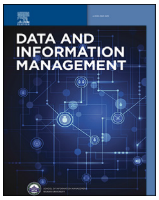
Data and Information Management 6 (2022) 100020

Contents lists available at ScienceDirect

Data and Information Management

journal homepage: www.journals.elsevier.com/data-and-information-management



The illusion of data validity: Why numbers about people are likely wrong

Bernard J. Jansen [a,*], Joni Salminen [b], Soon-gyo Jung [a], Hind Almerekhi [a,c]

[a] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[b] School of Marketing and Communication, University of Vaasa, Vaasa, Finland
[c] College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

ARTICLE INFO

Keywords:
People data
Measurement
Quantitative paradigm
Statistics

ABSTRACT

This reflection article addresses a difficulty faced by scholars and practitioners working with numbers about people, which is that *those who study people want numerical data about these people. Unfortunately, time and time again, this numerical data about people is wrong.* Addressing the potential causes of this wrongness, we present examples of analyzing people numbers, i.e., numbers derived from digital data by or about people, and discuss the comforting illusion of data validity. We first lay a foundation by highlighting potential inaccuracies in collecting people data, such as selection bias. Then, we discuss inaccuracies in analyzing people data, such as the flaw of averages, followed by a discussion of errors that are made when trying to make sense of people data through techniques such as posterior labeling. Finally, we discuss a root cause of people data often being wrong – the conceptual conundrum of thinking the numbers are *counts* when they are actually *measures*. Practical solutions to address this illusion of data validity are proposed. The implications for theories derived from people data are also highlighted, namely that these people theories are generally wrong as they are often derived from people numbers that are wrong.

1. Introduction

Let us begin our journey into people numbers by considering a well-known quote about science (Kelvin, 1883).

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science …. —Lord Kelvin

This insightful quote is a homage to the central role of numbers in the scientific method, stating quite explicitly that you must have numbers in order to have knowledge. Leaving aside the obvious irony that this quote is expressed not in numbers but in words, the position reflects the view of many, if not most, researchers in the scientific disciplines that *numbers are essential for scientific inquiry*. If there are no numbers, then it is not really science, and we see this need for numbers in studies of and about people in both the physical and online worlds.

The need for numbers is associated with a core tenet of science, namely the principle of falsifiability (Popper, 2002), which is the construct that science must be inherently disprovable, or it is not science. While the principle of falsifiability suffers a similar shortcoming of many theories outside the physical realm in that it cannot explain itself (i.e., to our knowledge, the principle of falsifiability is non-falsifiable), it is instrumental in the development and wording of hypotheses in such a way as aiming to disprove them. For the construction of hypotheses, numbers are quite beneficial. For example, take the hypothesis: *The Beatles is the greatest band of all time.* This statement does not pass the principle of falsifiability. It is non-falsifiable. There are no numbers. Now, take this reconstructed hypothesis: *The Beatles has more Billboard Number 1 hit songs than any other artist.* This hypothesis is falsifiable. It focuses on numbers. We just need to get all the hit songs from the top of the Billboard charts, sum by artist, rank the sums most to least, and determine which artist has the most number one hits. Spoiler – it *is* The Beatles! as of the time of this study (Billboard, 2022).

Although the qualitative approach is widely used (Silverman, 2020), numbers have a central place in the employment of the scientific method, and this centrality applies to research focused on people. Undeniably useful in many situations, the use of numbers can lead to a type of ‘illusion of validity’ (Kahneman & Tversky, 1973) when employing people numbers, i.e., numbers at scale derived from data created by or

[*] Corresponding author.
E-mail address: jjansen@acm.org (B.J. Jansen).

https://doi.org/10.1016/j.dim.2022.100020
Received 15 July 2022; Received in revised form 16 September 2022; Accepted 22 September 2022

about people. The illusion of validity is a cognitive bias of being overconfident in the accuracy of interpretations regarding a given dataset. In this article, we present another illusion, the illusion of validity concerning the data, in that you believe the outcome of your analysis because you believe in the accuracy of the data. However, when the underlying data about people is flawed, this belief results in *the illusion of data validity*.

We denote data about users, visitors, customers, audience members, or other population segments or subgroups (i.e., people) as 'people data', and numbers derived from this people data as 'people numbers'. People data is often collected at scale, usually online, often involving a proxy variable, and perhaps entailing labeling via human or automated means. *Online people data*, in particular, provides opportunities for modeling social phenomena using digitalized people data at scale. Our focus is primarily on this digital data about people, and the data can refer to various aspects of people, including their online behaviors and attitudes. People data is used for multiple purposes, such as opinion mining, sentiment analysis, trend detection, election prediction, user segmentation, and so on. Examples of people data include content generated by people, such as social media postings, online product reviews, and comments on online platforms, and content about people, such as website traffic numbers, metrics concerning online content, and analytics about people's online behaviors. While problems in survey data (e.g., respondent bias, confirmation bias, etc.: Bertrand & Mullainathan, 2001) and unstructured textual data (e.g., noise, need for preprocessing, and labeling: Kaisler et al., 2013) are known, in our experience, researchers and practitioners are much less aware of the challenges in structured people data, such as social media usage statistics (e.g., number of video views, likes, comments, shares) and website usage statistics (e.g., number of sessions, bounce rate, conversions, etc.). These are examples of people numbers based on or derived from people data. In combination with the application of the scientific method to the analysis of this people data, researchers and others tend to believe the findings, which are nearly unquestioned as being accurate (Chinn & Brewer, 2001). However, *what if the numbers derived from this people data are wrong?*

This question motivates our treatise, and the implications are quite stark if people data and people numbers are wrong. If your people numbers are wrong, it does not matter how rigorous your statistical method is, how state-of-the-art your algorithmic approach is, or how well you fine tune your machine learning (ML) model. These approaches can obviously provide clear and precise results. However, a clear and precise result does not mean a correct one – one can be clear, precise, and wrong! However, despite this obviousness, there is a tendency to ignore the potential flaws in people numbers, focus on tuning the model, perfecting the methods (Post & Votta, 2005), and trusting the results, thereby manifesting the 'illusion of data validity'. Our premise for this article is that, more often than not, people numbers are likely to be wrong. By wrong, we mean that these numbers are often presented and perceived as if they are precise and accurate when they are not.

Validity is often discussed in terms of methods or instruments (Kimberlin & Winterstein, 2008); specifically how correct a method is in measuring something, where correctness is the degree to which the results conform to the truth values. Therefore, validity is often understood as the quality of a method or instrument being correct. However, many data collection issues can degrade the validity of findings about people. Data validity is how accurate the data is, where accuracy is the degree to which the data conforms to true values (Boslaugh & Watters, 2008). If the data has high validity, it means that the values in the dataset correspond to true properties in the physical (or virtual) world (Diaconis & Efron, 1983). If the numbers do not correspond to the actual values, there is a data validity issue. The data validity is low.

There are several factors that can affect data validity, and we argue that much of the findings from people research are wrong because the employed people data has low data validity. We argue that the data does not accurately conform to the true values in the physical or virtual world, or to the perceived accuracy of those using this people data. Overall, the issue of data validity is often overlooked by those working with people data, especially when the dataset size is big (Heckman, 1979; Tufekci, 2014). We provide examples that support our view and discuss partial solutions. Aside from scholars, the issue of numbers being wrong matters to many organizational stakeholders, for example, those making predictions and decisions, or building ML or artificial intelligence (AI) systems based on these people numbers (Vecchio et al., 2018). Any inaccuracies in the people data can propagate decision-making bias, result in suboptimal resource allocation, and negatively affect performance. Importantly, the AI system can propagate, not correct, errors in people data, and given that there is ample evidence to show that most predictions about people are wrong (Anderson, 2000; Epstein, 2019; Ioannidis, 2005; Silver, 2015), this makes discussion and reflection on this matter extremely important.

In the next section, we discuss some foundational data issues that often emerged when there were no web analytics, AI, or digital data collection, and statistics were counted by hand or using rudimentary hardware. These data issues have not disappeared, even amidst digital data collection and analysis technologies – they still prevail, take new forms, and continue to teach us lessons about what matters in the pursuit of knowledge via people numbers.

2. Foundational data issues

To lay the groundwork for presenting our specific concern with people numbers, we highlight a non-exhaustive list of issues that threaten data validity. We highlight these issues in an effort to weaken the reader's faith in the validity of people numbers, which we believe is needed due to overconfidence in these numbers in industry and among researchers relying on 'data' as the answer (Siegel, 2010).

False Proxy: One often employs proxies in research. A proxy is a variable that serves in place of another variable that one seeks to investigate but, for some reason, cannot (Becker et al., 2016). Often, when you cannot access the variable that you want, you select something else that is easier to access, and this new variable represents an approximation of what you actually want to investigate. For example, eye fixations on a screen are proxies for cognitive attention (Salminen, Jansen, et al., 2018a) because cognitive attention itself is difficult to access. However, proxies can be false – called a *false proxy*. A false proxy is a seemingly objective proxy that inaccurately reflects the variable that you intend to investigate (Timberg, 2006). For example, online display advertising platforms cannot directly determine the viewers that are impacted by an ad. So instead, these platforms report how many times the display ad appeared on a screen, called an impression. But an impression does not determine impact, and in fact, it does not even determine views. Furthermore, false proxies have the additional problem of encouraging focus on the proxy instead of the real variable of interest.

Selection Effect: When conducting research, one has to make choices concerning sampling the population of interest. However, these choices introduce a potential problem in data collection called the *selection effect* or *selection bias* (Infante-Rivard & Cusson, 2018). When a sample is biased toward a specific subset of a target population, the data does not reflect that actual population. For example, suppose you start a business distributing online discount coupons for ordering from a restaurant website, and you get paid based on the number of coupons redeemed at the restaurant. Who are the best people to provide coupons to in order to maximize your return? The best group of people would be those who were already going to the restaurant website to place an order! Your success rate would be great analytically, but your effort of distributing the coupons would not increase the restaurant visits (at least to the extent you may believe). Many online advertising and discount coupons suffer from selection bias (Blake et al., 2015). These methods can, if not appropriately managed, target people who are already going to purchase the product anyway, giving the impression of great numerical returns. For other examples using three social media platforms at the time of this study, YouTube is known to be biased in favor of young males (Bughin,

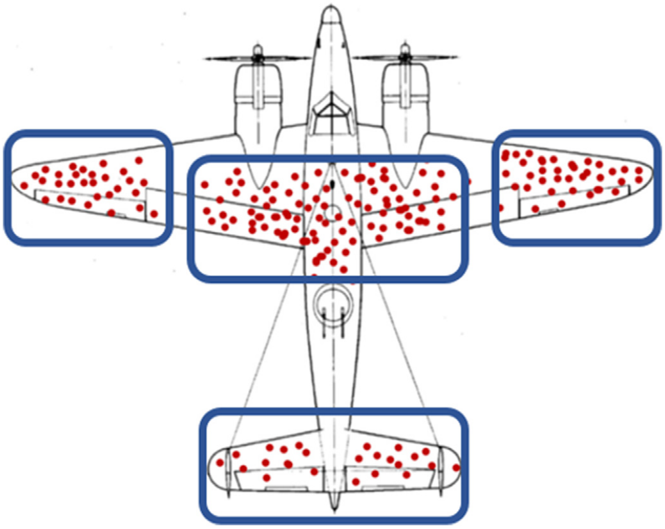


Fig. 1a. The red dots represent bullet holes in places where the planes had been shot. The blue boxes highlight the areas of interest. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

