Received: 4 October 2017 | Revised and accepted: 12 July 2018

DOI: 10.1002/widm.1280

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

OVERVIEW

There and back again: Outlier detection between statistical reasoning and data mining algorithms

Arthur Zimek[1] | Peter Filzmoser[2]

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

[2]Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Correspondence
Arthur Zimek, Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark.
Email: zimek@imada.sdu.dk

Outlier detection has been a topic in statistics for centuries. Over mainly the last two decades, there has been also an increasing interest in the database and data mining community to develop scalable methods for outlier detection. Initially based on statistical reasoning, however, these methods soon lost the direct probabilistic interpretability of the derived outlier scores. Here, we detail from a joint point of view of data mining and statistics the roots and the path of development of statistical outlier detection and of database-related data mining methods for outlier detection. We discuss their inherent meaning, review approaches to again find a statistically meaningful interpretation of outlier scores, and sketch related current research topics.

This article is categorized under:
Algorithmic Development > Statistics
Algorithmic Development > Scalable Statistical Methods
Technologies > Machine Learning

KEYWORDS
anomaly detection, outlier detection, outlier model, statistics and data mining

1 | INTRODUCTION

An outlier could be generally defined as being *"an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"* (Barnett & Lewis, 1994). Finding outliers (i.e., data objects that do not fit well to the general data distribution) is very important in many practical applications, including for example, credit card abuse detection in financial transactions data, the identification of measurement errors in scientific data, or the analysis of sports statistics data. In the light of the common metaphor grasping the task of data mining like mining for nuggets of information, outlier detection can be seen as being not merely interested in removing noise but also in finding interesting database objects deviating in their behavior considerably from the majority and, as such, providing new insights. Indeed, both aspects of outlier detection are like two sides of a coin as one person's noise may be another person's signal. The application scenarios given above highlight both interests in outliers, as measurement errors in scientific data should possibly just be removed (Li & Wong, 2001) whereas a case of credit card abuse is the solely interesting fact among a wealth of "just usual" data (that, in turn, could of course be interesting itself as well, e.g., for modeling a customer's interests and behavior—after removing outliers).

In the study, we will first give an overview on existing surveys and how this survey is taking a different point (Section 2). Then we will reason about the idea and meaning of outlierness and about consequences of identifying some entity as outlier (Section 3). We will give a short overview of different kind of categorizations of outlier detection methods (Section 4), before we come to an overview of statistical methods (Section 5) and of database-oriented data mining methods (Section 6). We round the survey up with a short discussion of evaluation methods (Section 7) and a vision of possible directions of

improvements based on the renewed statistical understanding and probabilistic interpretation of outlier scores (Section 8). Finally, we give an overview on some software packages (Section 9) and conclude the survey (Section 10).

2 | DIFFERENT POINTS OF VIEW TO SURVEY THE LITERATURE

In the past 10–15 years, quite a few surveys on outlier detection have been published, based in different research areas. Alas, the field is that large any survey can only follow some selected path. One could get the impression that even a survey on surveys could be of avail in order to identify the most helpful information required for a specific type of question. Here, we only touch on some recent sources for an overview on the topic, before we point out our own way of seeing things in this area.

2.1 | Some existing surveys

Most surveys so far focused on specific techniques, a specific research area, or a specific application domain.

Markou and Singh provided a pair of surveys on outlier detection, specialized to statistical approaches (Markou & Singh, 2003a) and neural network-based approaches (Markou & Singh, 2003b). They focus on *novelty* detection. A *novelty* can be considered as a specific type of outliers that occur after the training phase has been completed. The rationale is that a novelty does not fit well to the previously learned distributions. Although the survey does not clearly distinguish between supervised and unsupervised methods, the numerous discussed methods typically require a training phase of some sort. Overall, the techniques are seen from the pattern recognition point of view.

Hodge and Austin (2004) name different research areas that contributed outlier detection methods under different names as, for example, *"novelty detection, anomaly detection, noise detection, deviation detection or exception mining"*. (Time-honored literature, Anscombe & Guttman, 1960, reports yet other terms for outliers, for example, "wild", "straggler", "sport", "maverick", "aberrant", or "spurious" observation. Most of these terms for outliers, however, are outlying terms in today's literature on outliers.) Hodge and Austin (2004) give also, as an introduction, a coarse but extensive list of application problems. Methodologically, they discern between unsupervised, supervised, and semi-supervised methods. For unsupervised techniques (where our main interest is in this survey) they focus on clustering techniques and, hence, practically dismiss a huge bulk of work provided in the data mining community. Instead, they cover approaches coming from the artificial intelligence and pattern recognition community.

Agyemang, Barker, and Alhajj (2006) are more interested in the data mining aspect of outlier detection as finding the rare, interesting pattern in a huge amount of data, based on a listing of interesting application examples. Their survey elaborates the distinction between numeric and symbolic approaches and discerns different categories of algorithms in these approaches. They also touch on cluster-based techniques. In each of these categories, they give an extensive review of a multitude of existing approaches and sketch shortly the merits and shortcomings of the different categories.

Patcha and Park (2007) are especially interested in network intrusion detection. Nevertheless they survey a couple of general outlier detection methods as techniques that found use in network intrusion detection systems. Their systematic overview is based on the distinction of communities contributing the basic techniques, namely statistics, machine learning, and data mining. The interest of their survey, however, is more in the application of the basic techniques in specialized systems than in the characteristics of the basic techniques themselves.

Hadi, Rahmatullah Imon, and Werner (2009) discuss the matter from a recent statistical point of view and describe a broad variety of statistical approaches to outlier detection. They also discuss outlier detection as a statistical problem in a more general setting, yet only touch on database-related data mining and machine learning approaches. They had, of course, many precursors in providing surveys and discussions of research on handling of outliers in statistics. As still interesting and inspiring articles we recommend the works by Anscombe and Guttman (1960), Barnett (1978), and Beckman and Cook (1983). Among these, especially Beckman and Cook (1983) provide a thorough history of the evolution of the statistical theories of handling outliers. Gnanadesikan and Kettenring (1972) include a discussion of (multivariate) outliers in their more general overview.

Su and Tsai (2011) give a coarse overview on some categories of unsupervised outlier detection and roughly name some basic approaches to supervised and semi-supervised outlier detection.

Rousseeuw and Hubert (2011) focus on robust statistics for outlier detection. In this family of approaches, the basic principle of outlier detection is to fit a model to the data majority, where the model-fitting should be robust against outliers, and then to look for observations which significantly deviate from this model. In that way, these identified outliers are always related to an underlying model, such as a univariate location/scale model, a multivariate normal model, a regression model, a classification model, etc.

The most extensive recent survey on outlier detection in data mining has been provided by Chandola, Banerjee, and Kumar (2009). They discuss several interesting points. Besides discerning different techniques (as did most of the previous

surveys, although Chandola et al. additionally discuss information-theoretic and spectral approaches as categories), they also describe different application domains more thoroughly as previous surveys and list outlier detection approaches that found use in the different application areas. Furthermore, they detail different types of outliers, namely *point anomalies*, *contextual anomalies*, and *collective anomalies.* Often, a certain type of anomaly is typical for a certain application domain.

Similar to the survey of Chandola et al. in spirit but with a different background is the survey by Pimentel, Clifton, Clifton, and Tarassenko (2014). They focus on novelty detection in the sense of semi-supervised learning (where examples for the normal class are given, but not for the anomaly class; examples of that class thus appear as "novelty") yet without a strict distinction against other flavors of outliers and corresponding methods.

Specialized surveys discuss methods for specific scenarios such as discrete sequences (Chandola, Banerjee, & Kumar, 2012), graph-data (Akoglu, Tong, & Koutra, 2015), high-dimensional data (Zimek, Schubert, & Kriegel, 2012), or ensemble methods for outlier detection (Zimek, Campello, & Sander, 2013).

Among recent textbooks on data mining, especially Tan, Steinbach, and Kumar (2006) give a decent overview. Han, Kamber, and Pei (2011) survey outlier analysis as part of cluster analysis yet discuss some genuine database-related outlier detection methods. In a statistical context, the books by Rousseeuw and Leroy (2003), Hawkins (1980), and Barnett and Lewis (1994) remain classics.

2.2 | Focus and organization of this survey

Although we are indebted to all these survey articles and textbook overviews as inspiring guides to the literature, here we intend to make a different point that has not been made by previous authors.

We are interested in the statistical meaning of the different methods and how they relate to each other and to the original statistical intuition. While recent comprehensive surveys like the works by Chandola et al. (2009) and Pimentel et al. (2014) structure the literature according to specific techniques used (e.g., probabilistic, distance-based, clustering-based, information theoretic), here we rather see this as a continuum where the connection of different techniques to a probabilistic interpretation is tighter for some and looser for others. However, from an intuitive perspective (i.e., not focusing on the algorithmic techniques but rather on the goal that is to be achieved by using various algorithmic techniques), the various approaches are not falling apart in strict categories. At the end of the day the central question for any application of such outlier detection methods is how to statistically interpret the outlier score that has been provided by some method. This interpretation and its relationship to outlier scores of different methods are usually anything but obvious. Thus we aim at re-establishing the link between data mining and statistics.

3 | WHAT AN "OUTLIER" POSSIBLY MEANS

Reconsider the definition of an outlier we started with as *"an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"* (Barnett & Lewis, 1994). In about two decades of research in data mining many methods have been proposed to identify such outliers. Much attention has been spent on doing this ever faster, less attention has been attributed to the description *"appears to"*.

In this section, we would like to reason on the importance of this notion of outliers as *apparently* inconsistent data objects. If some method identifies some data object as an outlier, this merely means, the data object is *suspected* to be inconsistent.

3.1 | Correction of the data or correction of the model?

Inconsistency can mean that the data object is a contaminant from a different distribution than the model considered to describe the data. This is also the intuition of the classical definition of outliers by Hawkins (1980): *"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."* But inconsistency could also mean that the presupposed model is not describing the data as well as was assumed when selecting the model. Both conclusions can bear rather significant repercussions on the interpretation of the given observations.

Let us first consider the meaning of the first possibility implied by Hawkins' definition. A quite plausible illustration of the meaning of this definition can be found in Barnett (1978) short discussion of the legal case of Hadlum versus Hadlum, held in 1949. Mr. Hadlum suspected Mrs. Hadlum of having committed adultery based on the evidence of the birth of a child 349 days after Mr. Hadlum had left abroad for military service. Compared to an average human gestation period of 280 days, 349 days is an outlier *arousing suspicions that it was generated by a different mechanism.* However, the judges decided that a gestation period of 349 days was, while very improbable, still scientifically possible.

While we do not know if Mr. and Mrs. Hadlum lived happily ever after, what this example illustrates quite drastically is the ambiguity of any decision concerning the outlierness of a data point. The probability of the data point being a member of an unsuspicious distribution depends on certain assumptions on the generating mechanism, and it heavily depends on the domain whether a very low probability is just a very low probability or strongly suggests a different generating mechanism since a deviation by a certain degree is impossible. There may serious consequences be involved in such decisions. Accepting Mr. Hadlum's conjecture would result in a high probability value for the birth of the child with an average gestation period. The assignment of probabilities is only a first and presumptive step toward deciding cases though. It should be noted that the high probability value according to Mr. Hadlum's conjecture involves a fundamentally different interpretation of the data, as illustrated in Figure 1. The judges' assumption was that a gestation period with a very improbable duration may still be possible and the low probability of the event alone (without further proof of adultery) does not justify a considerably different interpretation of the data including severe legal consequences.

Nevertheless, assigning a probability value under certain assumptions is probably still the most viable way of helping decide on the outlierness of a data point. The trouble is in sufficiently being aware of the "certain assumptions" underlying any decision. One aspect relevant to this awareness is the specific domain of the data. As introduced above, to decide that a data point is an outlier still leaves the domain expert with two general possibilities: (a) remove the data point as a contaminant from a different generating mechanism or (b) assume it is a genuine member of the data distribution anyway and conclude that the scientist's assumptions concerning the distribution are flawed. This second possibility relates to scientific progress according to Popper's critical rationalism (Popper, 1934, 1959) and, hence, is actually also a very interesting possibility. Already Kruskal (1960) stated such thoughts:

> *"An apparently wild (or otherwise anomalous) observation is a signal that says: 'Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study.' Examples of such serendipity have been frequently discussed — one of the most popular is Fleming's recognition of the virtue of penicillium."*

Accordingly, it has been stated as a general recommendation by Beckman and Cook (1983): *"outliers should be treated generally as an indication that either the model or the cases may be in error, and they often provide useful diagnostic information."* An early reference for rejection of outliers can be found in the work of Bernoulli (1777). Bernoulli's comments suggest that it was common practice to discard discordant observations in the 18th century. A practice, however, of which he does not readily approve. Some illustrative remarks are, in Allens translation (Bernoulli & Allen, 1961) as follows:

> *"[...] astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest and, indeed, assigning to them the same reliability. This practice makes it more than clear that they are far from assigning to them the same validity to each of the observations they have made, for they reject some in their entirety, while in the case of others they not only retain them all but, moreover, treat them alike. I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be*

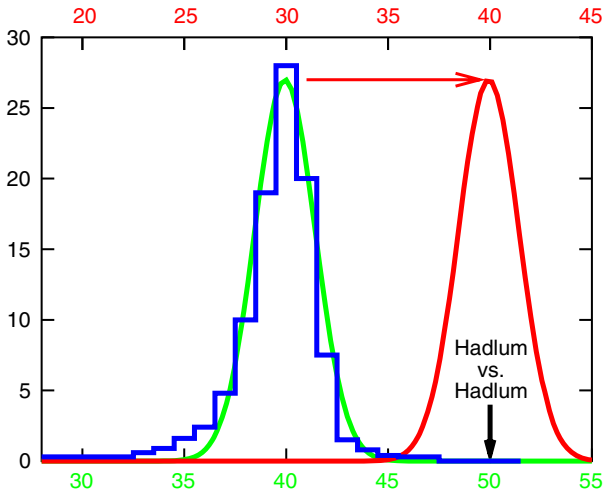

FIGURE 1 The histogram (blue) of human gestation periods (based on 13,634 cases, as reported by Barnett, 1978) with a fitted normal distribution (green), describing the hypothesis of the judges, and the alternative distribution (red) according to Mr. Hadlum's conjecture, assuming not an unusual value of the normal distribution but just a different distribution shifted by around 10 weeks, that is, a later start of the gestation period. This assumption of a totally different "generating mechanism" accommodates the alleged outlier perfectly

wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others."

