

Toward a Perspectivist Turn in Ground Truthing for Predictive Computing

Valerio Basile¹, Federico Cabitza², Andrea Campagner², and Michael Fell¹

¹ Università di Torino, Turin, Italy

² Università di Milano-Bicocca, Milan, Italy

Abstract. Most Artificial Intelligence applications are based on supervised machine learning (ML), which ultimately grounds on manually annotated data. The annotation process is often performed in terms of a majority vote and this has been proved to be often problematic, as highlighted by recent studies on the evaluation of ML models. In this article we describe and advocate for a different paradigm, which we call *data perspectivism*, which moves away from traditional gold standard datasets, towards the adoption of methods that integrate the opinions and perspectives of the human subjects involved in the knowledge representation step of ML processes. Drawing on previous works which inspired our proposal we describe the potential of our proposal for not only the more subjective tasks (e.g. those related to human language) but also to tasks commonly understood as objective (e.g. medical decision making), and present the main advantages of adopting a perspectivist stance in ML, as well as possible disadvantages, and various ways in which such a stance can be implemented in practice. Finally, we share a set of recommendations and outline a research agenda to advance the perspectivist stance in ML.

Keywords: Machine Learning · Reliability · Artificial Intelligence · Multi-rater Labelling · Observer Variability

1 Motivations and Background

Much of contemporary Artificial Intelligence is based on supervised machine learning, a methodology that leverages large datasets that are manually “annotated”, that is, associated with categorical labels in a case-wise manner (see Table 1). The common and undisputed pipeline for the preparation of these datasets includes the definition of the classification schema (i.e., the eligible labels), data collection and annotation³, and, critically, label aggregation. This latter step is often performed

³ Since annotation, interpretation and rating are terms that can be used indiscriminately to indicate both the process (i.e. the labelling) and its effect or result (i.e., the labels), here we prefer to speak of phenomenon and objects (to be classified); which is associated with a judgement and some labels, as result of an interpretation act by some raters, although all the synonyms mentioned above are equally good as long as they do not generate notational ambiguity. In the same vein, we speak of rating and rater, where other authors could use the terms label and annotator, respectively. To help the reader with this terminological fuss, we produced Table 1.

in terms of a majority vote. This has been proved to be problematic in many work settings, especially when the object to classify is so complex that most of the raters can get it wrong and the real experts are a minority [3, 7, 9, 12, 13], or the object is ambiguous as it is often the case in Natural Language Processing (NLP) tasks [4]. Furthermore, we argue that the issues with aggregation are not restricted to highly subjective phenomena, but are also relevant to fields where the phenomena to be classify are traditionally considered “objective”, like in clinical decision support .

Table 1. Glossary of the main concepts discussed in this article.

term	synonyms and related concepts	definition
annotation	labelling, interpretation, rating, labelling, judgment	Either the process or the result of labelling (i.e., associating labels with) a set of objects.
ground truth	gold standard, label set, target set, also dataset from which training, validation and test sets are extracted	set of the labels to be associated with the training objects. Ground truthing is the process where such a set is created.
label	category, class, target value, judgment, code, symbol, token	symbol (or set of symbols) by which to annotate (label, classify) objects in the training set. It can be any term out of a list, a category, a code, a level of measurement on an ordinal scale, a value.
model	algorithm	mathematical representation used to classify new objects and predict unknown outcomes on the basis of input unlabelled objects. The output of a learner.
object	instance, case, phenomenon, record, row, example	symbolic representation of an object of the phenomenon or reality of interest to be properly classified or predicted by the ML model.
phenomenon	object	what is represented as data in each record and is to be classified, either in the training set (by raters) or as output of the predictive model.
rater	annotator, observer, labeller	who is involved in labelling (i.e. ground truthing), that is who is supposed to classify training data to give as input to machine learning procedures.
reduction	derivation	the procedure of transforming a multi-rater (multi-label) ground truth into a single-label one (with only one label for each object, case), e.g., by majority voting.
supervised (learning)	(learning) by example	type of machine learning where labeled dataset (i.e., training sets) is used to train models that classify data or predict outcomes.

We propose and argue for a paradigm shift, which could move away from monolithic, majority-aggregated gold standard datasets, towards the adoption of methods that more comprehensively and inclusively integrate the opinions and perspectives of the human subjects involved in the knowledge representation step of modeling processes. Our proposal comes with important and still-to-investigate implications: first, supervised models equipped with full, non-aggregated annotations have been reported to exhibit a better prediction capability [2, 14, 43], in virtue of a better representation of the phenomena of interest; secondly, new tech-

niques for AI explainability can be devised that describe the classifications of the model in terms of multiple and alternative (if not complementary) perspectives [7, 36]; finally, we should consider the ethical implications of the above mentioned shift and its impact on cognitive computing, whereas the new generation of models can give voice to, and express, a diversity of perspectives, rather than being a mere reflection of the majority [36, 39].

2 Strong and weak data perspectivism

As anticipated above, in this paper we propose what we denote as a *perspectivist* approach in producing the ground truth (i.e., the ground truthing task) to be used in supervised classification tasks by systems developed with Machine Learning (ML) methods and techniques. This general stance can be articulated in two main versions, which could be connoted as either a *weak* or *strong* approach (see Figure 1).

