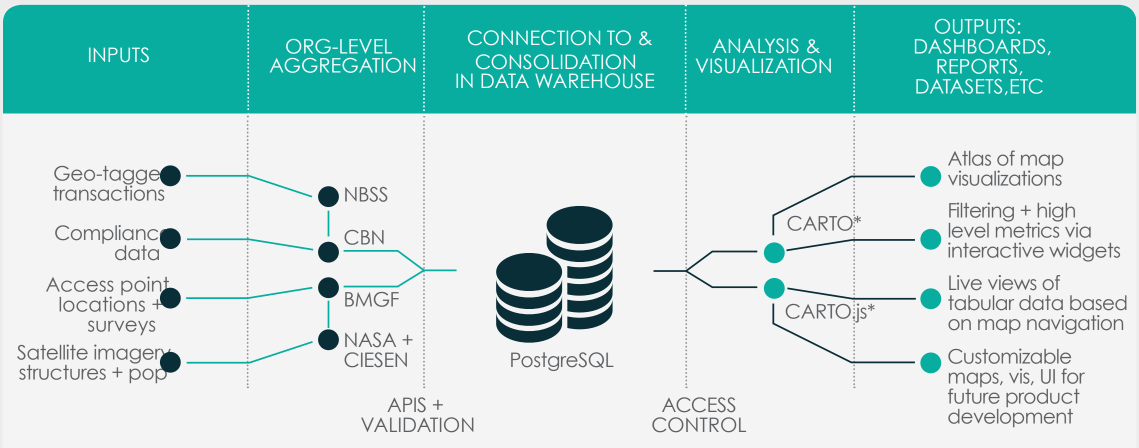
An **ineffective ability to guide supervision** and policy making. Lack of effective analytics resulting from **ineffective datasets for dissemination** to public and private users, which includes competition supervisors when monitoring risks and payment system operators, government agencies monitoring fiscal transfers, private sector identifying innovative products and channels and donor communities assessing program impacts.

**SupTech Innovation Solution:** Developing a supply-side core data rich environment by implementing **a transaction data warehouse** populated via **APIs with real-time** transaction and compliance **validated data** and stored in **relational SQL databases** allowing for it to be joined and queried. Implementing a **data stack** of demand-side data (statistics, financial indicators, satellite imagery etc.) on top of the supply-side data. **Data analysis and dissemination** is enabled through the development of **front-end tailored-made reports and dashboards** with **interactive data visualizations**.

**Benefits, challenges and risks – Supervisor:** Building a cost-effective yet comprehensive and appropriate data architecture. Developing a central analysis easy to use front-end yet maintaining individuality for users. The system created the ability to view key risk metrics and indicators in real time on a risk dashboard serving as a compliance tool and early warning system.

**Benefits, challenges and risks – Supervised Entities:** Learning to use a new front-end analysis system to generate reports applicable to individual context and correctly interpret visualizations generated. An integrative data model innovative solution enabled more effective sharing of intelligence and delivering timely accurate decision making.

Nigeria's Data Stack prototype


INPUTS
ORG-LEVEL AGGREGATION
CONNECTION TO & CONSOLIDATION IN DATA WAREHOUSE
ANALYSIS & VISUALIZATION
OUTPUTS: DASHBOARDS, REPORTS, DATASETS,ETC
Geo-tagged transactions
Compliance data
Access point locations + surveys
Satellite imagery structures + pop
NBSS
CBN
BMGF
NASA + CIESEN
PostgreSQL
CARTO*
CARTO.js*
Atlas of map visualizations
Filtering + high level metrics via interactive widgets
Live views of tabular data based on map navigation
Customizable maps, vis, UI for future product development
APIS + VALIDATION
ACCESS CONTROL

*CARTO is the platform used to render geo-spatial data.

