Advancing Perspectivist Ground Truthing with Social Science

Table of contents

In a 2006 marketing conference, Clive Humby stated “Data is the new oil”. He was positing that, like crude oil, data must be refined in order to gain value, in the form of analysis (Palmer, 2006). The suggestion then was that, by analyzing data, one gains insights which have value:

“Data is just like crude. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.”

Palmer (2006), describing Clive Humby’s talk

AI systems take this value one step further, automating tasks - pieces of work typically carried out by people. But like insights, AI systems are trained on data, and must also be placed in a context to be understood:

“First we must understand that data is not insight. It is easy to grab a single fact and extrapolate that one fact into an actionable direction. But without context a fact is just that a fact, it is not an insight.”

Palmer (2006), describing Clive Humby’s talk

This thesis argues that data used for requires more than refinement, but rather design prior to collection. Essentially, current demands for systems require not only refinement, but research and design prior to collection, as well as contemporary reporting standards. Though some works report on design principles, drawn from established practices from relevant fields, these have yet to be broadly adopted. The thesis synthesizes design principles from the social sciences with principles from the computational sciences, for the purposes of collecting data to be used in AI systems. It reports on a case-study aimed at applying these principles. Following the case study, this thesis notes the additional effort and cost in approaching data collection this way, and presents a suggestion for a change in infrastructure to support more modern approaches to data collection for AI systems.

# Introduction

AI systems run on data. Data are used to ‘train’ *models* - imperfect, simplified mathematical or computational representations of a phenomenon or process in the real world [[1]](#footnote-20). Where an AI system is a complete application with integrated components e.g. an interface, programmatic logic, and one or more models aimed at performing tasks typically requiring human intelligence[[2]](#footnote-21), the models themselves are embedded components that take inputs (e.g. media like text, images, audio, or data) and produce outputs (e.g. classifications, predictions or some form of decision). Model *behavior* - the model’s output or response to a given input - is determined by their *parameters* - internal settings or values[[3]](#footnote-22). *Algorithms* - step by step instructions, executed in order[[4]](#footnote-23) - are used to estimate model parameters from the ‘training’ data. As models used in AI systems are designed to perform tasks, their performance is often evaluated empirically by comparing their outputs to a reference. This reference is often a second form of data, referred to as the ‘ground truth’ or ‘gold standard’, which represents the ideal expected output of the system.

Thus, AI systems require two forms of carefully collected data: training data to estimate model parameters, and reference data to evaluate model outputs. Training data may be tabular data, but is often a form of media - text, audio, images, or video - whereas reference data often contains aggregated input from humans (Geiger et al., 2020, 2021; Muller et al., 2021; Sav et al., 2023). Data sets designed for training, reference or both, are often re-used, likely due to the ease of access - often an online form and a mere download - compared to the effort and cost required to design, collect, and evaluate such data sets. Human input may be collected explicitly - where a phenomenon of interest is set as a target, and humans are given a task whereby their responses produce data relevant to the target. Common scenarios include human judges annotating, labeling, or rating a) individual pieces of content of the same form as the training data, or b) generated system outputs for the presence/absence, or degree of the phenomenon of interest(Muller et al., 2021). Data may also be collected implicitly - where digital traces of human behavior, e.g. media consumption, form the target (Sav et al., 2023).

Qualities of both forms of data (training and reference) determine the quality of the system, whether used to train the system, to evaluate it, or both. Parameters are estimated from the training data, with the aim that observable patterns will be recognized in the data, and affect the behavior of the model used by the system. Thus, if the model accurately reflects the data, parameters will accurately reflect imperfections, inaccuracies, biases etc. in the data as well. Also, as models are evaluated by comparing their outputs to reference data, such that preferred models are those whose outputs most closely resemble the reference data (Birhane et al., 2022), imperfections in the reference are also reflected in the models preferred. Thus, the best possible performance in the real world directly corresponds to the degree to which training and reference data represent the phenomenon of interest, in the environment to which it is to be deployed. And thus, evaluating a model in an AI system involves two measurement problems: 1) measuring the phenomenon of interest in the media selected for inclusion in the reference and/or training data, and 2) measuring the similarity of the output of a model to the reference (Welty et al., 2019). While the second has received extensive attention, the first has not.

Human input is never identical even when submitted carefully - in other words, people do not perfectly agree (Cabitza et al., 2023). As most models are only compatible with a single reference point per piece of content, multiple inputs are often collected from different people with reference to items of content, and subsequently aggregated. This ignores the process that produced the reference data by assuming a singular ‘ground truth’ per piece of content and treating all disagreement as ‘noise’ rather than ‘signal’, leading to biased or inaccurate reference data (Aroyo & Welty, 2015; Cabitza et al., 2023). Thus, relevant distributions in the data used as the reference and/or training must resemble those in the environment to which the system will be deployed (Hullman et al., 2022), but also account for the fact that human input is never identical (Aroyo & Welty, 2015; Cabitza et al., 2023).

