Dalhousie Journal of Interdisciplinary Management
Volume 17 | 2023
ojs.library.dal.ca/djim

Data and its Misuse: The Efficacy of Objectivity

PRIYAM KALAVADIA

SCHOOL FOR RESOURCE AND ENVIRONMENTAL STUDIES

ABSTRACT

Mark Twain, the renowned American writer and humourist, is often quoted to have said “facts are stubborn things but statistics are pliable” (Kihuro, 2014). Although anecdotal, the point Mr. Twain makes resonates true in our modern information age society. The use of descriptive statistics is widespread in sports, humanities, academia, and probably of most consequence, news and media outlets. The mathematical properties of statistical analysis are inherently objective, however, its use (or misuse) can be hijacked by bad actors to compliment and give pseudo-rationality to propaganda and tailored societal messages. This misuse may not be deliberate, but to the layman, the message is what is perceived not the mechanism of how it has been portrayed. Data and information in this so-called information age is often branded with terminology that implies objectivity, though, can anything be objective when subject to human interpretation? The purpose of this paper is to question the inherent branding of objectivity in available data and information sources by evaluating mechanisms of representation.

Keywords: Data, Information, Objectivity, Classification

INTRODUCTION

We as individuals in the 21st century are inundated with information and stimuli from the ongoings of the world around us in an unprecedented magnitude due to the widespread acceptance of the world wide web and smart devices; data is the backbone of this so-called information age. Data is a difficult thing to define, and for the sake of this paper, we will be using the definition given by the famous information studies professor Marcia Bates. Bates (2006) defines data as all energy and material objects aside from the phenomena of entropy as potential information, and that all potential information on its own is objective until observed.

Data represents all stimuli overserved or not, while information is the conclusion reached through our interactions with data. Given this inherent objectivity of potential stimuli, how can it be used to formulate inappropriate conceptions? Does objectivity prevent it from being used as misinformation? Is information objective at all? The purpose and objective of this paper will be to contextually analyze data in society, how it is misused via symbolism and descriptive statistics, and how other mechanisms such as colour and visual aids result in the loss of objectivity in our grandiose and novel information age. The final question this paper aims to resolve is could data be perceived in an objective way, or is the term objectivity a fallacy.

Data And Its Role in Our Modern Society

What is data?

Under the assumption of Bates's (2006) definition, data can potentially be anything in the world. Although the word data is often conflated with computers, numbers, and scientific research, every single sentient life form in our universe is in some way a data management vessel. Consider even the most mundane activity of an ordinary day, going for a walk. The average person is bombarded with overwhelming amounts of stimuli that if one were to dissect and record every single piece of information that can be potentially perceived on a walk down the street, the final copy would be a piece of writing that rivals a Tolkien novel. Visual stimuli such as signage, cars, trees, grass, infrastructure, and other pedestrians all are taken in and used to formulate our path and decisions. Energy-based stimuli such as sound, smell, and touch can be data as well. Should I walk across the road? How fast? Should I stop to take in the sights? For how long? Data as

potential stimuli surrounds us in everything we do. The interpretation of this data brings us to the next point as data management vessels, information and knowledge, how and why do we make the decisions we do?

The Loss of Objectivity: the DIKW Pyramid

Rowley (2007) re-evaluates the efficacy of the DIKW pyramid (Data, Information, Knowledge, Wisdom). The second stage in this pyramid representation is how we as data management entities interpret the first stage, potential stimuli (data), as information. Potential stimuli that exist as energy and matter out in the world only become information when one contemplates it. For information to truly manifest in conscious thought, one has to discern relevant data available to formulate a coherent message (Rowley, 2007). Take the example of a yield sign at the point of two roads merging: the symbol of a yield sign is only significant if the entity observing it understands its meaning. A young child sitting in the passenger's seat who does not understand the meaning of the symbolism would ignore the sign as just stimuli. For data to become information, it must be observed and understood through symbolic meaning. This goes even further with the next step in the DIKW pyramid, knowledge (Rowley, 2007). Knowledge is the amalgamation of information from different facets used to make informed decisions in the context of capability, experience, skills and values (Rowley, 2007). It requires us to be able to interpret, categorize, store, and recollect data from the past to comprehend and discern what the best path of action is in the present.