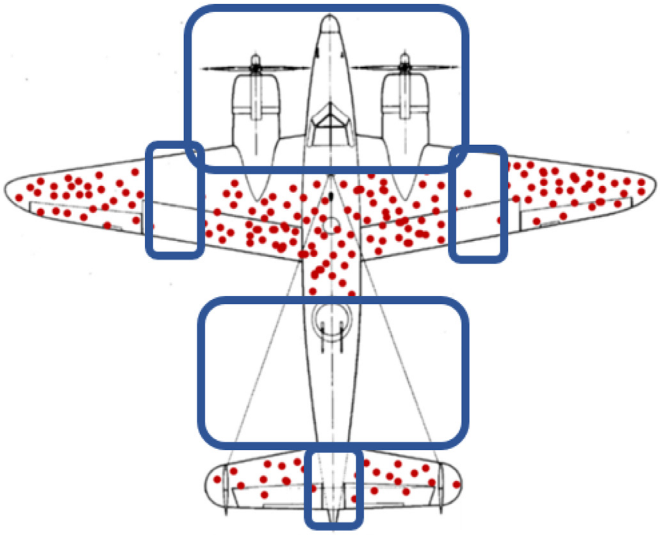


Fig. 1b. The recommendation was to armor the places where there were no bullet holes. The blue boxes highlight the areas of interest, e.g., figures reproduced after Survivorship bias (2022). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2010), Pinterest in favor of young females (Blank & Lutz, 2017), and TikTok in favor of Generation Z (Cervi, 2021). Nevertheless, despite these sampling biases taking place in platforms used for research, these biases are typically not rectified (or even recognized) when analyzing user behavior (Ruths & Pfeffer, 2014), and the consequence is a distortion of a statistical analysis due to a biased sampling method.

Missing Data: To conduct research, one typically relies on data. Data concerning one's research can be conceptually divided into two categories – *data you have* and *data you do not have*. The data you do not have is *missing data*. If you can identify the missing data, you can perhaps account for it through, for example, simulated data. For example, if people self-report their socio-demographic information (e.g., age and gender) to online platforms, this information can be directly used (albeit that it also potentially involves a self-reporting bias). When the person provides no such information, the platforms can auto-complete it using ML estimation (Ruths & Pfeffer, 2014), although the information is probabilities (Jansen et al., 2013). However, if you are unaware of missing data (i.e., unknown missing data), this can significantly affect your research results and implications, as well as foreseeing highly improbable events (Taleb, 2007). While not directly concerning people data, a classic missing data problem involves the United States (U.S.) Statistical Research Group during World War II (Ellenberg, 2015; Mangel & Samaniego, 1984). The U.S. Army Air Corps asked the research group to figure out how much to armor the planes so they would not get shot down and still be fast, and the Corps provided data on where planes were being shot (see Fig. 1a). What was the recommendation? The recommendation was to armor the planes in all areas with no bullet holes (see Fig. 1b). Here, the researcher correctly realized that there was a missing data problem (Rubin, 1976), where the data from the Corps were from the planes that made it back, and the missing data represented the planes that did not make it back! Missing data is, therefore, a data collection problem – *you did not collect the range of data you need to address your research.*

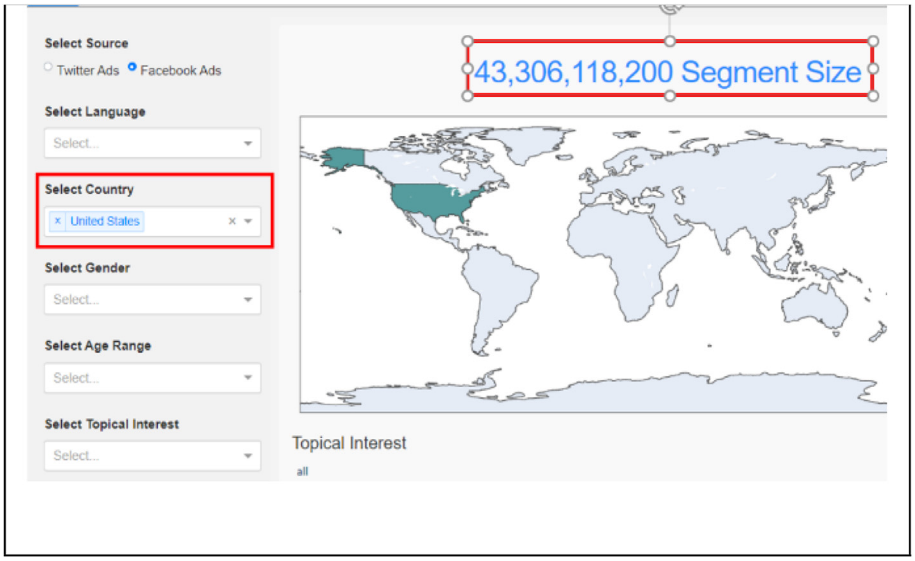
Dimensionality Curse: The curse of dimensionality (Bellman, 1961) states that the number of samples needed to estimate a function grows exponentially with respect to the number of input variables of the function. The *curse of dimensionality with people data* is that, as the number of features increases, the number of people represented by the set of features rapidly decreases (for experiment results in this area, see Chapman et al., 2008). The number of attributes and the number of people represented by those attributes are inversely correlated. To illustrate this point, the SegmentSizeEstimator[1] is a tool that generates population sizes based on Facebook or Twitter accounts, as of the time of this study. As shown in Fig. 2a, there were reportedly more than 43.3 billion Facebook accounts in the USA on the date we executed this search. However, as shown in Fig. 2b, by adding just four attributes (language, sex, age, interest), we are reduced to just 2.2 million accounts, a decrease of 99.9949%. The curse of dimensionality is therefore a data construction problem – *you have too many features for the size of your dataset.*

Flaw of Averages: When working with numbers representing many people, one often employs the average to describe the people represented by the numbers. However, this approach can lead to serious problems due to the *flaw of averages* (Brown et al., 2018), which is that findings based on the average are wrong on average. In fact, often when dealing with people, the average person does not exist! An interesting example of the flaw of averages is from a study in the 1950s. The U.S. Air Force designed cockpits of new aircraft based on the average physical measurements of 4000 pilots (Daniels, 1952). What was the result of this seemingly reasonable approach? Pilot complaints and several airplane crashes were, unfortunately, the result. Investigating the possible cause of crashes and complaints, the researcher involved in the original study of collecting pilot physical measurements examined the 10 most important metrics, and compared the averages to the 4000 individual pilots. What were the finding from this follow-on analysis? Not a single pilot in the 4000 sample perfectly matched the average! Quoting from Lieutenant Daniels:

> As an abstract representation of a mythical individual most representative of a given population, the ‘average man’ is convenient to grasp in our minds. Unfortunately, he doesn’t exist. Instead of being the easiest individual of a group to provide for, and the most common, the “average man” is in reality a very rare specimen and very hard to fit. (Daniels, 1952, p. 2, p. 2)

The average applies to a collection, group, population, segment, sample, etc., yet is nearly meaningless at the individual level and hides what is usually a distribution (Savage & Markowitz, 2012). The flaw of averages is wrong in most contexts but nearly always wrong when

[1] Segment size generated from https://acua.qcri.org/tool/SegmentSizeEstimator.



Select Source
○ Twitter Ads ● Facebook Ads

Select Language
Select...

Select Country
× United States

Select Gender
Select...

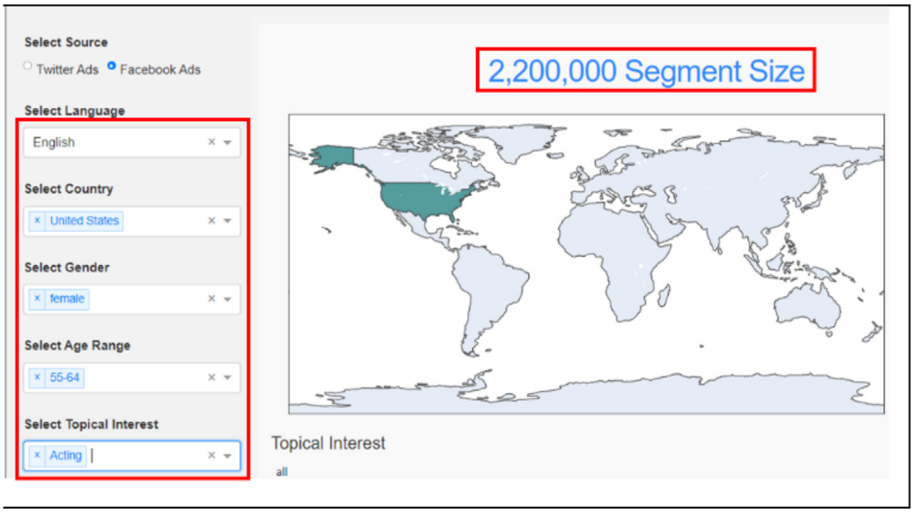
Select Age Range
Select...

Select Topical Interest
Select...

43,306,118,200 Segment Size

Topical Interest
all

Fig. 2a. Number of Facebook accounts in the U.S., reportedly, at the time of study, more than 43.43 billion accounts.



Select Source
○ Twitter Ads ● Facebook Ads

Select Language
English

Select Country
× United States

Select Gender
× female

Select Age Range
× 55-64

Select Topical Interest
× Acting

2,200,000 Segment Size

Topical Interest
all

Fig. 2b. Number of Facebook accounts in the U.S. at the time of study, with the addition of four attributes (English language, female gender, 55–64 age range, and topical interest in ‘Acting’). A reduction of 99.9949%.

dealing with many types of online people data. This is because many, but naturally not all, methods relying on the average assume a normal distribution, whereas much of the online people data follows a power law distribution. In power law distributions, the data is so skewed that a large percentage of the population would deviate from any ‘average’. The flaw of averages is a data analysis problem – *the average person does not exist.*

Deception of Statistics: You might think the flaw of averages is a straw person issue that can be remedied by using more aggregate metrics, such as standard deviation (as in ANOVA testing) or median (as in chi-squared testing). Although a soothing position, you would be incorrect, as aggregate metrics can also be extremely misleading. Statistics alone about a dataset do not adequately depict the data set. One example is Anscombe's quartet, comprising four data sets with practically identical descriptive statistics yet having different distributions and appearing very different when graphed (see Fig. 3). Statistician Francis Anscombe constructed the quartet in 1973 to demonstrate the importance of graphing data when analyzing it and the effect of outliers and other influential observations on statistical properties (Anscombe, 1973; Chatterjee & Firat, 2007; Matejka & Fitzmaurice, 2017). Of course, one has to be careful when relying only on graphical representations (Jones, 2006), as graphs can also be deceiving. The deception of statistics is a data analysis problem – *perusing only statistics does not tell the complete story*.

