Future Data-Driven Regulation

Philip Treleaven[1] and Sally Sfeir-Tait[1,2]
[1]University College London, [2]RegulAItion

Abstract

Banks spend $270 billion per year on compliance; some 10 percent or more of operating costs[1]. The global cost for the financial services sector is probably $trillions. Consequently, governments increasing see 'data-driven' automation as essential for efficient/effective regulation and conferring competitive economic advantage (BoE, 2020).

Data science technologies pioneered in the private sector are ripe for transforming regulation, compliance and supervision. Applications can include regulatory engagement through natural language text and speech *Chatbots* and intelligent assistants; support for Regulators via AI-based Robo-advisors and analytics; securing compliance records using blockchain distributed ledgers; augmenting supervision with learnings for aggregated data in a privacy preserving manner (federated learning); and computer-executable regulations: *codifying* regulators' handbooks to support automation.

This paper is written to help Regulators understand the opportunities of data science for *regulatory automation* (e.g. computer-executable regulatory handbooks); and emerging *regulatory challenges* (e.g. Regulator data access, algorithm interpretability). The paper also includes, for context, introductions to key 'RegTech' data science technologies; such as machine learning, blockchain, digital object identifiers and federated learning. Finally, it presents the UK Government-funded RegNet project, a privacy-preserving data analytics platform and infrastructure.

1. Regulatory Challenges

We are experiencing a 'perfect storm' of data technologies transforming radically the way business, organisations and society operate. Regulation is no exception.

Data-driven automation (pioneered by multinationals like Amazon, Google, Alibaba, Tencent etc.) offers great potential to transform regulation, surveillance and policy-making.

Regulators are facing a myriad of challenges to automate services, balance regulation of emerging technologies but also fascinate innovation, and respond rapidly to unforeseen 'black swan' events, like the impact of Covid-19:

- **Data challenges** – a deluge of data is already overwhelming Regulators. AI tools can help Regulators sift and prioritise compliance reports; and conduct market surveillance.
- **Privacy challenges** – Covid-19 has changed public perception of *privacy* versus *security*; with authorities increasingly using surveillance data from CCTV, mobile phones and ticketing to track infected people and contacts.
- **Resilience challenges** – pressure on Regulators to be flexible, adaptable and responsive to 'black swan' events.
- **Technology challenges** – innovations require new regulations; to counter 'innovation arbitrage', abuses and unintended consequences of technology (responsible AI). Examples include interpretability of machine learning algorithms that self-program but are unable to explain their decision-making. New alternative data sources, such as 'streamed' anonymous credit card transactions used for investments. New unregulated financial instruments such as Binary Options and Initial (crypto) Coin Offerings.

[1] International Banker (2018) https://internationalbanker.com/technology/the-cost-of-compliance/

- **Collaboration challenges**- pressure for Regulators to share sensitive data and intelligence, for example know your customer (KYC); and develop international standards for (financial) regulation across multiple jurisdictions.

To meet these challenges, Regulators need new ‘data-driven’ infrastructures covering registration, authorisation, guidance, supervision, reporting, surveillance and collaboration. Influential infrastructure technologies include: a) digital object identifiers (*DOIs*) for information management; b) *federated learning* for privacy-preserving analytics; and c) computer-executable (*computable*) regulatory handbooks for automation. Figure 1 illustrates challenges and potential technology solutions.

	Challenge	**Technology**
Data	Prioritising compliance PDF reports	AI Natural Language Processing (NLP)
Privacy	Privacy-preserving data access such as AML	Federated Learning
Resilience	Innovation arbitrage	AI-based market surveillance
Technology	Algorithm certification	AI algorithm interpretability
Automation	Compliance advice across multiple jurisdictions	*Computable* regulatory handbooks
Collaboration	International collaboration and standards	Secure national and international Regulator network

Figure 1: Regulatory challenge examples

2. Data-driven Regulation

Automated regulation is crucial to the future success of the financial services industry and especially the rapidly evolving new Financial Technology (FinTech) area. The vision of *Algorithmic Regulation* (Treleaven & Batrinca, 2017), modelled on Algorithmic Trading systems (Treleaven et al, 2013), is to stream compliance reports, social media data and other kinds of surveillance information from different sources to a platform where regulatory data are encoded using distributed ledger technology and automatically analysed using AI machine learning technology (see Figure 2).

