Meta-survey on outlier and anomaly detection

Madalina Olteanu [a], Fabrice Rossi [a,*], Florian Yger [b]

[a] *CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, Place du Maréchal de Lattre de Tassigny, 75016, Paris, France*
[b] *LAMSADE, CNRS, UMR 7243, Université Paris-Dauphine, PSL University, Place du Maréchal de Lattre de Tassigny, 75016, Paris, France*

ARTICLE INFO

ABSTRACT

The impact of outliers and anomalies on model estimation and data processing is of paramount importance, as evidenced by the extensive body of research spanning various fields over several decades: thousands of research papers have been published on the subject. As a consequence, numerous reviews, surveys, and textbooks have sought to summarize the existing literature, encompassing a wide range of methods from both the statistical and data mining communities. While these endeavors to organize and summarize the research are invaluable, they face inherent challenges due to the pervasive nature of outliers and anomalies in all data-intensive applications, irrespective of the specific application field or scientific discipline. As a result, the resulting collection of papers remains voluminous and somewhat heterogeneous.

To address the need for knowledge organization in this domain, this paper implements the first systematic meta-survey of general surveys and reviews on outlier and anomaly detection. Employing a classical systematic survey approach, the study collects nearly 500 papers using two specialized scientific search engines. From this comprehensive collection, a subset of 56 papers that claim to be general surveys on outlier detection is selected using a snowball search technique to enhance field coverage. A meticulous quality assessment phase further refines the selection to a subset of 25 high-quality general surveys.

Using this curated collection, the paper investigates the evolution of the outlier detection field over a 20-year period, revealing emerging themes and methods. Furthermore, an analysis of the surveys sheds light on the survey writing practices adopted by scholars from different communities who have contributed to this field.

Finally, the paper delves into several topics where consensus has emerged from the literature. These include taxonomies of outlier types, challenges posed by high-dimensional data, the importance of anomaly scores, the impact of learning conditions, difficulties in benchmarking, and the significance of neural networks. Non-consensual aspects are also discussed, particularly the distinction between local and global outliers and the challenges in organizing detection methods into meaningful taxonomies.

1. Introduction

Real world data-intensive applications are susceptible to the detrimental effects of noise, outliers and anomalies, which are commonly observed in large-scale data sets. Anomalies and outliers are generally defined as observations that deviate in an important way from other observations, a rather vague and flexible definition. Despite being the subject of study since the early stages of modern statistics (see e.g. [1]), effectively dealing with them remains a persistent challenge.

Over the course of several decades, thousands of research papers have been published on those topics, alongside numerous surveys, review papers and textbooks. Monographs focusing on statistical treatments of outliers can be traced back at least to Hawkings' seminal book

[2] while survey articles have spanned almost two decades, with some early works appearing in 2003 [3]. The richness of this survey literature has two significant implications.

Firstly, it poses difficulties for researchers in selecting relevant surveys to read and gauging their content. For instance, questions may arise regarding the continued relevance of a paper published two decades ago, as well as the expected background knowledge required when perusing a recent survey.

Secondly, this vast literature provides a unique perspective on the historical evolution of the outlier concept and the diverse approaches adopted by researchers from various communities, including database specialists, computer engineers, statisticians, machine learning experts, and more.

* Corresponding author.
E-mail addresses: Madalina.Olteanu@dauphine.psl.eu (M. Olteanu), Fabrice.Rossi@dauphine.psl.eu (F. Rossi), Florian.Yger@dauphine.psl.eu (F. Yger).

The objective of this paper is to analyze research papers that claim to be general reviews on anomaly and outlier detection. By doing so, it aims to address the questions raised by the extensive and diverse literature in this field, offering recommendations for reading and providing a historical perspective. To achieve this goal, the paper undertakes the first systematic meta-survey on anomaly and outlier detection, specifically focusing on survey papers rather than standard research papers (hence the *meta* aspect of the survey). To ensure the avoidance of selection bias, a classical systematic survey protocol, as outlined in [4], is followed.

Based on pilot study presented at the ESANN conference in 2022 [5], the meta-survey begins with a paper collection phase. Two specialized search engines are utilized to identify a substantial collection of nearly 500 papers related to outlier and anomaly detection. Through a manual analysis, a subset of 56 papers is carefully selected based on a strict definition of a "general survey on outlier and anomaly detection". Subsequently, a quality assessment is conducted, uncovering unexpected instances of plagiarism and leading to the refinement of the collection to 25 high-quality papers.

As an additional outcome of the paper collection process, a brief discussion is provided on specialized surveys. This discussion primarily focuses on the types of methods, application fields, and other factors that have been deemed sufficiently important or active to warrant the effort of reviewing corresponding sections of the literature.

Armed with this exhaustive collection of surveys, we can address several research questions. The first set of inquiries pertains to the field as a whole and encompasses historical aspects, methodological considerations, and paper structure. Specifically, we aim to assess the extent to which surveys are conducted in a systematic manner and how their findings are organized. We also delve into the interplay between surveys, examining aspects such as integration (e.g., citation of previous surveys) and the emergence of communities, both in terms of vocabulary and citations. Of particular interest is the temporal component of this integration, shedding light on how the field has evolved over time.

A second series of questions can be elucidated by examining in more details the content of the surveys themselves. We begin by discussing a selection of consensual topics, emphasizing the process of consolidating knowledge within these areas. For example, we explore how the definition and categorization of outliers have undergone increasing refinement over time. Additionally, we investigate the enduring presence of artificial neural networks in the field, particularly the transition from "shallow models" to the advent of "deep learning". The final research focus centers on non-consensual aspects, exploring topics where the literature presents divergent viewpoints on common issues.

By addressing these research questions, we aim to provide insights into the field of anomaly and outlier detection, shedding light on its historical development, knowledge consolidation, and areas of disagreement.

The paper is organized as follows. Section 2 recalls some important concepts about outliers and anomalies, in a historical perspective. It provides a background for the systematic meta-survey. Section 3 describes in details the meta-survey methodology and its implementation. Section 4 provides a global high level analysis of the selected surveys. Section 5 discusses the main consensual findings that can be gathered in from the surveys. Finally Section 6 is dedicated to the debated aspects for which different conflicting visions can be identified in the literature.

2. Outliers and anomalies

We discuss in this section three important aspects of the literature on outlier detection, focusing on historical aspects as well as on the Aggarwal's monograph [6]. This provides an important context for Sections 5 and 6. We discuss first general definitions of outliers and anomalies (Section 2.1). Then we explain the emphasis of classical statistical analysis on outlier removal and robust statistics (Section 2.2).

Finally, we present the idea that data models are the substance of outlier detection, even when they are implicit (Section 2.3). Notice that this Section is partially informed by the pilot study conducted in [5] and by the publications collected as part of the meta-survey, see Section 3 for details.

2.1. Elements of style

Historically, anomalies and outliers have been defined in plain English using arguably vague sentences such as

- an anomaly "*appears to deviate markedly from other members of the sample in which it occurs*" [7];
- 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*" [2];
- an outlier is "*an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*" [8].

Those informal definitions were used as guiding principles to build mathematically sound and operational definitions. Indeed the ground work on outlier detection was conducted by statisticians and a natural translation of the above principles involves hypotheses on the distribution of the data.

For instance Hawkins distinguishes in [2] two types of outliers: (i) either coming from the extreme cases of an heavy tailed distribution or (ii) from a "contaminant" distribution in a two distributions generating process, where one main distribution generates the "good" observations. A similar distinction is proposed in [9] with different names: the discordant observation is either a "natural variation" (as in (i)) and the model has weaknesses which corresponds to type (ii).

While it is quite common in the literature to use interchangeably anomaly and outlier (and in some papers, novelty), the models discussed above hint at a possible distinction. In the first case, an outlier is compatible with the chosen data distribution, albeit very uncommon. In the second case, outliers obey to a different distribution than the normal instances. While they still may be generated by the normal distribution, those outliers are in some sense "more abnormal" than in the first case. Thus following, e.g. Aggarwal in [6], we may use the term outliers for large deviations and anomalies for extreme ones (we can even distinguishing weak outliers from strong outliers, if we want to distinguish between rare and very rare cases).

It should also be noted that the distinction is somewhat historical. Older text books such as [2,8,10,11] use almost exclusively the term outlier (for instance [8] has approximately 3300 occurrences of outliers versus 47 uses of anomaly). [6] is more balanced with roughly 4 times more use of outliers than anomalies, but some recent surveys such as [12–14] use an inverse balance.

2.2. Estimation under contamination

In the pioneering statistical literature on outliers, the focus is on *outlier treatment* [15], with a distinction between *deterministic* outliers and *random* ones. The first case corresponds to errors in a broad sense. In this case, the goal is to detect the outlier/error to fix it or reject it. This is classically done with test of hypothesis.

The second case is more general and corresponds to situations where the offending observations cannot be explained by some mechanism (and hence be considered as an error). As in the first case, tests of hypothesis can be used to determine whether an observation is an outlier. The outcome includes rejecting the observation, modifying the assumed data distribution model to make the observation more plausible, or accommodating outliers in the model.

The common point between the two cases is the need for a model of normal behavior that is used to build the test, among other things. The core difficulty is that excepted in very specific cases, the normal behavior is at best partially known. In particular, under a parametric

model assumption, parameters must be estimated (generally by maximum likelihood). However, as pointed out in [16], we are facing a "chicken-and-egg" problem: if we knew the true model, it would be easy to identify outliers, while if we knew which observations have been generated by the true model (and thus are not outliers), it would be easy to estimate the parameters of the model (or more generally to assess the quality of our hypotheses on the data generation model).

This explains why outlier detection has been deeply linked to *robust statistics*. The goal of this field of statistics is to derive estimation procedures that are robust to arbitrary errors in the data [10,17]. Examples include robust estimations of the covariance matrix of multivariate data [18] and robust regression [11]. We refer the reader to Rousseeuw and Hubert recent survey on outlier treatment with robust methods [19] for details. Notice that even very basic statistical estimators can break when confronted with a small percentage of outliers. This is the case of mean and of the standard deviation which are used in a routine way as a preprocessing step to a lot of methods. There is therefore a lot of value in using robust approaches even when not directly considering the case of outlier detection and mitigation.

The emphasis of early approaches to outlier detection on outlier treatments can be seen as a manifestation of one of the two cultures discussed by Breiman in [20]. The goal is here to fit properly a model to a data set in order to analyze the model and interpret it. If a linear regression model is adequate, the values of its coefficients provide insights on the influence of the explanatory variables on the target variable. In this context, outliers are a nuisance. We just want to make sure they do not constitute a proof of inadequacy of the model to the real world and that they do not break our estimation procedures. In this end, we just want to get rid of them! However, in the more general data science context, the way outliers are handled can be quite different. For instance in intrusion detection [21–24], the normal behavior model is not interesting *per se* and serves only as a detector of intrusions. While the problem of robustness is still present, the detection performances are more important that the consistency of parameter estimates, for instance.

2.3. Models are everywhere

As pointed out in [6] one can summarize the whole field of anomaly detection as follows:

Virtually all outlier detection algorithms create a model of the normal patterns in the data, and then compute an outlier score of a given data point on the basis of the deviations from these patterns.

This is further summarized by the sentence: "*the data model is everything*" [6]. According to this view, the discussion above about the "chicken-and-egg" problem [16] of anomaly detection applies to the whole field, even to methods that do not have an obvious statistical interpretation. In general, the data model will be adjusted to the observations *blindly*, that is without knowing in advance whether a given observation is normal or not. Thus model fitting must be somehow "robust" to the presence of outliers (not necessarily in the robust statistics sense).

In addition, in the mixture/dual distribution point of view, when one can hypothesize a model both for the normal data and for the outliers, the problem of assigning observations to one of the two models is plagued by the imbalanced nature of the data [25–27]. By essence anomalies are rare and thus only simple models can be adjusted to them.

3. Methodology

This section presents the methodology used to select relevant surveys to include in the present work. This methodology is based on general principles of systematic reviews, see e.g. [4] and on a pilot study conducted for the ESANN 2023 conference [5].

The paper selection has been conducted using two specialized search engines, Google Scholar[1] (GS) and Semantic Scholar[2] (S2). Google Scholar does not provide an API and forbids the use of bots via its robot.txt file. As such, we use manually the site under the private mode of the Firefox browser[3] to avoid potential results tailoring. Semantic scholar was used via its dedicated API. The default ordering of both search engines was used.

3.1. Meta survey scope

We are interested in *general* surveys about outlier and anomaly detection. A publication is considered relevant if it fulfills the following conditions:

1. it must be written primarily in English (an abstract in another language does not prevent the inclusion into the meta survey);
2. it must discuss a significant number of prior works on anomaly detection in an organized way;
3. it must be a peer reviewed article published in a journal, in a collection book or in conference proceedings. This excludes explicitly submitted papers, technical reports and student oriented workshops. This also excludes monographs, text books and tutorial. They are used as a way to provide context for the meta-survey (some technical reports are also considered for this task);
4. it must be general, that is it should not target only a specific subset of the scientific literature based on restrictions such as: application field (e.g. computer security), classes of methods (e.g. deep learning methods), data types (e.g. anomalies in graphs) or learning conditions (e.g. streaming data). We made an exception by considering that categorical data and high dimensional data are so pervasive that surveys considering only this type of data are general enough to be included.