Source: Nigeria Payments and Transaction "Data Stack" Report (to be released)

**Diagnostic Questionnaire:** Guiding options for the respective questions are provided to assist in strategic development of an innovation roadmap. Some options are guides for a situational analysis of pain points whilst other provide innovations that can be considered for implementation. These do not always provide a linear progression but rather a set of options that can be combined in different manners to provide the best possible solution applicable for the organization. Guiding options with an element of linearity are marked with an arrow (↑). Not all the questions will align with each individual organization. There are therefore, in some cases, a category "other" created, where options or answers can be defined according to a best fit for individual context.

| Data Analysis | | | | | |
|---|---|---|---|---|---|
| ↑ What is the usefulness of the data given the level of granularity for the data sources? | Inappropriate or insufficient | Appropriate | Regularly reviewed | Designed Indicators | Other |
| Agency is collecting more data than it ethically or legally should be collecting. Alternatively, the agency is tracking the data only on the highest of levels (e.g. user count, total volume and value of transactions) | Sufficient granular level data is being gathered where needed and ethically appropriate. | Data collection, storage and governance processes are reviewed against best practices (e.g. NIST security protocols), appropriate protection frameworks (e.g. PCI, GDPR) and MOUs and other terms of use of datasets are complied with | Data is being captured with appropriate granularity, and indicators have been carefully crafted to measure progress against purpose of intelligence use | Define | |
| What type of analytics is conducted or needed? | Descriptive | Diagnostic | Prescriptive | Predictive | Applied Analysis[62] |
| Reports & descriptions: What went wrong or well?<br>(alerts/queries/searches tools) | Interactive visualizations: Why did it happen? What are my weak areas?<br>(statistical/ quantitative techniques) | Monitoring, alerts, notifications and recommendations: What is happening now? What is my next best action?<br>(regression analysis, multivariate statistics, pattern matching, data mining) | Future trends and predictions: What to expect in the future? How do I plan?<br>(graph analysis, simulation, complex event processing, neural networks, recommendation engines, heuristics, and machine and deep learning) | Artificial intelligence: How can the data, processes and analysis be adapted to enable self-learning analysis?<br>(combines deep learning, natural language processing and neural network algorithms to mimic the way a human brain works) | |
| ↑ What type of analytical reporting are in place? | Minimal. | Compiled. | Automated. | Dynamic and Interactive. | AI-augmented. |

[62] NOTE: Applied analysis are in the early stages of development and depends upon the context in which is applied.

| Data Analysis | | | | | |
|---|---|---|---|---|---|
| | Data is either not used to inform decisions and reporting, or is exclusively queried manually on an ad hoc basis, and added in only where needed. | Static reports are regularly created, but this is a manual process (e.g. using Excel, but requires requests to database specialists to get the numbers) | Static reports and static dashboards are intermittently generated and updated through a series of scheduled scripts, APIs, triggers, or other means of static automation. | Dashboards are available to explore the data in close to real-time, including visualization and/or report template customization. | Reports and decisions reflect descriptive, prescriptive and predictive analysis of large datasets and appropriate data visualisation. |
| What type of analytical tools is used to optimize data analytics?<br>e.g. Elastic Search<br>Apache Solr | Search | Development | Data Science | BI/ Visualisation | Intelligence Analytics |
| | E.g. C1 (continuous integration e.g. Gradle)<br>Source control (e.g. Git)<br>IDE (e.g. intelliJ) | E.g. Python, Spark, R<br>Front end prototyping tools, e.g. shiny or ability to create web apps | E.g. Visio/ Stata/ Tableau<br>SAS Visual analytics<br>Power BI<br>GoAML | E.g. Palantir<br>IBM Analyst Notebook<br>GoINT | |
| ↑ Are there partnerships in place to augment the data and/or provide related processing/ analytics tools?<br>No data partnerships exist. | None | In Progress | Ad Hoc | Integrated | Real-Time |
| | Partnerships exist but data is not regularly shared as part of the agreement. | Partnerships exist and have policies and technology in place to share data through a manual process or through flat files. | Partnerships exist and have policies in place to share data on demand through APIs or other automatic process. | Partnerships exist and have policies and technology in place to share data in real-time | |
| ↑ How are data matching exercises performed in FIU shared data sets? | Non-existent | Not practiced | Manually checked | Matching tools | Other |

