Environmental Science & Technology

pubs.acs.org/est

Feature

Data Analytics for Environmental Science and Engineering Research

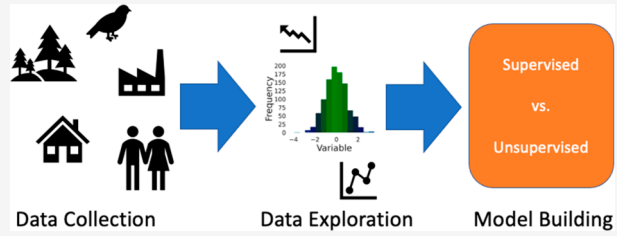
Suraj Gupta, Diana Aga, Amy Pruden, Liqing Zhang, and Peter Vikesland*

Cite This: https://doi.org/10.1021/acs.est.1c01026

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The advent of new data acquisition and handling techniques has opened the door to alternative and more comprehensive approaches to environmental monitoring that will improve our capacity to understand and manage environmental systems. Researchers have recently begun using machine learning (ML) techniques to analyze complex environmental systems and their associated data. Herein, we provide an overview of data analytics frameworks suitable for various Environmental Science and Engineering (ESE) research applications. We present current applications of ML algorithms within the ESE domain using three representative case studies: (1) Metagenomic data analysis for characterizing and tracking antimicrobial resistance in the environment; (2) Nontarget analysis for environmental pollutant profiling; and (3) Detection of anomalies in continuous data generated by engineered water systems. We conclude by proposing a path to advance incorporation of data analytics approaches in ESE research and application.



KEYWORDS: *environmental science and engineering, data analytics, machine learning, metagenomics, nontarget analysis, anomaly detection, water*

1. BACKGROUND

Data science is a rapidly evolving interdisciplinary field incorporating fundamentals from computer science, information science, mathematics, and statistics.[1] It combines principles and methodologies that facilitate and guide extraction of knowledge and insights from available data streams in a usable format that supports data-driven policy and decision making.[2] Moreover, data science provides the capacity to better delineate problems by improving alignment between the data that is available and the corresponding questions that can be addressed. The availability of voluminous amounts of data, powerful computational resources, affordable data storage, and highly efficient algorithms is enabling broader and deeper data analyses than previously possible.

Machine learning (ML) is an emerging data science subfield that includes algorithms and methodologies that can be used to find hidden patterns within data and aid in predictive model construction.[3] ML enables analysis of bigger and ever more complex data sets in more efficient and more accurate ways than otherwise possible.[4] In the past decade, ML has begun to significantly impact numerous disciplines[5] and Environmental Science and Engineering (ESE) has benefited from this surging interest in ML and its applications.[6] At its core, ESE is concerned with improving and maintaining the environment, with the ultimate goal of protecting human and ecological health. ESE encompasses diverse areas, such as water and wastewater treatment, air quality, environmental impact assessment, and hazardous waste management. ESE incorporates concepts from disciplines such as basic sciences, public health, engineering, biological sciences, and nanotechnology. Research and industrial work within ESE increasingly require collection of vast amounts of data that simultaneously reflect numerous data types (e.g., air and water quality measurements, flow measurements, spatial discretization, etc.) with wide spatial and temporal variability. Data of this nature encompasses a broad and dynamic range with masses reported from nanogram to teragram and flows ranging from microliters per second to millions of liters per day. Recent advances in environmental metagenomics and nontarget analysis further expand the types and volumes of data being generated within ESE.

The advent of novel data acquisition and handling techniques has opened the door to alternative and potentially more comprehensive means of environmental monitoring that will improve our capacity to understand and manage environmental systems. For instance, metagenomics and nontarget analysis are evolving as powerful ways to identify unknown contaminants for surveillance.[7,8] Further, identification of anomalous or unusual events in water/wastewater treatment processes using real time water quality data is a key tenet of the "digital water" movement that holds immense

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

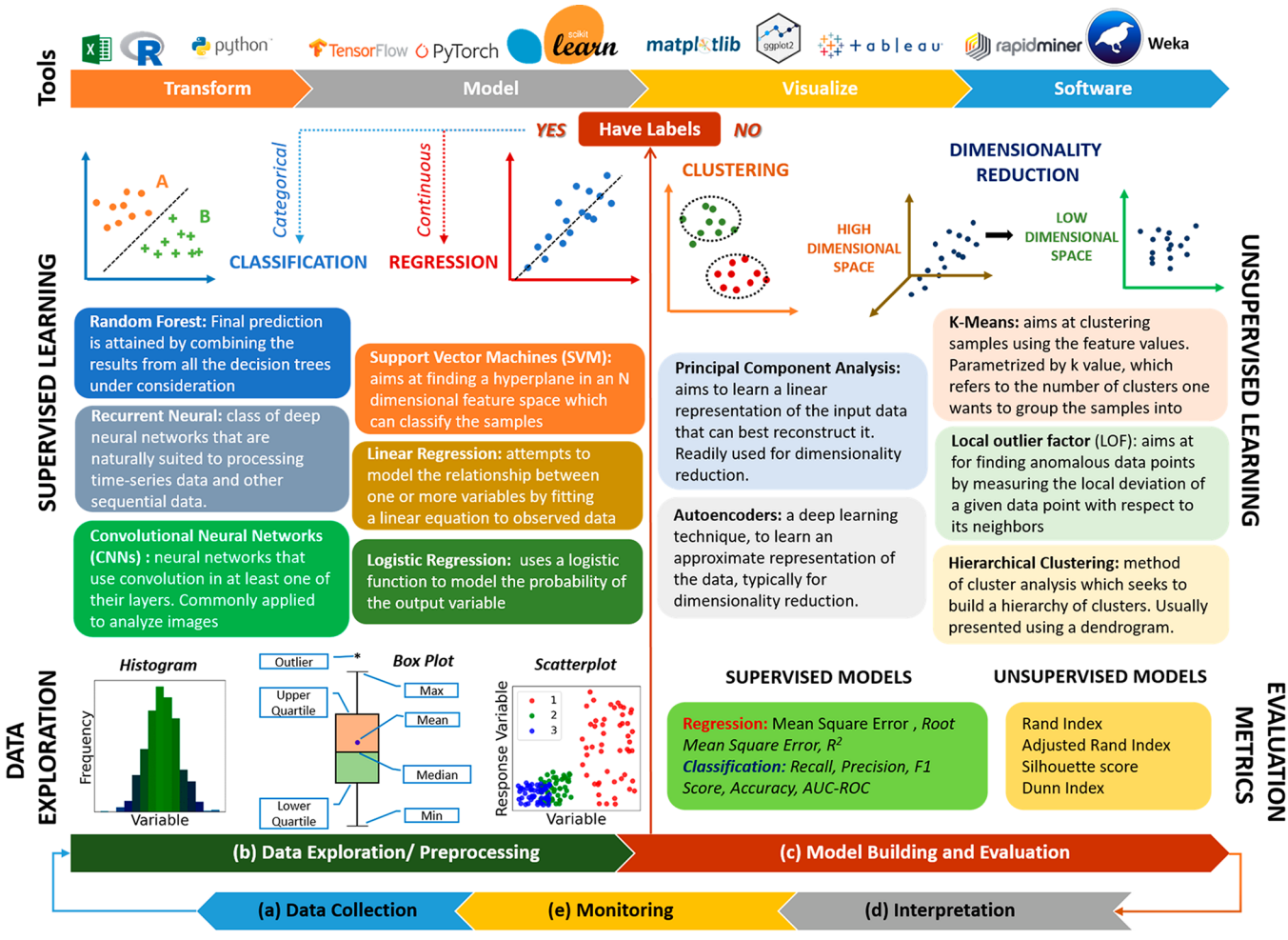


Figure 1. Data Analysis Framework and Methodologies. (a) Data Collection: Gather the data; (b) Data Exploration/Preprocessing: Fix the discrepancies, visualize and transform the data to the required form; (c) Model Building and Evaluation: Train the model and evaluate performance; (d) Interpretation: Use the model to make predictions, analyze and interpret the outcomes; (e) Monitoring: Based on the learned knowledge and domain expertise. The top thread on the figure represents the tools and software that are available to execute various steps in the end-to-end data analysis framework.

potential for the water industry.[9] However, interpreting these types of data is not trivial and requires appropriate application of ML techniques.

The aim of this Feature is to provide an overview of data analysis frameworks suitable for application within ESE. We first introduce a general data analysis framework, then highlight current applications of ML within ESE problem domains, and conclude with thoughts about the future.

2. GENERAL DATA ANALYSIS FRAMEWORK

A general data analysis framework includes data acquisition, data exploration, data preprocessing and visualization, algorithms for model building, and ways to evaluate and interpret the models and the results (Figure 1).