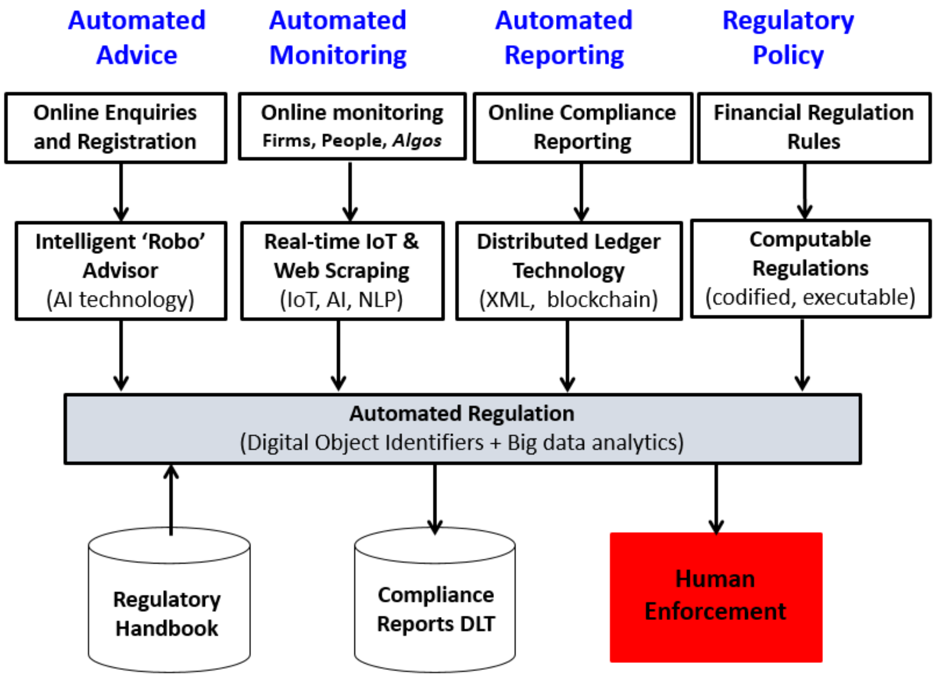
Automated Advice
Automated Monitoring
Automated Reporting
Regulatory Policy
Online Enquiries and Registration
Online monitoring Firms, People, *Algos*
Online Compliance Reporting
Financial Regulation Rules
Intelligent ‘Robo’ Advisor (AI technology)
Real-time IoT & Web Scraping (IoT, AI, NLP)
Distributed Ledger Technology (XML, blockchain)
Computable Regulations (codified, executable)
Automated Regulation (Digital Object Identifiers + Big data analytics)
Regulatory Handbook
Compliance Reports DLT
Human Enforcement

Figure 2: Algorithmic Regulation

As background, we next review the influential data science technologies.

3. Data Science Technologies

For Regulators, the data science technologies provide, on the one hand, unprecedented volumes of data and analytics tools for surveillance, but on the other 'revolutionary' innovations that present new regulatory challenges. To understand the opportunities offered by 'data-driven' regulation, it is necessary to understand the data science technologies contributing to the 'perfect storm'.

Our review divides the data science technologies into a) *data technologies* – the collection and analysis of huge volumes of historic and real-time information (e.g. financial, economic, social media, alternative); b) *algorithm technologies* – new forms of 'statistics', such as machine learning, computational statistics, and complex systems (e.g. deep neural networks, Monte Carlo simulation); c) *analytics technologies* – covering the application of the data technologies (e.g. natural language processing, sentiment analysis); and d) *infrastructure technologies* – providing the infrastructure for information management and automation (e.g. blockchain, computable regulations).

	Regulatory Automation	Regulatory Challenges
Data Technologies	▪ Big data ▪ Internet of Things (IoT) ▪ Chatbots	▪ Data privacy versus sharing ▪ Privacy-preserving access to data ▪ 'data' subversion
Algorithm Technologies	▪ Computational Statistics ▪ Artificial Intelligence & Machine Learning ▪ Complex Systems	▪ Combination of multiple models ▪ Algorithm interpretability ▪ Concept of 'conduct' and 'ethics' in algorithms
Analytics Technologies	▪ Natural Language Processing (NLP) ▪ Sentiment Analysis ▪ Behavioural Analytics ▪ Predictive Analytics ▪ Personalised Avatars	▪ Multiple natural languages ▪ 'deep' understanding of content ▪ Algorithm legal status and certification ▪ Public confidence and acceptability
Infrastructure Technologies	▪ Blockchain ▪ Digital object identifiers ▪ Federated Learning	▪ Regulator collaboration ▪ International regulatory standards

Figure 3: Regulatory Automation versus Regulatory Challenges

Data Technologies

Important data technologies include:

- **Big Data** – very large datasets of historic and real-time financial, economic, social media and alternative data; so complex that traditional data processing application software is inadequate to deal with them (Big data, 2020).
- **Internet of Things (IoT)** - the inter-networking of 'smart' physical devices, vehicles, buildings, etc. that enable these objects to collect and exchange data (Miraz et al, 2015).
- **Chatbots** – data provided by computer programs that simulates human conversation through voice commands or text chats or both (Ahmad et al, 2018); using natural language processing (NLP) and sentiment analysis to understand the conversation.