Although the true values of the relevant distributions are unknowable, cues to whether human input resembles useful measurements can be calculated (Jacobs & Wallach, 2021; Welty et al., 2019), and can account for a range of reasonable interpretations in the reference (Aroyo & Welty, 2015; Cabitza et al., 2023). The social sciences have developed sampling methods to represent distributions in populations which can be adapted to the collection of content (Groves et al., 2009), and syntheses show how to leverage variance in human input using knowledge from survey science (Beck et al., 2022), metrology (Welty et al., 2019), psychometrics (Jacobs & Wallach, 2021), and the perspective approach to ground-truthing (Cabitza et al., 2023). Yet, knowledge from these fields has broadly not been applied in the field of Machine Learning, leading to issues of representation and measurement (Hullman et al., 2022).

## Present Work

Present work firstly attempts to synthesize input from various fields on how to better gather reference data to be used the ‘ground truth’ for the models that underlie AI systems. Present work then attempts to demonstrate the potential of synthesizing knowledge from the social sciences (psychometrics, survey science) related to sampling and measurement, with extant work on more useful content collection for use in machine learning tasks (metrology, perspectivist ground-truthing). Building on prior work reviewed herein, its main contribution is a synthesized framework that can be used to ground challenging phenomena in various media, following principles from prior work. A secondary contribution is a case study spanning several manuscripts, of a complex evaluation data set creation project. A third contribution, is knowledge directly applicable to the grounding of personal values in text, such as our annotation procedure, analysis of reference data, and statistics of interest for planning and estimating the costs associated. A final contribution is immediately applicable results that work towards estimating personal values in song lyrics using language models.

Included in this thesis are two manuscripts that further motivate the case study: 1) the first reviews strengths and weaknesses of datasets used in the field of Recommender Systems, 2) the second reviews how poor data practices in the field of Signal Processing related to datasets whose interconnections were poorly reported, misleading results. The case study demonstrates the use of principles from the Social Sciences to solve problems of representation and measurement across 4 manuscripts: personal values in song lyrics. In a 5th manuscript, the same principles are applied to a second form of text, political speeches, expected to vary in terms of use of ambiguous language. Despite the moderate success in automatically estimating values in lyrics, this work demonstrates a failure with speeches. It includes recommendations for analyses to observe the potential for success or failure, and to estimate cost via less expensive pilot studies. This thesis follows the case study with work that 6) highlights the potential for shortcomings in the interpretation of AI system evaluations should a more epistemologically sophisticated framework for evaluation not be adopted, and 7) highlights an important component of scientific infrastructure needed for rigorous work on data sets for Machine Learning: the treatment of scientific work as open-source artifacts.

# Background

## Reference data comes from humans

AI runs on data generated by humans. Reference data often uses responses from humans in the form of labels or annotations of content. “[D]ata annotation is the practice of labeling a set of digital representations of objects” (Cabitza et al., 2023). Although few studies have systematically examined the frequency of use of data from humans, it has been shown that reference data includes human input either explicitly or implicitly very often. Geiger et al. (2021) systematically review 200 randomly sampled papers from 3 broad domains, Social Sciences & Humanities, Life & Biomedical Sciences and Physical & Environmental Sciences. Out of the 140 studies that were classification tasks, 73.05% (or 103 papers) used labels derived from human responses as the reference. Geiger et al. (2020) reviewed 164 papers whose classifiers were trained on Twitter data and observed that 65% of the works reviewed used human annotations for the purposes of training. They further note that this quantity did not include human annotations used for validation, or other meta-data e.g. hashtags contributed by humans. In some domains the contribution of humans is in the form of digital traces, as in the domain of Recommender Systems where it was observed that, out of the most highly cited papers between 2018 and 2022, 86% of the datasets used were transaction data released by vendors such as Amazon or Yelp (Sav et al., 2023). Whether human input is explicit or implicit, it is present in almost all reference data.

Furthermore, training/reference data sets are often re-used, in some cases treated as *benchmarks* - measurement instruments used to produce comparable quantitative assessments of models (Welty et al., 2019) - magnifying their impact. Geiger et al. (2021) observed that 56.31% of the classification tasks that were reviewed (or 58 papers) used only ‘external’ human labels, i.e. labels that were not collected specifically for the work in the paper, Geiger et al. (2020) observed that 33.3% of the papers used external annotations, and Sav et al. (2023) observed that just 4 data sets appeared in at least 10% of works reviewed, with the most commonly used data set appearing in 33% of the works reviewed. Examining the most highly cited papers in IEEE CVPR from 2020-2022, the initial papers announcing the benchmark, along with the training and reference data received citation counts in the tens of thousands: Imagenet (Deng et al., 2009) shows over 52k citations, COCO (Lin et al., 2014) shows over 29k, Pascal VOC (Everingham et al., 2010) shows over 15k, according to SCOPUS as of April 2025. Thus, these human input data sets have the potential for long lasting effects on work that follows.

