
Measuring Data

Margaret Mitchell
Hugging Face
Seattle, USA
meg@huggingface.co

Alexandra Sasha Luccioni
Hugging Face
Montreal, Canada
sasha.luccioni@hf.co

Nathan Lambert
Hugging Face
Berkeley, USA

Marissa Gerchick
Hugging Face*
Palo Alto, USA

Angelina McMillan-Major
University of Washington
Seattle, USA

Ezinwanne Ozoani
Hugging Face
Dublin, Ireland

Nazneen Rajani
Hugging Face
Palo Alto, USA

Tristan Thrush
Hugging Face
Palo Alto, USA

Yacine Jernite
Hugging Face
Brooklyn, USA

Douwe Kiela
Hugging Face
Palo Alto, USA

Abstract

We identify the task of *measuring data* to quantitatively characterize the composition of machine learning data and datasets. Similar to an object’s height, width, and volume, data *measurements* quantify different attributes of data along common dimensions that support comparison. Several lines of research have proposed what we refer to as measurements, with differing terminology; we bring some of this work together, particularly in fields of computer vision and language, and build from it to motivate *measuring data* as a critical component of responsible AI development. Measuring data aids in systematically building and analyzing machine learning (ML) data towards specific goals and gaining better control of what modern ML systems will learn. We conclude with a discussion of the many avenues of future work, the limitations of data measurements, and how to leverage these measurement approaches in research and practice.

1 Introduction

The size of datasets required for training machine learning (ML) models has quickly grown, as many recent models require amounts of data that are orders of magnitude larger than what was common only a few years ago. We are now at a point where many datasets are said to be “too large to document” for modern ML systems (Bender and Gebru et al., 2021). Combined with the modern *laissez-faire* approach to dataset development and usage, which is shallow and limited when it is done at all (Jo and Gebru, 2020), we are currently in the midst of creating, sharing, and training models based on datasets that we know very little about, contributing to behaviour from models that we can neither predict nor trace (Akyürek et al., 2022; Pruthi et al., 2020).

Measuring data is useful for (1) creating datasets from data sources; (2) documenting existing datasets; and (3) analyzing the outputs of systems as their own kind of data (cf. (Aka et al., 2021)), among other

*Work done while interning at Hugging Face.

uses. By measuring data from data sources, its measurements can be used to determine whether it should (or should not) be included in a dataset. Many measurements can be quickly and automatically calculated, and so help guide data collectors and dataset developers towards creating datasets that meet a given set of requirements or that have well-documented properties. For example, by comparing the changes in average sentence length or image size across incremental collection batches, a dataset developer can know whether some data sources may be more preferable to continue sampling from than others for the intended purposes of their dataset. Such real-time data selection can result in better performing models (Lee et al., 2021). Applied to datasets that are already built, measurements can be used to document the dataset’s characteristics and enable cross-dataset comparison. And applied to the output of machine learning models, such as a language model or the labels generated by an object detection system, measuring data can uncover patterns and biases that the model has learned (Aka et al., 2021; Meister and Cotterell, 2021a).

As we discuss below, recent work in fields ranging from philosophy to natural language processing have proposed methods for quantifying data and motivated the importance of this work, yet this research has not yet coalesced around a dedicated task. In the following sections, we detail several proposals for what we refer to as *data measurements*, and describe their relevance within ML pipelines.

2 Background and Prior Work

Despite the longevity, ubiquity, and importance of measurement, there is limited consensus on how to define the term, what sorts of things are measurable, or which conditions make measurement possible. Different fields have used different terms for measurements somewhat interchangeably, or with different nuances that are not precisely defined². Yet it is important to grapple with its definition to draw together the similar lines of research across fields and to create a common vocabulary and shared understanding around which to organize future interdisciplinary research.

Several fields of research, including those in the physical sciences, metrology, and psychometrics, have offered examples and guidance that can help to further clarify what the task of measuring data can entail. We briefly survey this previous work in this section, and in Section 4 connect these ideas to recently proposed methods for data quantification. Note that we use the term *dataset* as distinct from *data source*, with the distinction that a dataset is a set of data points created from a data source. Unless otherwise noted, we follow common practice of using the term *data* to refer to both datasets and data sources.

2.1 A Brief History of Measurement

The idea of *measurement* has existed as far back as we can trace human society. The earliest evidence of measurement is in the form of notches on bones, roughly 33,000 years ago (Vincent, 2022). Some of the earliest known examples of standardized measurement come from around the 30th century BCE, when the Egyptian cubit marked length based on the distance from the elbow to the tip of the middle finger (see Figure 1). Some of the first attributes of measurement included length and weight (Michell, 2005; Vincent, 2022). These are *extensive* attributes of an object, meaning that their value is directly reflective of the structure of the object: the more object there is, the higher the measurement value. Later research on measurement recognized *intensive* attributes, such as temperature, which are not dependent on the amount of the object and whose measurements can be taken *via* extensive attributes. For example, the intensive attribute of density is calculated by dividing the extensive attributes of mass and volume. In the early 20th century, this type of calculation would come to be known as a *derived measurement*, distinct from a *fundamental measurement* such as length.

Precise definitions of measurement were put forward throughout the 20th century. While definitions vary, “the assignment of numbers to items” is common across many definitions. (Campbell, 1928; Stevens, 1946; Roberts, 1985; Boslaugh and Watters, 2008; Díez, 2009). Further distinctions include that the assignment of numbers must be systematic (e.g., Roberts (1985)), that they facilitate

²Etymologically, while the word *measurement* in English derives from the ancient Greek term μέτρον (roughly, *MEH-tron*), *metric* is directly traceable to the use of the suffix -ικός in Greek: μετρικός (roughly *meh-tree-KOS*), which marks an adjectival form. An informal survey of our peers who speak a variety of Indo-European and Afro-Asiatic languages suggested that different terms for *metric* and *measurement* exist, yet the distinction between them is similarly blurred

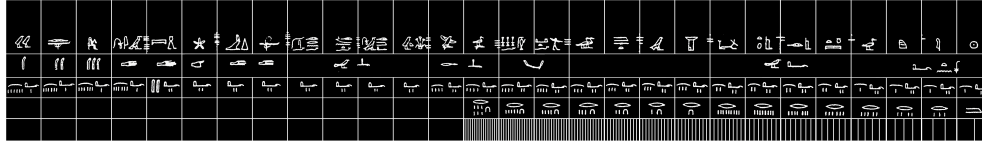


Figure 1: Ancient Egyptian Cubit Rod for measuring length. One of the earliest known objects for standardized measurement. Turin Museum, Wikimedia Commons, CC BY-SA 3.0