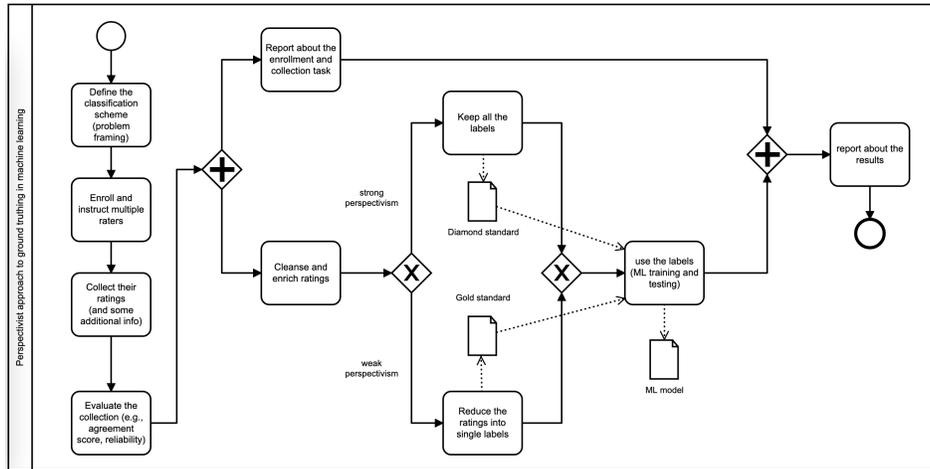


Fig. 1. A BPMN (Business Process Model and Notation) diagram of the ground truthing process in a perspectivist setting. Many tasks are common with more traditional ML pipelines, but the core distinction with these latter ones lies in the exclusive gateway (X) in the centre of the diagram. Parallel gateways (+) indicate opportunities for parallel activities, which we made explicit to emphasize the importance of the comprehensive reporting of the ground truthing process.

A *weak perspectivist* approach is adopted when researchers involved in ground truthing are not content to collect a single label for each object to be classified, that is to produce a *gold standard* label set; but rather aim to collect as many raters *and* annotations as possible, i.e., to build what in [14] has been called *diamond standard* (see Figure 1). We emphasize the distinction between raters and annotations because, as simply as it can be, raters could express more than

one label for a given object to classify [18, 31] (also as a way to express their indecision in case of strictly alternative labels), or they could expressly be asked to rank available labels in terms of pertinence or to associate each class with a confidence/probability level [16].

One could rightly wonder why ML researchers would want to collect such redundant information about the phenomena about which they design and develop decision support systems that are usually aimed at improving human decision making by proposing the one best label for each object to classify [3]. For now, we are not going to dispute that conceiving the output of such systems as single labels is the best way to improve human decision making⁴; even when the output must be a single label, ML researchers could want to collect multiple labels for single objects to classify when they deem relying on single judgments either too *limiting* (like in multi-label tasks, where labelling categories are not disjoint and they overlap to some extent) or too *error-prone* [47]. This could happen for two main reasons: 1) distrust in the raters involved, as it can be the case in crowdsourcing initiatives or questionnaire-driven studies where researchers cannot oversee the annotation task or ensure its accuracy [20, 47]; 2) the recognized variability (what later we will call their *inter-subjective* nature) of the phenomenon of interest [3], that is the recognition that different raters could classify the same object differently, and not necessarily because they are wrong. Indeed, ratings can differ not only because the raters are fallible, but also for a number of other factors, among which we recall: the intrinsic ambiguity and complexity of the phenomenon, including the so-called *cumulative mess*, that is the condition when the same object can legitimately be classified as many things at the same time [8]; the (in)stability of the phenomenon over time [20]; the complexity of the task, also in terms of number of distinct states or configurations of the phenomenon to detect [40] and of the concentration that is requested to the raters, and their necessary proficiency to detect and understand the phenomenon [26]; the raters' susceptibility to somehow systematic cognitive biases, both at individual and team level, like overconfidence, confirmation and availability bias, anchoring and halo effects [19], as well as to more contingent and context-driven external factors (what Kahneman and colleagues have denoted as *chance variability of judgments*, or "decision noise" [28]).

Thus, collecting multiple labels from a sample of individuals can help ML researcher get a sample of perceptions, opinions and judgments that could be maximally *representative* of the population of interest; also, and more practically speaking, doing so would help them draw from such a sample some *qualified* majority [12] for the sake of higher accuracy⁵.

Weak perspectivism would then advocate the production of gold standards by considering and taking into account multiple perspectives. This does not neces-

⁴ As a matter of fact, alternative ways, like in conformal prediction, could improve it more [5], especially if we consider improvement also beyond the mere dimension of error rate

⁵ With qualified majority we intend either a *statistically significant* majority (with respect to some hypothesis testing procedure), or an *overwhelming* majority, regardless of how this may be defined, and thus irrespective of what perspectives are considered and how they are distributed within the sample

sarily require to make each perspective symbolically explicit, in terms of multiple labels: a meeting where multiple experts are invited to share their opinion about an object would be perspectivist as long as all people involved could express their opinions and views in a discussion, which could then be summarized into single positions. Likewise, a weak perspectivist approach would require to collect multiple labels from a corresponding number of observers or raters but then would combine these labels and select one single label for each object to be annotated, mostly by some kind of majority voting (e.g., weighted voting). This process has been called *perspective reduction* in [14].