Although far more emphasis is placed on whether models achieve state of the art ‘performance’ or efficiency (Birhane et al., 2022), scholars over the past decade have attempted to draw attention to a lack of sophistication in how training and reference data are selected and evaluated (Aroyo & Welty, 2015). It has been argued that a focus on improving the data for a given task, will result in bigger gains than a focus on improving model[[5]](#footnote-26). Importantly, Hullman et al. (2022) show that optimizing for predictive accuracy does not absolve researchers from shortcomings in reference/training data, a situation exacerbated by the often re-use of data sets. A solution rather entails acknowledging that, whether deliberate or not, informed or not, organized or improvised, data are generated by process that may or may not be deliberately designed (Muller et al., 2021), yet would greatly benefit from design. Building on this, Welty et al. (2019) argue that datasets used to evaluate AI systems should be treated as measurement instruments in their own right. Drawing on the science of metrology, they propose that benchmark datasets ought to be evaluated using criteria analogous to those used for physical measurement tools.

## Common shortcomings of reference data design

Recent trends in Machine Learning (ML) — especially in deep learning — prioritize empirical performance over theoretical assumptions about the data generating process. A systematic analysis of highly cited ML works shows what is valued most: Performance, Generalization, Quantitative evidence, Efficiency, Building on past work, and Novelty Birhane et al. (2022). Unlike the social sciences (e.g. psychology), ML work often ignores attempts to model the process that gives rise to the data, and aims instead at predictive models whose outputs fall within some accepted estimated error bounds, resulting in poor or even biased reference data design (Hullman et al., 2022).

Considering the ever-presence of human influence on the reference data, best practices, considerations, and frameworks from the social sciences could inform designs, but have yet to be broadly applied in the computational sciences (Beck et al., 2022; Jacobs & Wallach, 2021). One reason for this gap may be that ML researchers prefer to work on building systems and evaluating their performance rather than researching, designing and executing ground-truthing projects (Muller et al., 2021; Sambasivan et al., 2021). Another may be a lack of focus on these topics in textbooks, and thus in education more broadly (Geiger et al., 2020). A third may be that the social and computational sciences have conceptually different focci: the computational sciences focus on the statistical model the system with substantially less emphasis on the content, whereas the social sciences treat the statistical model as a means to better understanding the relationships in the content (C. C. Liem et al., 2018). Psychology research thus contains many more research projects in which datasets are collected using responses from people, whereas datasets tend to be re-used extensively in machine learning work (Geiger et al., 2021). A further more practical complication is that work on these topics lacks the acknowledgement that ground-truthing is indeed a measurement problem, and lacks a central academic ‘home’: where psychology and economics have psychometrics and econometrics respectively, fields dedicated to studying field-specific measurement practices, the study of ground-truthing lacks a central banner under which academic work can accumulate and disseminate.

Decisions such as the selection of items for training data (Hullman et al., 2022), and the collection of human responses for reference data (Beck et al., 2022), are part of a design of a process that results in data (Muller et al., 2021). For example, an investigation of 15 data science workers, Muller et al. (2021) observed common phases, which include determining the annotation scheme - a) all possible labels that can be attributed to digital representations of objects along with any relevant guidelines, b) the actual process of collecting labels, and c) the process by which the annotations are then aggregated into a single label. They note the difficulty of this work: issues in the annotation schemes are often discovered as annotation projects progress, requiring varying degrees of improvised adjustment. For any of these components, decisions are made that impact the resulting reference data, whether or not they are being made by design.

Commonly observed shortcomings of data used in AI systems include: 1) representational biases in the content sampled for inclusion in training/evaluation datasets (Hullman et al., 2022), 2) a fallacious assumption of a single canonical ‘ground-truth’ when there are a range of reasonable interpretations (Aroyo & Welty, 2015; Cabitza et al., 2023), 3) measurement biases in the annotations collected (Beck et al., 2022; Hullman et al., 2022; Jacobs & Wallach, 2021), and 4) poor reporting of necessary information regarding the annotation-collection process (Geiger et al., 2021; Hullman et al., 2022).

An additional consideration that receives little attention is 5) the estimation of the number of annotations to gather, where fields that focus on gathering data from humans typically also have a strong emphasis on *a-priori* decisions, such as the pre-registration of calculated of target sample sizes estimated via statistical power analysis (Cohen, 1992), to mitigate sources of bias that come from the researcher. These considerations are absent in computational fields which appear to favor differing rules of thumb: e.g. in a well-cited textbook, Pustejovsky & Stubbs (2013) suggest to “have your corpus annotated by at least two people (more is preferable, but not always practical)”, whereas Artstein & Poesio (2008) suggest that “measuring reliability with only two coders is seldom enough, except for small-scale studies”.On the one hand, corpora tend to be very large, and resources are finite, making cost a primary factor in design decisions. On the other hand, rules of thumb lack clear substantiation in light of the both 1) the phenomenon being grounded and 2) the ambiguity of the media in which it is grounded. In other words, more variance is expected in annotation targets to the degree they are subjective or based on opinion (Beck et al., 2022), and more variance is expected in content to the degree to which it is ambiguous - i.e. can be interpreted in multiple ways - such as figurative language (Sandri et al., 2023). Further, some degree of variance will always be present when there are multiple annotations or ratings for a given piece of media independent of the target (Cabitza et al., 2023), and based on the range of reasonable interpretations of that target in that media (Aroyo & Welty, 2015).

