ANNUAL REVIEWS

Annual Review of Food Science and Technology

Emerging Applications of Machine Learning in Food Safety

Xiangyu Deng,[1] Shuhao Cao,[2] and Abigail L. Horn[3]

[1]Center for Food Safety, University of Georgia, Griffin, Georgia 30223, USA; email: xdeng@uga.edu

[2]Department of Mathematics and Statistics, Washington University, St. Louis, Missouri 63105, USA; email: s.cao@wustl.edu

[3]Department of Preventive Medicine, University of Southern California, Los Angeles, California 90032, USA; email: abigaillhorn@gmail.com

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Annu. Rev. Food Sci. Technol. 2021. 12:22.1–22.26

The *Annual Review of Food Science and Technology* is online at food.annualreviews.org

https://doi.org/10.1146/annurev-food-071720-024112

Copyright © 2021 by Annual Reviews.
All rights reserved

Keywords

public health, genomes, text data, transactional data, trade data, novel data streams

Abstract

Food safety continues to threaten public health. Machine learning holds potential in leveraging large, emerging data sets to improve the safety of the food supply and mitigate the impact of food safety incidents. Foodborne pathogen genomes and novel data streams, including text, transactional, and trade data, have seen emerging applications enabled by a machine learning approach, such as prediction of antibiotic resistance, source attribution of pathogens, and foodborne outbreak detection and risk assessment. In this article, we provide a gentle introduction to machine learning in the context of food safety and an overview of recent developments and applications. With many of these applications still in their nascence, general and domain-specific pitfalls and challenges associated with machine learning have begun to be recognized and addressed, which are critical to prospective use and future deployment of large data sets and their associated machine learning models for food safety applications.

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

1. INTRODUCTION

Foodborne illnesses remain a substantial and enduring burden on public health. An estimated 1 in 6 Americans (or 48 million people) is sickened by foodborne illness each year, causing 128,000 hospitalizations and 3,000 deaths (Scallan et al. 2011). In 2010, the Healthy People initiative of the United States issued its 2020 vision, which designated food safety as a focus area (Koh 2010). None of the vision's objectives for controlling six major foodborne pathogens by 2020 had been met as recently as 2019, according to surveillance data by the Foodborne Diseases Active Surveillance Network (FoodNet) (**Table 1**).

Over more than a century, major transformations in food production, distribution, and regulation have taken place, driven by and feeding into macrosocietal trends such as population increase, urbanization, and globalization (Doyle et al. 2015, Phillips 2006). Massive changes and advances in the food industry and supply chains have generated large volumes of data, especially in recent years, similar to in other sectors and industries. A plethora of data has been explored in innovative ways and at different stages along the farm-to-table continuum to improve the safety of the food supply. For instance, at preharvest, terrain and meteorological data were investigated for predicting pathogen contamination on produce farms (Strawn et al. 2013), and in the retail setting, paperless auditing and record keeping enabled 1.4 million monthly measurements of internal cooking temperatures of rotisserie chickens for food safety assurance (Yiannas 2015).

At the end of the food supply chain, consumer interactions with foods, including transaction, consumption, and experience feedback and sharing, also create copious amounts of data. These novel data streams (NDS) are increasingly propagated and accessible via digital platforms such as social media, search histories, crowdsourcing sites, and consumer reviews and commentary, as well as databases of product sales and consumption records. Mining of these data to inform food safety and public health is on the horizon (Harris et al. 2014, Maharana et al. 2019).

On the surveillance front, data-intensive systems play important roles in tracking foodborne illness cases and agents. Examples at the US federal level include PulseNet (Swaminathan et al. 2001), the National Antimicrobial Resistance Monitoring System (NARMS) (Gupta et al. 2004, Zhao et al. 2006), FoodNet (Scallan & Mahon 2012), and the National Outbreak Reporting System (Hall et al. 2013). Data collected by some of these systems have surged in the recent decade owing to the incorporation of genomic data on foodborne pathogens. Implementation of whole-genome sequencing (WGS) in surveillance and outbreak investigation has fueled an explosion of publicly available foodborne pathogen genomes in new systems such as GenomeTrakr (Allard

Table 1 Healthy People 2020 objectives and 2019 preliminary data

Pathogen[a]	Healthy People 2020 objective[b]	2019 preliminary data[c]
Campylobacter	8.5[d]	19.5
Salmonella	11.4	17.1
Shiga toxin–producing *Escherichia coli*	0.6	6.3
Listeria	0.2	0.3
Vibrio	0.2	0.9
Yersinia	0.3	1.4

[a]Pathogens with a Healthy People 2020 objective.
[b]Data from Healthy People (2010).
[c]Data from Foodborne Diseases Active Surveillance Network (Tack et al. 2020).
[d]Incidence rate, per 100,000 population.

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

et al. 2016), EnteroBase (Zhou et al. 2020), and the National Center for Biotechnology Information's Pathogen Detection (**https://www.ncbi.nlm.nih.gov/pathogens**). Routine use of WGS in public health microbiology has given rise to a data-driven area known as genomic epidemiology (Deng et al. 2016).

Recent advances in the data science approach to food safety have led to the discussion of Big Data (Marvin et al. 2017), a term that is not traditionally associated with food safety. To meet analytical challenges created by the deluge of data, machine learning has emerged as a promising tool for data-intensive analytics in food safety. In April 2019, the Food and Drug Administration (FDA) released a statement on "steps to usher the US into a new era of smarter food safety," in which artificial intelligence and machine learning applications in food safety were proposed (Sharpless & Yiannas 2019).

Artificial intelligence: a computer science branch that simulates behaviors commonly associated with intelligent beings, such as learning, problem-solving, and decision-making

Given the rapid emergence of machine learning applications in food safety, we aim to provide a comprehensive overview of the new field by introducing fundamentals of the methodology, reviewing recent and notable progress, and discussing challenges and potential pitfalls. Machine learning, as a general-purpose data analytics tool, has been used in other areas of agricultural and food science, such as food processing and quality evaluation, as reviewed elsewhere (Du & Sun 2006). In this review, we focus on domain-specific applications in food safety and public health.

2. INTRODUCTION TO MACHINE LEARNING FUNDAMENTALS

2.1. A Very Brief History of Machine Learning

Arthur Samuel (1959, p. 211) coined the term machine learning to demonstrate how a computer could acquire the ability to play and excel at the game of checkers in a way that, "if done by human beings or animals, would be described as involving the process of learning." Samuel's definition was later generalized as the field of study that gives computers the ability to learn without being explicitly programmed. The practice thereof dates back to the early eighteenth century; astronomers and geodesists introduced least-squares methods to describe planetary orbits based on measurements (data) to help sailors navigate the ocean (Stigler 1986). The theory and tools of modern machine learning were blueprinted by visionaries like Alan Turing after World War II (Turing 1950). Some of the most widely used algorithms and models, such as nearest neighbors, random forests, and neural networks, were invented between the 1960s and the 1990s. After the term Big Data was popularized in both the scientific community and the general public around the 1990s and 2000s, the explosive growth of machine learning benefited from vastly increasing data sizes, exponentially growing computer power, and new refinements of old tools, eventually leading to a myriad of breakthroughs. Notable milestones include recognition of handwritten digits (LeCun et al. 1989) and speech (Hochreiter & Schmidhuber 1997); classification of objects such as cats, dogs, and planes (Krizhevsky et al. 2012); and mastery of gameplay without human knowledge in the game of Go (Silver et al. 2017).

2.2. Why Machine Learning

As a subfield of artificial intelligence, machine learning differs from traditional algorithmic problem-solving by not attempting to program an exhaustive list of explicit instructions or rules. Instead, a machine learning system learns from examples and generalizes to new cases based on their closeness to learned examples (instance-based learning) or trains a model with data to learn its parameters through optimization and makes predictions using new (test) data (model-based learning).

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Training data (model training): a set of examples/data used to formulate a model and determine its structures or parameters

Transactional data: electronic records generated at point-of-sale, including checkout transactions, retail sales, credit card transaction history, and online grocery/delivery data

The data-driven and rule-agnostic characteristics of machine learning make it attractive for certain types of tasks. First, for problems that are difficult to pose mathematically (Shardanand & Maes 1995) or without explicit solution algorithms, machine learning may find a reasonable approximation. Second, some problems are so complex for existing methods that a prohibitively long list of rules would need to be programmed for their solutions. For example, for the game of Go, it is combinatorically unrealistic to find the optimal move through a brute-force search. Third, some tasks must cope with new data to which hard-coded rules are impossible to adapt, such as detecting novel spam in emails and on social media. Finally, machine learning can provide insights and verify heuristics on large-scale problems, such as the Go strategies and tactics innovated and rediscovered by DeepMind's AlphaGo (Baker & Fan 2017).

2.3. Major Categories of Machine Learning Methods

During training, machine learning systems may receive guidance or supervision. Based on the amounts of supervision provided, the learning can be categorized into supervised, unsupervised, semi-supervised, and reinforced.

In typical supervised learning tasks, such as classification and regression, the training data fed to the learning system are labeled with the desired outcome or the ground truth. To build a cat/dog classifier, a training set of many pet images must be assembled and labeled with the classes: cats, dogs, or neither. To develop a regressor that predicts a continuous numeric value, such as housing prices given a set of features (e.g., neighborhood, size, year built), many instances of houses are collected to fit a regression model, each including both features and a label: its price.

In unsupervised learning, training data are unlabeled, leaving the algorithm to unearth the hidden patterns. An example is the identification of customer groups through behavioral/transactional data without an a priori defined grouping. Another important application is anomaly and novelty detection. For example, a learning system shown mostly normal network traffic can learn to detect cyber intrusions.

Labeling of data can be labor intensive and is not readily available for large data sets. There are often few labeled instances among many unlabeled examples. As a combination of supervised and unsupervised learning, semi-supervised algorithms can weigh in on unlabeled data's contribution to feature-target relations, usually taking advantage of the assumption that nearby samples are likely to share the same labels (Zhu et al. 2003). For example, in automatic speech recognition, accented speech is commonly underrepresented in training data and problematic for supervised learning. Semi-supervised learning of tone and pitch accent has been shown to reduce the need for labeled training data for speech recognition (Levow 2006).

Unlike supervised learning, in which training data comes with specific answers to the question (class labels), reinforcement learning relies on a learning system (agent) to find the best strategy or path (policy) in a given situation. The learning is achieved through a trial-and-error process during which the agent is rewarded or penalized by the actions it takes, with the goal of maximizing the reward over time. Reinforcement finds plenty of use in robotics and gaming, from robots learning to walk (Haarnoja et al. 2018) to the AlphaGo program beating the world Go champion (Silver et al. 2017).