Do you need a real life example of the hazards of looking only at statistics and not the distributions? Look no further than the AAirPass program from American Airlines (Oyer, 2014), which ran from 1981 to 1993. For about $1,000,000 USD, you could get two business class airline tickets anytime to anywhere for life at no cost, earn miles, and have a dedicated company agent handle your flight arrangements. Because most people do not fly that much, and even the most frequent flyers would take a long time to rack up a million dollars’ worth of flying, by some statistical metrics (e.g., median or mean), the AAirPass was a good deal for American Airlines. However, what about outliers? Here, outliers are the people who fly a disproportionate number of times – *those who fly a lot!* People data typically have outliers, and airline people data has some extreme outliers (Thirumuruganathan et al., 2021). Also, changes to services can alter behavior (Jung, Salminen, & Jansen, 2022), in this case, by encouraging people to fly more since the flights are now free to those who purchased the passes. American Airlines reportedly sold just under 30 of these passes before ending the program; the AAirPass

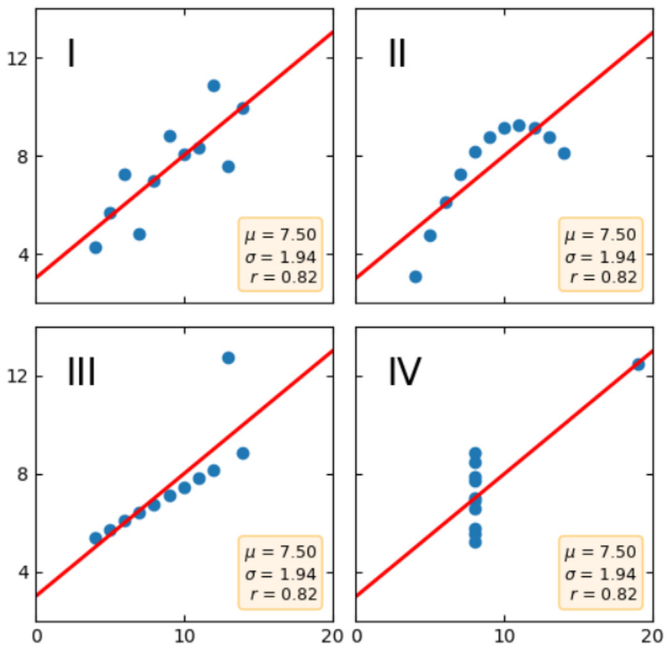


Fig. 3. A depiction of Anscombe's quartet, four data sets that have practically identical descriptive statistics yet have very different distributions.[21]

program has since cost the company tens of millions of dollars in costs and lost revenue, with some passengers racking up millions of miles per year, and, as of this writing, the program is still costing the company. The purchaser of just one pass cost the airline more than $21M USD (Briquelet, 2012), and others have reported it as one of the best purchases they had ever made (Freeman, 2018).

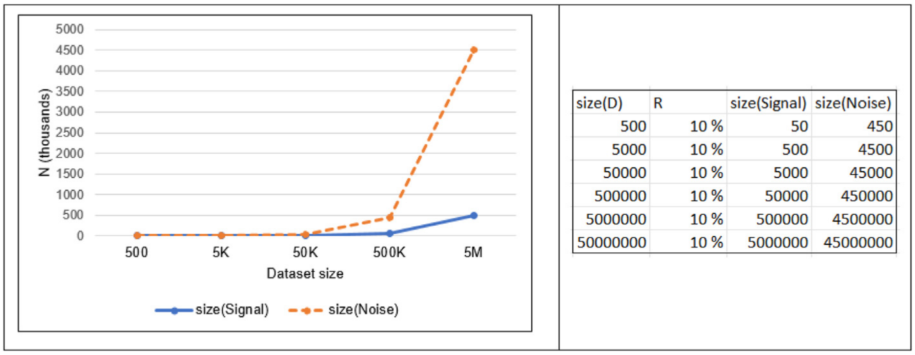
Simpson's Paradox: Simpson's Paradox (Simpson, 1951) is a phenomenon where a statistical relationship between two variables emerges, disappears, or reverses when the variables are examined for sub-populations. For example, one may have a treatment for a given disease that is non-effective when looking at the entire sample. However, when segmenting the sample, say by biological sex, the treatment is effective for all sexes! The paradox is that based on the entire sample, you would not recommend the treatment, but you would recommend the treatment for every sub-population in the sample. Several articles discuss Simpson's Paradox, along with the varieties and why it occurs (Ameringer et al., 2009; Blyth, 1972; Hernán et al., 2011; Wagner, 1982). Lerman (2017) provides multiple examples of how grouping the data can yield drastically different findings concerning social media behavior. Even though there are methods to test for Simpson's Paradox (Kievit et al., 2013; Lerman, 2017; Pearl, 2022), these are rarely used in research at the time of this writing, perhaps because they require ‘going out of your way’ to find potential problems in your data (a practice most people tend to avoid; for a discussion of this human behavior, see the many writings on either the Principle of Least Effort or satisficing).

GroundTruth or CrowdTruth Problem: Online people data is often in the form of text, such as social media posts. So, to work with it or convert it to numbers, researchers often label this textual people data via human annotation (Aroyo & Welty, 2013; D. Wu et al., 2013), such as by topic, sentiment, or toxicity, in order to map the labels against some numbers they can work with. Labeling is often done by having multiple labelers evaluate a snippet of online text to assign a label (Snow et al., 2008) concerning, for example, whether or not a social media post is toxic (Almerekhi et al., 2020, pp. 3033–3040). This is *posterior labeling* – classifying non-numerical people data using some theoretical construct after the fact to achieve some ‘ground truth’, relying on the supposed wisdom of the crowd (Surowiecki, 2005). However, crowd labeling is not crowd wisdom. Posterior labeling has many pitfalls (for a discussion of the myths of human annotation, see Aroyo & Welty, 2015), including ‘What if there is no ground truth?‘. Many labeling tasks try to fit labels to people data where the labels may be inappropriate, or the labels are subjective (Alonso et al., 2015; Wiebe et al., 1999). Another issue is quality control, and many publicly available ‘ground truth’ datasets are riddled with obvious errors (e.g., Chen, 2022). For example, in an evaluation of hate data sets in preparation for research (Salminen, Veronesi, et al., 2018b), there was a mislabeling of seven percent to thirty percent seen in the available ‘gold standard’ datasets (Waqas et al., 2019), including many highly used hate datasets. Others point out the issues of evaluating labeled data across models (Fortuna et al., 2021), and similar issues appear in other domains, like recommender systems (Dacrema et al., 2021). There is also the issue of random labels with majority voting schemes (Salminen et al., 2021). The GroundTruth or CrowdTruth problem is, therefore, a data creation problem – *posterior labeling can result in low data validity*.

Big Data Fallacy. The law of large numbers argues that the sample's mean approaches the sample population's actual average as a sample size increases (Kwak & Kim, 2017). This concept is often, either implicitly or explicitly, taken as a justification as to why ‘big data’ (i.e., millions or billions of samples) cannot be wrong. However, there are contrary arguments and evidence. The big data fallacy implies that more data does not translate to more information in equal measure (Wu, 2012). A reason for this can be a redundancy (i.e., replication of the same data, such as retweets) that decreases the signal-to-noise ratio (i.e., the amount of useful information compared to non-useful information) (Aizawa, 2003), but the growth of a dataset is problematic even when the signal-to-noise ratio remains constant. For instance, assume a signal-to-noise ratio, R, and a dataset D that consists of two subsets, S for signal and N for noise. As D increases and R remains constant, both S and N increase, but the absolute number of N quickly becomes overwhelming (see Fig. 4). The implication is that if an error occurs in a small sample of data, making the sample ‘big’ does not mystically eradicate this error. Lazer et al. (2014) refer to ‘big data hubris’ (p. 1203), meaning that some believe large enough datasets are enough to not consider whether the data meet the assumptions of a given method – if predictions on a test set work, then traditional concepts about data (e.g., distributions and variable dependencies) do not matter. However, this is a gross mistake that can lead to developing ‘part flu detectors, part winter detectors’ (Lazer et al., 2014), for example.

Subjective Objectivity: Data-driven decision making (Provost & Fawcett, 2013) is the process of making decisions based on data rather than intuition, and is (supposedly) employed by nearly every business, organization, and governmental agency. Metrics can suffer from an issue of performativeness, in that metrics can guide human action similarly to words (Lemmon, 1962) - numbers are transferred to reality through interpretative decision making. However, the use of data for driving decision making can give not true objectivity but the perception of objectivity, or *subjective objectivity*, which is assessing data based on opinion or feeling using one's own perspective or preference. For example, suppose a researcher conducts a large-scale survey analysis of faculty at several major universities, including collecting salary data. After the study, the researcher presents this result, ‘Senior professors earn 25% more than their junior counterparts.’ What do you expect the general reaction to be? Most would expect the reaction to be something along the lines of ‘That's reasonable’. The researcher then presents this result from the study. ‘Male professors earn 25% more than their female counterparts.’ What do you expect the general reaction to be? Most would expect the reaction to be something like, ‘That's an outrage!‘[3] This is an example of subjective objectivity – we believe data-driven decision making is objective. However, in reality, nearly every decision has objective and subjective elements, i.e., a degree of judgment (Kahneman et al., 2021), and the interpretation of numbers is nearly always “founded on personal impressions of phenomena” (Bowley, 1901, p. 6). For example, Zgraggen,

[3] We believe rightfully so.



size(D)	R	size(Signal)	size(Noise)
500	10 %	50	450
5000	10 %	500	4500
50000	10 %	5000	45000
500000	10 %	50000	450000
5000000	10 %	500000	4500000
50000000	10 %	5000000	45000000

Fig. 4. Fictitious data showing the effect of the absolute number of noisy samples increasing with the overall size of the dataset. R is signal-to-noise ratio.

Table 1
Foundational data validity issues and their definitions.

Data Validity Issue	Definition
False Proxy	A seemingly objective proxy that inaccurately reflects the variable that you intend to measure.
Selection Effect	A sample is biased toward a specific subset of a target population; the data does not reflect the actual target population.
Dimensionality Curse	As the number of features increases, the number of people represented by the set of features rapidly decreases.
Flaw of Averages	The assumption that the average applies to an individual; the average person does not usually exist.
Missing Data	The data concerning a phenomenon is not complete.
Deception of Statistics	The use of aggregate metrics does not fully describe a dataset.
Simpson's Paradox	A statistical relationship between two variables emerges, disappears, or reverses when the variables are examined for subpopulations.
GroundTruth or CrowdTruth Problem	Trying to fit labels to people data where the labels may be inappropriate or subjective.
Big Data Fallacy	More data does not translate to more information in equal measure.
Subjective Objectivity	Assessing data based on opinion or feeling using one's own perspective or preference without acknowledging this subjective element.

Zhao, Zeleznik, and Kraska (2018) investigated how people generate insights from data visualizations. In the experiment, more than sixty percent of the insights users generated were false. Similar to the Gestalt theory from cognitive psychology, where people see patterns in data that do not exist (Desolneux et al., 2007), subjective objectivity implies that data interpretation nearly always includes an interpretative component that is often not acknowledged.