3.2. Initial queries

The first paper selection was made using simplistic queries to cover a large spectrum of papers, favoring recall over precision. We used the four possible combinations of *outlier* or *anomaly* with *survey* or *review*. This choice was based on the results of the pilot study conducted in [5]. During this study, we observed firstly that historical works as well as textbooks (e.g. [6]) systematically use the words outlier or anomaly, or both, to describe the subject of interest. In addition we observed that all the general surveys we were aware of appeared at the top of the results of those queries. Those queries were therefore considered to be general enough to have a high recall. This is confirmed by the significant intersection between the results sets (see below) as well as the limited impact of the snowball search (see Section 3.3).

On each search engine, we selected the first 100 papers for each query. On S2, the four queries reported a total of 374 distinct paper ids (some articles may have multiple ids: we found 5 duplicated papers). On GS, we obtained 266 distinct paper "clusters" (a cluster contains several paper descriptions that refer to the same paper, but we still identified a duplicated paper). We identified 141 common papers and a combined collection of 492 unique papers (over a total of 800 initial results).

We read the title, abstract and in some cases the full paper in order to classify the 492 candidate documents into the following six classes:

---

[1] https://scholar.google.com
[2] https://www.semanticscholar.org/
[3] https://www.mozilla.org/en-US/firefox/new/

Table 1

Classification of the 492 papers identified by S2 and GS queries into the six classes defined in the main text (those figures concern the first phase only and do not include the papers obtained by the snowball search).

Type	Class	Number of papers
Excluded	C1	12
Not on outliers	C2	184
Text book or tutorial	C3	10
Not a survey	C4	69
Not a general survey	C5	170
General survey	C6	47

Table 2

Statistics on the earliest surveys discussed in the present paper: number of reference, number of words in the text (excluding references), total number of pages (including references).

Paper	Number of references	Number of words (text)	Number of pages
[34]	91	15,455	23
[3]	64	11,129	17
[33]	28	7,843	10
[32]	66	14,827	45
[35]	67	5,089	16
[36]	80	8,743	18
[29]	361	23,478	58
[16]	69	8,940	14

1. excluded documents based on "technical aspects": non English papers (4), non peer-reviewed documents (2), preprints (3), non existent documents (2) and a blatant case of plagiarism;
2. excluded documents based on their content: papers whose main subject is not outlier detection;
3. excluded documents based on their nature: monographs, text books and tutorial about outlier detection;
4. papers about outlier detection but that are not surveys (e.g. description of new methods, applications, etc.);
5. survey papers about specific aspects of outlier detection;
6. general survey papers about outlier detection (as claimed by the authors of the papers).

The breakdown of the 492 papers into the six classes is given by Table 1.[4]

3.3. Snowball search

In order to extend the coverage of our selection, we used a classical snowball search approach: references of 47 papers selected in the first phase were analyzed to find other general survey papers (class C6). As outlier detection is a long tradition in statistics, some papers and books were published before 2000. We decided to sort those publications in a historical group that was used to write Section 2: this gave us some historical background and a mean to discuss the temporal evolution of the way outliers are considered (only in a qualitative way as we cannot claim exhaustivity for those older publications). Finally, we also included text books and monographs from those references (class C3). After the snowball process, the class C6 consists in 56 papers: only 9 general survey papers were found during this part of the search process, which confirms the large coverage induced by the simplicity of the search queries.

3.4. Quality assessment

The quality of the papers selected during the queries and the snowball search is very uneven. There are in particular two major sources of quality problems. Firstly, plagiarism is non negligible: 13 papers among the 56 selected ones show different levels of plagiarism as detailed in Section 3.4.1. Secondly some papers are simply too short and cover a too small selection of papers to bring new insights on the field, as explained in Section 3.4.2.

3.4.1. Plagiarism

To assess plagiarism, we compared figures between papers and used in addition the SPECTRE embedding [28] provided by S2. This embedding assigns to each paper indexed on S2 a vector representation in dimension 768 based on its title, abstract and references. The embedding of a paper missing from S2 was computed using the pre-trained model provided by the authors. We also computed embeddings

for all papers based on their full content rather than only the abstract (leveraging again the pre-trained model). For each paper, we computed the five closest earlier papers in class C6 and in a selection of highly cited papers not in C6, using both embeddings. Then we compared in details each paper with its possible inspiration sources.

We found multiple use of figures from previous papers without proper credits as will as verbatim or almost verbatim use of their content. Overall, thirteen papers were found to exhibit various forms of plagiarism.

More precisely, we identified three papers published after 2009 that include figures "borrowed" without credit from the most cited survey paper to date [29]. In addition two papers were found to be obvious plagiarisms of [29] as they copied not only figures, but also most of the text with minimal editing. We spotted one instance of self-plagiarism as well as strong resemblance between two papers of different authors. Many papers include lengthy "quotes" from [29] with no proper attribution. We found also an obvious plagiarism of a lesser known survey [30], with subsequent articles borrowing both from [29,30]. [31] is another example of a source of tables and images copied without proper reference (Notice that [31] is an unpublished technical report and hence is excluded from C6). Some papers are also borrowing texts and images from the second most cited survey paper to date [32].

3.4.2. Contribution

Finally, we assessed the contribution of each paper to the state-of-the-art while taking into account the publication date. We consider that a general survey paper contributes to the state-of-the-art (SOTA) if it fulfills at least one of the following conditions:

1. it discusses recent papers, published after the previously published surveys or missed by them;
2. it addresses an important general problem of outlier detection such as anomaly categories or the rising importance of deep learning;
3. it provides a new point of view on the literature, e.g. by introducing a new taxonomy of methods or by analyzing existing methods with respect to e.g. their scalability.

To our surprise, a lot of the papers do not position themselves with respect to previous surveys, apart from citing some of them. Among 57 papers, only 14 explain explicitly their contribution compared to previous surveys.

Many papers include also a rather small selection of papers, most of them being already mentioned in previous surveys. As a reference, Table 2 shows some statistics on the earliest general surveys found in our search (prior 2010). Apart [33], all papers discuss more than 60 papers with a peak at 361 references in [29] (which was published in 2009). Those numbers set the bar quite high in terms of literature coverage and positioning for papers published after 2009.

---

[4] The list of the papers with the chosen classes are available here https://github.com/fabrice-rossi/outlier-anomaly-detection.

Table 3

Paper	Year	Citations per year	Title
MarkouSingh2003Novelty DetectionNeural [34]	2003	55.45	Novelty detection: a review—part 2: neural network based approaches
MarkouSingh2003Novelty DetectionStatistical [3]	2003	95.90	Novelty detection: a review—part 1: statistical approaches
Petrovskiy2003 OutlierDetection [33]	2003	6.75	Outlier detection algorithms in data mining systems
HodgeAustin2004 SurveyOutlier [32]	2004	227.05	A survey of outlier detection methodologies
BenGal2005 OutlierDetection [35]	2005	2.39	Outlier Detection
AgyemangBarkerEtAl2006 ComprehensiveSurvey [36]	2006	13.82	A comprehensive survey of numeric and symbolic outlier mining techniques
ChandolaBanerjeeEtAl2009AnomalyDetection [29]	2009	868.64	Anomaly detection: A survey
HadiImon2009DectectionOutliers [16]	2009	12.50	Detection of outliers
SuTsai2011OutlierDetection [38]	2011	4.83	Outlier detection
ZimekSchubertEtAl2012SurveyUnsupervised [39]	2012	77.00	A survey on unsupervised outlier detection in high-dimensional numerical data
AguinisGottfredsonEtAl2013BestPractice [40]	2013	116.50	Best-practice recommendations for defining, identifying, and handling outliers
Zhang2013AdvancementsOutlier [41]	2013	17.90	Advancements of outlier detection: A survey
PimentelCliftonEtAl2014ReviewNovelty [42]	2014	175.56	A review of novelty detection
ZimekFilzmoser2018ThereBack [43]	2018	28.80	There and back again: Outlier detection between statistical reasoning and data mining algorithms
Rokhman2019SurveyMixed [44]	2019	1.25	A survey on mixed-attribute outlier detection methods
TahaHadi2019AnomalyDetection [45]	2019	16.50	Anomaly detection methods for categorical data: A review
WangBahEtAl2019ProgressOutlier [46]	2019	70.25	Progress in outlier detection techniques: A survey
BoukercheZhengEtAl2020OutlierDetection [47]	2020	49.33	Outlier detection: Methods, models, and classification
CarrenoInzaEtAl2020AnalyzingRare [48]	2020	17.33	Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework
ThudumuBranchEtAl2020ComprehensiveSurvey [49]	2020	44.00	A comprehensive survey of anomaly detection techniques for high dimensional big data
Foorthuis2021NatureTypes [14]	2021	15.50	On the nature and types of anomalies: A review of deviations in data
NassifTalibEtAl2021MachineLearning [50]	2021	28.50	Machine learning for anomaly detection: A systematic review
RuffKauffmannEtAl2021UnifyingReview [13]	2021	192.50	A unifying review of deep and shallow anomaly detection
SamariyaThakkar2021ComprehensiveSurvey [12]	2021	5.50	A comprehensive survey of anomaly detection algorithms
SalehiMirzaeiEtAl2022UnifiedSurvey [51]	2022	42.00	A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges

Considering all the criteria described above, we identified only 25 papers[5] (among the remaining 44 papers that do not involve plagiarism) as contributing to the state-of-the-art (see the full list in Table 3). One paper [37] was excluded as it is written for a specific research community (medical research): it makes underlying assumptions about the data production process which limits strongly their generality. Its inclusion in class C6 is even debatable. Fig. 1 shows the number of papers published per year broken down into three categories.

While our sorting process remains expert based and partially questionable, it aligns relatively nicely with simple numerical characteristics of the papers. We represented each paper as a low dimensional vector using the following characteristics:

- size: total length (total number of characters), total text length (total number of characters excluding the references), number of pages;
- literature coverage: number of references and delay in years between the publication year of the survey and the publication year of the most recent paper cited;

- citations: the logarithm of the average number of citations per year of the paper since its publication according to Google Scholar[6];
- plagiarism: the Euclidean distance between the embedding of the survey and the embedding of its closest neighbor in the earlier papers, both using the abstract based embedding and using the full paper embedding.

A simple Principal Component Analysis can be used to capture roughly 73% of the variance of the data, as shown in the Scree plot on Fig. 2. Fig. 3 shows the projected papers: most of the papers that contributed to the SOTA have a high positive value on the first principal axis, while minor papers and papers with plagiarism have negative values on the same axis. The second principal axis can be used to separate partial papers with or without plagiarism. Fig. 4 shows the contribution of the variables to those axes. The first PC is mostly explained by size effects, including the number of citations received by the paper, but also by the freshness of the references. The second PC is more related to the proximity to other papers. We provide in Appendix A additional representations of the PCA results in order to explore the third component (which capture 10.8% of the variance). They confirm the interpretation derived from the first two components.

---

[5] A summary of our evaluation of the 56 papers in C6 is available here https://github.com/fabrice-rossi/outlier-anomaly-detection.

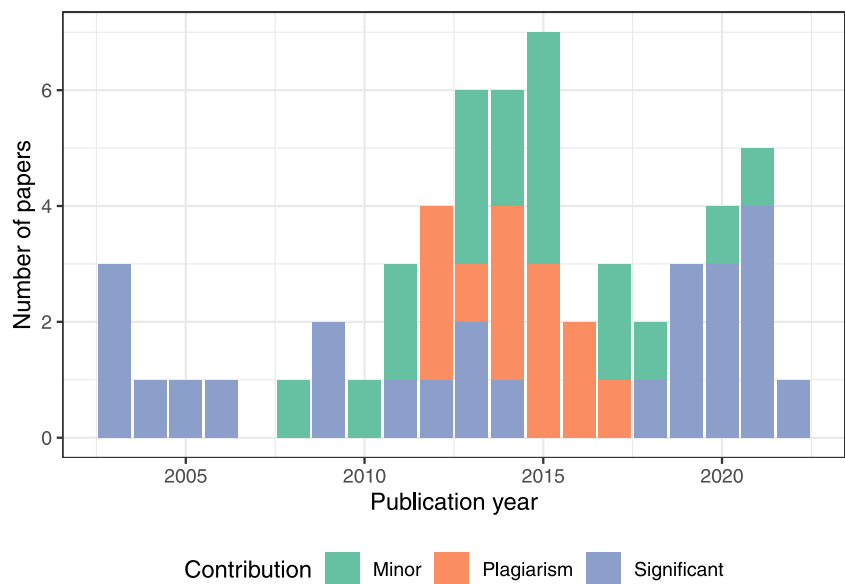[6] The citation numbers were collected in January 2023.

