A FOUNDATION FOR INDUCTIVE REASONING IN HARNESSING THE POTENTIAL OF BIG DATA

ROSSI A. HASSAD
Mercy College, New York
Rhassad@mercy.edu

ABSTRACT

Training programs for statisticians and data scientists in healthcare should give greater importance to fostering inductive reasoning toward developing a mindset for optimizing Big Data. This can complement the current predominant focus on the hypothetico-deductive reasoning model, and is theoretically supported by the constructivist philosophy and Gestalt theory. Big-Data analytics is primarily exploratory in nature, aimed at discovery and innovation, and this requires fluid or inductive reasoning, which can be facilitated by epidemiological concepts (taxonomic and causal) as intuitive theories. Pedagogical strategies such as problem-based learning (PBL) and cooperative learning can be effective in this regard. Empirical research is required to ascertain instructors' and practitioners' perceptions about the role of inductive reasoning in Big-Data analytics, what constitutes effective pedagogy, and how core epidemiological concepts interact with the evidence from Big Data to produce outcomes. Together these can support the development of guidelines for an effective integrated curriculum for the training of statisticians and data scientists.

Keywords: *Statistics education research; Big Data; Big-Data analytics; Reasoning; Inductive reasoning; Deductive reasoning; Constructivism; Gestalt theory*

1. INTRODUCTION

The quote: "We adore chaos because we love to produce order" from M. C. Escher (as cited in Lamontagne, 2003. p. 71) aptly reflects the intricacies, novelty, and challenges of Big Data, in a culture of pervasive digitization of information and the internet of things, where "N = All" (Harford, 2014, p. 17), and just about any relationship can be rendered statistically significant (McAbee, Landis, & Burke, 2017). Of foremost consideration therefore, are validity and confounding (Stricker, 2017; Ehrenstein, Nielsen, Pedersen, Johnsen, & Pedersen, 2017). The Big-Data concept has been exponentially evolving for close to two decades (Laney, 2001) with a general focus on automated decision support (Monteith & Glenn, 2016). It has infiltrated academia, industry, and government, becoming an inescapable phenomenon for growth and development (Taylor & Schroeder, 2015).

Albeit lagging other disciplines (Sacristán & Dilla, 2015), the upward trend in the use of Big Data holds for healthcare including clinical and public health, with the mantras of translational medicine (Jordan, 2015) and evidence-based practice (Lehane et al., 2018). These themes are the focus of biomedical informatics, which has been characterized as "big data restricted to the medical field" (Holmes & Jain, 2018, p. 1). Biomedical informatics has an interdisciplinary reach and encompasses data linkage, integration, and analysis (Kulikowski, 2012; Walesby, Harrison, & Russ, 2017), toward efficient and effective clinical decision making (Kyriacou, 2004). Areas of application include genomics, disease surveillance, medical imaging (McCue & McCoy, 2017), and to a lesser extent, treatment and vaccine development. One domain, in

Statistics Education Research Journal, 19(1), 238–258, http://iase-web.org/Publications.php?p=SERJ

which Big Data has been particularly impactful, is disease surveillance including Influenza, Ebola, and Zika (Chowell, Simonsen, Vespignani, & Viboud, 2016, p. 376):

> Infectious disease surveillance is one of the most exciting opportunities created by Big Data, because these novel data streams can improve timeliness, and spatial and temporal resolution, and provide access to "hidden" populations. These streams can also go beyond disease surveillance and provide information on behaviors and outcomes related to vaccine or drug use.

Indeed, effective surveillance and control of infectious diseases requires rapid and comprehensive data collection. Population movement and behavior are central in this regard, and "premium is placed on information that is more recent and granular" (Dolley, 2018, p. 3). Thus, access to large volumes of real-time digitized information from various sources of Big Data such as mobile phones (and GPS software), electronic health records, social media, and travel records are pivotal to stemming an epidemic, as was the case for the Ebola outbreak in West Africa (Fast & Waugaman, 2016). The advent of Big Data has given renewed emphasis to evidence-based practice.