Algorithm Technologies

Core algorithm technologies (Koshiyama et al, 2020) include:

- **Computational Statistics** - a large class of modern statistical methods that are computationally intensive (e.g. Monte Carlo methods).
- **Artificial intelligence** – AI, machine learning and other systems able to perform tasks normally requiring human intelligence, such as self-programming machine learning (ML) algorithms (e.g. Artificial Neural Networks).
- **Complex Systems** - system featuring a large number of interacting components whose aggregate activity is nonlinear (e.g. Agent-Based systems).

Analytic Technologies

Core analytics technologies include:

- **Backtesting** - assessing the viability of a model (e.g. trading strategy) by discovering how it performed using historical data.
- **Forecasting** – the process of making predictions of trends based on historic data; three basic strategies are qualitative techniques, time series analysis/projection, and causal models.
- **Algorithm interpretability** - in the context of AI and machine learning, the ability of the extent to which you are able to predict what is going to happen, given a change in input or algorithmic parameters. Closely related is explainability, the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms (Carvalho et al, 2019; Tjoa & Guan, 2019).
- **Natural Language Processing (NLP)** – the analysis and synthesis of natural language and speech (Hovy, 2019; NLP, 2020).
- **Sentiment Analysis** – using NLP, statistics, or machine learning methods to extract, identify, or characterize the sentiment content of text or speech (Sentiment, 2020; Text Mining, 2020).
- **Behavioural Analytics** – providing insight into the actions of people (Behaviour, 2020).
- **Predictive Analytics** - extracting information from existing data sets in order to determine patterns and predict future outcomes and trends (Kumar & Garg, 2018; Predict, 2020).
- **Personalised Avatars** - the embodiment of a person customised for interaction with the user with traits and physiognomies that 'resonates' with the user.

Infrastructure Technologies
Core automation technologies include:

- **Blockchain Technologies** – including distributed ledger (DLT) technology, distributed databases that secures, validates and processes transactional data; and smart contracts, a self-executing contract with the terms of the agreement between buyer and seller directly written into lines of code (Treleaven et al, 2017).
- **Digital Object Identifiers (DOI)** – a DOI is an identifier or handle, potentially persistent, used to identify objects uniquely, standardized by an international body (DOI, 2015; DOI, 2020).
- **Federated Learning** –- an infrastructure and machine learning technique that trains an algorithm across multiple decentralized data sources, without direct access to the data (cf. 'taking algorithms to data') (FL, 2020).
- **Computable Legal Rules** – a legal contract or regulation encoded in a computer-understandable notation (associated with a human-readable specification) executed by a computer (Surden, 2014).

We now look at the four categories of technology and their potential impact on regulation.

4. Data Impact on Regulation

Although artificial intelligence gets the media attention, the increasing availability of huge volumes of historic and streamed real-time data (i.e. Big data) is really driving the data revolution. One might say 'data is the new oil'.

When considering the impact of Big data on Regulation, important considerations are firstly the trend of collecting data in volume from ever increasing heterogeneous sources; secondly creating regulatory data models to consolidate the data sources for analytics; and thirdly the changing regulatory landscape, such as the suspension of data privacy to tackle Covid-19.

Generic considerations include:

- **Big data facilities** – consolidating historic and real-time data sets data from increasing set of heterogeneous sources, for example for surveillance.

- **Data characteristics** – often referred to as the 4 V's: a) *volume* – the size; b) *variety* – the heterogeneous nature; c) *velocity* – the speed of generation; and d) *veracity* - the trustworthiness of the data.
- **Data standards** - the rules for specifying data. Standard formats and tagging/typing are required in order to share, exchange, and understand data. Examples range from a regulatory XML (XBRL, 2020), to powerful data exchange formats (cf. FHIR (FHIR, 2019)) for international Regulator collaboration.
- **Data privacy v sharing** – Regulators have highly valuable and sensitive data, and need to comply with privacy legislation (e.g. EU GDPR (GDPR, 2020)). However, collaboration and analytics requires access to ample and relevant data. Data sharing has proven challenging. Privacy-preserving data access (e.g. Federated learning where the algorithm travels to the data, not the data to the algorithm), offers an interesting solution.   .
- **Digital object identifiers** – central to regulatory data are universal DOIs that are *unique*, *persistent* and *resolvable* identifiers for information management (e.g. AML). DOIs are discussed under infrastructure.

Big data facilities
'Data is a team game' with huge volumes coming from an increasing range of sources (Batrinca & Treleaven, 2015). As Regulators automate (cf. financial services), they will utilise a broadening range of data:

- **Business/economic data** – this covers business/economic reports and publications.
- **Transactional data**- - data generated from all the daily transactions that take place both online and offline. This includes business invoices and payment orders etc., and also
- **Social data** – social media data including Twitter, Facebook, Instagram, video uploads, blogs etc.
- **Online conferencing** –services including Skype, ZOOM, TEAMS and WebEx etc. used increasingly for remote working; a new way of life post-coronavirus.
- **Machine data** – data generated by IoT devices, industrial equipment, and sensors installed in CCTV cameras, machinery, etc.
- **Alternative data** – a 'catch all'; information gathered from non-traditional information sources, such as financial transactions, mobile devices, satellites, public records, and the Internet.