On the other hand, we would speak of *strong perspectivism* whenever the researchers' aim is to collect multiple labels, or multiple data about each class, about a specific object, and *keep them* all in the subsequent phases of training or benchmarking of the classification models. Doing so certainly impacts model training and evaluation, but can be realized in several ways, of varying complexity [14, 43, 46]. The easiest way, that is the most backward-compatible way that does not require ad-hoc implementations, is to replicate each object in the training set to reflect the number of times this object has been associated with a certain label by the raters [51]; nonetheless, also other methods have been proposed in the literature, that we will better describe in Section 4.

3 Between objectivity and subjectivity

From a conceptual point of view, the process of ground truthing can be understood as a process of collection of a set of *measures*, according to the theory of scales of measurement [17]. In this light, any mapping from a phenomenon or object of interest (also a portion of the reality) to a sign (mostly a symbol of conventional meaning) can be assimilated to a measurement; the related measure (symbol) that can be used to annotate its data representation is but the result of the process above that can be called in multiple, yet equivalent, ways: rating, judgement, perception, opinion, annotation.

As we said above, a perspectivist approach can be advocated especially for the annotation of those phenomena that we would consider as “highly subjective”, from a common sense stance, that is affected by high *observer variability*, as measured with some inter-rater agreement metrics, like kappa [21], alpha [30] or rho [12]. According to a common definition, a judgment is considered *subjective* when it is mainly “based on, or influenced by, personal feelings, tastes, or opinions.”; we usually contrast this concept with that of *objective*, a term that characterises judgements that, ideally, are not influenced by personal feelings or idiosyncrasies and which, on a practical level, the vast majority of people would see and label in the same way (barring clear errors or oversights).

In metrology, the science of measurement, the above terms are associated with slightly different meanings: A recent proposal [34] asserts that any measurement that aims to be dependable (as ground truth must be, by definition) should exhibit both the characteristics of *objectivity*, that is being related to an object, even to a “non-physical property such as chess playing ability”; and that of being *interper-*

sonal, that is independent of the subject or “identically interpretable by different measurers” [34].

In the perspectivist discourse, we recognize that the opposition obj/subjective is somehow too much affected by common sense and, worse yet, value judgments (in that subjective usually means “bad” or, to our practical aims, unreliable; and objective means good, and reliable). Therefore, we prefer to consider these two concepts as sort of ideal extremes of a full range of ways in which phenomena can present themselves to the interpretation and judgement of the members of a human community, what we call the *intersubjectivity spectrum* (see Figure 2).

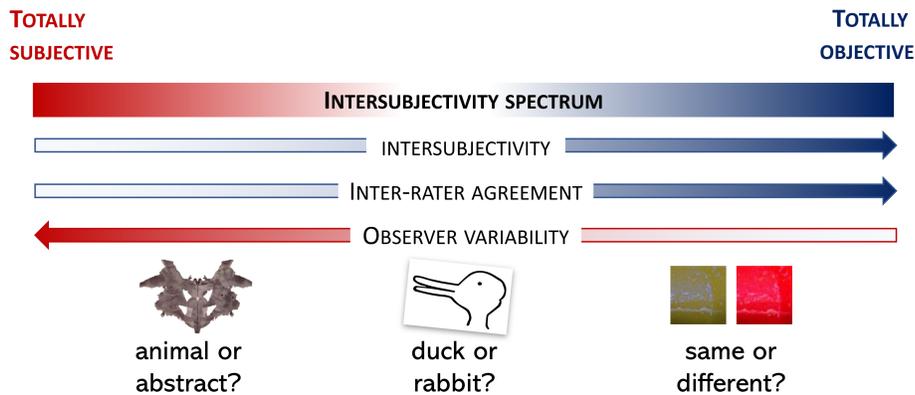


Fig. 2. The intersubjectivity spectrum and its relationship with inter-rater agreement.

In this range, the extremes are essentially abstract: no phenomenon, and hence opinion on it, can be said totally subjective, as human beings are all immersed in a culture and social ensemble (trivially, they may share a language and semantic conventions), even if we recognized the ineffable and undetermined nature of the phenomenon at hand; likewise, full objectivity is a mirage because any judgement (or measurement) is situated and affected by contingent and contextual conditions, even if executed by a sensor or machine.

Appraising where a certain phenomenon of interest can be located in this conceptual continuous range, that is assessing its level of intersubjectivity, can be preparatory to understand what kind of perspectivist approach, whether weak or strong, to adopt in one’s ML pipeline; and whether keeping multiple labels allows to preserve individual expression and intuition or, conversely, can undermine the reliability (and hence utility) of the ground truth. To this respect, highly intersubjective phenomena would be phenomena that are close to the objective end of the intersubjective spectrum and for which disagreement among the raters involved is limited or absent. On the other hand, low-intersubjectivity objects would be those for which the complexity and ambiguity of the phenomenon at hand, but also the experience, commitment and competence that are required of the involved raters (like in medical diagnosis) would make one expect low agreement scores [40]. This

can happen also in unsuspected domains, like medical diagnosis: just as a few examples, agreement in the grading of ovarian cancer slides has been found associated with kappa scores lower than 30% ($\kappa = .25$ [6]), like in ischemia detection from EEGs ($\kappa = .29$ [35]) and in phase lag detection in ECGs ($\kappa = .23$ [22]). Such a low agreement is anything but rare in classification tasks (even in medicine); being aware of this can call for the application of some advanced methods, where the single contributions collected are weighted so as to identify the most reliable ones in the subsequent phases of ML development.