Evidence-based practice (EBP) in healthcare has a long and controversial history (Tanenbaum, 2005), and is generally defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett, 1997, p. 1). EBP within the health professions has a predominant focus on the use of evidence from experimental studies (Krumholz, 2014; Meza & Yee, 2018), which can allow for directly implying causality, the cornerstone of diagnosis, treatment, and prognosis (Russo & Williamson, 2011). Big Data, however, is typically characterized by correlations rather than causal relationships, which could explain the health sector's slow embrace of Big Data (Sacristán & Dilla, 2015). Nonetheless, Big Data is touted as having the potential for personalized and precision medicine (Mayer, 2015). Indeed, data from correlational studies can contribute to causal reasoning and understanding, using fundamental evaluation and organizing frameworks, such as the Bradford-Hill criteria (Phillips & Goodman, 2004) and the epidemiological triad (Galea, Riddle, & Kaplan, 2009).

As is typical of paradigm shifts, Big Data is fraught with controversies, with some viewing it as a promise "for revolutionizing, well, everything" (Harvard, 2012, p. 17; Bansal, Chowell, Simonsen, Vespignani, & Viboud, 2016; Lee & Yoon, 2017) while others consider it to be just noise, a "big mistake" (Harford, 2014, p. 1), or nothing new (Bellazi, 2014). Although there is no consensus as to a definition of Big Data, what is obvious is that it is about size, that is, data that is too large to be processed using conventional statistical methods and software, and hence requires innovative expertise to generate actionable evidence (Dinov, 2016). Implicated here, are artificial intelligence (AI), machine learning (ML), and human cognition. Notably, such technological tools (AI and ML) lack "contextual reasoning ability" (Singh et al., 2013, p. 2), and this can result in spurious correlations and other outcomes.

2. PURPOSE, RATIONALE, AND SIGNIFICANCE

This paper argues that inductive reasoning is a necessary cognitive tool for efficient and effective Big-Data analytics, and should be given greater importance in training programs for statisticians and data scientists in healthcare toward developing a mindset for optimizing Big Data, and supporting evidence-based practice. In addition to augmenting our understanding of Big Data, incorporating strategies for inductive reasoning into the instructional repertoire can help to meet the diverse cognitive preferences of learners. This is discussed with attention to the current predominant focus on the hypothetico-deductive reasoning model, considered a closed or circular system of logic. It is noted that Big-Data analytics is primarily exploratory in nature, aimed at discovery and innovation, and that this requires fluid or inductive reasoning.

The constructivist philosophy of learning and Gestalt theory are posited as a theoretical framework that supports inductive reasoning as a basis for effective learning and practice. Two epidemiological concepts, the epidemiological triad (a taxonomic model) and the Bradford-Hill criteria (a causal reasoning model) are proposed and discussed as conceptual models, which can serve as intuitive theories for inductive reasoning in Big-Data analytics. A distinction is made between inductive reasoning as a creative process of the mind, and inductive inference, a rigid mathematically determined procedure. Artificial intelligence and machine learning are addressed with the view that the power of Big Data and advanced technology does not reduce the need for human reasoning.

Furthermore, the discussion supports the call for reform in statistics education (Ridgway, 2016), regarding what constitutes competent practice of statistics and data science. We are witnessing exponential growth in Big Data, and exploring its complex structure for novel insights requires diverse expertise beyond what is typically used with traditional data sets. Therefore, it is critical that the needs of Big-Data analytics be addressed by academic institutions and professional organizations when developing criteria for recruiting and preparing the next generation of statisticians and data scientists. Outreach strategies should address the varied knowledge and skill sets, as well as disciplinary background that can provide a strong foundation to become a competent statistician working with Big Data. Such messages should emphasize the importance of obtaining a holistic perspective of data including contextual understanding, rather than the status quo, which tends to focus almost exclusively on the hypothetico-deductive model and mathematical procedures.