2.1. Data Acquisition. Data acquisition is the process of collecting or acquiring data and storing it in a readily accessible form, such as a database (Figure 1a). Data acquisition requires thorough planning to ensure the data set can be effectively interrogated for the desired end purpose. Key elements of such a plan include deciding upon the data collection methodology, selecting variables of interest, determining the sampling frequency and the number of replicates, assessing how much data is needed to effectively build the model(s) and test the study hypotheses, and deciding upon means to properly document and store the data.

Work within ESE increasingly requires acquisition of vast amounts of data. Hence, it is necessary to frame and adopt succinct protocols to minimize biases and increase comparability and reproducibility across the discipline. To that extent, several publications and guidance documents have focused on good implementation practices for collection of robust data sets.[10,11,12,13]

2.2. Data Exploration/Visualization and Preprocessing. Data exploration or visualization is used to understand the data characteristics. Data exploration helps illustrate key aspects, such as sample size, missing values, distributions, initial patterns, correlations, or variables that appear to be sensitive to the system of interest. Exploration is commonly achieved by plotting and visualizing the data in a manner such that distributions and trends are readily apparent.[14]

Raw data are often incomplete, noisy, and inconsistent due to data redundancy, incompatibilities between multiple data sources, and divergent data collection protocols. Poor data quality can lead to inaccurate or misleading analyses.[15] Data preprocessing is intended to improve model quality and minimize computational resource usage.

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 1. Overview of Popular ML Algorithms

machine learning model	description	positives and negatives	exemplary applications within ESE
Logistic Regression (Logit model)[24]	• Supervised ML classification algorithm. • Statistical model that uses a logistic function to model the probability of the output variable. • Used to perform binary classification tasks where given samples are classified into two groups. • Can be extended to multiple output variables.	• Easy to implement. • Highly interpretable. • Does not require heavy computational resources. • Often used as the first model or benchmark to compare to more complex models.	• Estimate the probable presence of a type of vegetation using environmental variables.[25] • Predict contaminant sources[26] and water quality[27,28] in different water systems. • Anomaly detection in water treatment systems.[28]
Random Forest (RF)[29]	• A tree-based ensemble learning method that is used for supervised classification and regression problems. • The basic unit is a decision tree. • Samples and features are randomly sampled from the full data set and individual decision trees are made on the sampled data set. • The outcomes generated from the ensemble of decision trees are combined to develop the final model prediction.	• Robust, can handle linear and nonlinear data, outliers, as well as noisy data. • Not prone to overfitting when the model overtrains on the training data and fails to generalize on the testing or new data set.[30] • Depending on the size of the data set and the ensemble of trees, RFs can get complex and may require high computational resources and long times for both training and prediction.	• Predict relative abundance levels in sewage.[7] • Select important variables in predicting anomalous events in water distribution systems.[28] • Predict nutrient concentrations using water quality variables.[32] • Predict nanofiltration/reverse osmosis-membrane rejection of emerging contaminants.[31]
Support Vector Machines (SVM)	• Supervised learning algorithm. • Used for both classification and regression tasks. • Aims to find a hyperplane (in an N dimensional feature space) that has maximum distance between data points of all classes. The hyperplane classifies the samples or data points under consideration. • Maximizing the margin distance brings confidence in the classification of new data points. • Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. • Sets of mathematical functions (kernels) can be used to transform the input data as required.	• Uses a subset of training points in the decision function (support vectors) and is memory efficient. • Effective in high dimensional spaces and in cases where the number of samples is less than the number of dimensions. However, choosing kernel functions and regularization is crucial to avoid overfitting.	• Nonlinear time series forecasting model to predict water quality[33] • Predict the concentration of nitrate and nitrite in the mixed liquor of a wastewater treatment plant[34] • Phenolic pollutant concentration prediction in drinking water[35]
Artificial Neural Networks (ANNs)	• ANNs are computing systems inspired to analyze and process information the way the human brain does. • A simple ANN consists of an input layer, a hidden layer made of entities (called nodes) that receive weighted inputs from the input layer and then transform them using a nonlinear function that transmits the transformed value to the output layer where the final output is used to get predictions. • In conventional ANN architecture the information moves unidirectionally (i.e., from input to output layer). This architecture is also known as a feedforward network. • When the number of hidden layers is increased to achieve higher levels of abstraction in order to extract complex information from the data, the	• DL has emerged as the state-of-the-art ML method.[36−39] • DL models require large data sets for training to achieve robust performance. However, it is difficult to know the sample size required for the particular task beforehand. Some rules of thumb have been proposed based on a number of assumptions. For instance, the sample size needs to be 50−1000× the number of classes, or 10−100× the number of features, or 10−50× the number of weights in the network.[40] • For smaller data sets, alternative ML models may perform better. • ANN architectures exist that can be used to perform supervised, unsupervised, or semisupervised learning tasks. Some examples are recurrent neural networks (RNNs), convolutional neural networks (CNNs), and autoencoders.	• Predict effluent concentrations in a wastewater treatment plant.[42] • Anomaly detection in water treatment system.[28,43] • Air quality forecasting model.[44]

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 1. continued

machine learning model	description	positives and negatives	exemplary applications within ESE
	resulting architecture is known as a deep neural network or deep learning (DL). • ANNs tend to overfit. Conscientious model building, thorough hyperparameter tuning and extensive validation is often required.[41]	• Considered "black box" models as the internal functioning of ANNs can be hard to interpret. This can lead to result misinterpretation and the acceptance of poor models. A thorough understanding of various hyperparameters is necessary to understand the nuances of the model.	
Principal Component Analysis (PCA)[45,46]	• An unsupervised technique: PCA combines features and creates a new feature set of the same size where each new feature (i.e., the "components") is a combination of the original features. • New features are formed by finding the eigenvectors and corresponding eigenvalues that provide an estimation of how much variance each component explains.	• Widely used feature extraction method. Used to reduce the dimensionality of the data set, while preserving as much data variability as possible. • One can select a few components that explain the majority of the variability in the data and drop the remaining ones. • Can be used to reduce multicollinearity in the data as new components are independent of each other.	• Estimate the variance explained by given variables of a water quality data.[47] • Intercompare air pollution patterns.[44] • Establish the key compounds to distinguish between samples using high resolution mass spectrometry (HRMS) data.[48,49] • Estimate the importance of different variables and identify key-foulants in a membrane based drinking water treatment process.[50]
K-means Clustering (K-means)[51]	• Unsupervised ML algorithm that aims to cluster samples using feature values. • User defines a target k value, which refers to the number of clusters one wants to group the samples into. • The k value refers to the number of the centroid which refers to a location representing the center of a given cluster. • K-means starts with random centroids, and iteratively finds the most appropriate centroids and assigns the sample into respective clusters such that the sum of the squared distance between the samples and the centroid is minimized.	• Simple to implement and can easily scale to large data sets. • Disadvantage is that one needs to a priori define the k value, which may be difficult. • Fails for clusters with complex nonspherical geometric shapes. • Cluster centroids can get skewed in the presence of outliers hence it is advised to handle outliers before applying k-means.	• Delineate different air quality index threshold clusters using air pollution data.[52] • Analyze water quality data.[5,53,54] • Generate vulnerability maps for an aquifer using water quality data.[55] • Risk assessment of water pollution sources.[56]
Hierarchical Clustering	• Unsupervised ML algorithm that aims to cluster samples using feature values. • The hierarchical clustering algorithm in addition to breaking up the objects into clusters also shows the hierarchy or ranking of the distance and shows how dissimilar one cluster is from others. Hierarchical clustering is represented as dendograms. • Input is a measure of dissimilarity between samples. The choice of similarity (or distance) measures is the main influencing factor in hierarchical clustering. • Hierarchical clusters can be built either as agglomerative (bottom-up: each sample starts in its own cluster and clustering is done by merging each sample and moving up the hierarchy) or divisive (top-down: all samples are put in one cluster and splits are performed recursively moving down the hierarchy).	• No a priori information about the number of clusters is required. • Time intensive. • Could be difficult to identify the correct number of clusters based upon the dendogram.	• Feature clustering based on Pearson correlations in HRMS data.[57] • Estimate the similarity between the resistome profiles of different samples based on Bray−Curtis dissimilarity metric.[58] • Water quality assessment with hierarchical cluster analysis based on Mahalanobis distance.[59]

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

ESE data sets can incorporate numerous data types with wide spatial and temporal variability. Accordingly, if possible, a thorough data visualization/exploration and preprocessing process should be performed to gain initial insights and streamline the data for the subsequent model building steps. (Figure 1b). To aid this, depending on the data type, well established community guidelines could be leveraged. This process typically includes four steps: data cleaning, integration, reduction, and transformation (See the glossary for details).[16,17]

2.3. Model Building. *ML Models.* ML models fall into three classes: supervised, unsupervised, and semisupervised (Figure 1c). Supervised learning is a ML task where an algorithm is used to train a model using known data and then the learned model is used to predict the outcome of new or unforeseen instances.[18] In supervised learning, a sample in the data set has two components: (1) a set of variables (features) that define the sample and (2) class labels (outcomes) that one wants to predict. Two types of models are built using supervised learning algorithms: classification or regression models. A classification model is suggested when the output data can be categorized into specific groups or classes (i.e., discrete output variables). If the output variable is a continuous variable, a regression model is recommended.