Fig. 1. Number of general survey papers published per year sorted into papers with plagiarism, papers that have only a minor contribution to the SOTA and papers that improve the SOTA.
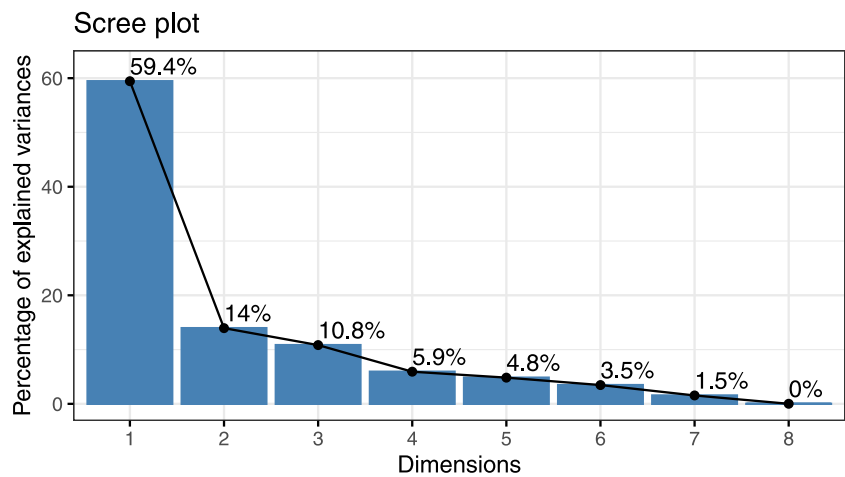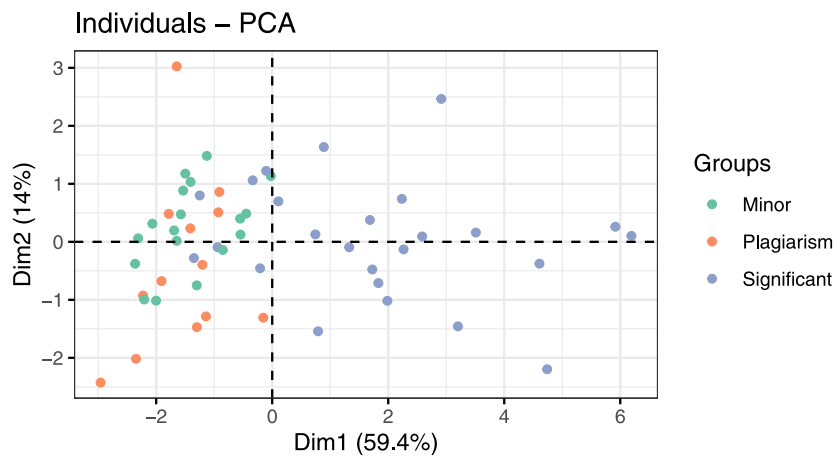

Fig. 2. Scree plot of principal component analysis results on the numerical characteristics of the survey papers.
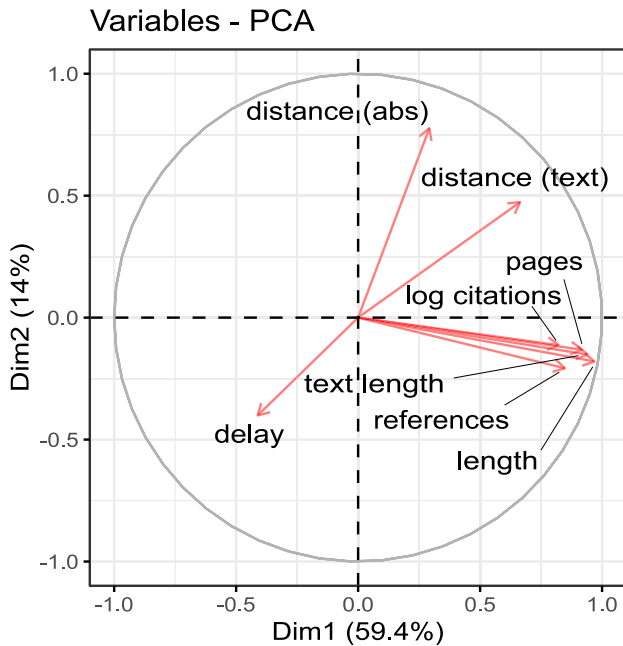

Fig. 3. Principal component analysis results on the numerical characteristics of the survey papers (first two principal components).

Variables - PCA

Fig. 4. Contributions of the variables to the first two principal components of the numerical characteristics of the survey papers.

Table 4
Number of papers from class C5 that focus on a specific aspect of outlier detection, grouped by specificity axe. The total is larger than 170 as 45 papers use two axes of specificity.

Specificity axis	Number of papers
Application fields	109
Class of methods	59
Data types	29
Learning conditions	18

3.5. Specific surveys

To get additional insights on the field, we did some further analysis of the papers in C5 (surveys on some specific aspects of outlier detection). Based on the content of the title, the abstract and, if needed, the full paper, we identified the restriction chosen by the authors with respect to four axes: application fields, class of methods, data types and learning conditions. Statistics on the specificity of those papers are given in Table 4.

Application fields are quite varied and we identified 39 different expressions used to design them ranging from very broad ones (such as *computer networks*) to very specific ones (for instance *water quality* or *smart grids*), see Table 5 for the main ones. The field of networked objects interpreted in a broad sense is largely dominating the application specific surveys and is itself structured in subfields. Some surveys cover the general case of networks of computers while others focus on networks of lower capacity systems such as wireless sensor networks or internet-of-things (IOT). Some surveys cover more generally intrusion detection, that is a specific case of anomalies in the context of networked computers. Finally some surveys cover other specific cases of security issues such as financial frauds.

Surveys specific to a particular class of methods are essentially dedicated to deep learning and to a less extent to robust statistics. We identified indeed 18 surveys focusing on neural networks (among 59 restricted on the method axis), among which 16 are about deep learning in general, one about Long Short-Term Memory models (LSTM, [52]) and one about Generative Adversarial Network (GAN, [53]). Moreover,

Table 5
Main application fields of specialized surveys. All other fields are specific to three papers at most.

Application field	Number of papers
Computer network (general case)	22
Sensor networks	19
Intrusion detection	16
Internet-of-things (IOT)	9
Financial frauds	6

among the 16 deep learning surveys, 12 are specific either to an application field (e.g. IOT) or to a type of data (for instance time series). We have also identified 7 papers on robust statistics, while the focus of the remaining 34 papers is more spread from k-means to clustering.

The data type specialization axis does not exhibit any dominating type apart from temporal data which are the main focus of 9 surveys out of 29. Finally, papers that cover a specific learning context are mainly dedicated to data streams (14 surveys among 18).

It appears clearly from this analysis that specialized surveys belong mainly to the field of computer science and computer engineering, while contributions from statistics seem to be relatively limited. Another clear outcome is the focus on deep learning which is sufficient popular to be combined with other restrictions. Those findings will be confirmed to some extent in the following section dedicated to the analysis of the selected papers.

Notice that results of the present section should be considered with caution and only as building blocks for a systematic survey on the identified subfields. Indeed because of the search strategy adopted we may have missed interesting specific surveys on e.g. robust statistics. Moreover, the analysis has been made on the raw results of the classification without applying the full methodology developed for the general surveys (snowball search and quality assessment). Thus the figures reported here are only indicative.

4. Global analysis of the selected survey papers

We discuss in this section general aspects of the selected papers, in particular their paper collection methodology, or the lack thereof (Section 4.1), their structure (Section 4.2), the structure of the field (Section 4.3) and their vocabulary (Section 4.4).

4.1. Methodology

It should first be noted that almost none of the surveys include a proper paper collection methodology. In fact only two of the papers in class C6 describe the way the papers were collected and selected [40,50]. [48] describes also a systematic literature collection but in less details and as a way to validate hypotheses about learning paradigms and their use in research papers.

The absence of a proper paper collection methodology is problematic as collection bias could be present: for instance the papers discussed in [51] give the impression that the contemporary research in outlier detection is almost uniquely conducted with deep learning approaches, whereas in the slightly older paper [47], deep learning papers are a minority.

In addition to those potential biases, the absence of an explicit paper collection methodology prevents its reproduction and increases the efforts needed to update the survey in the future. As most surveys consist in an organized collection of short summaries of selected papers, their long term value is potentially limited. It seems therefore important to be able to update them somehow, hence to follow a proper methodology.

4.2. Paper structure

Many survey papers are structured in a quite standard way: the authors identify a collection of interesting papers, arrange them in categories (potentially structured into a taxonomy) and then provide for each paper a short summary that contrasts it to other papers in the same category. In general, high level comparisons between categories are also provided.

Most of the surveys considered in the present paper do not depart from this general scheme. In our opinion, this type of structure has more drawbacks than advantages.

There is of course value in providing a short summary of recent papers: the number of papers produced each year is enormous and researchers most focus their attentions to a selection of them. To illustrate this consider the recent NeurIPS conferences.[7] They accepted 2344 papers in 2021 and 2672 papers in 2022. Based on the simple search tool available on the conference website, we can select papers about outlier and anomaly detection, 8 in 2021 and 14 in 2022. This would miss directly related papers that use a slightly different framing, for instance papers about out-of-distribution detection [51]. More generally, as the number of papers published (or simply made available on arXiv[8]) keeps increasing, researchers are likely to miss important papers that are only slightly departing from their main focus. In this context, summary oriented surveys can help researchers to remain aware of the progress in the state-of-the-art on subjects that are closely related to their main research interests.

However, the value of this type of surveys tend to decrease relatively quickly. For instance while [3,34] were very thorough surveys in 2003, their contemporary relevance is mainly historical and they illustrate by contrast with e.g. [13] the tremendous evolution of the field in almost 20 years. The recent survey [12] is currently very useful as a reference for the algorithmic complexity of a large collection of methods, but its interest will decrease over the years with the introduction of new and more efficient methods.

In addition, the organization of this type of surveys in broad categories is generally detrimental to the presentation of general issues (such as the problem induced by high dimensional data, see Section 5.2). Many of those surveys tend to get caught into details about the specific algorithms they are discussing at a given point while missing the point of agreement or the crucial differences between hypotheses. For instance, as pointed out in [47], many surveys distinguish *distance based* and *density based* outlier detection methods, while they are all essentially based on comparing distances to nearest neighbors and share therefore a lot of advantages and limitations.

The need for structure in any paper is fulfilled in this type of work by relying on categories (and taxonomies) in a way that may seem somewhat exaggerated. For instance, the idea that there may be a *statistical* method category is questionable as many *distance based* approaches can be derived from a Gaussian assumption or using a form of kernel density estimation. As pointed out in [43] most of the *nearest neighbors* based approached can be seen as density estimation based methods, as a consequence as non parametric statistical methods. See Sections 5.1 and 6.2 for longer discussions about taxonomies.

Some surveys considered in the present paper deviate from this generic structure, partially or completely. In our opinion they are the most interesting ones from a mid to long term perspective. We will include their findings in the discussion in Section 5. Their main contributions in addition to the papers they discussed can be summarized as follows:

- they discuss in details the nature of anomalies and outliers [14,40] by contrasting the different definitions proposed in the literature (Section 5.1), way beyond the traditional separation between point, contextual and collective anomalies proposed initially in [29];
- they provide some unifying views on methods that are generally discussed independently in summary based reviews [13,43,44];
- they show links between variants of outlier detection or extensions of the concepts, for instance distinguishing rare events detection from novelty detection [48] or discussing the general framework of out-of-distribution detection [51];
- they focus on major general issues such as large scale data [49] or high dimensional data [39] (Section 5.2).

4.3. Structure of the field

As shown on Fig. 1, there is a renewed need and interest for writing surveys on outlier detection. An initial series of surveys was published in the early 2000s, followed by a regular publication of one significant survey every two years. Since 2018, we are witnessing a significant increase in significant survey publications.

The field is relatively integrated in the sense that papers tend to cite previous surveys, as shown on Fig. 5. Highly cited papers, in particular [29,32] are not only cited by general papers but also by almost all the survey papers collected here.

Nevertheless, we can see that some older papers such as [33,35] tend to be forgotten. Several recent papers, such as [44,51] tend to cite only a small selection of recent surveys. This indicates a potential shift from the first phase of papers published before 2015 to the current phase which started around 2018. Older papers are replaced by more recent ones mainly as the former lose their summary oriented value, as discussed in the previous section. See Section 4.4 for additional remarks based on topic models.

Another potential source of the drop in important survey publication around 2015 is the disruption in the machine learning field induce by the re-emergence of neural networks with the explosion of deep learning (see Section 5.6). With such a disruption, it takes several years to propose novel approaches in this new paradigm and then a lag ensues regarding the publication of surveys. This phenomenon is also potential coupled with high progress rate of deep learning methods which may explain the large number of surveys that are both specific to deep learning and to something else (such as images or IOT) as discussed in Section 3.5.

4.4. Topic modeling

The vocabulary used in the selected papers has been investigated using a latent Dirichlet allocation (LDA) model [54]. The model has been trained on the text extracted from the pdf files, after some processing steps: suppression of special characters, numbers, and stop words, and lemmatisation. LDA has been trained both on 1-grams and 2-grams, and the number of topics has been selected using a trade-off between different criteria [55–58]. The model trained on 2-grams provided interesting results, with five topics illustrated in Fig. 6 and an associated clustering of the papers illustrated in Tables 6, 7, 8, 9, and 9. The clustering was obtained by using the topic distribution of each paper as its numerical representation. We are here in a particular case were each paper uses almost only one of the four topics as thus the clustering is obvious: we assign each paper to its dominating topic. Topics themselves are analyzed using two complementary illustrations: on the one hand, we extract top 0.1% of the most salient bigrams and represent their associated probabilities in each topic in Fig. 6, and on the other hand, we represent the 20 most frequent terms in each topic in Figs. 7–9.

Word saliency (bigrams in our case) has been introduced in [59] as a weighted Kullback–Leibler divergence between the posterior distribution of the topics conditionally to a specific word, and the marginal

---

Fig. 5. Citations between the survey papers: each row (resp. column) represents a paper, ordered by publication year (top to bottom, resp. left to right). Thin gray lines separate papers by publication years. The color of a paper identifier and of the corresponding row corresponds to the topic to which the paper is assigned (see Section 4.4). A colored square with a black border on a row shows that the row paper cites the column paper. The filling color is the topic of the cited paper. See Table 3 for the paper codes.

distribution of the topics. Hence, Fig. 6 illustrates the most informative bigrams in terms of how discriminant they are for the emerging topics. It may be completed by Fig. B.14 in the Appendix, which illustrates how these salient bigrams are similar to each other and how blocks of meaningful content emerge within topics and within documents.