Notably, the characterization and widespread perception of statistics as mathematics is associated with high levels of anxiety and fear among some students and prospective candidates (Luttenberger, Wimmer, & Paechter, 2018), and this can be a major barrier to pursuing a career as a statistician or data scientist. Hence, outreach strategies should highlight the necessary and important role of inductive reasoning or creative thinking in Big-Data analytics, as this can generate greater interest in the discipline from diverse groups, and help to mitigate the dearth of qualified personnel.

3. BACKGROUND

3.1 CONCEPTUALIZING BIG DATA

The term Big Data is undoubtedly a universal and evolving concept that has become a metaphor for discovery and innovation (Krumholz, 2014) through the application of advanced analytics to massive amounts of complex data. While the terms Big Data and Big-Data analytics are generally used in a synonymous sense, the latter refers to the modern tools, methods, and procedures used for making sense of Big Data. A key assumption is that the size and complexity of Big Data render it rich with insight and intelligence, which cannot be obtained using traditional data sets and technology (Church & Dutta, 2013; Bellazzi, 2014). However, there is no standard definition (Walesby, Harrison, & Russ, 2017; Lefèvre, 2018), and this is consistent with the view that Big Data is not monolithic in terms of composition and architecture, as well as technological and analytical needs, and hence its conceptualization could vary across disciplines, time, and contexts, in order to be relevant and beneficial (Auffray et al., 2016).

Nonetheless, there are core features known as the **V**s of Big Data (Russom, 2011). That is, Big Data constitutes massive amounts or **V**olumes of data, potentially on the scale of terabytes, petabytes, or exabytes (Panda, 2016; Chang, 2016), from a **V**ariety of sources and in various forms, mostly unstructured (Bamparopoulos, Konstantinidis, Bratsas, & Bamidis, 2016) but also semi-structured and structured, and is accumulated at high speed or **V**elocity, primarily through digitization of information. A fourth feature, referred to as **V**eracity, was subsequently recognized by Big-Data practitioners to give emphasis to data quality in terms of reliability and validity, recognizing the uncertainty and messiness of Big Data (Koutkias & Thiessard, 2014).

Moreover, Big Data may be static or dynamic, and its high magnitude, dimensionality, and complexity require innovative data-analytic procedures and technology to obtain useful information (Chen, Chiang, & Storey, 2012).

The largely unstructured or fluid nature of Big Data puts it primarily in the domain of exploratory data analysis (Chauhan & Kaur, 2015) including predictive analytics and data mining. Big Data does not typically refer to large conventional (relational) databases, with structured data existing as well-defined variables, collected to examine specific a priori research questions, objectives, or hypotheses. Also, although SQL (Structured Query Language), a standardized programming language, supports large volumes of data, it does not technically allow for optimal solutions for the existing and emerging needs of Big-Data analytics, in terms of speed, scalability, availability, and consistency (Binani, Gutti, & Upadhyay, 2016).

The various definitions of Big Data reflect a continuum from the mechanistic, which focuses on data size, algorithms, and technology, primarily AI-driven machine learning (Jain, 2017), to the pragmatic, which in addition to data size and technology, emphasizes data quality and decision-making (Bellazzi, 2014; McAfee & Brynjolfsson, 2012), as well as ethical and privacy considerations regarding linking, storing, transferring, and sharing data (Lefèvre, 2018; Walesby, Harrison, & Russ, 2017). The conventions that practitioners adopt will influence how they conceptualize and operationalize Big-Data analytics, and hence determine the quality and utility of the knowledge that is derived from Big Data. The core features, **V**olume, **V**ariety and **V**elocity, make Big Data inherently messy, with concerns about **V**eracity; hence uncertainty abounds in terms of the insight and meaning that can be derived from the data.

The major challenge for Big-Data analytics practitioners is to separate the signal (the useful information) from the noise or interference (Ehrenstein, Nielsen, Pedersen, Johnsen, & Pedersen, 2017), and this encompasses proficiency in interpreting and explaining observed correlations and patterns, as well as meaningful conceptualization of the data for exploratory analysis. In essence, Big-Data analytics is largely about decision-making under uncertainty and complexity, and best practice requires interdisciplinarity and collaboration (Panda, 2016; Mayer, 2015; Harford, 2014). More importantly in this regard, is a need to shift from the predominant rigid deductive reasoning approach to more inductive and fluid reasoning (Ferrer, O'Hare, & Bunge, 2009) as supported by the constructivist philosophy in terms of knowledge production and meaning-making (Malle, 2013). This need is particularly evident with pattern recognition using data visualization maps, a major Big-Data analytic technique. Inductive reasoning fosters thinking beyond existing schemas, including simplistic linear and reductionist approaches, which do not capture the realistic complexity of health and behavioral phenomena (Cavanagh & Lane, 2012; Krumholz, 2014).