4 Related work

In what follows, we report on the main research contributions that either inspired or adopt some form of perspectivism in the ML pipeline (see Figure 1). Annotation is a crossing task that regard different types of data, including texts (and any part of them), images (including segmented portions), and more or less structured data, like case records.

The scientific communities interested in ML, like NLP, Computer Vision (CV) and medical informatics, have traditionally relied on gold standard datasets to design, develop and evaluate supervised models: these datasets have usually been obtained by the annotation of a single rater or by means of the majority aggregation of few raters, and their reliability, and representativeness of the real task under consideration, have scarcely been questioned.

In the recent years, however, due to the huge increase in raw data availability, the increasing reliance on crowdsourcing and similar annotation protocols has highlighted the issue of observer variability in Machine Learning tasks [3, 13], an issue which was already well known in certain settings such as the computational linguistics [4] and medical ones [27, 33]: Beyer et al. [9] and Yun et al. [50] showed the majority-aggregated labels in the original ImageNet dataset to be not representative of the images in the dataset due to observer variability and the un-reliability of the annotation process; Svensson et al [44] noted the influence of observer variability on the performance of Machine Learning in a task of cancer detection and proposed ways to measure model performance in settings affected by variability; similarly, the impact of observer variability in NLP has been explored by Akhtar et al. [1, 2] in the task of hate speech detection.

While many works, and specifically those focusing on the crowdsourced learning setting, have adopted a weak perspectivist stance for the development of ML methods able to account for this observer variability [15, 24, 41, 45]; the need for ML methods explicitly taking into account a strong perspectivist approach, however, has only recently started to become a focus of research: Akhtar et al. [2] showed that a strong perspectivist approach to model training may also lead to performance increase; Sudre et al. [43] and Guan et al. [23] proposed a multi-task approach to deal with observer variability in medical imaging, showing how jointly learning the majority consensus and the individual raters' labels improves classification accuracy; Sharmanska et al [42] showed how accounting for disagreements among raters may help improving the performance in CV tasks with crowdsourced

labels; Peterson et al. [37] showed that accounting for raters’ disagreement and uncertainty may lead to generalizability and performance improvements in CV tasks; Plank et al. [38] proposed a loss based on inter-rater agreement metrics to address variability in part-of-speech tagging; Uma et al. [46] proposed the use of soft losses as a perspectivist approach for the training of ML models; Campagner et al [14] proposed a soft loss ensemble learning method, inspired by possibility theory and three-way decisions, for the training of ML models in perspectivist settings.

In a similar direction, recent works [7, 12, 39] explicitly explored the impact of strong perspectivism on the development and evaluation of supervised models, also from a more conceptual perspective. In particular, in [7, 12] experiments are presented in support of the thesis that disagreement in annotation may come from the subjectivity (i.e., low inter-subjectivity) of a task to varying extent [40], and therefore it should not be cast away as noise in the data, but rather it should be systematically accounted for at evaluation time. Furthermore, in [7, 39] the authors also show an additional advantage of strong perspectivism in supervised learning, namely its potential impact on the interpretability and fairness of the models. In an experiment on real data (an annotated corpus of hate speech), the authors of [7] show how individual labels can be used to cluster the raters by affinity, leading to the emergence of patterns that helps identifying socio-demographic aspects of the raters themselves, which are in principle opaque, especially in a crowdsourcing scenario; while in [39], the authors describe how a similar approach could be used to detect biases in the data and labels provided by raters. The related works, and the main ML techniques for perspectivist ground truths, are summarized in Table 2.

Table 2. A summary of methods proposed in the literature to train ML models based on perspectivist ground truths.

Method	Type	References	Description
Qualified Majority	Weak	[12, 28]	Collect the annotations of at least 12 raters; if there is a qualified (i.e., statistically significant) majority keep the majority label, otherwise discard the object
Object Replication/Weighting	Weak	[1, 38, 42, 51]	Replicate or weight each object in the training set depending on the number of labels provided by the raters, or on some measure of disagreement
Consensus Labeling	Weak	[15, 41]	Use a participatory consensus building process (e.g. Delphi method) to annotate the objects
Weighted Majority	Weak	[12, 24, 45]	Score the raters based on an estimate of their reliability; compute a weighted majority to annotate the object
Regularized Multi-task Learning	Strong	[23, 43]	Keep all labels; train a model that jointly predicts the majority label and all the individual raters’ labels
Ensemble Methods	Strong	[2, 14, 44, 51]	Sample multiple datasets from the original multi-rater one, according to some criteria, e.g., the distribution of labels, or the labelling rater; train an ensemble model encompassing several base models, one for each sampled dataset
Soft Loss Learning	Strong	[14, 37, 38, 46]	Associate each object in the training with a distribution, or set, over the labels provided by the raters; train a model using a modified loss function

5 Looking at the two sides of the same coin

As anticipated above, a perspectivist approach to ground truthing requires to preserve the classification multiplicity instead of getting rid of it by majority voting (if original labels have been produced) or consensus surveys (if original labels do not exist). Obviously, as also highlighted in Section 4 when discussing the related works setting the background for our proposal, this comes with some advantages and also some shortcomings, which we discuss in what follows.