2.4. Examples of Algorithms

Numerous machine learning algorithms have been developed that vary in sophistication to accommodate problems of different levels of complexity. Four representative and fundamental learning algorithms are summarized in **Figure 1**.

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

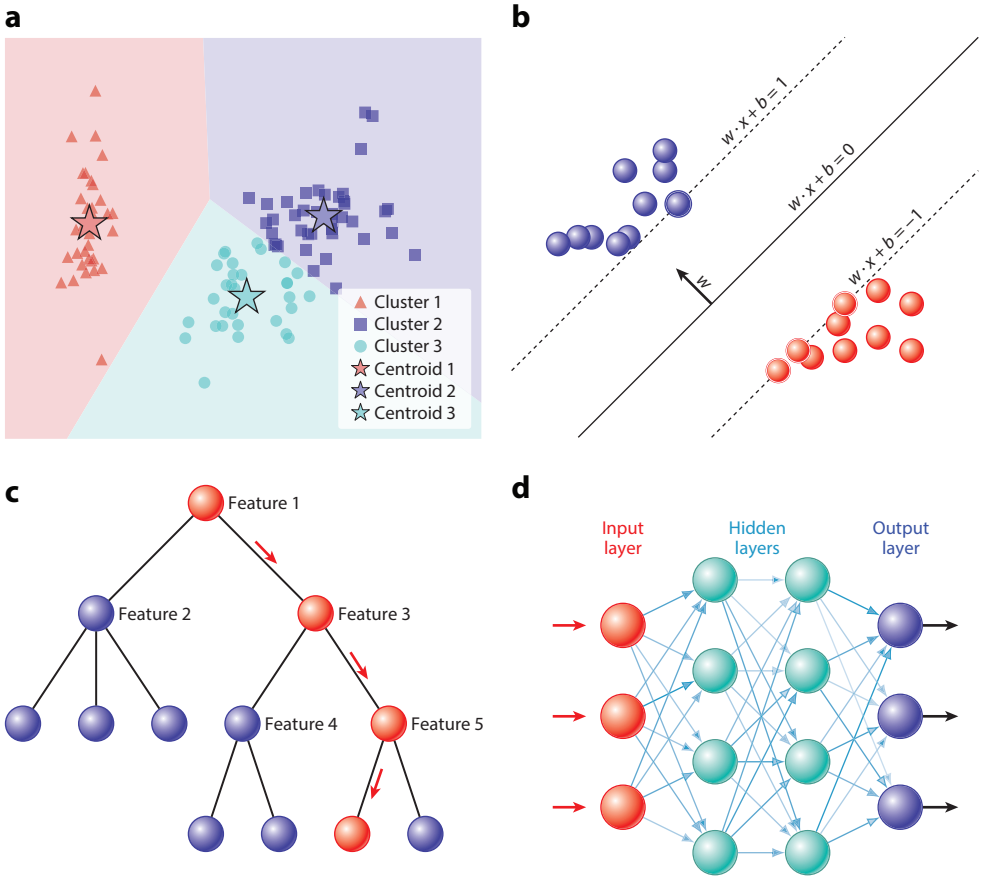

Figure 1

Examples of machine learning models. (*a*) A decision boundary plot of *k*-means clustering with three clusters, with new samples being grouped to a cluster by the colored region it lands on. (*b*) A line dividing two classes in a support-vector machine with a certain margin. *w* contains the trainable parameters, and *x* stands for the vector representation of a sample. (*c*) A decision tree with five features. A sample is classified into a certain class following the red arrow. (*d*) A neural network with two hidden layers; the arrow stands for a connection between units, with transparency indicating the connection strength. The Python source code to generate panel *a* was adapted from **https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html** under a BSD license. Panel *b* was adapted from **https://commons.wikimedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png**, under a Creative Commons license CC0. Panel *c* was adapted from **https://texample.net/tikz/examples/red-black-tree**. Panel *d* was adapted from **https://texample.net/tikz/examples/neural-network**, under a Creative Commons license 2.5.

K-means is an unsupervised algorithm that partitions similar observations into clusters dynamically. It uses geometric centers of observations (centroids) to prototype clusters, and an observation is then assigned to a cluster if it is closer to the cluster's centroid than any other centroid (**Figure 1*a***).

A support-vector machine (SVM) uses planes to best separate observations of different classes by representing them as points. A new observation is mapped onto the space, and its class is predicted according to the side of the plane on which it falls. SVM is particularly efficient for data in which different classes are well separated (**Figure 1*b***).

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Decision trees (**Figure 1*c***) attempt to split instances into different classes recursively through the interaction of different features. A new sample follows particular branches determined by the features to land on a leaf that provides its predicted class. It is possible to randomize this approach through averaging the results from a multitude of different trees (random forest) and to grow trees by following a more quantifiable criterion through the introduction of an objective function (gradient boosting). Both methods are examples of ensemble learning.

An artificial neural network (ANN) simulates biological neural networks by comprising layers of interconnected artificial neurons called processing units (**Figure 1*d***). These units or nodes receive input information and process it through a system that includes a linear combination of weights and input, a nonlinear activation function, and output signals to the next layer of nodes. Each ANN consists of one input layer, one or more intermediate layers called hidden layers, and one output layer. Together, they convert initial inputs into results for regression or classification tasks. An ANN containing many hidden layers is called a deep neural network, which is the core of deep learning.

2.5. Good Practices, Challenges, and Pitfalls

The computer science proverb "garbage in, garbage out" applies to all data-driven machine learning models, as they are highly dependent on how well the data are prepared. An array of feature engineering techniques can be applied to raw data, such as selecting useful features and dealing with missing information (e.g., data imputation/augmentation). It is a common tendency in machine learning practice to include many features at first to maximize the capture of underlying patterns and associations. However, beyond a certain point, the performance of the learned model deteriorates as the feature number increases, a phenomenon often referred to as "the curse of dimensionality" (Friedman 1997). Mathematically proven dimension reduction methods, such as principal component analysis, can be used to reduce the number of features. Evaluating the impact of subsets of features on model performance can be used to select features as well (Kohavi & John 1997). A major faulty practice in data preparation is data leakage that may lead to an overgeneralization of the model. The term leakage refers to the case in which the train/test data sets are not properly split so that the model sees the test set information while being trained.

The second common problem is data bias. For example, when using social network data to predict public opinion, one should consider the demographical and behavioral differences between the studied population and the target population. Overfitting/underfitting, or the commonly known bias–variance tradeoff problem, is another bias affecting supervised models. Affected predictions cannot capture all patterns of the testing data at the same accuracy as with the training data. For example, making a model's structure more sophisticated may increase accuracy in representing the relations in training data. However, by doing so, one risks overfitting, as predictions on the testing set may become very unstable and have large variance. Certain regularization techniques can be used to tame overfitting.

When a model is not performing as expected, one might be inclined to switch to another model or algorithm. However, it is also possible that the model has not seen enough data. The example of natural language disambiguation shows that different learning algorithms, including rather simple ones, can perform almost equally well when given enough data (Banko 2001). There is a notion that data matter more than algorithms for complex problems (Halevy et al. 2009).

Finally, one shall not "listen (only) to the data" (Nisbet et al. 2009, p. 739). A model's performance on a single measure, such as classification accuracy, is inadequate for determining its deployability. Spurious discoveries may occur from highly dimensional training sets, the inherent randomness of the model, or model overfitting. Interpretation of how a model arrives at its findings is therefore critical but sometimes challenging. Some black-box models may concurrently reduce

Ensemble learning: combining various components from different models to improve the generalization

Activation function: a function that decides how important a unit is in computing the output of a certain layer

Feature engineering: the process of transforming raw data into a data set that can be used as input for machine learning analysis

Dimension reduction: extraction of essential variables that sufficiently describe the original data, yet much more compact than the original

Test data: data that are unseen by the model and used to test whether the final model can generalize

Overfitting: the model localizes its approximation power on certain parts of data and fails to capture the whole trend

Underfitting: the model does not have the capacity or is not trained properly to capture the trend

Regularization: an Occam's razor to control the complexity of the model to prevent overfitting

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

variance while increasing predictive power, although at the cost of losing unbiasedness (Hastie et al. 2009) and parameter-prediction mechanics that are comprehensible by humans. There has been an argument in favor of inherently interpretable models, especially when leveraging machine learning in high-stakes decision-making (Rudin 2019).

3. MACHINE LEARNING APPLICATIONS USING GENOMIC DATA

In genetics, genomics, and medicine, machine learning holds promise for making biological discoveries and predictions from large genomic data sets. Machine learning models have been trained to recognize patterns and elements in DNA sequences, a process known as sequence annotation (Libbrecht & Noble 2015). Genomic signatures or biomarkers have been identified via machine learning techniques to assist in disease diagnosis, clinical decision-making, and drug discovery and development (He et al. 2019, Vamathevan et al. 2019). One could assume that machine learning analysis of foodborne pathogen genomes is an iteration or extension of established methodology and therefore straightforward owing to the relatively small genomes of the pathogens. However, domain-specific opportunities and challenges continue to arise as machine learning is increasingly used to tap into the rapidly growing resources of foodborne pathogen genomes and their metadata. Still in their infancy, such applications have been focused on antimicrobial resistance (AMR) prediction and genomic source attribution of certain pathogens.

3.1. Antimicrobial Resistance Prediction

Measurement of antimicrobial resistance or susceptibility traditionally relies on phenotypic assays that measure growth inhibition of an antibacterial agent on a population of pure culture bacteria. A common technique for antimicrobial susceptibility testing (AST) is broth dilution, which involves a range of antibiotic concentrations and determines the minimum inhibitory concentration (MIC) of a drug to inactivate or inhibit the growth of a particular bacterial isolate. Clinical breakpoints are assigned to divide AST results into categories in correlation with the likelihood of treatment outcome, including susceptible (high probability of a favorable outcome), resistant (low probability of a favorable outcome), and sometimes intermediate (Humphries et al. 2019, Turnidge & Paterson 2007).

Phenotypic AST is challenging to standardize across laboratories and time consuming for generating clinically actionable results. In clinical settings, accurate and rapid AST could inform timely clinical decision-making and improve antibiotic stewardship, whereas excessive prescription of antibiotics that the etiologic agents are resistant to can delay effective treatment and contribute to hospital spread of resistant strains. In public health monitoring of AMR, WGS has become routine. NARMS has started conducting WGS of all non-typhoidal *Salmonella* from clinical sources regardless of phenotypic resistance (NARMS 2015).