We summarize these threats to data validity in Table 1 immediately below.

3. Conceptually counting, mathematically measuring – why people numbers are often wrong

Even with the aforementioned validity issues of data collection, analysis, and interpretation, our premise is that there is a yet more fundamental problem with many people numbers. Specifically, we are referring to the situation where the people numbers are just wrong. In our experience, working with many people datasets, collecting people data, and generating people numbers from said data, people numbers can have a low validity in terms of representing what the data conceptually should represent, or the values of that data. Inappropriately, people numbers are often presented as being valid, precise, and arithmetic accurate. With this, researchers and others often perceive the people numbers are valid, precise, and arithmetic accurate when they are not. This presentation and perception fallacy results in findings that are inaccurate and implications from said findings that are faulty.

Let us take some examples from the 2020 U.S. elections. The U.S. holds major elections every four years for the president, congress, governors, and many other down ballot offices. The U.S. has two major political parties at the time of writing, the Democrats – the Blue party, and the Republicans – the Red party. Polling data prior to the 2020 elections, at all levels, showed a blue wave (i.e., the Blue Party would score major wins). Yet, the poll numbers were wrong in this regard, and the Blue Party eked out a narrow victory (Keeter et al., 2021). What about other data sources? Analysis of social media data prior to the same 2020 U.S. presidential election showed major support for the Red Party (Sharma et al., 2022). The social media analysis was also wrong in this regard, where the social media volume of posts was inversely correlated with votes. The Blue Party got more votes than the Red Party. Think of all the social media research and articles published from that data - what if most of it is wrong? In another case, economic data on purchases of presidential campaign merchandise prior to the same 2020 election showed major support for the Red Party (Bradsher, 2020). The economic data was also wrong in this regard, and the sales of merchandise were inversely correlated with votes; the Blue Party got more votes than the Red Party.

All of these examples are based on numbers derived from people data. This data was then used to make recommendations and predictions, and the predictions turned out to be wrong. So what could cause people data (drawn from multiple sources – polls, social media, economics, online analytics, etc.) about the same event all to be wrong?

Our premise is that there is an underlying data validity issue with most of this people data that directly affects the soundness of people numbers derived from them and, therefore, the decisions made (and theories) based on these numbers. The issue is that there is a misconception about what the people numbers are. Namely, these people numbers are counts, when in fact, people numbers are most often measures when online data or data at scale is involved. Counting is adding the number of items in a group to determine the quantity – an arithmetic operation. When counting, one ends with a precise numerical value of the number of items – a count. Measuring is assessing the amount of something – a calculating or estimating operation. A measure is a numerical value of the extent of something. Unlike a count, a measure has an error rate, and in situations like the close election presented earlier, error rates can matter a lot! With people numbers, thinking you are counting when you are really measuring or that the numbers presented are counts when they are really measures, can result in an illusion of data validity. The numbers appear valid, precise, and accurate, but they are not.

Consider the following question, *does a large social media platform, like Facebook at the time of the study, know the exact number of its users at a given time?* No, Facebook does not (Greenspan, 2019; Morse, 2019). Despite Facebook being able to digitally monitor its user base (a medium that one

[2] Plots generated from https://matplotlib.org/stable/gallery/specialty_plots/anscombe.html.

thinks would solve the counting problem), it does not know the exact number of its users at a given time. Neither does Twitter (Bradshaw, 2022), another large social media platform at the time of the study. This is because of a myriad of reasons: for example, new users join and old ones exit continuously, making it hard to track event accounts (which is a proxy variable for users) in real-time. Also, there is a non-trivial number of bot and fake accounts; some of these Facebook can detect and remove, but not all. Finally, what is a 'user' anyway? If a person has not logged in to an account for two months, should they still be included in the user base? What if they have not logged in for two years? What if the person is deceased? Such cases illustrate the difficulty of obtaining correct 'counts', even where people can be tracked digitally in a relatively controlled ecosystem.

4. Case study illustrating people numbers being wrong

To illustrate that numbers derived from people data are often not counts but instead measures, we present a case study of people numbers using website traffic metrics. The results of this research are reported in full in Jansen et al. (2022), but here, we present what is needed to highlight the 'count vs. measure' misconception and the errors resulting from this misconception. The case study is in the area of web analytics; a sub-area of analytics that examines web traffic data. Where do you get web traffic data? Approaches to collecting website analytics can be grouped by the focus of data collection, namely: (a) user-centric, (b) website-centric, and (c) network-centric (Jansen et al., 2022). User-centric data is gathered via a panel of users, which is tracked by software installed on the users' computers, such as a plugin for a web browser. Website-centric data is gathered via software on a specific website. Most websites use a site-centric approach for analytics data gathering, typically employing cookies and/or tagging pages on the website. Network-centric data is gathered via observing and collecting traffic in the network. There are various techniques for network-centric analytics gathering, with the most common being data purchased or acquired directly from Internet service providers.

Google Analytics is a website-centric analytics platform and is the most popular site analytics tool in use (W3Techs, 2020) at the time of the study. Google Analytics tracks and reports website analytics for a specific site. SimilarWeb is a service providing web analytics data for one or multiple websites, and an industry leader in this area at the time of the study. SimilarWeb uses a mix of user, site, and network-centric data collection approaches to triangulate data (Salkind, 2010; SimilarWeb, 2022b) but relies heavily on its user panel data.

Our motivational research question was, *how do the analytics compare between the two services?* For our analysis, using the Majestic Million (a creative commons list of 1 million websites), we identified 86 sites with both Google Analytics and SimilarWeb metrics (SimilarWeb, 2022a, 2022c; 2022a). We then collected the data for four core web analytics metrics – *total visits, unique visitors, bounce rate,* and *average session duration* – for each of the 86 websites. The definition of each metric is:

- *Total Visit* - Sum of times that at least one website page was loaded during a measurement period.
- *Unique Visitors* - Sum of actual people who have visited a website at least once during a period.
- *Bounce Rate* - A bounced visit is the act of a person immediately leaving a website with no interaction.
- *Average Session Duration* - The average length of time that visitors are on the website.

Since our data does not follow a normalized distribution, we transformed our data to a normal distribution via the Box-Cox transformation (Box & Cox, 1964). We used the log-transformation function, log (variable), and then executed a paired *t*-test on the four groups to statistically compare the differences between total visits, unique visitors, bounce rates, and average session duration on the transformed values. The reported values for total visits, unique visitors, bounce rates, and average session duration for Google Analytics and SimilarWeb differ significantly, as shown in Table 2 immediately below.

Table 2
Summary of results comparing Google Analytics and SimilarWeb for total visits, unique visitors, bounce rate, and average session duration. Difference uses Google Analytics as the Baseline. Results based on Paired T-Test for Hypotheses Supported.

Metric	Google Analytics	SimilarWeb	T-test and Result
Total Visits	(M = 6.82, SD = 0.31)	(M = 6.66, SD = 0.29)	t (85) = 6.43, p < 0.01.
Unique Visitors	(M = 6.56, SD = 0.26)	(M = 6.31, SD = 0.25)	t (85) = 12.60, p < 0.01
Bounce Rates	(M = 0.58, SD = 0.03)	(M = 0.63, SD = 0.02)	t (85) = −2,96, p < 0.01
Average Session Duration	(M = 2.15, SD = 0.05)	(M = 2.47, SD = 0.71)	t (85) = −8.59, p < 0.01

As shown in Table 2, the statistical testing results indicate a difference in the number of total visits, unique visitors, bounce rate, and average session duration between the two services.

However, underlying motivating questions remain, which are: *How close are the reported values to the 'true' values? Which approach is the most 'right'? Which approach is the most accurate? Which approach best reflects reality?* Regardless of the statistical testing results, these motivational questions are more challenging to address, as no gold standard exists for such an analysis. However, there is a 'true' number of visits, visitors, bounces, and average session duration. Compared to this reality, both analytics services are likely to be wrong in varying degrees of error – these numbers are measures, not counts! – and for different reasons intrinsic to how each service determines what is traffic and what are events.

We conducted a deductive analysis using a likelihood of error (Williamson et al., 2002) for each of our four metrics for our two services. For bounce rate, both Google Analytics and SimilarWeb are conceptually incorrect due to the practical issues of measuring a bounce visit. Concerning average session duration, again, for this metric, both Google Analytics and SimilarWeb are conceptually incorrect due to the practical issues of measuring the end of a session. For total visits, both are likely to be at least imprecise, as there is room for noise in the visits. Finally, regarding unique visitors, analytics services typically rely on a combination of cookies and tags to measure unique visitors, which generally results in overestimating unique visitors due to issues such as people clearing cookies, switching devices, or incognito browsing. So, again, the number from both services is likely to be wrong.

For all four major web traffic metrics, both approaches are inaccurate (i.e., wrong) due to conceptual reasons and, most likely, imprecise due to technical reasons. What is the problem? In our perspective, along with the challenges of collecting online data, it is a misconception to view web analytics data as 'counting'. In most cases, web analytics is not counting; instead, it is 'measuring'. Measuring has several sources of errors (e.g., sampling methods, measuring implementation – i.e., problems with who is doing the measuring, measurement instruments, and metric analysis). It is well known that there will be an error rate (±x) for nearly any measure (Bovbjerg, 2020). No measure or measurement tool is perfect, including online analytics services, and people numbers can be particularly messy. However, and this is a critical problem, *each of these platforms presents the numbers for these metrics as if they are precise counts*. Therefore, those viewing the number often proceed as if these numbers were precise counts, which they are not.

We and other researchers are often forced to use the numbers sourced from these black-box systems and application programming interfaces (APIs) when using data from the major platforms (e.g., Jansen & Schuster, 2011; Jiang et al., 2018; P. Wang et al., 2003; Y. Wang et al., 2021). The API and system output numbers appear precise, giving the impression of accuracy, and their accuracy is often bolstered by claims that the platform uses sophisticated AI or ML models to come up with the numbers. Yet, the accuracy of the numbers is rarely presented or investigated, as done by Jansen et al. (2022). The little evidence concerning this area should make you worried, to say the least. For example, at the time of this study, Facebook can predict a person's interests with seventy percent accuracy (Sabir et al., 2022), equating to a thirty percent error rate (which is an enormous error rate, given that three out of ten predictions is incorrect!). Google Ads, a major online advertising service at the time of writing, uses ML for targeting and budget optimization, but it can make a number of (hidden) errors, especially for low-resource languages (Hasan et al., 2020) and small business scenarios where the data is limited. Moreover, platforms frequently change their data processing rules with little or no communication with those accessing the data (Ruths & Pfeffer, 2014), even though the platform changes may affect the data-generating processes (Lazer et al., 2014) or the interpretation of that data (Jung, Salminen, & Jansen, 2022). There are many reasons to be worried about these errors, one of which is that decision-making is increasingly automated, as downstream AI and ML applications rely on the people numbers provided by the platforms without questioning them. Furthermore, even when humans make decisions, they increasingly depend on people numbers, with few questioning their validity.