TABLE 1

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (Invariantive)
NOMINAL	Determination of equality	<i>Permutation group</i> $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	<i>Isotonic group</i> $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	<i>General linear group</i> $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	<i>Similarity group</i> $x' = ax$	Coefficient of variation

Figure 2: S. S. Stevens’ measurement scales, from the seminal *On the Theory of Scales of Measurement*. (Stevens, 1946)

mathematical quantification (e.g., Boslaugh and Watters (2008)), that they apply to objects and their properties (e.g., Campbell (1928)), that they are accessible, proportional, and consistent (Cre (2011), and that they must be gradual, allowing for “more or less” comparisons (Díez, 2009). Different scales of measurement have also been defined relatively recently (see Figure 2), further deepening our understanding of how measurements can be applied in larger statistical analyses. Precise definitions of measurement paved the way for Measurement Theory, which examines the quantitative structure that abstract concepts, such as comprehension or (problematically) intelligence (Harrington, 1975), may take (Michell, 1997; Hox and Boeije, 2005).

2.2 Measurement in Machine Learning and Data Science

The task of measuring data has been at the core of Machine Learning from the early days of the field. In corpus linguistics research adopted into the modern-day ML subfield of natural language processing (NLP), one of the earliest large-scale corpora for English (Kucera et al., 1967) was carefully curated and documented with the support of a range of measures, such as by selecting document types to reflect a *prior* base distribution of English written text, or estimating the entropy of the data (Brown et al., 1992). However, as the scale of datasets used in machine learning has drastically increased in recent years, measuring practices have fallen behind as values of efficiency and approaches that prioritize quantity over considerations of other dataset qualities have become more common place (Scheuerman et al., 2021).

Addressing this phenomenon, recent work in machine learning has highlighted a critical need for more in-depth probing of the datasets we use for training models, stressing the importance of measures that can enable real-time assessment of data collection to help increase the scientific value and reusability of the data:

“A whole new science of data is needed, with HCI partnership, where sorely needed phenomenological goodness-of-data-metrics need to be developed. . . Such research is necessary for enabling better incentives for data, as it is hard to improve something we can not measure” (Sambasivan et al., 2021)

The idea of “goodness-of-data” metrics to measure data resembles recent proposals for quantifying data properties within machine learning research and data science more broadly using different

terminology. Examples from research on text and image data include “quality criteria” to analyze, evaluate, and compare the quality of vision and language datasets against one another (Ferraro et al., 2015); “task-independent metrics” to reflect states of the data without the contextual knowledge of the application (Pipino et al., 2002); “characteristic metrics”, which the authors define as “unsupervised measures” to quantitatively describe the properties of a collection of data or a dataset (Lai et al., 2020); “intrinsic metrics”, to quantify the properties of a dataset (Grusky et al., 2018; Bommasani and Cardie, 2020a); “statistical tendencies”, described as “present” in datasets (Meister and Cotterell, 2021a); “statistical properties” to surface issues with dataset contents (Paullada et al., 2021), and “bias measurements” to quantify dataset bias (Aka et al., 2021). Although these different works often do not directly connect their research to one another, they all provide proposals for quantifying datasets and their properties to aid data understanding and comparison. Many proposed quantification methods are also motivated by a shared desire to describe datasets without requiring “ground-truth” annotations commonly used in existing training and evaluation pipelines for machine learning models.

The utility of measuring data can be understood against the backdrop of increased “data-centric AI” research, with groups in academia and industry stressing the importance of “systematic methods to evaluate, synthesize, clean and annotate the data used to train and test [AI models]” (Liang et al., 2022). This has resulted in proposals for more systematic approaches to data collection (Whang et al., 2021), data-centric model training (Motamedi et al., 2021), benchmarking (Eyuboglu et al., 2022), and debugging (Rajani et al., 2022), as well as hubs that bring together resources and publications in the field (Data-centric AI, 2021). Recent tools in this vein include CleanLab (2022), DataLab (Xiao et al., 2022) and Snorkel (Ratner et al., 2017), which provide solutions for analyzing and curating quality ML datasets, but are limited in scope to absent methods for systematic evaluation of the data.

Taken together, this surge in data-centric work makes clear the need for systematic methods of data measurement. Similar thinking extends to tooling provided by and for companies, for example, IBM’s tool for data “quality analysis”, which uses a series of “quality scores” to enable formalized and systematic data preparation (Jariwala et al., 2022); Google’s “Know Your Data” tool, created specifically for AI practitioners, has also helped shed light on the contents and characteristics of popular datasets (Google Research, 2021). In terms of AI conferences, the DataPerf workshop was founded specifically for the development of tools and methodologies for dataset analysis (DataPerf Workshop, a,b) and the NeurIPS Datasets & Benchmarks Track was created in 2021 in order to incentivize dataset creation and analysis (Vanschoren and Yeung, 2021), indicating that the field is beginning to recognize the importance of better data quality and its contribution to better performance.

Prior work has also explored the possible negative consequences of using measurements without proper contextualization or qualification. For example, Gururangan et al. (2022) explore how unexamined definitions of what constitutes data “quality” may encode ideologies in ways that further suppress marginalized identities. Such phenomena call into question the possibility of having objective or universally relevant measurements of some implicitly defined data qualities (Scheuerman et al., 2021). In order to address these risks when attempting to measure *unobservable theoretical constructs* and particularly social constructs, Jacobs and Wallach (2021) propose a framework for measurement modeling in order to scope the assumptions, purpose, and validity of a specific measurement and avoid misuse or erroneous conclusions.

3 A Formulation of Measurement for Machine Learning Data

There is a substantial opportunity to modernize how the data itself is related to advances in ML research by quantifying the composition of the data. We refer to this task as *measuring data*. Intuitively, data measuring can be compared with measurement in the physical sciences. For example, geometry defines what a *distance* is and how it can be used to measure an item’s *length*; distance and length measurements can be applied to data in a comparable fashion at the level of the word, sentence, or document, resulting in measurements such as *word length*. Similarly, the measure of the *density* of an object, common in physics, or the *diversity* of specimens in an area, common in biology and ecology, have direct connection to quantifications introduced for machine learning data such as *image density* (Kingma and Welling, 2013) and *subset diversity* (Mitchell et al., 2020). These newly emerging measurement approaches for machine learning data can leverage the vast history of measurement to further focus the task of *data measurement*.