WGS-based AST has been widely studied as an enhancement or alternative to phenotypic AST (Ellington et al. 2017, Hendriksen et al. 2019). Initial efforts focused on a rule-based approach that requires a priori knowledge of AMR determinants. Typically, a curated panel of AMR genes is queried in pathogen genomes to yield binary classification as susceptible or resistant to corresponding drugs. Despite performing well on several enteric foodborne pathogens (Feldgarden et al. 2019, McDermott et al. 2016, Zankari et al. 2013), this approach is limited to known resistance determinants and unable to identify AMR conferred by novel or uncatalogued genes. Furthermore, AMR polygenically determined by multiple genes or mediated by minor mutations is difficult to identify.

Machine learning has naturally been attempted in hopes of overcoming some of the aforementioned limitations and capitalizing on the surge of WGS data in combating AMR (**Table 2**).

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Table 2 Selected studies on antimicrobial resistance prediction using WGS and machine learning

Organism	Machine learning model	Prediction type	Size of training set	Features	Number of drugs	Reference
Salmonella enterica	XGBoost	MIC determination	5,278	*k*-mer	15	Nguyen et al. 2019
S. enterica	LR, SCM	AMR classification	97	AMR genes (LR), *k*-mer (SCM)	7	Maguire et al. 2019
Klebsiella pneumonia	XGBoost, AdaBoost, bagging, random forest, SVM, extremely randomized trees	MIC determination	1,668	*k*-mer	20	Nguyen et al. 2018
Neisseria gonorrhoeae	Multivariate linear regression	MIC determination	670–681	SNPs, deletions, genes	5	Eyre et al. 2017
Streptococcus pneumonia	Mode MIC, random forest, elastic net	MIC determination	2,528	AMR genes	1	Li et al. 2016
Clostridium difficile, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *S. pneumonia*	SCM	AMR classification	111–556	*k*-mer	3–5 for each organism	Drouin et al. 2016
Escherichia coli, *Enterobacter aerogenes*, *Enterobacter cloacae*, *K. pneumonia*	LR	Susceptibility classification	78	AMR genes	12	Pesesky et al. 2016
Acinetobacter baumannii, *Staphylococcus aureus*, *S. pneumoniae*, *M. tuberculosis*	AdaBoost	AMR classification	99–1,350	*k*-mer	1–5 for each organism	Davis et al. 2016
M. tuberculosis	LR, SVM	AMR classification	652	SNPs	4	Niehaus et al. 2014

Abbreviations: AMR, antimicrobial resistance; LR, logistic regression; MIC, minimum inhibitory concentration; SCM, Set Covering Machine; SNP, single-nucleotide polymorphism; SVM, support vector machine; WGS, whole-genome sequencing.

Earlier studies established the feasibility of training machine learning models for AMR prediction in multiple clinically important bacteria (Macesic et al. 2017), such as *Mycobacterium tuberculosis* (Niehaus et al. 2014), *Staphylococcus aureus* (Davis et al. 2016), *Streptococcus pneumonia* (Davis et al. 2016, Drouin et al. 2016, Li et al. 2016), and *Neisseria gonorrhoeae* (Eyre et al. 2017). Studies on foodborne pathogens quickly ensued, taking advantage of the established infrastructure of AMR

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

monitoring and algorithmic processes in machine learning. These studies fall into two categories, as detailed below.

3.1.1. Categorical classification. Categorical interpretation of AST results is preferred by most clinicians (Turnidge & Paterson 2007) and amenable to classification by machine learning, one of the most common applications of supervised learning. Pesesky et al. (2016) reported a comparative study between machine learning and rule-based classification of AMR in 78 *Enterobacteriaceae* isolates, including *Escherichia coli*. Both types of classification relied on a set of known resistance proteins curated by the Resfams database (Gibson et al. 2015). The machine learning method trained a single logistic regression model for 6 classes of 12 antibiotics, whereas the rules-based method employed separately tailored rules for particular classes. The two classifiers achieved similar performance, agreeing with phenotypic AST for ∼90% of the isolates analyzed. Neither classifier was deemed accurate enough as a primary diagnostic by clinical standards, likely in part owing to the small training set with a bias toward highly resistant isolates. The study outlined future refinements of the method. In particular, it showed that lack or underrepresentation of certain resistance determinants in feature selection prevented effective learning of their contribution to AMR, especially when rare or novel AMR-conferring genes were involved.

Drouin et al. (2016) developed a reference-free approach for classification and biomarker identification of AMR in four bacterial species, including the foodborne pathogen *Clostridium difficile*. Using a *k*-mer representation of training genomes as input features (each genome is represented by a set of unique subsequences of length *k*), the model was agnostic of known AMR determinants and unbiased by any particular curation of AMR genes. Because its learning was not limited by existing knowledge of AMR mechanisms, the model could potentially identify novel AMR biomarkers.

By using *k*-mers, the study faced the curse-of-dimensionality challenge. Depending on the *k*-mer length and genome size, a *k*-mer representation of a genome is often highly dimensional, consisting of many *k*-mers that greatly outnumber the genomes they represent. In machine learning, the high dimensionality of feature space makes the model susceptible to overfitting. For AMR classification using *k*-mers, an overfit model fits a training set deceptively well by overly drawing on the noise and detail in the training data, such as random and spurious relationships between *k*-mers and AMR profiles specific to the training set, which degrades its performance when applied to a new test set. The study made extensions to the Set Covering Machine (SCM) algorithm to use the entire feature space consisting of all the *k*-mers, ranging from 12 to 123 million ($k = 31$). The SCM model was reported to not overfit and outperformed other models that reduced feature space via additional future selection. Furthermore, the authors performed a theoretical calculation on the upper bound of the error rate to show that the *k*-mer-based SCM model was not prone to overfitting even though there were far more features than examples (genomes).

By measuring the importance of individual *k*-mer features in AMR classification, the SCM model was found to rely on the fewest *k*-mers among tested machine learning models to make robust AMR predictions. This advantage, combined with the coverage of the entire feature space, allowed the SCM model to screen AMR biomarkers throughout genomes and converge on critical *k*-mers of biological relevance to AMR. The authors suggested that the potential to generate a minimal set of biomarkers could facilitate the interpretation of machine learning results by domain experts and the translation of these results into common clinical diagnostics such as polymerase chain reaction, as how to interpret and translate machine learning findings remains a barrier to practical applications.

The reference-free SCM model was later applied to a set of 97 non-typhoidal *Salmonella enterica* isolates sampled from chicken broiler farms in British Columbia, Canada (Maguire et al. 2019).

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Similarly, the model was able to learn and identify primary drivers of AMR for seven antibiotics studied. Also evaluated in this study were two gene-centric methods: a logistic regression classifier that used annotated AMR genes as features and a direct tallying of these genes to determine AMR in query genomes. The SCM and the logistic regression classifiers achieved more than 0.9 precision in classifying AMR to all seven antibiotics, outperforming the AMR gene-tallying method, which massively overpredicted AMR (precision 0 to 0.5). These results suggest limitations of AMR classification by direct AMR gene profiling, which may include incorrect AMR assignment to paralogs of AMR genes and inability to identify intragenic determinants of AMR such as promoter and regulatory sequences. Restricted by the training set, the authors noted the challenge of label imbalance where several antibiotics (labels) were underrepresented by resistant isolates and the inability to predict MIC owing to the small size of the set.

3.1.2. Minimum inhibitory concentration prediction. Compared with categorical classification, MIC measurement further enables granular characterization and precise monitoring of AMR. Although proposed in aforementioned studies (Drouin et al. 2016, Maguire et al. 2019, Pesesky et al. 2016), MIC prediction is difficult with training sets that are not big enough for robust regression modeling.

Li et al. (2016) and Eyre et al. (2017) predicted MIC levels by using training models with more than 2,500 *S. pneumoniae* genomes and more than 600 *N. gonorrhoeae* genomes, respectively. Both studies used curated AMR determinants as features, showing accurate MIC prediction by machine learning with well-characterized AMR genes and mutations.

Nguyen et al. (2019) took the *k*-mer-based, reference-free approach to MIC prediction in nontyphoidal *Salmonella* in one of the most extensive machine learning studies on AMR, as measured by training set size and the number of drugs investigated (**Table 2**). The training set included 5,728 *Salmonella* genomes with associated MIC to 15 antibiotics, which were collected by NARMS over 15 years (2002 to 2016) in the United States. The study used the extreme gradient boost (XGBoost) algorithm (Chen & Guestrin 2016), which is a leading model in the numerous Kaggle competitions (Niehaus 2016) that are popular among machine learning practitioners (**https://www.kaggle.com/competitions**). The algorithm is a scalable implementation of gradient-boosted decision trees, which combines numerous decision trees as weak learners to achieve stronger performance (ensemble learning). In an earlier study, the XGBoost model also outperformed other major machine learning models to predict MIC in *Klebsiella pneumoniae* independent of precompiled AMR genes or polymorphisms (Nguyen et al. 2018).

Overall, XGBoost regressors delivered a mean accuracy of 0.95 for MIC prediction within ±1 twofold dilution step. Measured by major errors (MEs; false resistant results) and very major errors (VMEs; false susceptible results) that are commonly used in AST evaluation, the predictive models met the FDA standards for automated systems with all 15 antibiotics for ME rates and 7 of the 15 for VME rates. To prevent overfitting, the authors created and monitored separate validation sets (10% of the data) that were nonoverlapping with training and test sets.

The study further shed insights on two largely unaddressed questions regarding using bacterial genomes as training data. First, the size of the training set is often arbitrarily determined; how big it should be for effective learning remains unclear. Through subsampling of the training set while maximizing genetic diversity within the subset, the authors showed that MIC prediction models could be generated with fewer than 500 genetically diverse genomes to yield more than 0.90 accuracy. This finding has implications for building practical machine learning models, as training large sets of genomes is computationally demanding, requiring 1.5 TB of random-access memory for the full set of 4,500 genomes in the study. Second, the extent to which the diversity and spatiotemporal representativeness of the training set may affect the learning of biological traits is

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

uncertain. When tracking evolving traits in microbial pathogens to inform current epidemiology, training models with currently circulating strains is presumably advantageous. Given the rapid rise of AMR in recent years, the authors evaluated MIC prediction in later isolates by learning from earlier isolates. Stable predictions were obtained from different partitions of training sets of prior years and test sets of later years, suggesting the model's sustainable utility over time.