To caveat and to be fair, we must bear in mind that online platforms were not designed for people research; they were designed to be computationally efficient and scalable, and to perform a commercial function. Practicality and research value can be at odds. For example, platforms may discard meta-data that would be valuable for research, or not save it to begin with (Paxson, 2004; Ruths & Pfeffer, 2014). As stated by Lazer et al. (2014, p. 1203), “most big data […] are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis”. Instead, the data is affected by algorithmic dynamics, i.e., changes made by engineers to tweak the system's commercial performance and by users using the service (Anderson et al., 2020). Changes in algorithms and user behavior are confounding variables that affect data quality validity (Jung, Salminen, & Jansen, 2022), and researchers are often unable to quantify this effect with the platform data alone.

We present this case study from the analytics area, but the issue of presenting numbers as counts when they are really measures occurs across most fields that rely on people numbers, especially numbers derived from large volumes of people data that are increasingly common. This sizeable people data is too often taken for granted by those who analyze the data, those who review the data, and those who write about the data. Yet, data scientists working with a perfect algorithm using flawed data – regardless of the data volume – will still end up with flawed results. Authors publishing results based on Twitter, Facebook, Google, or other large service datasets can still risk false predictions or inferences (Lazer et al., 2014), despite the size of their datasets. If the data is, for example, demographically biased to young adults, it does not matter if the absolute quantity is 100 observations or 1,000,000 observations – in both cases, the dataset represents young adults more than other groups, which is a selection bias. Nevertheless, judgments are often clouded by the sheer volumes of data, which is falsely believed to reflect the quality of the data. So, although ‘quantity does not equal quality’ sounds self-evident, it is often overlooked in the bewilderment of big numbers.

As an example, reviewers of research submissions might be less critical of data quality when the data is extracted in large numbers (e.g., ‘millions of clicks’) from a big company like Google, Facebook, or others, despite the possibility of the data having hidden errors. Even if the reviewer would question the data, the authors often have no way of ascertaining the data quality, and a standard response is along the lines of, “If Google is correct, so are we; Google is one of the biggest companies in the world, so they should be correct, right?“. As such, the data validity issue is often left at that – *an ‘appeal to authority’*. Note that while the people working in Google, for example, might know more details about data quality, most researchers publishing data sourced from the big technology companies operating the data ecosystem do not work in these companies, or they work in positions where they do not have intimate access to the precise data collection methodologies. Therefore, even if they acknowledge the numbers as measures, the researchers relying on this people data cannot directly assess its validity. So, this people data is, to a large extent, secondary data, even though in research, it is typically referred to and treated as primary data. This distinction matters quite a bit, as secondary data is typically considered less trustworthy than primary data in science (Venkatraman & Ramanujam, 1987).

5. Discussion and implications

“All empirical research stands on a foundation of measurement.” (Lazer et al., 2014, p. 1204). Nevertheless, even when we know that we are measuring, we researchers often present the results as if we are counting. This gives the allusion of an arithmetic process – *valid, precise, and accurate*, whereas, in reality, most people numbers are often not. These numbers are seemingly precise but not accurate in reality. For example, “how many visitors did I have on my website last month?” can be answered by web analytics software, but the answer is not precisely correct. How much off is it from reality? This information is often not provided. So, the issue is *not* that in the absence of error margins, logical deductions cannot be made; the issue is that researchers and stakeholders will make logical deductions, regardless. So, there is a need to fully understand the context of perceiving you are not counting but instead measuring. To shed light on this, we discuss three points: specifically, (a) the foundations for why you think you are counting, (b) why you don't count, and (c) what to do when you are measuring.

5.1. Why do we think we are counting when we are measuring?

Where does this illusion of data validity originate from? Why do we often view ourselves as ‘counting’ when the numbers are really from ‘measuring’ when working with people data? There are most likely many factors involved, but three reasons seem apparent – statistics, simplicity, and enumeration.

Statistics: One possible source is the sciences’ reliance on quantitative methods and algorithms, which are, at their core, statistical paradigms (Friedrich et al., 2021). Many people in the sciences start their academic careers focused on numbers (i.e., analytics) and quantitative approaches (i.e., statistical and/or algorithmic methods), and it is not difficult to see why. Working with numbers is both difficult and easy; difficult in that some of the statistical and algorithmic methods are challenging to master, but easy in that the results from these methods are relatively clear and precise (i.e., one always gets an output, as long as the input is in the correct form). So, if you have the numbers, these scientific methods are rigorous, and the results are clear and precise.

So, this counting versus measuring misconception may be grounded in statistics, on which many forms of analytics, algorithms, and ML models are based or based on the techniques that these methods employ. Statistics is “the science of counting” (Bowley, 1901, p. 7), and many statistical approaches rely on averages (Bowley, 1901, p. 7). Many (but not all, see (Reid, 2003)) foundational statistical techniques were developed for volume datasets where you are counting. However, much of the online people data is not small, and the volumes can be quite large. Even after years, decades actually, of working with uncertainty (Lindley, 2000) and algorithmic methods for large datasets, the ‘concept of the count’ still pervades the field of statistics and certainly the application of statistical techniques in other domains. It is not easy acknowledging this shift from counting to measuring. There is comfort in ignoring the data validity issues by assuming that our people numbers are accurate counts when it is clear they often are not, as shown in our web analytics example. When defined as precise counts, truth (as in an arithmetic number) is usually a fallacy in collecting and analyzing people numbers at any large scale (Galeano & Peña, 2019).

Simplicity: Another reason for the counting versus measuring misconception may be due to the need or the desire for simplicity. As seen with the web analytics tools in our case study example, people numbers presented as if they were accurate counts are clear, instead of addressing the messiness when indicating that the numbers are measurements (e.g., showing confidence intervals might appear 'complex' and 'messy' to stakeholders). Therefore, these measures are often passed off as precise arithmetic counts, which is a 'fallacy of precision' that is counterproductive for people number research. This presentation desire is mainly due to an aspiration for user friendliness and simplicity, where it is more convenient to show a precise number, even when that number does not truly reflect precise information. Moreover, many of the commonly used statistical analysis and visualization techniques employed do not like error rates, as the presentation of even simple metrics makes the discussion of such numbers ominously more complicated.

Enumeration: In grade school, we often learn to count with exercises such as images of discrete entities, such as apples. If there are three applies in the image, we learn to enumerate each one to get the total number – three. Like our apple, a person is a discrete entity. So it seems conceptually appropriate when we are working with people that the numbers are counts, and our work with these numbers is an enumeration (i.e., counting). However, in reality, when dealing with people data at scale, it is rarely enumeration. Instead, the scale of people is so large that the practicality prevents enumeration. Also, instead of working directly with people, we often use proxies, such as accounts, visits, posts, comments, unique visitors, clicks, and so on, which again seems like enumeration, but it is not. So, although our entities (i.e., people) are discrete, the numbers that we actually work with often are not.

5.2. Why Don't we count instead of measure?

Okay, then the obvious question becomes, *why don't we count rather than measure?* It turns out that counting is really hard to do with people data for at least two reasons – complexity and scale.

Complexity: Analysis of people data at first appears to be a simple operation that you can perform or which can be done automatically. However, even with relatively small volume datasets, people can make mistakes or can disagree on how to categorize them. These sources of error have by no means disappeared with the introduction of the technology that is often used on much larger people datasets. For example, two ML models, trained on the same people data, can disagree on which category a sample should belong to or the representation of that data (Jansen et al., 2021). As shown in our case study, two web analytics installations configured independently can report different numbers, causing decision makers to perhaps draw opposing conclusions. Different sentiment analysis tools rating the same text samples can give highly different results (Jung, Salminen, & JansenB, 2022). As another example, human annotation is often a highly subjective task. However, it is simpler to treat numbers inferred from the text as reflecting the meanings of that text, even though the summary scores are fallible. So, the determination of quantity, classifications, sentiment, toxicity, etc. is not straightforward, and although data about what people mean is often wrong, it is too complex to account for the errors when presenting the results.

Essentially, there are two general types of complexities that often prohibit you from counting: (a) technical and (b) paradigmatic. Technical problems deal with the challenging issues of data collection and imprecision. From our web analytics case study, for example, many technical issues limit your ability to actually 'count' what you want to tally. Research on the collection of online data (Aldous et al., 2022; Almerekhi et al., 2020, pp. 3033–3040; S. Jung et al., 2021; Salminen et al., 2022) has encountered many challenges. These include insufficient data collection support via APIs in collecting social media data that makes amassing people data hard or collected people data out of date, a difficulty to obtain sequentially critical data, which causes corrupted people data, and unpredictable changes in data collection support versions that result in inconsistent numbers. Our analysis of search logs, for example, shows that about ten percent of the query data is non-useable due to missing or corrupted data records (Jansen, 2006). The case gets even more complicated with paradigmatic choices, such as what label to give a social media post, the degree of sentiment expressed, etc. For such questions, there can be no definitive answers. While this disagreement should be a signal (Silver, 2015) that the phenomenon should be investigated, such disagreement is messy and complex, so it is often easier to consider it noise (see Aroyo & Welty, 2015 for a discussion of the complexities of Ground Truth vs. Crowd Truth). In summary, due to the complexities of people numbers, hiding the complexities of measures by presenting them as counts is much more comforting.

Scale: Even with small datasets, there is the possibility of errors, and the processing of people data at scale can get very difficult. At a large scale, counting becomes nearly insurmountable, and one cannot obtain absolute accuracy when people numbers surpass certain limits. At scale, arithmetic exactness is nearly unattainable: "[G]reat numbers are not counted correctly to a unit, they are estimated." (Bowley, 1901, p. 3). So, at scale, people numbers become the science of estimation (or the science of uncertainty: Lindley, 2000) – in other words, they are *measuring*.

As an illustrative example, let us look at a simple comparison task at different scales: *counting chickens*. Chicken data should be more straightforward to work with than people data, but the underlying scale issue is the same. There are insurmountable data validity issues when dealing with chickens at scale. The counting of chickens on a small family farm is an achievable task. The scale is small. What about counting all the chickens in the world? Due to the scale and temporal constraints, counting all the chickens worldwide is a practical impossibility. So, we are forced to measure. How would we measure the number of chickens in the world? Well, we could use a sampling technique. Take a small number of representative rural and urban areas and count the number of chickens in each of these areas. One could arrive at some chicken-per-square-meter numbers, which would then introduce error rates. With this number and the number of square meters, we could measure the number of chickens worldwide. This method has flaws, including that there are wild and feral chickens that the above method does not capture. However, when reporting the number of chickens, these methodological flaws are usually ignored, as the United Nations confidently states that there are 33 billion chickens in the world as of 2020 (United Nations, 2020); one has to dig quite deep to find the 'estimated' qualifier. The challenges, errors, and flaws with our chicken data are similar to the challenges, errors, and flaws with people data. Research with people numbers at scale is nearly always imprecise and inexact measuring instead of precise and exact counting, regardless of how confidently and how precisely the numbers are presented.