The main benefits of the perspectivist approach are:

1. To provide a theoretical backbone that recognizes and accepts the categorical irreducibility of some phenomena. This is especially relevant to those phenomena which exhibit a natural ambiguity, such as many tasks in NLP [4], or seemingly inconsistent clinical manifestations [13];
2. To extract valuable knowledge from what it is usually discarded as noise (cf. label noise [28]), i.e., disagreement. Such extra information is valuable for a decision support to be more useful in border-line and complex cases;
3. To avoid to ratify and legitimize the opinion of raters belonging to a majority group, re-iterating their truth in seemingly objective advice. Instead, the perspectivist approach aims at giving voice to the few who hold a minority view [36], or to those who are intimidated in collective debates;
4. To be able to build models that learn typical human error patterns (if it is plausible to define “errors” on the basis of minority stances) and use this information as a form of decision support;
5. To be able to develop models that can leverage label-uncertain [46] and fuzzy data [14] to improve performance, generalizability and robustness;
6. To allow for three-way, fuzzy, probabilistic methods that represent the subjectivity and uncertainty in the considered phenomena [13, 29], and provide decision makers with useful advice, that is methods that can improve trust, enhance user experience, and possibly mitigate the risk of automation bias and deskilling.

Since there is no rose without thorns, here we enumerate the main shortcomings that are associated with a perspectivist approach to supervised ML.

1. Need to involve multiple raters: this may represent an important bottleneck in terms of costs or time, and may thus result impractical or expensive in some domain (e.g., medicine) or when dealing with large datasets;
2. Incompatibility with standard ML approaches, which are usually not designed to take into account multiple perspectives or annotation, and need to design ad-hoc ML methods. While certain classes of learning algorithms (e.g. multi-label ones) may be able to handle multiple labels, it is not clear if these methods can be proficuously applied in the perspectivist ground truthing setting;
3. More complex validation/evaluation, due to the absence of a uniquely defined ground truth. While in some cases majority labels can be used as a benchmark ground truth [44], these may not be appropriate in strongly subjective or ambiguous settings.

6 Recommendations and a research agenda

This article aims to disseminate a renovated interest for an alternative approach to ground truthing with respect to the “reductionist” one where multiple ratings collected about a single object are reduced into single labels. As we saw in the previous section, Section 5, this approach entails both advantages and challenges, posed by the will to cope with information richness and manage complexity (also in terms of redundancy, uncertainty and inconsistency) instead of getting rid of it, in light of the research that we presented in Section 4. In this Section, in lieu of a conclusion, we proceed with two sections that shed light on the future: we present a set of agile recommendations for those willing to adopt a perspectivist approach to their ground truthing tasks; and then we propose a number of possible proposals calling for further research and contributions.

6.1 Recommendations

In what follows, we share some recommendations to embrace a perspectivist stance in ground truthing. While the impact of some of these practices must still be soundly evaluated, we also mention some of the main studies providing preliminary evidence supporting the recommendations, when available.

- Design annotation schemes that allow raters to associate objects with multiple labels, or also with a ‘none of these’ label, to account for multiple perspectives directly acknowledged by the single raters. Moreover, allow the raters to express a judgment of inadequacy of the available label set (see [3]);
- Involve *enough* raters. This can mean different things depending on the application domain: a number that allows for statistically significant majorities to emerge (e.g., at least 12 raters for dichotomous tasks) or a number of raters that allows to create a majority of superhuman accuracy, according to some estimate of the average accuracy of the raters (see [12, 28]);
- Involve heterogeneous raters, both in regard to origin and culture and to expertise and skills: different opinions are not always a source of noise, as asserted in [28], but rather of richness [3, 48].
- Be aware of the limits of majority voting, especially for complex objects to classify, like in the medical domain. Collecting more information about each judgement, like the rater’s confidence, should become common practice to help researchers detect these cases, and enable alternative *reductions* [14] with respect to simple majority, like weighted majority or any strong version of data perspectivism (see Table 2).
- Evaluate ML models also with respect to *robustness*, or their capability to adequately perform also on external datasets (that is data coming from settings other than where the training data were produced). If performance relevantly degrades on external datasets, this could suggest to adopt a perspectivist approach as a way to mitigate the risk that the models “overfit” the ratings of a non-representative sample of users.
- Report about the rater enrollment process and about the quality of their ratings in detail. In [11] the authors recommend to describe the process of ground

- truthing in terms of: a) Number of raters involved to produce the labels; b) The raters' profession and expertise (e.g., years from specialization or graduation); c) The incentive provided, if any; d) Particular instructions given to raters for quality control (e.g., which data were discarded and why); e) how long the labelling process took and, in the case of critical domains such as medicine and law, where and under what conditions it took place (e.g., controlled conditions, real-world interruptions); f) Any chance-adjusted measure of inter-rater agreement (e.g., Rho [12], Alpha [30], Kappa [21]); g) The Labelling technique (e.g., majority voting, Delphi method [32], consensus iteration).
- Collect additional information from the raters involved, so as to take actual account of their perspective and way of seeing the objects at hand. For instance, in [12] we recommended to collect the confidence expressed by the rater about each of their annotations in terms of an ordinal score, as well as other dimensions (e.g., complexity, difficulty, rarity, relevance) along a similar scale. This information can be used to detect the cases for which the raters' confidence was the lowest, thus suggesting that they are the most difficult ones to decide about; or those cases that, for their relevance, it is important that the ML model gets right (i.e., with a sufficient average accuracy) not to mislead its users.