| Data Analysis | | | | | |
|---|---|---|---|---|---|
| | No legal capacity to anonymously share data sets extracts | Matching tools available but not implemented due to privacy concerns | Data matching is done on a case by case basis through motivated requests | Automated technology implemented performing data matching yes/no exercises to identify cross-jurisdictional links | Define |

| Human Resources and Organizational Capacity | | | | | |
|---|---|---|---|---|---|
| Engineering Skills | Front-End developer | Database administrator/ big data specialist/ analytics engineer | Software engineer/ Technical lead | Systems administrator & architect/ IT security advisor | Infrastructure developer / Project manager/ Business analyst |
| | Able to script the processing of basic data via spreadsheet macros, command line scripts, etc. | Database & back-end web application knowledge; Understand how to set up and query a database to produce basic analytics, deploy a web application in a secure manner that allows it to interact directly with the world. | Mobile + front-end services; Understand how to build applications on top of the database and back-end application (business logic, algorithms), how to properly instrument these products with analytics, best practices for reliability/ security/ accessibility (particularly in low-connectivity environments) | APIs + separation of duties; Understand the basics of RESTful API design and implementation (e.g. object-orientation, CRUD operations, authentication like OAuth2 tokens, etc), and how these interact with clients of the API. | Understanding of how to leverage cloud services (e.g. serverless infrastructure, caching services, CDNs, microservices, sharing, reserved/ on-demand/ spot instances) to maximize performance to cost ratios. |
| ↑ Data Scientists Skills | Spreadsheets / Shell Scripts Only. | Statistical / Data mining techniques | Data platforms / architecture | ML model optimization | |
| | Able to do basic munging and transformations of data to come up with metrics. | Ability to manipulate data sets and build statistical models using statistical computer languages (R, Python, SQL) and advanced techniques (regression/ random forest/ text mining) | Able to do development in distributed data frameworks/ computing tools (Hadoop, Spark, Terradata etc.), creating data architecture and use web services | Deployment of features and products that leverage trained models, using cloud infrastructure, streaming data, etc | |
| ↑ Data Analyst Skills | Spreadsheets / Database design | Business Analysis | Descriptive | Predictive | |

| Human Resources and Organizational Capacity | | | | | |
|---|---|---|---|---|---|
| | Understanding relational database design and able to develop basic formulas, graphs and statistics | Ability to run queries, set up dashboards that quickly show off key metrics (e.g. Tableau, PowerBI, R Shiny, Data Studio, etc) | Able to produce Exploratory Data Analysis (Python notebooks / R markdown) and other basic data products, basic feature engineering and modelling (regressions, decision trees, etc) | Predictive Analytics, machine learning (tensorflow, etc) | |
| ↑ Which teams are using evidence-based, data-driven decision-making processes? | None | IT/Ops | Compliance & Analysts | Executive | All |
| | There are no teams making use of raw data or data products as a means to inform their decisions and processes. | The database administrators (DBAs) and dev/ops teams are making basic IT decisions based on the volume and types of data coming in. | The compliance teams are making use of lean analytics via exploratory data analyses, dashboards, reports, etc, to inform processes and roadmaps.<br>The analytics teams make use of raw/structured data to inform their case decisions. | The executive team is incorporating data into their discussions of strategy, policy design, regulatory guidelines, partnership development etc. | There is a formal data strategy in place that lays out which data sources and metrics are useful for each respective member of the team. |

| Human Resources and Organizational Capacity | | | | | |
|---|---|---|---|---|---|
| ↑ What is the management team’s level of understanding and set of capabilities when it comes to data? | None. | Spreadsheets. | Static Reports. | Dashboards. | Advanced Analytics. |
| | Management team tends to rely on physical or digital conversations, and a staff to prepare relevant metrics, trends, talking points, etc. | The management team is using spreadsheets to explore or review data. | The management team is using reports that are statically generated by analytic tools. | The management team has dashboards available for exploring relevant data and is regularly making use of them. | The management team has good data literacy skills and is actively involved in leveraging data science tools to perform exploratory data analysis, modelling, etc. |

# The authors

## LOUIS **DE KOKER**

Louis is a professor of law at La Trobe Law School, La Trobe University, Australia. Between 2014 and 2019 Louis led the Law and Policy Program of the Data to Decisions Cooperative Research Centre. The Law and Policy Program, a $1,8 million collaborative program between La Trobe Law School, UNSW Law and Deakin Law School, investigated policy and regulatory options to protect fundamental rights and ensure national security in a context where data technology is changing the paradigm.