5.3. What to do when you measure instead of count?

We now address what is to be done when you find yourself measuring, but it appears as if you are counting, and you want to alert the reader that the numbers are measures. You have already taken the first step in addressing the conundrum in that *you are cognizant of the discrepancy*. Now, what else is there to do? There is unlikely to be a single solution that will completely address the issue in all circumstances; however, there are five general approaches. The common theme is transparency – *identifying and explicitly acknowledging that the reported numbers are not precise arithmetic counts but measures with some error rate*.

Exploratory Data Analysis: Exploratory data analysis (EDA) is a philosophical approach of performing initial investigations on datasets (Tukey, 1977), enabling you to generate dataset characteristics where you use EDA to understand, summarize, and prepare the data for other data analysis. EDA entails examining the data for trends and outliers using visual and quantitative methods. Employing EDA, you can see the underlying structure of the data (e.g., see whether it is a power law distribution), identify the influential variables (e.g., see what variables are correlated), and highlight anomalies (e.g., determine outliers). An EDA of people data will significantly aid with performing triangulation, determining confidence intervals, defining margins of error, and calculating the impact of data uncertainty. By using EDA, for example, you can see how your data's aggregate statistics compare to the visual graph of the data, see what the distribution of the data is, how much data you have, which variables are correlated, and any key variables that you may be missing.

Triangulation of Data: In our context, triangulation uses more than one source for collecting people data. This is data triangulation (Noble & Heale, 2019). You may also want the conceptualization of collecting that data to be different, which is theory triangulation (Denzin, 2009). For example, we employed triangulation in our analytics case study by collecting data from two data sources (i.e., Google Analytics and SimilarWeb), with each data source approaching the data collection from different conceptual perspectives (i.e., site-centric for Google Analytic and panel reliant for SimilarWeb), which allowed for a comparison of the values between the two approaches. You can also use triangulation to compare technical aspects of the same conceptual data collection approach; this is method triangulation (Denzin, 2009). For example, by using our case study domain of web analytics, you could simultaneously examine two site centric analytics installations (e.g., Google Analytics and Matomo, which is an open-source analytics service at the time of writing) on a set of websites and compare their numbers. You may obtain different values using different methods or even multiple measurements using the same method. Therefore, triangulation does not remove the errors of validity or eliminate the fundamental issue of precision, but it does help to establish a range of measurement discrepancies.

Reporting of Error Rates: All measures have a degree of uncertainty that counting does not have, and the variability in measurement results is an inherent component of the measuring process. When measuring, you assume that some true value exists for what is being measured and attempt to find this ideal quantity to the best of your ability. This variability introduces errors, which represent the difference between the value you measured and the true value. For our purposes, we are concerned with systematic errors, which are errors inherent in the data measurement context. It is possible to determine the dispersion of these errors to some degree or range of accuracy, and this range is the error rate (Dror, 2020). A low error rate indicates that the measurements are more accurate, while a high error rate indicates that the values are less accurate. Calculating the error rate will give an idea of the confidence interval, which is a statement of the validity of a measure. In statistics, a confidence interval refers to the probability that the true value for a population parameter will fall between two set values. This statistical probability is often challenging to determine when focused on the data validity of people numbers. A practical alternate is heuristics, using the magnitude of the error rates. Uncertainty visualization of the error rate may improve users' estimates, particularly for information-rich interfaces (Hullman et al., 2019). Based on comparable statistical measures such as effect sizes (Cohen, 1992), reasonable rules of thumb are shown in Table 3. We acknowledge that these are heuristics, and the impact should be independently assessed for a given domain and dataset (Schäfer & Schwarz, 2019). However, as marks on the wall, we offer the following definitions in Table 3 immediately below.

Table 3
Implications of error rate range when reporting measures of people numbers.

Magnitude	Error Rate	Definition
Small	.2	Error rates of (±) .2 (i.e., 20%) or less are low, indicating that the measure is acceptably accurate and implying that the measure can generally be treated akin to a count with an acknowledged error rate that is small.
Medium	.5	Error rates of greater than .2 (i.e., 20%) and less than .7 (i.e., 70%) (±) are medium, implying that the measure has systemic or other measurement errors that impose qualifications on it being treated as a count. The error rate should be explicitly stated and acknowledged as a medium rate.
Large	.7	Error rates of (±) .7 (i.e., 70%) or more are large, implying that the measure has systemic measurement errors that generally prohibit it from being treated as a count. The error rate should be explicitly stated, and the numerical results should be qualified as possessing validity issues.

Estimation of Impact: A final determination is an estimation of the impact of using a measure instead of a count. In other words, you know the numbers are not counts, but does it matter? A minor±1% error in the reported measures is unlikely to cause any shift in the business decision's direction. Even with large error rates, the impact might be minimal. For example, in our case study, key metrics between the two analytics platforms were strongly correlated (Jansen et al., 2022). So, if the task was to rank the websites, the use of measures instead of counts would not have much substantial impact, as the differences or inaccuracies of the web analytics tools would not affect decision making negatively, as it appears the errors are systematic (i.e., the ranking using the two approaches would be about the same). There are other such examples, such as comparing researchers using Google Scholar,[4] a bibliographic service as we write this article. Google Scholar is known to have errors (Halevi et al., 2017). However, if comparing researchers and assuming the errors are systematic, the errors have minimal impact on scholar ranking. In other situations, the impact could be very different, and the error rates could matter a great deal. For example, business decision makers could make mistakes with even a small margin of error, which may have serious consequences. Therefore, the questions are: 'how can we know?" and 'how can we separate the cases where numbers are 'close enough'? This must be on a case-by-case basis, and experimentation would seem like a potential solution through which to identify such cases.

Acknowledging the Measures: Since many data analysis techniques rely on an assumption of numbers as counts rather than measures, it is an unreasonable expectation that these measures will not be used as counts. Therefore, what is reasonable? The use of the measures rather than counts should be acknowledged in the research article or report, such as through coverage in the limitations section. The analytics platforms should also be upfront about error rates. However, you should discuss the error rates and impacts of the errors as a limitation of the research. So, how do you report this limitation? The most straightforward way is to show the value within a range of possible values:

$$\text{measure}(\text{n}) = (\text{reported count value}\pm\text{uncertainty (x)}) = \text{n}\pm\text{x}$$

The reported count value is the 'count' value you used in your research, and uncertainty is the error rate you determined, knowing that it is a 'measure'. The measure is the distribution of the possible values (i.e., reported count value±uncertainty). In our use case of web analytics, for example, we can report that the number of unique visitors to the set of sites is some number (i.e., n), plus or minus some uncertainty value (± x). As an example, from our case study using a likelihood of error (Williamson et al., 2002), as explained in (Jansen et al., 2022):

> The total number of unique visitors for all 86 websites was 834.7 million (max = 138.1 million; min = 1,799; med = 4.3 million) reported by Google Analytics …. Based on the issues just outlined, it seems that, for Google Analytics, a 20% overestimate in monthly unique visitors to 30% overestimate for more extended periods seems reasonable. (Jansen et al., 2022). So, the measure of unique visitors is 834.7 million ± 166.9 million, where n = 834.7 million and x = ±166.9 million.

Explicit acknowledgment does not remove validity errors, but it does alert the reader to evaluate the research results with this limitation in mind – i.e., *it remains the reader that the numbers are not counts but measures*!

5.4. *Summary of implications*

For a concise presentation of the discussion given above, Table 4

[4] https://scholar.google.com/.

summarizes the reasons for thinking you are counting, reasons for why you want to count, and what actions to take when measuring instead.

6. Closing remarks

Data validity has been shown over dozens of academic studies (Onwuegbuzie & Johnson, 2006) to be crucial for the trustworthiness of findings about people. However, ensuring data validity in real-world situations remains far from a solved problem. In contrast, when working with people numbers, the numbers (as exact counts) are wrong and need to be taken as measures (or estimates), rather than immutable arithmetic truth. The focus is often on developing more complex models. Interestingly, in these cases where the data is inaccurate, simpler models would probably work better, as they generally have low variance (Brownlee, 2016). When reporting people numbers, unless efforts are taken to alert the reader that these numbers are not counts but measures, it is easy for the reader to get lulled into treating these people numbers as absolute, precise, accurate arithmetic facts. Measurements are neither as precise nor as accurate as counts. So, if you are presenting people numbers, do not present numbers from measuring as if they are from counting, and report the uncertainty of the numbers as a limitation. If you are consuming numbers about people, you should always take these people numbers with a 'grain of salt'[5] (West & Bergstrom, 2020).

There are several open questions in this domain of using people numbers. On the technical side, an interesting notion is developing online analytics systems that can self-monitor and recognize flaws, which with analytics systems at the time of writing is nearly non-existent. However, it seems reasonable that platforms such as Google Analytics, for example, could have capabilities for the self-diagnosis of flaws in implementation, data collection, data analysis, and metric reporting. While fundamentally, you might not be able to build systems that can completely supervise themselves without a knowledge of external business goals, it is a worthwhile objective to pursue the boundary – namely, the point where an analytics system can self-diagnose its successful implementation and basic data operations.

On the people side, constructing gold standard people data sets derived by human annotation in posterior labeling requires critical reevaluation. To view the resulting numbers based on labels from, usually, crowdsourced labelers as objective truth is flawed. The labels are relative and subjective and are more likely measures of consensus or disagreement after the fact. Furthermore, the labels are usually assigned well after and with limited knowledge of the context in which the textual expressions were made. Therefore, there needs to be further scrutiny into how ground truth datasets for AI and ML applications are constructed in these fields and if the concept of ground truth even applies to some forms of people data.

Additionally, an area that deserves substantial evaluation is the theory and theoretical constructs (Jansen & Rieh, 2010) that are derived from or supported by people numbers (Anderson, 2000). If the underlying people numbers are often wrong, then it makes conceptual sense that the theories derived from such people data are also wrong (Ioannidis, 2005). Furthermore, there is a risk that false results become accepted into the academic body of knowledge and become the bedrock for future work (Hopf et al., 2019; Moonesinghe et al., 2007). This area has not, to our knowledge, been well investigated (Graziano, 2016; Shtulman, 2017). The main question that is raised is 'how much error can there be in the people numbers before it invalidates the people theory?'. The answer to this question has major implications for domains such as advertising, marketing, tourism, hospitality, and other human-contrived domains that attempt to concoct domain-specific theories. We leave these open questions for future research.

We do not mean to be too harsh on those researchers and others collecting, analyzing, and reporting people numbers. The authors of this article are researchers in the domain of people data and often employ people numbers. Rather, instead of criticism, we opt for optimism, as "a faulty measurement made on logical principles is better than none, and may lead to others with progressive improvement." (Bowley, 1901, p. 5). Certainly, having some people numbers is better than having none, and we believe this is true, even if the reported numbers are measures with error rates or are vague (Tukey, 1962) and imprecise counts (Wang & Strong, 1996). However, at the same time, a reality check is also needed (Coombs, 1964). Working with people numbers can be messy, and those working with people numbers have an obligation to acknowledge this messiness and account for it. This reflection piece presents the rationale for such acknowledgment and offers the means to acknowledge it. So, if you are working with people numbers, we hope you find these reflections valuable in your research.

Table 4
Summary of discussion and implications of measuring versus counting.