When studying Figs. 6, 7, 8 and 9, it is worth noticing that topics appear to deal each with a different point of view: the first speaks mainly of *outlier detection* and *novelty detection*, the second both of *anomaly detection* and *outlier detection*, the third essentially of *outlier* detection, the fourth mostly of *novely detection*, while the fifth mostly of *anomaly detection*.

The papers have been sorted according to publication time in table to emphasize the fact that there is no obvious temporal structure in topics 1, 2 and 3, while topics 4 and 5 could be qualified as more recent ones, even if is contains the most cited survey [29] which is older than the other papers. This weak temporal structure can be seen also on Fig. 10 and appears to some extent in Fig. 5 (papers of topic 4 are the less cited ones because they are somewhat recent).

The first cluster of documents, associated to the first topic, contains the seven papers listed in Table 6. It contains the two historical surveys [3,34], as well as two papers by the same authors [39,43]. These

surveys are aiming at bridging gaps between the statistical and the machine learning community, hence their similarity. Some terms such as *rare event* or *learn scenario* are extremely specific. Terms such as *dimensional data*, or, to a lesser extent, *neural network*, *signal processing*, *time series* or *data stream* are also quite specific. For instance, the issue of high dimensional data is indeed addressed in detail in [39,41], but also in [38], and concern either the entire paper, or consistent sections of it.

The second topic alternates between *outlier detection* and *anomaly detection*, and is associated to the papers summarized in Table 7. These documents share common syntagms such as *algorithm based*, *detection algorithm*, *mining techniques* or *outlier mining*. *Machine learning* is also specific to this topic, and at a closer look, several of the papers in this cluster are stemming from the data mining community [36] or the machine learning one [50]. The papers in this cluster provide quite general surveys, with very similar taxonomies, the later ones inspired by the historical survey [32].

The third topic contains most of the salient terms, while the papers associated to it and summarized in Table 8 are speaking most specifically about *outlier detection*. On the one hand, this topic appears to focus more than the others on statistical approaches, hence the presence of
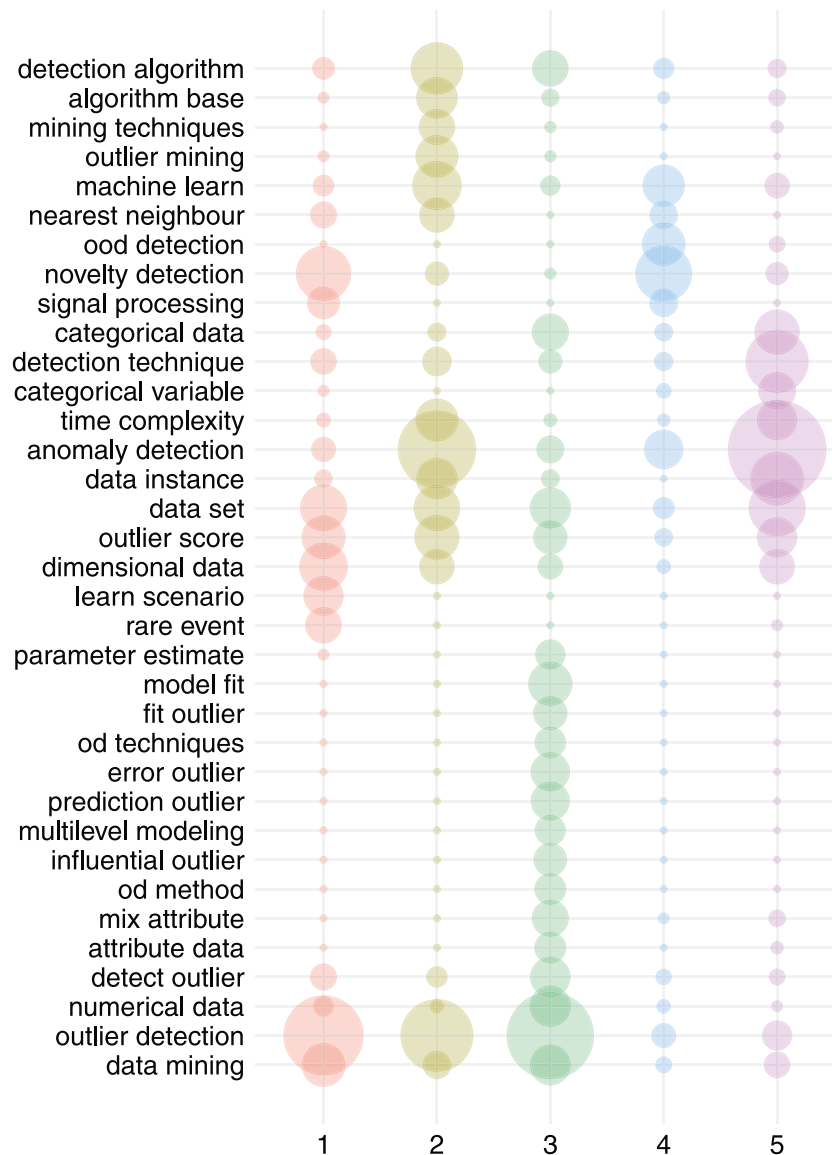
Fig. 6. Top 0.1% of the most salient bigrams in the corpus, for the LDA model. The surface of each disk is proportional to the frequency of the associated bigram within the topic.

bigrams such as *model fit, parameter estimate, fit outlier, predict outlier, multilevel modelling ...*. Indeed, surveys such as [16,35] or [46] propose taxonomies separating *statistical methods* from *machine learning* or *data mining methods*. On the other hand, surveys in this topic, such as [44] are addressing the issue of data types, and particularly other features than numerical ones. Hence, bigrams such as *numerical data, attribute data, mix attribute* appear with large frequencies.

Topic four contains only three papers, listed in Table 9, and very few salient terms: *novelty detection* and *ood detection*. It mixes papers that are quite different and very specific: [14] discusses anomaly types, [42] can be seen as a modernized version of the companion papers [3,34] as it leverages the signal processing literature, while [51] is the only survey that extends its scope to include subjects related to anomaly detection such as out-of-distribution detection (hence the saliency of *ood detection*).

The last topic concerns four documents listed in Table 10, among which the historical [29] and the recent and complete [13]. The other two are a survey focusing on outliers for categorical data [45], or outliers for big data [49]. They all appear to prefer using *anomaly detection* instead of *outlier* or *novelty detection*. In our opinion, all four are among the most important surveys for the literature, whether by

their historical and wide spectrum value, by their unifying view, or by addressing specific questions related to outlier detection. The salient bigrams for this topic are *categorical data* and *categorical variable*, *data set* or *data instance*, or *time complexity* and *dimensional data*.

Overall this topic based analysis provides an idea of the variability of the vocabulary used to describe outliers and anomalies, as well as, the methods used to detect them. However, this variability seems to be quite circumstantial and is somewhat explained by the pervasiveness of outlier detection in many relatively separated fields such as signal processing, statistics, data mining, etc. While this could be interpreted as indicative of a fragmented field, we will see in the next section that most of the surveys agree on numerous crucial points.

5. Consensual findings

We discuss in this section the high level findings that appear to be consensual throughout the selected surveys.

5.1. Definitions and taxonomies

The quotes of [2,8] -discussed in Section 2- are often used [6,31,43] to illustrate the definition of outliers and most reviews agree that the
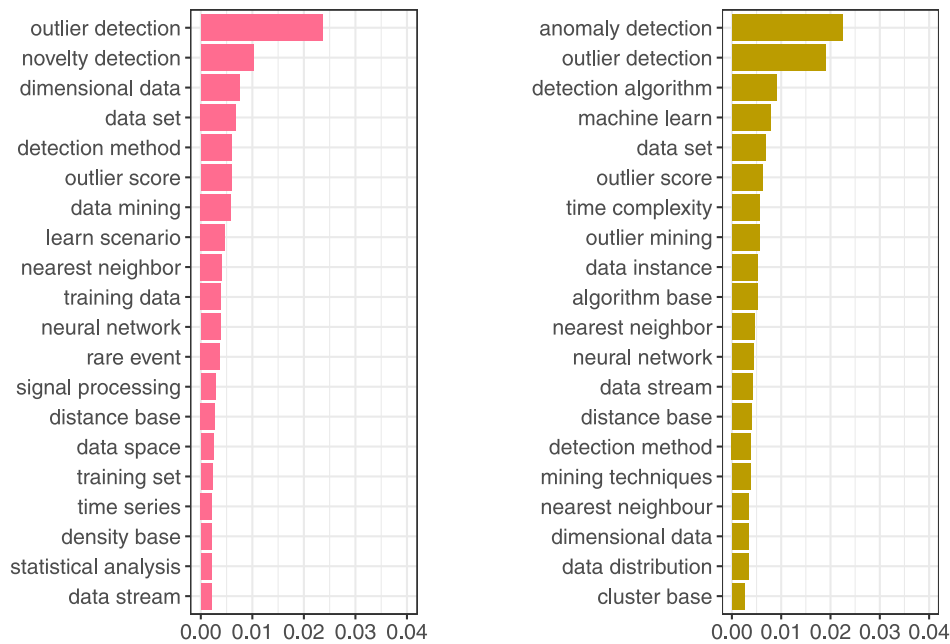
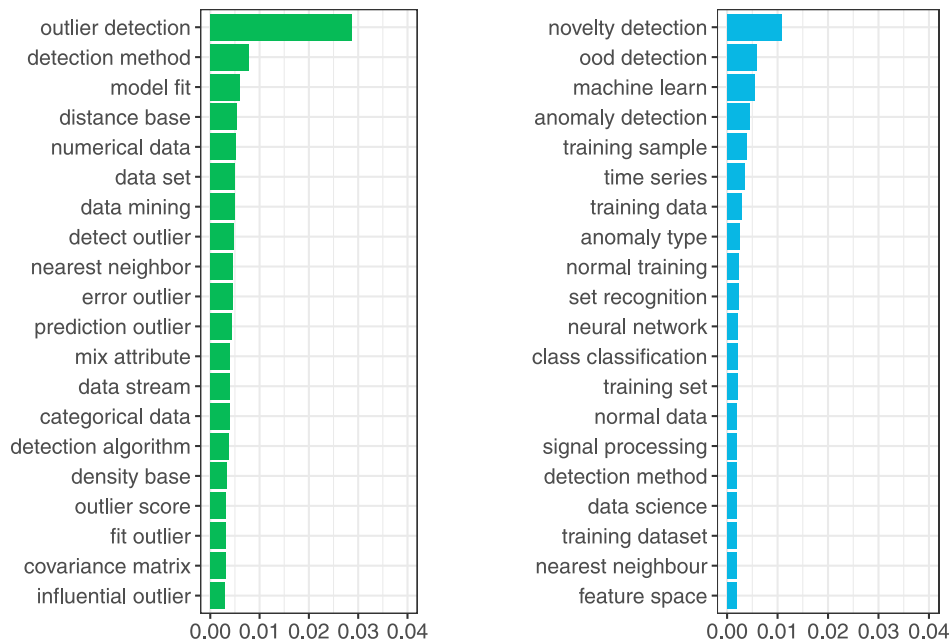Fig. 7. Most frequent 20 bigrams for topics 1 and 2.

Fig. 8. Most frequent 20 bigrams for topics 3 and 4.

definition is vague and application-dependent. Chandola et al. write in [29] "*Anomalies are patterns in data that do not conform to a well defined notion of normal behavior.*" which constitutes already a step towards more formal definitions.

Some review propose a definition of their own as [13], *an anomaly is an observation that deviates considerably from some concept of normality*. It is only recently that the notion of outlier was defined in a precise mathematical sense in [13].

The literature is shared across the fields of statistic, machine learning, signal processing and data mining. Each field having its application of interest and its own vocabulary. As most review insist on, its leads to a wide literature with many words and expressions covering related concepts from the classical outliers, anomaly, novelty to the less frequent rogue values, mavericks, dirty data, etc. In [48], the

authors tries to disentangle the definitions of rare event, anomaly, novelty and outlier. Hence, rare events are mostly found in problems of temporal nature. Then, their distinction between the outlier, anomaly and novelty is less consensual. According to their findings, outliers are found in unsupervised scenario whereas novelty and anomaly are found in supervised scenarios. In [13], Ruff et al. ma: papers of topics 4 and 5 tends to be less cited ke a subtle and more consensual difference between anomaly (from a distinct distribution), outlier (rare, low probability event) and novelty (instance from a new region).

Historically, the underlying generating mechanism of outliers was used to distinguish them from normal observations, as discussed in Section 2: there were basically good observations (with a heavy-tailed distribution for example) and contaminants. [31] distinguishes also between two types of outliers that coincide with Hawkins' definition

Table 6

Papers assigned to topic 1.

Paper	Year	Title
MarkouSingh2003Novelty DetectionNeural [34]	2003	Novelty detection: a review—part 2: neural network based approaches
MarkouSingh2003Novelty DetectionStatistical [3]	2003	Novelty detection: a review—part 1: statistical approaches
SuTsai2011OutlierDetection [38]	2011	Outlier detection
ZimekSchubertEtAl2012Survey Unsupervised [39]	2012	A survey on unsupervised outlier detection in high-dimensional numerical data
Zhang2013 AdvancementsOutlier [41]	2013	Advancements of outlier detection: A survey
ZimekFilzmoser2018ThereBack [43]	2018	There and back again: Outlier detection between statistical reasoning and data mining algorithms
CarrenoInzaEtAl2020AnalyzingRare [48]	2020	Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework

Table 7

Papers assigned to topic 2.