Unsupervised learning is done in the absence of pre-existing class labels. It is a ML category where the goal is to find hidden and unknown data patterns or to determine the data distribution across the data set.[19] Unsupervised learning algorithms are often used for data exploration and visualization. The two major types of unsupervised learning approaches are clustering and dimensionality reduction. Clustering algorithms are used to group similar samples together into clusters, whereas dissimilar samples are relegated to separate clusters. Dimensionality reduction algorithms are used to reduce the dimension of the feature set. Too many uninformative features can hinder predictive modeling; this is often referred to as the "curse of dimensionality". Some common supervised and unsupervised learning algorithms are listed in Table 1.

Semisupervised learning is a hybrid of supervised and unsupervised learning, where a set of labeled data is used in conjunction with a set of unlabeled training data. In many cases, getting true labels for a data set can be difficult. Semisupervised learning tackles this problem by making use of unlabeled data in the learning process and then leveraging this learned information along with the labeled data set for subsequent prediction.[20]

Model Optimization and Evaluation. Supervised learning is usually a three-step process. In the first step, the data are split into training and testing data sets with the split ranging between 60 and 80% for training and 40−20% for testing. The training data set is then iteratively separated into training and validation sets. The model is trained on the training set, while the validation set is used to tune the model hyperparameters that consist of the model architecture and the parameters that affect the speed and quality of the training process. Following this cross-validation, the learned model is tested against the test data set to evaluate model robustness and accuracy. To ensure unbiased model evaluation, the test data set cannot be used for training. Commonly used model evaluation metrics include accuracy, F1-score, precision, recall, area under the receiver operating characteristic (ROC) curve, mean absolute error (MAE), and mean squared error (MSE).[21]

Evaluating unsupervised models can be challenging because data do not have prior labels. Evaluation typically involves internal and external validation.[22] Internal validation is done by estimating inter- and intra-cluster distances to evaluate cluster quality. A cluster refers to a collection of data points that accumulate together based on certain similarities. Models minimizing intracluster distances while maximizing intercluster distances are preferred. Such a result suggests that these models are performing well at clustering similar data points together. Some commonly used metrics for this purpose are the Adjusted Rand Index and the Silhouette Coefficient. However, a good internal validation score does not guarantee the effectiveness of the obtained clusters for real applications. Hence, external validation is necessary to assess whether the data points are assigned to the correct clusters. This step usually requires human evaluation or comparison against benchmarks. For dimensionality reduction, task reconstruction error or loss is estimated to evaluate model performance. These methods seek to minimize the reconstruction error or loss−defined as the distance between the original data point and its projection onto a lower-dimensional subspace (its "estimate").[23]

The ESE community has begun to leverage various ML algorithms and techniques to analyze environmental data sets. While it is difficult to determine a priori which ML algorithm is best suited for a particular data set, an improved understanding of the positives and negatives and the inherent assumptions of the specific algorithms aid in the choice of the right technique(s). Table 1 summarizes algorithms popular within the ESE community. The table describes the algorithms, their advantages and disadvantages, the assumptions for specific data types, and gives examples where the algorithms have been used to address ESE related problems.

2.4. Data Interpretation. The final step in the model building cycle is to interpret the learnt model and develop reasonable explanations for the obtained analysis (Figure 1d). Data interpretation could mean extracting important variables contributing to the model prediction. These variables could help in verifying the hypotheses of the study or understanding what factors are critical in driving the observations in the study under consideration. At this stage, the literature and expert opinions are queried to make connections between the model outputs and knowledge about the relevant chemical, physical, and biological phenomena to make data-driven conclusions and decisions.

3. CURRENT APPLICATIONS OF MACHINE LEARNING IN ESE

In this section, we discuss current applications of ML algorithms within the ESE domain by presenting three case studies representing different application areas.

3.1. Metagenomic Data Analysis. High throughput shotgun (untargeted) metagenomic DNA sequencing offers a robust and effective way to access the microbial world and is now often used in ESE. Differentiating microbial communities in different environments,[60,61] studying the dissemination of antibiotic resistance in environmental systems,[62,63] defining bacterial communities in contaminated environments to explore bioremediation,[64] and characterizing microbial communities in wastewater treatment processes[65,66] are all examples of environments where metagenomic characterization is increasingly being applied.

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

In a typical metagenomic study, the first step is often to search sequenced DNA reads against a reference database to derive taxonomic or functional gene annotation. The obtained taxonomic or functional gene profiles are then normalized to provide relative abundance information corresponding to different species or genes in the sample that can be visualized and analyzed to answer specific questions about the sample. Continuous optimization and declining costs of sequencing platforms have made metagenomics related research more accessible than ever before. This trend has led to the generation of large volumes of publicly available sequencing data, but rendering these data into meaningful interpretations remains challenging.[67] Fortunately, the advent of ML methods has immensely accelerated advancement at every cross section of a typical metagenomic study.[68] Here, we highlight some applications of ML in antibiotic resistance-related studies.

One often faced challenge with metagenomic data is that of high dimensionality and low sample size (HDLSS; i.e., more variables than independent samples). For example, the number of genes annotated within a given environmental sample can easily exceed hundreds of thousands. Hence, many unsupervised techniques, such as principal component analysis (PCA) and non-metric multi-dimensional scaling (NMDS), are used as preprocessing steps to reduce dimensionality by removing uninformative features. A common application of these techniques is to examine the similarity/dissimilarity of different environmental samples by clustering based on species or gene composition.[69] Network-based ML (data is represented in the form of a graph where nodes are the data points to be clustered and edges represent the relationship or similarity between the data points) is another unsupervised approach to study protein−protein or gene−gene interactions to understand different functional pathways.[70]

Given the sample labels or response variables, and the gene composition, support vector machines (SVMs), ANNs, and ensemble methods (such as random forest (RF) or extremely randomized trees) have been extensively used to perform supervised learning and build predictive models. For example, identifying interesting patterns and important genes in a data set,[58] predicting relative antibiotic resistance abundance levels,[7] understanding the role of socioeconomic status in shaping the resistome or microbiome,[71] or the prediction of antibiotic resistance phenotype[72] are some of the unique problems that have been examined using the above-mentioned methods. Hidden Markov models (HMMs), which can be used in both supervised or unsupervised fashion, are one of the more readily used algorithms to detect antibiotic resistance gene variants or potential functional homologues[73] and have been applied for the discovery of novel ARGs from metagenomic data.[74]

Word embedding, which has gained a lot of popularity over the recent years, is a feature learning technique used in natural language processing (NLP). Word embedding is a term used for the representation of words or sentences in the form of numeric vectors that can be used for downstream ML tasks.[75] Raw DNA/protein sequences sliced into k-mers are analogous to the structure of a sentence and can be analyzed in a similar fashion as natural languages. Thus, there is a rapid thrust toward analyzing raw sequences using NLP based techniques.[76,77] MetaMLP,[78] based on a similar idea, uses a word embedding based classification model to predict ARG phenotype and has been found to perform 50× faster than DIAMOND (one of the fastest sequence alignment methods) with similar accuracy. Word embedding based models hold immense promise in analyzing metagenomic data as they have the ability to learn patterns directly from within raw sequences.

Similar to the antibiotic resistance example, other applications can also be envisioned. For instance, analyzing metagenomics data for taxonomic classification[79,80] or mobile genetic element classification,[81] where ML algorithms were able to outperform the traditional methods. In essence, ML algorithms have enhanced our ability to robustly analyze complex metagenomic data.

3.2. Nontarget Analysis of Environmental Samples. A rapidly developing approach in environmental analysis and toxicology involves the use of high resolution mass spectrometry (HRMS) based nontarget analysis (NTA) where data on accurate masses of molecular and fragment ions are collected without a priori information on the chemicals being analyzed. NTA can potentially aid in the screening and analysis of the vast and diverse universe of organic pollutants, a grand challenge faced by the ESE community.[82] Because the chemicals being detected are not predetermined, NTA provides an opportunity to comprehensively examine the occurrence, fate, and transport of chemical contaminants in different environmental niches with minimal bias.[82] Similarly, NTA can be applied in environmental metabolomics to facilitate understanding of the effects of chemical perturbations on exposed organisms (e.g., plants, animals, and humans), without focusing on a particular biochemical pathway.[83,84] A recent review highlighted studies that employed the different types of NTA techniques used in metabolomics to discover metabolite changes in plants induced by exposure to xenobiotics (e.g., pharmaceuticals, personal care products, pesticides, flame retardants, and engineered nanomaterials), to examine the effect of altered levels of nutrients in the environment on plant systems.[85] Here we present a brief summary of how the combination of NTA with ML can advance monitoring of water, wastewater quality, soil, and exposed organisms (e.g., humans, wildlife).