Reasons Why You Think You're Counting	Reasons Why You Want to Count	Actions to Take When You Are Using Measures Instead of Counts
Statistics Simplicity Enumeration	Complexity Scale	Exploratory Data Analysis Triangulation Reporting of Error Rate Gauging the Impact Acknowledging the Measure

As a limitation, we must mention that there might be people numbers derived from people data that are counts or very close to actual counts. There are organizations, for example, that employ people data as part of their services, control nearly the entire data processing stream, and have specialists who clean and disambiguate data. One example might be the number of official marriages in a given county – let us take the U.S. – which would be a large scale people numbers problem that is still probably close to counts. Why? Well, in the U.S., marriage licenses are processed in low level municipalities using a fairly regulated method. The scale at the municipality level is small – sometimes just a few hundred or thousand people. The results of all these smaller jobs at the municipality level are reported to states and eventually to the federal government for aggregation. The census data in some countries is reportedly very close to counts (US Census Bureau, 2022), although sampling has sometimes been used (US Census Bureau, 2020), and there are reports of over/under counting (US Census Bureau, 2022) of population subgroups (Jacobsen & Mather, 2020). In the commercial sector, one could picture an international fast food company, say selling hamburgers, getting global hamburger sales data by aggregating the daily counts at each of its thousands of outlets (Stice, 2019), which is again, a highly regulated process that is really aggregated smaller jobs. However, outside of these tightly coupled processes that can be divided into many small counting procedures, fractal processes, and focused on a narrow primary data variable, people numbers are usually not counts but measures.

Let us end our journey concerning people numbers where we began, with the well-known quote by Kelvin (1883).

> When you can **measure** what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science …. Lord Kelvin [**emphasis added**]

Note the verb that Kelvin uses – *measure*! When you can *measure* something and express it in numbers, you know something about it. Kelvin was a researcher of the physical world, considering electricity, navigation, heat, and temperature. He obviously measured in his fields of research, and his numbers were probably nearly always measurements. Error rates were most likely an inherent part of his work. Numbers in the

[5] Grain of salt: An English language idiom that means to view something with scepticism.

scientific method are needed, and we believe that Kelvin is correct – *numbers are essential for scientific inquiry*. Then again, we would add that it is best if the numbers are presented and interpreted correctly.

Declaration of competing interest

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

Acknowledgments

We would like to thank Natasa Milic-Frayling, Sanjay Chawla, Marie Christine Rufener, and Ghanim Al-Sulaiti, all from the Qatar Computing Research Institute at the time, for their comments on an earlier version of this work. We thank the anonymous reviewers for their helpful comments and the journal editors, Professor Dan Wu and Professor Tinting Jiang, for their support of this manuscript. The opinions expressed in this manuscript are those of the authors.

References

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Aldous, K. K., An, J., & Jansen, B. J. (2022). Measuring 9 emotions of news posts from 8 news organizations across 4 social media platforms for 8 months. *ACM Transactions on Social Computing (TSC), 4*(4), 1–31.

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020). *Are these comments triggering? Predicting triggers of toxicity in online discussions*. Proceedings of The Web Conference, 2020.

Alonso, O., Marshall, C. C., & Najork, M. (2015). Debugging a crowdsourced task with low inter-rater agreement. In *Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries*. https://doi.org/10.1145/2756406.2757741, 101–110.

Ameringer, S., Serlin, R. C., & Ward, S. (2009). Simpson's paradox and experimental research. *Nursing Research, 58*(2), 123–127. https://doi.org/10.1097/NNR.0b013e318199b517

Anderson, C. (2000). *The end of theory: The data deluge makes the scientific method obsolete. Wired*. https://www.wired.com/2008/06/pb-theory/.

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020). Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of the web conference 2020* (pp. 2155–2165). ACM. https://doi.org/10.1145/3366423.3380281.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17–21.

Aroyo, L., & Welty, C. (2013). *Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard*. WebSci2013, 2013.

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine, 36*(1), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

Becker, T. E., Atinc, G., Breaugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2016). Statistical control in correlational studies: 10 essential recommendations for organizational researchers. *Journal of Organizational Behavior, 37*(2), 157–167. https://doi.org/10.1002/job.2053

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Oxford University Press.

Bertrand, M., & Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *The American Economic Review, 91*(2), 67–72.

Billboard. (2022). *Greatest of all time artists*. https://www.billboard.com/charts/greatest-of-all-time-artists/.

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica, 83*(1), 155–174.

Blank, G., & Lutz, C. (2017). Representativeness of social media in great britain: Investigating Facebook, linkedin, twitter, pinterest, Google+, and instagram. *American Behavioral Scientist, 61*(7), 741–756.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364–366.

Boslaugh, S., & Watters, D. P. A. (2008). *Statistics in a nutshell: A desktop quick reference*. O'Reilly Media, Inc.

Bovbjerg, M. L. (2020). Random error. In *Foundations of epidemiology*. Oregon State University. https://open.oregonstate.education/epidemiology/chapter/random-error/.

Bowley, A. L., & Arthur, L. S. (1901). *Elements of statistics. P. S. King*. http://archive.org/details/elementsstatist03bowlgoog.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B, 26*(2), 211–252.

Bradshaw, T. (2022). *Twitter admits overstating audience figures for 3 years*. Financial Times. https://www.ft.com/content/3184a769-4288-44e8-abea-300d27c1aa95.

Bradsher, K. (2020). *Forget the polls: This Chinese indicator is flashing 'trump.' the New York times*. https://www.nytimes.com/2020/10/28/business/trump-china-election.html.

Briquelet, K. (2012). *'Free'quent flier has wings clipped after American Airlines takes away his unlimited pass*. May 13 https://nypost.com/2012/05/13/freequent-flier-has-wings-clipped-after-american-airlines-takes-away-his-unlimited-pass/. New York Post.

Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences, 115*(11), 2563–2570. https://doi.org/10.1073/pnas.1708279115

Brownlee, J. (2016). *Gentle introduction to the bias-variance trade-off in machine learning. Machine learning mastery*. March 17 https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/.

Bughin, J. R. (2010). *The world according to YouTube: Explaining the rise of online participative video. Peer-to-Peer Networks and Internet Policies*, 179–191.

Cervi, L. (2021). Tik tok and generation Z. *Theatre, Dance and Performance Training, 12*(2), 198–204. https://doi.org/10.1080/19443927.2021.1915617

Chapman, C. N., Love, E., Milham, R. P., ElRif, P., & Alford, J. L. (2008). Quantitative evaluation of personas as information. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 52, pp. 1107–1111). https://doi.org/10.1177/154193120805201602, 16.

Chatterjee, S., & Firat, A. (2007). Generating data with identical statistics but dissimilar graphics: A follow up to the Anscombe dataset. *The American Statistician, 61*(3), 248–254.

Chen, E. (2022). *30% of google's emotions dataset is mislabeled*. https://www.surgehq.ai//blog/30-percent-of-googles-reddit-emotions-dataset-is-mislabeled.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction, 19*(3), 323–393. https://doi.org/10.1207/S1532690XCI1903_3

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Coombs, C. H. (1964). *A theory of data*. Wiley. xviii, 585.

Dacrema, M. F., Boglio, S., Cremonesi, P., & Jannach, D. (2021). A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems, 39*(2), 1–49.

Daniels, G. S. (1952). *The "Average" man. Air Force aerospace medical research lab wright-patterson AFB OH*. https://apps.dtic.mil/sti/citations/AD0010203.

Denzin, N. K. (2009). *The research act: A theoretical introduction to sociological methods*. Routledge.

Desolneux, A., Moisan, L., & Morel, J.-M. (2007). *From gestalt theory to image analysis: A probabilistic approach* (Vol. 34). Springer Science & Business Media.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248*(5), 116–131.

Dror, I. (2020). The error in 'error rate': Why error rates are so needed, yet so elusive. *Journal of Forensic Sciences, 65*. https://doi.org/10.2139/ssrn.3593309

Ellenberg, J. (2015). *How not to Be wrong: The power of mathematical thinking*. Penguin Books.

Epstein, D. (2019). *Range: Why generalists triumph in a specialized world*. Riverhead Books.

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management, 58*(3), Article 102524. https://doi.org/10.1016/j.ipm.2021.102524

Freeman, T. (2018). *Mark Cuban reveals the best thing He ever bought, which only 25 people in the world still own. Maxim*. April 16 https://www.maxim.com/news/mark-cuban-american-airlines-ticket-2018-4/.

Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., … Friede, T. (2021). Is there a role for statistics in artificial intelligence? *Advances in Data Analysis and Classification*, 1–24.

Galeano, P., & Peña, D. (2019). Data science, big data and statistics. *Test, 28*(2), 289–329. https://doi.org/10.1007/s11749-019-00651-9

Graziano, M. (2016). *Most theories of consciousness are worse than wrong. The atlantic*. https://www.theatlantic.com/science/archive/2016/03/phlegm-theories-of-consciousness/472812/.

Greenspan, A. (2019). *PlainSite: Reality check*. Facebook, Inc. https://www.plainsite.org/realitycheck/facebook.html.

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—review of the Literature. *Journal of Informetrics, 11*(3), 823–834. https://doi.org/10.1016/j.joi.2017.06.005

Hasan, M. A., Tajrin, J., Chowdhury, S. A., & Alam, F. (2020). Sentiment classification in bangla textual content: A comparative study. In *2020 23rd international conference on computer and information technology (ICCIT)* (pp. 1–6).

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161. https://doi.org/10.2307/1912352

Hernán, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology, 40*(3), 780–785.

Hopf, H., Krief, A., Mehta, G., & Matlin, S. A. (2019). Fake science and the knowledge crisis: Ignorance can be fatal. *Royal Society Open Science, 6*(5), Article 190161. https://doi.org/10.1098/rsos.190161

Hullman, J., Qiao, X., Correll, M., Kale, A., & Kay, M. (2019). In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

Infante-Rivard, C., & Cusson, A. (2018). Reflection on modern methods: Selection bias—a review of recent developments. *International Journal of Epidemiology, 47*(5), 1714–1722. https://doi.org/10.1093/ije/dyy138

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jacobsen, L. A., & Mather, M. (2020). *How will we Measure the Accuracy of the 2020 census? PRB*. https://www.prb.org/resources/how-will-we-measure-the-accuracy-of-the-2020-census/.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research, 28*(3), 407–432. https://doi.org/10.1016/j.lisr.2006.06.005

Jansen, B. J., Jung, S., & Salminen, J. (2021). The effect of hyperparameter selection on the personification of customer population data. *International Journal of Electrical and Computer Engineering Research, 1*(2). https://doi.org/10.53375/ijecer.2021.31. Article 2.

Jansen, B. J., Jung, S., & Salminen, J. (2022). Measuring user interactions with websites: A comparison of two industry standard analytics approaches using data of 86 websites. *PLoS One, 17*(5), Article e0268212. https://doi.org/10.1371/journal.pone.0268212

Jansen, B. J., Moore, K., & Carman, S. (2013). Evaluating the performance of demographic targeting using gender in sponsored search. *Information Processing & Management, 49*(1), 286–302.