Paper	Year	Title
HodgeAustin2004 SurveyOutlier [32]	2004	A survey of outlier detection methodologies
AgyemangBarkerEtAl2006Comprehensive Survey [36]	2006	A comprehensive survey of numeric and symbolic outlier mining techniques
BoukercheZhengEtAl2020OutlierDetection [47]	2020	Outlier detection: Methods, models, and classification
NassifTalibEtAl2021MachineLearning [50]	2021	Machine learning for anomaly detection: A systematic review
SamariyaThakkar2021 ComprehensiveSurvey [12]	2021	A comprehensive survey of anomaly detection algorithms

Table 8

Papers assigned to topic 3.

Paper	Year	Title
Petrovskiy2003OutlierDetection [33]	2003	Outlier detection algorithms in data mining systems
BenGal2005OutlierDetection [35]	2005	Outlier Detection in: Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers
HadiImon2009 DectectionOutliers [16]	2009	Detection of outliers
AguinisGottfredsonEtAl2013BestPractice [40]	2013	Best-practice recommendations for defining, identifying, and handling outliers
Rokhman2019SurveyMixed [44]	2019	A survey on mixed-attribute outlier detection methods
WangBahEtAl2019ProgressOutlier [46]	2019	Progress in outlier detection techniques: A survey

Table 9

Papers assigned to topic 4.

Paper	Year	Title
PimentelCliftonEtAl2014ReviewNovelty [42]	2014	A review of novelty detection
Foorthuis2021NatureTypes [14]	2021	On the nature and types of anomalies: A review of deviations in data
SalehiMirzaeiEtAl2022UnifiedSurvey [51]	2022	A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges

Table 10

Papers assigned to topic 5.

Paper	Year	Title
ChandolaBanerjeeEtAl2009AnomalyDetection [29]	2009	Anomaly detection: A survey
TahaHadi2019AnomalyDetection [45]	2019	Anomaly detection methods for categorical data: A review
ThudumuBranchEtAl2020 ComprehensiveSurvey [49]	2020	A comprehensive survey of anomaly detection techniques for high dimensional big data
RuffKauffmannEtAl2021UnifyingReview [13]	2021	A unifying review of deep and shallow anomaly detection

Fig. 9. Most frequent 20 bigrams for topic 5.

[2], errors and events (generated by a different mechanism) to be identified for further investigation. This tradition is still followed by recent surveys, for instance by [14] for whom there is noise or actual signals.
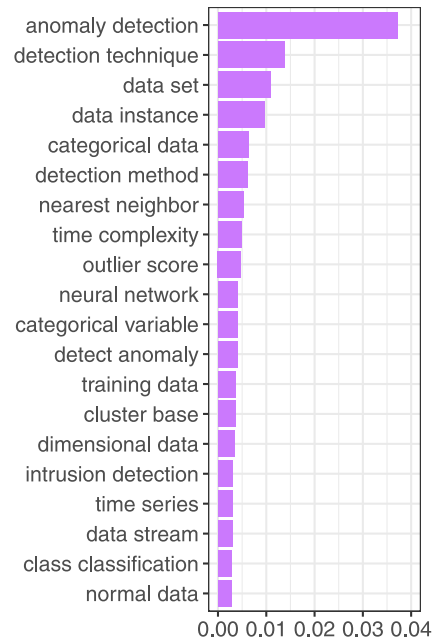
This simple taxonomy is later enriched by [40], that proposes 5 main types of outliers (with a total of 13 subtypes). It covers the types of [2,31] but also adds interesting types such as model outliers (having a large residual and possibly influencing the model) and cluster analysis outlier. This view on the model being at the center of outlier detection is later developed in [43]. Some reviews also used the model either parametric, semi-parametric and non-parametric [31] or shallow and deep [13] as an axis of analysis.
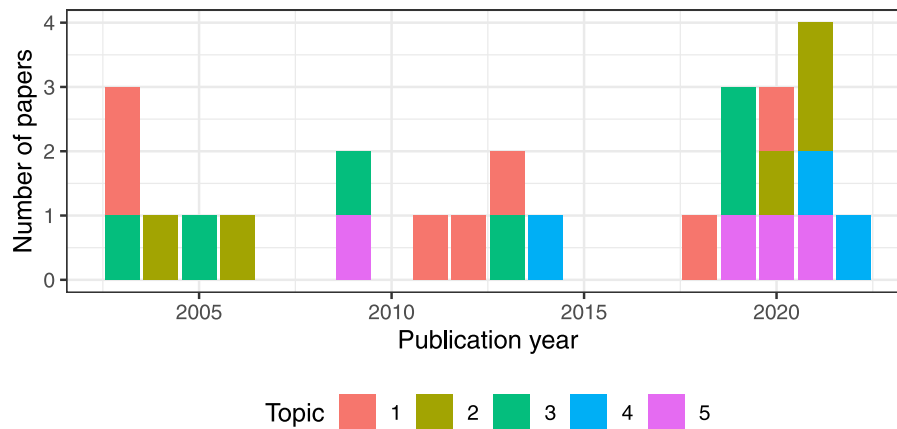
Fig. 10. Topic distribution by publication year: each bar counts the number of papers published this year. The bar content is colored according to the topic to which is paper published this year is assigned.

Influenced by the need to adapt to complex data, a more modern view has emerged. The simpler case was opposing global outliers to local outliers [31] (see Section 6.1 on this aspect). Those definition are refined in [29] that distinguish point anomaly, contextual anomaly and collective anomaly. The simplest case is the one of *point anomalies* where a single observation can be classified in isolation as an anomaly. This type of anomaly is the main focus of most of the methods. In statistical and machine learning terms, it is associated to the classical independence hypothesis between observations. When the observations are statistically dependent, the notion of anomaly should be revised. Indeed, the expected value of an observation is in this case dependent from the values of other observations. Then the status of an observation (normal or anomalous) cannot be decided in isolation. Such anomalies are called *contextual anomalies* or *conditional anomalies*. In [14], only atomic and aggregated outliers are discussed (leaving aside the context). [13] adds to the types discussed in [29] two more types suited to the deep neural network case, low-level sensory anomaly and high-level semantic anomaly (which could be thought as subtypes of contextual anomalies). Low and high refer to the feature level in a deep learning perspective. In images for example, a low-level sensory anomaly would be the pixel or texture level whereas a high-level semantic anomaly would be the presence of an object in the image.

Data types were another axes of analysis of outliers. Originally, the main challenges were to extend the algorithms to multivariate cases and time-series [2,8,9]. Hence, a basic axis for a taxonomy would oppose univariate to multivariate cases. It is now completed with more complex data type such as categorical data [45,60] and text, time-series and discrete sequences or spatial data [31]. This list is completed with graphs and networks in [6,29]. It must be noted that this discussion is completely absent of [13] as they handle data through a feature map which could be adapted to complex data.

In summary, a consensus emerges from the literature on three axes of taxonomy. The outliers can be categorized according to

- the underlying *generative mechanism*, characterizing outliers as errors, interesting or influential,
- the underlying *independence hypothesis*, leading to the definition of point, contextual and collective anomaly,
- the *data type*, from simple univariate data to multivariate and structured data (including categorical data, times series, spatial data and graphs).

Globally, while taxonomies on anomaly types have evolved through time, surveys tend to agree on the main separations (e.g. contextual versus isolated). Many surveys do not even mention them owing to consensus associated to the definition proposed in [29]. This is in stark contrast with taxonomies on the methods themselves which are highly variable as discussed in Section 6.2.

5.2. The high dimensionality issue

While extremely important, the impact of high dimensionality on outlier detection performances is discussed in several reviews, but not in the majority of them and not always by considering the effects of the curse of dimensionality, beyond the complexity burden. For example, many surveys considered in this study mention *distance-based* or *density-based* methods, but only rarely mention how data defined in a high-dimensional feature space may impact – negatively – the performances of these methods.

Notice that any discussion on high dimensionality should distinguish the dimension of the description space, that is the number of features used to describe the entities under study, from the intrinsic dimensionality of the data (see e.g. [61]). Outlier detection in the description space is directly impacted by the *curse of dimensionality* while methods that try first to reduce somehow the dimensionality prior to the application of a classical outlier detection approach face this curse during the dimensionality reduction phase. Following [13], it could be argued that *reconstruction models*, from principal component analysis to auto-encoders, put their effort in the dimensionality reduction phase, will density estimation models and one-class approaches try to address directly the original data. In this section, we discuss the way surveys present the difficulties of both approaches.

One of the first surveys on unsupervised outlier detection, [31] (unfortunately unpublished) proposes a taxonomy where the case of high-dimensional data sets is specifically considered, as different from the *simple data set* baseline. After recalling that *in high-dimensional spaces the data is sparse, the convex hull more difficult to discern and the notion of proximity less meaningful*, the survey focuses on subspace-based methods and on some specific distance-based methods, and discusses the computational complexity, the efficiency, and the difficulty to tune pre-defined parameters for these two families of methods.

The question of outlier detection in the context of high-dimensional data has been studied in detail in [39]. Several illustrations allow to assess the main issues related to high-dimensionality and to derive several consequences for outlier detection tasks: concentration of scores, noise attributes, definition of reference sets, bias of scores, interpretation and contrast of scores, exponential search space, data-snooping bias, and eventually hubness. These problems challenge the correctness of the methods, and the evaluation criteria for assessing the validity of outlier detection. Traditional methods, for instance, based on distance computations in the description space and classified according to different taxonomies as *distance-based*, *density-based*, *nearest-neighbour based* or even *clustering based* are thus generally severely impacted by high-dimensionality. The rest of the paper discusses methods suited for high-dimensionality, either from an *efficiency* or an *effectiveness* point

of view. Several classes of methods – approximate neighborhood computations, ensemble methods, angle-based methods, subspace-based methods – are critically discussed, while stressing the difficulties of actually evaluating the different methods, particularly from a qualitative point of view.

The conclusion of [39] which dates back to more than ten years ago states that

> *the area of outlier detection specialized for high-dimensional data offers lots of opportunities for improvement. There are just a few approaches around in the literature so far, yet there are many directions to go and problems still to solve. The researcher should, though, be aware of the existing attempts of solution and the associated pitfalls.*

With this in mind, it is at least surprising – and actually quite troublesome – that many subsequent surveys do not discuss specifically high-dimensionality issues, propose brief discussions lacking of perspective, or, worse, continue to present *distance-based* and *dissimilarity-based* methods without mentioning the curse of dimensionality or considering it from the point of view of computational burden only.

The issues related to high-dimensionality and some of the problems stressed in [39] have been discussed however in some recent surveys, such as [46,47,49]. [46] is mainly interested in the computational burden and misses a thorough discussion on the effectiveness of the methods. Furthermore, the taxonomy of the methods is not very helpful for assessing how they deal – or not – with dimensionality issues. [47] picks up on [39] and review more recent methods, while preserving a taxonomy of methods according to their efficiency or effectiveness in the high-dimensional framework. [49] also draw extensively from [39], by recalling some of the issues implied by the curse of dimensionality, and reviewing some subspace-based methods.

Nevertheless, the question of high dimensionality appears to be still an open one, and recent surveys are still quite void off thorough discussions on the topic, beyond, as we mentioned, the computational complexity. The fact that Aggarwal's text book [6] dedicates a simple chapter to the issue and bases his discussion only on subspace methods is also revealing: true high dimensional problems remain very difficult and outlier detection is not anomalous with respect to this difficulty. A possible explanation is the recent focus on deep learning approaches which use very frequently a low dimensional latent representation as shown in [51] and are thus targeting high dimensional issues via a form of intrinsic dimensionality recovery.

5.3. On the importance of anomaly scores

Anomaly detection methods can output either a score that measures to what extent an observation is anomalous or a binary label that directly says whether the observation should be considered anomalous or not. These two possible outcomes are at least mentioned by almost all the surveys, starting early ones such as [32].

From a machine learning point of view, binary labeling is attractive as it is a simple task: from a score it is always possible to derive a classification via a simple thresholding while the reverse is false. Thus labeling should reach better performances than scoring considering the same resources (both in terms of data size and of computational burden). Interestingly, while many survey papers cover one-class approaches [62] the only one to interpret it in terms of machine learning efficiency is [13], under Vapnik's simplicity principle. This is probably because many surveys miss the fact that one-class methods are estimating a level set of the probability density of the normal data, even if this aspect in only implicit in the construction of the method.

While appealing on a machine learning point of view, labeling methods are somewhat less convenient in practice. Firstly two recent surveys emphasize the need for interpretable and explainable decisions [13,43]. According to a recent survey on the subject [63], scoring can be seen as a form of minimal step in this direction.

In addition, scoring can be tuned *a posteriori* to the operational conditions: the scoring threshold between outliers and normal data can be adapted to e.g. the human resources available to investigate the anomalies. Using metrics such as the area under the ROC curve (AUC) (or possibly better ones depending on the trade-off between precision and false alarm rate, see [13]), one can evaluate the performances of the scoring approach for the full range of decision threshold.

