sustainability

MDPI

Review

Leveraging Big Data Tools and Technologies: Addressing the Challenges of the Water Quality Sector

Juan Manuel Ponce Romero, Stephen H. Hallett * and Simon Jude

School of Water, Energy and Environment, Cranfield University, Cranfield MK43 0AL, UK; j.ponceromero@cranfield.ac.uk (J.M.P.R.); s.jude@cranfield.ac.uk (S.J.)
* Correspondence: s.hallett@cranfield.ac.uk; Tel.: +44-(0)786-1234-750111

Received: 1 September 2017; Accepted: 20 November 2017; Published: 23 November 2017

Abstract: The water utility sector is subject to stringent legislation, seeking to address both the evolution of practices within the chemical/pharmaceutical industry, and the safeguarding of environmental protection, and which is informed by stakeholder views. Growing public environmental awareness is balanced by fair apportionment of liability within-sector. This highly complex and dynamic context poses challenges for water utilities seeking to manage the diverse chemicals arising from disparate sources reaching Wastewater Treatment Plants, including residential, commercial, and industrial points of origin, and diffuse sources including agricultural and hard surface water run-off. Effluents contain broad ranges of organic and inorganic compounds, herbicides, pesticides, phosphorus, pharmaceuticals, and chemicals of emerging concern. These potential pollutants can be in dissolved form, or arise in association with organic matter, the associated risks posing significant environmental challenges. This paper examines how the adoption of new Big Data tools and computational technologies can offer great advantage to the water utility sector in addressing this challenge. Big Data approaches facilitate improved understanding and insight of these challenges, by industry, regulator, and public alike. We discuss how Big Data approaches can be used to improve the outputs of tools currently in use by the water industry, such as SAGIS (Source Apportionment GIS system), helping to reveal new relationships between chemicals, the environment, and human health, and in turn provide better understanding of contaminants in wastewater (origin, pathways, and persistence). We highlight how the sector can draw upon Big Data tools to add value to legacy datasets, such as the Chemicals Investigation Programme in the UK, combined with contemporary data sources, extending the lifespan of data, focusing monitoring strategies, and helping users adapt and plan more efficiently. Despite the relative maturity of the Big Data technology and adoption in many wider sectors, uptake within the water utility sector remains limited to date. By contrast with the extensive range of applications of Big Data in in other sectors, highlight is drawn to how improvements are required to achieve the full potential of this technology in the water utility industry.

Keywords: water; pollutants; Chemicals Investigation Programme; Water Framework Directive; SAGIS; environmental risk

1. Introduction

The context and environment surrounding the water utility sector is complex and dynamic. Continuing evolution of the chemical/pharmaceutical industry, diversity of stakeholders, the continuous changes in the political context, and economic fluctuation represent just some of the key drivers. Additionally, there is a growing public awareness of environmental issues, leading to a perceived need for a fair apportionment of liability within the water industry, which in turn has led to

more stringent legislation and enforcement [1–3]. This poses serious challenges within water utilities, especially when addressing potential pollutants reaching the Wastewater Treatment Plants (WWTPs), with chemicals in the waste stream being diverse in nature, and disparate in sources.

The sources of effluents treated by WWTPs include residential, commercial and industrial points of origin, as well as diffuse sources such as agricultural and hard surface (e.g., urban, amenity and industrial) runoff. Consequently, effluents arising contain a wide range of organic and inorganic pollutants, such as herbicides, pesticides, phosphorous, pharmaceuticals, and a new generation of chemicals of emerging concern (e.g., PBDE (PolyBrominated Diphenyl Ethers), in use as flame retardants) [3–5]. These compounds can be present either in dissolved form, or in association with organic matter [6], and the associated risks pose significant environmental and removal challenges.

A key concern facing water utilities and regulators is therefore the understanding of the complex sources, pathways, and behaviour of such contaminants arriving at the WWTPs, in order to support evidence-based decision making and investment planning. This has resulted in industry-wide initiatives, such as the Chemicals Investigation Programme (CIP) in the UK [7]. The CIP seeks to provide a baseline for monitoring the chemicals considered likely to reach water treatment plants, thus providing insight into the behaviour, sources, and control measures applicable to them. Such approaches draw together a broad spectrum of data sources and types, used to inform the modelling approaches.

This paper examines the challenges associated with the applicability of existing computational analytical approaches, applied to such data sources, which may currently be limiting the value of the information that can be derived. The term Big Data is taken to relate to a collection of hardware and software tools designed specifically to address complexity inherent in data. These tools are able to manipulate concurrently substantive datasets characterized by their different volumes, varieties, velocities, and veracities, which are difficult to process and manage using traditional data management techniques. The need to process data with one or more of these characteristics forms the core of Big Data science, colloquially referred to as the "4 Vs" [8]. We consider how adoption of a new generation of Big Data tools and techniques, specifically designed to be able to handle complex datasets, could prove beneficial within the water utility sector. This paper also discusses how Big Data approaches can help to establish new relationships between chemicals, the environment, and human health, providing a better understanding of the contaminants in wastewater (origin, pathways, and persistence). We highlight how the water industry is seeking to utilize Big Data approaches to add value to, and extend the utility of, existing legacy datasets, combined with other contemporary data sources, so extending the lifespan of extant data held. We focus on investment planning and monitoring strategies, helping users adapt and plan more efficiently, for example to changes in water legislation, to the shared benefit of all stakeholders.

From a review of the literature, only a limited reported usage of advanced Big Data techniques was apparent within the water utility sector, with few examples reported of direct use [9–11], and with these being limited to concerns over the storage and handling alone of large volumes of data. The integration of the other tenets of Big Data, such as the integration of data sources with different levels of variety, velocity, and veracity [8,12], or the utilization and analysis of unstructured data, seems to remain relatively unexplored in the industry to-date. Due to both the nature and relative novelty of the technologies considered, many of the references used in this article arise from practitioner and grey literature.

2. Big Data and the Water Sector

Big Data is becoming ubiquitous today, offering a wide range of opportunities and innovation, in addition to improvements in analytical insights. Big Data techniques embrace and extend traditional informatics approaches, allowing a level of data processing that would traditionally have been unachievable. Rising complexity in new sources of data derive in part from associated reduction in

the costs of both the generation and storage of data over the last decade, which in turn has led to the creation, and growing ease of access to substantive datasets [13].

Modern information, generated and handled by organizations and corporations cannot be compared with the period before the turn of the 21st Century. In 2012, Oracle predicted a global increase of data generation of 40% per year, growing exponentially from an approximated 0.2 ZB of data held in 2008 to an estimated 44 ZB in 2020 [14–16]. To emphasize this shift in data generation, Eric Schmidt, the then Executive Chairman of Google, noted: "Every two days, we create as much information as we did from the dawn of civilization up until 2003" (Technomy Conference, Lake Tahoe, CA, USA, 4 August 2010). The growing willingness of private and public-sector organizations to grant access to the data they hold, encouraging its widespread use, in many occasions free of charge, is facilitating access to an amount of information never possible in the past, for example open data initiatives via web portals such as $_{\text{data.gov}}$ (USA) and $_{\text{data.gov.uk}}$ (UK). However, the increase in volume and sources has been accompanied at the same time by an increasing complexity impacting on the efficient use of this data. Big Data is still an emerging science that is in constant evolution, but its successful use has greatly accelerated the processing of data and improved decision making. It already has proven value in Medicine [17,18], Natural Sciences [19,20], Engineering [21], Social Sciences [22], and Legislative [9,23] fields. Examples may be drawn from the visual analysis of air quality [19], the cost-effective allocation of CO_2 emissions [24], the evaluation and identification of cost-effective acid mine drainage management [25], the forecasting of risk in criminal justice decision [26], the reduction of readmission risk in hospital patients [27], and the improvement of city governments services [28].

The water utility sector has all the necessary components in place for a widespread application of Big Data technology, from the control and improvement of potable water quality to the management of wastewaters. The sector has access to vast stores of data concerning water quality, e.g., dissolved substances, climate, consumer preference information and usage patterns, and catchment-based land use activities to name a few. Being able to effectively manage, transform, fuse, and analyse these datasets as a whole could result in a competitive advantage and a decrease in the uncertainty surrounding decision making. However, the level of uptake seems to be lower than in other sectors, where Big Data and machine learning are now used extensively in supporting evidence-based decision making.