Data Privacy, Access, Sharing & Collaboration
Although data privacy has to date dominated public debate, the deployment of *social* data to combat Covid-19 in China, South Korea and Singapore are game changers. For example, CCTV cameras' face recognition and public temperature monitors, named passenger seat location on trains/planes, mobile phone location and tracking, and APPs alerting people to proximity of Covid-19 'carriers'.

As countries start to rebuild devastated economies, and put in place procedures for the next pandemic, the impact on compliance and regulation data and privacy rules are likely to be profound.

Data Regulation
In summarising the impact of 'Big data' on regulation automation:

- **Regulatory data** – Regulators are consolidating historic and real-time data sets data from increasing sets of heterogeneous sources (e.g. compliance, business, social, alternative).
- **Regulatory standards** – Regulators need 'data model' standards (cf. XML, FHIR) for management of information from firms to regulators, within a Regulator and between Regulators.

- **Regulatory collaboration** – collaboration with privacy-preserving data access is required at a number of levels: within a Regulator, between international Regulators in a specific sector (e.g. Finance), and between all national Regulators (e.g. Finance, Healthcare, Telecoms, Legal services) in one country.

Next, we review algorithms, such as machine learning, and their impact on Regulation.

5. Algorithms[2] impact on Regulation

The terms algorithm, artificial intelligence (AI) and machine learning are used interchangeably. However, data science algorithms cover three broad domains: Computational Statistics (e.g. Monte Carlo methods), Artificial Intelligence (e.g. Artificial Neural Networks), and Complex Systems (e.g. Agent-Based systems). See Figure 4.

- **Computational Statistics** - computationally intensive statistical methods.
- **AI Algorithms** - mimicking a new form of human learning, reasoning, knowledge, and decision-making
  - Knowledge or rule-based systems
  - Evolutionary algorithms
  - Machine learning
- **Complex Systems** - system featuring a large number of interacting components whose aggregate activity is nonlinear.

Figure 4: Algorithm taxonomy (Koshiyama et al, 2020)

Computational Statistics

Computational Statistics models refers to computationally intensive statistical methods including Resampling methods (e.g., Bootstrap and Cross-Validation), Monte Carlo methods, Kernel Density estimation and other Semi and Non-Parametric methods, and Generalized Additive Models.

AI and Machine Learning

AI algorithms are a continuum of epistemological models spans three main communities:

- **Knowledge-based** or heuristic algorithms (e.g. rule-based) - where knowledge is explicitly represented as ontologies or IF-THEN rules rather than implicitly via code;
- **Evolutionary** or metaheuristics algorithms - a family of algorithms for global optimization inspired by biological evolution (e.g. Genetic Algorithms, Genetic Programming, etc.); and
- **Machine Learning** algorithms - a type of AI program with the ability to learn without explicit programming, and can change when exposed to new data; mainly comprising *Supervised*, *Unsupervised*, and *Reinforcement Learning*.

Complex Systems

Lastly, a complex system is any system featuring a large number of interacting components (e.g. agents, processes, etc.) whose aggregate activity is nonlinear (not derivable from the summations of the activity of individual components). Examples include Cellular automata, Agent-based models, Network-based models, and Multi-Agent systems.

With regard to regulation, machine-learning algorithms are having the greatest impact.

Machine Learning

The most impactful are ML algorithms; broadly a combination of the *classic trio* of Supervised, Unsupervised and Reinforcement Learning, with the *disruptors*: Deep Learning, Adversarial Learning, Transfer and Meta Learning. This interaction constantly yields new models (e.g., Long Short-Term Memory, Generative Adversarial Networks) and applications (e.g., Natural Language Processing, Object Recognition, Forecasting etc.).

[2] For a comprehensive review see Koshiyama et al, (2020) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527511

Supervised, Unsupervised, Reinforcement ML
Classic ML subdivides into:

- **Supervised learning** - learns or infers a pattern (function) from labelled training data consisting of a set of training examples.
- **Unsupervised learning** - learns or infers a pattern in a data set with no pre-existing labelled training data and with a minimum of human supervision.
- **Reinforcement learning** - enables an algorithm using a system of reward and punishment to learn by trial and error using feedback from its own actions and experiences.