Louis is currently a co-lead of La Trobe LawTech (LT$^2$), a multidisciplinary law and technology research group at La Trobe Law School. Louis' financial crime research focuses on managing the relationship between financial inclusion and anti-money laundering and counter terrorist financing objectives. He has undertaken various university research engagements with the bodies such as the World Bank and AusAID and has worked closely with the Consultative Group to Assist the Poor (CGAP) and with the Financial Integrity Working Group of the Alliance for Financial Inclusion (AFI). This work extended to a range of developing countries including Ghana, Indonesia, Jordan, Kenya, Kyrgyzstan, Malaysia, Namibia, Nigeria, Palau, Uganda and the Ukraine. His publications have been cited in publications and research papers of international bodies such as the World Bank, the Basel Committee on Banking Supervision, the International Labour Organisation, the G20's Global Partnership for Financial Inclusion and the World Economic Forum.

## ROCHELLE **MOMBERG**

Rochelle is currently a strategy, research and educational independent consultant with a focus on financial and digital innovation projects. She has undertaken various research engagements which includes digital identity verification and proofing regulatory compliance and financial inclusion thought leadership strategy projects.

Between 2016 and 2019 Rochelle led the Regulatory stream of the Digital Frontiers Institute where she conceptualized and developed Anti-Money Laundering and Digital Identity courses and taught students across the globe. The Digital Frontiers Institute non-profit organization, funded by The Omidyar Network, FSD Africa, and the Bill & Melinda Gates Foundation, builds human capacity in digital services.

Between 2012 and 2016 Rochelle project managed and performed strategy and research on the Illicit Financial Flows Project of the Financial Intelligence Centre of South Africa. The project estimated the size of illicit financial flows and the illicit economy (including tax evasion, trade mispricing and transfer pricing, money laundering and corruption) and provided policy advice to the Director of the FIC and the Minister of Finance. She performed the modelling to quantify the extent of the illicit narcotics sector within South Africa and developed educational sessions to generate a greater Bitcoin understanding. The work extended to driving the project agenda forward with global stakeholders such as the World Bank, the OECD and Global Financial Integrity.

# Key references

1. BFA Global (2018), The State of RegTech: The Rising Demand for "Superpowers", BFA Global RegTech for Regulators Accelerator (R$^2$A) survey,

2. Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018a), Financial Authorities in the Era of Data Abundance: RegTech for Regulators and SupTech Solutions, BFA Global RegTech for Regulators Accelerator (R$^2$A) white paper.

3. Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018b), The RegTech for Regulators Accelerator (R$^2$A) Process Giving Financial Authorities Superpowers, BFA Global RegTech for Regulators Accelerator (R$^2$A) white paper.

4. Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018c). An AML SupTech Solution for the Mexican National Banking and Securities Commission (CNBV), Project Retrospective and Lessons Learned, BFA Global RegTech for Regulators Accelerator (R$^2$A) case study.

5. Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018d). An API-based Prudential Reporting System for the Bangko Sentral ng Pilipinas (BSP), Project Retrospective and Lessons Learned, BFA Global RegTech for Regulators Accelerator (R$^2$A) case study.

6. Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018e). A Chatbot Application and Complaints Management System for the Bangko Sentral ng Pilipinas (BSP), Retrospective and Lessons Learned, BFA Global RegTech for Regulators Accelerator (R$^2$A) case study.

7. Simone di Castri, Stefan Hohl, Arend Kulenkampff and Jermy Prenio (2019), The suptech generations, Financial Stability Institute Insights on policy implementation No 19, Bank for International Settlements.