2.1. Development and Expansion of the Big Data Technology

Big Data technology receives considerable public interest. One implication is that the computational technologies available are being developed in a sustained and prolific manner. Big Data techniques have significantly reduced the cost and infrastructure requirements arising due to the development of implicit techniques, allowing the horizontal scalability of resources. This horizontal scalability consists of the distribution of computational processes along different processing "threads", which are allocated to multiple machines working collectively and simultaneously in parallel. The adoption of concurrent, cheaper and less technologically advanced hardware, compared with traditional single-thread processing, lowers the time needed for the processing of large datasets. Data storage has also become more robust, improving and facilitating redundant storage, preventing data loss, whilst allowing fault-tolerance and increased overall reliability [12,17].

The use of computational infrastructure on demand, also known as the "elastic cloud" or "infrastructure as a service" (IaaS), is another technological approach that has permitted the rapid expansion of Big Data applications. This relies on the increase or decrease of the size of an allocated cluster, storage, or processing capability, in accordance with the immediate needs of the tasks being undertaken. This drives down significantly the cost of the whole infrastructure, as it permits the allocation of resources only "as and when" necessary [29,30].

Big Data has generated enormous interest within the private sector, leading to the emergence of a wide range of products and technologies. In addition, there is also a noted shift towards "service

as a business" (SaaB), where companies share part, or all of the entire codebase of their products, or allow the use of their software free of charge, making profit solely from the services associated with the software [31]. Hosting their product, training, consulting, or technical help then become some of the services offered to generate revenue. This business model has been used for years by companies such as SUSE [32] or RedHat [33], which have been offering open-source Linux distributions free to use. The reason behind this reside in an awareness of the value of the contributions made by the community, which in turn help accelerate the development of the products, and also help reveal the early detection and patching of software errors and vulnerabilities.

Other factors, such as the popularization of the IoT (Internet of Things) and the growing interest in the scientific applications of Big Data have also contributed to the increased pace of development of the technology. A vast community, focused on the development of free and open-sourced tools, has arisen around this field, boosting exponentially the development of the technology. SCADA-based control systems are an area that have been widely exploited by the water sector over many years, serving the operation of complex plant systems [34,35]; such systems add to the body of available data.

Together, these circumstances have generated a plethora of linked software technologies, all in constant evolution. There are a wide range of options available for achieving given goals, with solutions appearing and being withdrawn rapidly, or being fused with other existing software tools. This is a sign not only of the high interest in this technology, but also of its high complexity and rapid evolution. A wide range of alternatives is available in both free and paid-for versions, with a substantive documentation accessible. Figure 1 illustrates this complexity, displaying a small selection of the different tools used in contemporary Big Data analytics. Many of the tools noted, including Cassandra and Hbase, belong to the *Apache Software Foundation*, which provide support to open-source software projects. The community support of such projects allows for rapid innovation, and highlights the strong interest in this field.

2.2. Current Use in the Water Sector

Searches of literature in scientific databases (Scopus®, Web of Science®) reveals limited documented uses of Big Data approaches in the water industry. The implementation of smart meters and smart sensors being one area generating large amounts of near-real-time information, permitting water companies to deliver better services and improved infrastructure, drawing upon Big Data techniques [36–41]. However, the application of these technologies across other business activities, such as the better characterization of water effluents, or the better understanding of the water quality, seem as yet relatively unexplored. This was highlighted by the low number of results obtained during the search of certain combination of keywords in peer reviewed publications (Table 1). Relevancy was established in each case through consultation of the article metadata and content where made available. In the specific case of the use of Big Data in the evaluation of the risks posed by the presence of certain chemicals in water, there seem to be no relevant publications to date. The cause for the low return in results can be either due to the non-existence of this type of use within the industry, or due to corporate unwillingness to disclose such information.

In seeking to ascertain those publications in academic journals, specific keywords were applied in Scopus® and Web of Science®, and the limited set of results recorded. The same search terms were also used in Google Scholar®, obtaining thousands of results, due to the inclusion of other kinds of publications (grey literature) and websites in addition to the exclusively peer-reviewed academic journals. Nonetheless, the relative novelty of this science makes the grey literature an important source as it can provide an insight on the enormous industrial interest in this field. It is observed that, due to the nature of the evolution of this technology, the number of publications in grey literature is greater than those appearing in scientific publications.

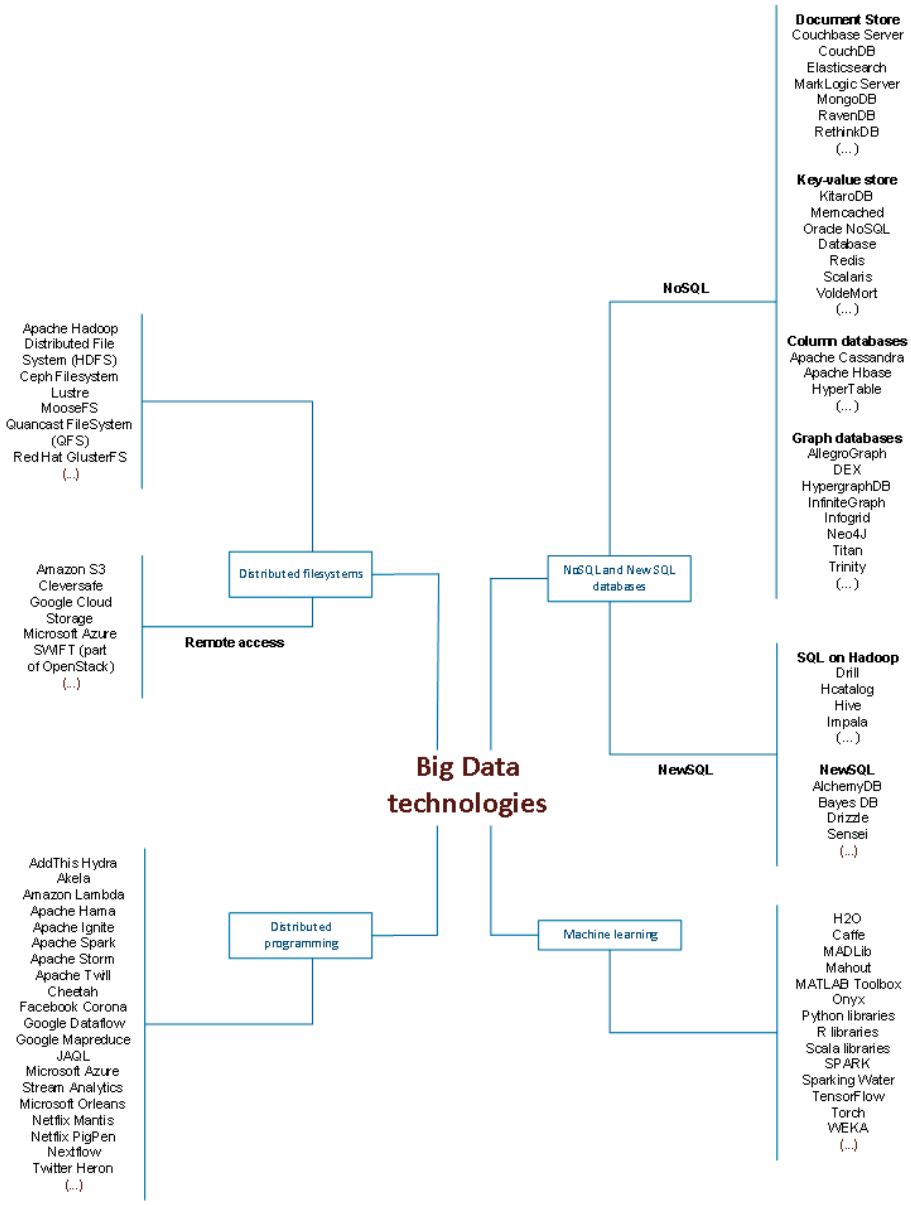

Figure 1. Some examples of software solutions available for the use of Big Data.

Table 1. Search results for keyword combinations within two scientific databases (Scopus®, Web of Science®) (March–May 2017).