Deep Learning, Adversarial Learning, Transfer/Meta Learning
Next, the new forms of learning 'disrupting' classic ML subdivide into:

- **Deep Learning** - deep learning algorithms attempt to model high-level abstractions in data by using multiple processing layers, with complex structures or otherwise, composed of multiple non-linear transformations (Goodfellow, et al, 2016).
- **Adversarial Learning** - adversarial machine learning is a technique employed in the field of machine learning which attempts to 'fool' models through malicious input (Huang et al, 2011).
- **Transfer/Meta Learning** – these two learning paradigms are tightly connected, as their main goal is to encapsulate knowledge learned across many tasks and transfer it to new, unseen ones. In Transfer learning, knowledge is transfer from a trained model; in Meta Learning the learning method (learning rule, initialization, architecture etc.) is abstracted and shared across tasks (Flennerhag et al, 2018).

The combination of these classic and disruptive ML yields powerful new algorithms such as *Long-Short Term Memory* (LSTMs) – a type of deep recurrent neural network capable of learning arbitrary long-term dependencies; and *Generative Adversarial Networks (*GANs*)* – an architecture comprised of two networks, pitting one against the other (thus the 'adversarial'. Algorithms are becoming increasingly complex.

Algorithm Regulation – *Responsible AI*
New AI algorithms are constantly emerging, with each 'strain' mimicking a new form of human learning, reasoning, knowledge, and decision-making. The current main disrupting forms of learning include Deep Learning, Adversarial Learning, Transfer and Meta Learning. Since ML algorithms effectively self-program and evolve dynamically, financial institutions and Regulators are becoming increasingly concerned with ensuring there remains a modicum of human control, focusing on Algorithmic *Interpretability/Explainability, Robustness* and *Legality*. From a regulatory perspective, the concern is that, in the future, an ecology of (trading) algorithms across different institutions may 'conspire' and become unintentionally *fraudulent* (cf. LIBOR) or subject to subversion through compromised datasets (e.g. Microsoft Tay). New and unique forms of systemic risks can also emerge, potentially coming from excessive algorithmic complexity.

In summarising, the impact of 'algorithms' on automated regulation:

- **Algorithm innovation** – new forms of machine learning have produced powerful models (e.g., Long Short-Term Memory, Generative Adversarial Networks); and applications' innovations (e.g., Natural Language Processing, Adversarial examples, Deep Fakes, etc.).
- **Algorithm interpretability** – given ML algorithms self-programme, companies and Regulators are increasingly concerned about understanding the decisions made by algorithms and the extent to which the internal mechanics of an AI system are explainable in human terms.
- **Algorithm conduct** – although 'deep learning' is a popular topic in AI, ML algorithms have yet to embody any (human-like) system of values; concepts of conduct, legality and ethics.

- **Algorithm certification** – do algorithms need regulation? Two issues debated are firstly the legal status of algorithms (i.e. should algorithms be 'artificial persons' in law); secondly should a regulatory body be established to certify algorithms?

6. Analytics impact on Regulation

Analytics is the application of computational statistics, AI and complex systems to analysis large and varied data sets to uncover hidden patterns to help make informed decisions. This includes natural language programming (NLP), sentiment analyse, plus behavioural and predictive analytics.

Natural Language Processing

NLP is the understanding of humans' natural language through text or speech. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages. Natural Language processing is a difficult problem in computer science, due to the ambiguous nature of the human language. It requires understanding both the syntax and semantics; the words and how the concepts are connected to deliver the intended message. For a Regulator, NLP can automate their call-centre function, analyse compliance reports and monitor surveillance.

Sentiment Analysis

In partnership with NLP, sentiment analysis is the computational process of understanding the meaning and interpretation of words and sentence structure in a piece of text or speech. For example, identifying and categorizing opinions expressed to determine whether the meaning conveyed is positive, negative, or neutral.

Techniques cover a) syntax analysis for segmenting and tagging words and units, including Word segmentation, Parsing, Lemmatization, Morphological segmentation etc.; and semantic analysis for understanding meaning, including Named entity recognition (NER), Word sense disambiguation, Natural language generation etc. (Sentiment, 2020).

Behavioural/Predictive Analytics

Closely related to NLP and sentiment analysis are behavioural and predictive analytics that provide insight into the actions of people and organisations (Behavioural, 2020). Behavioural analytics centres on understanding how people act and why, providing insight on decision-making. Predictive analytics centres on 'forecasts' of what might happen in the future with an acceptable level of reliability, and includes what-if scenarios and risk assessment. A Regulator might *score* firms and individuals on their propensity for compliance breaches using Behavioural/predicting analytics.

Regulatory Analytics

In summary, a 'deluge' of data is already overwhelming Regulators. What AI-based analytics can provide is powerful tools to do the 'heavy lifting'.