8. Dirk Broeders and Jermy Prenio (2018), Innovative technology in financial supervision (suptech) – the experience of early users, Financial Stability Institute Insights on policy implementation No 9, Bank for International Settlements.

9. Toronto Centre (2018), SupTech: Leveraging technology for better supervision, Practical leadership and technical guidance TC Note.

10. World Bank (2018), From Spreadsheets to Suptech: Technology Solutions for Market Conduct Supervision. Discussion Note.

11. World Economic Forum (2019), The new physics of financial services.

# Additional references and readings

1. Adam Hughes (2018), What is an API? Everything You Need to Know.
2. AI4People (2018), Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.
3. Garrett Alley (2018), Data integration vs Data pipeline.
4. Arner, Douglas, Jànos Barberis, and Ross Buckey (2016), FinTech, RegTech, and the reconceptualization of financial regulation. Northwestern Journal of International Law & Business. 37: 371.
5. Bank of International Settlements (BIS) (2015), Central Bank's use of and interest in 'big data', Irving Fisher Committee.
6. Bank of International Settlements (BIS) (2019), The use of big data analytics and artificial intelligence in central banking, Irving Fisher Committee.
7. Banerjee (2018), Introduction to Named Entity Recognition: A tool which invariably comes handy when we do Natural Language Processing tasks.
8. Barker (2019), 12 open source tools for natural language processing.
9. Basel Committee on Banking Supervision (BCBS) (2013), Principles for effective risk data aggregation and risk reporting.
10. Basel Committee on Banking Supervision (BCBS) (2018), Sound Practices: implications of fintech developments for banks and bank supervisors.
11. Bilyk (2019) The APP Solutions: Natural language processing tools and libraries.
12. Brownlee (2019), Machine learning mastery: A tour of ML algorithms.
13. Deloitte (2017), Conversational Chatbots.
14. Denise Dias (2017), FinTech, RegTech and SupTech: What They Mean for Financial Supervision. Toronto Centre Note.
15. See Simone di Castri, Matt Grasser, and Arend Kulenkampff (2020), The "Datastack": A Data and Tech Blueprint for Financial Supervision, Innovation, and the Data Commons.
16. European Banking Authority (EBA) (2018), Recommendations on outsourcing to cloud service providers, EBA/REC/2017/03.
17. Ernst & Young (2018), Data governance: securing the future of financial services - Financial Services Leadership Summit,
18. Fedak (2018), Towards Data Science: 5 Heroic Tools for Natural Language Processing.
19. Financial Conduct Authority (FCA) (2016), Call for input on supporting the development and adopters of RegTech.
20. Financial Conduct Authority (FCA) (2018), Algorithm Trading Compliance in Wholesale Markets.
21. Financial Stability Board (2017), Artificial intelligence and machine learning in financial services: Market developments and financial stability implications.
22. Luciano Floridi (2019) AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations,
23. Gartner (2019), Top Trends on the Gartner Hype Cycle for Artificial Intelligence.
24. Giorgio Gasparri (2019), Frontiers in AI: Risks and Opportunities of RegTech and SupTech Developments.
25. Government of Canada (2019), Data Sovereignty and Public Cloud.
26. Harvard Business Review (2018), A Simple Tool to start making decision with the help of AI.
27. Mustafa Hashmi, Pompeu Casanovas, and Louis de Koker L. (2019), Legal Compliance Through Design: Preliminary Results. CEUR Workshop Proceedings. 59-72.
28. International RegTech Association (IRTA) (2018), Principles for Regtech firms.
29. IGI Global, Disseminator of Knowledge.
30. Julapa, Jagtiani, Larry Wall and Todd Vermilyea (2018), The Roles of Big Data and Machine Learning in Banking Supervision, Banking Perspectives, The Clearing House.org.
31. Mark Balkenende (2018), TheNewStack: The Big data debate – Batch versus Stream Processing.
32. MAS media release (2017), MAS Sets up Data Analytics Group.
33. Nick Maxwell and David Artingstall (2017), The Role of Financial Information-Sharing Partnerships in the Disruption of Crime, Royal United Services Institute for Defence and Security Studies. London. 5.
34. McKinsey (2017), Why you need a digital data architecture to build a sustainable digital business.
35. Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu (2019), Spatiotemporal Anomaly Detection using Deep Learning for Real-time Video Surveillance.
36. Oliver Wyman (2018), Supervising Tomorrow.
37. Quantexa (2017), Chief Data Office Strategy Brief: Data Driven Decisions to Complex Business Problems.
38. Randy Quarles (2018), 2018 Financial Markets Conference – Keynote: A Conversation on Machine Learning in Financial Regulation, Federal Reserve Bank of Atlanta.
39. RUSI (2018), Sharpening the Money-Laundering Risk Picture: How data analytics can support Financial Intelligence, Supervision and Enforcement.
40. SnapLogic (2019), Data pipeline architecture.
41. United Nations, About goAML.
42. UNODC (2017), UNODC's Software for Financial Intelligence Unit.
43. Dirk Andreas Zetzsche, Douglas Arner, Ross Buckley, and Rolf Weber (2019). The Future of Data-Driven Finance and RegTech: Lessons from EU Big Bang II, European Banking Institute Working Paper Series 2019/35,
44. Yeh, Chih-Kuan, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang (2017), Learning deep latent space for multi-label classification.