### Representational bias

When sampling content to include in training/test datasets, samples for the training/test sets will ideally be drawn from the same distribution as the content in which they will eventually be deployed. *Representation bias* in content selected for training and/or evaluation datasets refers to the degree to which relevant distributions in data used as reference and/or training data resemble distributions in the environment to which a system is deployed (Hullman et al., 2022). If data used for training under-represents parts of the input space of an algorithm that then estimates parameters from that input space, the model resulting will have higher error rates for those under-represented parts of the input space when deployed. If content is selected without appropriate design aimed at representing the population from which samples are drawn, the overall distribution will not represent the population it was drawn from. Thus, optimizing for predictive accuracy using very large datasets does not ‘absolve’ researchers from having to consider the data generating process, and this includes sampling pieces of content to be annotated.

Approaches to representation problems can come from Sampling Theory, which frames the problem as one of selecting elements of a population, from which a sample must be drawn, and where the aim is that measurements of interest in the sample resemble measurements of interest in the population (Groves et al., 2009). This framework is typically applied to selecting people for inclusion in survey studies, whereby their responses to questions lend themselves to inference about a target population. Although there is no ‘one-size-fits-all’ solution to sampling, this thesis makes use of *stratified random sampling* as a general strategy: namely, the identification of groups of elements within a population that may affect the measurements in question, and the random sampling of elements within the groups, with approximately equal observations. In principle, this allows for the representation of the groups in population, on the measurement of interest, with some margin of error (Groves et al., 2009).

### The perspectives of annotators

The field of machine learning tends to treat all annotation variance as noise rather than signal. Often multiple ratings per piece of content are collected, aggregated, and only then shared, forming a singular ‘ground truth’ for the aspect of the content being labelled or rated. The quality of annotations is typically assessed using inter-annotator agreement, where more agreement is typically thought to indicate higher quality data Aroyo & Welty (2015). Thus, it is assumed that there is a singular canonical truth for each aspect / content pair, comprised of aggregated human responses, visible as the general agreement of human response, and which forms a target to which we align our automated systems. To illustrate more accurate representation of human responses, however, Aroyo & Welty (2015) operationalize their term ‘crowd truth’ as the ‘gold standard’ being the probability that a sentence contains an element, based on the probability that an annotator annotated that sentence with that element i.e. the label isn’t represented as ‘present’ or ‘not present’, but as a probability that an annotator labelled it as such. The probability that they may label it as such may in part explained by certain characteristics of theirs, such as their backgrounds, personal experiences etc. (Beck et al., 2022).

Disagreement is common and never fully reducible (Cabitza et al., 2019). Cabitza et al. (2023) show that this is the case whether the task is typically thought of as subjective, e.g. NLP tasks (Aroyo & Welty, 2015), but also in tasks thought to be far less so, e.g. medical cases (Cabitza et al., 2019). Disagreement, observable as variance in the human input, is often removed via 1) adjusting annotator training and instruction so as to reduce variance in the human inputs at the time of collection, 2) adjusting annotations via discussion post-collection, thus allowing annotators to establish conventions, discuss views, and re-think their responses, or 3) completely post-hoc at the time of modeling, via methods like majority voting, without input from the annotators. Each method of reducing variance - e.g. thorough training for crowd-sourced workers, regular annotator meetings to resolve disagreements, or taking a mean of ratings or majority vote - may result in different data independent of content, or the phenomenon of interest being annotated in the content.