The impact of 'analytics' on automated regulation (see Figure 2) spans:

- **Automated guidance** – Chatbots and NPL interfaced to a Regulator's handbook, can provide automated call-centre functions, and automation of the registration and authorisation process for applicants.
- **Automated monitoring** – Regulators need to 'harvest' online dataset from heterogeneous sources (e.g. compliance, business, social, alternative) to build surveillance profiles on companies, individuals and their associates. Analytics can be used for surveillance to alert Regulators to regulatory breaches (e.g. AML), market abuses (e.g. insider dealing) and emerging problems (e.g. Libor, PPI).
- **Automated reporting** – this addresses the burden of compliance reporting, both for firms and Regulators. For example, using analytics to sift compliance reports and prioritise them for Regulators. Secondly, using analytics to manage Regulator-firm interactions. Thirdly, doing real-time compliance and governance reporting with associated analytics.

- **Automated prediction** – for the future (inspired by the film *Minority Report*) predictive analytics might identify firms, individuals and products that require heightened surveillance (e.g. Binary Options, Initial Coin Offerings).

Lastly, the key to unlocking regulatory analytics (discussed below) is *computable* regulatory handbooks. Making handbooks computer-readable, verifiable, and computer-executable.

7. Infrastructure impact on Regulation

New infrastructures are required for regulatory automation, standards and international collaboration. Contributing technologies include: a) blockchain technology for secure storage; b) digital object identifiers (DOIs) for information management; c) federated learning for privacy-preserving analytics; and d) computer-executable (*computable*) regulations for automating regulatory handbooks.

Blockchain Technology

Elements of blockchain technology, originally conceived for Bitcoin, have far-reaching potential in other areas (DLT, 2016), equally for regulation. The core technologies are:

- **Distributed Ledger Technology (DLT)** – a decentralized database where transactions are kept in a shared, replicated, synchronized, distributed bookkeeping record, which is secured by cryptographic sealing. The key distinction between 'distributed ledgers' and 'distributed databases' is that nodes of the distributed ledger cannot/do not trust other nodes – and so must independently verify transactions before applying them.

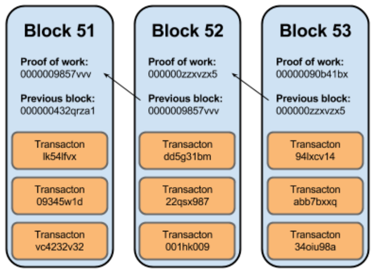

Figure 5: Blockchain Distributed Ledge

- **Smart Contracts/Regulation** - are simply the rules, possibly computer programs that attempt to codify contracts and regulation with the intent that the records managed by the distributed ledger are authoritative with respect to the existence, status and evolution of the underlying legal rights and obligations they represent. Smart regulation technology has the potential to automate regulations, laws and statutes (e.g. www.fca.org.uk/publication/discussion/digital-regulatory-reporting-pilot-phase-2-viability-assessment.pdf).

Digital Object Identifiers

DOIs are set to have a major impact for information management. As context, unique identifiers are central to the management of digital (e.g. ISBN, URLs) and physical objects (e.g. telephone numbers, barcodes, US SSN, IP addresses). Digital Object Identifiers (DOI) (DOI, 2020), pioneered by Bob Khan co-inventor of the Internet, unifies management of information and digital objects across the Internet, with the Handle system (Handle, 2019) resolving the physical location of an object. DOIs originated in the requirement to define uniquely publications, make them persistent, and accessible via the Internet. DOIs are *unique* identifiers used for a class of objects; *persistent* with a long-lasting reference to a digital object; and *resolvable* by identifying the location of a digital object. The *DOI Handle* System (www.handle.net) provides secure name resolution over the Internet, designed to enable a broad set of communities to use the technology to identify digital content independent of location.

For Regulators, DOIs offer a uniform schema for uniquely identifying firms, individuals and associates, together with compliance and surveillance reports. They underpin privacy-preserving data access and analytics, and collaboration across national and international regulators.

Federated Learning

Also pivotal is federated learning (FL), a collaborative machine learning technique that trains an algorithm across multiple decentralized data sets, without the need to centralise data and hence compromise privacy. Traditional machine learning models require centralizing of the training data and analytics on one machine, data centre or Cloud. The current FL focus is decentralised neural network models, such as Google's TensorFlow Federated (www.tensorflow.org/federated), trained across millions of mobile devices.

Computable Legal Rules

Computer-executable legal rules is a general term used for legal specifications that a computer can read, understand, verify and execute. They span executable legal contracts, regulations and statutes. Numerous research areas and international consortia are investigating executable legal specifications. These include international legal contract consortia, such as the Oasis LegalXML group (www.legalxml.org), CEN MetaLex (www.metalex.eu), Accord Project (www.accordproject.org) and Stanford's CodeX Project (http://compk.stanford.edu/). It also includes blockchain *smart contracts*; more accurately smart transactions, where the terms of the agreement between buyer and seller being directly written into lines of code.

For Regulators, a *computable* regulatory handbook is pivotal for automation. Making a handbook readable, understandable, verifiable and executable will allow other AI-based applications to be built on top.