These recommendations are complementary to those proposed by the *Perspectivist Data Manifesto*⁶, a collaborative initiative to promote a perspectivist research agenda in the AI community. In particular, the signatories of the above manifesto also recommend to create and distribute non-aggregated datasets, in order to foster the discussion on the principles of data perspectivism among the research community and to facilitate experimental research in this direction.

6.2 Open problems for a possible research agenda

Finally, in what follows we delineate some possible open problems and research directions that we deem relevant to advance the perspectivist stance in predictive computing:

1. First, and in connection with the manifesto mentioned in the previous section, we recommend the creation and dissemination of benchmark datasets that could be used to evaluate perspectivist ML models, possibly with respect to different data types (like images, texts, structured data) and for different classification tasks (e.g., detection, risk stratification, forecasting). Such datasets are necessary for the evaluation of novel algorithmic proposals, and they could be used also as benchmarks for setting up challenges, which have recently proven to be important drivers for the development of novel techniques and approaches;
2. Disagreement (and possibly errors) in the multi-rater labels are part and parcel of the perspectivist stance to ground truthing. It is of interest to develop techniques and approaches that are able to exploit the multiple labels to understand and model how the raters err, or on which types of objects they

⁶ <https://pdai.info>

disagree more, so as to develop learning models with the ability to predict the chance that the raters would err on a new object;

3. The usual assumption in ML development is that a model that learnt the majority label from an ensemble of raters could actually provide a better prediction than the majority prediction of an ensemble of models that learnt the single raters' labels. Further research could address this conjecture and investigate if good results (that is useful and reliable predictions) could be achieved also by preserving multiplicity at ground truth level and reducing variability/uncertainty at prediction level through some computational technique.
4. The perspectivist stance in ground truthing amounts to associating multiple labels to each object. However, especially in low-intersubjectivity tasks, a certain amount of divergent labels is expected in, and may be intrinsic to, the task, and a part of these could be due to errors, inattention or negligence. Thus, it would be interesting to develop ML models that are effectively able to disentangle subjectivity from error, and also to characterize (from a learning theoretic perspective) when this disentanglement is possible at all;
5. In the literature, different algorithmic approaches able to account for a perspectivist ground truth have been proposed (see Table 2): these include data augmentation strategies based on the replication of objects associated with multiple labels, ensemble learning methods [14], or soft loss approaches [46]. It would be interesting to assess, on real-world problems and applications, the performance of these (and other) approaches, so as to understand their properties and the appropriateness of different algorithms for specific tasks;
6. While some ways to evaluate such ML models have been proposed in the literature (e.g., by evaluating the performance of ML models on objects associated with “overwhelming majorities” [14, 44], or by adopting soft loss metrics [46]), these proposals should be unified to develop novel metrics that can better take into account the (chance-adjusted) agreement rate between the raters, or the presence of chance effects and label noise;
7. Similarly to the above point, we need further research on the impact of observer variability on the predictive performance of ML models. Simulating this kind of variability would allow to test for the robustness of these models and get an estimate of the extent the model is “overfitting” on either opinions that are not representative of the user population or on partial aspects of the phenomenon of interest.
8. The main concept underlying data perspectivism in ground truthing is that collecting multiple opinions and perceptions about a single phenomenon or object to classify – and preserving this multiplicity in ML pipelines – is valuable, and can convey value to decision makers and users at prediction time through apt decision support [7, 14]. However, a dataset characterized by multiple perspectives, that is by some relevant extent of disagreement between the raters, could be seen as rich as it is noisy. While standard measures of inter-rater agreement (or similar measures, like entropy) can capture this “noise” aspect within a *diamond standard*, these measures are not fully capable to represent the informational content and value of such multi-faceted information [14]. Therefore, this approach needs the definition of a theoretical framework (also in terms of the definition of suitable metrics) by which to evaluate the com-

promise between richness (in that the more different perspectives, the better) and reliability (in that multiplicity does not indicate confusion but complementarity): though we recognize this as a faint approximation of this concept, such an approach could in principle be based on some apt combination of an inter-rater agreement and an information-theoretic (e.g. conditional entropy, or some variants of entropy for imprecise probabilities [49]) measure. In doing so, learning-theoretic characterizations of perspectivist ML can be developed, which could enable to understand when a given dataset can be reliable “ground” for sound decision support; or, conversely, when its quality needs to be improved with some intervention of *decision hygiene* [28], like, e.g., aggregating judgements with methods that leverage professional expertise;

9. As previously described in [7, 39] (see also Section 4), the perspectivist stance could also have potential applications and impacts in terms of model interpretability and algorithmic fairness, e.g. by enabling the modeling and detection of biases and discrimination induced by (some groups of) raters. Further research should thus be devoted toward the investigation of possible applications of perspectivist ML in eXplainable and fair AI;
10. Finally, we believe it would be interesting to consider settings in which raters are able to express more information than a single label [3] – for example, by providing a ranking of the possible labels or expressing their confidence in the labels that they propose [12]. Given the similarity of these settings to the problems typically investigated in the field of computational social choice [10] further research should investigate the approaches proposed in that context, and how they could be applied to design perspectivist ML methods that could deal with more general, structured labeling representations.