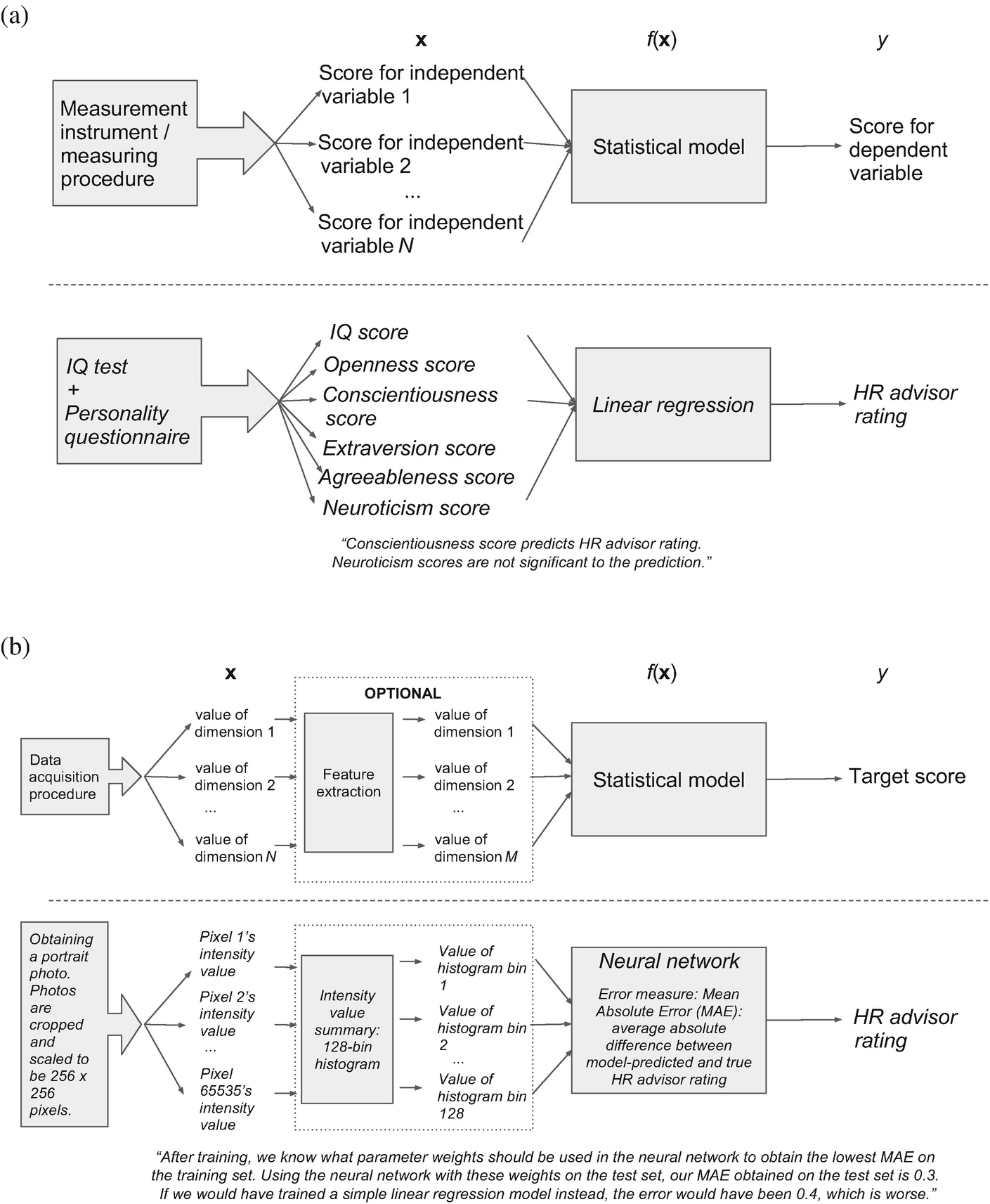
Further, variance in observed disagreement can be signal rather than noise. This signal may help to better understand the content being annotated: annotations may vary based on the ambiguity of the stimuli themselves, both in terms of the mode (audio vs. image vs. video vs. text), specific medium (Tweet vs. podcast transcript), or even the specific piece of content being annotated (Aroyo & Welty, 2015). Thus, not all pieces of content are equally unambiguous, and more ambiguous content is likely to result in greater variances in human input. This signal may help better understand the phenomenon of interest being annotated in the content: for at least some phenomena, the assumption that there is a single ground-truth to approximate with annotations doesn’t hold (Aroyo & Welty, 2015; Beck et al., 2022; Cabitza et al., 2023). More broadly, variances in the ratings may inform a finite “range of reasonable interpretations” of the phenomenon of interest being annotated in each piece of content, rather than a singular point. This signal may also help to better understand the background of annotators: people’s ethnic and/or cultural backgrounds may determine how they interpret content, and thus characteristics of the annotators may explain variance in the annotations. For example, although we expect hate speech exists, people’s perceptions of what constitutes hate speech may vary Beck et al. (2022). Showing that perceptions vary by identifiable characteristics, e.g. gender identity, ethnicity etc. may help unearth biases, whereby a single group perspective appears ‘objective’ (Cabitza et al., 2023).

Taking an approach to gathering reference data that attempts to account for the perspectives of the annotators is referred to as the *perspectivist*[[6]](#footnote-29) approach. It can apply to both the data annotation but also the modelling phase of ML projects, where benefits to ML models has been shown in a number of contexts (Cabitza et al., 2023). Although typically focused on the annotation of language data, perspectivist approaches can be broadly applied to annotations in reference data: *weak* perspectivist approaches involve taking perspectives into account while designing and collecting annotations e.g. by ensuring heterogenous raters and gathering enough ratings, as well as sharing and reporting the disaggregated data, but ultimately reducing annotations to a single label or rating for modeling. *Strong* perspectivist approaches involve taking perspectives into account for ground truthing and modelling phases.

Taking the perspectivist approach has a number of clear benefits, but also costs. It involves substantially more effort required to design the process that will result in annotations, higher costs in terms of the number of annotations and annotators needed in order to examine sources of variance, and challenges validating the data. In addition there are thus far few perspectivist modelling approaches that make full use of the variance in inputs (Cabitza et al., 2023). However, the perspectivist approach better reflects the reality that collecting annotations is a process that generates data with a number of relevant components (Hullman et al., 2022; Jacobs & Wallach, 2021). Further, it is a more complete report of the data resulting from the annotation process: the inclusion of the varying inputs in turn allows for better understanding of the content being annotated, the annotators annotating it, and the phenomenon of interest being annotated, which in turn allows for the development of models that make use of this information (Cabitza et al., 2023). This thesis accounts for annotator perspectives by collecting data using stratified sampling (Groves et al., 2009) among annotators, using cross-classified multilevel models to assess whether participant characteristics have statistically significant effects on their ratings (Doedens et al., 2022), and reporting disaggregated data (Cabitza et al., 2023).