3.2 EXTRACTING VALUE FROM BIG DATA

The value of Big Data goes beyond its size, to its multidimensionality and complexity, the unraveling of which is the domain of data mining and machine learning (ML), driven by algorithms premised on assumptions about the latent knowledge that can be uncovered (Lee & Yoon, 2017; Bellazzi, 2014). Machine learning, a cognitive technology application, is a variant of artificial intelligence (AI), in which computer systems are initially "trained" or "supervised" by data scientists "to think" using large and varied data sets (NIH, 2018) and utilizing techniques such as pattern recognition, correlation, regression, multidimensional scaling, and cluster analysis. The computer algorithms "learn" from examples to deduce outcomes or objects based on predictive features or characteristics of labeled input information, much like deep or conceptual learning. This capability can then be transferred to other contexts to gain insight. It is an iterative process that allows computer models to independently adapt and explore when exposed to new data, rather than having to be specifically programmed. The IBM Watson is the foremost example to date of ML (Chen, Argentinis, & Weber, 2016).

Indeed, what we program (as in computerized algorithms) determines what we get, visualize or produce in terms of correlations and patterns. This evokes the philosophical principles of epistemology and ontology (Kitchin, 2014), that is, what constitutes knowledge, and how is it organized and represented. Certainly, the very lure of Big-Data analytics is that it can allow for important discoveries not achievable with traditional data and technology, however, this also demands a high level of conceivability or creative thinking to discover the unexpected. Therefore, focusing on the technical aspects or mechanics of Big Data, such as size and technology is necessary but not sufficient if we are to realize the potential of Big Data. Big Data might be huge but its "power" does not eclipse the need "for vision or human insight." (McAfee & Brynjolfsson, 2012, p. 7).

A case in point is Google's data-aggregating tool, Google Flu Trends (GFT), considered to be the first application of Big-Data analytics in the public health field (Zou, Zhu, Yang, & Shu, 2015). It was designed to estimate influenza-like illness from online search activity, but the estimation algorithm consistently and significantly resulted in overestimation. Notably in 2012 and 2013 the GFT predicted more than double the actual prevalence of flu reported by the Centers for Disease Control and Prevention (CDC) (Lazer, Kennedy, King, & Vespignani, 2014). And this Big-Data failure was attributed by experts to the purely mechanical nature of the Google algorithm, which was free of theory (Harford, 2014) and not adapted (Butler, 2013) to reflect temporally relevant factors. A major unrecognized confounding factor was the widespread dramatic and fear-mongering news stories about the Flu, particularly in 2012, which resulted in an unduly high level of related search activity by those with and without flu symptoms (Harford, 2014). This clearly has implications for the way we think about and reason with Big Data. Big Data needs new thinking (Krumholz, 2014).

3.3 REASONING WITH BIG DATA

Undoubtedly, advanced computer technology including artificial intelligence is desirable in a world obsessed by bigger, faster, and now. But conceptualizing, explaining, and meaning-making require particular cognitive expertise (Krumholz, 2014); this "requires man, not machine" (Jenkins, 2013, p. 4). This consideration and need is not new to statistics or for that matter any scientific discipline, but is of key importance to Big Data, given the high level of complexity and multidimensionality, and hence uncertainty about what it represents, and the intelligence that can be derived. Human reasoning is complex (Johnson-Laird, 2010), and falls within the purview of cognitive psychology.