3.2. Genomic Source Attribution of Foodborne Pathogens

According to the Centers for Disease Control and Prevention, approximately 95% of foodborne illnesses in the United States are sporadic, non-outbreak cases whose food exposures and contamination sources are challenging to determine. With source information for most foodborne infections being largely unknown, it is difficult to understand foodborne illness epidemiology and develop intervention measures to prevent and mitigate such illnesses. Major foodborne pathogens, such as *Salmonella* and *E. coli*, are zoonotic enteric bacteria whose primary reservoirs include livestock and wild animals. Unlike AMR, for which many genetic determinants have been identified and characterized, mechanistic understanding of zoonotic host specificity and tropism is still limited. The lack of genetic markers prevents a rule-based approach to source prediction but creates an opportunity for machine learning investigations that do not necessarily require a priori knowledge of genetic determinants of bacterial host preference or adaptation.

3.2.1. Prediction of host specificity and potential. Lupolova et al. (2016) used an SVM classifier to predict host specificity of *E. coli* O157. Trained by 185 genomes from human and cattle infections using pan-genome content, up to 85% of human and 91% of cattle isolates were correctly classified into their respective isolation hosts. Interestingly, such a host dichotomy was not revealed by clustering analysis using multidimensional scaling also based on pan-genome content or by phylogenetic analysis using core genome single-nucleotide polymorphisms (SNPs). These results were interpreted as the particular ability of the machine learning approach to derive host-specific information from *E. coli* O157 genomes, which the authors proposed to imply distinctive zoonotic potentials of the pathogen. Host-specificity prediction by the SVM classifier was then expanded to *Salmonella* (Lupolova et al. 2017). Prediction consistent with the isolation host was made for 67–90% *Salmonella* Typhimurium isolates from human, avian, swine, and bovine sources. Although *Salmonella* Typhimurium is known for a broad zoonotic host range, only a small subset of the analyzed isolates had high probability scores for multiple hosts. A similar observation was made in *E. coli*, and these findings were inferred as marked host restriction, with only a minority showing a generalist capacity to colonize different hosts.

Host-range-restricted *Salmonella* strains have been associated with invasive, extraintestinal infection and loss of genes or gene functions (pseudogene formation) that are dispensable for inhabiting their hosts (Thomson et al. 2008). Using atypical mutations indicating a functional change of protein, Wheeler et al. (2018) trained a random forest classifier that learned to identify invasive and host-adapted populations of *S. enterica*, including both established serotypes and emergent lineages within particular serotypes. Furthermore, genes most informative for the classification were found to indicate degradation of metabolic pathways, a common theme underlying host adaptation.

3.2.2. Zoonotic source attribution. Zhang et al. (2019) applied a random forest classifier to zoonotic source attribution of *Salmonella* Typhimurium using genomic surveillance data in the United States. The classifier was trained with more than 1,200 genomes that had been collected from human cases and zoonotic sources by three major US laboratory surveillance programs

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

(PulseNet, NARMS, and GenomeTrakr) between 1949 and 2014. Estimated to be 83% accurate in predicting bovine, poultry, swine, and wild bird sources, the classifier correctly attributed isolates from seven out of eight major zoonotic outbreaks of *Salmonella* Typhimurium in the United States during 1994–2013 to their respective livestock sources. Notably, the study attempted to explain machine learning predictions through biological interpretation of the results. Phylogenetic analysis delineated livestock lineages showing steady rates of mutation, which allowed inference of their recent but pronounced association with livestock production systems. Genomic investigation revealed elevated accumulation of lineage-specific pseudogenes as signals of potential host adaptation in animal-associated lineages as they diverged from generalist populations. Metabolic profiling identified possible metabolic acclimation in certain animal isolates, which also indicates host adaptation. The study further showed that robust source prediction could be made by using a set of 50 key genetic features from the entire feature space of more than 3,000 SNPs, indels, and genes; multiple key features had been experimentally implicated to contribute to bacterial interaction with animal hosts. These findings suggest a rationale for machine learning application in food safety problem-solving when the underlying mechanism is complex or not well understood. Furthermore, some learned features may provide insights to study the mechanism.

The potential of genomic source attribution by machine learning has since led to several studies with a focus on methodology. Lupolova et al. (2019) performed a technical review and comparative study on different machine learning models for host prediction and statistical methods for feature selection. Guillier et al. (2020) streamlined zoonotic source attribution of *Salmonella* Typhimurium by using accessory genes alone as features and developed a software pipeline accordingly. Munck and colleagues assembled four European genome data sets for developing genomic source attribution tools (Munck et al. 2020a) and reported a logistic boost model using core genome multilocus sequencing typing as input to achieve accurate zoonotic attribution of *Salmonella* Typhimurium in Denmark (Munck et al. 2020b).

3.3. Challenges and Potential Pitfalls of Machine Learning Applications in Food Safety Using Genomic Data

AST and source attribution represent related but different types of phenotypic inference from genomic data. When machine learning inference is intended to identify genetic variations causally associated with specific phenotypes, it can be considered as a subset of microbial genome-wide association studies (mGWAS). Adapted from GWAS methods used in human genetics, mGWAS face challenges and pitfalls specific to bacterial species, including genome-wide linkage disequilibrium and strong population structuring, such as distinct lineages and clonal groups (Eyre et al. 2017, San et al. 2019). Such genetic and population traits can lead to identification of genotype–phenotype associations that are correlational but not causal. The still-nascent use of machine learning in food safety genomics has only begun to consider such challenges and pitfalls, either during feature selection (Lupolova et al. 2019) or at results confirmation (Drouin et al. 2016).

Phenotypes like AMR can be readily tested in the laboratory. Because AMR is often conferred by single or a few genes, functional confirmation of AMR biomarkers is relatively straightforward. In comparison, source association is a more complex phenomenon shaped by an interplay of bacterial, host, and environmental variables. Experimental assay for host specificity and precise confirmation of biomarker functions can be difficult (e.g., owing to lack of proper animal models). Adding to the complexity is the utility of surrogate markers in the prediction of complex phenotypes such as source association. For instance, phylogenetic markers of a lineage that is geographically restricted to a source can reliably indicate the source, although they play no functional roles in the source association. Although surrogate markers have little or limited use in AMR prediction for which functional confirmation of AMR determinants is typically anticipated, they

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

may be exploited for source attribution, especially when exact causation of host or environmental tropism is difficult to ascertain. In fact, surrogate subtyping markers have traditionally been used for microbial source tracking (Scott et al. 2002).

In machine learning practices, training set design greatly affects the outcome of the investigation. Differences in designing genome training sets have led to contradictory conclusions in source attribution of *Salmonella* Typhimurium (Wheeler 2019). Lupolova et al. (2017) reported more than 90% accuracy in assigning genomes to human hosts and a surprisingly high prevalence of host-restricted strains, including perceived human-adapted populations, based on machine learning prediction of host specificity. These findings are contrary to the established epidemiology that most human infections of *Salmonella* Typhimurium originate from animal reservoirs. Assuming the predominance of zoonotic transmission, Zhang et al. (2019) argued against treating humans as a source class and reservoir of *Salmonella* Typhimurium and challenged the notion of commonly circulating human-adapted strains. They proposed that the accuracy of human host prediction in the Lupolova study was biasedly inflated by an overrepresentation of closely related human isolates in the training set, of which 85% had another human isolate as their closest neighbor on the phylogeny (Zhang et al. 2019). Using a diverse set of human isolates to minimize training set bias caused by phylogenetic and epidemiological redundancy, Zhang et al. (2019) showed that US human isolates are indistinguishable from livestock isolates at the genomic level, contesting the existence of distinct human host signals in *Salmonella* Typhimurium genomes and suggesting mixed origins of human infections in various zoonotic sources.

Machine learning practices in clinical and public health settings will likely be challenged by the need for standardized methods. A major barrier to the deployment of WGS-based AST is the lack of methodology standardization, such as a resistance gene database and quality-control metrics (Ellington et al. 2017). Similarly, the design of training sets, the choice of machine learning algorithms, and the strategy for validation all pose difficulties for standardization. A recent compilation of European genome sets for developing WGS-based source attribution methods represents an initial effort toward standardization (Munck et al. 2020a).

Trade data: may come from supply-chain or trans-shipment logistics records, aggregated trade statistics, or modeled representations of supply structures

Text data: loose, unstructured language-based content that may come from social media, webpages, news, scientific literature, or company records

4. MACHINE LEARNING APPLICATIONS USING NOVEL DATA STREAMS

This section is focused on various sources of NDS (Althouse et al. 2015) that can be applied, together with machine learning or related data scientific and computational approaches, to support food safety research and practice. NDS enable passive, continuous, automatic data collection to provide enhanced timeliness, detail, breadth, and scalability of observations that primary data systems set up specifically for food safety may not (see the sidebar titled Novel Data Streams Versus Primary Data for Food Safety) (Althouse et al. 2015, Bansal et al. 2016, Oldroyd et al. 2018, Timmins et al. 2018). Importantly, these large digital sources of data hold significant structure (multiple dimensions or features) that must be teased out through extensive processing and analysis, lending themselves to analysis using machine learning techniques. Three major NDS data sources in combination with machine learning analysis are reviewed here: text, transactional, and trade data. Text data have seen the most application in association with machine learning techniques and is covered in greatest detail. Transactional and trade data, although promising, have only a few initial demonstrated applications and are covered more briefly.

4.1. Text Data

Text data represent loose or unstructured information in the form of natural-language text that can provide real-time information for monitoring and responding to food safety contamination events

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

NOVEL DATA STREAMS VERSUS PRIMARY DATA FOR FOOD SAFETY

The concept of novel data streams (NDS) was originally introduced by Althouse et al. (2015) and defined in the context of public health surveillance as those data streams that are directly initiated by the user and not already maintained by public health departments or other health professionals. Similar definitions have been introduced as secondary or found data (Connelly et al. 2016, Timmins et al. 2018). We adapt this definition to the context of food safety as data streams that are collected passively through user-initiated content and not designed with the specific intention of supporting food safety applications and thus not already maintained by public health or food safety authorities. To comply with machine learning techniques, these NDS must also be digital. NDS are contrasted with primary data generated specifically for food safety or tracking purposes, including biological/chemical monitoring data (World Health Organ. 1995), livestock data (Gates et al. 2015), inspection data (FDA 2020a), alert and recall data (FDA 2020b), outbreak surveillance data (Cent. Dis. Control 1997), proprietary company supply-chain track-and-trace data, and genomic data (see Section 3).

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

or threats (Greis & Nogueira 2017). For foodborne outbreak investigation purposes, sources of relevant text data include user-generated public posts on social media or review platforms, public web data such as articles in the news media or professional organization websites, and private company web-based data (Tao et al. 2020). Research and applications involving text data for food safety have focused on the use of text mining and natural language processing techniques to supplement traditional surveillance systems with reports of foodborne illness or food safety incidents.