While there are various methods that can be used for NTA in ESE, HRMS is the most popular because of its capacity for sensitive detection of low levels of contaminants and metabolites in complex environmental samples. However, the power of NTA using HRMS has been limited by problems associated with sample preparation and data analysis. In analyzing environmental samples using MS, sample concentration and cleanup are critical because the signal intensity of the MS features depend not only on the concentrations of the chemicals, but also on the amounts and the nature of the matrix present in the sample extracts. The reproducibility of the ionization of compounds in MS can be compromised significantly by matrix effects. Therefore, it is important to have an appropriate number of replicate samples and properly selected blank samples when conducting NTA. Solid-phase extraction for sample cleanup is commonly used, but this approach can introduce bias because highly polar contaminants may be lost during sample preparation. Unlike in target analysis where variations due to matrix effects and sample losses can be corrected using stable isotope-labeled reference compounds as surrogates, this approach cannot be used in NTA because the purpose of NTA is to identify unknown contaminants that were not included in the target list. To normalize for instrumental variation and matrix effects, internal standards with varying polarity can be added to the sample extract prior to analysis. Recently, several compounds of

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

diverse structure, log K_{ow}, chromatographic behavior, and ionization efficiency have been proposed for inclusion in a quality control mixture usable for NTA.[86] However, the proposed mixture has limitations, such as for the detection of hydrophilic compounds and molecular formula generation for compounds containing fluorine. Finally, due to the substantial amount of MS data acquired indiscriminately under full-scan and the subsequent MS/MS fragmentation of molecular and fragment ions, data processing to identify relevant features is daunting. Therefore, NTA workflows with built-in filters and criteria are being developed to facilitate prioritization of MS features that are useful for chemical structure annotation. In this regard, advanced data processing tools, high throughput statistical packages, and user-friendly visualization programs are needed to fully interrogate the rich data sets acquired by NTA.

NTA based on HRMS holds great promise for the comprehensive monitoring of the occurrence and fate of contaminants in water and wastewater.[87] It also allows researchers to retrospectively analyze stored HRMS data to screen for suspected contaminants (i.e., suspect screening) that may have been missed during target analysis.[88] Using advanced computational strategies, such as hierarchical cluster analysis, common patterns in the occurrence of contaminants in the environment and their emission pathways can be predicted based on time series analysis of the aggregated NTA data.[89] NTA combined with cluster analysis was successfully applied to reveal previously unmonitored chemical contaminants in soil and sediment samples.[90] The application of NTA and advanced postacquisition data treatment will continue to enhance our ability to discover emerging contaminants in the environment, including those that bioaccumulate and pose risks to humans and wildlife. For instance, new polyhalogenated compounds were detected for the first time in blubber samples from marine mammal sentinel species using both LC-HRMS and GC-HRMS for analysis, and an open-source data mining software in the R programming environment that detected halogenated signatures in full scan HRMS.[91] Finally, NTA combined with personal passive samplers and proper sample preparation techniques can be used to unlock the composition of chemical mixtures that humans are exposed to on a daily basis, which can be used in investigating the human exposome.[92]

Akin to metagenomic data, NTA data pose a similar HDLSS challenge, as each mass spectrum constitutes a large number of peaks representing potential compounds present in a given sample. Hence, data preprocessing is crucial and inevitable. Various data preprocessing steps are performed such as detecting peaks, subtracting peaks that were found in blank or control samples, componentization (i.e., grouping of signals that probably belong to one unique molecular structure) and removing noise using replicate measurements.[93] Following preprocessing, various supervised and unsupervised ML algorithms can be applied to engineer, select, and extract relevant features. However, before any further analysis, data normalization, and data scaling are two crucial steps that require consideration.

The most common algorithms used for NTA are linear projection methods, such as PCA and supervised partial least-squares discriminant analysis (PLS-DA).[94,95,48] PCA aids in sample comparison by removing variance from the sample set. Supervised PLS-DA along with relevant metadata can be used to extract features pertaining to the specific questions that are being asked. However, a major challenge is that feature detection based on peak intensity can be misleading because of erroneous peak assignments. Also, relevant information can be lost when differentiating the actual contaminant signals and background noise when using intensity information for feature selection. Hence, several alternative approaches that aim at using raw data signals (retention time × mass-to-charge ratio) that bypass the peak detection step have been proposed and implemented for improved extraction of information and underlying patterns in the data set.[96−98] This includes techniques such as transforming the raw data signal to distance matrices for dissimilarity analysis,[96−98] and using clustering to extract features.[99]

A number of other data analysis algorithms exist, but have not yet become common in NTA environmental analysis. Clustering techniques such as k-means and hierarchical clustering can be used to identify similar samples. Algorithms like RF, SVMs and ANNs can be applied to classify samples based on different categories and would be advantageous in illustrating nonlinear relations in the data. These algorithms have shown excellent promise in analyzing HRMS data in other fields[100−103] and this promise can certainly be extended to environmental sample analysis.

3.3. Anomaly Detection in Engineered Water Systems (EWS). EWS is an umbrella term for systems of water collection, treatment, distribution, storage and their operation. Recently, there has been a major thrust toward digitalization of the water sector.[104] In particular, the evolution of cyberinfrastructure and of online process control instrumentation has led to the development of advanced process control solutions such as supervisory control and data acquisition (SCADA) systems.[105,106] Such advances have enabled water utilities to continuously monitor water quality, identify problems, and effectively oversee maintenance issues both remotely and more locally. These systems entail collection of a large volume of raw data that could be, in conjunction with appropriate data analysis techniques, transformed into valuable information that can be leveraged to make proactive decisions to optimize overall performance.[107] In particular, there is surging interest in using ML techniques to identify unusual patterns in raw EWS data as a means of discovering unexpected activities—this is broadly termed anomaly detection.[108]

A typical anomaly detection task in EWS aims to differentiate between natural, expected variations in water quality and unusual or suspicious variations caused by contamination or failure somewhere in the system. The challenge is to characterize these normal water quality variations as it requires analysis of long-term data that encompasses inherent background variability.[109] This characterized normal response helps to flag anomalous events in the data. Various ML algorithms have been applied to address this problem (Table 1).

Within EWS, one primary form of the generated data is a time series, where each time point can be considered a discrete sample. The time series consists of measurements of indicators collected over time from one or several sensors, which form the feature set. Hence, each sample can be represented as a multidimensional vector, where each dimension represents a feature. Historic data is then used to train the model. Given appropriate analysis of the supporting data set, one can frame and solve problems to address a multitude of aims such as detecting leaks, sensor failures, abrupt changes in water quality, or contamination events.[110−112]

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

There are a number of different anomaly detection techniques available.[113] In the unsupervised approach, unlabeled samples are clustered using an algorithm; such as k-means, density-based clustering algorithms, or expectation-maximization (EM) clustering. The concept is that the normal data points cluster to form high density clusters, whereas anomalous data points cluster separately or distant from the normal data points/clusters and are located in low-density regions. This way one can carry out additional investigation using new samples and classify them relative to the normal data. In the supervised approach, labels (normal or anomalous) are assigned to each sample based on past information and expert knowledge of anomalous behavior. In this way, the problem is converted into a binary classification problem. Though, it can easily be extended to a multiclass classification to categorize different types of anomalies. SVM, ANN, Bayesian Network, Logistic Regression models, and their variants are well explored algorithms in this space.[113] A special case of SVM, One-Class SVM (OCSVM) is a widely used method for anomaly detection. In OCSVM, the entire training data set is considered as one-class (e.g., normal class) and the new data points are classified as similar (normal) or different (anomalous) to the training data. Because considering all data points from one class is equivalent to having no label, OCSVM is considered as unsupervised learning method.