Bernoulli's subsequent reasoning is a treasure chest of arguments for supporting the discussion preeminently laid out by Popper: whether or not discording observations should not rather help to correct the theory at stake instead of being rejected. In the same spirit Bessel, a German astronomer, reasoned in a geodetic work published 1838 (Bessel, 1838) that it was difficult (and seemed impractical in his own work) to define a criterion when to reject some observation. Instead, all completed observations should contribute equally to the result in order to avoid any arbitrariness of the results.

A related consideration was presented by Faloutsos (2010) to prove job security for data miners. Faloutsos showed (based on previous work Faloutsos & Megalooikonomou, 2007; Seshadri et al., 2008), that it will never be known if a derived model is ultimately valid. In an outlier detection scenario, we illustrate this phenomenon in Figure 2, where new points become known over time. Considering the first bulk of points (Figure 2a), a linear model appears to be well fitting the data, except, perhaps, one suspicious point. After new data has become known, the deficiency of the model is revealed, a more complex (polynomial) model is required (Figure 2b). The previously suspicious point (i.e., an alleged outlier) has become the first hint to the requirement of a better model.

3.2 | Consequences for the application of outlier detection methods

If we let aside these problems for the time being, we will assume to have an accurate (explicit or implicit) model for usual data (inliers). Then we want to identify data objects not generated by the corresponding mechanism. There are, however, still some issues to think about before we can sensibly talk about outlier detection.

First, there is some ambiguity in the use of terms. Collett and Lewis (1976) discern between

- data objects appearing *"in the eyes of the analyst"* (i.e., according to subjective feeling) *surprising* or *suspicious*, and
- *discordant* data objects, that is, an observation that is *"on some objective statistical criterion inconsistent with the rest of the sample"*.

A slightly different flavor is in the distinction provided by Beckman and Cook (1983) between

- a *discordant observation*, that is, *"any observation that appears surprising or discrepant to the investigator"*;
- a *contaminant*, that is, *"any observation that is not a realization from the target distribution"*;
- an *outlier*: *"a collective to refer to either a contaminant or a discordant observation"*.

In other words, we have different types of data objects occasionally termed "outliers" and it will be helpful to keep this distinction in mind.

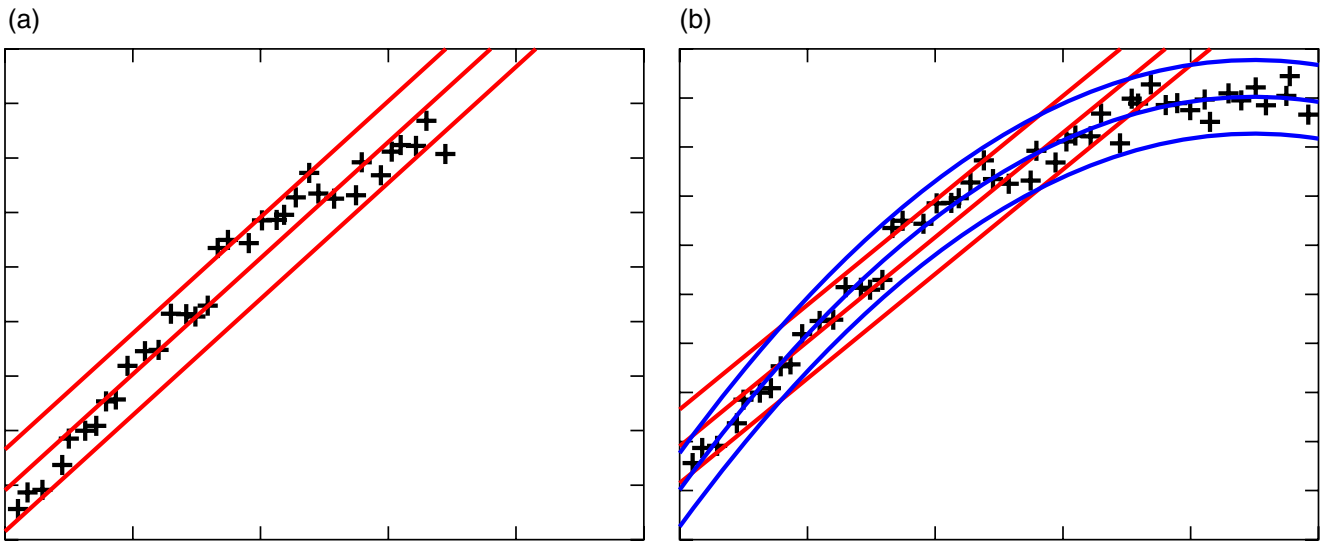

FIGURE 2 Is some data point an outlier or is the model wrong? (a) Linear model and some outlier and (b) more points and adapted (more complex) model

Let us, hence, designate, corresponding to Hawkins (1980), those objects as *"(true) outliers"* that have been *"generated by a different mechanism"* than the remainder or major part of the data or than the whatsoever defined reference set. Then there are different but possibly partly overlapping sets of objects:

- objects that *appear to be outliers* (independent of whether or not they actually are (true) outliers, that is, "discordant observations" in the sense of Beckman & Cook, 1983);
- objects that *are actually (true) outliers* (independent of whether or not they appear—according to some subjective feeling, or according to some specific data mining algorithm, or according to some objective statistical criterion—to be outliers, that is, "contaminants" in the sense of Beckman & Cook, 1983).

As a typical application area, this principal problem of outlier detection can also be illustrated by the example of network intrusion detection discussed by Patcha and Park (2007). Usually, intrusion detection systems are trained to detect anomalous behavior. However, not every anomalous behavior is a malicious intrusive activity, and not every truly intrusive activity comes along with anomalous behavior.

In more theoretical terms, we can figure this distinction as in Table 1 as the intersections of two partitions: true inliers and true outliers as one partition, apparent inliers and apparent outliers as another partition. The recognition of true outliers as apparent outliers relates to the set of true positives, unrecognized true outliers are false negatives, true inliers that have not been recognized as outliers (apparent inliers) relate to the set of true negatives, true inliers that have been tagged as apparent outliers are false positives.

A relatively simple outlier detection scenario (cf. Figure 3) assumes a Gaussian process generating the normal data (inliers) and a smaller but broader uniform distribution generating outliers (we could describe this distribution as background noise). It can be expected that a certain amount of outliers will be covered by the inlier distribution (resulting in false negatives) while there may be points in the tails of the Gaussian (i.e., true inliers) that are only recognizable as outliers (resulting in false positives), depending on the choice of a rejection threshold (the "dividing line" critically discussed by Bernoulli). If we insist, despite Bernoulli, in drawing a dividing line, a natural choice of this threshold could be the left and the right intersection point of both density functions. Moving the threshold will either decrease the number of false positives on the expense of increasing the number of false negatives or vice versa. Hence there is, in general, a maximum level of true positives any outlier detection can sensibly reach, depending on the overlap of inlier- and outlier-distributions. We can, however, assume that the confidence in rejecting observations as outliers (i.e., ideally, the probability of truly being an outlier) increases with the distance from the mean of the Gaussian.

Thus, in the example of network intrusion detection, if the system reports anomalous behavior, the interpretation and the final decision whether or not this is a network intrusion is still in the responsibility of the security administrator. In general terms the designation of any object by any outlier detection model as an outlier (perhaps with a certain outlier score) always needs to be interpreted in the specific context of an application.

There are some important conclusions from these observations for research on outlier detection methods. First, it will always be difficult to find reliably *all* outliers and to receive outlier signals *only* for outliers. Second, even if there is a high score or some other strong signal for a data object being an outlier, it is up to the domain scientist to decide on the actual outlierness. Third, having decided that an outlier is present, the question remains what to make out of it: reject it or accommodate it? What kind of error is responsible for the outlier? What costs are incurred with rejecting an inlier as an outlier or with

TABLE 1 True inliers/outliers versus apparent inliers/outliers

	Apparent outliers	**Apparent inliers**
True outliers	True positives	False negatives
True inliers	False positives	True negatives

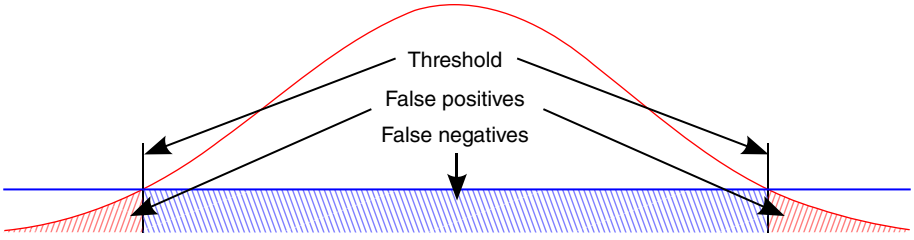

FIGURE 3 A simple outlier detection scenario: There is a maximum level of true positives any outlier detection can possibly reach, dependent on the overlap of outlier- (here the blue uniform distribution) and inlier distributions (here the red Gaussian distribution)

retaining an outlier as an inlier? In general, finally, we can conclude that there is no indisputable way of evaluating outlier detection methods. And this is probably the most important observation for data mining researchers when they attempt to evaluate a new outlier detection algorithm in comparison to existing approaches. There are, of course, some disputable ways of evaluating outlier detection methods (Emmott, Das, Dietterich, Fern, & Wong, 2013; Marques, Campello, Zimek, & Sander, 2015; Swersky, Marques, Sander, Campello, & Zimek, 2016). A thorough discussion of problems involved in such evaluations has been provided by Campos et al. (2016).

4 | CATEGORIES OF OUTLIER DETECTION METHODS

Letting aside a categorization of methods by algorithmic techniques (as provided in most other surveys), from a systematic point of view approaches can be roughly classified along the following different axes.

4.1 | Global versus local outliers

The distinction between *global* and *local* outliers (and correspondingly, between global and local outlier detection methods, introduced by Breunig, Kriegel, Ng, & Sander, 2000) refers to the scope of a database being considered when a method decides on the "outlierness" of a given object. While some methods take always the complete database into account, others consider only a local selection of database objects, for example, the k nearest neighbors of a point. Schubert, Zimek, and Kriegel (2014b) discussed the aspect of locality in outlier detection in depth.

Chandola et al. (2009) discern certain types of outliers as guiding categories, namely *point anomalies*, *contextual anomalies*, and *collective anomalies.* While the former exhibit outlier characteristics individually, for the latter two, outlier characteristics depend on the context, for example, the spatial or temporal neighborhood, or on the collective appearance of objects with corresponding characteristics, that is, a single object is unsuspicious while a collection of similar objects becomes spurious.

This distinction can also be seen as an instantiation of the "degree of locality"-categorization by Schubert et al. (2014b). Independently, Schubert et al. and Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2014) extend the classic notion of the "locality" (the spatial relation between the observations) to other contexts such as temporal or spatial–temporal relations. This way, different instantiations of "locality" can also cover "contextual anomalies" or "collective anomalies" defining "context sets" and "reference sets" (Schubert, Zimek, & Kriegel, 2014b) accordingly.

4.2 | Labeling versus scoring methods

At a different axis, one can distinguish *labeling* versus *scoring* outlier detection methods. The former are leading to a binary decision of whether or not a given object is an outlier whereas the latter are rather assigning a degree of "outlierness" to each object characterizing "how much" this object is an outlier. Often the binary decision (deriving a label) is implicitly or explicitly also based on scores or probability estimates and constitutes an attempt to "draw a dividing line" (Bernoulli (1777)). The statistical equivalent is a test, resulting in a decision on whether or not to reject the hypothesis that some observation is an outlier.

4.3 | Supervised versus unsupervised methods

Another classification of outlier approaches discerns between *supervised* and *unsupervised* approaches. A supervised approach is based on a set of observations where the status of being an outlier or not is known and the differences between those different types of observations are learned. Supervised approaches can be considered as very imbalanced classification problems (since the class of outliers has inherently relatively few members only). *Semi-supervised* approaches have attracted interest especially in outlier detection. The scenario here is that usual (normal) data are available in abundance while unusual (outlying) data are rare. What is more, in many application scenarios, like network intrusion detection, fraud detection, or fault detection in sensitive machines like airplanes, a realistic method should be able to detect new, unexpected, and unforeseen behavior, that is, data it could never have been trained to recognize as a specific class while any method could have been trained to recognize the normal data or behavior. This typification of outlier detection approaches is nicely discussed by Hodge and Austin (2004). Since both, semi-supervised (Warrender, Forrest, & Pearlmutter, 1999; Dasgupta & Nino, 2000; Dasgupta & Majumdar, 2002) and supervised (Abe, Zadrozny, & Langford, 2006; Hido, Tsuboi, Kashima, Sugiyama, & Kanamori, 2008, 2011; Phua, Alahakoon, & Lee, 2004; Steinwart, Hush, & Scovel, 2005; Zhu, Kitagawa, & Faloutsos, 2005) approaches, learn to discern the class of inliers versus the class of outliers, these methods often are also labeling methods, while unsupervised approaches are more often scoring methods. Labeling methods, however, could assign a confidence

estimate to their decision, and scoring methods could provide a threshold which score is to be considered (heuristically) good enough to decide that the corresponding object is an outlier. Labeling methods classify data into two classes, outliers versus inliers. Yet they can fall in either category unsupervised, semi-supervised, or supervised.

Note that unsupervised methods can be applied in supervised scenarios, although the application requires careful implementation following some principles to avoid certain pitfalls, as dicussed by Swersky et al. (2016).

4.4 | Parametric versus nonparametric methods

Finally, we can discern *parametric* from *nonparametric* approaches. Parametric methods assume a particular family of distributions (e.g., Gaussian distributions) to describe the (normal) data and fit the presupposed model to the data by learning the parameters of the model (e.g., for a Gaussian distribution: mean and standard deviation). Nonparametric approaches to outlier detection do not fit a presupposed model and do not assume a particular family of distributions. It should be noted, though, that nonparametric methods are not necessarily (and typically are not) parameter-free. Most nonparametric methods require the user to provide parameters. They are called nonparametric since they do not learn the parameters of some specific distribution (i.e., they do not fit a particular distributional model). However, nonparametric methods typically still have some implicit assumptions that might or might not be suitable to the data at hand.