3.1 Data Measurements

An intuition behind *what a data measurement is* is well-explained in Lai et al. (2020), as something to “quantitatively describe or summarize the properties of a data collection”. The authors continue:

“These metrics generally do not use ground truth labels and only measure the intrinsic characteristics of data. The most prominent example is descriptive statistics that summarizes a data collection by a group of unsupervised measures such as mean or median for central tendency, variance or minimum-maximum for dispersion, skewness for symmetry, and kurtosis for heavy-tailed analysis.”

Following this and the prior work discussed in Section 2, we define a *data measurement* as something that:

- quantifies the magnitude of a characteristic or property of the data
- is calculated from the data’s composition
- can be composed of, or derived from, units

Within the context of data, a *unit* can be a single atomic instance with which a dataset is constructed. For text, this might be the character, word, or sentence. For images, this might be the image or pixel. Following the terminology introduced in Section 2.1, such units demarcate *extensive* attributes of data for *fundamental* measurements: the more there is, the higher the measurement value. What can be recognized as *intensive* attributes and *derived* measurements for data have also been proposed. We turn to examples of these measurement types in Section 4, illustrating how these are used in different modalities.

4 Examples of Data Measurements

In the paragraphs below, we describe a set of measurements that have been proposed in different threads of research, contextualized with respect to the specific task of measuring data.

Several approaches are domain- and modality-agnostic, denoted as *General Data Measures*, in order to differentiate from measurements applied specifically in the context of language 🗣️ or computer vision 🖥️, denoted in *Modality-specific Measures*. Of particular note are measurements that rely on embeddings of the data instances in a metric space, as such representations allow users to more directly leverage the full range of existing measurements in those spaces. Such embeddings may be obtained for example by modeling co-occurrences between data instances through Singular Value Decomposition (Golub and Reinsch, 1970) or Word2Vec (Mikolov et al., 2013), or by using pre-trained embedding models developed on external datasets. We return to a discussion of the implications of using external sources for dataset measurements in Section 6.

As motivated in Section 1, all of the measurements that we describe can be useful to better understand ML data and datasets, their contents and their characteristics, prior to model training, iteratively during the data collection process, or even after the model is trained, since the output of a model can itself be treated as data and measurements on this data provide quantifications of characteristics that model has learned. The measurements we discuss are both old and new, spanning work on applied statistics within data science as well as research in language and image data analysis. They serve as an example of the kinds of measurements that can be used in the task of measuring data and are summarized in Table 4. We do not proclaim to be extensive – there are many measurements that we do not describe in the sections below – but we do aim to illustrate the breadth and multiplicity of data measurements that are relevant to the ML community.

We broadly categorize the measurements as DISTANCE, DENSITY, DIVERSITY, TENDENCY, and ASSOCIATION. We describe them and summarize their connection to machine learning data in the sections below.

4.1 Distance

As something that provides *extensive* values of *fundamental* attributes, distance is one of the most basic measurements. Within statistics and data science, *distance* quantifies separation between data

Table 1: Examples of different data measurements proposed in image- and language-based data science and machine learning, alongside analogs in the physical sciences.

	DISTANCE	DENSITY	DIVERSITY	TENDENCY	ASSOCIATION
Physical Sciences	Length	Mass-per-volume	Biodiversity	Mean, Median, Mode	Correlation
General Data Measures	Euclidean Distance	Data Density	Gini Diversity	Burstiness	
	Cosine Similarity	KNN Density	Vendi Score		
	Earth Mover’s Distance				
	Kullback-Leibler Divergence				
Modality-Specific Data Measures	Word Mover’s Distance (language 🐝)	Information Density (language 🐝)	Text Diversity (language 🐝)	Perplexity (language 🐝)	Pointwise Mutual Information
	Levenshtein Distance (language 🐝)	Idea Density (language 🐝)	Lexical Diversity (language 🐝)	Fit to Zipf’s Law (language 🐝)	
	Inception Distance (vision 🐱)	The Inception score (vision 🐱)	Image Diversity (vision 🐱) Subset Diversity (vision 🐱)		

instances, variables, distributions, or samples. Distance measurements are fundamental to machine learning given that much of ML is about learning representations of data that are easy to handle and manipulate, generally by embedding them in a *metric space* that defines distances between points.

General Data Measures

Euclidean Distance: In its most basic, one-dimensional form, this measurement is familiar across measurement systems for thousands of years and across fields: the length between two points. It can be applied in spaces with two or more dimensions, and so is commonly used when comparing vectors. It can be calculated from the Cartesian coordinates of points using the Pythagorean theorem, and is symmetric. Applied to data, Euclidean distance can be used to capture everything from how many words are between a re-occurrence of a word, to the divergence between the distribution over words in different parts of a dataset.

Cosine Similarity: A measure of similarity between two sequences or vectors, this measurement corresponds to the cosine of the angle between two vectors. This measurement is used in many fields, including NLP, where each word can be assigned a different coordinate and a document can then be represented as the vector of the numbers of occurrences of each word. In this way, cosine similarity can be used to measure of how similar two documents are likely to be in terms of the words they contain.

Earth Mover’s Distance (Rubner et al., 2000) represents the minimal cost for transforming one distribution of into another of the same domain. Cost can be defined in different ways depending on the specific applications, for instance either as the Euclidean distance of some measure of the quantity of work necessary to carry out the transformation. Earth Mover’s distance has been used in applications ranging from information retrieval and pattern recognition to calculate the distances between data from different modalities with diverse types of features (e.g. (Peleg et al., 1989; Orlova et al., 2016)).

Kullback-Leibler Divergence (1951): This measure aims to quantify the extent to which one probability distribution is different from another. It is a non-symmetric measure of the difference between the two distributions, i.e. the expected number of bits that are necessary to encode points from a given distribution when using another distribution (as opposed to the original distribution directly). Widely used in Bayesian statistics, KL divergence is also useful in applications ranging from magnetic resonance imaging to analyses of protein and genome structures (Belov and Armstrong, 2011).

Modality-specific Measures

Different variations of the general data measurements have also been defined in NLP and computer vision to better reflect the particularities of their modalities and the repercussions that this can have on the way the measurements are calculated.

Word Mover's Distance (Kusner et al., 2015) (language 🗣️): leverages word embeddings to calculate the minimum cumulative distance from one sequence of words to another. It is particularly useful for measuring the similarity of two documents in a way that general distance measurements do not given that it leverages the geometry of the word space itself.