Owing to the success of ML in detecting anomalies in other domains, such as network security, researchers have started exploring its potential in water quality anomaly detection.[43,114,115] Based on the studies published so far, DL algorithms have shown promise in detecting anomalies and have outperformed conventional techniques on a number of occasions. However, it should be noted that DL methods could be slow to train depending on the depth of the network and the amount of data available.[116]

Many studies have adopted a batch learning approach where the model is trained on historical data and then the new data is categorized using this trained model. With a continuous incoming time-series data stream and the possibility of novel anomalies, retraining the models with every new data set can quickly become impractical and difficult to reliably execute. Continual or active learning frameworks that continuously learn as the new data stream comes in and that can identify anomalous behavior in real time are required to circumvent these issues.[117,118] Variants of Latent Dirichlet Allocation, Markov Models, and ANN based architectures are algorithms that have been applied in other domains to achieve continual online learning.[119−121] However, thes approaches remain underexplored for anomaly detection in water systems as these frameworks are not trivial and have their own implementation challenges.[118,119]

Ultimately, anomaly detection will be most valuable if it can learn continually as the data stream comes in and yield real-time reporting that informs immediate corrective action. It is crucial that the models being utilized are fast and accurate. Hence, going forward, there is a need to shift the focus toward hybrid approaches (combinations of different algorithms) to build powerful models that are able to detect multiple types of anomalies in real-time. Such models will enhance EWS anomaly detection and advance public health.[9,118,119]

4. PATH FORWARD

There is increasing interest in and a growing body of research documenting how data analytics is being used to address ESE problems. As illustrated in this Feature, the power of data analytics has been widely recognized, as has been the vast need for its application. Anomaly detection has particularly promising applications for water professionals and practitioners. Notably, broader application of metagenomics and NTA can revolutionize environmental monitoring efforts. However, the application of data analytics in ESE practice remains in its infancy and concerted efforts are required to make data analytics an integral part of ESE research and education. Such a goal would most effectively be achieved if there were an agreed-upon plan of action. Coordinated efforts on several fronts are needed to help the ESE community reap the potential benefits of data analytics.

First, there is a need to encourage collaborations between data scientists and ESE practitioners that involve diversely engaged and integrated research teams from multiple disciplines. Such collaborations will facilitate cross-disciplinary communication and simultaneous skills building. For example, data scientists come from a culture highly supportive of data sharing (e.g., GitHub, Bitbucket, public databases). The ESE community should embrace this culture and incorporate data sharing both locally, regionally, and globally. Recent efforts to address the COVID-19 pandemic,[122,123] the global dissemination of antimicrobial resistance,[7] and the development of globally vetted data analysis approaches within the NTA community are all steps in this direction. Such collaborations have the potential to yield profound data-driven insights and conclusions that would be impossible to achieve within normal academic and professional silos. Working on their own, data scientists may create models that answer specific questions, but lack the interpretive training required to contextualize their results into real world applications. Simultaneously, ESE practitioners with data may not possess the corresponding tools or insights required for proper analysis. Strategic collaborations will lead to improved data analytics, data interpretation, and improved decision making.

Second, there is a lack of user-friendly data analytics tools and platforms devised for ESE problems. Although understanding the theory behind ML algorithms is not a difficult task for many trained scientists, a major hurdle that ESE researchers face is in the coding and implementation of these models. With the continual advancement of data science, programming is becoming an increasingly necessary skill, without which models may be improperly developed or may lack efficiency. User-friendly tools may circumvent the problems with coding learning curves and may be essential for widespread practitioner adoption. While many tools (Microsoft power BI, Weka, Tableau, Azure) are available for general data analysis, application-specific tools/platforms would greatly aid in performing end-to-end analyses.

Third, the field needs to incorporate data analytics-oriented curricula. This could include the addition of data analytics focused modules within existing coursework, the introduction of new data science and statistics courses within ESE degree programs, and the development of relevant capstone projects for ESE applications. Such experiential learning would help students understand the intricacies of different algorithms and provide hands-on experience in applying various data analytics techniques in the context of specific ESE problems. Further, encouraging students to take part in data science internships would be an excellent means of developing such expertise.

Fourth, introducing data analytics workshops and making the relevant resources available is valuable for students,

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

researchers, and professionals. A plethora of online resources, such as Coursera, LinkedIn Learning, and Edx, are available for low or no cost, offering high-quality content in data science. Expanding access to and creation of open source learning platforms could prove instrumental in instigating the data-driven problem-solving approach.

Finally, we know that data analytics is not new, but continues to evolve at a rapid pace. These advancements furnish us with new and powerful ways to analyze the data that can help holistically tackle the challenges faced by ESE community, and ultimately inform decision-making and policy formulation at scales never previously imagined. Hence, it is critical to be abreast of new developments in data analytics and to continue making efforts to harvest their full potential.

AUTHOR INFORMATION

Corresponding Author

Peter Vikesland − *Via Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0003-2654-5132; Email: pvikes@vt.edu

Authors

Suraj Gupta − *The Interdisciplinary PhD Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, Virginia 24061, United States*

Diana Aga − *Department of Chemistry, University at Buffalo, The State University of New York, Buffalo, New York 14226, United States;* orcid.org/0000-0001-6512-7713

Amy Pruden − *Via Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0002-3191-6244

Liqing Zhang − *Department of Computer Science, Virginia Tech, Blacksburg, Virginia 24061, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c01026

Notes

The authors declare no competing financial interest.

Biography

Peter J. Vikesland is the Nick Prillaman Professor of Civil and Environmental Engineering at Virginia Tech. He received his B.A. from Grinnell College in Chemistry in 1993 and M.S. and Ph.D. in Civil and Environmental Engineering from University of Iowa in 1995 and 1998. Vikesland's research focuses on the fate of nanomaterials in the environment, nanotechnology enabled sensor platforms for environmental quality assessment, the environmental dissemination of antibiotic resistance, and the application of data analytics in environmental science and engineering. He is a past President of the Association of Environmental Engineering and Science Professors (AEESP), is a U.S. National Science Foundation CAREER awardee, and is the recipient of the 2018 Walter Weber Research Innovation Award from AEESP.

ACKNOWLEDGMENTS

This study was supported by National Science Foundation (NSF) awards OISE (#1545756), ECCSS NNCI (#1542100), and CSSI (#2004751). Additional support was provided by the Center for Science and Engineering of the Exposome at the Virginia Tech Institute for Critical Technology and Applied Science (ICTAS) and the Virginia Tech Graduate School supported Genetics, Bioinformatics, and Computational Biology (GBCB) and Sustainable Nanotechnology (VTSuN) programs.

GLOSSARY

Accuracy	Correct predictions/total predictions.
Autoencoders	A deep learning technique that aims to build a model where the output targets are set equal to the input variables. It seeks to learn an approximate representation of the data and reconstruct it. It is an unsupervised learning approach used for dimensionality reduction.
Bayesian Network	Bayesian networks are a kind of probabilistic graphical model that utilizes Bayesian inference for probability estimations. Bayesian networks are used to model conditional dependence, and hence causation, using a directed graph.
Data cleaning	Identification of outliers or filling in of missing values. Outlier identification can be done by using Z-score, quartile values, or hypothesis testing. Missing data can be handled in multiple ways such as filling the value by computing the summary statistics of the given variable, using predicted values computed by an ML algorithm, manual curation, or by ignoring the missing record.

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Data integration	The combination of multiple data sources into a single coherent form.
Data reduction	Aggregation or elimination of redundant information to reduce the data set size.
Data transformation	Conversion of raw data by normalizing to a common scale to ensure consistency and comparability.
Density Based Clustering	Work by recognizing "dense" groups of points, permitting it to learn clusters of discretionary shape and distinguish anomalies in the information.
Dimensionality Reduction	The process of reducing the number of random variables or attributes under consideration.
EM Clustering	Similar to k-means except it assigns the samples into clusters based on the probabilities estimated using the EM algorithm. The objective is to maximize the overall probability of the data for the given (final) clusters.
Ensemble Methods	Ensemble methods are algorithms that combine predictions from several base models to obtain one optimal model.
F1-Score	F1-Score is a harmonic mean of recall and precision.
HMM	HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process.
k-mer	A substring with a length *k* in a biological sequence of nucleotides.
Naïve Bayes	Naive Bayes methods are a family of simple probabilistic classifiers based on Bayes rule. The method is called "Naive" for its assumption of conditional independence among the features.
Natural Language Processing (NLP)	A subfield of computer science, artificial intelligence, and linguistics that deals with the interaction of computers with human languages.
NMDS	NMDS is an ordination technique based on distance or dissimilarity matrix. NMDS represents pairwise dissimilarity between samples in a low-dimensional space.
PLS-DA	PLS-DA is a linear classification model that is able to predict the class of new samples.
R	Programming language for Statistical Computing
ROC-Curve	A receiver operating characteristic (ROC) curve is a plot of true positive rate (TPR-*y* axis) against the false positive rate (FPR-*x* axis). It measures the performance of a classifier.
SCADA	Supervisory control and data acquisition (SCADA) is a system designed to gather and analyze real time data. It is used in water/wastewater treatment plants to monitor and manage processes.

REFERENCES

(1) Tukey, J. W. The future of data analysis. *Ann. Math. Stat.* **1962**, *33* (1), 1−67.