Searched Term	Scopus®		Web of Science®	
	Results	Relevant	Results	Relevant
"Big Data" AND "sewage"	6	1	2	1
"Big Data" AND "pollutants" AND "water"	5	3	3	1
"Big Data" AND "risk" AND "chemicals" AND "water"	5	0	0	0

3. Key Considerations

To understand the current situation within the water industry, it is necessary to address the context within which the different stakeholders operate. Modern society has added a complexity to the water sector in hitherto unimaginable ways. Population increase, scarcity, and depletion of natural resources, increases in industrial processes, stringent legislation, and privatization of water companies are just some of the influencing factors that need to be considered. Figure 2 identifies some of the key

drivers in the UK water sector. As may be seen in Table 2, there are many stakeholder organizations influencing in the UK water industry, each having different drivers, priorities, and constraints, which frequently conflict. Simultaneously, the systems and conditioning factors are also continuously evolving, potentially hindering long-term solutions.

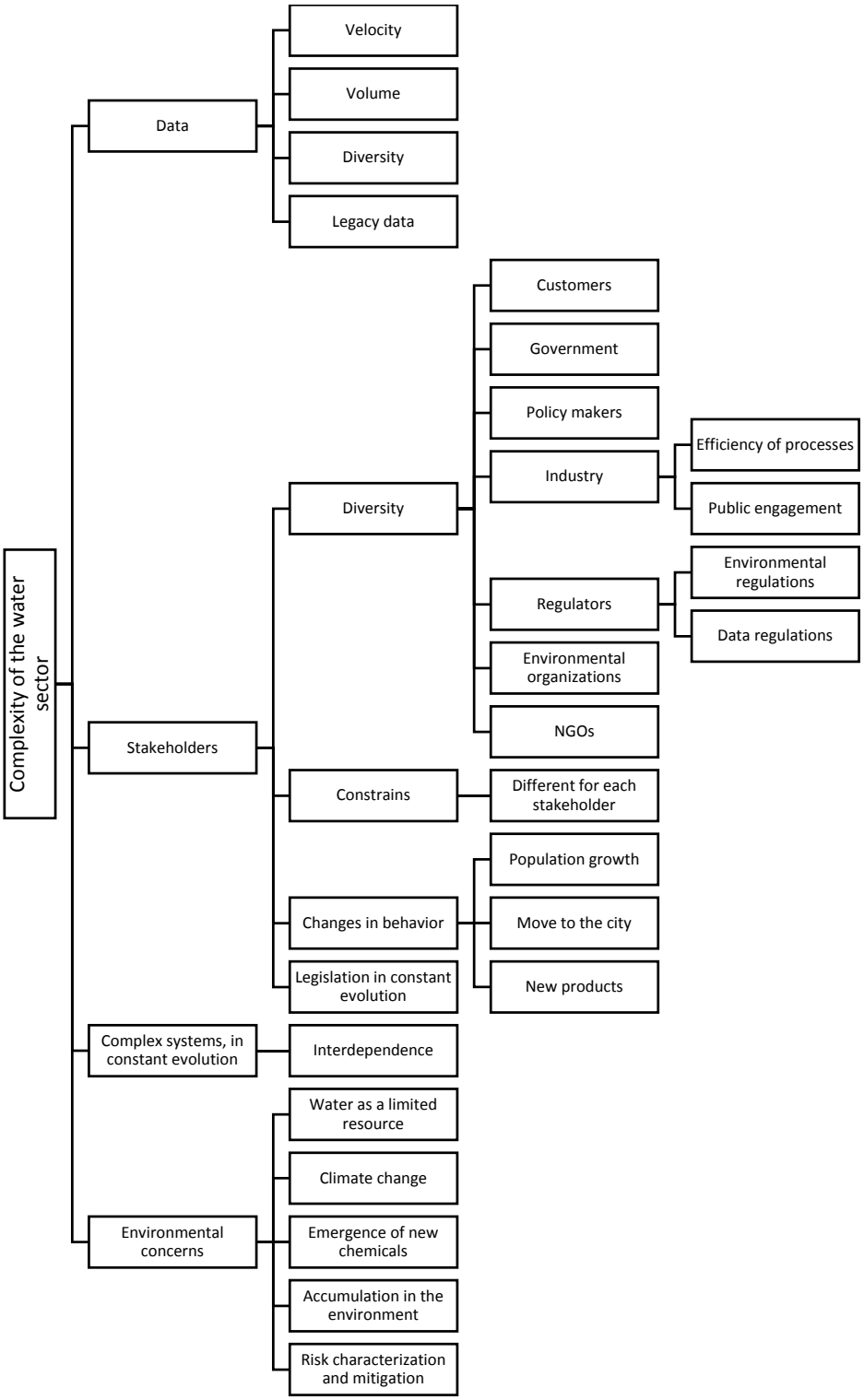

Figure 2. Some of the factors affecting to the complexity of the water sector, which need to be carefully considered in order to understand the current situation of the water sector.

3.1. Stakeholders Constraints—Public, Shareholders, Regulators, Government, Environmental

The diversity of the stakeholders involved (defined by Freeman [42] as "those groups and individuals who can affect and are affected, or are affected by, the accomplishment of organizational purpose") provides an additional factor contributing to the complexity of the UK water utility sector. Each of the key UK stakeholders (Table 2) plays an important role in the development, direction and implementation of policies and practices. It is important to highlight how different geographic regions in the UK also have different stakeholders with equivalent responsibilities. For instance, the environment agencies acting as Competent Authorities are the Environment Agency (EA) in England and Wales, the Scottish Environment Protection Agency (SEPA) in Scotland, and the Northern Ireland Environment Agency (NIEA) in Northern Ireland. Knowledge of the main drivers for each of these groups is important in gaining an understanding of the operation of the water sector.

Table 2. Classification of the key influencers in the UK water sector and some examples of membership of each class.

Role	Name	Region
General interest	Customers	General
Government	Welsh Assembly Government (WAG)	Wales
	Countryside Council for Wales	Wales
	Joint Nature Conservation Committee (JNCC)	UK
	Natural Resources Wales (NRW)	Wales
	Scottish Natural Heritage (SNH)	Scotland
	Department for Environment, Food and Rural Affairs (DEFRA)	UK
Industry	Anglian Water	East of England
	Dŵr Cymru Welsh Water	Wales
	Northumbrian Water	North East England
	Severn Trent Water	West Midlands, East Midlands
	South West Water	South West England
	Southern Water	South East England
	Thames Water	Greater London, Thames Valley
	United Utilities	North West England
	Wessex Water	South West England
	Yorkshire Water	Yorkshire and the Humber
Industry/Government	Northern Ireland Water	Northern Ireland
	Scottish Water	Scotland
NGO	The Canal & River Trust	England and Wales
Regulatory	Environment Agency (EA)	England and Wales
	Natural England	England
	Northern Ireland Environment Agency (NIEA)	Northern Ireland
	Northern Ireland Environment Agency (NIEA)	Northern Ireland
	Scottish Environment Protection Agency (SEPA)	Scotland

3.2. Complex Systems and Their Continuous Evolution

Directly or indirectly, chemical usage is embedded in most human activities. In many cases, chemicals are discharged to the environment once the associated activity or process is concluded. The presence or accumulation of such chemicals may pose a danger for humans or natural wildlife. As a main receptor and transport agent for chemicals, water bodies can cause diffusion of these substances, in certain geographical cases even with a trans-national dimension. This issue is aggravated with the continuous increase in the use and emergence of new chemicals [43,44], and the complexity of the physical, biological, and chemical interactions taking part in the ecosystem and organisms. In some cases, the magnitude of the consequences are unknown, and the process to assess consequent risk can take a significant time (sometimes >10 years, as outlined in a report from the American National Research Council (NRC) [43]). This represents a challenge for scientists and engineers, industry, and policy makers, all of whom require a strong evidence-base, able to support the decisions taken. Environmental legislation and its evolution in time is further representative of this complexity.

With the need for a fair apportionment, growing public interest, and the evolution of legislative frameworks, the continuously evolving fraction and composition of the chemicals reaching the water bodies presents a challenge in itself. Some of these chemicals can be present at concentrations less than current detection limits, or they can accumulate in the environment, while others can pose a yet unknown risk for humans and/or the environment [4,45–48]. For example, the PBDE limit proposed by the EU of 0.000000049 μg/L is several orders of magnitude below the levels achievable by even advanced treatment [5].