Levenshtein Distance (1965) (language 🗣️): is a measurement for calculating the distance between strings, defined as the minimum number of single-character edits (i.e. insertions, deletions) that will convert one word into the other. It has been used in language applications ranging from speech recognition to spelling correction given its simplicity and efficiency.

Inception Distance (vision 🖼️): Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018), which leverage an Inception model (Szegedy et al., 2015) trained on ImageNet, were created specifically to assess the quality of images generated by ML models. They do so by calculating the distance between a distribution of generated images with that of a set of real images, considered the ground truth. In that sense, they relate to the tendency measurements that we define in Section 4.4, given that they calculate the distances between image distributions, not the images themselves.

4.2 Density

Density is a qualification of compactness; in the context of datasets, density can be calculated either across a whole dataset or on a subset of it. It indicates how well a data space represents concepts, such as by quantifying how many variations of an exemplar are present. Intuitively, it tells us how well we expect a model to handle a part of the space, or how coherent (possibly implicit) categories of the data are. Density measures have the advantage of being relatively easy to compute across even large datasets, with the disadvantage of being hard to interpret. They can be useful for initial, high-level analyses of datasets and as indicators to guide further fine-grained analyses which can go deeper in terms of the intrinsic characteristics of outliers and representative samples in the dataset (e.g., (Vig et al., 2021; Kannan et al., 2017)). Relevant measures for density in datasets include:

General Data Measures

KNN Density (Loftsgaarden and Quesenberry, 1965; Zhu et al., 2008): Average similarity between examples, which can be used for determining whether an unlabeled example is an outlier compared to other examples and finding patterns within a dataset. Similar to density in the physical sciences, this is a derived measurement that leverages distance; as such, it can also be considered a Distance measurement.

Data Density (Lai et al., 2020): An estimate of the number of samples that fall within a unit of volume in an embedding space. Similarly to density measures in the physical sciences, this measurement utilizes a fundamental measurement – volume – to derive a further value. This measurement is particularly useful in approaches involving data visualization, since it can help reflect how many data points are present in a defined area (such as a region of a map).

Modality-specific Measures

With the advent of larger and larger datasets, both computer vision and natural language processing have leveraged density measurements to get a bird's eye view of their contents. We describe some of these modality-specific measurements below, given that in practice their definitions can highly depend on their context of application.

Idea Density (Covington, 2009) (language 🗣️): Concentration of ideas (or propositions) within a sentence, based on existing research in psycholinguistics which showed that text with a lower density of ideas is easier to understand (Kintsch and Keenan, 1973). Computing this automatically allows distinguishing between documents that are conceptually simpler (i.e., those written for nonspecialist audiences) versus more technical ones. Different methods may be used to identify an “idea” or “proposition”, which can range from simple word matching within the dataset itself to more complex methods using external models.

Image Density (vision 🖼️): There are many approaches that aim to approximate the probability density functions (PDFs) of images using approaches ranging from modeling the low-level statistics of natural images (Olshausen and Field, 1996) to more complex approaches such as Variational Auto-Encoders (VAEs) (Kingma and Welling, 2013). While these approaches can work well for simpler and smaller images (such as the digits from the MNIST dataset Krizhevsky et al. (2009)), it remains difficult to model high-resolution images with many objects.

4.3 Diversity

Records of measurements quantifying how diverse a sample is are more recent, appearing over the last century in subfields of biology, notably for the purpose of measuring ecological diversity (cf. (Harris, 1916; Magurran, 1988)). Diversity is also referred to as heterogeneity in many sciences (Rényi et al., 1961; Nunes et al., 2020) which have their own approaches to measure it, Hill numbers in ecology (Hill, 1973), Hannah–Kay indices in economics (Hannah and Kay, 1977). In these domains, “Diversity Indices” are common, which operate over proportions, entropy (cf. Shannon Index), and probabilities with respect to randomness (cf. Simpson Index). The development and usage of diversity measurement within Machine Learning is much more recent (cf. Vendi Score).

General Data Measures

Gini Diversity Index (1912): Known by several different names in fields including economics, sociology, and psychology, this measure has been applied within Machine Learning for classification decision trees to select data splits (Raileanu and Stoffel, 2004; Strobl et al., 2007), similar to the usage of KL-Divergence discussed in subsection 4.1, although it remains under-utilized in other approaches and applications.

Vendi Score (Friedman and Dieng, 2022): In response to the lack of generic measures of diversity in ML, recent work by Friedman and Deng has proposed a reference-free way to calculate dataset diversity. The Vendi score does this by leveraging the exponential of the Shannon entropy of the eigenvalues of a similarity matrix calculated based on a user-defined similarity function. As such, it can be applied in fields as spanning from molecular modeling to computer vision and NLP.

Modality-specific Measures

Lexical Diversity (language 🗣️): Early research on some of the first “large” corpora, such as the British National Corpus, also provided many statistical measurements – for example, regarding concordances and collocations (Leech, 1992). This includes measurements such as n-gram diversity (or word diversity), which is essentially defined as the number of distinct n-grams or words divided by the total vocabulary (Li et al., 2015). While the recent expansion of dataset size has witnessed a loss of these measurements, they are useful for many reasons, ranging from detecting anomalous sequences to determining the maximum input size for models trained on the data.

Lexical Diversity (Lai et al., 2020) (language 🗣️): This measure provides a signal regarding how dispersed a cluster of text is, based on a set of embeddings. It can be used as an indicator of dataset homogeneity and to identify subsets or groups within a textual dataset (e.g. data from a given domain or source).

Inception Score (Salimans et al., 2016) (vision 🖼️): has been used to evaluate the quality of generated images by calculating their diversity based on the entropy of the marginal distribution of class labels as they are predicted by a classifier trained on the ImageNet dataset (Deng et al., 2009), meaning that it is limited by the classes in ImageNet as well as the quality of its images.

Subset Diversity (Mitchell et al., 2020) (vision 🖼️): The proportion of human characteristics subject to power differentials in a cluster of data, applied within the context of subset selection of images. This measurement applies specifically to data that represents individuals with respect to characteristics

such as gender and race, and requires that these characteristics be known. This can be used to compare the representativity of datasets compared to the populations that they are meant to reflect.