4.5 | Discussion

These categories (global vs. local, labeling vs. scoring, supervised vs. unsupervised vs. semi-supervised, parametric vs. nonparametric) are categories regarding the general behavior or high-level properties. Often, basic techniques (statistical tests, distance-based, density-based approaches, and so on) are used for categorization of approaches. This is again an orthogonal category, and most techniques can be used to define methods falling in any of the previously given high-level categories. In the following, we survey outlier detection approaches according to the fundamental techniques used and the scientific background, discussing methods from the area of statistics (Section 5) and from the area of database and data mining research (Section 6), but putting these into relation to the high-level categories. Overall, we focus on unsupervised approaches, that is, the normal behavior is not known in advance but is considered to be represented by the major part of a database. Possibly there exist several different normal patterns and abnormality is attributed to objects not belonging to any major group or pattern. Neither is a spurious mechanism known producing outliers.

5 | STATISTICAL APPROACHES

In general, statistical research has been providing two major methodological approaches for dealing with the possible presence of outliers in data. First, the outliers should be identified for further study. The identification can lead to (a) rejection (removal) of spurious data; (b) recognition of important new information or even (c) revision of the model describing the data by incorporating allegedly outlying elements; and (d) refinement of the experimental setup. Second, modeling of the data is designed in a more robust way to deal with the possible presence of outliers without actually being interested in the identification of specific outliers. The latter is an important aspect also for many data mining approaches (robustness), the former is more directly the motivation of data mining research on outlier detection and, hence, shall be in focus here. Both high-level approaches are however not always completely distinct since a method of robustification or of accommodation of outliers may provide a method of identification of outliers as a by-product, or vice versa.

The fundamental problem stated by Bernoulli as *"I see no way of drawing a dividing line between those [observations] that are to be utterly rejected and those that are to be wholly retained"* (Bernoulli, 1777; Bernoulli & Allen, 1961) has been the motivation for researchers in statistics over decades, beginning with the first attempt to formalize a test by Peirce (1852).

5.1 | The statistical model

In general, statistical methods to outlier detection (identification, rejection) are based on presumed distributions of objects. The classical textbook of Barnett and Lewis (1994) discusses numerous tests for different distributions. The tests are optimized for each distribution dependent on the specific parameters of the corresponding distribution, the number of expected outliers, and the space where to expect an outlier. A commonly used rule of thumb, known as the "$3 \cdot \sigma$-rule", suggests that points deviating more than three times the standard deviation (*SD*) from the mean of a normal distribution may be considered outliers (Knorr & Ng, 1997).

A major problem of these classical approaches is obviously the required assumption of a specific distribution in order to apply a specific test. There are tests for univariate as well as multivariate data distributions but all tests assume a single, known

data distribution to determine an outlier. A classical, simple approach is to fit a Gaussian distribution to a data set, or, equivalently, to use the Mahalanobis distance, also known as quadratic form distance, based on the covariance matrix Σ as a measure of outlierness. Alternatively, the assumption of a mixture model consisting of one Gaussian distribution (of non-outliers) and one uniform distribution (of outliers) facilitates a simple expectation–maximization procedure (as an example see the work of Eskin, 2000). Sometimes, the data are assumed to consist of k Gaussian distributions and the means and *SD*s are computed data driven. A straightforward solution is to apply, for example, the expectation maximization (EM)-clustering algorithm (MacQueen, 1967) to derive models for k clusters C_i ($i = 1, \ldots, k$) and assign to each data object x the value $1 - \max_{i=1}^{k} \Pr(x|\, C_i)$ as outlier score.

However, mean, *SD*, or covariance are rather sensitive to outliers and the potential outliers are still considered for the computation step.

Possible effects of including outliers in parameter estimation are known as *masking* and *swamping*: Outliers *mask* their own presence by influencing the values of parameters as mean or covariance (resulting in false negatives), or *swamp* regular points to appear as outlying due to the influenced parameters (resulting in false positives). Figure 4 offers a visual explanation.

5.2 | Robust parameter estimation

There are proposals of more robust estimations of the mean and the *SD* in order to tackle the problem of outliers influencing the model estimation. Some of these still aim at identification (rejection) of outliers, others are more interested in robustification of methods or accommodation of outliers. Early examples were published in the late seventies and early eighties (Campbell, 1980, 1982; Maronna, 1976), yet research on this topic is still an issue in statistics (Hardin & Rocke, 2004; Rousseeuw & Van Driessen, 1999). A recent discussion of different techniques was presented by Rousseeuw and Hubert (2011). Robust methods generally are designed to be less influenced by potential outliers, for example, by assigning different weights to different objects, emphasizing near-by or central objects with a higher weight and limiting the influence of far-away or peripheral objects with smaller weights. Problems remain there, too, as the robustification requires a clue on the closeness of points before the distance measure (based, e.g., on a Mahalanobis distance) is available. Related discussions consider robustification of PCA (Delannay, Archambeau, & Verleysen, 2008; Kriegel, Schubert, & Zimek, 2008).

For multivariate outlier detection it is usually assumed that the data majority is generated from a multivariate normal distribution with a certain mean and covariance. Outliers are assumed to follow a different process, and thus they are supposed to be generated from a distribution which usually is not even specified. The observed data are then originating from a mixture of both distributions. This model is also called the ϵ-contamination model, and it goes back to the pioneering work of Peter Huber on robust location estimation (Huber, 1964), which can be considered as the starting point for a formal approach to the theory of robust statistics. Such a model might appear to be very specific, in particular, because multivariate normality is assumed for the data majority, and this is not common in the area of computer science. It might even be limiting for the application domain, and for more complex data structures one would need to generalize the model. However, having an underlying model is convenient for defining an outlier cutoff value, since this can be based on the model distribution, and consequently also for the evaluation, because outliers are observations that do not originate from the model distribution. The evaluation needs to be carried out differently if no underlying model is assumed.

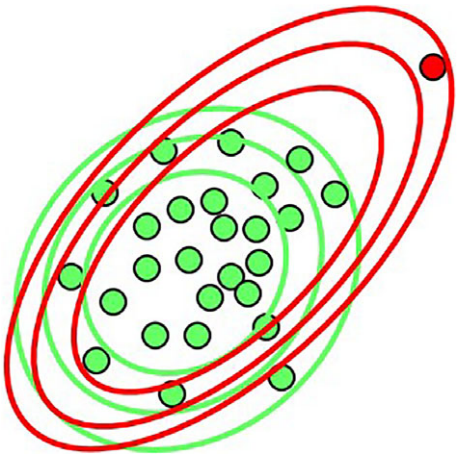

FIGURE 4 A distribution model (green density contours) computed for the inliers (green points) reveals the outlier (red point) as far of. If the outlier, however, was taken into account when fitting the distribution model to the data (red density contours), the outlier itself might be well covered by the model (it is masked), while some inlier might now appear as being too far off (the lower right inlier is swamped)

When using the ε-contamination model, multivariate outlier detection methods are typically employing Mahalanobis distances. Clearly, the ingredients for Mahalanobis distances, the multivariate mean and covariance, need to be estimated robustly in order to avoid that these estimates are affected by the outliers themselves. In one of the first papers devoted to this problem, the authors proposed either the MVE (Minimum Volume Ellipsoid) or the MCD (Minimum Covariance Determinant) estimator for robust location and covariance estimation (Rousseeuw & van Zomeren, 1990). Nowadays, various alternatives for robust location and covariance estimation are available (Maronna, Martin, & Yohai, 2006). Several aspects can be considered for the selection of appropriate estimators, such as the (theoretical) robustness properties (e.g., the breakdown point), the statistical efficiency (usually referring to an assumed multivariate normal distribution), the computational complexity, and of course the availability of an algorithm for the computation.

Assuming multivariate normality, the squared Mahalanobis distances are approximately chi-square distributed with p degrees of freedom, where p is the number of variables of the multivariate data. This approximation is also employed if robust estimators are used to compute Mahalanobis distances (Rousseeuw & van Zomeren, 1990), although also more thorough investigations on the distribution of the robust distances have been carried out (Hardin & Rocke, 2005). An outlying observation leads to an exceptionally high value of the (squared) Mahalanobis distance, and together with the assumed distribution of the squared distances one typically uses a quantile (e.g., $q_{0.975}$) of this distribution as outlier cutoff. More clearly, an observation is considered as multivariate outlier if its squared robust Mahalanobis distance exceeds the value of the quantile 0.975 of the distribution χ^2_p.

One could argue that using a quantile of the distribution as an outlier cutoff will still declare a small fraction (0.025) of observations as outliers, although they originate from the model distribution, and not from the distribution referring to the outliers. In other words, observations which are in the extremes of the model distribution should be distinguished from "real" outliers (contaminants) which are generated by an entirely different mechanism. This idea is the basis for an adaptive outlier cutoff, which looks for the supremum of the difference between the empirical distribution function of the squared Mahalanobis distances and the theoretical chi-square distribution in the upper tails of these distributions (Filzmoser, Garrett, & Reimann, 2005). In addition, the cutoff can be adjusted by the actual sample size and dimension of the data.

5.3 | Statistical testing

Another approach for multivariate outlier identification is based on statistical testing. The principle behind these multivariate outlier tests is to use high-breakdown estimators with good performance under the null hypothesis, stating that no outliers are present in the data. The common assumption is that the data are generated from a multivariate normal distribution, and the test is typically based on robustly estimated (squared) Mahalanobis distances (Becker & Gather, 1999). The test proposed in the work of Cerioli (2010) achieves high power because of an improved outlier cutoff value which better approximates the distribution of the robust distances, and an improvement of the robust covariance estimator based on re-weighting.

5.4 | High-dimensional, low-sample size data

The previously mentioned outlier detection procedures are not applicable for data with more variables than observations, typically high-dimensional data. These methods are typically based on the MCD estimator (similar for other affine equi-variant estimators), which cannot be computed in this case due to singularity of the covariance matrix. Therefore, and according to the need of robust estimators for high-dimensional low-sample size data in particular in applications of bioinformatics, different proposals of such robust covariance estimators and algorithms for their computation have been developed, such as the orthogonalized Gnanadesikan-Kettenring (OGK) estimator (Maronna & Zamar, 2002), the Stahel-Donoho estimator (Donoho, 1982; Stahel, 1981), an estimator based on spatial signs (Locantore et al., 1999), or an estimator employing the kurtosis as a measure for outlyingness (Peña & Prieto, 2001). These estimators (at least those which are computationally feasible) have been compared for outlier detection in simulation studies, and a new proposal of a fast algorithm for outlier detection in high-dimensional data has been made (Filzmoser, Maronna, & Werner, 2008).

Another class of covariance estimators can be used for high-dimensional outlier detection, namely those based on regularization. Since outlier detection methods based on the Mahalanobis distance involve the inverse covariance matrix, the so-called precision matrix, it is desirable to directly estimate this matrix. A popular nonrobust estimator is the graphical lasso (GLASSO; Friedman, Hastie, & Tibshirani, 2008) Its robust estimation leads to a different concept of robustness, namely the concept of cell-wise contamination (Alquallaf, Van Aelst, Yohai, & Zamar, 2009) in contrast to row-wise contamination. The idea is that down-weighting complete rows (observations) in high-dimensional data would lead to a severe loss of information, and thus down-weighting only single outlying cells of an observation is preferable. This concept has been used for robust precision matrix estimation (Öllerer & Croux, 2015; Tarr, Müller, & Weber, 2016). In the context of cell-wise outlier identification, progress has been made recently concerning a fast algorithm and diagnostic tools to investigate the outliers available (Rousseeuw & Van den Bossche, 2016).

5.5 | Nonparametric statistical methods

Typically, statistical approaches are parametric approaches, assuming a specific family of distributions to describe the normal data and estimating the parameters of this distribution. For two- or three-dimensional data, depth-based approaches are a nonparametric alternative. Depth-based approaches organize data objects in convex hull layers, expecting outliers from data objects with shallow depth values only (Johnson, Kwok, & Ng, 1998; Ruts & Rousseeuw, 1996; Tukey, 1977). These approaches do not require assumptions on specific data distributions except that they assume the normal data to follow *a single* distribution. They are, however, infeasible for data spaces of high dimensionality (in this case, "high" means roughly more than three dimensions) due to the inherent exponential complexity of computing convex hulls. On the other hand, the concept of matrix depth (Chen, Gao, & Ren, 2015) seems promising also in the high-dimensional case.

Another related but more generally applicable approach is Support Vector Data Description (Tax & Duin, 2004), enclosing normal data by a hyperplane, possibly working in a transformed data space using a kernel function. This is, however, an example of semi-supervised methods, that require normal data without contamination by outliers in order to fit a model.

In both variants, parametric and nonparametric, statistical approaches are usually global methods, that is, they compare the outlierness of an outlier candidate against all other objects.

6 | DATABASE-ORIENTED OUTLIER MODELS

Database-oriented research joined machine learning and statistical learning to shape the area of data mining with the particular focus on efficiency and scalability. The seminal outlier detection methods in the database area were therefore, though motivated by the statistical modeling of outliers as deviating from the rest of the data, shifting the focus from a model-driven thinking about outlierness to an algorithm-driven thinking about the *efficient* identification of outliers. Here, however, we are not interested in the algorithmic aspect but in the subsisting outlier models, that were often substantially simplified compared to the statistical origin.

6.1 | Deviation-based outliers

Deviation-based outlier detection groups objects, captures some characteristics of the group, and considers those objects outliers that deviate considerably from the general characteristics of the groups. Foundations to this reasoning have been laid out by Thompson (1935), problems associated with this reasoning have been expounded by Pearson and Chandra Sekar (1936).

The basic approach was suitable for rejecting one outlier, but not several outliers. Hence, the more recent deviation-based approaches propose heuristics to select groups of outliers, mostly based on random groupings (Arning, Agrawal, & Raghavan, 1996; Chakrabarti, Sarawagi, & Dom, 1998; He, Deng, Xu, & Huang, 2006; Sarawagi, Agrawal, & Megiddo, 1998). The forming of groups at random is admittedly rather arbitrary and so are the results depending on the selected groups. Forming groups at random, however, avoids exponential complexity. The commonly pursued greedy approaches more or less also rely on the assumption that there is one single usual distribution from which the outliers can be found as deviating. This is related to the problem of identifying several outliers simultaneously. While statistical tests for rejection of outliers formerly have been applied consecutively in order to reject several outliers, approaches for simultaneous rejection have also been proposed. The seminal paper by Rosner (1975) proposed an approach technically similar to those pursued later by deviation-based outlier detection.