4.1.1. Data types. Sources of NDS text data that have been applied together with machine learning techniques in food safety applications can be categorized into user-generated public post data and web-based data. Text data from user-generated posts include posts made on social media networking sites such as Twitter (Devinney et al. 2018, Harris et al. 2017, Harrison et al. 2014, Kuehn et al. 2014, Sadilek et al. 2017) and Facebook, crowdsourced consumer review sites such as Yelp (Effland et al. 2018, Nsoesie et al. 2014, Schomberg et al. 2016) and Amazon (Maharana et al. 2019), and participatory systems such as IWasPoisoned.com (Quade & Nsoesie 2017). Post data may also include proprietary content such as company message and feedback boards, user forums and blogs (Kate et al. 2014), and query data such as Google search history (Sadilek et al. 2018). Post data come with multiple features available for mining and analyzing in food safety applications. The text content of a post may include natural language writing along with a title or hashtags, which represent user-labeled keywords. Text data can be analyzed to determine the sentiment (positive, negative, or neutral) of posts or content. Non-text metadata that may come with a post and contribute to analysis include temporality, geotags marking the location a post is made or pinned to and which have been used to associate a user with a suspected/implicated origin of contamination (e.g., restaurant), or the cost and rating of a location.

Web data sources include articles in the news media and academic or professional organization websites. Web data offer more varied information for analysis and have been used to build food safety event surveillance systems (e.g., outbreaks, recalls) that monitor, aggregate, and rank relevant content and keywords across multiple websites and spatial locations (Chen et al. 2016, Kate et al. 2014). Comprehensive reviews of consumer-generated data for foodborne illness surveillance (Oldroyd et al. 2018) and general use of text data in food science (Tao et al. 2020) are available elsewhere.

4.1.2. Methods and approach. A multistep analysis approach is required to transform high-dimensional unstructured text content into actionable food safety information (**Figure 2**). The

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

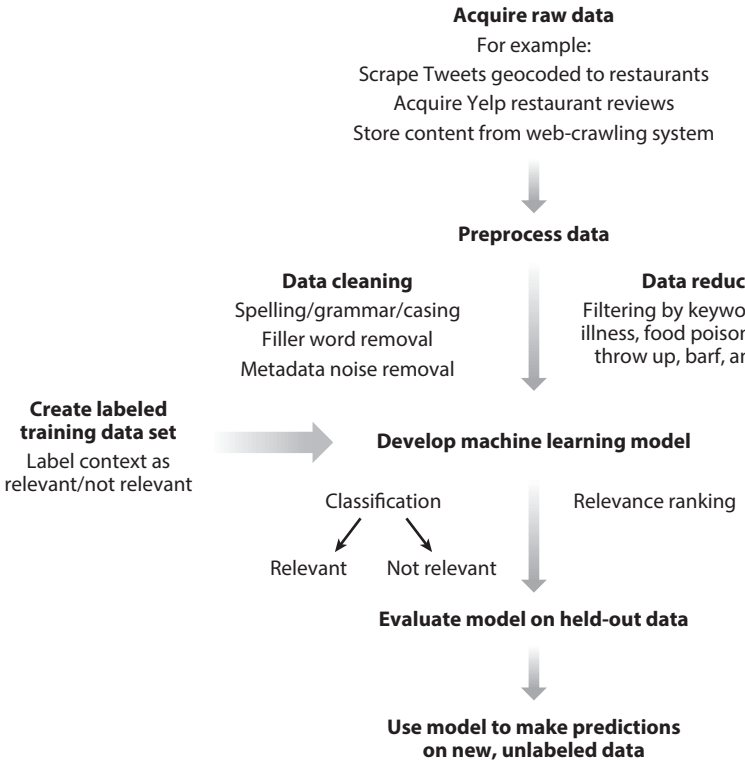

Figure 2

Schematic framework of text data machine learning analysis approach.

workflow may begin with data-scientific techniques to obtain data, followed by data-processing steps, including data cleaning or preprocessing (such as spelling/grammar corrections, filler word removal, and metadata noise removal), and data reduction (filtering), before machine learning algorithms to learn from and make predictions on data can finally be applied (Tao et al. 2020, Zhai & Massung 2016). Keyword identification is often a first step used to extract or filter data that may have food safety–related content. A set of specific keywords or phrases is often constructed a priori and may include words such as illness, food poisoning, vomit, throw up, barf, and nausea. A commonly applied machine learning task is to classify post or website data as relevant or not relevant to food safety events. Classification algorithms used in these contexts include decision trees, SVMs, naïve Bayes, and neural networks (Oldroyd et al. 2018).

To create a monitoring system that searches and sorts information across multiple websites, machine learning algorithms are employed to identify and rank relevant content where a simple keyword search would be insufficient (Kate et al. 2014). Ranking methods applied in this task have included cosine similarity (Drury & Roche 2019); a scoring function based on features including geographic region, time period, and food-related keywords (Chen et al. 2016); and text classification approaches, e.g., the ranking support vector machine (Rank-SVM) (Kate et al. 2014).

4.1.3. Applications and successes. Research and applications involving text data for food safety have focused on the use of text mining and natural language processing techniques to supplement traditional surveillance systems with reports of foodborne illness or food safety incidents. Social

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

media, reviews, news, and other web data can be monitored to capture near-real-time food safety violations or illness reports that are not reported through the official channels (i.e., filing reports with the local health department) and would otherwise go unrecognized. Key advantages of these sources over traditional data streams include the near-real-time availability of the data compared with the release of official reports of illness or outbreaks, which may be delayed by weeks, and the wider reach of data, which is especially helpful for capturing reports from younger users who are overrepresented on social media platforms but underrepresented in national foodborne illness outbreak statistics (Kuehn 2014, Oldroyd et al. 2018). The latter helps address the critical issue of underreporting of foodborne disease incidence (Oldroyd et al. 2018, Scallan et al. 2011).

Over the past decade, multiple studies have investigated the use of public social media data for foodborne illness surveillance. Some notable examples involve surveillance systems piloted in conjunction with local and federal health agencies, including the analysis of tweets in Chicago (Harris et al. 2014), St. Louis (Harris et al. 2017), Las Vegas (Sadilek et al. 2017), and New York City (Harrison et al. 2014) and Yelp reviews in New York City (Effland et al. 2018) and San Francisco (Schomberg et al. 2016). In an innovative application of machine learning and the joint analysis of text data with other types of NDS, a team involving Google researchers and the Chicago and Las Vegas health departments combined aggregated Google search queries with smartphone location data to identify restaurants violating health codes (Sadilek et al. 2018). This system, which first identifies search terms indicative of foodborne illness and then identifies restaurants visited by the users placing those search terms, demonstrated a threefold improvement over Twitter-based systems in identifying potential health code–violating venues. In addition to food safety incidents at restaurant outlets, unsafe food products have been identified by analyzing Amazon reviews with text classification methods (Maharana et al. 2019).

Public website data have also been used to develop systems for monitoring emerging food safety issues through web-crawling systems. Comprehensive information aggregation databases for food safety information have been developed, including the MedISys system of the European Commission (Rortais et al. 2010); a news media article content ranking and prioritization system by Singapore's National Environment Agency in collaboration with IBM Research (Kate et al. 2014); and a food safety event database system for greater China (Chen et al. 2016).

Many of these systems were designed to identify outbreaks otherwise missed and/or locations where an outbreak has occurred or is occurring. In practice, there have been a few successes in identifying small outbreaks of foodborne illness that traditional surveillance techniques have missed (Kuehn 2014), and there is early evidence of large-scale outbreak detection involving national restaurant chains (Quade & Nsoesie 2017). However, the main applications have been found in identifying restaurants that are likely to have food safety violations, whether or not an outbreak is actively occurring. Additionally, although not requiring sophisticated machine learning techniques, systems have been used to interact with posters to gather more information about the foodborne illness being reported, including the date and time of the foodborne illness event, restaurant details, and user contact information (Harris 2017, Kuehn 2014).

4.2. Transactional Data

Transactional data are a form of NDS that find application during foodborne disease outbreak investigations. Conventional epidemiological investigation processes involve patient interviews to identify commonalities across case patients followed by microbiological tests on suspected samples (Smith et al. 2015, World Health Organ. 2008). Transactional data provide an objective history of consumption records that have been demonstrated to support, complement, and even supplement conventional investigation techniques in generating hypotheses about the causative food vehicle

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

at early stages of an investigation, and/or to identify the location of contamination in retail or restaurants or elsewhere in the supply chain.

Most transactional data applications have involved analysis of individual consumer checkout data collected from known case-patients or businesses where contaminated food is thought to have been purchased, including individual barcode scanner receipts and loyalty card, credit card, and employee access card history. Since 2006, these data sets have helped to identify upward of 20 outbreak culprits (Moller et al. 2018). The methods applied have involved standard statistical techniques to compare purchase records across individuals to identify commonalities, e.g., odds ratios to compare histories from case-patients with control groups of shoppers. Because data records often must be collected on an individual basis, these applications do not lend themselves to automatization and machine learning techniques.

Machine learning techniques may find application with aggregated sales data, which exist in the form of aggregated store-based or spatially aggregated retail sales or loyalty card data. A few examples have illustrated the use of this data in outbreak surveillance and in outbreak investigations, helping identify the causative food vehicle. Kaufman et al. (2014) developed an approach to determine the contaminated food item that caused an outbreak by relating reported outbreak locations with the spatial patterns of sales of hundreds of individual products sold in retail supermarkets (Hu et al. 2016). Food items with sales patterns more closely resembling the outbreak distribution are deemed likely to be the causative food vehicle.

Methodologically, the approach involves a probability model of sales and maximum-likelihood estimation to identify a set of the most likely contaminated food products. Whereas the product identification method itself is a theory-driven probability model, binary classification learning methods are applied to characterize the accuracy of the approach and learn structure in its performance. Unsupervised clustering algorithms are applied to identify similar product distribution patterns to identify groups of food products that cannot be distinguished and may confound investigations. Using synthetic (simulated) foodborne outbreak patterns, Kaufman et al. (2014) evaluated the method on postal-code-aggregated weekly sales data of 580 food products in Germany. This approach has been expanded to account for time, consumer mobility, and noise and applied to a real-world outbreak in Norway (Norström et al. 2015).

4.3. Trade Data

Trade data, traditionally recorded or logged for company operations or statistical analysis, have recently been found to have innovative applications in food safety risk assessment. Here, we define trade data in the context of NDS as publicly available detailed records, aggregated records statistics, or modeled representations of flows characterizing the production, consumption, or movement of food through complex supply chains across countries, within a country, or between sectors of the supply system. Some examples of data sources include federal trade statistics and international trade, production, and consumption data. These data sources enable mapping of supply relationships, allowing for sophisticated, often network-theoretic food safety risk assessment analyses.