Finally, a score can be used to rank the instances rather than to split them into outliers and normal ones. Notice however that ranking is putting the weight of the decision on the analyst shoulders, as they will have to decide where to stop in the list of ranked observations. Even worse, the stopping decision could be driven in this case by operational considerations in an opaque way (and even possibly changing depending on external events). While those are valid considerations, they should be explicitly stated. As stated in [43],

> *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 is usually anything but obvious.*

In summary, thresholding scores into decisions is part of the model fitting process and should not be ignored. It is interesting to see that while there is a consensus between the two surveys that discuss to some extent the topic [13,43], it is generally completely disregarded in the other surveys.

5.4. Learning conditions

It should be first pointed out that the learning conditions to be discussed here are somewhat orthogonal to previously discussed scoring or labeling as both can apply to either supervised and unsupervised conditions (see [43]).

There is a very clear consensus in the literature that anomaly detection is done in majority in an unsupervised learning context as far as the nature of the observations is concerned: the detection algorithm works without knowing the true nature of the observations. The recent survey [50], which uses a sound paper collection methodology, reports that 58% of the papers it reviewed can be associated to a specific learning paradigm. Among them, 46% used an unsupervised learning paradigm. In almost all the surveys included in the present paper, supervised methods occupy only a small part of the discussion (for instance 4 pages out of 58 in [29]). This is also the case in the main text book on the subject which dedicates only a chapter to supervised models [6]. Notable exceptions are [48,51] discussed below.

However, the anomaly detection paradigm should not be confused with the learning paradigm of the main task the analyst is trying to solve. Indeed, as recalled in Section 2.2, early statistical approaches treat outliers as a nuisance that impairs estimation, in particular in a supervised context. For instance most of [9] is dedicated to the effect of anomalies on the estimation of generalized linear models (see also [11]). So the anomalous versus normal nature of the examples is unknown, but the learning paradigm could be supervised.

In the selected surveys, evocation of this subtlety seems to be related to older papers with the exception of [40] (which is still in the first phase of surveys, before 2018, see Section 4.3). This is quite natural as the main evolution of the field is to shift the attention from being robust to outliers to detecting them. In addition, data mining oriented surveys, such as [29], tend to interpret robust methods from the point of view of outlier detection. For instance edited linear regression techniques such as Rousseeuw's *least trimmed squares* [64] whose original aim is fitting a linear model robustly in presence of outliers, is presented as a way to detect outliers (by their large residuals). Statisticians are clearly aware of this shift as exemplified in the surveys by [16] which dedicates a significant space to statistical methods but explicitly restricts the

discussion to the detection setting in the unsupervised context. We refer the reader to [40] for thorough discussion on the methodological aspects of the "outliers as a nuisance" paradigm (albeit limited to the field of organizational science).

Nevertheless, while quite uncommon, the case of supervised learning is generally discussed in the surveys we selected. Indeed in some application contexts such as fraud or computer intrusion detection, it may be possible to collect a data set with labeled examples combining (a lot of) normal examples and (a small set of) anomalous examples. In this case, the problem is a standard but difficult supervised learning one. The difficulties come from the imbalanced nature of the data [25–27], as collecting examples of anomalous behavior is generally difficult, and from the ill-posed nature of the classification: while the normal data class is well defined, the anomalous data form a collection of unrelated examples that can exhibit vastly different characteristics. In addition, the normal data set can be contaminated by undetected anomalies [13].

Among the surveys studied here, only one is explicitly dedicated to the supervised learning paradigm [48]. It distinguishes anomaly detection as the supervised case from outlier detection as the unsupervised one, a distinction that we did not encounter elsewhere in the literature. It also discusses variation over anomaly detection in the context of rare event detection and novelty detection. We believe that this survey is introducing an unfortunate confusion between one-class learning [62] and supervised learning, especially compared to the thorough discussion on the subject in [13] for instance, but it has the merit of showing that anomaly detection is related closely to other problems, especially in the supervised context. Those problems are surveyed in [51] which extends the discussion to the more general setting of out-of-distribution and open-set detection. This corresponds in particular to situations where a supervised model is trained on a subset of the classes it will be facing in the deployment phase.

Between those two extreme cases, with zero or full supervision, different levels of partial supervision have been explored relatively recently in the context of anomaly detection. While the concept is briefly mentioned as early as in [32], it restricted to the idea of having examples of the normal data (labeled as such) and a collection of unlabeled data, a framework known as LPUE, Learning from positive and unlabeled examples [13]. More importantly, early surveys do not discuss papers using weak supervision beyond citing a few examples, even in surveys that dedicate a (short) section to them (e.g. [38]). A systematic coverage starts only in relatively recent papers such as [46] (and well as in [6]). The importance of semi-supervised learning in a classical sense, i.e. when labels are also available for the outliers, should not been understated as even a small number of such labels can improve strongly the detection performances [13]. There is for this reason a tendency to try and break out of the unsupervised setting by some form of outlier "generation", especially in the deep learning community [51].

5.5. Benchmarking

Many of the reviews insist on the fact that benchmarks and open datasets are needed for the development of the field [13,43,45,65]. Indeed, repurposing classification datasets as it is done in some surveys may induce biases and limit the results and the conclusions. Using downsampling techniques in the evaluation procedures for building training and validation sets is also troublesome due to the scarcity and the large variance of the outliers. [65] suggest that not only data, but also samples used for training should be made open and available for reproducibility purposes.

It should be also mentioned that benchmarking is not difficult because of the lack of meaningful datasets only. The nature of the outliers themselves and the nature of the quantified outlierness by each of the methods make the task harder. The third aspect to consider is the lack of general and well understood evaluation metrics. Whereas AUC

for instance appears to be plebiscited by the few surveys containing benchmarks, it has its own limits and biases [13].

Beyond the issue of evaluation metrics for the effectiveness of outlier detection, the question of interpretability of the results stands out as at least as important. As mentioned in [43], *the quest for a truly general and superior method is futile*, whereas [13] recall that *all models are wrong*. Since outlier detection is essentially an unsupervised task, benchmarking should be used for analyzing and understanding the strengths and the weaknesses of each method, but also how and when certain models are wrong, especially when confronted to datasets with different characteristics. [13] for example use methods neuralisation and apply explainable AI techniques to get some insights on the interpretability of the methods.

Considering the above, one may wonder whether a general *all-purpose* benchmarking is actually meaningful or useful. Taking into account the diversity in the nature of outliers, as stressed for instance in [14], it would be rather uneasy to build general datasets including meaningful anomalies, and none of the reviews is actually providing any guidelines for building benchmarks of interest.

5.6. Neural network models

A possibly surprising finding of our meta-survey is the prevalence and staying power of artificial neural networks through this 20 years period. The earliest survey considered here is the two parts one by Markou and Singh [3,34], in which one part is entirely dedicated to neural networks. [33] and latter [32] discuss also neural networks. Auto-encoders are already popular as anomaly detection techniques as exemplified by [66,67].

This early presence is followed by a reduced interest phase as papers in the late 2000s and the early 2010s tend to mention neural networks only in passing or not at all, as [16,38–41]. As already hypothesized in Section 4.3, this is probably related to the decrease in popularity of neural networks before the explosion of deep learning. In recent surveys, the absence of neural networks seems to be related to the particular case of categorical data [44,45]. [14] is also a particular as it focuses only on anomaly types. By contrast, the absence of neural network methods from [12] is less easy to interpret.

As expected, deep learning is present in all recent surveys. In particular [51] while it presents itself as a generic survey could almost have been excluded from our study as its discusses almost only deep learning approaches. This is in stark contrast with [13] which goes beyond the separation between shallow and deep models to provide a unifying view for learning an outlier detection model.

A synthetic representation of the evolution of coverage of neural networks over the years is provided by Fig. 11. As explained above, we have a quite clear decrease of interest in the early 2010s. The figure emphasizes another interesting pattern: the resurgence of neural networks can be first attributed to the appearance of deep learning methods (in 2019 and 2020), but recent papers tend to include again shallow artificial neural networks.

As discussed in Section 3.5, deep learning is popular enough to have generated 18 surveys on its use in outlier detection (among them [68,69] are highly cited, despite the fact the first one is only an unpublished technical report). Based on [51], one can also argue that deep learning allowed to unify problems related to outlier detection (namely out-of-distribution and open-set detection) under a general umbrella that could be described as the detection of non conformity of the data to the hypothesized distribution assumption.

6. Debated topics

We discuss in this section two selected topics for which the surveys are not in agreement.
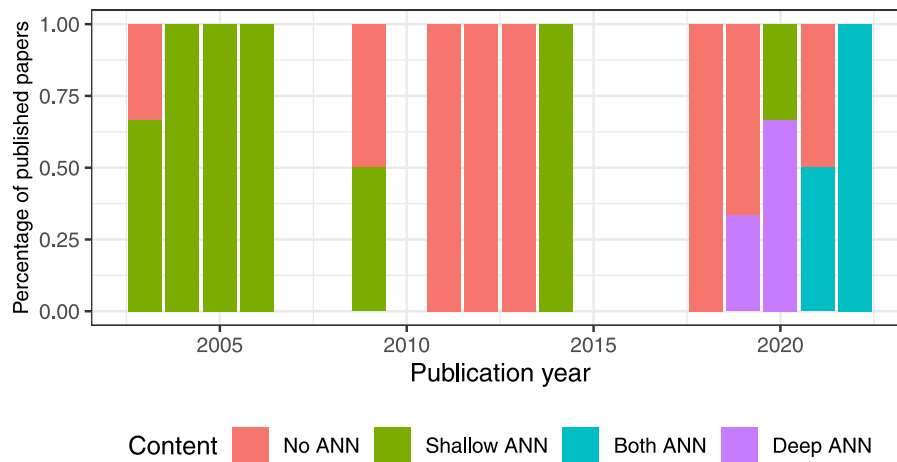
Fig. 11. Inclusion of artificial neural network (ANN) papers per year: bars represent the distribution of the type of ANN papers considered by the surveys published during a given year.

6.1. Local versus global

The idea that there are local and global outliers is quite popular in the field and appeared relatively early [31], possibly as a consequence of the introduction of the Local Outlier Factor (LOF) by [70]. Other surveys mention this distinction, mainly [12,43,46,47]. The classical definition of local versus global anomalies, taken from [31] is

A global outlier is an anomalous data point with respect to all other points in the whole data set, but may not [be one] with respect to points in its local neighborhood. A local outlier is a data point that is significantly different with respect to other points in its local neighborhood, but may not be an outlier in a global view of the data set.

Obviously a global outlier must be a local one as a point anomalous with respect to *all* the points in the whole data set, is anomalous with respect to *any* subset of those points. However the definition introduces a paradox: if a point is significantly different from its neighbors, i.e. by essence the points that are the closest ones in the data set, then it must be significantly different from all the other points. To resolve the paradox, one must consider two different criteria: one criterion is used to define a neighborhood and another one is used to characterize local differences. This is done in the initial paper on LOF [70] which considers k-nearest neighbors to define the local subset of points and a local density estimator to characterize each point.

More generally, as argued in [71], the notion of "local outliers" should be refined in order to distinguish the comparison scope and the characterization scope. The *comparison scope* associates to a given a point a subset of the full data set and uses it as the basis of the decision to consider the point as anomalous or not. The *characterization scope* denotes the subset of the data set used to build a characterization of a point. For instance LOF is twice local. Firstly, a point is only compared to its neighbors. Secondly, each point is characterized by a density computed on its neighbors. As shown in [71] numerous methods are considered *local* but use in practice different combination of local and global aspects (for instance a local comparison scope paired with a global characterization scope).

While most of the literature focuses on the comparison scope, [13] argues that the locality should refer to the characterization scope rather than to the comparison scope. Using any complex *global* density estimation model (or one-class model), one can find small convex regions of low density and thus identify outliers that would be miss by simpler models, without the need for the estimation of a local model for each data point. In other words, *global* models can be used to detect *local* outliers.

A parallel but somewhat similar discussion is provided in [14] in which the opposition between local and global outliers is rephrased in terms of contextual versus non-contextual ones. This is also considered in [71] which shows that another way to circumvent the paradox of the local outlier definition is to use some features of the points to define the comparison scope and the rest of the features to define the characterization scope. A typical example is given by outlier detection in time series, where time is used to define the comparison scope and while the values of the series are used to find possible outliers.

6.2. On taxonomies

As described in Section 4.2 most survey papers in our selection use at least a classification of the methods they discuss into several categories to organize their presentations. Many of them arranged those categories into a hierarchical structure, providing a taxonomy of outlier detection methods. One of the most advanced of such taxonomies is proposed in [31]. Perhaps unsurprisingly, the consensus between those taxonomies and categorizations is minimal.

We believe that this is a consequence of shoehorning a very diverse set of methods into a collection of vaguely defined boxes. We already cite [47] which remarks that *distance based* and *density based* outlier detection methods are very frequently separated while they are all essentially based on comparing distances to nearest neighbors. In some surveys, finding the rationale of the categories is difficult: for instance [46] as a *learning based* category from which clustering methods and ensemble methods are excluded!

There are nevertheless some agreements. For instance, if we look past the differences in names, some of the categories used in [13, 42] align nicely. They both use a reconstruction based category and they somewhat agree on the one-class/level set approaches (called *domain based* in [42]) and on the density estimation methods (called *probabilistic* in [42]).