Another point of view, essentially equivalent to deviation-based approaches, has been brought forward based on information theory. The complexity of a data set, as measured by entropy or some related measure of information, is assumed to be increased by outliers. Thus, for a data set $\mathcal{D}$, the minimal subset $O \subseteq \mathcal{D}$ is sought that maximally reduces the data complexity of $\mathcal{D} \setminus O$. Since there is not necessarily a unique optimum to this problem, different solutions pursue different optimization methods and use different measures of complexity (such as Kolmogorov complexity, entropy, relative uncertainty; Chandola et al., 2009; Smets & Vreeken, 2011).

6.2 | Density-based outliers

Density-estimation (Scott, 2008; Silverman, 1986) is the major nonparametric counterpart to the statistical approach of fitting a model of density distribution to some data set. Given a data sample, the underlying density distribution is estimated locally, basically by counting how many data objects are present in the sample in some local volume. The most basic form of a density estimate is a histogram, a technique that is useful in statistics for univariate data but becomes quickly infeasible for an increasing number of variables (dimensions). For higher dimensional (multivariate) data, the basic variants are to either check how large the volume V is that is required to cover k-nearest neighbors around some given estimation point p or to fix a volume V (typically by fixing the radius of a ball, in database terminology: by ϵ-range queries), centered at p, and count how many

points are present in this volume. In both cases, $\frac{\text{number of points}}{V}$ is an estimate of the local density, where either the number of points is fixed (= k) or the volume V is fixed to allow comparison of different density estimates. Further variants can assign different weights to neighbors depending on their distance to p (e.g., by choosing different kernels for a kernel density estimation, KDE). Selecting values for k or V remains a challenge.

In the literature, often "distance-based" and "density-based" methods are distinguished. This can be seen as an artifact due to the algorithmic design but the distinction is misleading when we consider the models: outlier models from both pseudo categories are based on density estimates.

One could argue that, while we can meaningfully assess distances on many kind of non-Euclidean data, it might not be in all cases necessarily clear how to interpret density estimates derived from non-Euclidean data. However: firstly, "density-based" methods have been successfully applied in non-Euclidean data (Schubert, Zimek, & Kriegel, 2014b), and secondly, the use made from the distance assessment in "distance-based" methods is effectively not different in any way from a density estimate, as we detail below. We thus subsume both here under the category "density-based."

The misconception is perhaps a consequence of the naming of the seminal method in the database literature. Knorr and Ng called their method DB-outlier ("DB" for: distance-based; Knorr & Ng, 1997, 1998; Knorr, Ng, & Tucanov, 2000). Their method is motivated by distribution-based approaches but uses a quite simplified model that can be understood as using local density-estimates, centered around points. This model relies on the choice of two thresholds, ε and π. In a database $\mathcal{D}$, an object $x \in \mathcal{D}$ is an outlier if at least a fraction π of all data objects in $\mathcal{D}$ has a distance above ε from x. A meaningful value for the minimum distance ε depends on the range of attribute values in the data space, on applied normalization procedures, and on the dimensionality of the data. For the threshold π, a meaningful value depends on the size of the database and assumptions on distributions of database objects.

Originally, the distance-based outlier model is a labeling approach, just deciding whether or not the threshold is exceeded. Omitting the threshold π and reporting the fraction π_x of data objects $o \in \mathcal{D}$ where distance$(x, o) > \varepsilon$ results in an outlier score for x in the range [0, 1]:

$$\text{score}(o) = \frac{|\{x \in \mathcal{D} : \text{distance}(x,o) > \varepsilon|\}}{|\mathcal{D}|}. \tag{1}$$

We can interpret this as the curried form of the original DB-outlier method (Schubert, Zimek, & Kriegel, 2014b). The additional parameter π can then be seen as fixing a decision threshold for the score. Scores that are larger than the threshold are qualifying the associated object as outlier.

Instead of this fraction, we could equivalently use the number of points within the ε-range:

$$\text{score}(o) = \frac{|\{x \in \mathcal{D} : \text{distance}(x,o) \leq \varepsilon)\}|}{V_\varepsilon}. \tag{2}$$

The resulting ranking would be inverted but otherwise equivalent. This way it becomes quite obvious that this constitutes a density-estimate centered at o. The lower this estimated density, the more we consider o being an outlier. (And we can drop the denominator if we are only interested in the ranking and not in the density-estimate as such, as the volume V_ε is always the same.)

The alternative formulation of a density-estimate would be to drop the parameter ε and to report the distance required to include a certain amount of points. This is the essence of the notion of $k$nn-outliers (kth nearest neighbor), that have been introduced in the database literature by Ramaswamy, Rastogi, and Shim (2000). For each object o, the distance to its kth nearest neighbor (let us denote this distance as k-dist) is used as an outlier score and the objects are ranked according to these scores, that is, objects with a larger distance to their k-th nearest neighbor are more prominent outliers. As a variant, formulated by Angiulli and Pizzuti (2002), the sum of distances to all points within the set of k nearest neighbors (called the 'weight') is proposed as an outlier degree. In both cases, the inverse of the (average of) distances constitutes a proper density-estimate, and would deliver the same (though inverted) ranking:

$$\text{score}(o) = \frac{k}{k\text{-dist}(o)} \tag{3}$$

or

$$\text{score}(o) = \frac{k}{\frac{1}{k}\sum_{i=1}^{k} i\text{-dist}(o)}. \tag{4}$$

Variants of this global comparison of variations in "local" density take reverse neighborhood into account (Hautamäki, Kärkkäinen, & Fränti, 2004; Radovanović, Nanopoulos, & Ivanović, 2014). Angle-based outlier detection (Kriegel, Kröger,

Schubert, & Zimek, 2008) takes the distances as weight into account. The outlier degree for each point is based on the angles to all other pairs of points. The variance of these angles, weighted by the involved distances, is the outlier factor.

These methods have in common that they use local density estimates but are not interested in the actual family of the density-distribution. They are thus *nonparametric* approaches. Despite the locality of density estimates, they share with the statistical approaches the property of being *global* in the sense that the outlierness of some given object is evaluated in direct comparison to all other objects: All local density estimates are compared with each other on a global scale.

This property was challenged as being not sufficiently adaptive to more complex data that might comprise clusters of different local density. This is the motivation for the so-called local outlier detection methods, with the seminal method LOF (Local Outlier Factor) by Breunig et al. (2000). The LOF compares the density (as estimated by a value called the local reachability density, lrd) of each object o of a database $\mathcal{D}$ with the density of the k nearest neighbors of o.

The density estimate lrd is defined as the inverse average reachability distance from the neighbors

$$\mathrm{lrd}(p) := 1 \Big/ \frac{\sum_{o \in \mathrm{kNN}(p)} \text{reach-dist}_k(p,o)}{|\,\mathrm{kNN}(p)\,|}, \tag{5}$$

where the (asymmetric) reachability distance is given by:

$$\text{reach-dist}_k(p,o) = \max\{k\text{-dist}(o), d(p,o)\}. \tag{6}$$

The effect of this definition of a local density estimate is a certain smoothing to reduce variability. The final LOF score for some point p is the comparison of the locally relevant lrd values, that is, the density estimate of the neighbors of p in relation to the density estimate of p:

$$\mathrm{LOF}_k(p) = \frac{1}{|\,\mathrm{kNN}(p)\,|} \sum_{o \in \mathrm{kNN}(p)} \frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}_k(p)} \tag{7}$$

A LOF value of approximately 1 indicates that the corresponding object is located within a region of homogeneous density (i.e., a cluster). The LOF score achieves its highest values when the local density estimate (lrd) of the test point is small relative to the estimates of its nearest neighbors. Thus, the higher the LOF value of an object o is, the more distinctly is o considered an outlier.

We can now with Schubert, Zimek, and Kriegel (2014a) identify several ingredients or building blocks of outlier detection methods based on density estimates: (a) the density estimation method or kernel (i.e., the method to build a local model), (b) the assessment of the neighborhood or distance (the context or considered information for building a local model), and (c) the comparison of the local density estimate to other density estimates that can differ both in the comparison method (i.e., how to compare the local model to other models) and in the reference set used for the comparison.

Several extensions and refinements of the basic LOF model have been proposed in the literature. They can typically be put into relation to LOF by identifying in which of these ingredients they vary and how they vary. For example the connectivity-based outlier factor (Tang, Chen, Fu, & Cheung, 2002) and Influenced Outlierness (Jin, Tung, Han, & Wang, 2006) are both changing the definition of neighborhood to be taken into account for the density estimate and the comparison of local density estimates (i.e., they change both context and reference). In these both, as well as in many other variants, the lrd is replaced by simpler (or sometimes more complex) variants of density estimation. Using just the k-dist as density estimate instead of lrd has been defined as an explicit method "SimplifiedLOF" by Schubert et al. (2014b), but has been used as such implicitly (and probably unintentionally and unaware of the simplification of LOF) several times before in various papers. Local density factor (Latecki, Lazarevic, & Pokrajac, 2007) and kernel density estimation (KDE) outlier score (Schubert, Zimek, & Kriegel, 2014a) replace the lrd explicitly by KDE in different formulations.

Papadimitriou, Kitagawa, Gibbons, and Faloutsos (2003) propose another local outlier detection schema named local outlier integral (LOCI) based on the concept of a multigranularity deviation factor (MDEF). The main difference between the LOF and the LOCI outlier model is that the MDEF of LOCI uses ε-neighborhoods rather than k nearest neighbors. The authors propose an approximate algorithm computing the LOCI values of each database object for any ε value. The results are displayed as an outlier plot per object, that is, they observe essentially for each object how its density estimates behave when changing the kernel bandwidth.

Many variants of local outlier detection and their relationship to LOF as well as an adaptive interpretation of the concept of *locality* have been discussed by Schubert et al. (2014b).

LOF and its variants can find *local* outliers, that exhibit a lower density than *their neighbors* (as opposed to the *global* outliers, that need to be prominent when compared to *all* objects). While global outliers typically would also be local outliers (and thus would also be ranked prominently by LOF), the reverse does not necessarily hold. By using density-estimation techniques, these methods are not assuming particular families of distributions and are altogether *nonparametric* approaches. They

assume, however, that normal data follow somehow homogeneous density-distributions, that could be grasped as clusters. Yet the clustering structure of the dataset is not modeled explicitly.

6.3 | Cluster-based outliers

Cluster-based outlier detection methods define outliers based on an explicit clustering of a data set. Objects that are not covered by the clusters (or by a sufficient number of clusters in case of applying multiple clustering procedures or a clustering ensemble) are then deemed outliers. The density-based methods (Section 6.2) rely at most on an *implicit* cluster-assumption: the fact that outliers are assumed where the local density is smaller than other local density estimates could mean that higher local densities are encountered where we have clusters in the data set. This remains a relative notion of clustering though, cluster-based outlier detection relies on some explicit clustering-step and therefore binds the notion of outlierness to some specific clustering model. So could the clustering method DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), that allows explicitly for noise objects that do not belong to any cluster, be used as an outlier detection method. Regarding those noise objects as outliers would be equivalent to defining an absolute density threshold to distinguish between inliers and outliers. Objects that exhibit a local density below that threshold are outliers. Thus in this model, outliers are global, and we have a binary decision (label). The GLOSH method (Campello, Moulavi, Zimek, & Sander, 2015) relates potential outliers to a hierarchy of density estimates, thus dropping the strict global threshold and retrieving a ranking of outliers that exhibits both global and local properties and is adaptive to local variations in the density. As density-based clustering is nonparametric, so the outlier model based on such clustering approaches is also nonparametric.

If the clustering model is parametric, such as in using EM clustering for outlier detection (Eskin, 2000) or an adapted version of k-means (Chawla & Gionis, 2013), we get a stronger connection to statistical approaches (Section 5). The method k-means (Chawla & Gionis, 2013), for example, would roughly relate to a repeated, iterative, and somehow robustified estimate of standard normal distributions.

Also the method of Paulheim and Meusel (2015) models the data explicitly and measures the deviation of each point from its expected location (as predicted by some regression learning algorithm). They do not allow for different clusters, though, but assume a single (i.e., global) normal pattern. Thus their method is suitable to identify global outliers.

Cluster-based outlier detection has been extended to the notion of outliers in subspaces, where subspace clustering algorithms are applied several times. The identification of outliers is then based on ranking the points according to the number of times they do not belong to any cluster in the different subspace-clustering results (Müller, Assent, Iglesias, Mülle, & Böhm, 2012; Müller, Assent, Steinhausen, & Seidl, 2008).

6.4 | Adaptations to special types of data

As data mining often has to deal with data of particular characteristics, many specialized outlier detection methods have been proposed.

- Fundamental challenges of high-dimensional data for outlier detection have been discussed by Zimek et al. (2012), and various methods dedicated to outlier detection in high-dimensional data have been proposed (Dang, Assent, Ng, Zimek, & Schubert, 2014; de Vries, Chawla, & Houle, 2012; Keller, Müller, & Böhm, 2012; Kriegel, Kröger, et al., 2008; Kriegel, Kröger, Schubert, & Zimek, 2009b, 2012; Müller, Schiffer, & Seidl, 2010, 2011; Nguyen, Gopalkrishnan, & Assent, 2011; Pham & Pagh, 2012).
- Another large subtopic is outlier detection in spatial data, trajectory data, or some other notion of separating context and indicator variables (Chawla & Sun, 2006; Hayes & Capretz, 2015; Janeja, Adam, Atluri, & Vaidya, 2010; Kou, Lu, & Chen, 2006; Leach, Sparks, & Robertson, 2014; Lee, Han, & Li, 2008; Liang & Parthasarathy, 2016; Liu, Lu, & Chen, 2010; Lu, Chen, & Kou, 2003; Shekhar, Lu, & Zhang, 2003; Song, Wu, Jermaine, & Ranka, 2007; Sun & Chawla, 2004). The spatial neighborhood can be interpreted as a special case of locality for local outlier detection. We refer to Schubert et al. (2014b) for further discussion.
- Outlier detection in graph data (Perozzi, Akoglu, Sánchez, & Müller, 2014; Sánchez, Müller, Irmler, & Böhm, 2014; Sánchez, Müller, Laforet, Keller, & Böhm, 2013; Wang & Davidson, 2009) includes, for example, community outliers (Gao et al., 2010). A recent survey on this area has been provided by Akoglu et al. (2015). Also the context considered in a graph can be seen as a special case of locality for local outlier detection (Schubert, Zimek, & Kriegel, 2014b)
- Detecting outliers in time series has also found much interest (Abraham & Box, 1979; Fox, 1972; Jagadish, Koudas, & Muthukrishnan, 1999; Takeuchi & Yamanishi, 2006; Tsay, 1988) and comes with special challenges. Further discussion can be found in the survey by Chandola et al. (2009).