# Boxes and figures

## Glossary

| WORD/ TERM/ACRONYM | Working definition |
|---|---|
| Algorithm | A step by step method, procedure or formula for solving a problem, based on conducting a sequence of specified actions. A computer program can be viewed as an elaborate algorithm. |
| Alternative data | Social media, voice, biometric and image data sources |
| Artificial Intelligence (AI) | The theory and development of computer systems able to perform tasks that traditionally require human intelligence. AI can ask questions, discover and test hypotheses, and make decisions automatically based on advanced analytics operating on extensive data sets. AI has several sub-categories, such as Machine Learning, Expert Systems, NLP etc. |
| Application Programming Interface (API) | The means by which a piece of computer software communicates with another. A set of rules and specifications followed by software programs to communicate with each other, and an interface between different software programs that facilitates their interaction.<br>Often, these pieces of software are designated as a client and a server: the client initiates a request, for example to create, read, update, or delete a record in a database, while the server completes that action and returns an appropriate response. For example, these requested actions can include: authentication to a service, machine learning predictions on a given data set, web browser / mobile app functionality, transfer of data across a network, and many more. |
| Anomalies and Outliers | "An observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism".[63] Depending on the context, outliers are also termed exceptions, discordant observations, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants.[64] |

[63] D. M. Hawkins (1980), Identification of outliers, Vol. 11, Springer.

[64] V. Chandola, A. Banerjee, V. Kumar (2009), Anomaly detection: A survey, ACM Computing Surveys (CSUR) 41 (3).

| WORD/ TERM/ACRONYM | Working definition |
|---|---|
| Anomaly Detection | A form of statistical analysis in which data or behaviors are flagged that differ significantly from those typically exhibited by the population. This functionality can be automated and/or enhanced by applying machine learning techniques to a historical dataset that contains previously labeled anomalies in order to train a model that can infer future anomalies (i.e. supervised machine learning). |
| Big Data | Refers to the large volume of and/or complicated data sets that can be generated, analyzed and stored using a variety of data elaboration techniques, information systems and digital tools.<br>This capability is driven by the increased availability of structured data, the ability to process unstructured data, increased data storage capabilities and advances in computing power. |
| Big Data Analytics | Analytics focused on, for instance, discovering patterns, correlations, and trends in the data, or customer preferences. It can be based on AI or other technologies. |
| Chatbots[65] | A computer program designed to simulate conversation with human users and is widely used for online customer services at FSPs and beyond. More recent chatbots use the AI sub-category of Machine Learning for improved performance. |
| Cloud Computing | Refers to the use of an online network ("cloud") of hosting processors to increase the scale and flexibility of computing capacity. This model enables convenient on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage facilities, applications and services) that can be rapidly released with minimal management effort or service provider interaction. |
| Cognitive Computing | Refers to simulating human thought processes in a computerized model. Using self-learning algorithms that use data mining, pattern recognition and natural language processing, the computer can mimic the way the human brain works. |