### Measurement bias

The social sciences treat data from people as imperfect observations of a latent variable called a *construct* - like the effectiveness of a teacher, or recidivism i.e. the risk that someone will repeat a crime, or personality from the field of Psychology (Cronbach & Meehl, 1955). Social and computational sciences traditionally have different focci: where the social sciences emphasize an interpretable meaning of and , where and are not always directly observable, the computational sciences instead focus on the statistical procedure that correlates in terms of (see Fig 1.). Applying the latent variable approach to the gathering of annotations, Jacobs & Wallach (2021) suggest there is a ‘measurement error model’ (a term they borrow from the field of Economics), that links the unobservable latent variable, and properties that we can observe - in our case, the data produced when people label or annotate. Thus, the annotations we observe can not be the ‘ground truth’ as such a thing is unknowable. Rather, each annotation is an imperfect indication that can be used to estimate the ground truth.

 Fig. 1: C. C. Liem et al. (2018)

As one may measure one’s height with a ruler, one may acknowledge that no measurement is perfect, but estimate one’s latent ‘height’ via multiple measurements Jacobs & Wallach (2021). Similar to the ruler being an instrument to measure height, the social sciences - e.g. Psychology, Survey Science and Cognitive Science - research and develop *instruments*: standardized, systematic procedures designed to compare individuals (Cronbach, 1960). These are often surveys or standardized tasks, are designed to measure one or more constructs, and undergo assessment prior to being considered acceptably valid and useful (Cronbach & Meehl, 1955).

Survey science focuses specifically on surveys. It seeks to minimize the influence of sources of ‘noise’ via survey design: *Task Structure* involves refining specific wording and response options, including deciding on the inclusion of “I Don’t Know” or otherwise neutral response options, *Order Effects* involves strategies to randomly present content, as judgements of a specific piece of content are affected by perceptions of immediately previous pieces of content, and *Annotator Effects* which involves strategies to appropriately account for differences in perception based on the backgrounds, experiences and opinions of the annotators (Beck et al., 2022).

A number of other fields provide frameworks for assessing the quality of a measurement instrument, including psychometrics (Jacobs & Wallach, 2021), and metrology - the science of measurement - (Welty et al., 2019). For instance, the concept of *reliability* asks whether similar inputs consistently produce similar outputs, either across annotators (inter-rater reliability) or over time (test-retest reliability). The related concept of *precision* in metrology, separates the similarity of measurements from an instrument into *repeatability*, the similarity of measurements given that the operator, equipment, calibration, environment, and time between measurements are held constant, and *reproducibility*, the similarity of measurements given that the aforementioned are not held constant (Welty et al., 2019).

Beyond consistency, *validity* addresses whether the instrument is actually measuring what it claims to measure (Cronbach & Meehl, 1955; Jacobs & Wallach, 2021). This includes checks for face validity (does it seem plausible?), content validity (does it cover the full scope of the concept?), and structural or substantive validity (do the internal patterns make sense given extant theory?). Other forms such as convergent and discriminant validity test whether the measure behaves as expected relative to related or unrelated constructs, while predictive, hypothesis, and consequential validity consider what the measurement enables: does it support useful predictions, align with theoretical expectations, or have appropriate consequences in applied contexts? The ultimate conclusion is thus an estimate of construct validity: does the instrument measure the construct it intends to? (Cronbach & Meehl, 1955)

Although there is no one-size-fits-all solution to estimating the quality of an instrument, these various tools provide insights into whether the measurements appear to have qualities fitting of good measurements. This case study in this thesis builds on extant work by starting with a validated questionnaire. It then makes use of estimates of inter-rater reliability, precision, structural validity, to assess the quality of measurements.