Supply-chain data collected and available publicly as part of the Chinese Ministry of Commerce's Important Product Traceability System have been studied to identify features of these network structures that are conducive to successful traceback to the source in the event of an outbreak (Lu et al. 2019). Chinese supply data together with import–export data accessed from a public data source, ImportGenius (**www.importgenius.com**), form the basis of a related application to predict food import firms likely to fail FDA site inspections (Levi et al. 2019). Logistic regression, gradient-based boosting on decision trees, and neural networks are all evaluated in

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

the prediction task, trained on FDA import inspection records. The models are trained to select nonlinear interactions of features of producers, suppliers, and supply-chain network structural relationships that are most predictive of risk, the combination of which improves identification of firms likely to fail FDA site inspections by >40% from current approaches.

Because exact, fine-grained data on subnational supply networks of food commodities are often not available, trade structures may also be modeled. Supply-chain networks depend on many factors, such as production locations, population centers, and storage and transport infrastructure (Balster & Friedrich 2019, Venkatramanan et al. 2017). Innovative uses of data on these features, in combination with sophisticated modeling techniques and machine learning algorithms, have been applied to develop models of food-flow network structures. A model of the food supply in Germany differentiating between the 402 German administrative districts has been developed using demand modeling methods from transport engineering (Balster & Friedrich 2019). In this approach, available production and consumption statistics provide the inputs at the origins and the outputs at the destinations of the flow network. The flows from origin to destination are then estimated using gravity models, calibrated with data on cross-regional, cross-sectorial trade flows coming from a national transport survey through an iterative, mass-balanced optimization algorithm. A similar approach to subnational food-flow modeling using machine learning has been developed in the United States to model food flows across the 3,142 counties and 7 aggregated commodity categories (Lin et al. 2019). A supervised machine learning approach involving logistic and gamma regression is incorporated to train the model to ensure that the properties of the estimated networks follow known structural properties of observed empirical food-flow networks.

When applied together with additional modeling techniques, these modeled supply structures have opened up applications for food safety risk analysis. A network-theoretic approach for locating the source of large-scale outbreaks of foodborne disease has been developed (Horn & Friedrich 2019) and applied in combination with the spatially aggregated German food supply network model (Balster & Friedrich 2019). The approach is evaluated on recent outbreaks of foodborne disease in Germany and has been expanded to approach the problem of identifying the food vehicle source of an outbreak through a statistical learning task involving hierarchical clustering (A. Horn, M. Fuhrmann, T. Schlaich, A. Balster, A. Kaesbohrer, M. Filter, H. Friedrich, unpublished results).

4.4. Challenges, Potential Pitfalls, and Future Development of NDS with Machine Learning in Food Safety

The passive generation and collection of NDS, although leading to enhanced timeliness, detail, breadth, and scalability of observations, also creates novel challenges, including biased samples, privacy concerns, and security issues. Sample bias and related issues that arise with NDS generation are compounded by the uncertain nature of the foodborne disease reporting context, which lends itself to unique challenges and pitfalls when combined with automated machine learning tasks, especially for applications such as crowd-sourced surveillance. Although researchers are aware of these issues, addressing them is an active area of research and the majority of the methods and applications reviewed here are still in research and development stages.

Still, NDS hold great promise for delivering innovations that improve health and practice when combined with machine learning techniques for model development and feature selection. These include real-time or near-real-time access to data; objective data-derived measures where commonly interview-based measures are used; increased breadth and/or resolution of spatial and temporal dimensions of data; scalability and increased populations sample size or catchments; and access to measures not possible with existing data streams.

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

4.4.1. Novel data streams and data access, bias, privacy, and security. Text data from user-generated NDS sources have critical biases to be aware of. Social media users, for example, represent a convenience sample of a younger, predominantly urban population with specific race/ethnicity biases and are not a representative sample of society; additionally, platform penetration is known to vary by geographic region (Altenburger & Ho 2019, Oldroyd et al. 2018, Tufekci 2014). Relying on consumer reports of food poisoning may perpetuate bias through both medical uncertainty and societal inequities. User attribution of foodborne illness to a specific food item or consumption location is notoriously difficult given the incubation periods of foodborne pathogens, multiplicity of consumed foods, and inaccuracies of recall (Gertler et al. 2017). Research has shown that consumer stereotypes around ethnic foods may affect the likelihood that consumers implicate such restaurants for food poisoning in social media data (Altenburger & Ho 2019, Zukin et al. 2015).

Proprietary platforms or websites may provide limited data access for researchers, or access only at a high cost. For instance, less than 1% of Twitter's data is available using its application programming interface. This may lead to biased samples if the sample size is insufficient or imbalanced and may be unfeasible for smaller, low-budget public health authorities to access.

There are generally limited privacy or security issues when analyzing data freely accessible on the web or public, such as voluntarily reported cases of potential foodborne illness that are posted to social media platforms. Still, privacy is a concern when it comes to data on individual users (Lazer et al. 2009, Sapienza & Palmirani 2018, Tene & Polonetsky 2013). To protect privacy, sensitive information, for example, usernames and addresses in sales, loyalty card, and location data, is often anonymized before sharing with researchers. A common challenge is that data shared under specific research agreements may not be shared with the general public or with a journal publishing an article, preventing validation or reproducibility studies.

4.4.2. Pitfalls and limitations of machine learning in novel data stream data analysis. A common challenge involving text data is the lack of specificity in foodborne disease–related keywords. Words such as nausea, sick, or diarrhea can apply to many diseases and ailments, leading to high false-positive rates. A related challenge is dealing with slang, sarcasm, and irony in posts. Promising work lies ahead in applying machine learning approaches developed for sarcasm recognition, rooted in sentiment analysis and pattern recognition, to these problems (Bouazizi & Ohtsuki 2016, Oldroyd et al. 2018, Pinheiro et al. 2015).

Many of the methods reviewed here require machine learning in combination with principled mathematical or predictive mechanistic modeling, which introduces a structure or logic into a problem approach. Predictive modeling techniques, such as agent-based models (Zoellner et al. 2019), networks (Garre et al. 2019, Horn & Friedrich 2019, Manitz et al. 2014), and other probabilistic or simulation-based models, can be used to investigate food safety questions with limited input data and/or without a training or learning component. And in many cases, technologies that do not require advanced analytics but do require efficient information technology systems for identifying, capturing, and assimilating data in real-time are more practical in outbreak surveillance and response settings, such as the supply data input, mapping, and visualization tools deployed during the 2011 enterohemorrhagic *E. coli* outbreak (Weiser et al. 2013, 2016).

We conclude that although NDS and the techniques reviewed here can supplement and/or complement existing data and analysis techniques to address food safety challenges, they are not a replacement for investigative work. None of the applications reviewed here are fully automatic, and most have yet to be rigorously evaluated in prospective applications (versus in historical training data) to ensure internal and external validation and prevent overfitting (Altenburger & Ho 2019, Althouse et al. 2015).

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

4.4.3. Prospects and future developments. Many promising applications for the use of machine learning with NDS have not yet been seen in food safety but could be modeled after applications in related fields. Loyalty cards (Aiello et al. 2019), restaurant sales, and online grocery (Huyghe et al. 2017) or delivery (Schulz et al. 2019) data sets, which have been used in consumer behavior and nutrition applications, could be applied to identify likely outbreak food vehicle sources, as Kaufman et al. (2014) pioneered using retail sales data. The prediction task could be improved by incorporating additional data, such as product-specific features (e.g., shelf life, probable consumption date, the likelihood that a particular product contains a particular pathogen), or factors influencing the purchase of particular items (e.g., weather, holidays, or sporting events), features commonly employed in retail consumption demand forecasting models. Aggregated credit card transactions, which have been used to develop machine learning models of consumer shopping trajectories (Krumme et al. 2013, Singh et al. 2015), could be applied to identify locations (e.g., markets, restaurants) where contaminated products may have been purchased. Location data from cell phone call data records or GPS logged by smartphone applications have found many applications in studying the spread of human-to-human infectious diseases (Oliver et al. 2020), but few examples exist so far involving foodborne disease (Sadilek et al. 2017, Teyhouee et al. 2017). Approaches operating on social media and search queries could be expanded from foodborne illness surveillance to other areas of food safety, including product recalls, allergens, or food safety regulations.

Finally, although this section has focused on applications of NDS and machine learning, there are many promising opportunities for combining NDS with primary food safety databases. Environmental data, including weather reports or satellite imagery, a source of NDS not yet mentioned, can be combined with agricultural data to predict food safety events or hazards in agriculture (Pang et al. 2020). An exciting emerging area involves integration across data sources, such as the algorithm developed by Google using location and web search text data and trained on restaurant inspection data (Sadilek et al. 2017), or between genomic and supply data (Dallman et al. 2016). In a long-term vision, it is possible to imagine the combination of many of these data sources, NDS and primary, together in large-scale, end-to-end predictive modeling systems.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objective of this review.

ACKNOWLEDGMENTS

X.D. is supported in part by US Department of Agriculture National Institute of Food and Agriculture Hatch project 1006141. S.C. is supported in part by National Science Foundation award DMS-1913080. A.H. is supported by National Institutes of Health Ruth L. Kirschstein National Research Service Award Institutional Training Grant T32 5T32CA009492-35.

LITERATURE CITED

Aiello LM, Schifanella R, Quercia D, Prete LD. 2019. Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Sci.* 8:14

Allard MW, Strain E, Melka D, Bunning K, Musser SM, et al. 2016. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* 54:1975–83

Altenburger KM, Ho DE. 2019. When algorithms import private bias into public enforcement: the promise and limitations of statistical debiasing solutions. *J. Inst. Theor. Econ.* 175:98–122

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Althouse BM, Scarpino SV, Meyers LA, Ayers JW, Bargsten M, et al. 2015. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Sci*. 4:17

Baker L, Fan H. 2017. Innovations of AlphaGo. *DeepMind Blog*, April 10. **https://deepmind.com/blog/article/innovations-alphago**

Balster A, Friedrich H. 2019. Dynamic freight flow modelling for risk evaluation in food supply. *Transp. Res. E* 121:4–22

Banko M. 2001. Scaling to very very large corpora for natural language disambiguation. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 26–33. Stroudsberg, PA: Assoc. Comput. Linguist.

Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. 2016. Big data for infectious disease surveillance and modeling. *J. Infect. Dis.* 214:S375–S79

Bouazizi M, Ohtsuki T. 2016. A pattern-based approach for sarcasm detection on Twitter. *IEEE Access* 4:5477–88

Cent. Dis. Control. 1997. Foodborne Diseases Active Surveillance Network, 1996. *Morb. Mortal. Wkly. Rep.* 46:258–61

Chen S, Huang D, Nong W, Kwan HS. 2016. Development of a food safety information database for Greater China. *Food Control* 65:54–62

Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–94. New York: Assoc. Comput. Mach.