The inherent objectivity of potential stimuli is lost at this stage. How we interpret data is a very personal mechanism. We compare new stimuli to our past experiences, basal instincts, and probably the most confounding factor, our feelings. Feelings may seem trifling and arbitrary, however, the power and influence that they have on our understanding of the world and decision-making processes should not be understated. Decision-making processes are not immune to the influence of emotion, although data-derived rational processes are still driving factors, emotions are not epiphenomenal as consequentialists seem to believe but are heavily integrated (Loewenstein et al., 2001). Consider the "risk-as-feelings" model derived from Loewenstein et al.'s (2001) article in the *Psychological Bulletin* in figure 1. Emotional factors are part of our understanding process, not secondary to it.

Figure 1.

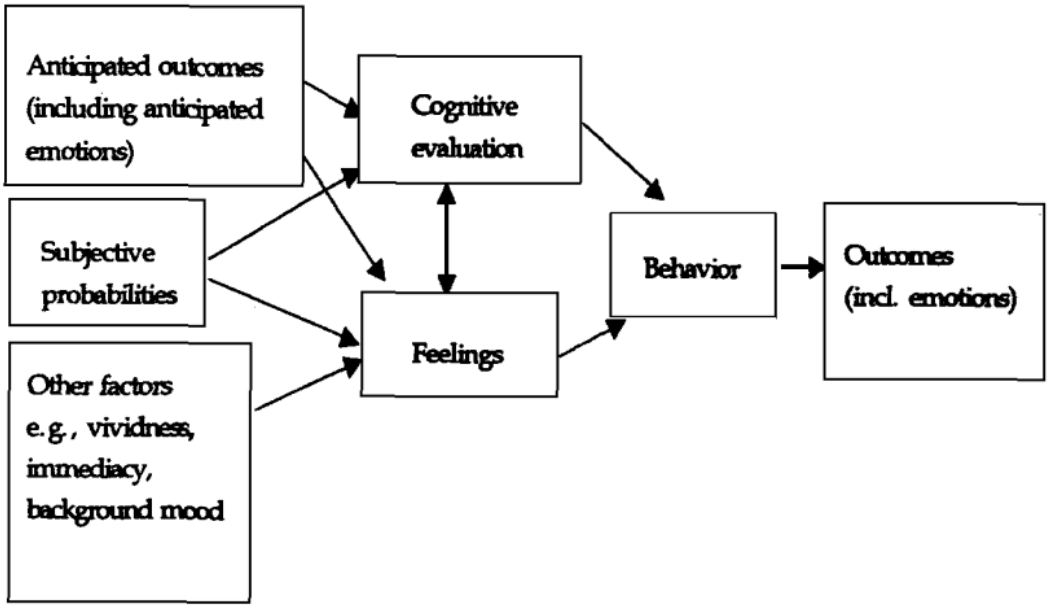

Fig.1 from the works of Loewenstein et al. (2001) the risk-as-feelings model highlights the influence that emotions have on decision-making based on available knowledge, information and data.

Data, Information and Knowledge in the Information Age

The advent of worldwide connectivity vectors such as the internet and smart devices has launched us into a brand-new world of data. We are not solely limited by stimuli from our immediate surroundings, rather we are launched into a globalist perspective as a mere observer of virtually all activity from almost all locations. Although the term information age is a bit of a misnomer as humans have been collecting, storing, and disseminating data since the beginning of time, the impacts of connectivity-based technology should not be undermined. In addition to this technological advancement of entering a global ecosystem, our emotions are also a vital factor as described by Loewenstein et al. (2001). It comes to little surprise that our emotions can be hijacked and utilized by media and news outlets to prey on our mechanisms of thought and decision-making to promulgate inappropriate messages from seemingly objective data.