According to Jones (2016, p. 5) "cognitive models help us to know where to search and what to search for as the data magnitude grows." Additionally, the high degree of uncertainty and dimensionality calls for emphasis on metacognition or reflective thinking to help to guard against decision-making pitfalls such as confirmatory bias and failure to recognize spurious correlations (Lane & Corrie, 2007). And lest we forget, when it comes to Big Data, the "curse of dimensionality" is ever present (Dinov, 2016, p. 9). Other well-established cognitive biases must also be considered (Tversky &, Kahneman, 1974).

Kitchin (2014, p. 10) notes that "there is an urgent need for wider critical reflection on the epistemological implications of Big Data and data analytics" toward embracing inductive reasoning (Krumholz, 2014). The mainstream hypothetico-deductive approach (frequentist and Bayesian) to research and data analysis reflects a closed or circular system of logic (Gelman & Shalizi, 2013), and this defeats the purpose of Big-Data analytics, which is largely exploratory, that is, the goal is to generate novel and meaningful insights and hypotheses "born from the data" rather than from theory (Kelling et al., 2009, p. 613). Tukey (as cited in Jones, 1987, p. 606) characterized exploratory data analysis as "flexible realistic data analysis that tries to let the data speak freely."

4. INDUCTIVE REASONING

There is some semantic ambiguity surrounding the term *reasoning* also commonly used interchangeably with *inference*. While statistical inference associated with hypothesis testing is inductive (Evans, 1993), it is not inductive reasoning. Rather, it is a rigid algorithmic procedure for determining a population estimate from a sample statistic, and requires no creative processes of the mind (or reasoning), in the true psychological sense. Bakker, Kent, Derry, Noss, and Hoyles (2008, p. 139) refer to hypothesis testing as "mostly a formalized form of inductive inference," and according to Fisher (1935, p. 39) although it involves "uncertain inferences," it does not mean that they are not "mathematically rigorous inferences." Lehmann (2012, p. 1069), reported that Neyman characterized the mere acceptance and rejection of a hypothesis as void of "cognitive significance." Specifically, Neyman (1957, p. 13) observed, that this process does not "involve reasoning," rather it is "an automatic adherence to a pre-assigned rule," a position also supported by Fisher (1935, p. 40) in noting that a probabilistic statement of uncertainty "is no guarantee of its adequacy for reasoning of a genuinely inductive kind."

Cognitive scientists have characterized inductive reasoning as the ability to be creative and generate novel ideas and discovery (De Koning, Sijtsma, & Hamers, 2003; Sternberg & Gardner, 1983; Klauer & Phye, 2008). According to Hayes and Heit (2018, p. 1): "Inductive reasoning entails using existing knowledge to make predictions about novel cases." The call for greater emphasis on inductive reasoning in the context of Big-Data analytics represents an appeal for intellectual flexibility (Kyriacou, 2004), i.e., an open system of logic. The positivist perspective that knowledge is objective, and can only be derived from hypothetico-deductive reasoning models (Bonell, Moore, Warren, & Moore, 2018), is limiting and myopic. The very theories and models used for deductive reasoning were initially developed using qualitative research and inductive reasoning (Tonidandel, King, & Cortina, 2018). Indeed, both modes of reasoning are important, toward augmenting our understanding of health and behavioral phenomena through Big Data, and optimizing its value.

There is an abundance of research indicating that humans are naturally predisposed to perform inductive reasoning over deductive reasoning (Nisbett, Krantz, Jepson, & Kunda, 1983; Kemp & Tenenbaum, 2009), especially when faced with challenging situations, problems, and uncertainty. They intuitively construct and explore hypotheses to arrive at a solution (Arthur, 1994). This is supported by the popular category-based induction model (Hayes, Ruthven, & Newell, 2007; Hayes & Newell, 2009), which posits that amidst uncertainty, humans perform a global inspection of the information or task (Klauer & Phye, 2008), comparing (and contrasting) characteristics of objects and items, identifying similarities (and differences), and heuristically formulating, confirming or refuting generalizations (Klauer, 1996; Tversky, 1977). Core cognitive tasks associated with induction are "categorization, probability judgment, analogical reasoning, scientific inference, and decision making" (Hayes, Heit, Swendsen, 2010, p. 278), as well as "concept formation" (Nisbett, Krantz, Jepson, & Kunda, 1983, p. 339).