Connelly R, Playford CJ, Gayle V, Dibben C. 2016. The role of administrative data in the big data revolution in social science research. *Soc. Sci. Res.* 59:1–12

Dallman T, Inns T, Jombart T, Ashton P, Loman N, et al. 2016. Phylogenetic structure of European *Salmonella* Enteritidis outbreak correlates with national and international egg distribution network. *Microb. Genom.* 2:e000070

Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, et al. 2016. Antimicrobial resistance prediction in PATRIC and RAST. *Sci. Rep.* 6:27930

Deng X, den Bakker HC, Hendriksen RS. 2016. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food. Sci. Technol.* 7:353–74

Devinney K, Bekbay A, Effland T, Gravano L, Howell D, et al. 2018. Evaluating Twitter for foodborne illness outbreak detection in New York City. *Online J. Public Health Inform.* 10(1):e120

Doyle MP, Erickson MC, Alali W, Cannon J, Deng X, et al. 2015. The food industry's current and future role in preventing microbial foodborne illness within the United States. *Clin. Infect. Dis.* 61:252–59

Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, et al. 2016. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genom.* 17:754

Drury B, Roche M. 2019. A survey of the applications of text mining for agriculture. *Comput. Electron. Agric.* 163:104864

Du C, Sun D. 2006. Learning techniques used in computer vision for food quality evaluation: a review. *J. Food Eng.* 72:39–55

Effland T, Lawson A, Balter S, Devinney K, Reddy V, et al. 2018. Discovering foodborne illness in online restaurant reviews. *J. Am. Med. Inform. Assoc.* 25:1586–92

Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, et al. 2017. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin. Microbiol. Infect.* 23:2–22

Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, et al. 2017. WGS to predict antibiotic MICs for *Neisseria gonorrhoeae*. *J. Antimicrob. Chemother.* 72:1937–47

Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, et al. 2019. Validating the AMRFinder Tool and Resistance Gene Database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* 63(11):e00483-19

Food Drug Adm. (FDA). 2020a. *Inspection classification database search*. Rep., US Dep. Health Hum. Serv., Washington, DC. **https://www.accessdata.fda.gov/scripts/inspsearch/**

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Food Drug Adm. (FDA). 2020b. *Recalls, market withdrawals, & safety alerts*. Rep., US Dep. Health Hum. Serv., Washington, DC. **https://www.fda.gov/safety/recalls-market-withdrawals-safety-alerts**

Friedman JH. 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.* 1:55–77

Garre A, Fernandez PS, Brereton P, Elliott C, Mojtahed V. 2019. The use of trade data to predict the source and spread of food safety outbreaks: an innovative mathematical modelling approach. *Food Res. Int.* 123:712–21

Gates MC, Holmstrom LK, Biggers KE, Beckham TR. 2015. Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: challenges and opportunities. *Front. Public Health* 3:74

Gertler M, Czogiel I, Stark K, Wilking H. 2017. Assessment of recall error in self-reported food consumption histories among adults—particularly delay of interviews decrease completeness of food histories—Germany, 2013. *PLOS ONE* 12:e0179121

Gibson MK, Forsberg KJ, Dantas G. 2015. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* 9:207–16

Greis NP, Nogueira ML. 2017. A data-driven approach to food safety surveillance and response. In *Food Protection and Security: Preventing and Mitigating Contamination during Food Processing and Production*, ed. S Kennedy, pp. 75–99. Food Sci. Technol. Nutr. Swaston, UK: Woodhead Publ.

Guillier L, Gourmelon M, Lozach S, Cadel-Six S, Vignaud ML, et al. 2020. AB_SA: accessory genes-based source attribution - tracing the source of *Salmonella enterica* Typhimurium environmental strains. *Microb. Genom.* 6:mgen000366

Gupta A, Nelson JM, Barrett TJ, Tauxe RV, Rossiter SP, et al. 2004. Antimicrobial resistance among *Campylobacter* strains, United States, 1997–2001. *Emerg. Infect. Dis.* 10:1102–9

Haarnoja T, Ha S, Zhou A, Tan J, Tucker G, Levine S. 2018. Learning to walk via deep reinforcement learning. arXiv:1812.11103 [cs.LG]

Halevy A, Norvig P, Pereira F. 2009. The unreasonable effectiveness of data. *IEEE Intell. Syst.* 24:8–12

Hall AJ, Wikswo ME, Manikonda K, Roberts VA, Yoder JS, Gould LH. 2013. Acute gastroenteritis surveillance through the National Outbreak Reporting System, United States. *Emerg. Infect. Dis.* 19:1305–9

Harris JK, Hawkins JB, Nguyen L, Nsoesie EO, Tuli G, et al. 2017. Using Twitter to identify and respond to food poisoning: the Food Safety STL Project. *J. Public Health Manag. Pract.* 23(6):577–80

Harris JK, Mansour R, Choucair B, Olson J, Nissen C, et al. 2014. Health department use of social media to identify foodborne illness—Chicago, Illinois, 2013–2014. *Morb. Mortal. Wkly. Rep.* 63:681–85

Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, et al. 2014. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013. *Morb. Mortal. Wkly. Rep.* 63:441–45

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin: Springer

He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* 25:30–36

Healthy People. 2010. *Healthy People 2020 Food Safety Objectives*. Washington, DC: HHS. **https://www.healthypeople.gov/2020/topics-objectives/topic/food-safety/objectives**

Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. 2019. Using genomics to track global antimicrobial resistance. *Front. Public Health* 7:242

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput.* 9:1735–80

Horn A, Friedrich H. 2019. Locating the source of large-scale outbreaks of foodborne disease. *J. R. Soc. Interface* 16:20180624

Hu K, Renly S, Edlund S, Davis M, Kaufman J. 2016. A modeling framework to accelerate food-borne outbreak investigations. *Food Control* 59:53–58

Humphries RM, Abbott AN, Hindler JA. 2019. Understanding and addressing CLSI breakpoint Revisions: a primer for clinical laboratories. *J. Clin. Microbiol.* 57:e00203-19

Huyghe E, Verstraeten J, Van Kerckhove A. 2017. Clicks as a healthy alternative to bricks: how online grocery shopping reduces vice purchases. *J. Mark. Res.* 54:61–74

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Kate K, Chaudhari S, Prapanca A, Kalaganam J. 2014. FoodSIS: a text mining system to improve the state of food safety in singapore. In *KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1709–18. New York: Assoc. Comput. Mach.

Kaufman J, Lessler J, Harry A, Edlund S, Hu K, et al. 2014. A likelihood-based approach to identifying contaminated food products using sales data: performance and challenges. *PLOS Comput. Biol.* 10:e1003692

Koh HK. 2010. A 2020 vision for healthy people. *N. Engl. J. Med.* 362:1653–56

Kohavi R, John GH. 1997. Wrappers for feature subset selection. *Artif. Intell.* 97:273–324

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, ed. F Pereira, CJC Burges, L Bouttou, KQ Weinberger. San Diego, CA: Neural Inf. Proc. Syst. Found.

Krumme C, Llorente A, Cebrian M, Pentland AS, Moro E. 2013. The predictability of consumer visitation patterns. *Sci. Rep.* 3:1645

Kuehn BM. 2014. Agencies use social media to track foodborne illness. *JAMA* 312:117–18

Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, et al. 2009. Computational social science. *Science* 323:721–23

LeCun Y, Boser B, Denker JS, Henderson D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1:541–51

Levi R, Renegar N, Springs S, Zaman T. 2019. Supply chain network analytics guiding food regulatory operational policy. SSRN. **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3374620**

Levow G. 2006. Unsupervised and semi-supervised learning of tone and pitch accent. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 224–31. New York: Assoc. Comput. Linguist.

Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE Jr., et al. 2016. Penicillin-binding protein transpeptidase signatures for tracking and predicting β-lactam resistance levels in *Streptococcus pneumoniae*. *mBio* 7:e00756-16

Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16:321–32

Lin X, Ruess P, Marston L, Konar M. 2019. Food flows between counties in the United States. *Environ. Res. Lett.* 14:084011

Lu X, Horn A, Su J, Jiang J. 2019. A universal measure for network traceability. *Omega* 87:191–204

Lupolova N, Dallman TJ, Holden NJ, Gally DL. 2017. Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microb. Genom.* 3:e000135

Lupolova N, Dallman TJ, Matthews L, Bono JL, Gally DL. 2016. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *PNAS* 113:11312–17

Lupolova N, Lycett SJ, Gally DL. 2019. A guide to machine learning for bacterial host attribution using genome sequence data. *Microb. Genom.* 5:e000317

Macesic N, Polubriaginof F, Tatonetti NP. 2017. Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. *Curr. Opin. Infect. Dis.* 30:511–17

Maguire F, Rehman MA, Carrillo C, Diarra MS, Beiko RG. 2019. Identification of primary antimicrobial resistance drivers in agricultural nontyphoidal *Salmonella enterica* serovars by using machine learning. *mSystems* 4:e00211-19

Maharana A, Cai K, Hellerstein J, Hswen Y, Munsell M, et al. 2019. Detecting reports of unsafe foods in consumer product reviews. *JAMIA Open* 2:330–38

Manitz J, Kneib T, Schlather M, Helbing D, Brockmann D. 2014. Origin detection during food-borne disease outbreaks - a case study of the 2011 EHEC/HUS outbreak in Germany. *PLOS Curr.* 6. **https://doi.org/10.1371/currents.outbreaks.f3fdeb08c5b9de7c09ed9cbcef5f01f2**

Marvin HJ, Janssen EM, Bouzembrak Y, Hendriksen PJ, Staats M. 2017. Big data in food safety: an overview. *Crit. Rev. Food Sci. Nutr.* 57:2286–95

McDermott PF, Tyson GH, Kabera C, Chen Y, Li C, et al. 2016. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. *Antimicrob. Agents Chemother.* 60:5515–20

Møller FT, Mølbak K, Ethelberg S. 2018. Analysis of consumer food purchase data used for outbreak investigations, a review. *Eurosurveillance* 23:1700503

Munck N, Leekitcharoenphon P, Litrup E, Kaas R, Meinen A, et al. 2020a. Four European *Salmonella* Typhimurium datasets collected to develop WGS-based source attribution methods. *Sci. Data* 7:75

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Munck N, Njage PMK, Leekitcharoenphon P, Litrup E, Hald T. 2020b. Application of whole-genome sequences and machine learning in source attribution of *Salmonella* Typhimurium. *Risk Anal*. 40:1693–705

Natl. Antimicrob. Resist. Monit. Syst. Enteric Bact. (NARMS). 2015. *NARMS 2015 Human Isolates Surveillance Report*. Silver Spring, MD: NARMS

Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, et al. 2018. Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci. Rep.* 8:421

Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, et al. 2019. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* 57:e01260–18

Niehaus D. 2016. *Tree Boosting with XGBoost—Why Does XGBoost Win "Every" Machine Learning Competition?* Trondheim, Nor.: Nor. Univ. Sci. Technol.