3.3. Environmental Concerns

Knowledge of the nature of the pollutants present in the influents reaching WWTPs is relevant not only for the preservation of water bodies, but also for the protection of land. WWTPs apply different processes to effluents to promote the removal of both particulate matter and dissolved pollutants, to meet quality standards. This generates large amounts of biosolids, rich in nutrients, and of great economic and environmental interest for agricultural or amenity use. However, a wide range of substances, some of them undesirable, persist [49–51], and can be transferred to land and the food chain after application [4,52], examples being phosphorus, metals, and endocrine disruptors. Economic and environmental restrictions for the use of alternative disposal options (e.g., combustion and disposal at sea) have led to an increase in the proportion of wastewater sludge recycled to agricultural land. Ofwat (the UK Water Services Regulation Authority) have estimated a rise from approximately 44% in 1992, to 80% in 2010 [53], whilst Cooper et al. [54], estimated the amount of sewage sludge recycled to agriculture as approximately 71% in 2013.

The adoption of European legislation concerning water and environment, and more specifically the Water Framework Directive (WFD), has significantly affected the control measures used for the protection of surface waters. The WFD [55] and other Directives identify a series of substances present in water that were previously under insufficient control, setting out "strategies against pollution of water", and defining the gradual procedure to remediate this problem. The substances identified were classified as priority (PS) or priority hazardous substances (PHS), and environmental quality standards (EQS) were also set for surface waters. The WFD state the necessity of compliance with the EQS for all the priority substances and the other pollutants listed so as to achieve "*good chemical status*" [56]. Competent Authorities (displayed in Table 3 for the UK) are responsive of the survey and monitoring of the water bodies to identify pressures, for detection of long-term trends, and classification of their status. Assessments are based on the evaluation of a range of biological communities rather than the sole measurement of chemicals. Although it is recognized as a more integrative and effective method for evaluating and measuring the ecological quality, there is still a series of gaps hindering this task. Some of them are expressed in the River Basin Management Plan for the UK, one of the Parliament's POSTnotes [57]:

- A large amount of data is required to obtain a high level of certainty in the decision making, something that is not always possible.
- The monitoring of water bodies is needed at a larger scale to identify the causes and effects of failures to meet the "good" status.
- Seasonal river fluctuations happen naturally, and it is expected to be more frequent in the future due to climate change. Lower river flows mean lower dilution of pollutants, which can lead to more severe impacts of the contaminants. It can also generate a series of technical difficulties identifying sources and impact of contaminants. In addition, pollutants assessed in an annual base can be subject to an underestimation of the effects.
- Historical industrial sites can also be a source of surface and groundwater pollutants. However, site location and characteristics of the pollutants emitted are not always known.
- Storm drains can carry a variety of contaminants derived from urban run-off and misconnections of domestics and commercial sewers, which eventually will derive into the waterbodies.

Hering et al. [58] also states some of the other gaps derived from this approach, such the heterogonous response of these biological communities to stress and restoration, and the need of long-term monitoring data to understand the aquatic ecosystems.

Table 3. Regulations and competent authorities responsible for the transposition of the WFD into the British legislation.

Region	Legislation	Competent Authority
England and Wales	The Water Environment (Water Framework Directive) (England and Wales) Regulations 2003 (Statutory Instrument 2003 No. 3242)	Environment Agency (EA)
Scotland	The Water Environment (Water Framework Directive) Regulations (Northern Ireland) 2003 (Statutory Rule 2003 No. 544)	Scottish Environment Protection Agency (SEPA)
Northern Ireland	The Water Environment and Water Services (Scotland) Act 2003 (WEWS Act)	Northern Ireland Environment Agency (NIEA)

4. Key Opportunities and Challenges

As it has been mentioned previously, Big Data techniques have been proved useful for supporting decision making and the prediction of outcomes. The development of a new generation of sensors, more accurate, smaller, cheaper to manufacture, and able to transmit the information in almost real time, is a contributing factor to the ubiquity of devices generating data of use for the water industry [59–63]. However, despite having access to a broad range of data sources and technical resources, the water utility sector appears to make very limited use of it for the improvement of water quality and source apportionment. A wider application of these techniques can provide context to data already available by water companies and regulators (e.g., CIP data, climate data, water quality data, and customer data), thus allowing the extraction of additional valuable information. The use of these techniques will support the provision of evidence required in the decision-making process. This can be achieved with the use of analysis techniques such as machine learning, which is able to extract additional and more accurate patterns and relationships from data. Results can provide better understanding of the processes occurring and help support a rationale for an improved decision making, to the benefit of industry, regulators, and customers alike. Table 4 displays some of the potential applications of Big Data for contributing in the decision-making process and the apportionment of liability in the water industry.

Table 4. Potential contributions of Big Data in addressing characterization and apportionment of liability.

Contribution Factor		Opportunities for the Use of Big Data
Points of origin	Residential	Characterization of the population: age, gender, health status, density, behaviour, etc. This can be used to infer the presence of certain drugs, medicines and other chemicals in water, and to explain the presence of certain volatile chemicals.
	Commercial	Identification of commercial activity: food services, car cleaning, storage of cleaning products, storage of paints, etc. Different activities might use different chemicals, each posing risks to the environment, and which can reach the sewage system by varied pathways. Certain services such as Google Maps® already hold extensive databases with the location and type of commercial activity.
	Industrial	Characterization of the industrial activity and the kind of chemicals in use, historical presence of industry and discharges/leaks. This can be used not only for the apportionment of the pollutant contribution to wastewater, but also for establishing background levels of chemicals, and achieve a more accurate risk assessment.

Table 4. *Cont.*

Contribution Factor		Opportunities for the Use of Big Data
Diffuse sources	Agricultural	Characterization of activity: crop, livestock, type of operation (extensive/intensive), soil classification, historical chemical composition of water bodies. This will not only contribute to a more equitable apportionment of responsibilities, but also to the prediction of outcomes according to external factors such as rainfall.
	Hard surface run-off (urban, amenities, industrial, etc.)	Identification of sources and factors with influence on the run-off: urban areas, population density, soil classification, traffic information, weather data, pipe bursts, industrial activity, open-air activities (concerts, camping areas, etc.) which can impact on the composition of the run-off water. A classification of the permeability of surface can be also used for modelling at large scale contribution of chemicals.

4.1. Improved Efficiencies, Streamlined Processes

The information holdings of water companies, government, and environmental agencies include large datasets comprising diverse information from different sources, e.g., analytical observations, land measurements, weather data, satellite multispectral imagery, and geo-referenced asset location, maintenance and performance and customer demographics, as well as comprising differing types of data—geospatial, temporal qualitative and quantitative in nature. Such resources may be collectively being classified as Big Data. Contemporary and emergent Big Data techniques can help overcome some of the challenges inherent in storage, maintenance, and analysis of such a diversity of data.

4.2. Improved Evidence Based Decision Making in the Water Industry

The correct application of these techniques offers the potential to provide insight as to the source, behaviour, prevalence, and destiny of chemicals reaching sewage and, ultimately, water treatment plants. Decision-making can be informed through incorporation and analysis of the wide spectrum of external datasets such demographic or thematic environmental information (e.g., soil, geology and meteorology), which in turn can enlighten new relationships between elements not initially considered. These approaches can also lead towards the extraction of additional outcomes from existing data, drawing upon the use of analytical techniques, as well as the means to convey, communicate and display the results in a manner whereby each of the target groups can be easily understood. This can be used to achieve more efficient approaches for the control of chemicals, and to ensure that current legislation is realistic and effective.

4.3. Use of Machine Learning for the Better Analysis of Hold Data

In the process of knowledge acquisition aimed to provide a better understanding of the natural world, observed natural phenomenon must be analysed to establish patterns that can support explanations. This procedure commences with the detection of regularities in data, continuing with the formulation of hypotheses that can characterize those regularities, finishing with the testing of the hypotheses against new data to evaluate legitimacy. This procedure is known as data mining [64]. It is time-consuming, and it can become challenging depending on the nature (e.g., volume, kind, or veracity) of the data holdings in question. However, machine learning algorithms can be used to support the rapid characterization of patterns in data [65], being of great interest for the analysis of data of large volume or complexity.

Samuel [66] offered an early definition of machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed". The definition has evolved over time, currently being understood as "the process by which a computer can work more accurately as it collects and learns from the data it is given" [67]. Its use is increasing with the popularization of open-source software, free and fast access to data, a diminution of the price of the infrastructure required to analyse and store the data, and advances in computational techniques.