It seems to use that while the use of categories to organize the presentation of collections of methods is almost mandatory, identifying meaningful categories is a somewhat ill-posed and quite difficult problem. Indeed we personally are convinced that the categories proposed in [13] (and to some extent the ones in [42]) are interesting because they emphasize the main quality metric shared by the methods: density estimation quality, level set quality or reconstruction quality. This provides a high level view on the field and is somewhat orthogonal to the implementation itself. However, this type of categorization is probably far less useful from a practical point of view. If a data scientist is facing an outlier detection problem in e.g. the context of fraud detection, how can they chose between a reconstruction error approach or a density estimation one?
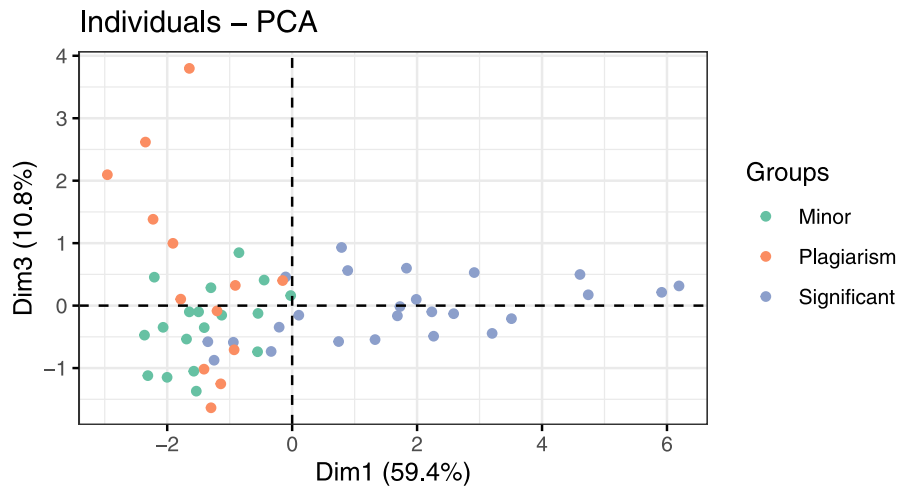
Fig. A.12. Principal component analysis results on the numerical characteristics of the survey papers: projection on the first and third principal components.
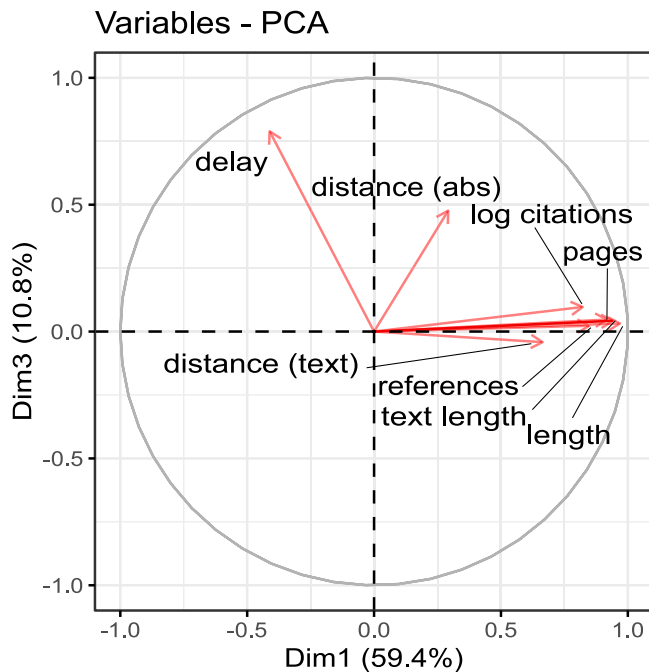

Fig. A.13. Contributions of the variables to the first and third principal components of the numerical characteristics of the survey papers.

7. Conclusion

At the end of the day, what does one draw from the reading of a large collection of surveys on anomaly and outlier detection?

A very disappointing conclusion is both the large presence of plagiarism and the almost total lack of paper collection methodology even in high quality surveys. Both problems are probably not specific to the field of outlier detection. Nevertheless is seems important to emphasize that the paper collection bias is a well documented problem in surveys and that the contribution of surveys to a field could only improve by following standard procedures as outlined in [4] for instance.

Those problems set apart, our meta-survey shows firstly that although the literature is apparently extremely abundant, the number of surveys actually contributing to the state of the art is rather limited, after one has cleared the field. Secondly, it appeared to us that survey approaches consisting in briefly summarizing a list of methods and

proposing a – usually arbitrary – taxonomy are neither really useful for the practitioners, nor meant to last overtime. In our opinion, taxonomies centred around the anomaly types and/or the data characteristics are more useful in practice and may be more easily updated. Thirdly, we observed that important aspects such as the computational complexity, the impact of high dimensionality, the interpretability of the outlier score as a probability measure, or a more unified view of the methods, should be more consistently discussed when reviewing the state of the art. Nevertheless, only a small number of surveys actually consider these issues, and bring perspective to the field.

If one had to read one paper – or a couple of – to get a unified and thorough view on anomaly detection, we suggest the recent survey [13]. Is is, to our knowledge, the first to propose a formal mathematical definition of the notion of outlier and to review a broad area of the field – including deep neural networks – with a probabilistic perspective on the methods. An alternative reading may be [43] which provides a high-level perspective on outlier detection, and attempts to bring a statistical view on the methods, and a probabilistic interpretation of the anomaly scores. Since the specific problems related to the high dimensionality are not specifically addressed in the two previous surveys, we also suggest [39], which provides the most detailed discussion on the curse of dimensionality in the framework of anomaly detection. To get a better understanding of the diverse reality that is sometimes hidden under the generic concept of outliers, we recommend the very detailed discussion on anomaly types provided by [14]. Eventually, Aggarwal's monograph [6] represents an important reading, and contains a consolidated summary of the literature until the mid 2010's.

Beyond those recommendations, we want to emphasize that numerous open questions appearing in several of the surveys we selected should be mentioned, such as the need for benchmark datasets and frameworks, and the challenges related to the interpretability and the visualization of outliers. As it was discussed in a couple of surveys already, while it is most probably useless to look for a universal effective method, due to the different natures of outliers and to the unsupervised framework in most situations, further investigation should be done in the interpretability aspects.

Finally, the role of artificial neural networks, and especially of deep learning, in anomaly detection appears to be on the risen as discussion in Sections 3.5 and 5.6. The number of surveys we found without specifically targeting them and the growing importance of those models in recent generic surveys ask for a systematic review dedicated to deep learning surveys. This could be done using a similar methodology as the one used in the present paper. We would recommend to adapt the search queries to the trends observed in our surveys, in particular in terms of the generalized framework explored in [51]: one should not restrict the search to outlier and anomaly, but rather expend
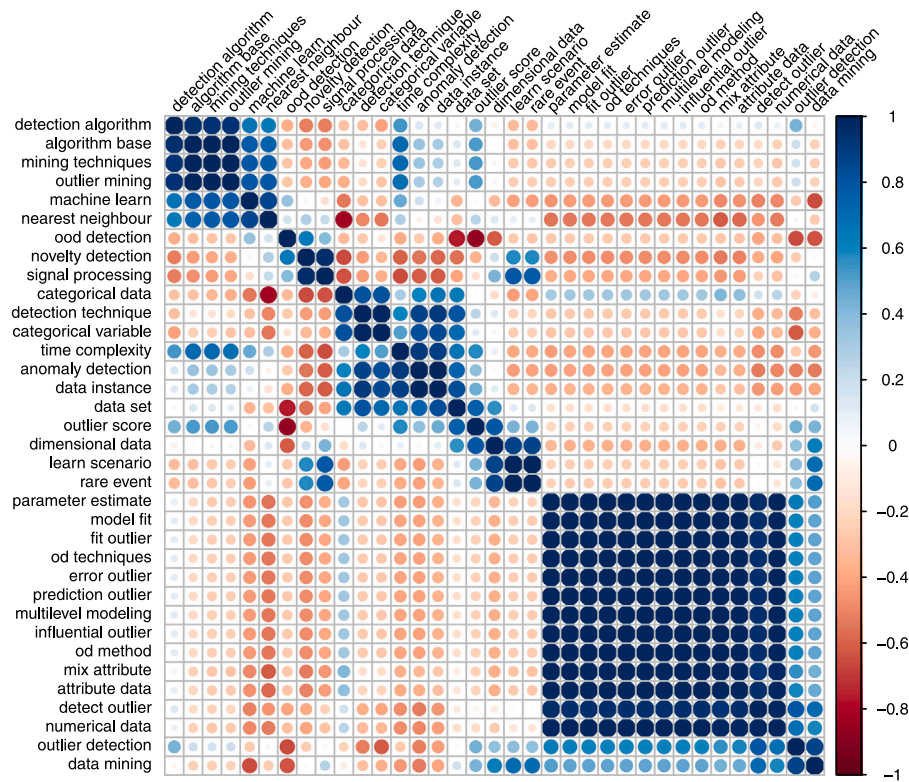
Fig. B.14. Correlation matrix of the most salient 0.1% bigrams, according to the LDA model. Bigrams are ordered according to a hierarchical clustering.

it to include expressions such as "out-of-distribution" and "novelty". Moreover, the popularity of some particular type of deep models such as Generative Adversarial Network (GANs) should be acknowledged: one should not only search for surveys about "deep learning" but also about GAN (see for instance [72]) or probably in the near future about Transformers [73].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Additional PCA representation

The PCA results presented in Section 3.4 can be complemented by an analysis of the third component (motivated by the scree plot on Fig. 2). As shown on Figs. A.12 and A.13, the third principal component (PC) plays a very similar role to the one of the second component, bringing some separation between papers with plagiarism and papers with minor contribution. While the second PC tends to oppose distances to closest papers to delay between publications and references, the third PC opposes the distances computed on the full text to delays and distances computed on the abstract. Distances play a stronger role in the second PC while the third one focuses more on the delay. This shows

more clearly that plagiarism detection can be detected in some situation based on the freshness of the references which cannot be hidden by a simple rephrasing of the paper. Overall the third component confirms that our selection process aligns with simple characteristics of the papers.

Appendix B. Additional LDA illustrations

Fig. B.14 contains an illustration of the correlation matrix computed for the most salient 0.1% bigrams in the final corpus, and according to the LDA model outputs. Correlations are measuring the cosine-similarity on the bigram profiles, as given by their relative frequencies within each topic. Within the matrix, bigrams are ordered according to a hierarchical clustering. As one may easily see, bigrams are naturally grouped into blocks with similar contents.

References

[1] F. Edgeworth, On discordant observations, London, Edinb. Dublin Philos. Mag. J. Sci. 23 (143) (1887) 364–375, http://dx.doi.org/10.1080/14786448708628471.

[2] D. Hawkins, Identification of Outliers, in: Monographs on Applied Probability and Statistics, Springer Dordrecht, 1980, http://dx.doi.org/10.1007/978-94-015-3994-4.

[3] M. Markou, S. Singh, Novelty detection: A review—Part 1: Statistical approaches, Signal Process. 83 (12) (2003) 2481–2497, http://dx.doi.org/10.1016/j.sigpro.2003.07.018, URL https://www.sciencedirect.com/science/article/pii/S0165168403002020.

[4] B. Kitchenham, Procedures for Performing Systematic Reviews, Join technical report KU TR/SE-0401 & NICTA TR 0400011T.1, Department of Computer Science, Keele University & National ICT Australia Ltd., 2004, URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5.

[5] M. Olteanu, F. Rossi, F. Yger, Challenges in anomaly and change point detection, in: 30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, (ESANN 2022), Bruges, Belgium, 2022, pp. 277–286, http://dx.doi.org/10.14428/esann/2022.ES2022-6.

[6] C.C. Aggarwal, Outlier Analysis, second ed., Springer, 2017, http://dx.doi.org/10.1007/978-3-319-47578-3.

[7] F.E. Grubbs, Procedures for detecting outlying observations in samples, Technometrics 11 (1) (1969) 1–21, http://dx.doi.org/10.1080/00401706.1969.10490657, URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657.

[8] V. Barnett, T. Lewis, Outliers in Statistical Data, second ed., John Wiley & Sons Ltd., 1978.

[9] R.J. Beckman, R.D. Cook, Outlier ... ...s, Technometrics 25 (2) (1983) 119–149, http://dx.doi.org/10.1080/00401706.1983.10487840.

[10] P.J. Huber, Robust Statistics, in: Wiley Series in Probability and Statistics, Wiley New York, 1981, http://dx.doi.org/10.1002/0471725250.

[11] P.J. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection, in: Wiley Series in Probability and Statistics, Wiley, 1987, pp. 1–335, http://dx.doi.org/10.1002/0471725382.

[12] D. Samariya, A. Thakkar, A comprehensive survey of anomaly detection algorithms, Ann. Data Sci. (2021) http://dx.doi.org/10.1007/s40745-021-00362-9, URL https://link.springer.com/article/10.1007/s40745-021-00362-9.

[13] L. Ruff, J.R. Kauffmann, R.A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T.G. Dietterich, K.-R. Müller, A unifying review of deep and shallow anomaly detection, Proc. IEEE 109 (5) (2021) 756–795, http://dx.doi.org/10.1109/JPROC.2021.3052449.

[14] R. Foorthuis, On the nature and types of anomalies: A review of deviations in data, Int. J. Data Sci. Anal. 12 (4) (2021) 297–331, http://dx.doi.org/10.1007/s41060-021-00265-1, URL https://link.springer.com/article/10.1007/s41060-021-00265-1.

[15] J.N. Miller, Tutorial review–outliers in experimental data and their treatment, Analyst 118 (1993) 455–461, http://dx.doi.org/10.1039/AN9931800455.