(2) Hayashi, C., What is data science? Fundamental concepts and a heuristic example. In *Data science, classification, and related methods*; Springer, 1998; pp 40−51.

(3) Bishop, C. M. *Pattern recognition and machine learning*; Springer, 2006.

(4) Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing* **2016**, *2016* (1), 67.

(5) Agarwal, R.; Dhar, V., Big data, data science, and analytics: The opportunity and challenge for IS research. In *INFORMS*: **2014**.25443

(6) Wilcox, C.; Woon, W. L.; Aung, Z. *Applications of machine learning in environmental engineering*; Citeseer, 2013.

(7) Hendriksen, R. S.; Munk, P.; Njage, P.; Van Bunnik, B.; McNally, L.; Lukjancenko, O.; Röder, T.; Nieuwenhuijse, D.; Pedersen, S. K.; Kjeldgaard, J. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **2019**, *10* (1), 1124.

(8) Sobus, J. R.; Wambaugh, J. F.; Isaacs, K. K.; Williams, A. J.; McEachran, A. D.; Richard, A. M.; Grulke, C. M.; Ulrich, E. M.; Rager, J. E.; Strynar, M. J. Integrating tools for non-targeted analysis

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

research and chemical safety evaluations at the US EPA. *J. Exposure Sci. Environ. Epidemiol.* **2018**, *28* (5), 411−426.
(9) Dogo, E. M.; Nwulu, N. I.; Twala, B.; Aigbavboa, C. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. *Urban Water J.* **2019**, *16* (3), 235−248.
(10) Wilde, F. D.; Radtke, D. B.; Gibs, J.; Iwatsubo, R. T. *National field manual for the collection of water-quality data: US Geological Survey Techniques of Water-Resources Investigations*, Book 9, **1998**.
(11) National Research Council. *Confronting the nation's water problems: The role of research*; National Academies Press, 2004.
(12) Bharti, R.; Grimm, D. G., Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics* **2021**.22178
(13) Veenaas, C., Developing tools for non-target analysis and digital archiving of organic urban water pollutants. **2018**.
(14) Keim, D. A. Visual exploration of large data sets. *Commun. ACM* **2001**, *44* (8), 38−44.
(15) Chen, J.; Chen, Y.; Du, X.; Li, C.; Lu, J.; Zhao, S.; Zhou, X. Big data challenge: a data management perspective. *Frontiers of Computer Science* **2013**, *7* (2), 157−164.
(16) Chakrabarti, S.; Cox, E.; Frank, E.; Güting, R. H.; Han, J.; Jiang, X.; Kamber, M.; Lightstone, S. S.; Nadeau, T. P.; Neapolitan, R. E. *Data mining: know it all*; Morgan Kaufmann, 2008.
(17) Gibert, K.; Sànchez-Marrè, M.; Izquierdo, J. A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. *AI Communications* **2016**, *29* (6), 627−663.
(18) Russell, S.; Norvig, P. *Artificial intelligence: a modern approach*. **2002**.
(19) Hinton, G. E.; Sejnowski, T. J.; Poggio, T. A. *Unsupervised Learning: Foundations of Neural Computation*; MIT press, 1999.
(20) Van Engelen, J. E.; Hoos, H. H. A survey on semi-supervised learning. *Machine Learning* **2020**, *109* (2), 373−440.
(21) Brink, H.; Richards, J. W.; Fetherolf, M.; Cronin, B. *Real-world machine learning*; Manning Shelter Island, NY, 2017.
(22) Kassambara, A. *Practical guide to cluster analysis in R: Unsupervised machine learning*; Sthda, ; Vol. *1*.
(23) Ghodsi, A. Dimensionality reduction a short tutorial. *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada* **2006**, *37* (38), 2006.
(24) Wright, R. E., Logistic regression. **1995**.
(25) Álvarez-Arbesú, R.; Felicísimo, A. M. GIS and logistic regression as tools for environmental management: a coastal cliff vegetation model in Northern Spain. *WIT Transactions on Information and Communication Technologies* **2002**, 26.
(26) Liu, L.; Sankarasubramanian, A.; Ranjithan, S. R. Logistic regression analysis to estimate contaminant sources in water distribution systems. *J. Hydroinf.* **2011**, *13* (3), 545−557.
(27) Thoe, W.; Gold, M.; Griesbach, A.; Grimmer, M.; Taggart, M. L.; Boehm, A. B. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. *Water Res.* **2014**, *67*, 105−117.
(28) Muharemi, F.; Logofătu, D.; Andersson, C.; Leon, F., Approaches to building a detection model for water quality: a case study. In *Modern Approaches for Intelligent Information and Database Systems*; Springer, 2018; pp 173−183.
(29) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R news* **2002**, *2* (3), 18−22.
(30) Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R. A tutorial review: Metabolomics and partial least squares-discriminant analysis-a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* **2015**, *879*, 10−23.
(31) Lee, S.; Kim, J. Prediction of Nanofiltration and Reverse-Osmosis-Membrane Rejection of Organic Compounds Using Random Forest Model. *J. Environ. Eng.* **2020**, *146* (11), 04020127.
(32) Castrillo, M.; García, Á. L. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Res.* **2020**, *172*, 115490.
(33) Tan, G.; Yan, J.; Gao, C.; Yang, S. Prediction of water quality time series data based on least squares support vector machine. *Procedia Eng.* **2012**, *31*, 1194−1199.
(34) Yang, Y. H.; Guergachi, A.; Khan, G. Support vector machines for environmental informatics: application to modelling the nitrogen removal processes in wastewater treatment systems. *Journal of Environmental Informatics* **2006**, *7* (1), 14−25.
(35) Feng, J.; Pan, L.; Cui, B.; Sun, Y.; Zhang, A.; Gong, M. New Approach for Concentration Prediction of Aqueous Phenolic Contaminants by Using Wavelet Analysis and Support Vector Machine. *Environmental Engineering Science* **2020**, *37* (5), 382−392.
(36) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A. Mastering the game of go without human knowledge. *Nature* **2017**, *550* (7676), 354−359.
(37) Mikolov, T.; Deoras, A.; Povey, D.; Burget, L.; Černocký, J. *Strategies for Training Large Scale Neural Network Language Models*; IEEE, 2011; pp 196−201.
(38) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. In *Imagenet: A Large-Scale Hierarchical Image Database*; Ieee, 2009; pp 248−255.
(39) Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12* (Aug), 2493−2537.
(40) Alwosheel, A.; van Cranenburgh, S.; Chorus, C. G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of choice modelling* **2018**, *28*, 167−182.
(41) Yang, J.; Xu, J.; Zhang, X.; Wu, C.; Lin, T.; Ying, Y. Deep learning for vibrational spectral analysis: Recent progress and a practical guide. *Anal. Chim. Acta* **2019**, *1081*, 6−17.
(42) Guo, H.; Jeong, K.; Lim, J.; Jo, J.; Kim, Y. M.; Park, J.-p.; Kim, J. H.; Cho, K. H. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *J. Environ. Sci.* **2015**, *32*, 90−101.
(43) Inoue, J.; Yamagata, Y.; Chen, Y.; Poskitt, C. M.; Sun, J. *Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning 2017*, 2017; IEEE: *2017*; pp 1058−1065.
(44) Voukantsis, D.; Karatzas, K.; Kukkonen, J.; Räsänen, T.; Karppinen, A.; Kolehmainen, M. Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2. 5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Sci. Total Environ.* **2011**, *409* (7), 1266−1276.
(45) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2* (1−3), 37−52.
(46) Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc., A* **2016**, *374* (2065), 20150202.
(47) Dalal, S. G.; Shirodkar, P. V.; Jagtap, T. G.; Naik, B. G.; Rao, G. S. Evaluation of significant sources influencing the variation of water quality of Kandla creek, Gulf of Katchchh, using PCA. *Environ. Monit. Assess.* **2010**, *163* (1−4), 49−56.
(48) Schollée, J. E.; Schymanski, E. L.; Avak, S. E.; Loos, M.; Hollender, J. Prioritizing unknown transformation products from biologically-treated wastewater using high-resolution mass spectrometry, multivariate statistics, and metabolic logic. *Anal. Chem.* **2015**, *87* (24), 12121−12129.
(49) Masiá, A.; Campo, J.; Blasco, C.; Picó, Y. Ultra-high performance liquid chromatography-quadrupole time-of-flight mass spectrometry to identify contaminants in water: an insight on environmental forensics. *Journal of Chromatography A* **2014**, *1345*, 86−97.
(50) Peiris, R. H.; Hallé, C.; Budman, H.; Moresoli, C.; Peldszus, S.; Huck, P. M.; Legge, R. L. Identifying fouling events in a membrane-based drinking water treatment process using principal component analysis of fluorescence excitation-emission matrices. *Water Res.* **2010**, *44* (1), 185−194.
(51) MacQueen, J. In *Some methods for classification and analysis of multivariate observations 1967*; Oakland, CA, 1967; pp 281−297.