[65] Medium (2019), What is a chatbot.

| WORD/ TERM/ACRONYM | Working definition |
|---|---|
| Cryptography | Address the encryption and decryption of private communications through the internet and computer systems. Cyber cryptographic algorithms are used to transfer electronic data over the internet so that no third-party can read the data. The strength of the code is judged according to four parameters: confidentiality; integrity; non-repudiation & authentication. |
| Data Dictionary | Data dictionary encompasses the concepts, attributes and allowed formats of individual data points, and the relationship between data points. |
| Data Governance | The overall management of the availability, usability, integrity and security of data used in an enterprise. Defines how data is accessed and treated and what is appropriate. The application of policies, people, processes and technology to create a consistent and appropriate use of an organization's data[66]. |
| Data Lake | A storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data. The data structure and requirements are not defined until the data is needed. |
| Data Mining | The process used to turn raw data into useful information by extracting or uncovering patterns and finding anomalies and relationships in large datasets to enable predictions of future trends involving methods at the intersection of machine learning, statistics, and database systems. Data mining is the analysis step of the "knowledge discovery in databases" (KDD) process. |
| Data Management | The implementation (process of moving an idea from concept to reality) of architectures, tools and processes to achieve data governance objectives. An administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users.<br>Key applications include data validation, consolidation, visualization and cloud computing. |
| Data Protection Regulation | EU law promulgated in 2015, and came into force in member states in 2018 |

[66] Data Republic (2019), Understanding the difference between data governance and data management.

| WORD/ TERM/ACRONYM | Working definition |
|---|---|
| Data Sovereignty | The concept that data is subject to a country's laws when it is stored within certain borders. |
| Data Warehouse (DW) | A central repository for all the (integrated) modeled/ structured data collected by an enterprise's various operational systems, be they physical or logical. DW emphasizes the capture of data from diverse sources in a relational database for query and analysis rather than for transaction processing. |
| Decision Trees | A supervised machine learning technique in which classification of data is accomplished through a series of decision criteria, which are automatically determined during the training of the machine learning model. Benefits to this technique include interpretability of the way decisions are made, and ease of incorporation as a feature of a website or mobile app. Examples include rudimentary credit scoring models and chatbot backends. |
| Deep Learning | A type of Machine Learning (ML) where a model of algorithms is applied to neural networks to detect patterns and predict outcomes, thereby mimicking the neuron networks of the human brain. The word "deep" relates to the numerous layers of virtual neurons used to process data. |
| Distributed Ledger Technology (DLT) | DLT such as blockchain are a means of recording information through a distributed ledger, i.e., a repeated digital copy of data at multiple locations. These technologies enable nodes in a network to securely propose, validate and record state changes (or updates) to a synchronized ledger that is distributed across the network's nodes. |
| FinTech | Technologically enabled financial innovation that could result in new business models, applications, processes or products with an associated material effect on financial markets and institutions and the provision of financial services. |
| Information System (IT) Infrastructure | T infrastructure refers to the composite hardware, software, network resources, data centers, facilities and related equipment for the existence, operation and management of an enterprise IT environment. It allows an organization to deliver IT solutions and services to its employees, partners and/or customers and is usually internal to an organization and deployed within owned facilities and relates to its IT Architecture. |