Regulation Infrastructures

In summary, a major challenge facing a Regulator, group of national/international Regulators, or any compliance department is managing privacy-preserving access to their data and collaborative analytics. The contributory infrastructure technologies, as discussed are blockchain, universal DOIs, federated learning and *computable* regulatory handbooks.

What is required is a 'RegNet' – *a regulatory Internet for data*; unifying information management amongst Regulators across the Internet (e.g. digital object identifiers), privacy-preserving data access and analytics (e.g. federated learning), and digital automation (e.g. computable regulations). See Figure 6. Arguably each Regulator requires:

- **Regulator infrastructure** – an internal privacy-preserving data and analytics infrastructure together with a computable regulatory handbook.
- **National infrastructure** – at the next level countries are starting to integrate their Regulators. An example is the UK Government Better Regulation Executive (www.gov.uk/government/groups/better-regulation-executive). The solution is for all Regulators to use common internal platforms that supports collaboration at a national level.
- **International infrastructure** – already there is close collaboration between the financial Regulators of the USA, UK, Canada, Australia etc. on regulatory standards. A secure trans-national infrastructure is an obvious next step.

8. UK Government-funded RegNet Project

As discussed, Regulators require new infrastructures for regulatory automation, standards and collaboration. An illustration is the Innovate UK funded RegNet project whose principal role is a privacy-preserving data analytics platform, for Services sector companies (e.g. Insurance, Legal Services); not just regulation. Besides privacy-preserving analytics, it is also a secure collaboration infrastructure (see Figure 6).

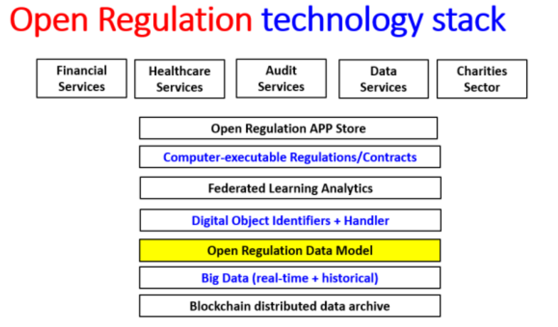
Open Regulation technology stack
Financial Services
Healthcare Services
Audit Services
Data Services
Charities Sector
Open Regulation APP Store
Computer-executable Regulations/Contracts
Federated Learning Analytics
Digital Object Identifiers + Handler
Open Regulation Data Model
Big Data (real-time + historical)
Blockchain distributed data archive

Figure 6: RegNet Technology Stack

For Regulators, the Blockchain + DOI layers support information management both within a single Regulator and between Regulators. The federated learning layer support privacy-preserving analytics across distributed regulatory data sets. The *computable* regulations/contracts support automation by encoding a regulatory handbook. Finally, the top layer can even support APPs for an Open Regulation initiative (cf. UK Open Banking).

9. Future Data-driven Regulation

In conclusion, governments increasingly see 'data-driven' automation as essential for efficient and effective regulation.

Data-driven Regulator

Data-driven automation (pioneered by multinationals like Amazon, Google, Alibaba, Tencent etc.) offer role models for transforming regulation, compliance, market surveillance and policy-making. The unification of data science technologies enables *regulatory automation*; and handling the emerging *regulatory challenges*. Pivotal is a regulatory infrastructure like RegNet that supports privacy-preserving analytics and collaboration.

National Digital Infrastructure

Given the overlap of Regulators, a shared national regulatory infrastructure is also beneficial. As illustrated by Estonia's *e-Estonia* government initiative, a national infrastructure can have a profound impact on business, society and innovation. With the success of the UK Open Banking initiative, the Government might consider an Open Regulation initiative.

International Collaboration and Standards

Lastly, given the huge cost of compliance and regulation there is growing pressure on jurisdictions to collaborate and develop international standards for regulations, handbooks and reporting. Made more urgent by 'Black swan' events, like the economic impact of Covid-19, which have a profound impact on society, government and regulation.

10. Authors

Philip Treleaven is Director of the UK Centre for Financial Computing (www.financialcomputing.org) and Professor of Computing at UCL. The Centre has over 70 PhD students and 600 associated Masters' students working on AI and computational finance. It collaborates with the major Regulators and financial institutions.

Sally Sfeir-Tait is an Honorary Senior Research Fellow in UCL Computer Science and Chief Executive Officer, RegulAItion Ltd. Sally coordinates UCL's RegTech collaborations and projects.

11. References

(Ahmad et al , 2018) Ahmad, N., Che, M., Zainal, A., Rauf, M., Adnan, Z., Review of Chatbots Design Techniques, International Journal of Computer Applications August 2018

(Batrinca & Treleaven, 2015) Batrinca, B., Treleaven, P., Social Media Analytics: A Survey of Techniques, Tools and Platforms, AI & SOCIETY, February 2015, Volume 30, Issue 1, pp 89-116.