[16] A.S. Hadi, A.H.M.R. Imon, M. Werner, Detection of outliers, WIREs Comput. Stat. 1 (1) (2009) 57–70, http://dx.doi.org/10.1002/wics.6, URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.6.

[17] P.J. Huber, E.M. Ronchetti, Robust Statistics, second ed., in: Wiley Series in Probability and Statistics, Wiley New York, 2009, http://dx.doi.org/10.1002/9780470434697.

[18] M. Hubert, P.J. Rousseeuw, S. Van Aelst, Multivariate outlier detection and robustness, in: C. Rao, E. Wegman, J. Solka (Eds.), Data Mining and Data Visualization, in: Handbook of Statistics, vol. 24, Elsevier, 2005, pp. 263–302, http://dx.doi.org/10.1016/S0169-7161(04)24010-X, URL https://www.sciencedirect.com/science/article/pii/S016971610424010X.

[19] P.J. Rousseeuw, M. Hubert, Anomaly detection by robust statistics, WIREs Data Min. Knowl. Discov. 8 (2) (2018) e1236, http://dx.doi.org/10.1002/widm.1236, URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1236.

[20] L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author), Statist. Sci. 16 (3) (2001) 199–231, http://dx.doi.org/10.1214/ss/1009213726.

[21] A. Patcha, J.-M. Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends, Comput. Netw. 51 (12) (2007) 3448–3470, http://dx.doi.org/10.1016/j.comnet.2007.02.001, URL https://www.sciencedirect.com/science/article/pii/S138912860700062X.

[22] C.F. Tsai, Y.F. Hsu, C.Y. Lin, W.Y. Lin, Intrusion detection by machine learning: A review, Expert Syst. Appl. 36 (2009) 11994–12000.

[23] Y. Yu, A survey of anomaly intrusion detection techniques, J. Comput. Sci. Coll. 28 (1) (2012) 9–17.

[24] A.L. Buczak, E. Guven, A survey of data mining and machine learning methods for cyber security intrusion detection, IEEE Commun. Surv. Tutor. 18 (2) (2016) 1153–1176, http://dx.doi.org/10.1109/COMST.2015.2494502.

[25] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284, http://dx.doi.org/10.1109/TKDE.2008.239.

[26] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Prog. Artif. Intell. 5 (4) (2016) 221–232, http://dx.doi.org/10.1007/s13748-016-0094-0.

[27] Y. Ma, H. He, Imbalanced Learning: Foundations, Algorithms, and Applications, John Wiley & Sons, 2013, http://dx.doi.org/10.1002/9781118646106.

[28] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D.S. Weld, SPECTER: Document-level representation learning using citation-informed transformers, in: ACL, 2020.

[29] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (3) (2009) http://dx.doi.org/10.1145/1541880.1541882.

[30] J. Xi, Outlier detection algorithms in data mining, in: 2008 Second International Symposium on Intelligent Information Technology Application, Vol. 1, 2008, pp. 94–97, http://dx.doi.org/10.1109/IITA.2008.26.

[31] Y. Zhang, N. Meratnia, P. Havinga, A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets, Technical Report Paper P-NS/TR-CTIT-07-79, Centre for Telematics and Information Technology (CTIT), Netherlands, 2007, URL https://research.utwente.nl/en/publications/a-taxonomy-framework-for-unsupervised-outlier-detection-technique.

[32] V. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2) (2004) 85–126, http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9.

[33] M.I. Petrovskiy, Outlier detection algorithms in data mining systems, Program. Comput. Softw. 29 (4) (2003) 228–237, http://dx.doi.org/10.1023/A:1024974810270.

[34] M. Markou, S. Singh, Novelty detection: A review—Part 2: Neural network based approaches, Signal Process. 83 (12) (2003) 2499–2521, http://dx.doi.org/10.1016/j.sigpro.2003.07.019, URL https://www.sciencedirect.com/science/article/pii/S0165168403002032.

[35] I. Ben-Gal, Outlier detection, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer US, Boston, MA, 2005, pp. 131–146, http://dx.doi.org/10.1007/0-387-25465-X_7.

[36] M. Agyemang, K. Barker, R. Alhajj, A comprehensive survey of numeric and symbolic outlier mining techniques, Intell. Data Anal. 10 (2006) 521–538, http://dx.doi.org/10.3233/IDA-2006-10604, 6.

[37] D. Cousineau, S. Chartier, Outliers detection and treatment: A review, Int. J. Psychol. Res. 3 (1) (2010) 58–67, http://dx.doi.org/10.21500/20112084.844, URL https://revistas.usb.edu.co/index.php/IJPR/article/view/844.

[38] X. Su, C.L. Tsai, Outlier detection, WIREs Data Min. Knowl. Discov. 1 (3) (2011) 261–268, http://dx.doi.org/10.1002/widm.19, URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.19.

[39] A. Zimek, E. Schubert, H.P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat. Anal. Data Min.: ASA Data Sci. J. 5 (5) (2012) 363–387, http://dx.doi.org/10.1002/sam.11161, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11161.

[40] H. Aguinis, R.K. Gottfredson, H. Joo, Best-practice recommendations for defining, identifying, and handling outliers, Organ. Res. Methods 16 (2) (2013) 270–301, http://dx.doi.org/10.1177/1094428112470848.

[41] J. Zhang, Advancements of outlier detection: A survey, EAI Endorsed Trans. Scalable Inf. Syst. 1 (1) (2013) http://dx.doi.org/10.4108/trans.sis.2013.01-03.e2.

[42] M.A. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249, http://dx.doi.org/10.1016/j.sigpro.2013.12.026, URL https://www.sciencedirect.com/science/article/pii/S016516841300515X.

[43] A. Zimek, P. Filzmoser, There and back again: Outlier detection between statistical reasoning and data mining algorithms, WIREs Data Min. Knowl. Discov. 8 (6) (2018) e1280, http://dx.doi.org/10.1002/widm.1280, URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1280.

[44] N. Rokhman, A survey on mixed-attribute outlier detection methods, CommIT (Commun. Inf. Technol.) J. 13 (1) (2019) 39–44, http://dx.doi.org/10.21512/commit.v13i1.5558, URL https://journal.binus.ac.id/index.php/commit/article/view/5558.

[45] A. Taha, A.S. Hadi, Anomaly detection methods for categorical data: A review, ACM Comput. Surv. 52 (2) (2019) http://dx.doi.org/10.1145/3312739, URL https://dl.acm.org/doi/10.1145/3312739.

[46] H. Wang, M.J. Bah, M. Hammad, Progress in outlier detection techniques: A survey, IEEE Access 7 (2019) 107964–108000, http://dx.doi.org/10.1109/ACCESS.2019.2932769, URL https://ieeexplore.ieee.org/document/8786096.

[47] A. Boukerche, L. Zheng, O. Alfandi, Outlier detection: Methods, models, and classification, ACM Comput. Surv. 53 (3) (2020) http://dx.doi.org/10.1145/3381028, URL https://dl.acm.org/doi/10.1145/3381028.

[48] A. Carreño, I. Inza, J.A. Lozano, Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework, Artif. Intell. Rev. 53 (5) (2020) 3575–3594, http://dx.doi.org/10.1007/s10462-019-09771-y.

[49] S. Thudumu, P. Branch, J. Jin, J.J. Singh, A comprehensive survey of anomaly detection techniques for high dimensional big data, J. Big Data 7 (1) (2020) 42, http://dx.doi.org/10.1186/s40537-020-00320-x, URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00320-x.

[50] A.B. Nassif, M.A. Talib, Q. Nasir, F.M. Dakalbab, Machine learning for anomaly detection: A systematic review, IEEE Access 9 (2021) 78658–78700, http://dx.doi.org/10.1109/ACCESS.2021.3083060, URL https://ieeexplore.ieee.org/document/9439459.

[51] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M.H. Rohban, M. Sabokrou, A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges, Trans. Mach. Learn. Res. (2022) URL https://openreview.net/forum?id=aRtjVZvbpK.

[52] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735, arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

[53] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Vol. 2, NIPS '14, MIT Press, Cambridge, MA, USA, 2014, pp. 2672–2680.

[54] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022, http://dx.doi.org/10.1016/B978-0-12-411519-4.00006-9.

[55] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (suppl_1) (2004) 5228–5235, http://dx.doi.org/10.1073/pnas.0307752101.

[56] J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, A density-based method for adaptive LDA model selection, Neurocomputing 72 (7–9) (2009) 1775–1781, http://dx.doi.org/10.1016/j.neucom.2008.06.011.

[57] R. Arun, V. Suresh, C. Veni Madhavan, M. Narasimha Murthy, On finding the natural number of topics with latent dirichlet allocation: Some observations, in: Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14, Springer, 2010, pp. 391–402, http://dx.doi.org/10.1007/978-3-642-13657-3_43.

[58] R. Deveaud, E. SanJuan, P. Bellot, Accurate and effective latent concept modeling for ad hoc information retrieval, Document Numérique 17 (1) (2014) 61–84, http://dx.doi.org/10.3166/dn.17.1.61-84.

[59] J. Chuang, C.D. Manning, J. Heer, Termite: Visualization techniques for assessing textual topic models, in: Proceedings of the International Working Conference on Advanced Visual Interfaces, 2012, pp. 74–77.

[60] K.T. Divya, N.S. Kumaran, Survey on outlier detection techniques using categorical data, Int. Res. J. Eng. Technol. (IRJET) 3 (12) (2016) 899–904.

[61] R.J. Durrant, A. Kabán, When is "nearest neighbour" meaningful: A converse theorem and implications, J. Complexity 25 (4) (2009) 385–397, http://dx.doi.org/10.1016/j.jco.2009.02.011, URL https://www.sciencedirect.com/science/article/pii/S0885064X09000260.

[62] S.S. Khan, M.G. Madden, One-class classification: taxonomy of study and review of techniques, Knowl. Eng. Rev. 29 (3) (2014) 345–374, http://dx.doi.org/10.1017/S026988891300043X.

[63] E. Panjei, L. Gruenwald, E. Leal, C. Nguyen, S. Silvia, A survey on outlier explanations, VLDB J. 31 (5) (2022) 977–1008, http://dx.doi.org/10.1007/s00778-021-00721-1.

[64] P.J. Rousseeuw, Least median of squares regression, J. Amer. Statist. Assoc. 79 (388) (1984) 871–880, http://dx.doi.org/10.1080/01621459.1984.10477105, URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477105.

[65] G.O. Campos, A. Zimek, J. Sander, R.J.G.B. Campello, B. Micenková, E. Schubert, I. Assent, M.E. Houle, On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, Data Min. Knowl. Discov. 30 (4) (2016) 891–927, http://dx.doi.org/10.1007/s10618-015-0444-8.

[66] S. Song, D. Shin, E. Yoon, Analysis of novelty detection properties of autoassociators, 2001, pp. 577–584, http://dx.doi.org/10.1016/B978-008044036-1/50068-8,

[67] S. Hawkins, H. He, G. Williams, R. Baxter, Outlier detection using replicator neural networks, in: Y. Kambayashi, W. Winiwarter, M. Arikawa (Eds.), Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 170–180.

[68] R. Chalapathy, S. Chawla, Deep learning for anomaly detection: A survey, 2019, arXiv preprint arXiv:1901.03407.

[69] G. Pang, C. Shen, L. Cao, A.V.D. Hengel, Deep learning for anomaly detection: A review, ACM Comput. Surv. 54 (2) (2021) http://dx.doi.org/10.1145/3439950.

[70] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, LOF: Identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00, Association for Computing Machinery, New York, NY, USA, 2000, pp. 93–104, http://dx.doi.org/10.1145/342009.335388.

[71] E. Schubert, A. Zimek, H.P. Kriegel, Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection, Data Min. Knowl. Discov. 28 (1) (2014) 190–237, http://dx.doi.org/10.1007/s10618-012-0300-z.

[72] X. Xia, X. Pan, N. Li, X. He, L. Ma, X. Zhang, N. Ding, GAN-based anomaly detection: A review, Neurocomputing 493 (2022) 497–535, http://dx.doi.org/10.1016/j.neucom.2021.12.093, URL https://www.sciencedirect.com/science/article/pii/S0925231221019482.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Madalina Olteanu is a Professor of applied mathematics and data science at Paris Dauphine University PSL, and member of CEREMADE research center. She (co)authored more than 50 papers published in international journals or peer-reviewed conferences. Her research interests are mainly focused on the study of spatial and temporal patterns in complex data, and particularly in humanities and social sciences. She uses and investigates techniques related to time series analysis, change-point detection, feature selection in unsupervised learning, and self-organizing maps.

Fabrice Rossi is a Professor of data science at Paris Dauphine University PSL. He is a member of the CEREMADE. He has (co) authored more than 150 peer reviewed research papers published in international journals and in proceedings of conferences. His research focuses on the science of data, more specifically on automatic learning and the exploratory analysis of data. In parallel with theoretical, methodological and industrial developments, he has built multidisciplinary collaborations, particularly in epidemiology, history, archaeology and sociology.

Florian Yger is an associate professor in computer science at Paris Dauphine University PSL, and member of the LAMSADE research center since 2015. He is a member of the Prairie (PaRis Artificial Intelligence Research InstitutE) where he holds a junior chair (Chaire Tremplin) and a visiting researcher at RIKEN AIP since 2017. He contributes to the problem of representation learning for images, times series and structured data. His work is applied to biomedical signals, image processing and cybersecurity.