Machine learning algorithms (Table 5) can be grouped by their learning style into three groups, namely supervised machine learning (SML), semi-supervised machine learning (SSML) and unsupervised machine learning (UML). The difference between these lies in the use or not of annotated training data. SML algorithms are provided with characterized training data representing how judgements might be provided by an expert. The goal is to minimize error in future classification judgements with respect the given inputs, predicting or forecasting target values. After training and validation is complete, SML machine learning approaches can be used effectively in prediction of new observations. UML uses unlabelled data, not being any classification or categorization present among the input observations. The aim of these kinds of algorithms is to infer hidden structure from the input data, finding generalities in the data that can be of use. However, in this case there is not an a-priori basis to judge goodness of results from UML, the goal being to establish interesting and useful generalities in data [65]. SSML algorithms make use of a combination of labelled and unlabelled input data, existing a desired prediction problem, but expecting the model to also learn the structures to organize the data. Large amounts of unlabelled data is used in conjunction with labelled data to construct better classifiers [68]. Supervised machine learning is generally more appropriate for classification and regression tasks, while unsupervised learning approaches suit clustering and association mining [64,67].

Table 5. Example algorithms used in machine learning.

Learning Style	Example Algorithms
Supervised Machine Learning	Decision trees, neural networks, Naïve Bayes, Support Vector Machines, K-Nearest Neighbours, Logistic regression, Adaboost [69,70], Generalized Linear Models, Linear and Quadratic Discriminant Analysis, Neural Network Models, etc.
Unsupervised Machine Learning	Single Link, Complete Link, CobWeb, K-Means, Expectation Maximization, Artificial Neural Networks, Support Vector Machines, Gaussian Mixture Models, Neural Network Models, etc.

4.4. Fusion—Integration of All Sources

The continual improvement in the effectiveness of machine learning algorithms, along with the ease in the generation of data, is allowing their increasing use in decision support, with noted examples in Medicine [71], Biology [72], and the Social Sciences [26]. However, expert judgement is still needed to incorporate assumptions beyond those required for the prediction methods, usually not testable but influencing the results, and the verification of the outcomes [73]. It is also necessary to use a risk assessment method able to integrate different sources of decision process and the expert judgement. In this case, Weight of Evidence has been proven useful in environmental risk assessment [74–77], and its use in combination with Big Data techniques will allow a holistic approach able to assist decision makers in the process of risk assessment.

The use of different datasets implies handling data with different characteristics (e.g., type, volume, and confidence level). As a result, it is necessary to augment the use of databases with different approaches to those of traditional SQL databases to handle this diversity. NoSQL databases (referred as not only SQL [12]), such as "MongoDB", allow the storage of data in "documents", with non-normalized data models, without predefined structure, and easily horizontally-scalable.

4.5. CIP as an Example of Decision Challenge

The need to complying with the EQS set by European Union Directives is particularly challenging for wastewater. The large number of chemicals involved, the analytical challenge of working with very low concentrations (ng/L), and the complexity of reactions and interactions difficult the process [3]. In the UK, the CIP is a strategy being developed by the industry and regulators, and driven by the UK Water Industry Research (UKWIR), for better understanding the challenge and risks that the presence

of certain pollutants in water represent. It comprehends the characterization of a series of chemicals of especial interest likely to be found in effluents from WWTPs [3,7].

The first phase of the CIP was undertaken between 2009 and 2012, with the objective of monitoring those chemicals considered most likely to reach the water treatment plant, so providing an insight into the behaviour, sources, and applicable control measures. Final effluents from 162 WWTPs, distributed throughout England, Scotland and Wales, were sampled either 14 or 28 times over a period of one year. A total of 70 determinants were targeted, 64 of which were trace contaminants situation. Results obtained have contributed with useful findings and provided a deeper insight of the current status of the sector [2–4]. The substantive and diverse data this study has generated has been used for guiding rational decision making within the water industry [78]. However, there are a series of factors which limit the amount of information which can be obtained by using this data alone, including:

- The vast amount and associated variety of measurements: The number of individual measurements is such that the correct handling and transformation presented a challenge. Storing, classification, and handling of this data requires the use of adequate methods for its effective use.
- The values obtained lack context: The complexity of the relationships between the different elements and processes taking part in the environment requires consideration of more than just the concentration of a chemical to evaluate its true impact. For instance, rainfall will affect the water level of the water bodies, thus affecting the concentration/dilution of chemicals and their effect/efficacy. It is necessary also to consider the natural eutrophication levels for the water body considered, as some rivers are naturally more eutrophicated than others. This can be achieved by considering the historical water quality status for each location.
- Unknown repercussions of the presence of some of the chemicals: While the impact of the presence of some chemicals in the environment is well known, science still lacks enough evidence to determine the direct or indirect effect for some combinations, or their long-term presence.
- The presence of a chemical or group of chemicals can influence the behaviour of others: Some chemicals can alter the behaviour or the impact of others, enhancing or reducing consequent effects. For example, high levels of phosphorus can lead to higher levels of eutrophication, which in turn affects biodegradation rates of other contaminants.

The use and incorporation of external datasets, such as meteorological data, demographics, and historical industry presence, is required to provide information complementary to the measurements, thus solving current gaps and extending the period in which this data can be of use. Drawn from the literature, Table 6 displays some of the technical solutions to the limitations noted. Databases such EU Agri4Cast or the UK Census data are freely available, and have been chosen to contextualize the results obtained in the first phase of the CIP. Their use provides the support needed to aid identification and control of sources, and the optimization of current remediation processes, which are some of the needs highlighted by Gardner et al. [3] after the study of the data obtained. This not only extends the period in which the data hold can be of use, but it can also unveil more effective strategies for the control of trace substances. Table 6 further suggests the use of both Supervised and Unsupervised Machine Learning approaches for use in predicting the concentration or presence of pollutants in other locations of the UK, while algorithms drawing on the latter approach can also be used for finding new ways by which the chemicals act.

Table 6. Limitations identified in the CIP data and some proposed solutions. * Materials comprehend both software and datasets.

Limitation	Proposed Solution	Materials * of Interest
Vast amount and associated variety of measurements	NoSQL database	Non-relational database
Measures lacking context (for the particular case of the UK)	Agro-meteorological data	Agri4Cast [79]
	Climate data and climate change projections	UKCP09 [80]
	Historical flood data	Remotely Sensed Flood Estimates [81] and Historic Flood Map [82]
	Historical mining locations	Inventory of Closed Mining Waste Facilities [83]
	Historical pollution levels	The Environment Agency "What's in Your Backyard?" service [84]
	Historical river water quality status	Environment Agency Water Quality Archive (WIMS) [85] and Environment Agency's Historic River Quality Data [84]
	Population data	UK Census (Age distribution, sex distribution, general health, population density, social class distribution) [86]
	River levels	River and sea levels in England [87] and Wales river levels [88]
	Sensitive areas	Eutrophic lakes [89], eutrophic rivers [90], and nitrates rivers [91]
	Soil classification	LandIS [92]
Unknown repercussion of the presence of chemicals	Machine Learning	Supervised and Unsupervised Machine Learning algorithms
Influence of the presence some chemicals on others	Machine Learning	Supervised and Unsupervised Machine Learning algorithms

4.6. SAGIS as an Example of Decision Challenge

SAGIS (Source Apportionment GIS) is a collaborative source apportionment model sponsored by UKWIR, the Environment Agency, and the Scottish Environment Protection Agency. SAGIS is being used in the UK to provide the necessary rationale to support fair apportionment for the water industry under the "polluter pays" principle of the EU, ensuring that the WFD is implemented effectively. It is built on the Environmental Agency's SIMCAT model, and combines GIS, export coefficient databases, and water quality models [93]. The outputs generated are used in the prediction of determinants (namely, nitrate; ortophosphate; diethylhexyl phthalate (DEPH); benzo-a-pyrene; fluoranthene; naphthalene; nonylphenol; triclosan; ethinylestradiol (EE2); PBDE (sum of BDE congeners 28, 47, 99, 100, 153 and 154); tributyltin (TBT); and total and dissolved phases of cadmium, copper, mercury, nickel, lead, and zinc), also included are contributions to modelled water concentrations, and the comparison of simulated outputs with observed river monitoring data.