- Generalizing from time series to sequences can again take very different flavors of outlierness (Barua & Sander, 2014; Sadoddin, Sander, & Rafiei, 2016). A dedicated survey on outlier detection in sequence data has been provided by Chandola et al. (2012).
- Also related and in a sense even more general is the problem of outlier detection in streaming data (Angiulli & Fassetti, 2010; Assent, Kranen, Baldauf, & Seidl, 2012; Franke et al., 2009; Kontaki, Gounaris, Papadopoulos, Tsichlas, & Manolopoulos, 2016; Pokrajac, Lazarevic, & Latecki, 2007; Yamanishi, Takeuchi, Williams, & Milne, 2000, 2004). Sadik and Gruenwald (2013) give an overview on research issues in this challenging setting for outlier detection.
- A further specific data structure is directional or circular data. Examples are wind directions of arrival times (on a 24-hr clock). Different models are used for circular data, such as the Von Mises distribution, the wrapped normal, or wrapped Cauchy distribution, or the Cardioid distribution (Mardia & Jupp, 2000). In order to identify outliers, weighted likelihood estimation can be used, which is a modified form of maximum likelihood estimation, where each score is associated a weight. The weight can be derived from a minimum distance estimator like the Power Divergence Measure (Cressie & Read, 1988). Details to this approach are provided by Agostinelli (2007).

Many other areas with special requirements could be mentioned, such as categorical or ordinal data (Akoglu, Tong, Vreeken, & Faloutsos, 2012; Das & Schneider, 2007; Otey, Ghoting, & Parthasarathy, 2006; Smets & Vreeken, 2011; Tang, Pei, Bailey, & Dong, 2015; Yu, Qian, Lu, & Zhou, 2006), binary data (Smets & Vreeken, 2011), or uncertain data (Jiang & Pei, 2011; Liu, Xiao, Cao, Hao, & Deng, 2013). Also mining events or trends can be seen as variant of mining anomalies (Schubert, Weiler, & Kriegel, 2014, 2016). The amount of literature is growing fast, and to cover such specialized areas would require dedicated specialized surveys. However, just as different categories of clustering can be seen as being actually closely related to each other and even to different data mining tasks such as frequent pattern mining (Zimek & Vreeken, 2015) also these different scenarios for outlier detection could be seen as variants or different flavors of the same fundamental problem. As pointed out by Schubert et al. (2014b), instantiating the essential building blocks of a general algorithmic approach (e.g., by dedicated distance measures or by a specific notion of locality to grasp context and reference sets for the outlier definitions) can relate well-understood models, such as the LOF model, to very different problems such as spatial outliers, outliers in video streams, or graph outliers—a point of view that can probably also be extended to time series, to sequences, to streaming data, or to other special cases.

6.5 | Algorithmic variants for improving efficiency

As we have seen (Section 6.2), the work of Knorr and Ng (1997, 1998) was originally motivated by statistical reasoning but simplifies the approach to outlier detection considerably. Such simplifications are motivated by the need for scalable methods for large data sets. Their simplification and efficient algorithmic design, in turn, inspired many new outlier detection methods within the database and data mining community over the last two decades. For most of these approaches, however, the connection to a statistical reasoning is not obvious any more.

Discussing such variants and efficiency techniques would take a survey in its own and would deviate too much from the focus we chose for this article. A selection of prominent general examples could be a long list of references (Angiulli & Fassetti, 2009; Angiulli & Pizzuti, 2002, 2005; Bay & Schwabacher, 2003; Bhaduri, Matthews, & Giannella, 2011; de Vries et al., 2012; Fan, Zaïane, Foss, & Wu, 2006; Ghoting, Parthasarathy, & Otey, 2008; He et al., 2006; Jin, Tung, & Han, 2001; Knorr et al., 2000; Kollios, Gunopulos, Koudas, & Berchthold, 2003; Nguyen & Gopalkrishnan, 2009; Pei, Zaïane, & Gao, 2006; Ramaswamy et al., 2000; Sugiyama & Borgwardt, 2013; Wang, Parthasarathy, & Tatikonda, 2011) and yet would be a close to arbitrary selection from the literature. Some discussion of efficiency issues in outlier detection can be found elsewhere, though. Zimek et al. (2012) describe some such approximation techniques (e.g., different kinds of projections) for the special challenges of high-dimensional data. A helpful more general categorization of fundamental techniques for such approximations (in database-terminology: filter-refinement-techniques) is given by Orair, Teixeira, Wang, Meira Jr, and Parthasarathy (2010).

Many of these variants target the problem of delivering the top-n outliers (Angiulli & Fassetti, 2009; Angiulli & Pizzuti, 2005; Bay & Schwabacher, 2003; Ghoting et al., 2008; Jin et al., 2001; Kollios et al., 2003; Kriegel, Kröger, et al., 2008; Nguyen & Gopalkrishnan, 2009; Orair et al., 2010; Ramaswamy et al., 2000), where the user should specify in advance some number n of outliers to retrieve. The algorithms then use typically filter-refinement approaches (Kröger & Renz, 2009), looking for candidate outliers by some approximation of the outlierness characteristic (e.g., using an approximate distance) and refining only those candidates that still have a chance to be placed among the top-n *exact* outliers in the final ranking of outlier scores.

A problem for the usability of these approaches could be that the user in many applications does not know in advance how many outliers should be retrieved. But even if there is some estimate (or upper bound) on the expected number $|O|$ of outliers

in some application, the rankings are typically not good enough to report all $|O|$ outliers in the top $|O|$ positions. So n should be chosen larger than $|O|$. How much larger depends on the data set, on the suitability of the outlier model used for this data set, and on the importance of finding *all* or at least most of the outliers. If finding all outliers is important, it does not seem uncommon that n must be a multiple of $|O|$ in the order of $10 \cdot |O|$ (Campos et al., 2016; This problem also has repercussions for the evaluation of outlier detection results, we will therefore come back to this in Section 7.)

Relating these approaches to the point we are taking in this survey, we could say that the community was focusing on computing some property ("outlier score") ever more efficiently but lost a clear connection to a statistical notion of this property. Database-oriented solutions model the "original" statistical meaning only approximately, and many efficiency-tuned variants are based on approximations of such database-oriented models being in turn approximations (often vague ones) of the original statistical meaning of outlierness. From this observation, we do not conclude that such "approximations of approximations" are meaningless. But it should be acknowledged that it is not necessarily of the utmost importance to approximate approximations as good as possible. Rather, an approximate solution can help to emphasize certain characteristics just as the copy of a copy would typically give a weaker impression of all the details of the original picture but might actually increase the contrast of the most striking patterns. If these are the important characteristics for outlier detection, an approximation can be better than the "original" ("exact") outlier score. This has been discussed and exemplified by Kirner, Schubert, and Zimek (2017).

In conclusion, rather than tuning for efficiency some algorithmic approaches to outlier detection while losing the connection to a solid statistical notion of outlierness ever more, we suggest to take the original statistical notion of outliers again and again as the target for approximate and efficiently tuned database-oriented methods for outlier detection.

7 | EVALUATION OF OUTLIER DETECTION RESULTS

Most of the outlier detection methods that we discussed deliver as a result a complete ranking of the database according to the method's measure of outlierness or could be used that way. For example, if a method is based on a statistical test, the significance level could be used as a measure of outlierness to define a ranking. Methods that are based on a classifier's decision could often also deliver the classifier's confidence or class probability estimate. Methods that focus on the top-n outliers only would typically also deliver a complete ranking where however the ranking below rank n is not reliable (as the measure of outlierness has not necessarily been refined).

Rankings of outlier scores, in turn, can be translated into hard decisions only if we have thresholds available, which outlier score is high enough to "really" signify outlierness. This is rarely possible. Recall the discussion of the meaning of outlierness in Section 3: outlier detection methods are supposed to report objects that are *suspicious.* Many more objects can be expected to look suspicious that eventually do not qualify as actually being outliers while some "real" outliers might not look suspicious enough. For any realistic problem, we can therefore not expect to have all "real" outliers and no actual inliers reported on the top positions of the ranking.

If we set these considerations aside for a moment, let us assume that a target number of outlier candidates n is specified in advance. Then perhaps the seemingly most natural evaluation measure for the outlier ranking is the precision at n ($P@n$), defined as the proportion of correct results in the top n ranks (Craswell, 2009a). Let us assume we have a database $\mathcal{D}$, containing N objects, namely $|O|$ outliers $O \subset \mathcal{D}$ and $|I|$ inliers $I \subseteq \mathcal{D}$ ($\mathcal{D} = O \cup I$, $O \cap I = \emptyset$). Then a formal definition of $P@n$ is given by:

$$P@n = \frac{|\{o \in O \mid \text{rank}(o) \leq n\}|}{n}, \tag{8}$$

assuming that the outlier ranking is unique (otherwise, ties have to be broken arbitrarily but consistently).

Although seemingly a natural choice for an evaluation measure, $P@n$ and related measures such as recall, accuracy, and F1 are actually quite problematic if we take again into account the above considerations. On top of that it is rather unclear how to fairly choose the parameter n—or some cut-off value to decide where in the ranking to distinguish between outliers (ranked higher) and inliers (ranked lower).

Choosing n equal to the number of outliers in the ground truth, $n = |O|$, results in the R-Precision measure (Craswell, 2009b). However, when the number of outliers $n = |O|$ is very small relatively to N, typical values of $P@n$ can be deceptively low even for reasonably good rankings, and thus not very informative as such. This is often the case, since outliers are supposed to be rare. In a typical scenario, we have $|O| \ll |I|$, $|I| \approx N$, and no guarantee that $|O| > 0$. Let us assume, for example, a data set with 10 outliers and 1 million inliers. An algorithm that assigns the true outliers to the (quite high) ranks 11–20 will nevertheless have a P@10 of 0, but a P@20 of 0.5 (which is deceptively low despite the rather good ranking). Note that, in

such a case, 0.5 is already the maximum $P@n$ for all n. Furthermore, we could not distinguish the quality in terms of P@20 between this result, ranking all 10 outliers on ranks 11–20, and a perfect result, ranking all outliers on ranks 1–10.

It is therefore more meaningful to use measures that average across different values of n. One possibility is to use the average precision (Zhang & Zhang, 2009):

$$\mathrm{AP} = \frac{1}{|O|}\sum_{o \in O} P@\mathrm{rank}(o). \tag{9}$$

The values of $P@n$ are averaged over the ranks of all outlier objects $o \in O$, thus one has not to choose a particular value for n. Note that, this way, we can also distinguish the two scenarios (rank 1–10 vs. rank 11–20). However, this measure does not seem to be used quite frequently, perhaps because it is hard to interpret in absolute terms as its behavior still depends strongly on the number of outliers in a dataset.

For both, precision at n and average precision, Campos et al. (2016) suggested adjustments for chance. An adjustment for chance is helpful if a measure is to be interpreted in absolute terms. If the performance of methods is compared over different data sets with different proportions of outliers, without such an adjustment the comparison can be misleading.

However, the most popular and meaningful evaluation measure in the literature on unsupervised outlier detection is based on a curve known as the Receiver Operating Characteristic (ROC). This measure also avoids the choice of n, as the curve is obtained by plotting, for all possible choices of n, the true positive rate (the proportion of outliers correctly ranked among the top n) versus the false positive rate (the proportion of inliers ranked among the top n). If the evaluated ranking was random, the curve can be expected to remain close to the diagonal. A perfect ranking (i.e., all outliers are ranked ahead of any inliers) produces a curve consisting of a vertical line at false positive rate 0 and a horizontal line at the top of the plot (indicating a true positive rate of 1 for every false positive rate > 0). This measure nicely addresses the problem that renders other measures problematic: the problem of imbalance between the amount of outliers (positive class) and inliers (negative class) ($O \ll I$).

Using the ROC, we interpret the outlier detection result as not reporting any inliers but only outliers, yet in a certain order (i.e., we do not have true or false negatives, we are only encountering objects for the first column of Table 1). Each step further down in the ranking contributes therefore either a true positive (an outlier, according to ground truth) or a false positive (an inlier, according to ground truth). The false positive *rate* is normalized by the maximal number of false positives (i.e., the number of inliers), and the true positive *rate* is normalized by the maximal number of true positives (i.e., the number of outliers).

For the comparison on a larger amount of data sets and parameter settings, the visual comparison of curves can be avoided by summarizing each ROC curve by a single value, the area under the ROC curve (thus called ROC AUC). The ROC AUC value ranges between 0 and 1. A perfect ranking of the database objects would result in a ROC AUC value of 1, whereas an inverted perfect ranking would result in a ROC AUC value of 0. A random ranking of the database objects would result in a ROC AUC value close to 0.5.

The ROC AUC value can be interpreted as the average of the recall at n (i.e., the true positive rate over the n top-ranked objects), where n is taken over the ranks of all inlier objects in I.

There is also a probabilistic interpretation (Hanley & McNeil, 1982): If we choose a pair (o, i) at random from $O \times I$, the ROC AUC value of a ranking is the probability that o and i are ranked in the correct order (i.e., o appears in the ranking before i).

Note that these evaluation measures require the availability of external ground truth (i.e., labels identifying outliers vs. inliers). They are therefore useful only for the evaluation of outlier detection methods on benchmark data (such as provided by Campos et al., 2016) or in supervised learning scenarios where also labeling approaches are used and could be evaluated with techniques as in classification.