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(52) Kingsy, G. R.; Manimegalai, R.; Geetha, D. M. S.; Rajathi, S.; Usha, K.; Raabiathul, B. N. In *Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data*; IEEE, 2016: pp 1945−1949.
(53) Zou, H.; Zou, Z.; Wang, X. An enhanced K-means algorithm for water quality analysis of the Haihe River in China. *Int. J. Environ. Res. Public Health* **2015**, *12* (11), 14400−14413.
(54) Areerachakul, S.; Sanguansintukul, S. *Clustering analysis of water quality for canals in Bangkok, Thailand*: Springer, 2010; pp 215−227.
(55) Javadi, S.; Hashemy, S. M.; Mohammadi, K.; Howard, K. W. F.; Neshat, A. Classification of aquifer vulnerability using K-means cluster analysis. *J. Hydrol.* **2017**, *549*, 27−37.
(56) Li, C.; Sun, L.; Jia, J.; Cai, Y.; Wang, X. Risk assessment of water pollution sources based on an integrated k-means clustering and set pair analysis method in the region of Shiyan, China. *Sci. Total Environ.* **2016**, *557*, 307−316.
(57) Brunner, A. M.; Bertelkamp, C.; Dingemans, M. M. L.; Kolkman, A.; Wols, B.; Harmsen, D.; Siegers, W.; Martijn, B. J.; Oorthuizen, W. A.; Ter Laak, T. L. Integration of target analyses, non-target screening and effect-based monitoring to assess OMP related water quality changes in drinking water treatment. *Sci. Total Environ.* **2020**, *705*, 135779.
(58) Gupta, S.; Arango-Argoty, G.; Zhang, L.; Pruden, A.; Vikesland, P. Identification of discriminatory antibiotic resistance genes among environmental resistomes using extremely randomized tree algorithm. *Microbiome* **2019**, *7* (1), 123.
(59) Du, X.; Shao, F.; Wu, S.; Zhang, H.; Xu, S. Water quality assessment with hierarchical cluster analysis based on Mahalanobis distance. *Environ. Monit. Assess.* **2017**, *189* (7), 1−12.
(60) Bibby, K.; Crank, K.; Greaves, J.; Li, X.; Wu, Z.; Hamza, I. A.; Stachler, E. Metagenomics and the development of viral water quality tools. *npj Clean Water* **2019**, *2* (1), 1−13.
(61) Chu, B. T. T.; Petrovich, M. L.; Chaudhary, A.; Wright, D.; Murphy, B.; Wells, G.; Poretsky, R. Metagenomics reveals the impact of wastewater treatment plants on the dispersal of microorganisms and genes in aquatic sediments. *Appl. Environ. Microbiol.* **2018**, *84* (5), No. e02168-17.
(62) Petrovich, M. L.; Zilberman, A.; Kaplan, A.; Eliraz, G. R.; Wang, Y.; Langenfeld, K.; Duhaime, M.; Wigginton, K.; Poretsky, R.; Avisar, D. Microbial and Viral Communities and Their Antibiotic Resistance Genes Throughout a Hospital Wastewater Treatment System. *Front. Microbiol.* **2020**, *11*, 153.
(63) Dai, D.; Rhoads, W. J.; Edwards, M. A.; Pruden, A. Shotgun metagenomics reveals taxonomic and functional shifts in hot water microbiome due to temperature setting and stagnation. *Front. Microbiol.* **2018**, *9*, 2695.
(64) Malla, M. A.; Dubey, A.; Yadav, S.; Kumar, A.; Hashem, A.; Abd_Allah, E. F. Understanding and designing the strategies for the microbe-mediated remediation of environmental contaminants using omics approaches. *Front. Microbiol.* **2018**, *9*, 1132.
(65) Giwa, A. S.; Ali, N.; Athar, M. A.; Wang, K. Dissecting microbial community structure in sewage treatment plant for pathogens' detection using metagenomic sequencing technology. *Arch. Microbiol.* **2020**, *202*, 1−9.
(66) Guo, J.; Ni, B.-J.; Han, X.; Chen, X.; Bond, P.; Peng, Y.; Yuan, Z. Data on metagenomic profiles of activated sludge from a full-scale wastewater treatment plant. *Data in brief* **2017**, *15*, 833−839.
(67) Laudadio, I.; Fulci, V.; Stronati, L.; Carissimi, C. Next-generation metagenomics: Methodological challenges and opportunities. *Omics: a journal of integrative biology* **2019**, *23* (7), 327−333.
(68) Soueidan, H.; Nikolski, M. Machine learning for metagenomics: methods and tools. *arXiv preprint arXiv:1510.06621* **2015**. DOI: 10.1515/metgen-2016-0001
(69) Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **2007**, *62* (2), 142−160.
(70) Yang, P.; Yu, S.; Cheng, L.; Ning, K. Meta-network: optimized species-species network analysis for microbial communities. *BMC Genomics* **2019**, *20* (2), 187.
(71) Collignon, P.; Beggs, J. J.; Walsh, T. R.; Gandra, S.; Laxminarayan, R. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. *Lancet Planetary Health* **2018**, *2* (9), No. e398-e405.
(72) Arango-Argoty, G.; Garner, E.; Pruden, A.; Heath, L. S.; Vikesland, P.; Zhang, L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **2018**, *6* (1), 1−15.
(73) El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47* (D1), D427−D432.
(74) Berglund, F.; Marathe, N. P.; Österlund, T.; Bengtsson-Palme, J.; Kotsakis, S.; Flach, C.-F.; Larsson, D. G. J.; Kristiansson, E. Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome* **2017**, *5* (1), 134.
(75) Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12* (ARTICLE), 2493−2537.
(76) Ng, P., dna2vec: Consistent vector representations of variable-length k-mers. *arXiv preprint arXiv:1701.06279* **2017**.
(77) Woloszynek, S.; Zhao, Z.; Chen, J.; Rosen, G. L. 16S rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses. *PLoS Comput. Biol.* **2019**, *15* (2), No. e1006721.
(78) Arango-Argoty, G. A.; Heath, L. S.; Pruden, A.; Vikesland, P.; Zhang, L. MetaMLP: A fast word embedding based classifier to profile target gene databases in metagenomic samples. *bioRxiv* **2019**, 569970.
(79) Liang, Q.; Bible, P. W.; Liu, Y.; Zou, B.; Wei, L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics* **2020**, *2* (1), No. lqaa009.
(80) Fiannaca, A.; La Paglia, L.; La Rosa, M.; Renda, G.; Rizzo, R.; Gaglio, S.; Urso, A. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinf.* **2018**, *19* (7), 198.
(81) Krawczyk, P. S.; Lipinski, L.; Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* **2018**, *46* (6), No. e35-e35.
(82) Hollender, J.; van Bavel, B.; Dulio, V.; Farmen, E.; Furtmann, K.; Koschorreck, J.; Kunkel, U.; Krauss, M.; Munthe, J.; Schlabach, M. High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. *Environ. Sci. Eur.* **2019**, *31* (1), 42.
(83) Soria, N. G. C.; Montes, A.; Bisson, M. A.; Atilla-Gokcumen, G. E.; Aga, D. S. Mass spectrometry-based metabolomics to assess uptake of silver nanoparticles by Arabidopsis thaliana. *Environ. Sci.: Nano* **2017**, *4* (10), 1944−1953.
(84) Matich, E. K.; Ghafari, M.; Camgoz, E.; Caliskan, E.; Pfeifer, B. A.; Haznedaroglu, B. Z.; Atilla-Gokcumen, G. E. Time-series lipidomic analysis of the oleaginous green microalga species Ettlia oleoabundans under nutrient stress. *Biotechnol. Biofuels* **2018**, *11* (1), 1−15.
(85) Matich, E. K.; Soria, N. G. C.; Aga, D. S.; Atilla-Gokcumen, G. E. Applications of metabolomics in assessing ecological effects of emerging contaminants and pollutants on plants. *J. Hazard. Mater.* **2019**, *373*, 527−535.
(86) Knolhoff, A. M.; Premo, J. H.; Fisher, C. M., A Proposed Quality Control Standard Mixture and Its Uses for Evaluating Nontargeted and Suspect Screening LC/HR-MS Method Performance. *Anal. Chem.* **2021**.931596
(87) Samanipour, S.; Kaserzon, S.; Vijayasarathy, S.; Jiang, H.; Choi, P.; Reid, M. J.; Mueller, J. F.; Thomas, K. V. Machine learning combined with non-targeted LC-HRMS analysis for a risk warning system of chemical hazards in drinking water: A proof of concept. *Talanta* **2019**, *195*, 426−432.
(88) Angeles, L. F.; Islam, S.; Aldstadt, J.; Saqeeb, K. N.; Alam, M.; Khan, M. A.; Johura, F.-T.; Ahmed, S. I.; Aga, D. S. Retrospective suspect screening reveals previously ignored antibiotics, antifungal