Jansen, B. J., & Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology, 61*(8), 1517–1534. https://doi.org/10.1002/asi.21358

Jansen, B. J., & Schuster, S. (2011). Bidding on the buying funnel for sponsored search and keyword advertising. *Journal of Electronic Commerce Research, 12*(1), 1–18.

Jiang, T., Yang, J., Yu, C., & Sang, Y. (2018). A clickstream data analysis of the differences between visiting behaviors of desktop and mobile users. *Data and Information Management, 2*(3), 130–140. https://doi.org/10.2478/dim-2018-0012

Jones, G. E. (2006). *How to lie with charts* (2nd ed.).

Jung, S., Salminen, J., & Jansen, B. (2021). Persona analytics: Implementing mouse-tracking for an interactive persona system. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1–8).

Jung, S., Salminen, J., & Jansen, B. J. (2022). Engineers, aware! Commercial tools disagree on social media sentiment: Analyzing the sentiment bias of four major tools. *Proceedings of the ACM on Human-Computer Interaction, 6*(EICS), 1–20. https://doi.org/10.1145/3532203

Jung, S., Salminen, J., & Jansen B, J. (2022). The effect of hiding dislikes on the use of YouTube's like and dislike features. In *14th ACM web science conference* (pp. 202–207). https://doi.org/10.1145/3501247.3531546, 2022.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment.* Brown Spark: Little.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. https://doi.org/10.1037/h0034747

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In *2013 46th Hawaii international Conference on system sciences* (pp. 995–1004). https://doi.org/10.1109/HICSS.2013.645

Keeter, S., Hatley, N., Lau, A., & Kennedy, C. (2021). *What 2020's election poll errors tell us about the accuracy of issue polling. Pew research center methods.* https://www.pewresearch.org/methods/2021/03/02/what-2020s-election-poll-errors-tell-us-about-the-accuracy-of-issue-polling/.

Kelvin, W. T. (1883). *Electrical units of measurement. Being one of the series of lectures delivered at the.* Institution of Civil Engineers. *session 1882-83*.

Kievit, R., Frankenhuis, W., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology, 4*, 513.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology, 70*(2), 144–156.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science, 343*(14 March), 1203–1205.

Lemmon, E. J. (1962). On sentences verifiable by their use. *Analysis, 22*(4), 86–89. https://doi.org/10.1093/analys/22.4.86

Lerman, K. (2017). *Computational social scientist beware: Simpson's paradox in behavioral data*. CSS. https://doi.org/10.1007/s42001-017-0007-4, 2018.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society. Series D (The Statistician), 49*(3), 293–337.

Mangel, M., & Samaniego, F. J. (1984). Abraham wald's work on aircraft survivability. *Journal of the American Statistical Association, 79*(386), 259–267. https://doi.org/10.2307/2288257

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 1290–1294).

Moonesinghe, R., Khoury, M. J., Janssens, A., & Cecile, J. W. (2007). Most published research findings are false—but a little replication goes a long way. *PLoS Medicine, 4*(2), Article e28. https://doi.org/10.1371/journal.pmed.0040028

Morse, J. (2019). *50 percent of Facebook users could be fake, report claims*. Mashable. https://me.mashable.com/tech/1903/50-percent-of-facebook-users-could-be-fake-report-claims.

Noble, H., & Heale, R. (2019). Triangulation in research, with examples. *Evidence-Based Nursing, 22*(3), 67–68. https://doi.org/10.1136/ebnurs-2019-103145

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools, 13*(1), 48–63.

Oyer, P. (2014). *Everything I ever needed to know about economics I learned from online dating.* Harvard Business Review Press.

Paxson, V. (2004). Strategies for sound Internet measurement. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement* (pp. 263–271). https://doi.org/10.1145/1028788.1028824

Pearl, J. (2022). Comment: Understanding Simpson's paradox. In *Probabilistic and causal inference: The works of judea Pearl* (pp. 399–412).

Popper, K. (2002). *The logic of scientific discovery* (2nd ed.). Routledge.

Post, D. E., & Votta, L. G. (2005). Computational science demands a new paradigm. *Physics Today, 58*(1), 35–41. https://doi.org/10.1063/1.1881898

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics, 31*(6), 1695–1731.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064.

Sabir, A., Lafontaine, E., & Das, A. (2022). Analyzing the impact and accuracy of Facebook activity on facebook's ad-interest inference process. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW1), 1–34.

Salkind, N. (2010). Triangulation. In *Encyclopedia of research design*. SAGE Publications, Inc. https://doi.org/10.4135/9781412961288.n469.

Salminen, J., Jansen, B. J., An, J., Jung, S., Nielsen, L., & Kwak, H. (2018a). Fixation and confusion: Investigating eye-tracking participants' exposure to information in personas. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 110–119). https://doi.org/10.1145/3176349.3176391

Salminen, J., Jung, S., & Jansen, B. (2022). Developing persona analytics towards persona science. *27th International Conference on Intelligent User Interfaces*, 323–344.

Salminen, J., Kamel, A. M., Jung, S., & Jansen, B. J. (2021). The problem of majority voting in crowdsourcing with binary classes. In *Proceedings of 19th European conference on computer-supported cooperative work*. https://doi.org/10.18420/ecscw2021_n12. Zurich, Switzerland.

Salminen, J., Veronesi, F., Almerekhi, H., Jung, S., & Jansen, B. J. (2018b). Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 fifth international Conference on social networks analysis, Management and security (SNAMS)* (pp. 88–94). https://doi.org/10.1109/SNAMS.2018.8554954

Savage, S. L., & Markowitz, H. M. (2012). *The flaw of averages: Why we underestimate risk in the face of uncertainty.* Wiley.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00813.

Sharma, K., Ferrara, E., & Liu, Y. (2022). Characterizing online engagement with disinformation and conspiracies in the 2020 U.S. Presidential election. In *Proceedings of the international AAAI conference on web and social media* (Vol. 16, pp. 908–919).

Shtulman, A. (2017). *Scienceblind: Why our intuitive theories about the world are so often wrong.* Hachette UK.

Siegel, D. A. (2010). *The mystique of numbers: Belief in quantitative approaches to segmentation and persona development. CHI Extended Abstracts*. https://doi.org/10.1145/1753846.1754221

Silver, N. (2015). *The signal and the noise: Why so many predictions fail–but some don't.* Penguin Books.

Silverman, D. (2020). *Qualitative research.* SAGE.

SimilarWeb. (2022a). *Connecting Google Analytics and similarweb. Similarweb knowledge center*. https://support.similarweb.com/hc/en-us/articles/208420125-Connecting-Google-Analytics-and-Similarweb.

SimilarWeb. (2022b). *Our data | similarweb. Similarweb*. https://www.similarweb.com/corp/ourdata/.

SimilarWeb. (2022c). *What does connecting my Google analytics account with SimilarWeb mean? – knowledge center—SimilarWeb*. https://support.similarweb.com/hc/en-us/articles/208420125-What-does-connecting-my-Google-Analytics-account-with-SimilarWeb-mean.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B, 13*(2), 238–241.

Snow, R., O`connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 254–263).

Stice, J. (2019). *The truth about how many burgers McDonald's has sold*. Mashed.Com. March 21 https://www.mashed.com/148375/the-truth-about-how-many-burgers-mcdonalds-has-sold/.

Surowiecki, J. (2005). *The wisdom of crowds. Anchor*.

Survivorship bias. (2022). *Survivorship bias. Wikipedia*. https://en.wikipedia.org/w/index.php?title=Survivorship_bias&amp;oldid=1110053063.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable* (Vol. 2). Random house.

Thirumuruganathan, S., Jung, S., Ramirez Robillos, D., Salminen, J., & Jansen, B. J. (2021). Forecasting the nearly unforecastable: Why aren't airline bookings adhering to the prediction algorithm? *Electronic Commerce Research, 21*(1), 73–100.

Timberg, C. (2006). *How AIDS in africa was overstated*. April 6 http://www.washingtonpost.com/wp-dyn/content/article/2006/04/05/AR2006040502517.html.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media. Eighth international AAAI conference on weblogs and social media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1–67. https://doi.org/10.1214/aoms/1177704711

Tukey, J. (1977). *Exploratory data analysis*.

United Nations. (2020). Chickens | gateway to poultry production and products | food and agriculture organization of the united Nations. *Food and Agriculture Organization of the United Nations*. https://www.fao.org/poultry-production-products/production/poultry-species/chickens/en/.

US Census Bureau. (2020). *Developing sampling techniques—history—U.S. Census Bureau*. https://www.census.gov/history/www/innovations/data_collection/developing_sampling_techniques.html.

US Census Bureau. (2022). *Census Bureau releases estimates of undercount and overcount in the 2020 census*. https://www.census.gov/newsroom/press-releases/2022/2020-census-estimates-of-undercount-and-overcount.html.

Vecchio, P. D., Mele, G., Ndou, V., & Secundo, G. (2018). Creating value from social big data: Implications for smart tourism destinations. *Information Processing & Management, 54*(5), 847–860. https://doi.org/10.1016/j.ipm.2017.10.006

Venkatraman, N., & Ramanujam, V. (1987). Measurement of business economic performance: An examination of method convergence. *Journal of Management, 13*(1), 109–122. https://doi.org/10.1177/014920638701300109

W3Techs. (2020). *Usage statistics and market share of Google analytics for websites*. October 2020 https://w3techs.com/technologies/details/ta-googleanalytics.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician, 36*(1), 46–48.

Wang, P., Berry, M. W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 743–758.

Wang, Y., Liu, X., Ju, Y., Börner, K., Lin, J., Sun, C., & Si, L. (2021). Chinese E-romance: Analyzing and visualizing 7.92 million alibaba valentine's day purchases. *Data and Information Management, 5*(4), 363–371. https://doi.org/10.2478/dim-2021-0006

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Waqas, A., Salminen, J., Jung, S., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS One, 14*(9), Article e0222194. https://doi.org/10.1371/journal.pone.0222194

West, J. D., & Bergstrom, C. T. (2020). *Calling bullshit: The art of scepticism in a data-driven world*. Penguin UK.

Wiebe, J., Bruce, R., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the association for computational linguistics* (pp. 246–253).

Williamson, K., Burstein, F., & McKemmish, S. (2002). Chapter 2—the two major traditions of research. In K. Williamson, A. Bow, F. Burstein, P. Darke, R. Harvey, G. Johanson, S. McKemmish, M. Oosthuizen, S. Saule, D. Schauder, G. Shanks, & K. Tanner (Eds.), *Research methods for students, academics and professionals* (2nd ed., pp. 25–47). Chandos Publishing. https://doi.org/10.1016/B978-1-876938-42-0.50009-5.

Wu, M. (2012). *The big data fallacy and why we need to collect even bigger data | TechCrunch*. https://techcrunch.com/2012/11/25/the-big-data-fallacy-data-%e2%89%a0-information-%e2%89%a0-insights/.

Wu, D., He, D., Qiu, J., Lin, R., & Liu, Y. (2013). Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese. *Journal of Information Science, 39*(2), 169–187.

Zgraggen, E., Zhao, Z., Zeleznik, R., & Kraska, T. (2018). Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–12). https://doi.org/10.1145/3173574.3174053