In a real application on unknown data without labeled examples, internal evaluation measures (i.e., measures that rely only on the data and do not take external knowledge into account) would be useful. So far, however, only one internal evaluation measure for outlier rankings is known, IREOS (Marques et al., 2015), that is restricted to the evaluation of top-n results (i.e., we have to give the parameter n). IREOS assesses the separability or classification-hardness of the detected outliers. This measure is computationally quite expensive, as it trains classifiers with varying parameters to separate each outlier individually from the rest of the database.

8 | BACK TO THE FUTURE: OUTLIER SCORES AS PROBABILITY ESTIMATES

So far, we followed the path of ongoing research from the statistical roots to efficient database solutions and observed the loss of connection between the two fields. Is there a way back to the roots that simultaneously is promising for future development? Can modern efficient database methods be reintegrated into statistically sound models?

By design most outlier detection models make explicit or implicit assumptions. Eventually, any outlier score provided by an outlier model should help the user to decide on the actual outlierness. For most approaches, however, the outlier score does not translate easily to an outlier probability. Indeed, the scores provided by varying methods differ widely in their scale, their range, and their meaning. In some cases, high values of an outlier score mean that the corresponding database object is *not at all* an outlier. In other cases a higher value indicates more "outlierness." In some cases the minimum occurring outlier score is around 1, in other cases 1 is the maximum value. For many methods the scaling of occurring values of the outlier score even differs within the same method from data set to data set, that is, outlier score x in one data set means, we have an outlier, in another data set x is not at all an extraordinary score. Obviously this makes the comparison of different outlier detection models easily leading astray both the data miner who wants to evaluate the performance of a newly proposed method and the nonexpert user who wants to know which method is best suited for a given application.

However, the ranges and scales of outlier scores are usually just side products of some formula used in the specific approach to assign an outlier score and, hence, are not necessarily properties of the method.

There have been some attempts to provide outlier scores with a clearer statistical meaning. LoOP (an adaptation of LOF; Kriegel, Kröger, Schubert, & Zimek, 2009a), SOS (using reverse neighborhood relationships; Janssens, Huszár, Postma, & van den Herik, 2012), and COP (using subspace projections; Kriegel et al., 2012) are examples for methods that aim to deliver not just outlier scores without a clear interpretation but to provide, as scores, probability estimates (based on different model assumptions). An alternative is to relate the outlier score to the number of *SD*s (Papadimitriou et al., 2003; Paulheim & Meusel, 2015). Other approaches (Gao & Tan, 2006; Kriegel, Kröger, Schubert, & Zimek, 2011; Schubert, 2013) aim at converting outlier scores (produced by some outlier detection method) into outlier probabilities. Such meta-methods typically take some outlier score distribution (the outlier scores provided by some previously applied outlier detection method) as input and fit some distribution model to the distribution of outlier scores.

Although this fistful of methods can only be seen as initial steps in the direction of providing probabilistically interpretable outlier scores and thus leading the efficient database solutions back to their statistical roots, the direction is interesting and important, as it might be useful in various potential applications that we sketch in the following.

8.1 | Score-based evaluation

Probably the main reason why there has been little attention to the outlier scores as such is that they—so far—have not been used much. Common evaluation measures such as Precision at n and the area under the ROC curve (ROC AUC) (see Section 7) do not evaluate the scores, but only the ranking of the objects. This also reflects a usage scenario for outlier detection, where the system would rank the objects by outlierness and the operator then—as time permits—inspects the top-ranked anomalies manually.

However, there are clear requirements for such outlier scores from a statistical point of view, as formulated by Hawkins (1980):

> *"A sample containing outliers would show up such characteristics as large gaps between 'outlying' and 'inlying' observations and the deviation between outliers and the group of inliers, as measured on some suitably standardized scale."*

The question is therefore how to achieve "a suitably standardized scale" for measuring outlierness and how to quantify "large gaps" in the context of efficient, database-oriented, multivariate unsupervised outlier detection methods. Procedures for normalization, standardization, or probabilistic interpretation of outlier scores (Gao & Tan, 2006; Kriegel et al., 2011; Schubert, 2013) are tackling this aspect. The related question of how to evaluate outlier *scores* rather than just outlier rankings has been brought forward and discussed by Schubert, Wojdanowski, Zimek, and Kriegel (2012). Their solution is basically a similarity measure for score vectors that takes into account the class imbalance between outliers and inliers. It has been used for the improvement of outlier ensembles (see Section 8.2) rather than for the mere evaluation. For (supervised) evaluation, such an approach would be useful if the ground truth would not only provide labels (outlier vs. inlier) but that would also ascribe how prominent some outlier is or that would even annotate something like "true" outlier probabilities. This is rarely the case in such purity. However, the method of Schubert et al. (2012) could also find use in cost-based evaluation, for example, to put more emphasis on important outliers that might, however, be hard to detect—a scenario that remains to be studied.

8.2 | Ensembles for outlier detection

Ensemble learning, that is, the combination of several learners or models to some meta-learner, meta-predictor, or meta-model, has a rich tradition and solid theory in the context of classification (Brown, Wyatt, Harris, & Yao, 2005; Dietterich, 2000; Kuncheva & Whitaker, 2003; Rokach, 2010; Valentini & Masulli, 2002) and has also been transfered to the area of

unsupervised learning. Many studies discuss ensemble clustering (Ghosh & Acharya, 2011; Gionis, Mannila, & Tsaparas, 2007; Iam-On & Boongoen, 2015; Nguyen & Caruana, 2007; Strehl & Ghosh, 2002). For outlier detection, more methods than theory are available, although the idea has been already present almost half a century ago (Dempster & Gasko-Green, 1981; Gnanadesikan and Kettenring (1972)), when Gnanadesikan and Kettenring (1972) stated (p. 109):

> *"The complexity of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier protection procedure. A more reasonable approach seems to be to tailor detection procedures to protect against specific types of situations, e.g., correlation distortion, thus building up an arsenal of techniques with different sensitivities. This approach recognizes that an outlier for one purpose may not necessarily be one for another purpose! However, if several analyses are to be performed on the same sample, the result of selective segregation or outliers should be a more efficient and effective use of the available data."*

Primarily, this is an argument in favor of having various techniques instead of aiming to develop the one and only, one-size-fits-all technique, and to value different techniques for their diversity, which will also mean that they will perform differently well on different data.

The next step, however, would be to combine different techniques to an ensemble. In the data mining literature, ensemble techniques for outlier detection have been explicitly studied for more than a decade by now (Gao & Tan, 2006; Kirner et al., 2017; Kriegel et al., 2011; Lazarevic & Kumar, 2005; Liu, Ting, & Zhou, 2012; Nguyen, Ang, & Gopalkrishnan, 2010; Rayana & Akoglu, 2016; Schubert et al., 2012; Zhang et al., 2017; Zimek, Gaudet, Campello, & Sander, 2013). Only to a minor part, these studies were looking into combining actually different techniques (algorithms) for outlier detection, to a major part they were studying other sources of diversity such as different subspaces, subsamples, noise, approximations, parameters, or randomized procedures. Fundamental challenges and patterns have been discussed in position papers (Aggarwal, 2012; Zimek, Campello, & Sander, 2014).

If outlier scores are combined, some form of normalization is essential (Kriegel et al., 2011). However, some of these methods use the outlier scores explicitly to improve the construction of ensembles (Schubert et al., 2012). This relates to combining not only the prediction but also the confidence, as it has been recommended for ensemble classification (Ting & Witten, 1999).

8.3 | Statistical test on the score distribution

If the distribution of outlier scores is interpretable as an outlier probability distribution we are back to the statistical standard approach of applying a statistical test on the scores. This requires a good transformation of outlier scores to outlier probabilities, though, and the methods discussed above can only be seen as first steps in this direction.

8.4 | Explanation of outliers

Recently, there is a growing interest in methods for deriving *explanations* of outliers, that is, to give the users of some outlier detection method further aid in understanding and evaluating the result with respect to their domain.

Irrespective of whether it is being tackled as an add-on to some particular outlier detection method, as a by-product of some outlier detection method, or independently of any outlier detection method, the problem of *explaining outliers* has been treated with similar solutions in the literature under different names. The problem has been named, for example, "finding intensional knowledge" (Knorr & Ng, 1999), "detection of outlying subspaces" (Zhang, Lou, Ling, & Wang, 2004), "outlier explanation" (Micenková, Ng, Dang, & Assent, 2013; Paulheim & Meusel, 2015), "interpreting outliers" (Dang et al., 2014; Dang, Micenková, Assent, & Ng, 2013), "outlying aspect mining" (or "outlying aspect discovery"; Duan et al., 2015; Vinh et al., 2015, 2016), "outlying properties" (Angiulli, Fassetti, Manco, & Palopoli, 2017; Angiulli, Fassetti, & Palopoli, 2009), or "characterizations" (Angiulli, Fassetti, & Palopoli, 2013).

Although the terms "explanation," "interpretation," and "intensional knowledge" are more general, all approaches in the literature so far deliver essentially subspaces as explanations—be it subsets of attributes (i.e., axis-parallel subspaces) or combinations of attributes (i.e., arbitrarily-oriented subspaces), or error vectors (i.e., a weighted combination of attributes, where the weights are part of the "explanation"), where again all attributes or just some attributes could contribute to the error vector. This direction of research might also benefit from a renewed statistical interpretability of outlier scores and vice versa.

9 | SOFTWARE

Most outlier detection methods mentioned in Section 5 are implemented in the statistical software environment R (R Core Team, 2017) freely available at https://cran.r-project.org/, for example, in the add-on packages **robustbase** (Maechler et al., 2016), **rrcov** (Todorov & Filzmoser, 2009), or **mvoutlier** (Filzmoser & Gschwandtner, 2017).

More and more authors provide individual implementations of their own methods (such as some of the methods discussed in Sections 6-8). Some provide also implementations of competitors or test beds involving several methods. Prominent languages are Python (esp. scikit-learn; Pedregosa et al., 2011) and Java. GLOSH (Campello et al., 2015), for example, is provided by McInnes, Healy, and Astels (2017), as part of the HDBSCAN scikit-learn contribution. Some very popular algorithms are available in various implementations and languages that come effectively with different efficiency, see, for example, the overview and experimental comparison of LOF implementations by Kriegel, Schubert, and Zimek (2017).

The ELKI framework (Schubert et al., 2015; https://elki-project.github.io/) provides efficient (Kriegel et al., 2017) implementations for many outlier detection methods from a database background, including ensemble methods and meta-methods for the normalization of outlier scores. A list of algorithms implemented in ELKI is available at https://elki-project.github.io/algorithms/.

10 | CONCLUSIONS

In this survey, we gave an overview on statistical methods and on data mining methods for outlier detection. Based on reflections on the meaning of outlierness, we discussed their relationships and differences. We could summarize the literature that statistical methods are typically model-driven while data mining methods are typically algorithm-driven, focusing on efficiency. Model-driven methods are straightforward to evaluate using statistical tests but require the assumed model to fit the data which restricts the applicability if the data are not yet well understood. Algorithm-driven method development, oriented toward efficiency and applicability on large data sets and various data types, lost the connection with the original statistical notion and a probabilistic interpretability of their models and results. They are more flexible in their applicability. At the same time, however, this makes the evaluation of these methods challenging which, in turn, puts the usefulness of their broad applicability in doubt.

That much about the "there". Regarding the "...and back again" we reasoned on the interdisciplinary vision, and pointed out some first steps taken in the literature, to gain a renewed statistical notion of efficient data mining methods for outlier detection. We outlined potential lines of development based on such a reunion and on the resulting probabilistic interpretability of outlier detection results.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

Density-based clustering

ORCID


Arthur Zimek http://orcid.org/0000-0001-7713-4208

REFERENCES

Abe, N., Zadrozny, B., & Langford, J. (2006). Outlier detection by active learning. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 504–509). Philadelphia, PA.

Abraham, B., & Box, G. E. P. (1979). Bayesian analysis of some outlier problems in time series. *Biometrika*, *66*(2), 229–236.

Aggarwal, C. C. (2012). Outlier ensembles. *ACM SIGKDD Explorations*, *14*(2), 49–58.

Agostinelli, C. (2007). Robust estimation for circular data. *Computational Statistics & Data Analysis*, *51*(12), 5847–5866.

Agyemang, M., Barker, K., & Alhajj, R. (2006). A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, *10*, 521–538.

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph-based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, *29*(3), 626–688.

Akoglu, L., Tong, H., Vreeken, J., & Faloutsos, C. (2012). Fast and reliable anomaly detection in categorical data. In *Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM)* (pp. 415–424). Maui, HI.

Alqullaf, F., Van Aelst, S., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, *37*(1), 311–331.

Angiulli, F., & Fassetti, F. (2009). DOLPHIN: An efficient algorithm for mining distance-based outliers in very large datasets. *ACM Transactions on Knowledge Discovery from Data*, *3*(1), 1–57.

Angiulli, F., & Fassetti, F. (2010). Distance-based outlier queries in data streams: The novel task and algorithms. *Data Mining and Knowledge Discovery*, *20*(2), 290–324.

Angiulli, F., Fassetti, F., Manco, G., & Palopoli, L. (2017). Outlying property detection with numerical attributes. *Data Mining and Knowledge Discovery*, *31*(1), 134–163.

Angiulli, F., Fassetti, F., & Palopoli, L. (2009). Detecting outlying properties of exceptional objects. *ACM Transactions on Database Systems*, *34*(1), 1–62.

Angiulli, F., Fassetti, F., & Palopoli, L. (2013). Discovering characterizations of the behavior of anomalous subpopulations. *IEEE Transactions on Knowledge and Data Engineering*, *25*(6), 1280–1292.

Angiulli, F. & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)* (pp. 15–26). Helsinki, Finland.

Angiulli, F., & Pizzuti, C. (2005). Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, *17*(2), 203–215.

Anscombe, F. J., & Guttman, T. (1960). Rejection of outliers. *Technometrics*, *2*(2), 123–147.

Arning, A., Agrawal, R., & Raghavan, P. (1996). A linear method for deviation detection in large databases. In *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 164–169). Portland, OR.

Assent, I., Kranen, P., Baldauf, C., & Seidl, T. (2012). AnyOut: Anytime outlier detection on streaming data. In *Proceedings of the 17th International Conference on Database Systems for Advanced Applications (DASFAA)* (pp. 228–242). Busan, South Korea.