Misuse of Symbolism and Statistics in The Information Age

Categorization and Symbolism

Human society's tendencies to categorize the world around us have been a great vector in our technological and scientific advancement, but those same tendencies may sometimes be misguided. Consider language, the basis of communication and one of the more significant

reasons why humankind has transcended beyond other species. Language is a standardization of the utilization of symbols to describe material and energy-based stimuli (data) (Dickins & Dickins, 2001). Words act as both auditory and visual (when written) symbolism and allow us to describe the world (Zhirenova et al., 2016). This process is never-ending, as new words, slang, and expressions are being continually added to the already vast vocabulary collection. Due to this never-ending progression, it sometimes takes time for the world to catch up with new words and meanings thus leading to some unintended miscategorising. Such as in the current social zeitgeist, expanded pronoun and gender categorization has been highlighted as an issue. However, in the name of inclusivity, the magnitude of individual categories has expanded tremendously. So much so that some who may not be as versed in the digital media world may risk offence through unintended obliviousness. Words, thus, as a representation of our understanding of stimuli, fall under the classification of information. Although to the observer of words, it can still be data until it is understood.

Regardless, our need for categorization has allowed us to reach the so-called information age that we are in. Language, words, and associated symbolism lead to further categorization techniques such as the earliest mathematical formulations of statistical analysis. Statistical analysis, in its own right, is objective as it simply describes how stimuli can fall under different categories without the input of human inference. However, different methods of descriptive statistical analysis can be used to end up with different categorization end-products. Choosing which classification method to use and how to display this descriptive statistical analysis is left to human devices. Miscategorising is just as big of a problem here as it is with choosing the right

words. This can be quite dangerous as it gives a façade of objectivity to a subjective practice. This façade, when interpreted by the common man, allows them to formulate what they think is a fact about something that may be severely misrepresented.

Misuse of Descriptive Statistics by Classification: The Florida Example

The most egregious mechanism of statistical misuse is how data is classified. Classification systems are used to display data in categorized groups. There are many different ways this categorization can be implemented and each way has its perks and setbacks. The data itself also plays a huge role as to what classification system should be used. If you are working with data that presents itself in a normal (or Gaussian) distribution it may be best to use standard deviation classification systems or, conversely, if the data is skewed to one pole or the other, transformations or other classification methods may be the preferred method (ESRI, n.d.). The mismatch of classification systems to data results in completely different outputs that can easily tailored to a certain and even seemingly opposing societal messages. How do you know which one to use? Well, that depends entirely on what message you want to project. For a more tangible example, refer to figure 2. below, which helps portray the impact of different classification types on the same dataset. Figure 2. is from the textbook *Thematic Cartography and Geovisualization* (Slocum et al., 2009).

Figure 2.