SAGIS aims to incorporate all major point and diffuse source applicable to a range of substances, providing evidences for the rational identification of effective programs of measures to meet required EQS, of sources of contamination (not only those arising from sewage treatment) and the evaluation of the significance for each of them. Data inputs used in SAGIS include river network, hydrology and river monitoring data, water body boundaries, rainfall, land use and agricultural data, on-site wastewater treatment system, sewage treatment works, combined sewer overflow (CSO) and storm tank locations, soil erosion, highway run-off and pollution inventory, industrial discharges to river, and sewer (including extensive sewerage treatment works data) databases. However, the paucity of key data available has necessitated only a limited validation to date of the modelling outcomes for organic compounds.

Further incorporation of Big Data and machine learning may enhance the functionality of the model, allowing a deeper understanding of the sources of chemicals and incorporation of new

substances leading to more accurate predictions. For example, the presence of certain groups of pharmaceutical chemicals can be inferred from the age group of the population contributing to the effluents. In the same way, historical data can be used to fill the current gaps in the WFD, providing the necessary background for the judgement in the heterogeneous response of biological communities to stress.

5. Discussion and Conclusions

Modern technologic advances have reduced significantly the difficulty and costs for the generation and storage of data, being the current generation ratios beyond comparison with those in the past. In addition to it, some companies and governments are facilitating access to some of their datasets, encouraging their use. Furthermore, popularization and increasing of public interest is also contributing to the fast development of this technology and the creation of a complex and rapidly evolving "ecosystem" of hardware and software.

Big Data with machine learning have proven useful in supporting decision making and the prediction of outcomes in different fields. Despite having access to a large number of resources (datasets and sensors), there is a low number of peer-reviewed documented applications in the improvement of water quality and the study of sources and pathways of pollutants in water.

The water sector is a complex system, with many causal affects between variables that influence the water quality and the decision-making process. Some of the challenges being faced are the need for a fair apportionment, a growing public interest, the evolution of legislative frameworks, and the continuously evolving fraction and composition of the chemicals reaching the water bodies. Furthermore, the presence of chemicals in concentrations below detection levels, with an accumulative nature in the environment, or with an unknown risk for humans and/or the environment, suppose an additional obstacle for the appropriate risk assessment.

In the particular case of the UK, the adoption of European legislation (and more specifically the Water Framework Directive) has significantly affected the control measures for the protection of surface waters. The need for a rationale able to support a fair apportionment of liability within the water sector has led to the development of the CIP and SAGIS. Both approaches have generated useful results, but they have also highlighted a series of gaps which can be addressed through the use of Big Data.

The first phase of the CIP (from 2009 to 2012) consisted in the monitoring of those chemicals most likely to reach water treatment plants and the characterization of final effluents of 162 WWTPs from England, Scotland, and Wales. The aim of this phase was to gain a better understanding of the sources, behaviour, and control measures, used in guiding rational decision making within the water industry. However, measurements obtained lack the context required to unveil the true impact. Big Data approaches can be used for providing background considerations of the results (Table 6), thus permitting the identification of sources, provision of better assessment of current remediation processes, and an increase in the period of time over which this data can be of use.

SAGIS outputs are used in the prediction of determinants, contributions to modelled water concentrations, and comparison of simulated outputs with observed river monitoring data. It provides rationale support to ensure the effective implementation of the WFD and a fair apportionment work for the water industry under the "polluter pays" principle of the EU. However, improvement is needed in the modelling approaches used for organic compounds. Incorporation of Big Data and machine learning can provide a deeper understanding of sources of chemicals, and unveil relationships between chemicals which could be of use for the enhancement of the model.

The British legislative context is going through a crucial stage after the Brexit referendum in 2016. Although currently the exact nature of the future relationship between the UK and EU remains to be determined, current European legislation is already being transcribed into British law. Consequently, ratifying or modifying such laws will require supporting evidence, specific to the British context. Current legislation is not lacking in areas requiring improvement in the evaluation of ecological quality, with specificity for each location required and accounting for numerous factors

(e.g., historical industrial legacy, weather, and adapted species) each exerting significant influence. Big Data approaches have proven to be effective in the merging of diverse and large datasets in other fields, providing more accurate and precise outputs and its application therefore promises to offer a strong evidential-based supporting role pertaining to the water sector in both legislative and decision making.

Acknowledgments: This work was supported by the Natural Environment Research Council [grant number NE/M009009/1]. The authors are grateful to the Natural Environment Research Council (NERC). We are also grateful to Atkins Ltd. for their support and in particular Arthur Thornton and Carlos Constantino. The authors are grateful to the UK Water Industry Research (UKWIR) for their support. We thank Michael Hann for his advice.

Author Contributions: The final manuscript has been approved by all three authors. Juan Manuel Ponce Romero conceived of the presented idea, and compiled the manuscript. Stephen H. Hallett and Simon Jude supervised this work, verified the analytical methods used, and provided substantial inputs to the text. All three authors discussed the results and contributed to the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. European Commission. Introduction to the New EU Water Framework Directive—Environment—European Commission. Available online: http://ec.europa.eu/environment/water/water-framework/info/intro_en.htm (accessed on 15 March 2017).
2. Gardner, M.; Jones, V.; Comber, S.; Scrimshaw, M.D.; Coello-Garcia, T.; Cartmell, E.; Lester, J.; Ellor, B. Performance of UK wastewater treatment works with respect to trace contaminants. *Sci. Total Environ.* **2013**, *456*, 359–369. [CrossRef] [PubMed]
3. Gardner, M.; Comber, S.; Scrimshaw, M.D.; Cartmell, E.; Lester, J.; Ellor, B. The significance of hazardous chemicals in wastewater treatment works effluents. *Sci. Total Environ.* **2012**, *437*, 363–372. [CrossRef] [PubMed]
4. Jones, V.; Gardner, M.; Ellor, B. Concentrations of trace substances in sewage sludge from 28 wastewater treatment works in the UK. *Chemosphere* **2014**, *111*, 478–484. [CrossRef] [PubMed]
5. House of Commons Science and Technology Committee. *Water Quality: Priority Substances First Report of Session*; House of Commons Science and Technology Committee: London, UK, 2013.
6. Charriau, A.; Lesven, L.; Gao, Y.; Leermakers, M.; Baeyens, W.; Ouddane, B.; Billon, G. Trace metal behaviour in riverine sediments: Role of organic matter and sulfides. *Appl. Geochem.* **2011**, *26*, 80–90. [CrossRef]
7. UKWIR. The UKWIR Chemicals Investigation Programme—A Mid-Programme Update. Available online: https://www.ukwir.org/site/web/news/news-items/ukwir-chemicals-investigation-programme (accessed on 1 March 2017).
8. IBM. Big Data & Analytics Hub Extracting Business Value from the 4 V's of Big Data. Available online: http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data (accessed on 1 March 2017).
9. Coombes, P.; Barry, M. A systems framework of big data providing policy making—Melbourne's water future. In Proceedings of the OzWater14 Conference, Brisbane, Australia, 29 April–1 May 2014.
10. Cordier, M.O.; Garcia, F.; Gascuel-Odoux, C.; Masson, V.; Salmon-Monviola, J.; Tortrat, F.; Trepos, R. A machine learning approach for evaluating the impact of land use and management practices on streamwater pollution by pesticides. In Proceedings of the MODSIM05—International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, Melbourne, Australia, 12–15 December 2005; pp. 2651–2657.
11. Starzyk, J. Water Resource Planning and Management using Motivated Machine Learning. In Proceedings of the 10th IHP/IAHS George Kovacs Colloquium, Paris, France, 2–3 July 2010; pp. 214–220.
12. Mitchell, I.; Wilson, M. *Linked Data—Connecting and Exploiting Big Data*; White Paper; Fujitsu: London, UK, 2012; pp. 1–21.
13. Vitolo, C.; Elkhatib, Y.; Reusser, D.; Macleod, C.J.A.; Buytaert, W. Web technologies for environmental Big Data. *Environ. Model. Softw.* **2015**, *63*, 185–198. [CrossRef]