To conclude, the perspectivist approach is essentially aimed at *caring about* the representativeness and reliability of the ground truth of ML systems. More specifically, and programmatically, data perspectivism fosters wariness toward aggregated gold standards: these reference datasets express single-truth assumptions that can fall short of capturing the necessary complexity of the phenomena for which we want to have support from computational means [8]. After all, ML systems are complicated machines that essentially reiterate past judgments and legitimate them [25] by putting the perceptions of very few (relatively speaking) raters to the attention of countless users and decision makers. There is no guarantee that new objects will be comparable to (or equatable with) those with which such models had been trained, not to speak of the contingent context. We believe that adapting the single-truth assumption of ML to the perspectivist paradigm is not only more fair towards minority opinions, but it also has the potential to yield a more accurate and explainable quantitative evaluation of the trained models, as shown by the recent works that are surfacing exploring this research direction [46, 7]. Such a stance creates the necessary room for asking questions such as *whose opinion is the model relying on in its prediction?*, and *what opinions do we want to project into the interpretation of the unexpected new?*; and for reflecting on the implications of these matters on the ethical nature and impact of the decisions made with the support of predictive models.

References

1. Akhtar, S., Basile, V., Patti, V.: A new measure of polarization in the annotation of hate speech. In: Alviano, M., Greco, G., Scarcello, F. (eds.) *AI*IA 2019 – Advances in Artificial Intelligence*. pp. 588–603. Springer International Publishing, Cham (2019)
2. Akhtar, S., Basile, V., Patti, V.: Modeling annotator perspective and polarized opinions to improve hate speech detection. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. vol. 8, pp. 151–154 (2020)
3. Aroyo, L., Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* **36**(1), 15–24 (2015)
4. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. *Computational Linguistics* **34**(4), 555–596 (2008)
5. Balasubramanian, V., Ho, S.S., Vovk, V.: *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes (2014)
6. Barnard, M.E., Pyden, A., Rice, M.S., Linares, M., Tworoger, S.S., Howitt, B.E., Meserve, E.E., Hecht, J.L.: Inter-pathologist and pathology report agreement for ovarian tumor characteristics in the nurses’ health studies. *Gynecologic oncology* **150**(3), 521–526 (2018)
7. Basile, V.: It’s the end of the gold standard as we know it. leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In: Baldoni, M., Bandini, S. (eds.) *AIxIA 2020: Advances in Artificial Intelligence*. XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 24-27, 2020, Revised and Selected papers, vol. 12414. Springer Nature Switzerland AG (2021)
8. Bechmann, A., Bowker, G.C.: Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society* **6**(1), 2053951718819569 (2019)
9. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? *arXiv preprint arXiv:2006.07159* (2020)
10. Brandt, F., Conitzer, V., Endriss, U., Lang, J., Procaccia, A.D.: *Handbook of computational social choice*. Cambridge University Press (2016)
11. Cabitza, F., Campagner, A.: The need to separate the wheat from the chaff in medical informatics. *International Journal of Medical Informatics* **152** (2021)
12. Cabitza, F., Campagner, A., Sconfienza, L.M.: As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai. *BMC Medical Informatics and Decision Making* **20**(1), 1–21 (2020)
13. Cabitza, F., Locoro, A., Alderighi, C., Rasoini, R., Compagnone, D., Berjano, P.: The elephant in the record: on the multiplicity of data recording work. *Health informatics journal* **25**(3), 475–490 (2019)
14. Campagner, A., Ciucci, D., Svensson, C.M., Figge, M.T., Cabitza, F.: Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences* **545**, 771–790 (2021)
15. Chang, J.C., Amershi, S., Kamar, E.: Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. pp. 2334–2346 (2017)
16. Chen, X., Bennett, P.N., Collins-Thompson, K., Horvitz, E.: Pairwise ranking aggregation in a crowdsourced setting. In: *Proceedings of the sixth ACM international conference on Web search and data mining*. pp. 193–202 (2013)
17. Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., Tyrer, P.: Rating scales, scales of measurement, issues of reliability: resolving some critical issues for clinicians and researchers. *The Journal of nervous and mental disease* **194**(8), 557–564 (2006)