| WORD/ TERM/ACRONYM | Working definition |
|---|---|
| Information System (IT) Architecture | Refers to the conventions, rules, and standards used as technical framework to design or integrate various components of the information system infrastructure. |
| Latent Spaces | A latent space is a low dimensional vector space generatively learned from a high dimensional input data space. The generative nature of a latent space ensures that distribution of classes and not the boundary between classes is learned. |
| Machine Learning | A form of AI, a method of designing a sequence of actions to solve a problem that optimize automatically through experience and with limited or no human intervention by focusing on the giving computers the ability to learn without being specifically programmed for such through hand-inputted codes. It uses a variety of techniques, including neural networks and deep learning.<br>ML has progressed from rules-based (logic-based algorithm) methods to data-based (big data analytic) methods. |
| Mature technology | A technology that has been in use for long enough that most of its initial faults and inherent problems have been removed or reduced by further development. In some contexts, it may also refer to technology that has not seen widespread use, but whose scientific background is well understood. |
| Model Training, Testing and Validation | The process of optimizing a machine learning model toward a goal, using data that's been split in order to prepare it well for future data it has not yet seen. Training data is used to explore what useful features can be extracted, test data is used to tune the architecture of the model, and validation data is used solely to calculate the success metrics (e.g. accuracy, precision, recall, etc.) |
| Natural Language Processing (NLP) | A form of AI, that is the ability of a computer program to understand human language as it is spoken. Is a form of statistical analysis in which words are digitized and language is modeled in order to enable human interaction with computers via text or voice (i.e. conversational user interfaces, or CUIs). Examples: chatbots, voice assistants.<br>Machine learning techniques can be applied such that the computer continues to grow its model of language as it increases its history of interactions. For instance, NLP technology applied to regulation, that uses natural language, enable machine readable regulation. |

| WORD/ TERM/ACRONYM | Working definition |
|---|---|
| Neural networks | Is a series of algorithms that endeavors to recognize underlying relationships in diverse and multiple data sets through a process that mimics the way the human brain operates. |
| Optical Character Recognition | A specific form of computer vision that focuses on the transcription of image data into textual data. Examples include license plate readers, OCR-enabled scanners and mobile apps, passport and other identification card readers, and file conversion tools. |
| RegTech | A new field within the financial services industry that utilizes information technology to enhance regulatory processes. It puts a particular emphasis on regulatory monitoring, reporting and compliance. |
| Relational Database | A set of formally described tables from which data can be accessed or reassembled in many ways without having to reorganize the database tables. The standard user and application programming interface (API) of a relational database is the Structured Query Language (SQL). |
| Robotic Process Automation | An emerging form of business process automation technology based on the notion of software robots |
| Taxonomy (of data) | Taxonomy is the classification of data (e.g., into categories and sub-categories) according to a pre-determined conceptual framework. |
| Semantic Interoperability | Is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems. |
| Sentiment Analysis | A specific form of natural language processing that focuses on the emotional content in a given corpus of text or transcribed speech. Examples include social media data mining to get an edge in securities trading and tagging urgency of customer service data. |
| Situational analysis | Analyzing an organization's internal and external environment to understand the organization's capabilities and business environment. Involves targeting the specific objectives in the business and identifying the factors that support or hinder those objectives. |

| WORD/ TERM/ACRONYM | Working definition |
|---|---|
| Structured data | Data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis. A data structure is a kind of repository that organizes information for that purpose. |
| Supervised learning | The Data mining task of inferring a function from labeled training data. The training data consist of a set of training examples to provide a learning basis for future data processing. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).<br>It is typically done in the context of classification, when we want to map input to output labels, or regression, when we want to map input to a continuous output. Common algorithms in supervised learning include logistic regression, naive bayes, support vector machines, artificial neural networks, and random forests. In both regression and classification, the goal is to find specific relationships or structure in the input data that allow us to effectively produce correct output data. |
| SupTech | The use of technologically enabled innovation by supervisory authorities. |
| Unsupervised Learning | A type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses, in other words it extracts patterns directly from the raw data.<br>The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. Other methods include representation learning and density estimation. In situations where it is either impossible or impractical for a human to propose trends in the data, unsupervised learning can provide initial insights that can then be used to test individual hypotheses. |
| Unstructured data | Data in non-standardized formats that cannot be organized in traditional databases with searchable fields for easy sorting, extraction and analysis. |

## Contact

Email **info@bfaglobal.com**

Website **www.bfablogal.com/R2A**

Twitter **@BFAGlobal | @R2Accelerator**

LinkedIn **https://www.linkedin.com/company/bfaglobal**