14. Oracle Corporation. *Mastering Big Data: CFO Strategies to Transform Insight into Opportunity*; A FSN & Oracle White Paper; FSN Publishing Limited: Herts, UK, 2012.
15. Computer Sciences Corp Data Universe Explosion & the Growth of Big Data | CSC. Available online: http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode (accessed on 26 April 2016).
16. IDC. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Available online: https://www.emc-technology.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf (accessed on 14 April 2016).
17. Gunarathne, T.; Wu, T.-L.; Choi, J.Y.; Bae, S.-H.; Qiu, J. Cloud computing paradigms for pleasingly parallel biomedical applications. *Concurr. Comput. Pract. Exp.* **2011**, *23*, 2338–2354. [CrossRef]
18. Young, S.D. A "big data" approach to HIV epidemiology and prevention. *Prev. Med.* **2015**, *70*, 17–18. [CrossRef] [PubMed]
19. Du, Y.; Ma, C.; Wu, C.; Xu, X.; Guo, Y.; Zhou, Y.; Li, J. A Visual Analytics Approach for Station-Based Air Quality Data. *Sensors* **2016**, *17*, 30. [CrossRef] [PubMed]
20. Gerstein, M. Genomics: ENCODE leads the way on big data. *Nature* **2012**, *489*, 208. [CrossRef] [PubMed]
21. Lee, J.; Kao, H.-A.; Yang, S. Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. *Procedia Cirp* **2014**, *16*, 3–8. [CrossRef]
22. Pijanowski, B.C.; Tayyebi, A.; Doucette, J.; Pekin, B.K.; Braun, D.; Plourde, J. A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment. *Environ. Model. Softw.* **2014**, *51*, 250–268. [CrossRef]
23. Kim, G.-H.; Trimi, S.; Chung, J.-H. Big-Data Applications in the Government Sector. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
24. An, Q.; Wen, Y.; Xiong, B.; Yang, M.; Chen, X. Allocation of carbon dioxide emission permits with the minimum cost for Chinese provinces in big data environment. *J. Clean. Prod.* **2017**, *142*, 886–893. [CrossRef]
25. Betrie, G.D.; Tesfamariam, S.; Morin, K.A.; Sadiq, R. Predicting copper concentrations in acid mine drainage: A comparative analysis of five machine learning techniques. *Environ. Monit. Assess.* **2013**, *185*, 4171–4182. [CrossRef] [PubMed]
26. Berk, R.A. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*; Springer: New York, NY, USA, 2012; ISBN 1461430852.
27. Bayati, M.; Braverman, M.; Gillam, M.; Mack, K.M.; Ruiz, G.; Smith, M.S.; Horvitz, E. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLoS ONE* **2014**, *9*, e109264. [CrossRef] [PubMed]
28. Glaeser, E.L.; Hillis, A.; Kominers, S.D.; Luca, M. Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. *Am. Econ. Rev.* **2016**, *106*, 114–118. [CrossRef]
29. Bhardwaj, S.; Jain, L.; Jain, S. Cloud Computing: A Study of Infrastructure as a Service (IAAS). *Int. J. Eng. Inf. Technol.* **2010**, *2*, 60–63.
30. Subashini, S.; Kavitha, V. A survey on security issues in service delivery models of cloud computing. *J. Netw. Comput. Appl.* **2011**, *34*, 1–11. [CrossRef]
31. Kindström, D. Towards a service-based business model—Key aspects for future competitive advantage. *Eur. Manag. J.* **2010**, *28*, 479–490. [CrossRef]
32. SUSE LLC. SUSE Enterprise Linux. Available online: https://www.suse.com/ (accessed on 1 March 2017).
33. Red Hat Inc. Red Hat. Available online: https://www.redhat.com/en (accessed on 1 March 2017).
34. Olsson, G. Instrumentation, control and automation in the water industry—State-of-the-art and new challenges. *Water Sci. Technol.* **2006**, *53*, 1–16. [CrossRef] [PubMed]
35. Jamieson, D.G.; Shamir, U.; Martinez, F.; Franchini, M. Conceptual design of a generic, real-time, near-optimal control system for water-distribution networks. *J. Hydroinform.* **2007**, *9*, 3–14. [CrossRef]
36. Marvin, S.; Chappells, H.; Guy, S. Pathways of smart metering development: Shaping environmental innovation. *Comput. Environ. Urban Syst.* **1999**, *23*, 109–126. [CrossRef]
37. Alahakoon, D.; Yu, X. Advanced analytics for harnessing the power of smart meter big data. In Proceedings of the 2013 IEEE International Workshop on Inteligent Energy Systems (IWIES), Vienna, Austria, 14 November 2013; pp. 40–45.
38. Beal, C.D.; Flynn, J. Toward the digital water age: Survey and case studies of Australian water utility smart-metering programs. *Util. Policy* **2015**, *32*, 29–37. [CrossRef]

39. Kitchin, R. The real-time city? Big data and smart urbanism. *GeoJournal* **2014**, *79*, 1–14. [CrossRef]
40. McKenna, E.; Richardson, I.; Thomson, M. Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy* **2012**, *41*, 807–814. [CrossRef]
41. Boyle, T.; Giurco, D.; Mukheibir, P.; Liu, A.; Moy, C.; White, S.; Stewart, R. Intelligent Metering for Urban Water: A Review. *Water* **2013**, *5*, 1052–1081. [CrossRef]
42. Freeman, R.E. *Strategic Management: A Stakeholder Approach*; Cambridge Unviersity Press: Cambridge, UK, 1984; ISBN 978-0-521-15174-0.
43. Abt, E.; Rodricks, J.V.; Levy, J.I.; Zeise, L.; Burke, T.A. Science and decisions: Advancing risk assessment. *Risk Anal.* **2010**, *30*, 1028–1036. [CrossRef] [PubMed]
44. Hristozov, D.R.; Zabeo, A.; Foran, C.; Isigonis, P.; Critto, A.; Marcomini, A.; Linkov, I. A weight of evidence approach for hazard screening of engineered nanomaterials. *Nanotoxicology* **2014**, *8*, 72–87. [CrossRef] [PubMed]
45. Janssen, C.R.; Heijerick, D.G.; De Schamphelaere, K.A.C.; Allen, H.E. Environmental risk assessment of metals: Tools for incorporating bioavailability. *Environ. Int.* **2003**, *28*, 793–800. [CrossRef]
46. Hutchinson, T.H.; Brown, R.; Brugger, K.E.; Campbell, P.M.; Holt, M.; Länge, R.; McCahon, P.; Tattersfield, L.J.; van Egmond, R. Ecological risk assessment of endocrine disruptors. *Environ. Health Perspect.* **2000**, *108*, 1007–1014. [CrossRef] [PubMed]
47. Chang, H.-S.; Choo, K.-H.; Lee, B.; Choi, S.-J. The methods of identification, analysis, and removal of endocrine disrupting compounds (EDCs) in water. *J. Hazardous Mater.* **2009**, *172*, 1–12. [CrossRef] [PubMed]
48. Linkov, I.; Ames, M.R.; Crouch, E.A.C.; Satterstrom, F.K. Uncertainty in Octanol−Water Partition Coefficient: Implications for Risk Assessment and Remedial Costs. *Environ. Sci. Technol.* **2005**, *39*, 6917–6922. [CrossRef] [PubMed]
49. Singh, R.P.; Agrawal, M. Potential benefits and risks of land application of sewage sludge. *Waste Manag.* **2008**, *28*, 347–358. [CrossRef] [PubMed]
50. Lundin, M.; Olofsson, M.; Pettersson, G.; Zetterlund, H. Environmental and economic assessment of sewage sludge handling options. *Resour. Conserv. Recycl.* **2004**, *41*, 255–278. [CrossRef]
51. Fytili, D.; Zabaniotou, A. Utilization of sewage sludge in EU application of old and new methods—A review. *Renew. Sustain. Energy Rev.* **2008**, *12*, 116–140. [CrossRef]
52. Eriksson, E.; Lundy, L.; Donner, E.; Seriki, K.; Revitt, M. Sludge management paradigms: Impact of priority substances and priority hazardous substances. In Proceedings of the 12th International Conference on Urban Drainage, Porto Alegre, Brazil, 11–16 September 2011.
53. Ofwat. *Water 2020: Our Regulatory Approach for Water and Wastewater Services in England and Wales Appendix 2 Moving Beyond Waste—Further Evidence and Analysis*; Ofwat: Birmingham, UK, 2016.
54. Cooper, J.; Carliell-Marquet, C. A substance flow analysis of phosphorus in the UK food production and consumption system. *Resour. Conserv. Recycl.* **2013**, *74*, 82–100. [CrossRef]
55. European Parliament. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy. Available online: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02000L0060-20141120 (accessed on 14 March 2016).
56. European Parliament. EUR-Lex—32008L0105—EN. Available online: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008L0105 (accessed on 27 October 2015).
57. Parliamentary Office of Science and Technology. *River Basin Management Plans*; Parliamentary Office of Science and Technology: London, UK, 2008.
58. Hering, D.; Borja, A.; Carstensen, J.; Carvalho, L.; Elliott, M.; Feld, C.K.; Heiskanen, A.-S.; Johnson, R.K.; Moe, J.; Pont, D. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. *Sci. Total Environ.* **2010**, *408*, 4007–4019. [CrossRef] [PubMed]
59. Hill, D.; Kerkez, B.; Rasekh, A.; Ostfeld, A.; Minsker, B.; Banks, M.K. Sensing and Cyberinfrastructure for Smarter Water Management: The Promise and Challenge of Ubiquity. *J. Water Resour. Plan. Manag.* **2014**, *140*, 1814002. [CrossRef]
60. Ingildsen, P. *Smart Water Utilities: Complexity Made Simple*; IWA Publishing: London, UK, 2015; ISBN 1780407572.