Current experimental research findings challenge the notion that inductive and deductive reasoning follow different neural pathways, that is, a dual-process model (Johnson-Laird, 2010; Hayes, Heit, & Swendsen, 2010), by showing that a single process or dimension accounts for both types of reasoning (Hayes, Stephens, Ngo, & Dunn, 2018). This is consistent with a long-standing theory indicating that inductive reasoning (or intuition, using heuristic models) and deductive reasoning (using analytical models) may occur on a cognitive continuum from purely intuitive to purely analytical, with varying degrees of both in between (Hammond, 1981; 2000). Theoretical support for inductive reasoning as a key dimension of effective learning and practice, particularly in the context of Big-Data analytics, includes constructivism (Malle, 2013; Fosnot, 2013), which characterizes learning as an iterative and meaning-making experience, and Gestalt theory (Rock & Palmer, 1990), which explains how humans are inclined to perceive and organize information.

5. THEORETICAL SUPPORT FOR INDUCTIVE REASONING WITH BIG DATA

Theoretical and philosophical foundations of learning serve to explain and predict the learning process and outcomes and hence facilitate effective curricular design and pedagogy. In this regard, behaviorism, cognitivism, and constructivism are recognized as the major theoretical frameworks (Ertmer & Newby, 2013), the latter being the dominant approach in the disciplines of Science, Technology, Engineering and Mathematics (STEM) (Kelley & Knowles, 2016), including statistics and data science. Another theory, which is becoming increasingly relevant to the teaching, learning, and practice of Big-Data analytics, is Gestalt (Weissmann, 2010).

5.1 CONSTRUCTIVISM

The constructivist philosophy of learning focuses on what constitutes knowledge and how we come to know, and is premised on the assumption that “knowledge is constructed in the mind of the learner” (Bodner, 1986, p. 873). It emphasizes a flexible, student-centered, active learning process of integrating information, identifying patterns, and making connections, toward a meaning-making experience (Fosnot, 2013) that fosters transferrable knowledge and skills. In other words, constructivism is associated with deep and meaningful learning including conceptual understanding. Moreover, constructivist methods have been equated to inductive teaching and learning, with specific reference to “inquiry learning, problem-based learning, project-based learning, case-based teaching, discovery learning, and just-in-time teaching” (Prince & Felder, 2006, p. 123; Dennick, 2016).

There are two major constructivist perspectives (Powell & Kalina, 2009); the cognitive, and the social or socio-cultural (Vygotsky, 1978). The socio-cultural model is a variant of cognitive constructivism, and is more mainstream in educational reform (Cobb, 1996). It posits that the extent of learning is influenced by the quality of the social interaction (collaboration, negotiation, and cooperation) within the learning context or community. According to social constructivism, understanding “is a function of the content, the context, the activity of the learner, and, perhaps most importantly, the goals of the learner” (Savery & Duffy, 1996, p. 136). This is usually contrasted with behaviorism, viewed as a rigid, passive, instructor-centered, stimulus-response system of learning, which is linked to rote or surface learning. The popularity and plausibility of constructivism is its cognitive underpinning.

Indeed, constructivism is just as controversial as it is popular, with labels such as “a fad” (Sjøberg, 2010, p. 485), and “the cult of constructivism” (Beatty, 2009, p. 464) especially regarding educational reform. This view is generally attributed to variability in the understanding of constructivism, including mis-characterizations (Fosnot, 2013). One common misconception is simply “viewing a series of tasks and laboratory activities as being equivalent to scientific inquiry” and “hands-on instruction,” without attention to “minds-on” experiences (Kelley & Knowles, 2016, p. 5), that is, actively engaging students in critical thinking and reasoning rather than emphasizing procedural tasks.