4.4 Tendency Measures

We introduce the term “tendency” to categorize summary statistics that quantitatively describe features from a collection of information or a distribution over measurements. Tendency values are calculated from distributions over measurements, including count, such as what is captured in descriptive and sufficient statistics. Within machine learning, measurements of central tendency are fundamental, with both objective functions and evaluation scores often utilizing means (averages). Applied to data, tendency measurements can provide information about the general nature of the distributions captured in the data, for example, the length of different instances. Common tendency measures include measures of *central tendency* (mean, median, mode) and *dispersion* (standard deviation, variance, the minimum and maximum values of a variable of interest, kurtosis). Other example measures include:

General Data Measures

Burstiness (Goh and Barabási, 2008): Many complex systems can be characterized using intermittent, heterogeneous patterns (“bursts”), properly representing these bursts, their frequency, and amplitude can help predict their future behavior. Machine learning datasets that have a temporal aspect (e.g., news articles, product/restaurant reviews, etc.), are the input to such a metric, where modeling the burst patterns can contribute to better understanding the dataset itself.

Skewness: While assumptions of normal distribution are still commonplace in ML, there are many cases where those assumptions do not hold, especially for datasets gathered ‘in the wild’. Measuring skewness (i.e. the extent to which the probability distribution of a variable deviates from the normal distribution) can help orient subsequent modeling approaches and indicate whether additional steps are necessary, for instance in terms of class balancing or data normalization (Prati et al., 2009; Provost, 2000).

Modality-specific Measures

Tendency measures can operate on different units depending on the characteristics of the datasets where they are put into use. For instance, when applied to text datasets, they can operate on levels ranging from characters and words to sentence lengths and frequencies. Applied to image datasets, tendency measures can operate on either the representations of the image pixels or latent representations of the images.

Fit to “Zipf’s Law” (Meister and Cotterell, 2021a; Luccioni et al., 2022) (language 🗣️): A familiar quantity in corpus linguistics and NLP, measurements based on Zipf’s law quantify how closely the quantities of an item in a dataset or data batch match a Zipfian distribution, which is an inverse rank–frequency distribution. Natural human languages adhere to Zipf’s law to various extents, with different coefficients (Gelbukh and Sidorov, 2001; Bentz and Ferrer Cancho, 2016). This coefficient, as well as the distance between the ideal Zipfian distribution and the observed distribution, function as measurement for the naturalness of data (Luccioni et al., 2022). Recent work has also demonstrated that Zipf’s Law arises naturally across domains with underlying, unobserved variables (Aitchison et al., 2016), suggesting it may be extended to datasets in different domains. Detecting groups of items that do not correspond to the Zipfian distribution can help identify outliers and undesirable artifacts in a dataset (such as, e.g., HTML tags).

Perplexity (language 🗣️): Historically, perplexity was used in information theory to reflect how well a probability distribution predicts a sample (Thomas and Joy, 2006). It was subsequently leveraged to evaluate the performance of language models (Brown et al., 1992; Bengio et al., 2000; Melis et al., 2017) as well as datasets, for tasks such as perplexity sampling (De la Rosa et al., 2022) and data selection (Toral et al., 2015). Perplexity can be a useful tool for detecting anomalies in datasets (e.g. sentences that have been incorrectly parsed, encoding errors, etc.), although its limitations have been documented compared to other measures such as Zipf’s law (Meister and Cotterell, 2021a).

4.5 Association-based Measures

Measures of association provide quantifications of the relationships between items in a dataset. These were largely developed in the last century, with the most common association measurement. Within

machine learning, association measurements provide insight into features that may be redundant with respect to one another, proxy variables, and artefacts of the data that pair concepts together (i.e., whether or not their pairing is reflective of their relationship in the world more generally). One class of association measures that are commonly used throughout work in data science and machine learning are *correlations*, which quantify the degree to which variables are linearly related. Other kinds of associations that move beyond the linear relationship include those based on mutual information, such as normalized pointwise mutual information.

General Data Measures

Data Correlations: Common across many fields for years, correlation measures can help quantify how different variables coordinate with one another, i.e., the strength of association and the direction of their relationship. The higher the value, the stronger the relationship between them. Correlations can be calculated for words within text data, or to signals within vision and audio data, where they can also function as a quantification of similarity. There are many choices for measuring correlations: some of the most commonly used measurements are the Pearson correlation coefficient (Benesty et al., 2009) and the Spearman rank-order correlation coefficient (Ramsey, 1989; Kokoska and Zwillinger, 2000), which can help represent inter-dataset relationships and dependencies.

Pointwise Mutual Information (PMI) (Aka et al., 2021): PMI provides values for the strength of association between different terms, where a term can be a word in text data, or a label in image data. This can be less scalable compared to some of the density- and distribution-based measures described above, as it requires exhaustive pairwise calculations if applied to all terms in a dataset, but is tractable when limited to a set of terms (and their co-occurrences with all other terms). PMI can contribute towards disentangling relationships between terms and detecting patterns and anomalies. For example, normalized PMI can be used to identify toxic stereotypes that appear in a dataset as strong associations between identity terms and other items (such as “woman” and “smile”) (Aka et al., 2021).

4.6 Other Types of Measurements

While we have provided an initial categorization of dataset measurements proposed in previous work, there are other existing measures that do not fit into the categories defined above, or that have yet to be applied specifically in machine learning and will be useful to further explore in future research. These include:

Redundancy Measurements: Work on redundancy within machine learning datasets has shown the negative impacts of duplicate data on language models (Lee et al., 2021), however, quantifying (and removing) duplicated items in a dataset has yet to become a norm in ML dataset creation. Straightforward measurements for this kind of redundancy could include the count of unique instances with duplicates, the total number of duplicates in the dataset, and counts of duplicates by type. Similar to several of the measures described above, entropy can also be used to measure redundancy, such as via the Generalized Entropy Index.

Readability (Flesch, 1979; Halliday, 1989) (language 🐝): Readability aims to represent how difficult a text is to a reader, and is frequently used in domains such as language didactics (To et al., 2013) and corpus linguistics (Pitler and Nenkova, 2008). It has been proposed as a metric for evaluating natural language generation (Novikova et al., 2017) but has yet to be used for dataset analysis, which could help determine the complexity and linguistic density of texts used for training models.

Noise: Related to the density measures defined above are measures for noise, which can quantify the random variation of density. These have been defined for audio and images (Avila-Vázquez, 2014). They have been defined for text as well (cf. (Subramaniam et al., 2009)), although they tend to require additional models of language. Similar proposals to density in text processing include *conceptual cohesion* (Binkley et al., 2009), *semantic similarity* (Han et al., 2013), and *semantic coherence* (Bommasani and Cardie, 2020b).