Barnett, V. (1978). The study of outliers: Purpose and model. *Applied Statistics*, *27*(3), 242–250.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester: John Wiley&Sons.

Barua, S., & Sander, J. (2014). Mining statistically significant co-location and segregation patterns. *IEEE Transactions on Knowledge and Data Engineering*, *26*(5), 1185–1199.

Bay, S. D. & Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 29–38). Washington, DC.

Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, *94*, 947–955.

Beckman, R. J., & Cook, R. D. (1983). Outlier.....s. *Technometrics*, *25*(2), 119–149.

Bernoulli, D. (1777). Diiudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda. *Acta Academiae Scientiarum Imperialis Petropolitanae*, 3–23.

Bernoulli, D., & Allen, C. G. (1961). The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. *Biometrika*, *48*(1–2), 3–18.

Bessel, F. W. (1838). *Gradmessung in Ostpreußen und ihre Verbindung mit Preußischen und Russischen Dreiecksketten*. Berlin: Königliche Akademie der Wissenschaften.

Bhaduri, K., Matthews, B. L., & Giannella, C. R. (2011). Algorithms for speeding up distance-based outlier detection. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 859–867). San Diego, CA.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)* (pp. 93–104). Dallas, TX.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion*, *6*, 5–20.

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(3), 231–237.

Campbell, N. A. (1982). Robust procedures in multivariate analysis II. Robust canonical variate analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *31*(1), 1–8.

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, *10*(1), 1–51.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., … Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, *30*, 891–927.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, *105*(489), 147–156.

Chakrabarti, S., Sarawagi, S, & Dom, B. (1998). Mining surprising patterns using temporal description length. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)* (pp. 606–617). New York City, NY.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58.

Chandola, V., Banerjee, A., & Kumar, V. (2012). Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *24*(5), 823–839.

Chawla, S. & Gionis, A. (2013). k-means--: A unified approach to clustering and outlier detection. In *Proceedings of the 13th SIAM International Conference on Data Mining (SDM)* (pp. 189–197). Austin, TX.

Chawla, S., & Sun, P. (2006). SLOM: A new measure for local spatial outliers. *Knowledge and Information Systems (KAIS)*, *9*(4), 412–429.

Chen, M., Gao, C., & Ren, Z. (2015). Robust covariance matrix estimation via matrix depth. *arXiv:150600691*.

Collett, D., & Lewis, T. (1976). The subjective nature of outlier rejection procedures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *25*(3), 228–237.

Craswell, N. (2009a). Precision at n. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 2127–2128). Boston, MA: Springer.

Craswell, N. (2009b). R-Precision. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (p. 2453). Boston, MA: Springer.

Cressie, N., & Read, T. R. C. (1988). Cressie-read statistic. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*, Supplementary Volume (pp. 37–39). New York: John Wiley & Sons.

Dang, X. H., Assent, I., Ng, R. T., Zimek, A., & Schubert, E. (2014). Discriminative features for identifying and interpreting outliers. In *Proceedings of the 30th International Conference on Data Engineering (ICDE)* (pp. 88–99). Chicago, IL.

Dang, X. H., Micenková, B., Assent, I., & Ng, R. (2013). Local outlier detection with interpretation. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)* (pp. 304–320). Prague, Czech Republic.

Das, K. & Schneider, J. G. (2007). Detecting anomalous records in categorical datasets. In *Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 220–229). San Jose, CA.

Dasgupta, D & Majumdar, N. S. (2002). Anomaly detection in multidimensional data using negative selection algorithm. In *Proceedings of the 2002 Congress on Evolutionary Computation (CEC)* (pp. 1039–1044). Honolulu, HI.

Dasgupta, D. & Nino, F. (2000). A comparison of negative and positive selection algorithms in novel pattern detection. In *Proceedings of the 2000 I.E. International Conference on Systems, Man, and Cybernetics (ICSMC)* (pp. 125–130). Nashville, TN.

de Vries, T., Chawla, S., & Houle, M. E. (2012). Density-preserving projections for large-scale local anomaly detection. *Knowledge and Information Systems*, *32*(1), 25–52.

Delannay, N., Archambeau, C., & Verleysen, M. (2008). Improving the robustness to outliers of mixtures of probabilistic PCAs. In *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* (pp. 527–535). Osaka, Japan.

Dempster, A. P., & Gasko-Green, M. (1981). New tools for residual analysis. *The Annals of Statistics*, *9*(5), 945–959.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems (MCS)* (pp. 1–15). Cagliari, Italy.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators.* (PhD thesis). Harvard University.

Duan, L., Tang, G., Pei, J., Bailey, J., Campbell, A., & Tang, C. (2015). Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery*, *29*(5), 1116–1151.

Emmott, A. F., Das, S., Dietterich, T., Fern, A., & Wong, W. K. (2013). Systematic construction of anomaly detection benchmarks from real data. In *Workshop on Outlier Detection and Description (ODD), held in conjunction with the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 16–21). Chicago, IL.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the 17th international conference on machine learning (ICML)* (pp. 255–262). Stanford, CA: Stanford University.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X.. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 226–231). Portland, OR.

Faloutsos, C. (2010). KDD innovation award talk. Retrieved from http://www.cs.cmu.edu/∼christos/TALKS/10-KDD-award/Faloutsos10IA.pdf

Faloutsos, C., & Megalooikonomou, V. (2007). On data mining, compression, and Kolmogorov complexity. *Data Mining and Knowledge Discovery*, *15*(1), 3–20.

Fan, H., Zaïane, O. R., Foss, A., & Wu, J. (2006). A nonparametric outlier detection for efficiently discovering top-N outliers from engineering data. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* (pp. 557–566). Singapore.

Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computer & Geosciences*, *31*, 579–587.

Filzmoser, P & Gschwandtner, M. (2017). mvoutlier: Multivariate outlier detection based on robust methods. R package version 2.0.8. Retrieved from https://CRAN.R-project.org/package=mvoutlier

Filzmoser, P., Maronna, R., & Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, *52*(3), 1694–1711.

Filzmoser, P., Ruiz-Gazen, A., & Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, *55*(1), 29–47.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *34*(3), 350–363.

Franke, C., Karnstedt, M., Klan, D., Gertz, M., Sattler, K. U., & Chervakova, E. (2009). In-network detection of anomaly regions in sensor networks with obstacles. *Computer Science – Research and Development*, *24*(3), 153–170.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y., & Han, J. (2010). On community outliers and their efficient detection in information networks. In *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 813–822). Washington, DC.

Gao, J. & Tan, P. N. (2006). Converting output scores from outlier detection algorithms into probability estimates. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)* (pp. 212–221). Hong Kong, China.

Ghosh, J., & Acharya, A. (2011). Cluster ensembles. *WIREs Data Mining and Knowledge Discovery*, *1*(4), 305–315.

Ghoting, A., Parthasarathy, S., & Otey, M. E. (2008). Fast mining of distance-based outliers in high-dimensional datasets. *Data Mining and Knowledge Discovery*, *16*(3), 349–364.

Gionis, A., Mannila, H., & Tsaparas, P. (2007). Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 1–30.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*(1), 81–124.

Hadi, A. S., Rahmatullah Imon, A. H. M., & Werner, M. (2009). Detection of outliers. *WIREs Computational Statistics*, *1*(1), 57–70.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Waltham, MA: Morgan Kaufmann.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Hardin, J., & Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, *44*(4), 625–638.

Hardin, J., & Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, *14*, 910–927.

Hautamäki, V., Kärkkäinen, I., & Fränti, P. (2004). Outlier detection using k-nearest neighbor graph. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)* (pp. 430–433). Cambridge, England.

Hawkins, D. (1980). *Identification of outliers*. London: Chapman and Hall.

Hayes, M. A., & Capretz, M. A. M. (2015). Contextual anomaly detection framework for big sensor data. *Journal of Big Data*, *2*, 2.

He, Z., Deng, S., Xu, X., & Huang, J. Z. (2006). A fast greedy algorithm for outlier mining. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* (pp. 567–576). Singapore.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2008). Inlier-based outlier detection via direct density ratio estimation. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)* (pp. 223–232). Pisa, Italy.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, *26*(2), 309–336.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*, 85–126.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Iam-On, N., & Boongoen, T. (2015). Comparative study of matrix refinement approaches for ensemble clustering. *Machine Learning*, *98*(1–2), 269–300.

Jagadish, H. V., Koudas, N., & Muthukrishnan, S. (1999). Mining deviants in a time series database. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB)* (pp. 102–113). Edinburgh, Scotland.

Janeja, V. P., Adam, N. R., Atluri, V., & Vaidya, J. (2010). Spatial neighborhood based anomaly detection in sensor datasets. *Data Mining and Knowledge Discovery*, *20*(2), 221–258.

Janssens, J., Huszár, F., Postma, E., & van den Herik, J. (2012). *Stochastic outlier selection*. Tilburg: Tilburg centre for Creative Computing.

Jiang, B. & Pei, J. (2011). Outlier detection on uncertain data: Objects, instances, and inferences. In *Proceedings of the 27th International Conference on Data Engineering (ICDE)* (pp. 422–433). Hannover, Germany.

Jin, W., Tung, A. K., & Han, J. (2001). Mining top-n local outliers in large databases. In *Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 293–298). San Francisco, CA.

Jin, W., Tung, A. K. H., Han, J., & Wang, W. (2006). Ranking outliers using symmetric neighborhood relationship. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* (pp. 577–593).

Johnson, T., Kwok, I., & Ng, R. (1998). Fast computation of 2-dimensional depth contours. In *Proceedings of the 4th ACM International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 224–228). New York City, NY.

Keller, F., Müller, E., & Böhm, K. (2012). HiCS: High contrast subspaces for density-based outlier ranking. In *Proceedings of the 28th International Conference on Data Engineering (ICDE)* (pp. 1037–1048). Washington, DC.

Kirner, E., Schubert, E., & Zimek, A. (2017). Good and bad neighborhood approximations for outlier detection ensembles. In *Proceedings of the 10th International Conference on Similarity Search and Applications (SISAP)* (pp. 173–187). Munich, Germany.

Knorr, E. M. & Ng, R. T. (1997). A unified notion of outliers: Properties and computation. In *Proceedings of the 3rd ACM International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 219–222). Newport Beach, CA.

Knorr, E. M. & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)* (pp. 392–403). New York City, NY.

Knorr, E. M. & Ng, R. T. (1999). Finding intensional knowledge of distance-based outliers. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB)* (pp. 211–222). Edinburgh, Scotland.

Knorr, E. M., Ng, R. T., & Tucanov, V. (2000). Distance-based outliers: Algorithms and applications. *The VLDB Journal*, *8*(3–4), 237–253.

Kollios, G., Gunopulos, D., Koudas, N., & Berchthold, S. (2003). Efficient biased sampling for approximate clustering and outlier detection in large datasets. *IEEE Transactions on Knowledge and Data Engineering*, *15*(5), 1170–1187.

Kontaki, M., Gounaris, A., Papadopoulos, A. N., Tsichlas, K., & Manolopoulos, Y. (2016). Efficient and flexible algorithms for monitoring distance-based outliers over data streams. *Information Systems*, *55*, 37–53.

Kou, Y., Lu, C. T., & Chen, D. (2006). Spatial weighted outlier detection. In *Proceedings of the 6th SIAM International Conference on Data Mining (SDM)* (pp. 614–618). Bethesda, MD.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2008). A general framework for increasing the robustness of PCA-based correlation clustering algorithms. In *Proceedings of the 20th International Conference on Scientific and Statistical Database Management (SSDBM)* (pp. 418–435). Hong Kong, China.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009a). LoOP: Local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)* (pp. 1649–1652). Hong Kong, China.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009b). Outlier detection in axis-parallel subspaces of high dimensional data. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* (pp. 831–838). Bangkok, Thailand.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2011). Interpreting and unifying outlier scores. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM)* (pp. 13–24). Mesa, AZ.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2012). Outlier detection in arbitrarily oriented subspaces. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM)* (pp. 379–388). Brussels, Belgium.

Kriegel, H. P., Schubert, E., & Zimek, A. (2017). The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, *52*(2), 341–378.

Kriegel, H. P., Schubert, M., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 444–452). Las Vegas, NV.

Kröger, P., & Renz, M. (2009). Multi-step query processing. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 1858–1862). Boston, MA: Springer.

Kruskal, W. H. (1960). Some remarks on wild observations. *Technometrics*, *2*(1), 1–3.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, *51*, 181–207.

Latecki, L. J, Lazarevic, A., & Pokrajac, D. (2007). Outlier detection with kernel density functions. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)* (pp. 61–75). Leipzig, Germany.

Lazarevic, A. & Kumar, V. (2005). Feature bagging for outlier detection. In *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 157–166). Chicago, IL.

Leach, M. J. V., Sparks, E. P., & Robertson, N. M. (2014). Contextual anomaly detection in crowded surveillance scenes. *Pattern Recognition Letters*, *44*, 71–79.

Lee, J. G., Han, J., & Li, X. (2008). Trajectory outlier detection: A partition-and-detect framework. In *Proceedings of the 24th International Conference on Data Engineering (ICDE)* (pp. 140–149). Cancun, Mexico.

Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(1), 31–36.

Liang, J. & Parthasarathy, S. (2016). Robust contextual outlier detection: Where context meets sparsity. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 2167–2172). Indianapolis, IN.

Liu, B., Xiao, Y., Cao, L., Hao, Z., & Deng, F. (2013). SVDD-based outlier detection on uncertain data. *Knowledge and Information Systems*, *34*(3), 597–618.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39.

Liu, X., Lu, C. T., & Chen, F. (2010). Spatial outlier detection: Random walk based approaches. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)* (pp. 370–379). San Jose, CA.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., & Cohen, K. L. (1999). Robust principal components for functional data. *TEST*, *8*, 1–73.

Lu, C. T., Chen, D., & Kou, Y. (2003). Algorithms for spatial outlier detection. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM)* (pp. 597–600). Melbourne, FL.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematics, Statistics, and Probabilistics: Vol. 1* (pp. 281–297). Berkeley, CA.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., … Anna di Palma, M. (2016). *Robustbase: basic robust statistics.* R package version 0.92-7. Retrieved from http://robustbase.r-forge.r-project.org/.