Niehaus KE, Walker TM, Crook DW, Peto TEA, Clifton DA. 2014. Machine learning for the prediction of antibacterial susceptibility in *Mycobacterium tuberculosis*. In *IEEE-EMBS International Conference on Biomedical Health and Informatics*, pp. 618–21. Piscataway, NJ: IEEE

Nisbet R, Miner G, Elder J. 2009. *Handbook of Statistical Analysis and Data Mining Applications*. Cambridge, MA: Academic

Norström M, Kristoffersen AB, Görlach FS, Nygård K, Hopp P. 2015. An adjusted likelihood ratio approach analysing distribution of food products to assist the investigation of foodborne outbreaks. *PLOS ONE* 10:e0134344

Nsoesie EO, Kluberg SA, Brownstein JS. 2014. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. *Prev. Med.* 67:264–69

Oldroyd RA, Morris MA, Birkin M. 2018. Identifying methods for monitoring foodborne illness: review of existing public health surveillance techniques. *JMIR Public Health Surveill*. 4:e57

Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, et al. 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Sci. Adv.* 6:eabc0764

Pang H, Mokhtari A, Chen Y, Oryang D, Ingram DT, et al. 2020. A predictive model for survival of *Escherichia* coli O157:H7 and generic *E. coli* in soil amended with untreated animal manure. *Risk Anal*. 40:1367–82

Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, et al. 2016. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data. *Front. Microbiol.* 7:1887

Phillips L. 2006. Food and globalization *Annu. Rev. Anthropol.* 35:37–57

Pinheiro V, Pontes R, Furtado V. 2015. A #hashtagtokenizer for Social Media Messages. *Int. J. Comput. Linguistics Appl*. 6:141–58

Quade P, Nsoesie EO. 2017. A platform for crowdsourced foodborne illness surveillance: description of users and reports. *JMIR Public Health Surveill.* 3(3):e42

Rortais A, Belyaeva J, Gemo M, van der Groot E, Linge JP. 2010. MedISys: an early-warning system for the detection of (re-)emerging food- and feed-borne hazards. *Food Res. Int.* 43(5):1553–56

Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1:206–15

Sadilek A, Caty S, DiPrete L, Mansour R, Schenk T, et al. 2018. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *npj Digit. Med.* (1):36

Sadilek A, Kautz H, DiPrete L, Labus B, Portman E, et al. 2017. Deploying nEmesis: preventing foodborne illness by data mining social media. In *Proceedings of the Twenty-Eighth AAAI Conference on Innovative Applications (IAAI-16)*, pp. 3982–89. Menlo Park, CA: Assoc. Adv. Artif. Intell.

Samuel AL. 1959. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* 3:210–29

San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, et al. 2019. Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls. *Front. Microbiol.* 10:3119

Sapienza S, Palmirani M. 2018. Emerging data governance issues in big data applications for food safety. In *Electronic Government and the Information Systems Perspective. EGOVIS 2018*, ed. A Kő, E Francesconi, pp. 221–30. Lect. Notes Comput. Sci. 11032. Cham, Switz.: Springer

Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, et al. 2011. Foodborne illness acquired in the United States—major pathogens. *Emerg. Infect. Dis.* 17:7–15

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Scallan E, Mahon BE. 2012. Foodborne Diseases Active Surveillance Network (FoodNet) in 2012: a foundation for food safety in the United States. *Clin. Infect. Dis.* 54(Suppl. 5):S381–84

Schomberg JP, Haimson OL, Hayes GR, Anton-Culver H. 2016. Supplementing public health inspection via social media. *PLOS ONE* 11(3):e0152117

Schulz E, Bhui R, Love BC, Brier B, Todd MT, Gershman SJ. 2019. Structured, uncertainty-driven exploration in real-world consumer choice. *PNAS* 116:13903–8

Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J. 2002. Microbial source tracking: current methodology and future directions. *Appl. Environ. Microbiol.* 68:5796–803

Shardanand U, Maes P. 1995. Social information filtering: algorithms for automating "word of mouth." In *CHI '95: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ed. IR Katz, pp. 210–17. New York: ACM Press

Sharpless N, Yiannas F. 2019. *Statement from acting FDA Commissioner Ned Sharpless, M.D., and Deputy Commissioner Frank Yiannas on steps to usher the U.S. into a new era of smarter food safety*. Press Rel., April 30. **https://www.fda.gov/news-events/press-announcements/statement-acting-fda-commissioner-ned-sharpless-md-and-deputy-commissioner-frank-yiannas-steps-usher**

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550:354–59

Singh VK, Bozkaya B, Pentland A. 2015. Money walks: implicit mobility behavior and financial well-being. *PLOS ONE* 10:e0136628

Smith K, Miller B, VIerk K, Williams I, Hedberg C. 2015. *Product tracing in epidemiologic investigations of outbreaks due to commercially distributed food items – utility, application, and considerations*. Rep., Minn. Integr. Food Saf. Cent. Excel., Minneapolis, MN. **http://mnfoodsafetycoe.umn.edu/wp-content/uploads/2015/10/Product-Tracing-in-Epidemiologic-Investigations.pdf**

Stigler SM. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Belknap

Strawn LK, Fortes ED, Bihn EA, Nightingale KK, Grohn YT, et al. 2013. Landscape and meteorological factors affecting prevalence of three food-borne pathogens in fruit and vegetable farms. *Appl. Environ. Microbiol.* 79:588–600

Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, CDC PulseNet Task Force. 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis.* 7:382–89

Tack DM, Ray L, Griffin PM, Cieslak PR, Dunn J, et al. 2020. Preliminary incidence and trends of infections with pathogens transmitted commonly through food—Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2016–2019. *Morb. Mortal. Wkly. Rep.* 69:509–14

Tao D, Yang P, Feng H. 2020. Utilization of text mining as a big data analysis tool for food science and nutrition. *Compr. Rev. Food Sci. Food Saf.* 19:875–94

Tene O, Polonetsky J. 2013. Big data for all: privacy and user control in the age of analytics. *Northwest. J. Technol. Intellect. Prop.* 11:239–73

Teyhouee A, McPhee-Knowles S, Waldner C, Osgood N. 2017. Prospective detection of foodborne illness outbreaks using machine learning approaches. In *Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2017*, ed. D Lee, YR Lin, N Osgood, R Thomson, pp. 302–8. Lect. Notes Comput. Sci. 10354. Cham, Switz.: Springer

Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, et al. 2008. Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res.* 18:1624–37

Timmins KA, Green MA, Radley D, Morris MA, Pearce J. 2018. How has big data contributed to obesity research? A review of the literature. *Int. J. Obes.* 42:1951–62

Tufekci Z. 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA: AAAI Press

Turing AM. 1950. Computing machinery and intelligence. *Mind* LIX:433–60

Turnidge J, Paterson DL. 2007. Setting and revising antibacterial susceptibility breakpoints. *Clin. Microbiol. Rev.* 20:391–408

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)

Annu. Rev. Food Sci. Technol. 2021.12. Downloaded from www.annualreviews.org
Access provided by Columbia University on 02/22/21. For personal use only.

Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, et al. 2019. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18:463–77

Venkatramanan S, Wu S, Marathe A, Marathe M, Eubank S, et al. 2017. Towards robust models of food flows and their role in invasive species spread. In *2017 IEEE International Conference on Big Data (Big Data)*. Piscataway, NJ: IEEE

Weiser AA, Gross S, Schielke A, Wigger JF, Ernert A, et al. 2013. Trace-back and trace-forward tools developed ad hoc and used during the STEC O104:H4 outbreak 2011 in Germany and generic concepts for future outbreak situations. *Foodborne Pathog. Dis.* 10:263–69

Weiser AA, Thöns C, Filter M, Falenski A, Appel B, Käsbohrer A. 2016. FoodChain-Lab: a trace-back and trace-forward tool developed and applied during food-borne disease outbreak investigations in Germany and Europe. *PLOS ONE* 11:e0151977

Wheeler NE. 2019. Tracing outbreaks with machine learning. *Nat. Rev. Microbiol.* 17:269

Wheeler NE, Gardner PP, Barquist L. 2018. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLOS Genet*. 14:e1007333

World Health Organ. 1995. *Global Environment Monitoring System: Food Contamination Monitoring and Assessment Programme (GEMS/Food): Compilation of Analytical Quality Assurance Study Reports*, *1994*. Geneva: World Health Organ.

World Health Organ. 2008. *Foodborne Disease Outbreaks: Guidelines for Investigation and Control*. Geneva: World Health Organ.

Yiannas F. 2015. How Walmart's SPARK keeps your food fresh. *Walmart Newsroom*, Jan. 12. **https://corporate.walmart.com/newsroom/sustainability/20150112/how-walmarts-spark-keeps-your-food-fresh**

Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agerso Y, et al. 2013. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J. Antimicrob. Chemother.* 68:771–77

Zhai CX, Massung S. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York/Williston, VT: Assoc. Comput. Mach./Morgan & Claypool

Zhang S, Li S, Gu W, den Bakker H, Boxrud D, et al. 2019. Zoonotic source attribution of *Salmonella enterica* serotype Typhimurium using genomic surveillance data, United States. *Emerg. Infect. Dis.* 25:82–91

Zhao S, McDermott PF, Friedman S, Abbott J, Ayers S, et al. 2006. Antimicrobial resistance and genetic relatedness among *Salmonella* from retail foods of animal origin: NARMS retail meat surveillance. *Foodborne Pathog. Dis.* 3:106–17

Zhou Z, Alikhan NF, Mohamed K, Fan Y, Agama Study G, Achtman M. 2020. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res.* 30:138–52

Zhu X, Ghahramani Z, Lafferty J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML'03: Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ed. T Fawcett, N Mishra, pp. 912–19. Menlo Park, CA: AAAI Press

Zoellner C, Jennings R, Wiedmann M, Ivanek R. 2019. EnABLe: an agent-based model to understand *Listeria* dynamics in food processing facilities. *Sci. Rep.* 9:495

Zukin S, Lindeman S, Hurson L. 2015. The omnivore's neighborhood? Online restaurant reviews, race, and gentrification. *J. Consum. Cult.* 17:459–79

Review in Advance first posted on January 20, 2021. (Changes may still occur before final publication.)