Homogeneity and inclusion: Similarly, many of the authors who have developed diversity measures have also introduced measures of homogeneity (Lai et al., 2020) and inclusion (Mitchell et al., 2020), which are complementary to the diversity measures and can help represent different aspects of (non-)correspondence between the dataset and the populations that it represents, or those that will be impacted by model predictions.

5 Discussion

5.1 On the Limitations of Measurements

In this work, we have attempted to describe and categorize different approaches for measuring data within the context of machine learning. While all of the measures that we have mentioned can provide crucial information about the composition and contents of datasets, they also come with limitations, both technically and in terms of problematic social biases. Measures and measurements can also be used as a tool of oppression, an issue that must be contended with as we advance work on measuring data that will influence models used throughout society.

Examining the limitations of measurement through a technical lens, one of the most salient limitations for several of the measurements we describe is that they are usually applied by leveraging an additional source other than the data being measured. For example, FID and KID leverage models trained on ImageNet, whereas perplexity is calculated based on a learned model of natural language. As such, values for these measurements are relative to the external models that are used and do not function to quantify the dataset in-and-of-itself. This means that these measurements are only directly comparable across different datasets when the same models are used – this is a known issue with perplexity, whose limitations in terms of evaluating language models have been pointed out (Meister and Cotterell, 2021b). While this can be addressed to some extent by using a standard set of models for computing different measurements, this would not address another severe limitation of these kinds of measures: the issue of bias inherent to the models used for them. For instance, the well-documented biases of the ImageNet categories (see Crawford and Paglen (2021); Luccioni and Rolnick (2022)) will influence the results of metrics that rely on these models. This is also true for measures that leverage embedding models (e.g., Word Mover’s Distance), since they depend upon the latent representations of concepts from the embeddings, which have been shown to contain problematic biases (Bolukbasi et al., 2016).

Examining limitations through a social lens, measurements of data, as with measurements of people, have the potential to be influenced by specific values and abstracted away from their original intent. For example, while students’ test scores are intended to measure their performance on a specific task, these scores are often used as a proxy to measure a students’ individual ability to learn and succeed in later education, and usually do not account for important environmental factors including class size or teacher compensation (Chaudhary, 2009). In fact, assessments are often tied to shifting power dynamics involving implicit discrimination and limiting resources and can be used to justify the usage of algorithms as mechanisms of oppression (Noble, 2018). Measurements provide views on the data, but these views already encode prior beliefs and assumptions about what can be measured and what ought to be measured. If measures are reported without the right contextualization, they can mislead and provide an undeserved veneer of objectivity to subjective or incomplete evaluations (Paulus et al., 2017). As such, it is important not to abstract the actual methods that are used to derive a measurement and to directly reference definitions and explain any assumptions used to perform a calculation up front (Jacobs and Wallach, 2021). This can be done in tools such as datasheets (Gebru et al., 2021) and data statements (Bender and Friedman, 2018), where measurements can complement other sources of information regarding datasets, such as data sources and metadata.

There are also critical lessons that can be learned throughout history on how measurements have been abused by those in power to shortchange the general public. For example, contributors to the *Cahiers de doléances* published at the start of the French Revolution (1789) detailed the need for standardized measurements that would create fair pay – a grievance that led to the creation of the Metric system López (2020).

5.2 Future Work

We have discussed measurements with examples largely drawing from the ML subfields of computer vision and natural language processing, where data collection norms are common and notions of measurement are relatively straightforward to connect. However, this only begins to scratch the surface of the space of possible data measurements within machine learning research. For instance, in reinforcement learning (RL), where the goal is to enable an agent to maximize a reward signal via trial and error, there are relatively few curated data practices (Sutton and Barto, 2018) or standardized quantification strategies. Yet every task comes with a specific reward function that quantifies the agent’s performance. These may be possible to conceptualize as types of measurements, which

could help with comparing agent actions across different systems. In this example, measuring the data as it returns from the agent has the potential to open new research avenues in how an agent can be improved. Progress in this vein is beginning (cf. the improved statistical analyses proposed in Agarwal et al. (2021)), but substantial opportunity exists to create a comprehensive notion of measurement in line with those developing across other ML sub-fields.

We have also not addressed measurements outside of datasets comprised of a set of static instances. A dataset comes with collection methodology, norms, judgment protocols, prevailing conditions for measurement, and so is more than a set of instances that can be measured as described here. Datasets also often come with metadata (such as time stamps of collection) that present additional avenues of measurement. Other aspects of datasets require documentation and analyses well beyond quantification, such as annotation contexts Bender and Friedman (2018); Gebru et al. (2021).

Our presentation of dataset measurements have also not touched on temporal sequences, a fundamental characteristic of time-series data. There are also many other measurements introduced in the physical sciences (e.g., physics-based measures such as energy, brightness, flow) that could potentially be extended to data. Analysing multimodal datasets, such as the recent large dataset LAION (Schuhmann et al., 2021), presents unique measurement challenges that we also do not address: Different modalities of a dataset can interact to result in novel harms and biases, as illustrated by Birhane et al. (2021). This makes it important to find novel ways of combining existing uni-modal measurements, as well as building upon these to create new measurements that span multiple modalities.

6 Conclusion

We have defined the process of *measuring data* for work in dataset curation and data analysis. The goal of measuring data and datasets is to derive *data measurements*: quantifications of data to understand its composition and the magnitude of different characteristics. The existence of similar lines of research on what we identify as measurement, across fields, suggests the opportunity to foster an interdisciplinary research area on measuring data and datasets, with a shared understanding and common vocabulary. This can aid in understanding what data or a dataset represents, how it compares to other datasets, and how it might be directly improved. We expect that further discussion within research communities will be needed to systematize the definitions and usage of these terms as data measurement develops into an established task.

Previous work makes clear that this task should be approached through perspectives of how to better understand both qualitative and quantitative aspects of data and how to characterize the constructs that the data represents. This is relevant to processes for developing ML and AI systems, as well as for auditing such systems. Several fields and many researchers have provided helpful insights into different kinds of measurements that can be applied to data; here, we have pulled them together alongside work in fairness and ethics advocating for critical dataset analysis. We believe these lines of work together identify the task of measuring data, motivate its importance, and provide methods for calculating measurements.

Similar to previous work in fairness and ethics in machine learning, we believe concerted efforts in terms of more mindful data and dataset analysis are necessary to pave the way towards higher quality models and tools within the ML community. With the ability to measure data, we gain control over what models learn, and can start to account for problematic artifacts of the relatively uncurated data collection approach used to do, such as overrepresentation of a limited set of viewpoints, or problematic social prejudices encoded as facts. Existing research has proposed several ways to support and empower different kinds of analyses, and many measures and metrics described in previous work are complementary, representing different views of a given dataset. Further work is necessary to directly connect these approaches and to develop the task of measuring data more fully. This would fundamentally advance the rigor in creating datasets, and our ability to understand what they represent.