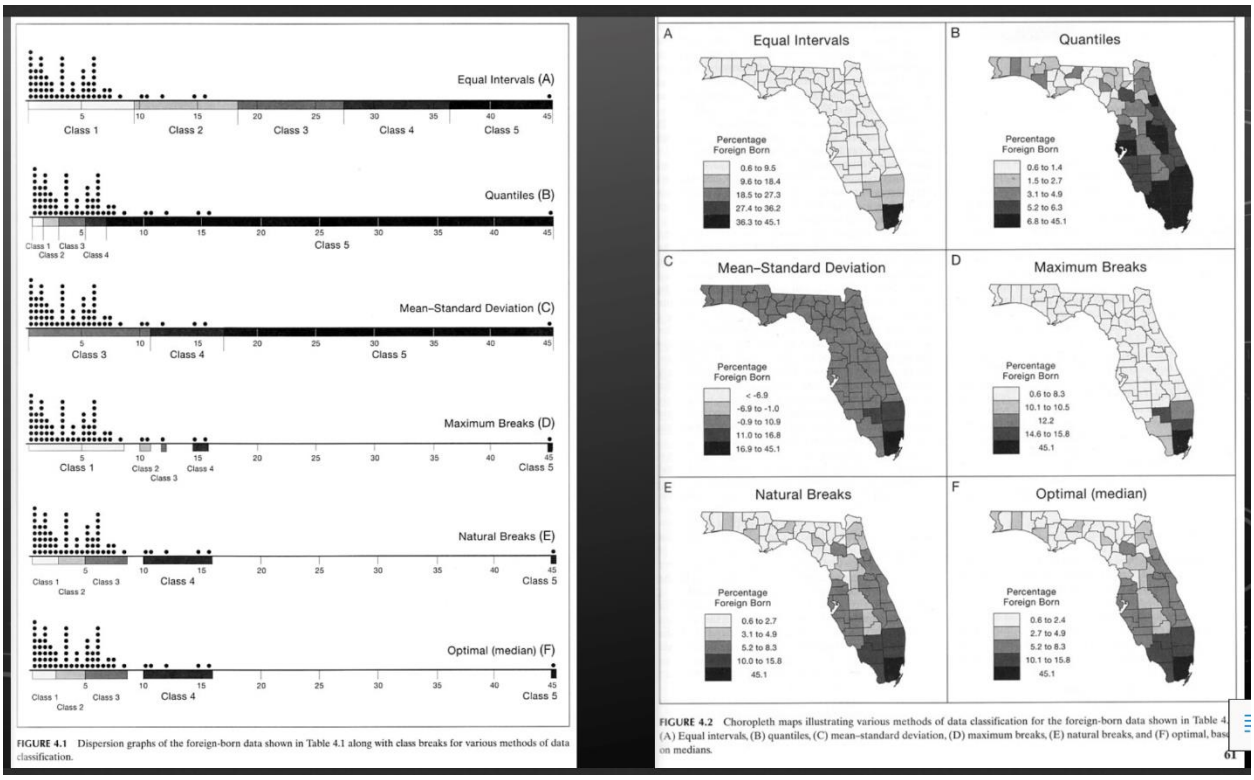

Fig. 2 This graphic from Slocum et al.'s work *Thematic Cartography and Geovisualization* (2009) allows for better visualization of the impact of different classification systems in how data is categorized and how it affects the end product images.

This particular data is meant to symbolize how many residents in the state of Florida are from foreign birth (Slocum et al., 2009). This kind of data can easily be politicized by any particular political entity to drive home policy that revolves around immigration. What about immigration are they trying to portray? Well, that is highly dependent on what kind of classification system they use. The graphic on the left side of figure 2. helps highlight what exactly is happening to the data.

Consider the classification system equal intervals, the lowest and highest numbers from the data are taken and divided into equal categories (ESRI, n.d.). The number of categories is left to the cartographer or statistician. Regardless, due to the nature of this data being skewed, the result of using this classification system is that the pole that it skews away from will be underrepresented. In this case, it would be data points near the high end of the range. The resulting graphic, map A, portrays this underrepresentation as most of Florida is light-coloured (low foreign-born residents). This graphic would be very useful to surreptitiously drive home the point that most residents of Florida are from American-born backgrounds and that all immigrants congregate near the Miami region.

Quantile classification dictates that the data itself, not the range, is divided into groups based on equal incidences (ESRI, n.d.). Since most of the plot points are on the low end of the range for this data set, they will be underrepresented. If you check the ranges presented in each category, they are vastly different. The group that this data skews towards has a difference of 0.8 percent (0.6 to 1.4), while the group from which the data skews away has a difference of 38.3 percent (6.8 to 45.1). This graphic, map B, makes it seem like that data is pretty evenly dispersed, which in reality is not. The number of immigrants on the northeast coastline is nowhere close to the number of immigrants near the southern tip.

Mean standard deviation might be the most absurd classification system used here. Mean standard deviation classification is very useful when the data is in a neat and tidy normal distribution (ESRI, n.d.). Standard deviation calculates how far away from the average the data sways, and if used with a skewed data set, the pole that the data skews away from will be highly

over-represented (ESRI, n.d.). This is further highlighted by the graphic, map C: the entire state is grey. This is highly inappropriate as distance away from the mean in the skewed end will be small, and distance away from the mean on the other end will be enormous. Although if the common observer were to see this representation without the knowledge of how the data was sorted, they would assume that the entire state of Florida is teeming with immigrants, which is far from the truth.