Yet a broader perspective of constructivism views it as a multi-theoretical framework, which embodies some aspects of behaviorism in addition to cognitivism. Gestalt psychology, which addresses principles of perception and pattern recognition, and is a forerunner of cognitive psychology (Tobias, 2010), is also implicated, as is the generative learning theory (Wittrock, 1992), which preceded constructivism, and posits that learners actively integrate new ideas into existing schemata or mental models to generate meaning and understanding. In this regard, “knowledge is being actively constructed by the individual and knowing is an adaptive process, which organizes the individual’s experiential world” (Karagiorgi & Symeou, 2005, p. 18; Hendry, 1996).

Accordingly, prior knowledge plays an important role in the constructivist learning context, for which there is strong empirical support, with prior knowledge accounting for about two thirds of the variance in learning (Tobias, 2010, p. 52; Sjøberg, 2010; Prince & Felder, 2006).

5.2 GESTALT THEORY

Of particular relevance to Big-Data analytics, is Gestalt psychology, which general concept is subsumed under constructivism. The Gestalt principles posit that the visual system organizes information based on innate laws of grouping, that is, humans are inclined to perceive objects and images based on similarity, proximity, continuation, and closure (Rock & Palmer, 1990), with a general proclivity for symmetry and simplicity (Chater & Vitányi, 2003). This is of key importance as scientists navigate and try to make sense of massive amounts of data. While these laws or principles generally result in accurate representations, spurious ones can also occur (Rock & Palmer, 1990). Gestalt, a German word, connotes shape, form, or configuration; Max Wertheimer, the pioneer of Gestalt psychology, stated the following on Gestalt, which is generally understood as "the whole is more than the sum of the parts" (Nesbitt & Friedrich, 2002, p. 738), as well as, the meaning of the part is determined by our perception of the whole (Wertheimer, 1924, as cited in Weissmann, 2010, p. 2137):

> The fundamental of the Gestalt "formula" might be expressed in this way: There are wholes, the behavior of which is not determined by that of their individual elements, but where the parts are themselves determined by the innate nature of the whole.

A main criticism of Gestalt psychology is that human perception is more than just grouping and segmenting; it also involves meaning and interpretation (Pinna, 2010; Boudewijnse, 2004), the foundation of constructivism, as noted by Vygotsky (1978, p. 33).

> A special feature of human perception – which arises at a very young age – is the *perception of real objects*. This is something for which there is no analogy in animal perception. By this term I mean that I do not see the world simply in color and shape but also as a world with sense and meaning. I do not merely see something round and black with two hands; I see a clock and I can distinguish one hand from the other.

Indeed, pattern recognition and visualization maps have become synonymous with Big-Data analytics, and rightly so if we are to derive its benefits. Green (1998, p. 1) argues that:

> Humans are poor at gaining insight from data presented in numerical form. As a result, visualization research takes on great significance, offering a promising technology for transforming an indigestible mass of numbers into a medium which humans can understand, interpret and explore.

Gestalt principles provide a framework for the effective design, organization, and presentation of information, as there seems to be a natural human preference for visual representations (Olshannikova, Ometov, Koucheryavy, & Olsson, 2015); a picture speaks a thousand words. But development in this area is lacking when it comes to realizing the potential of Big-Data analytics, and according to Guberman (2015, p. 25):

> The main cause of stagnation in this field has been the neglecting of knowledge accumulated in Psychology of Perception in general and in Gestalt Psychology in particular. Too much emphasis has been put on mathematics and engineering and too little on laws of human perception, which must be imitated.

According to Gestalt theory, we have an innate tendency to "constellate" elements that seem similar (Weissmann, 2010, p. 2139), and, being mindful of these automatic perceptions, can reduce the likelihood of bias in interpretation, and spurious representations.

5.3 INTUITION, PRIOR KNOWLEDGE, AND INDUCTIVE REASONING

Inductive reasoning is driven by intuition, which is informed by prior knowledge and is necessary for creativity and discovery (Tenenbaum, Griffiths, & Kemp, 2006). The critics however, contend that it is a blind and unscientific process, claiming the absence of theory and hypothesis (Kyriacou, 2004). Nevertheless, scientists come to the data with various conceptual models or schemas, and not as a tabula rasa or blank slate. Kitchin (2014, p. 5) states that "an inductive strategy of identifying patterns within data does not occur in a scientific vacuum and is discursively framed by previous findings" while according to Kemp and Tenenbaum (2009, p. 50) "induction draws on systems of rich conceptual knowledge."