7 Acknowledgements

Thank you to Leon Derczynski and James Vincent for conversations that fundamentally shaped our thinking on this topic.

References

- World in the Balance: The Historic Quest for an Absolute System of Measurement*. 2011.
- R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.
- L. Aitchison, N. Corradi, and P. E. Latham. Zipf’s law arises naturally when there are underlying, unobserved variables. *PLOS Computational Biology*, 12(12):1–32, 12 2016. doi: 10.1371/journal.pcbi.1005110. URL <https://doi.org/10.1371/journal.pcbi.1005110>.
- O. Aka, K. Burke, A. Bauerle, C. Greer, and M. Mitchell. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–335, 2021.
- E. Akyürek, T. Bolukbasi, F. Liu, B. Xiong, I. Tenney, J. Andreas, and K. Guu. Tracing knowledge in language models back to the training data. *arXiv preprint arXiv:2205.11482*, 2022.
- R. Avila-Vázquez. How is SNR calculated in images?, 08 2014.
- D. I. Belov and R. D. Armstrong. Distributions of the kullback–leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology*, 64(2):291–309, 2011.
- E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.
- E. Bender and Gebru, Timnit, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.
- J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.
- Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.
- C. Bentz and R. Ferrer Cancho. Zipf’s law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, pages 1–4. University of Tübingen, 2016.
- D. Binkley, H. Feild, D. Lawrie, and M. Pighin. Increasing diversity: Natural language measures for software fault prediction. *Journal of Systems and Software*, 82(11):1793–1803, 2009.
- M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.
- A. Birhane, V. U. Prabhu, and E. Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.
- T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.
- R. Bommasani and C. Cardie. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, 2020a.
- R. Bommasani and C. Cardie. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online, Nov. 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.649. URL <https://aclanthology.org/2020.emnlp-main.649>.

- S. Boslaugh and D. P. A. Watters. *Statistics in a Nutshell: A Desktop Quick Reference*. O'Reilly Media, Inc., 2008.
- P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, and R. L. Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992.
- N. Campbell. *An Account of the Principles of Measurement and Calculation*. 1928.
- L. Chaudhary. Education inputs, student performance and school finance reform in michigan. *Economics of Education Review*, 28(1):90–98, 2009. ISSN 0272-7757. doi: <https://doi.org/10.1016/j.econedurev.2007.11.004>. URL <https://www.sciencedirect.com/science/article/pii/S0272775708000514>.
- Clean Lab. Clean lab, 2022. URL <https://github.com/cleanlab/cleanlab>.
- M. A. Covington. Idea density — a potentially informative characteristic of retrieved documents. In *IEEE Southeastcon 2009*, pages 201–203, 2009. doi: 10.1109/SECON.2009.5174076.
- K. Crawford and T. Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, pages 1–12, 2021.
- Data-centric AI. Data-centric ai resource hub, 2021. URL <https://datacentricai.org/>.
- DataPerf Workshop. Dataperf: Benchmarking data for better ML, a. URL <https://dataperf.org/>.
- DataPerf Workshop. Dataperf: Benchmarking data for data-centric AI, b. URL <https://sites.google.com/view/dataperf2022>.
- J. De la Rosa, E. G. Ponferrada, P. Villegas, P. G. d. P. Salas, M. Romero, and M. Grandury. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *arXiv preprint arXiv:2207.06814*, 2022.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- J. A. Díez. History of measurement theory. *J. Díez, Hidstory of measurement theory*, 3, 2009.
- F. Estates. *Cahiers de doléances*, 1789.
- S. Eyuboglu, B. Karlaš, C. Ré, C. Zhang, and J. Zou. dcbench: a benchmark for data-centric ai systems. In *Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2022.
- F. Ferraro, N. Mostafazadeh, T.-H. Huang, L. Vanderwende, J. Devlin, M. Galley, and M. Mitchell. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1021. URL <https://aclanthology.org/D15-1021>.
- R. Flesch. How to write plain english. *University of Canterbury*. Available at http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml. [Retrieved 5 February 2016], 1979.
- D. Friedman and A. B. Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.
- T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- A. Gelbukh and G. Sidorov. Zipf and heaps laws’ coefficients depend on language. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 332–335. Springer, 2001.
- C. Gini. *Variabilità e Mutabilità*. 1912.

- K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, jan 2008. doi: 10.1209/0295-5075/81/48002. URL <https://doi.org/10.1209/0295-5075/81/48002>.
- G. H. Golub and C. H. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14:403–420, 1970.
- Google Research. Know your data, 2021. URL <https://knowyourdata.withgoogle.com>.
- M. Grusky, M. Naaman, and Y. Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065. URL <https://aclanthology.org/N18-1065>.
- S. Gururangan, D. Card, S. K. Drier, E. K. Gade, L. Z. Wang, Z. Wang, L. Zettlemoyer, and N. A. Smith. Whose language counts as high quality? measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*, 2022.
- M. A. K. Halliday. *Spoken and written language*. Oxford University Press, USA, 1989.
- L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese. UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 44–52, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL <https://aclanthology.org/S13-1005>.
- L. Hannah and J. A. Kay. *Concentration in modern industry: Theory, measurement and the UK experience*. Springer, 1977.
- G. M. Harrington. Intelligence tests may favour the majority groups in a population. *Nature*, 258(5537):708–709, 1975.
- J. A. Harris. The variable desert. *The Scientific Monthly*, 3:41–50, 1916.
- M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- M. O. Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973.
- J. Hox and H. Boeije. Encyclopedia of social measurement. *Data Collection, Primary vs. Secondary, I*, pages 593–599, 2005.
- A. Z. Jacobs and H. Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.
- A. Jariwala, A. Chaudhari, C. Bhatt, and D.-N. Le. Data quality for AI tool: Exploratory data analysis on IBM API. *International Journal of Intelligent Systems & Applications*, 14(1), 2022.
- E. S. Jo and T. Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316, 2020.
- R. Kannan, H. Woo, C. C. Aggarwal, and H. Park. Outlier detection for text data. In *Proceedings of the 2017 siam international conference on data mining*, pages 489–497. SIAM, 2017.
- D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- W. Kintsch and J. Keenan. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive psychology*, 5(3):257–274, 1973.