(Behaviour, 2020) Behavioural Analytics, Wikipedia, https://en.wikipedia.org/wiki/Behavioral_analytics

(BoE, 2020) Bank of England, Transforming data collection from the UK financial sector, www.bankofengland.co.uk/paper/2020/transforming-data-collection-from-the-uk-financial-sector

(Carvalho et al, 2019) Carvalho, D., Pereira, E., Cardoso, J., Machine Learning Interpretability: A Survey on Methods and Metrics, Electronics 2019, 8, 832, www.mdpi.com/2079-9292/8/8/832

(DigitalTwin, 2020) Digital Twin, Wikipedia, https://en.wikipedia.org/wiki/Digital_twin

(DLT, 2016) Distributed Ledger Technology: beyond blockchain, UK Government Chief Scientific Advisor, www.gov.uk/government/uploads/system/uploads/attachment_data/file/492972/gs-16-1-distributed-ledger-technology.pdf

(DOI, 2015) Digital Identifier Handbook, www.doi.org/doi_handbook/1_Introduction.html

(DOI, 2020) Digital Object Identifiers, Wikipedia, https://en.wikipedia.org/wiki/Digital_object_identifier

(FHIR, 2019) Fast Healthcare Interoperability Resources, www.hl7.org/fhir/summary.html

(FL, 2019)Federated Learning, Wikipedia, https://en.wikipedia.org/wiki/Federated_learning

(Flennerhag, 2018) Flennerhag, S., Moreno, P. G., Lawrence, N. D., & Damianou, A. Transferring knowledge across learning processes. arXiv preprint arXiv:1812.01054.

(GDPR, 2020) What is GDPR? The summary guide to GDPR compliance in the UK, www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018

(Goodfellow, et al, 2016) Goodfellow, I., Bengio, Y., & Courville, A. Deep learning. MIT press.

(Handle, 2019) Fact Sheet DOI System and the Handle System, www.doi.org/factsheets/DOIHandle.html

(Hovy, 2019) Hovy, E., Natural Language Processing: A Brief Review, www.isi.edu/natural-language/teaching/cs544/spring12/current/notes/setup1-NLP-ov.pdf

(Huang et al, 2011) Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B., & Tygar, J., Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (pp. 43-58). ACM.

(Jones & Knaack, 2019) Jones, E., Knaack, P., Global Financial Regulation: Shortcomings and Reform Options, www.researchgate.net/publication/331173347_Global_Financial_Regulation_Shortcomings_and_Reform_Options

(Koshiyama et al, 2020) Koshiyama, A., Firoozye, N., Treleaven, P., Algorithms in Future Capital Markets, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527511

(Kumar, 2018) Kumar, V., Garg, M., Predictive Analytics: A Review of Trends and Techniques, International Journal of Computer Applications, https://www.researchgate.net/publication/326435728

(Miraz et al, 2015) Miraz, M., Ali, M., Excell, P., Picking, R., A review on Internet of Things (IoT), Internet of Everything (IoE) and Internet of Nano Things (IoNT), http://dx.doi.org/10.1109/ITechA.2015.7317398

(NLP, 2020) Natural Language Processing, Wikipedia, https://en.wikipedia.org/wiki/Natural_language_processing

(Predict, 2020) Predictive Analytics, Wikipedia, https://en.wikipedia.org/wiki/Predictive_analytics

(Sentiment, 2020) Sentiment Analysis, MonkeyLearn, https://monkeylearn.com/sentiment-analysis/

(Sugden, 2014) Surden, H., Computable Contracts Explained, http://www.harrysurden.com/wordpress/archives/203

(Text Mining, 2020)Text Mining, Wikipedia, https://en.wikipedia.org/wiki/Text_mining

(Tjoa & Guan, 2019) Tjoa., E., Guan, C., A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI Erico, https://arxiv.org/ftp/arxiv/papers/1907/1907.07374.pdf

(Treleaven et al, 2013) Treleaven, P., Galas, M., Lalchand, V., Algorithmic Trading Review, Communications of the ACM, Vol. 56 No. 11, Pages 76-85.

(Treleaven et al, 2017) Treleaven, P., Gendal Brown, R.,Yang, D., Blockchain Technology in Finance, IEEE COMPUTER, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8048631

(Treleaven & Batrinca, 2017) Treleaven, P., Batrinca, B., Algorithmic regulation: automating financial compliance monitoring and regulation using AI and blockchain, Journal of Digital Transformation, www.capco.com/Capco-Institute/Journal-45-Transformation/Algorithmic-regulation

(XBRL, 2020) eXensible Business Reporting Language, https://en.wikipedia.org/wiki/XBRL

(XML, 2020) Introduction to XML, W3 Schools, www.w3schools.com/xml/xml_whatis.asp