### Inadequate reporting

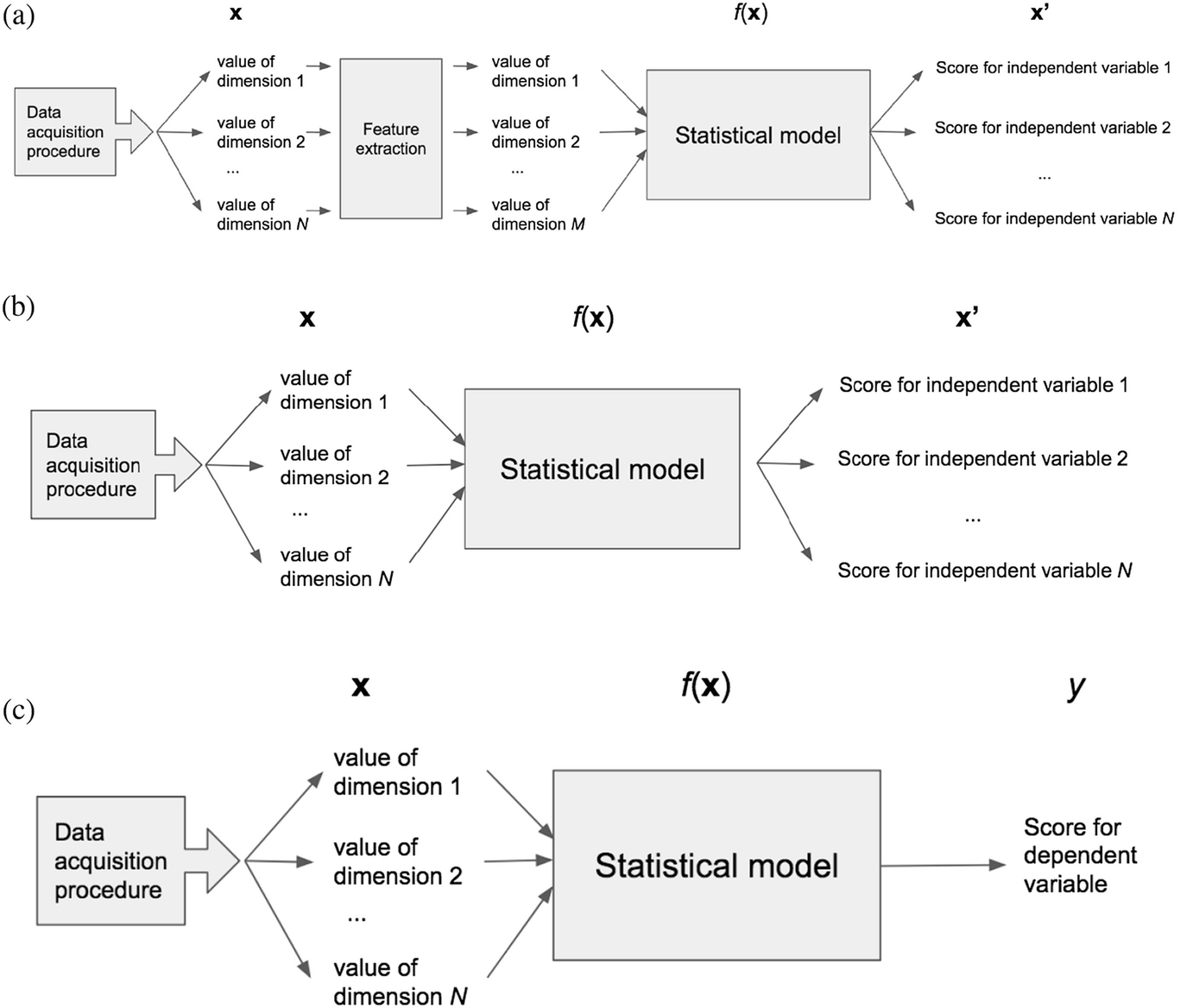
Despite the central role that human-labeled data play in machine learning, studies often provide insufficient documentation about how these data were created. In their systematic review, Geiger et al. (2021) and Geiger et al. (2020) found that many ML papers fail to adequately describe the processes used to create ‘ground truth’ labels, leaving unclear what exactly is being measured or how. As a result, the reference data often function as opaque black boxes, preventing meaningful scrutiny of what a model has learned. This is particularly concerning as they show that such datasets are reused across multiple studies, amplifying the impact of unreported or poorly understood annotation processes. Building on this, Hullman et al. (2022) argues that when reference data are under-specified, it becomes impossible to determine what data-generating process a trained model actually represents. Without transparency about how labels were created, we cannot evaluate the model’s fitness for deployment, its fairness, or its generalizability.

Geiger et al. (2021) provide concrete recommendations for improving transparency and accountability when reporting annotation procedures. They argue that researchers should report who the annotators are (including demographics or expertise where relevant), how they were recruited, and what task instructions were given. Additionally, they recommend documenting the labeling environment, such as tools or interfaces used, and whether annotators worked independently or collaboratively. Reporting should also include measures of inter-annotator agreement, details on how disagreement was handled (e.g., aggregation method, arbitration), and any quality control mechanisms applied. These reporting practices not only support reproducibility and critical evaluation but also help surface the social and epistemic assumptions embedded in the annotation process — assumptions which directly shape the model’s understanding of the world. Cabitza et al. (2023) further emphasize that adequate reporting should include details such as the number and expertise of raters, their incentives, instructions provided, time spent per annotation, inter-rater agreement metrics, the method used to aggregate annotations, and any confidence measures. These elements are necessary not only for reproducibility but also for evaluating the quality and appropriateness of the annotation process itself. Without them, assessments of model performance risk being built on shaky foundations, misrepresenting both the model and the phenomenon it is intended to capture.

# Present Work

This thesis proposes a framework for designing reference data that synthesizes principles from psychometrics, survey science, metrology, and perspectivist approaches. It responds to the shortcomings identified in prior work by emphasizing deliberate design prior to data collection. Rather than assuming fixed annotation targets or relying on post hoc aggregations strategies, this approach frames annotation as a measurement problem and applies tools from the social sciences to guide decisions about sampling, instrument development, and interpretation.

Chapters 2 and 3 are published manuscripts resulting from a project directed by this thesis, whereby students were assigned to assess the quality of datasets in two fields: Recommender systems (Sav et al., 2023), and Signal Processing (C. C. S. Liem et al., 2024). Largely inspired by (**gieger2020garbage?**), and (Geiger et al., 2021), rather than reporting on whether or not machine learning papers report characteristics of the data that they use for training and evaluation, our project aimed to recover all initial data-set reports as the unit of analysis. These papers show important shortcomings in the quality of the datasets, and issues arising from poor reporting practices.