- S. Kokoska and D. Zwillinger. *CRC standard probability and statistics tables and formulae*. Crc Press, 2000.
- A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- H. Kucera, W. N. Francis, W. F. Twaddell, M. L. Marckworth, L. M. Bell, and J. B. Carroll. Computational analysis of present-day american english. 1967.
- S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.
- Y.-A. Lai, X. Zhu, Y. Zhang, and M. Diab. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. *arXiv preprint arXiv:2003.08529*, 2020.
- K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.
- G. N. Leech. 100 million words of English: The British National Corpus (BNC). *Language Research*, 1/4, 1992.
- V. Levenshtein. Двоичные коды с исправлением выпадений, вставок и замещений символов. In Доклады Академии наук, volume 163, pages 845–848. Российская академия наук, 1965.
- J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.
- W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.
- D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.
- A. S. Luccioni and D. Rolnick. Bugs in the data: How imagenet misrepresents biodiversity. *arXiv preprint arXiv:2208.11695*, 2022.
- S. Luccioni, Y. Jernite, and M. Mitchell. Introducing the data measurements tool: an interactive tool for looking at datasets, 2022. URL <https://huggingface.co/blog/data-measurements-tool>.
- V. López. How the french revolution created the metric system. 2020. URL <https://www.nationalgeographic.com/history/history-magazine/article/french-revolution-toppled-king-forged-metric-system>.
- A. E. Magurran. *Ecological diversity and its measurement*. Princeton university press, 1988.
- C. Meister and R. Cotterell. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, 2021a.
- C. Meister and R. Cotterell. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*, 2021b.
- G. Melis, C. Dyer, and P. Blunsom. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*, 2017.
- J. Michell. Quantitative science and the definition of measurement in psychology. *British journal of Psychology*, 88(3):355–383, 1997.
- J. Michell. Measurement theory. *Encyclopedia of Social Measurement*, 2005.

- T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- M. Mitchell, D. Baker, N. Moorosi, E. Denton, B. Hutchinson, A. Hanna, T. Gebru, and J. Morgenstern. Diversity and inclusion metrics in subset selection. *CoRR*, abs/2002.03256, 2020. URL <https://arxiv.org/abs/2002.03256>.
- M. Motamedi, N. Sakharnykh, and T. Kaldewey. A data-centric approach for training deep neural networks with less data. *arXiv preprint arXiv:2110.03613*, 2021.
- S. U. Noble. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press, 2018.
- J. Novikova, O. Dušek, A. C. Curry, and V. Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.
- A. Nunes, T. Trappenberg, and M. Alda. The definition and measurement of heterogeneity. *Translational psychiatry*, 10(1):1–13, 2020.
- B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network: computation in neural systems*, 7(2):333, 1996.
- D. Y. Orlova, N. Zimmerman, S. Meehan, C. Meehan, J. Waters, E. E. Ghosn, A. Filatenkov, G. A. Kolyagin, Y. Gernez, S. Tsuda, et al. Earth mover’s distance (emd): a true metric for comparing biomarker expression levels in cell populations. *PLoS one*, 11(3):e0151859, 2016.
- A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.
- F. M. Paulus, N. Cruz, and S. Krach. The impact factor fallacy. *Frontiers in Psychology*, 9, 2017.
- S. Peleg, M. Werman, and H. Rom. A unified approach to the change of resolution: Space and gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):739–742, 1989.
- L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.
- E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195, 2008.
- R. C. Prati, G. E. Batista, and M. C. Monard. Data mining with imbalanced class distributions: concepts and methods. In *IICAI*, pages 359–376, 2009.
- F. Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI’2000 workshop on imbalanced data sets*, volume 68, pages 1–3. AAAI Press, 2000.
- G. Pruthi, F. Liu, S. Kale, and M. Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020.
- L. E. Raileanu and K. Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93, 2004.
- N. Rajani, W. Liang, L. Chen, M. Mitchell, and J. Zou. Seal : Interactive tool for systematic error analysis and labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2022. URL <https://arxiv.org/abs/2210.05839>.
- P. H. Ramsey. Critical values for spearman’s rank order correlation. *Journal of educational statistics*, 14(3):245–253, 1989.
- A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017.

- A. Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA, 1961.
- F. S. Roberts. *Measurement theory*. 1985.
- Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover’s distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.
- T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL <https://doi.org/10.1145/3411764.3445518>.
- M. K. Scheuerman, A. Hanna, and E. Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3476058. URL <https://doi.org/10.1145/3476058>.
- C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946. doi: 10.1126/science.103.2684.677. URL <https://www.science.org/doi/abs/10.1126/science.103.2684.677>.
- C. Strobl, A.-L. Boulesteix, and T. Augustin. Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis*, 52(1):483–501, 2007.
- L. V. Subramaniam, S. Roy, T. A. Faruque, and S. Negi. A survey of types of text noise and techniques to handle noisy text. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND ’09, page 115–122, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584966. doi: 10.1145/1568296.1568315. URL <https://doi.org/10.1145/1568296.1568315>.
- R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- M. Thomas and A. T. Joy. *Elements of information theory*. Wiley-Interscience, 2006.
- V. To, S. Fan, and D. Thomas. Lexical density and readability: A case study of english textbooks. *Internet Journal of Language, Culture and Society*, (37):61–71, 2013.
- A. Toral, P. Pecina, L. Wang, and J. Van Genabith. Linguistically-augmented perplexity-based data selection for language models. *Computer Speech & Language*, 32(1):11–26, 2015.
- J. Vanschoren and S. Yeung. Announcing the neurips 2021 datasets and benchmarks track, 2021.
- J. Vig, W. Kryściński, K. Goel, and N. F. Rajani. Summvis: Interactive visual analysis of models, data, and evaluation for text summarization. *arXiv preprint arXiv:2104.07605*, 2021.
- J. Vincent. *Beyond Measure: The Hidden History of Measurement*. Faber & Faber, 2022.
- S. E. Whang, Y. Roh, H. Song, and J.-G. Lee. Data collection and quality challenges in deep learning: A data-centric ai perspective. *arXiv preprint arXiv:2112.06409*, 2021.
- Y. Xiao, J. Fu, W. Yuan, V. Viswanathan, Z. Liu, Y. Liu, G. Neubig, and P. Liu. Datalab: A platform for data analysis and intervention. *arXiv preprint arXiv:2202.12875*, 2022.

J. Zhu, H. Wang, T. Yao, and B. K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK, Aug. 2008. Coling 2008 Organizing Committee. URL <https://aclanthology.org/C08-1143>.