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. New York, NY: John Wiley & Sons.

Markou, M., & Singh, S. (2003a). Novelty detection: A review — Part 1: Statistical approaches. *Signal Processing*, *83*, 2481–2497.

Markou, M., & Singh, S. (2003b). Novelty detection: A review — Part 2: Neural network based approaches. *Signal Processing*, *83*, 2499–2521.

Maronna, R., Martin, D., & Yohai, V. (2006). *Robust statistics: Theory and methods*. Toronto, Canada: John Wiley & Sons Canada Ltd.

Maronna, R., & Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics*, *44*(4), 307–317.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, *4*(1), 51–67.

Marques, H. O., Campello, R. J. G. B., Zimek, A., & Sander, J. (2015). On the internal evaluation of unsupervised outlier detection. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management (SSDBM)* (pp. 7:1–12). San Diego, CA.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

Micenková, B., Ng, R. T., Dang, X. H., & Assent, I. (2013). Explaining outliers by subspace separability. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM)* (pp. 518–527). Dallas, TX.

Müller, E., Assent, I., Iglesias, P., Mülle, Y., & Böhm, K.. (2012). Outlier ranking via subspace analysis in multiple views of the data. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM)* (pp. 529–538). Brussels, Belgium.

Müller, E., Assent, I., Steinhausen, U., & Seidl, T. (2008). OutRank: Ranking outliers in high dimensional data. In *Proceedings of the 24th International Conference on Data Engineering (ICDE) Workshop on Ranking in Databases (DBRank)* (pp. 600–603). Cancun, Mexico.

Müller, E., Schiffer, M., & Seidl, T. (2010). Adaptive outlierness for subspace outlier ranking. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM)* (pp. 1629–1632). Toronto, Canada.

Müller, E., Schiffer, M., & Seidl, T. (2011). Statistical selection of relevant subspace projections for outlier ranking. In *Proceedings of the 27th International Conference on Data Engineering (ICDE)* (pp. 434–445). Hannover, Germany.

Nguyen, H. V., Ang, H. H., & Gopalkrishnan, V. (2010). Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA)* (pp. 368–383). Tsukuba, Japan.

Nguyen, H. V. & Gopalkrishnan, V. (2009). Efficient pruning schemes for distance-based outlier detection. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)* (pp. 160–175). Bled, Slovenia.

Nguyen, H. V., Gopalkrishnan, V., & Assent, I. (2011). An unbiased distance-based outlier detection approach for high-dimensional data. In *Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA)* (pp. 138–152). Hong Kong, China.

Nguyen, N. & Caruana, R. (2007). Consensus clusterings. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)* (pp. 607–612). Omaha, NE.

Öllerer, V., & Croux, C. (2015). Robust high-dimensional precision matrix estimation. In K. Nordhausen & S. Taskinen (Eds.), *Modern nonparametric* (pp. 325–350). Robust and multivariate methods, Heidelberg, Germany: Springer.

Orair, G. H., Teixeira, C., Wang, Y., Meira, W., Jr., & Parthasarathy, S. (2010). Distance-based outlier detection: Consolidation and renewed bearing. *Proceedings of the VLDB Endowment*, *3*(2), 1469–1480.

Otey, M. E., Ghoting, A., & Parthasarathy, S. (2006). Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery*, *12*(2–3), 203–228.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & Faloutsos, C. (2003). LOCI: Fast outlier detection using the local correlation integral. In *Proceedings of the 19th International Conference on Data Engineering (ICDE)* (pp. 315–326). Bangalore, India.

Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*, 3448–3470.

Paulheim, H., & Meusel, R. (2015). A decomposition of the outlier detection problem into a set of supervised learning problems. *Machine Learning*, *100*(2–3), 509–531.

Pearson, E. S., & Chandra Sekar, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, *28*(3/4), 308–320.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pei, Y., Zaïane, O., & Gao, Y. (2006). An efficient reference-based approach to outlier detection in large datasets. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)* (pp. 478–487). Hong Kong, China.

Peirce, B. (1852). Criterion for the rejection of doubtful observations. *The Astronomical Journal*, *2*(45), 161–163.

Peña, D., & Prieto, F. J. (2001). Cluster identification using projections. *Journal of the American Statistical Association*, *96*(456), 1433–1445.

Perozzi, B., Akoglu, L., Sánchez, P. I., & Müller, E. (2014). Focused clustering and outlier detection in large attributed graphs. In *Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 1346–1355). New York, NY.

Pham, N. & Pagh, R. (2012). A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 877–885). Beijing, China.

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: Classification of skewed data. *ACM SIGKDD Explorations*, *6*(1), 50–59.

Pimentel, M. A. F., Clifton, D. A., Clifton, L. A., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249.

Pokrajac, D., Lazarevic, A., & Latecki, L. J. (2007). Incremental local outlier detection for data streams. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 504–515). Honolulu, HI.

Popper, K. R. (1934). *Logik der Forschung. Zur Erkenntnistheorie der modernen naturwissenschaft*. Wien: Julius Springer.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson & Co.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radovanović, M., Nanopoulos, A., & Ivanović, M. (2014). Reverse nearest neighbors in unsupervised distance-based outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, *27*(5), 1369–1382.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)* (pp. 427–438). Dallas, TX.

Rayana, S., & Akoglu, L. (2016). Less is more: Building selective anomaly ensembles. *ACM Transactions on Knowledge Discovery from Data*, *10*(4), 42:1–42:33. https://doi.org/10.1145/2890508

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*, 1–39.

Rosner, B. (1975). On the detection of many outliers. *Technometrics*, *17*, 221–227.

Rousseeuw, P., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 73–79.

Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection*. New York, NY: Wiley-Interscience.

Rousseeuw, P. J. & Van den Bossche, W. (2016). Detecting deviating data cells. *ArXiv e-prints.*

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*(411), 633–639.

Ruts, I., & Rousseeuw, P. J. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis*, *23*, 153–168.

Sadik, M. S., & Gruenwald, L. (2013). Research issues in outlier detection for data streams. *ACM SIGKDD Explorations*, *15*(1), 33–40.

Sadoddin, R., Sander, J., & Rafiei, D. (2016). Finding surprisingly frequent patterns of variable lengths in sequence data. In *Proceedings of the 16th SIAM International Conference on Data Mining (SDM)* (pp. 27–35). Miami, FL.

Sánchez, P. I., Müller, E., Irmler, O., & Böhm, K. (2014). Local context selection for outlier ranking in graphs with multiple numeric node attributes. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management (SSDBM)* (pp. 16:1–16:12). Aalborg, Denmark.

Sánchez, P. I., Müller, E., Laforet, F., Keller, F., & Böhm, K. (2013). Statistical selection of congruent subspaces for mining attributed graphs. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM)* (pp. 647–656). Dallas, TX.

Sarawagi, S., Agrawal, R., & Megiddo, N. (1998). Discovery-driven exploration of OLAP data cubes. In *Proceedings of the 6th International Conference on Extending Database Technology (EDBT)* (pp. 168–182). Valencia, Spain.

Schubert, E. (2013). *Generalized and efficient outlier detection for spatial, temporal, and high-dimensional data mining.* (PhD thesis). Ludwig-Maximilians-Universität München, Munich, Germany.

Schubert, E., Koos, A., Emrich, T., Züfle, A., Schmid, K. A., & Zimek, A. (2015). A framework for clustering uncertain data. *Proceedings of the VLDB Endowment*, *8*(12), 1976–1979.

Schubert, E., Weiler, M., & Kriegel, H. P. (2014). SigniTrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds. In *Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 871–880). New York, NY.

Schubert, E., Weiler, M., & Kriegel, H. P. (2016). SPOTHOT: Scalable detection of geo-spatial events in large textual streams. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management (SSDBM)* (pp. 8:1–8:12).

Schubert, E., Wojdanowski, R., Zimek, A., & Kriegel, H. P. (2012). On evaluation of outlier rankings and outlier scores. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM)* (pp. 1047–1058). Anaheim, CA.

Schubert, E., Zimek, A., & Kriegel, H. P. (2014a). Generalized outlier detection with flexible kernel density estimates. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM)* (pp. 542–550). Philadelphia, PA.

Schubert, E., Zimek, A., & Kriegel, H. P. (2014b). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, *28*(1), 190–237.

Scott, D. W. (2008). *Multivariate density estimation: Theory, practice, and visualization*. Hoboken, NJ: John Wiley & Sons.

Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C, & Leskovec, J. (2008). Mobile call graphs: Beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 596–604). Las Vegas, NV.

Shekhar, S., Lu, C. T., & Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, *7*(2), 139–166.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Smets, K. & Vreeken, J. (2011). The odd one out: Identifying and characterising anomalies. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM)* (pp. 804–815). Mesa, AZ.

Song, X., Wu, M., Jermaine, C. M., & Ranka, S. (2007). Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, *19*(5), 631–645.

Stahel, W. A. (1981). *Breakdown of covariance estimators*. E.T.H. Zürich, Switzerland: Fachgruppe für Statistik.

Steinwart, I., Hush, D., & Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, *6*, 211–232.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.

Su, X., & Tsai, C. L. (2011). Outlier Detection. *WIREs Data Mining and Knowledge Discovery*, *1*(3), 261–268.

Sugiyama, M., Borgwardt, K. M. (2013). Rapid distance-based outlier detection via sampling. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS)* (pp. 467–475). Lake Tahoe, NV.

Sun, P. & Chawla, S. (2004). On local spatial outliers. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM)* (pp. 209–216). Brighton, England.

Swersky, L., Marques, H. O., Sander, J., Campello, R. J. G. B., & Zimek, A. (2016). On the evaluation of outlier detection and one-class classification methods. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1–10). Montreal, Canada.

Takeuchi, J., & Yamanishi, K. (2006). A unifying framework for detecting outliers and change points from time series. *IEEE Transactions on Knowledge and Data Engineering*, *18*(4), 482–492.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Addison Wesley.

Tang, G., Pei, J., Bailey, J., & Dong, G. (2015). Mining multidimensional contextual outliers from categorical relational data. *Intelligent Data Analysis*, *19*(5), 1171–1192.

Tang, J., Chen, Z., Fu, A. W. C., & Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. In *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* (pp. 535–548). Taipei, Taiwan.

Tarr, G., Müller, S., & Weber, N. C. (2016). Robust estimation of precision matrices under contamination. *Computational Statistics & Data Analysis*, *93*, 404–420.

Tax, D. M. J., & Duin, R. P. W. (2004). Support vector data description. *Machine Learning*, *54*(1), 45–66.

Thompson, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics*, *6*(4), 214–219.

Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research*, *10*, 271–289.

Todorov, V., & Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, *32*(3), 1–47.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, *7*, 1–20.

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Valentini, G. & Masulli, F. (2002). Ensembles of learning machines. In *Proceedings of the 13th Italian Workshop on Neural Nets* (pp. 3–22). Vietri, Italy.

Vinh, N. X., Chan, J., Bailey, J., Leckie, C., Ramamohanarao, K., & Pei, J. (2015). Scalable outlying-inlying aspects discovery via feature ranking. In *Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* (pp. 422–434). Ho Chi Minh City, Vietnam.

Vinh, N. X., Chan, J., Romano, S., Bailey, J., Leckie, C., Ramamohanarao, K., & Pei, J. (2016). Discovering outlying aspects in large datasets. *Data Mining and Knowledge Discovery*, *30*(6), 1520–1555.

Wang, X. & Davidson, I. (2009). Discovering contexts and contextual outliers using random walks in graphs. In *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM)* (pp. 1034–1039). Miami, FL.

Wang, Y., Parthasarathy, S., & Tatikonda, S. (2011). Locality sensitive outlier detection: A ranking driven approach. In *Proceedings of the 27th International Conference on Data Engineering (ICDE)* (pp. 410–421). Hannover, Germany.

Warrender, C., Forrest, S., & Pearlmutter, B. (1999). Detecting intrusions using system calls: Alternative data models. In *Proceedings of the 1999 I.E. Symposium on Security and Privacy* (pp. 133–145). Oakland, CA.

Yamanishi, K., Takeuchi, J. I., Williams, G., & Milne, P. (2000). On-line unsupervised outlier detection using finite mixture with discounting learning algorithms. In *Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 320–324). Boston, MA.

Yamanishi, K., Takeuchi, J. I., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixture with discounting learning algorithms. *Data Mining and Knowledge Discovery*, *8*, 275–300.

Yu, J. X., Qian, W., Lu, H., & Zhou, A. (2006). Finding centric local outliers in categorical/numerical spaces. *Knowledge and Information Systems*, *9*(3), 309–338.

Zhang, E., & Zhang, Y. (2009). Average precision. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 192–193). Boston, MA: Springer.

Zhang, J., Lou, M., Ling, T. W., & Wang, H. (2004). HOS-Miner: A system for detecting outlying subspaces of high-dimensional data. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)* (pp. 1265–1268). Toronto, Canada.

Zhang, X., Dou, W. C., He, Q., Zhou, R., Leckie, C., Ramamohanarao, K., et al. (2017). LSHiForest: A generic framework for fast tree isolation based ensemble anomaly analysis. In *Proceedings of the 33rd International Conference on Data Engineering (ICDE)* (pp. 983–994). San Diego, CA.

Zhu, C., Kitagawa, H., & Faloutsos, C. (2005). Example-based robust outlier detection in high dimensional datasets. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)* (pp. 829–832). Houston, TX.

Zimek, A., Campello, R. J. G. B., & Sander, J. (2013). Ensembles for unsupervised outlier detection: Challenges and research questions. *ACM SIGKDD Explorations*, *15*(1), 11–22.

Zimek, A., Campello, R. J. G. B., & Sander, J. (2014). Data perturbation for outlier detection ensembles. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management (SSDBM)* (pp. 13:1–13:12). Aalborg, Denmark.

Zimek, A., Gaudet, M., Campello, R. J. G. B., & Sander, J. (2013). Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 428–436). Chicago, IL.

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, *5*(5), 363–387.

Zimek, A., & Vreeken, J. (2015). The blind men and the elephant: On meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Machine Learning*, *98*(1–2), 121–155.

How to cite this article: Zimek A, Filzmoser P. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *WIREs Data Mining Knowl Discov*. 2018;e1280. https://doi.org/10.1002/widm.1280