The primary case study of this thesis spans the published manuscripts in chapters 4-6, and the manuscript under review in chapter 7. It works towards path (b) in (C. C. Liem et al., 2018), shown in: , and showcases a design for a challenging ground-truthing project, in terms of the complexity of the phenomenon of interest, ambiguity in the media that selected and annotated. It incorporates design choices to address the aforementioned shortcomings into a singular framework, guided by best practices in the social sciences, which it then extends, by demonstrating the potential for grounding a 10-dimensional latent construct, in two forms of text.

To mitigate representational bias, the framework uses stratified sampling to select both the media being annotated and the annotators themselves, ensuring that relevant subgroups within a target population are included. To address measurement bias, it treats the target construct as latent and uses a previously validated instrument to guide the annotation task. It then estimates the reliability and structural validity of the measurements, when the instrument is adapted for use in gathering annotations. The ambiguity of the selected media — in this case, song lyrics and political speeches — is treated as a design variable rather than an obstacle, allowing the study to explore how human interpretation varies in response to content characteristics as well as annotator characteristics. Further, the number of ratings required per item is estimated a priori, providing a clear rationale for the scope and cost of the data collection.

Finally, this framework operationalizes perspectivist ground-truthing by acknowledging and preserving variation in annotator responses. It collects annotator characteristics, reports the annotation process in detail, and publishes disaggregated annotations to support further analysis. These choices are not merely methodological preferences but are essential to generating ground truth data that reflect the complexity of real-world phenomena. In doing so, the thesis offers both a critique of current practices and a working example of a more rigorous, transparent, and epistemologically grounded approach to reference data design in AI.

Chapter 8 is a manuscript that highlights and acknowledges limitations in the manner in which machine learning models are evaluated, and their outputs interpreted (Altmeyer et al., 2024). My contribution is specifically sections regarding bias and anthropomorphization.

Chapter 9 is a manuscript that highlights and acknowledges limitations in the manner in which science broadly treats its outputs (C. C. S. Liem & Demetriou, 2023). Finally I conclude in chapter 10.

## References

Altmeyer, P., Demetriou, A. M., Bartlett, A., & Liem, C. (2024). Position: Stop making unscientific AGI performance claims. *arXiv Preprint arXiv:2402.03962*.

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15–24.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Beck, J., Eckman, S., Chew, R., & Kreuter, F. (2022). Improving labeling through social science insights: Results and research agenda. *International Conference on Human-Computer Interaction*, 245–261.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 6860–6868.

Cabitza, F., Locoro, A., Alderighi, C., Rasoini, R., Compagnone, D., & Berjano, P. (2019). The elephant in the record: On the multiplicity of data recording work. *Health Informatics Journal*, *25*(3), 475–490.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101.

Cronbach, L. J. (1960). *Essentials of psychological testing, 2nd edition*. Harper.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Doedens, P., Ter Riet, G., Boyette, L.-L., Latour, C., Haan, L. de, & Twisk, J. (2022). Cross-classified multilevel models improved standard error estimates of covariates in clinical outcomes–a simulation study. *Journal of Clinical Epidemiology*, *145*, 39–46.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*, 303–338.

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). " garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *arXiv Preprint arXiv:2107.02278*.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (Vol. 561). John Wiley & Sons.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

Liem, C. C. S., & Demetriou, A. M. (2023). Treat societally impactful scientific insights as open-source software artifacts. *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, 150–156. <https://doi.org/10.1109/ICSE-SEIS58686.2023.00020>

Liem, C. C. S., Taşcılar, D., & Demetriou, A. M. (2024). A quest through interconnected datasets: Lessons from highly-cited ICASSP papers. *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, 1–8. <https://doi.org/10.1109/CBMI62980.2024.10859219>

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 197–253.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part v 13*, 740–755.

Muller, M., Wolf, C. T., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., Sharma, A., Brimijoin, K., Pan, Q., Duesterwald, E., et al. (2021). Designing ground truth and the social life of labels. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Palmer, M. (2006). *Data is the new oil*. <https://web.archive.org/web/20170202072553/http://ana.blogs.com/maestros/2006/11/data_is_the_new.html>

Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for machine learning*  (First Edition). O’Reilly Media.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Sandri, M., Leonardelli, E., Tonelli, S., & Ježek, E. (2023). Why don’t you do it right? Analysing annotators’ disagreement in subjective tasks. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2428–2441.

Sav, A.-G., Demetriou, A. M., & Liem, C. C. (2023). Annotation practices in societally impactful machine learning applications: What are popular recommender systems models actually trained on? *Perspectives@ RecSys*.

Welty, C., Paritosh, P., & Aroyo, L. (2019). Metrology for AI: From benchmarks to instruments. *arXiv Preprint arXiv:1911.01875*.

1. https://en.wikipedia.org/wiki/Scientific\_modelling [↑](#footnote-ref-20)
2. https://en.wikipedia.org/wiki/Artificial\_intelligence [↑](#footnote-ref-21)
3. https://en.wikipedia.org/wiki/Statistical\_parameter [↑](#footnote-ref-22)
4. https://en.wikipedia.org/wiki/Algorithm [↑](#footnote-ref-23)
5. https://www.youtube.com/watch?v=06-AZXmwHjo [↑](#footnote-ref-26)
6. https://pdai.info/ [↑](#footnote-ref-29)