61. Reis, S.; Seto, E.; Northcross, A.; Quinn, N.W.T.; Convertino, M.; Jones, R.L.; Maier, H.R.; Schlink, U.; Steinle, S.; Vieno, M.; et al. Integrating modelling and smart sensors for environmental and human health. *Environ. Model. Softw.* **2015**, *74*, 238–246. [CrossRef] [PubMed]
62. Kerkez, B.; Gruden, C.; Lewis, M.; Montestruque, L.; Quigley, M.; Wong, B.; Bedig, A.; Kertesz, R.; Braun, T.; Cadwalader, O.; et al. Smarter Stormwater Systems. *Environ. Sci. Technol.* **2016**, *50*, 7267–7273. [CrossRef] [PubMed]
63. Eggimann, S.; Mutzner, L.; Wani, O.; Schneider, M.Y.; Spuhler, D.; Moy de Vitry, M.; Beutler, P.; Maurer, M. The Potential of Knowing More: A Review of Data-Driven Urban Water Management. *Environ. Sci. Technol.* **2017**, *51*, 2538–2553. [CrossRef] [PubMed]
64. Bell, J. *Machine Learning Hands-On for Developers and Technical Professionals*; Computer Scientist; Wiley: Indianapolis, IN, USA, 2015; ISBN 9781118889060.
65. Smith, T.C.; Frank, E. Introducing Machine Learning Concepts with WEKA. *Methods Mol. Biol.* **2016**, *1418*, 353–378. [CrossRef] [PubMed]
66. Samuel, A.L. 4.3.3 Some Studies in Machine Learning Using the Game of Checkers Some Studies in Machine Learning Using the Game of Checkers. *IBM J.* **1959**, *3*, 210–229. [CrossRef]
67. Cielen, D.; Meysman, A.; Ali, M. *Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools*; Manning Publications Co.: Greenwich, CT, USA, 2016; ISBN 9781633430037.
68. Zhu, X. Semi-Supervised Learning Literature Survey. Ph.D. Thesis, University of Wisconsin, Madison, WI, USA, 2007.
69. Maglogiannis, I.G. *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*; IOS Press: Amsterdam, The Netherlands, 2007; ISBN 1586037803.
70. Harrington, P. *Machine Learning in Action*; Manning Publications: Shelter Island, NY, USA, 2012; ISBN 1617290181.
71. Shipp, M.A.; Ross, K.N.; Tamayo, P.; Weng, A.P.; Kutok, J.L.; Aguiar, R.C.T.; Gaasenbeek, M.; Angelo, M.; Reich, M.; Pinkus, G.S.; et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **2002**, *8*, 68–74. [CrossRef] [PubMed]
72. Weston, J.; Leslie, C.; Ie, E.; Zhou, D.; Elisseeff, A.; Noble, W.S. Semi-supervised protein classification using cluster kernels. *Bioinformatics* **2005**, *21*, 3241–3247. [CrossRef] [PubMed]
73. Athey, S. Beyond prediction: Using big data for Policy Problems. *Science* **2017**, *355*, 483–485. [CrossRef] [PubMed]
74. Regmi, N.R.; Giardino, J.R.; Vitek, J.D. Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA. *Geomorphology* **2010**, *115*, 172–187. [CrossRef]
75. Weed, D.L. Weight of Evidence: A Review of Concept and Methods. *Risk Anal.* **2005**, *25*, 1545–1557. [CrossRef] [PubMed]
76. Linkov, I.; Loney, D.; Cormier, S.; Satterstrom, F.K.; Bridges, T. Weight-of-evidence evaluation in environmental assessment: Review of qualitative and quantitative approaches. *Sci. Total Environ.* **2009**, *407*, 5199–5205. [CrossRef] [PubMed]
77. Burton, G.A.; Chapman, P.M.; Smith, E.P. Weight-of-Evidence Approaches for Assessing Ecosystem Impairment. *Hum. Ecol. Risk Assess.* **2010**, *8*, 1657–1673. [CrossRef]
78. Jarvie, H.P.; Neal, C.; Williams, R.J.; Neal, M.; Wickham, H.D.; Hill, L.K.; Wade, A.J.; Warwick, A.; White, J. Phosphorus sources, speciation and dynamics in the lowland eutrophic River Kennet, UK. *Sci. Total Environ.* **2002**, *282*, 175–203. [CrossRef]
79. European Commission. Agri4Cast Resources Portal. Available online: http://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx (accessed on 1 August 2017).
80. Environment Agency. Met Office UKCP09. Available online: http://ukclimateprojections.metoffice.gov.uk/21678 (accessed on 1 August 2017).
81. Environment Agency. Remotely Sensed Flood Estimates. Available online: https://data.gov.uk/dataset/remotely-sensed-flood-estimates (accessed on 1 August 2017).
82. Environment Agency. Historic Flood Map. Available online: https://data.gov.uk/dataset/historic-flood-map1 (accessed on 1 August 2017).
83. Environment Agency. Inventory of Closed Mining Waste Facilities. Available online: https://data.gov.uk/dataset/inventory-of-closed-mining-waste-facilities2 (accessed on 1 August 2017).

84. Environment Agency. What's in Your Backyard? Available online: http://maps.environment-agency.gov.uk/wiyby/wiybyController?ep=maptopics&lang=_e (accessed on 1 August 2017).
85. Environment Agency. Water Quality Archive (WIMS). Available online: http://environment.data.gov.uk/water-quality/view/landing (accessed on 1 August 2017).
86. Office for National Statistics UK 2011 Census. Available online: https://www.ons.gov.uk/census/2011census (accessed on 1 August 2017).
87. Environment Agency. River and Sea Levels in England. Available online: https://flood-warning-information.service.gov.uk/river-and-sea-levels (accessed on 1 August 2017).
88. Cyfoeth Naturiol Cymru—Natural Resources Wales Wales River Levels. Available online: https://naturalresources.wales/evidence-and-data/maps/check-river-levels/?lang=en (accessed on 1 August 2017).
89. Environment Agency. Sensitive Areas—Eutrophic Lakes. Available online: https://data.gov.uk/dataset/sensitive-areas-eutrophic-lakes (accessed on 1 August 2017).
90. Environment Agency. Sensitive Areas—Eutrophic Rivers. Available online: https://data.gov.uk/dataset/sensitive-areas-eutrophic-rivers (accessed on 1 August 2017).
91. Environment Agency. Sensitive Areas—Nitrates. Available online: https://data.gov.uk/dataset/sensitive-areas-nitrates2 (accessed on 1 August 2017).
92. Cranfield University. LandIS—Land Information System. Available online: http://www.landis.org.uk/ (accessed on 1 August 2017).
93. Comber, S.D.W.; Smith, R.; Daldorph, P.; Gardner, M.J.; Constantino, C.; Ellor, B. Development of a Chemical Source Apportionment Decision Support Framework for Catchment Management. *Environ. Sci. Technol.* **2013**, *47*, 9824–9832. [CrossRef] [PubMed]

Leveraging big data tools and technologies: Addressing the challenges of the water quality sector

Ponce Romero, Juan Manuel

2017-11-23

Ponce Romero JM, Hallett SH, Jude S, Leveraging big data tools and technologies: addressing the challenges of the water quality sector, Sustainability, Vol. 9, Issue 12, 2017, Article number 2160
http://dx.doi.org/10.3390/su9122160