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

compounds, and metabolites in Bangladesh surface waters. *Sci. Total Environ.* **2020**, *712*, 136285.
(89) Albergamo, V.; Schollée, J. E.; Schymanski, E. L.; Helmus, R.; Timmer, H.; Hollender, J.; De Voogt, P. Nontarget screening reveals time trends of polar micropollutants in a riverbank filtration system. *Environ. Sci. Technol.* **2019**, *53* (13), 7584−7594.
(90) Chiaia-Hernández, A. C.; Scheringer, M.; Müller, A.; Stieger, G.; Wächter, D.; Keller, A.; Pintado-Herrera, M. G.; Lara-Martin, P. A.; Bucheli, T. D.; Hollender, J. Target and suspect screening analysis reveals persistent emerging organic contaminants in soils and sediments. *Sci. Total Environ.* **2020**, *740*, 140181.
(91) Cariou, R.; Méndez-Fernandez, P.; Hutinet, S. b.; Guitton, Y.; Caurant, F.; Le Bizec, B.; Spitz, J. r. m.; Vetter, W.; Dervilly, G., Nontargeted LC/ESI-HRMS Detection of Polyhalogenated Compounds in Marine Mammals Stranded on French Atlantic Coasts. *ACS ES&T Water* **2021**.1309
(92) Travis, S. C.; Kordas, K.; Aga, D. S. Optimized workflow for unknown screening using gas chromatography high-resolution mass spectrometry expands identification of contaminants in silicone personal passive samplers. *Rapid Commun. Mass Spectrom.* **2021**, *35* (8), No. e9048.
(93) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. *Nontarget screening with high resolution mass spectrometry in the environment: ready to go?*; ACS Publications, 2017.
(94) Müller, A.; Schulz, W.; Ruck, W. K. L.; Weber, W. H. A new approach to data evaluation in the non-target screening of organic trace substances in water analysis. *Chemosphere* **2011**, *85* (8), 1211−1219.
(95) Schollée, J. E.; Schymanski, E. L.; Hollender, J., Statistical approaches for LC-HRMS data to characterize, prioritize, and identify transformation products from water treatment processes. In *Assessing Transformation Products of Chemicals by Non-Target and Suspect Screening- Strategies and Workflows Vol. 1*; ACS Publications, 2016; pp 45−65.
(96) Sirén, K.; Fischer, U.; Vestner, J. Automated supervised learning pipeline for non-targeted GC-MS data analysis. *Analytica Chimica Acta: X* **2019**, *1*, 100005.
(97) Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011**, *83* (4), 1079−1087.
(98) Johnsen, L. G.; Amigo, J. M.; Skov, T.; Bro, R. Automated resolution of overlapping peaks in chromatographic data. *J. Chemom.* **2014**, *28* (2), 71−82.
(99) Smirnov, A.; Jia, W.; Walker, D. I.; Jones, D. P.; Du, X. ADAP-GC 3.2: graphical software tool for efficient spectral deconvolution of gas chromatography-high-resolution mass spectrometry metabolomics data. *J. Proteome Res.* **2018**, *17* (1), 470−478.
(100) Chen, C.; Husny, J.; Rabe, S. Predicting fishiness off-flavour and identifying compounds of lipid oxidation in dairy powders by SPME-GC/MS and machine learning. *Int. Dairy J.* **2018**, *77*, 19−28.
(101) Taghadomi-Saberi, S.; Mas Garcia, S.; Allah Masoumi, A.; Sadeghi, M.; Marco, S. Classification of bitter orange essential oils according to fruit ripening stage by untargeted chemical profiling and machine learning. *Sensors* **2018**, *18* (6), 1922.
(102) Yang, Q.; Xu, L.; Tang, L.-J.; Yang, J.-T.; Wu, B.-Q.; Chen, N.; Jiang, J.-H.; Yu, R.-Q. Simultaneous detection of multiple inherited metabolic diseases using GC-MS urinary metabolomics by chemometrics multi-class classification strategies. *Talanta* **2018**, *186*, 489−496.
(103) Smolinska, A.; Hauschild, A. C.; Fijten, R. R. R.; Dallinga, J. W.; Baumbach, J.; Van Schooten, F. J. Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis. *J. Breath Res.* **2014**, *8* (2), 027105.
(104) Garrido-Baserba, M.; Corominas, L.; Cortés, U.; Rosso, D.; Poch, M. The fourth-revolution in the water sector encounters the digital revolution. *Environ. Sci. Technol.* **2020**, *54* (8), 4698−4705.
(105) Vujnović, G.; Perišić, J.; Božilović, Z.; Milovanović, M.; Getman, R. V.; Radovanović, L. USING SCADA SYSTEM FOR PROCESS CONTROL IN WATER INDUSTRY. *Acta Technica Corviniensis-Bulletin of Engineering* **2019**, *12* (2), 67−72.
(106) Olsson, G. ICA and me-a subjective review. *Water Res.* **2012**, *46* (6), 1585−1624.
(107) Zhao, H.; Hou, D.; Huang, P.; Zhang, G. Water quality event detection in drinking water network. *Water, Air, Soil Pollut.* **2014**, *225* (11), 2183.
(108) Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)* **2009**, *41* (3), 1−58.
(109) Hasan, J.; Goldbloom-Helzner, D.; Ichida, A.; Rouse, T.; Gibson, M. *Technologies and Techniques for Early Warning Systems to Monitor and Evaluate Drinking Water Quality: A State-of-the-Art Review*; Environmental Protection Agency Washington Dc Office of Water, 2005.
(110) Izquierdo, J.; López, P. A.; Martínez, F. J.; Pérez, R. Fault detection in water supply systems using hybrid (theory and data-driven) modelling. *Mathematical and Computer Modelling* **2007**, *46* (3−4), 341−350.
(111) Murray, S.; Ghazali, M.; McBean, E. A. Real-time water quality monitoring: assessment of multisensor data using Bayesian belief networks. *Journal of Water Resources Planning and Management* **2012**, *138* (1), 63−70.
(112) Raciti, M.; Cucurull, J.; Nadjm-Tehrani, S., Anomaly detection in water management systems. In *Critical Infrastructure Protection*; Springer, 2012; pp 98−119.
(113) Ahmed, M.; Mahmood, A. N.; Hu, J. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications* **2016**, *60*, 19−31.
(114) Yuan, Y.; Jia, K. *A Water Quality Assessment Method Based on Sparse Autoencoder 2015*; IEEE, 2015; pp 1−4.
(115) Muharemi, F.; Logofătu, D.; Leon, F. Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication* **2019**, *3* (3), 294−307.
(116) Alom, M. Z.; Taha, T. M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M. S.; Hasan, M.; Van Essen, B. C.; Awwal, A. A. S.; Asari, V. K. A state-of-the-art survey on deep learning theory and architectures. *Electronics* **2019**, *8* (3), 292.
(117) Russo, S.; Lürig, M.; Hao, W.; Matthews, B.; Villez, K. Active learning for anomaly detection in environmental data. *Environmental Modelling & Software* **2020**, *134*, 104869.
(118) Ahmad, S.; Lavin, A.; Purdy, S.; Agha, Z. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* **2017**, *262*, 134−147.
(119) Stocco, A.; Tonella, P. *Towards Anomaly Detectors that Learn Continuously*; IEEE, 2020; pp 201−208.
(120) Hoffman, M.; Bach, F. R.; Blei, D. M. In *Online learning for latent dirichlet allocation* 2010, Citeseer: pp 856−864.
(121) Mongillo, G.; Deneve, S. Online learning with hidden Markov models. *Neural computation* **2008**, *20* (7), 1706−1716.
(122) Bogler, A.; Packman, A.; Furman, A.; Gross, A.; Kushmaro, A.; Ronen, A.; Dagot, C.; Hill, C.; Vaizel-Ohayon, D.; Morgenroth, E. Rethinking wastewater risks and monitoring in light of the COVID-19 pandemic. *Nature Sustainability* **2020**, *9*, 1−10.
(123) Bivins, A.; North, D.; Ahmad, A.; Ahmed, W.; Alm, E.; Been, F.; Bhattacharya, P.; Bijlsma, L.; Boehm, A. B.; Brown, J. *Wastewater-Based Epidemiology: Global Collaborative to Maximize Contributions in the Fight Against COVID-19*; ACS Publications, 2020.

https://doi.org/10.1021/acs.est.1c01026
Environ. Sci. Technol. XXXX, XXX, XXX−XXX