18. Dumitrache, A., Aroyo, L., Welty, C.: Capturing ambiguity in crowdsourcing frame disambiguation. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 6 (2018)
19. Eickhoff, C.: Cognitive biases in crowdsourcing. In: Proceedings of the eleventh ACM international conference on web search and data mining. pp. 162–170 (2018)
20. Eickhoff, C., de Vries, A.P.: Increasing cheat robustness of crowdsourcing tasks. Information retrieval **16**(2), 121–137 (2013)
21. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
22. Foreman, B., Mahulikar, A., Tadi, P., Claassen, J., Szaflarski, J., Halford, J.J., Dean, B.C., Kaplan, P.W., Hirsch, L.J., LaRoche, S., et al.: Generalized periodic discharges and ‘triphasic waves’: a blinded evaluation of inter-rater agreement and clinical significance. Clinical Neurophysiology **127**(2), 1073–1080 (2016)
23. Guan, M., Gulshan, V., Dai, A., Hinton, G.: Who said what: Modeling individual labelers improves classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
24. Heinecke, S., Reyzin, L.: Crowdsourced PAC Learning under Classification Noise. In: Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing. vol. 7, pp. 41–49. AAAI (2019)
25. Hildebrandt, M.: The issue of bias. the framing powers of ml. In: editor, T. (ed.) Machine Learning and Society: Impact, Trust, Transparency. MIT Press (2021)
26. Jagabathula, S., Subramanian, L., Venkataraman, A.: Identifying unreliable and adversarial workers in crowdsourced labeling tasks. The Journal of Machine Learning Research **18**(1), 3233–3299 (2017)
27. Jewett, M.A., Bombardier, C., Caron, D., Ryan, M.R., Gray, R.R., Louis, E.L.S., Witchell, S.J., Kumra, S., Psihramis, K.E.: Potential for inter-observer and intra-observer variability in x-ray review to establish stone-free rates after lithotripsy. The Journal of urology **147**(3), 559–562 (1992)
28. Kahneman, D., Rosenfield, A., Gandhi, L., Blaser, T.: Noise: How to overcome the high, hidden cost of inconsistent decision making. Harvard Business Review **94**, 38–46 (2016)
29. Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Medicine **4**(1), 1–6 (2021)
30. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications (2018)
31. Li, S.Y., Jiang, Y., Chawla, N.V., Zhou, Z.H.: Multi-label learning from crowds. IEEE Transactions on Knowledge and Data Engineering **31**(7), 1369–1382 (2018)
32. Linstone, H.A., Turoff, M.: The delphi method: Techniques and applications. In: The Delphi method: Techniques and applications. Addison-Wesley Publishing (1975)
33. Lok, C.E., Morgan, C.D., Ranganathan, N.: The accuracy and interobserver agreement in detecting the ‘gallop sounds’ by cardiac auscultation. Chest **114**(5), 1283–1288 (1998)
34. Maul, A., Mari, L., Wilson, M.: Intersubjectivity of measurement across the sciences. Measurement **131**, 764–770 (2019)
35. Mehta, S., Granton, J., Lapinsky, S.E., Newton, G., Bandayrel, K., Little, A., Siau, C., Cook, D.J., Ayers, D., Singer, J., et al.: Agreement in electrocardiogram interpretation in patients with septic shock. Read Online: Critical Care Medicine— Society of Critical Care Medicine **39**(9), 2080–2086 (2011)
36. Noble, J.A.: Minority voices of crowdsourcing: Why we should pay attention to every member of the crowd. In: proceedings of the ACM 2012 conference on computer supported cooperative work companion. pp. 179–182 (2012)

37. Peterson, J.C., Battleday, R.M., Griffiths, T.L., Russakovsky, O.: Human uncertainty makes classification more robust. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9617–9626 (2019)
38. Plank, B., Hovy, D., Søgaard, A.: Learning part-of-speech taggers with inter-annotator agreement loss. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 742–751 (2014)
39. Rizos, G., Schuller, B.W.: Average jane, where art thou?—recent avenues in efficient machine learning under subjectivity uncertainty. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 42–55. Springer (2020)
40. Salminen, J.O., Al-Merekhi, H.A., Dey, P., Jansen, B.J.: Inter-rater agreement for social computing studies. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 80–87. IEEE (2018)
41. Schaekermann, M., Goh, J., Larson, K., Law, E.: Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW), 1–19 (2018)
42. Sharmanska, V., Hernández-Lobato, D., Miguel Hernandez-Lobato, J., Quadrianto, N.: Ambiguity helps: Classification with disagreements in crowdsourced annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2194–2202 (2016)
43. Sudre, C.H., Anson, B.G., Ingala, S., Lane, C.D., Jimenez, D., Haider, L., Varsavsky, T., Tanno, R., Smith, L., Ourselin, S., et al.: Let’s agree to disagree: Learning highly debatable multirater labelling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 665–673. Springer (2019)
44. Svensson, C.M., Figge, M.T., Hübner, R.: Automated classification of circulating tumor cells and the impact of interobserver variability on classifier training and performance. *Journal of Immunology Research* **2015** (2015)
45. Tao, F., Jiang, L., Li, C.: Label similarity-based weighted soft majority voting and pairing for crowdsourcing. *Knowledge and Information Systems* **62**, 2521–2538 (2020)
46. Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: A case for soft loss functions. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 8, pp. 173–177 (2020)
47. Vaughan, J.W.: Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.* **18**(1), 7026–7071 (2017)
48. Wu, B., Chen, W., Sun, P., Liu, W., Ghanem, B., Lyu, S.: Tagging like humans: Diverse and distinct image annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7967–7975 (2018)
49. Yager, R.R.: Interval valued entropies for dempster–shafer structures. *Knowledge-Based Systems* **161**, 390–397 (2018)
50. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. arXiv preprint arXiv:2101.05022 (2021)
51. Zhang, J., Wu, M., Sheng, V.S.: Ensemble learning from crowds. *IEEE Transactions on Knowledge and Data Engineering* **31**(8), 1506–1519 (2018)