Maximum breaks, natural breaks (Jenks), and optimal (median) classification systems are similar mechanisms in how they divide up the data. Maximum breaks take the largest numerical gaps between the data sets and use that to divide up the data into categories (ESRI, n.d.). Since the data here is skewed it is not surprising to find the map mostly light coloured. Natural breaks or the Jenks natural breaks classification is a combination of two classification systems. It uses large numerical gaps all while minimizing deviation from the mean (ESRI, n.d.). When this classification system is used, the resulting map will look fairly evenly distributed. Although somewhat similar in their mechanism, natural and maximum breaks produce outputs with almost opposite underlying messages when used with skewed data. Additionally, optimal (median) classification is very similar to natural breaks (the output is virtually identical). This classification system is another algorithmic-based way of dividing the classes, similar to natural breaks, however, instead of focusing on minimizing deviation, optimal (median) classification is based on the medians of the subclasses (ESRI, 2020).

It is easy to see from figure 2. how different classification systems are utilized to transform the same underlying data set to convey different messages. The underlying data set in its own

inherent right is objective, however, human interaction with data nullifies its objectivity, as discussed above. The cartographer or data management professional who choses one classification over another has decided to interpret the data in one way or another based on his tendencies and biases. It is even more discouraging that when such figures are used in media and news outlets, the type of classification system used is very rarely disclosed. Thus, the common observer, who may not be an aspiring statistician, will assume that the data portrayed is objective and in turn internalize and formulate their meanings to it based on their own experiences, instincts, and feelings.

COVID-19, The Georgia Example

Figure 2. above, shows a very visual depiction of how misuse can be administered. This sort of statistical "sleight of hand" is ever prominent in the real world. Consider the statement by Mark Monmonier (2018) from his published book *How to Lie with Maps*,

> "[readers] must watch out for statistical maps carefully contrived to prove the points of self-promoting scientists, manipulating politicians, misleading advertisers, and other propagandists. Meanwhile, this is an area in which the widespread use of mapping software has made unintentional cartographic self-deception" (p. 153).

A rather egregious example of misrepresenting data in recent times comes from Georgia's Department of Public Health from 2020 in figure 3 below. The two figures look seemingly identical in terms of colour and visual representation, yet, the data set is grossly different. However, as pointed out by Monmonier (2018), this misrepresentation is not always intentional. During the

novel, confusing, and fear-inducing time of the SARS coronavirus 19 pandemic, news and informational media outlets have become of utmost importance. The information presented has a real gravitas since lives and public health are on the line. As shown in figure 3. retrieved from The Map Room, a website dedicated to geospatial technology, the Georgia Department of Public Health's website portrayed the daily number of infected citizens per one hundred thousand (Crowe, 2020).

Figure 3.

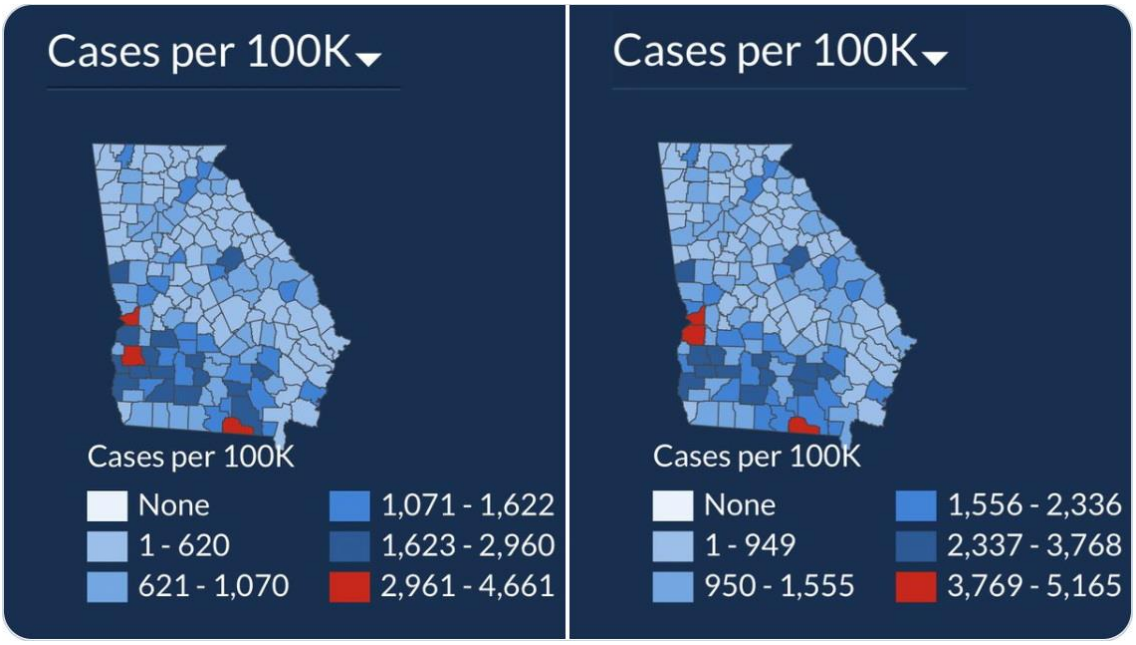

Fig. 3 two maps were taken from the state of Georgia's Department of Public Health, one earlier on in the pandemic (left) and one much later (right). Although incidence had gone up by 49%, the maps are visually similar (Crowe, 2020).

When looking at just the map and not the legend below both maps seem virtually identical. This may lead the reader to believe that the state has the pandemic under control, and that there is no reason to worry about the rising rates. Of course, when you look a little more closely you can see that the range of the scale bar has shifted. The red category in the first map has now shifted to be mostly blue. Two things could have potentially happened to produce such a mistake. The first one is rather appalling; the state deliberately contrived this statistical error to stem public uncertainties by placating fears of rising infection rates. The utilitarian reasoning behind a deception such as this could be understandable in hindsight, however, the loss of public trust is not a matter that should be taken lightly. The classification system used here seems most like a standard deviation system as the legend is continuous (thus any breaks systems are out), the colour scheme is not evenly spread (thus quantiles are out), and ranges are not uniformly divided (equal intervals are also out). This could have been avoided adding more groups for rising numbers and not reclassifying the data under the same number of groups with the same colour scheme. This blunder may have been a simple oversight or something much more conspiratorial. Regardless, this data was misrepresented and may have been used (intentionally or not) to invoke feelings that lead to a belief of infection rates not on being the rise.

The Use of Colour and Other Visual Aids

As mentioned earlier and complimented by Loewenstein et al. (2001), emotions are not to be understated in our cognitive processes. The use of colour schemes and other visual aids may seem arbitrary when discussing data however according to a colour-emotion study by

Takahashi and Kawabata (2017), colours based on hue and lightness show to induce emotional responses such as joy, sadness, anger and more. The use of colour is not an arbitrary choice, it is often chosen based on careful deliberation. Consider Figure 2., although purely an academic example, the figures choose to represent light colours as American-born and darker colours as foreign-born. Takahashi and Kawabata's (2017) study shows that darker colours are shown to produce more negative emotional responses. The reader may not be conscious of the associations being made between immigration and negative emotions, but it may still be effective in driving home a certain message about the data. Without explicitly showing anti-immigration sentiments the subtlety of colour choice and the emotional connection is a tool that can be used to further subvert objectivity from data. The example in Figure 3. also exemplifies this point in using bright red to denote areas most affected by the pandemic. Red, according to Takahashi and Kawabata's (2017) study is shown to exemplify a strong stimulus-response. In conjunction with the current pandemic and its atmosphere of heightened public uncertainty and fear, this stimulus-response would strengthen that emotional outlook. Residents in those most affected areas, after observing this figure put forward by the state's Department of Public Health, would be in an exacerbated state of fear and panic.

Conclusion

Consider the quote by the sociobiologist and author E. O. Wilson regarding the nature of human society, "[we are bound by] paleolithic emotions, medieval institutions and god-like technology" (An Intellectual Entente, 2009). Although, we have come so far in our technological

advancements and through closing the gap from tribes to global communities by increased connectivity, we are still under the thumb of our evolutionary history and instincts. Emotions drive us, whether we are conscious of them or not. Objective data is a farcical and theoretical term. As soon as potential data is observed and understood, it is subject to our interpretation, thus losing any notion of objectivity. We bring with us emotional baggage from past experiences, basal instincts, and feelings in how we interpret the data that surrounds us. Information and knowledge as discussed from Rowley's (2007) re-evaluation of the DIKW pyramid, are personal terms subject to our biases. Wisdom was not even broached as a topic since even the definition is muddied by subjective opinions. Symbolism via words lends weight to this subjective idea as well, since language is ever-evolving to include new definitions and ideas. It is not concrete, rather a continuum that resembles public opinion and political trends. The advent of mathematical formulations of statistical analysis was in theory a way of combating this inherent human tendency of subjectivity by removing the human influence all together through representation with objective numbers and equations. Even then, subjectivity creeps back. Choosing which descriptive statistical classifications to use is a subjective choice. How do you know which one is the right choice? What do you want the numbers to say? The misuse of stats to lend notions of objectivity to an entity can be dangerous. To the common observer or reader, stats give the information an aura of truth, but which truth? Colour schemes and additional visual aids further exacerbate this farcical idea of objectiveness in information gathering.

What does this mean for informational professionals? Well, until we can override our evolutionary history and flawed cognitive functions and essentially become computers capable

of only Boolean-like decision making, there are limited solutions. Education and awareness of our subjective tendencies and potential misinformation through bad actors is the only play in the playbook. On the other hand, I would have it no other way. What makes us human is our nuanced inquisitive subjective nature. It allows us to see the beauty in the data driven-world around us and drives our creative processes. It is not for the informational professionals to decide what kind of information is right or wrong, just to educate on how it is presented.

References

An Intellectual Entente. (2009). Harvard Magazine. Retrieved 10 December 2021, from https://www.harvardmagazine.com/breaking-news/james-watson-edward-o-wilson-intellectual-entente.

Bates, M. J. (2006). Fundamental forms of information. *Journal of the American Society for Information Science, 57*(8), 1033–1045.

Crowe, J. (2020). *Georgia's COVID-19 maps: Bad faith or bad design*?. The Map Room. Retrieved 10 December 2021, from https://www.maproomblog.com/2020/07/georgias-covid-19-maps-bad-faith-or-bad-design/.

Dickins, T. E., & Dickins, D. W. (2001). Symbols, Stimulus Equivalence and the Origins of Language. Behavior and Philosophy, 29, 221–244. http://www.jstor.org/stable/27759429

ESRI: Data classification methods. (n.d.). Pro.arcgis.com. Retrieved 10 December 2021, from https://pro.arcgis.com/en/pro-app/latest/help/mapping/layer-properties/data-classification-methods.htm.

ESRI: Standardize Field (Data Management). (2020). Pro.arcgis.com. Retrieved 10 December 2021, from https://pro.arcgis.com/en/pro-app/2.7/tool-reference/data-management/standardizefield.htm.

Kihuro, M. (2014). *Facts are stubborn things, but statistics are pliable*. The East African. Retrieved 10 December 2021, from

https://www.theeastafrican.co.ke/tea/oped/comment/facts-are-stubborn-things-but-statistics-are-pliable--1328564.

Loewenstein, G., Weber, E., Hsee, C., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127*(2), 267-286. https://doi.org/10.1037/0033-2909.127.2.267

Monmonier, M. (2018). *How to lie with maps* (3rd ed., p. 153). University of Chicago Press.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163-180. https://doi.org/10.1177/0165551506070706

Slocum, T.A., McMaster, R. B., Kessler, F. C., Howard, H. H. (2009). *Thematic Cartography and Geovisualization*. 3rd Ed. Harlow, Essex : Pearson Education ltd.

Takahashi, F., & Kawabata, Y. (2017). The association between colors and emotions for emotional words and facial expressions. *Color Research & Application*, *43*(2), 247-257. https://doi.org/10.1002/col.22186

Zhirenova, S.A., Satemirova, D.A., Ibraeva, A.D., Tanzharikova, A.V. (2016) "The Cognitive Content of the World of Symbols in a Language," *International Journal of Environmental & Science Education*, 11(9), pp. 2841–2849. Available at: https://doi.org/10.12973/ijese.2016.725a.