Such prior mental models, schemata or knowledge repertoire are particularly relevant to Big-Data analytics where the ideas and hypotheses for exploration are vast and the analyst just cannot explore everything, hence, the need for a plausible basis for conceptualizing (Wise & Shaffer, 2015) at all stages of the statistical process. Unless data scientists possess at least a macro understanding of the domain of application, for example, surveillance for infectious diseases, genomics, medical imaging, etc., they could be considerably disadvantaged in terms of performing efficient and meaningful data analysis, and optimizing the potential of Big Data.

It has been said that "visualization tools work best when scientists have an idea of what to visualize in a pile of data" (Harvard, 2012). Indeed, this encapsulates inductive reasoning, the primary cognitive process for generalizing, and formulating scientific hypotheses, which are core elements of exploratory data analysis and discovery (Feeney & Heit, 2007), the key goals of Big-Data analytics including machine learning. This is reinforced by Tenenbaum & Kemp (2007, p. 1) and Mitchell (1997):

> Formal analyses in machine learning show that meaningful generalization is not possible unless a learner begins with some sort of inductive bias: some set of constraints on the space of hypotheses that will be considered.

Accordingly, equipping students and practitioners with conceptual frameworks (Imenda, 2014) – taxonomic and causal (Tenenbaum, Griffiths, & Kemp, 2006) – for organizing and evaluating data in the healthcare context can help to transform Big Data into "smart data" by generating novel ideas and discovery.

The risk of not using conceptual frameworks to guide exploratory analysis of Big Data is evident in scientific reports of false positive genomic associations (Kyriacou, 2004), and gross overestimates of flu cases using the Google Flu Trends tool (Harford, 2014). Hence, in a world of Big Data, where the stakes and potential benefits are high, the preparation and readiness of data scientists should not be left to assumptions and chance. Inductive reasoning applies to all phases of data analysis. Therefore, training programs for statisticians and data scientists should give greater importance to the development and application of inductive reasoning toward fostering a mindset for optimizing the value of Big Data. Of critical importance in this regard is the following (Kemp & Tenenbaum, 2009, p. 20):

> A psychological account of induction should answer at least two questions – what is the nature of the background knowledge that supports induction, and how is that knowledge combined with evidence to yield a conclusion?

With regard to the nature of the prior or background knowledge that supports inductive reasoning, and how that knowledge interacts with the evidence to produce an outcome or decision, Tenenbaum, Griffiths, & Kemp (2006) found empirical support for using either taxonomic (classification) or causal reasoning models, based on the properties of the inductive context (Kemp & Tenenbaum, 2009).

6. EPIDEMIOLOGICAL CONCEPTS AS BASIS FOR INDUCTIVE REASONING WITH BIG DATA

Given the uncertainty associated with Big Data, which is usually observational or non-experimental (Christensen & Davis, 2018), classification and causal reasoning models are particularly necessary for deriving understanding and meaning (Johnson, Rajeev-Kumar, & Keil, 2016). Toward this end, epidemiology, which is considered the twin discipline of statistics (Bhopal, 2016) offers concepts and frameworks that can serve as a firm basis for reasoning with Big Data in healthcare. A common definition of epidemiology is "the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to the prevention and control of health problems" (Last, 2001). In other words, it is concerned with patterns of distribution, determinants, and risk of diseases and other health-related states in the population (Bhopal, 2016), and is the core underpinning of evidence-based medicine (EBM).

Moreover, epidemiology is multidisciplinary in nature: it includes unifying and organizing frameworks for conceptualizing, modeling, and explaining variability, as well as facilitating causal analysis and understanding (Ahrens, Krickeberg, & Pigeot, 2005) and generating plausible hypotheses (Stroup, Goodman, Cordell, & Scheaffer, 2004). Such models operate as intuitive theories, which Tenenbaum, Griffiths, & Kemp (2006, p. 309) define as:

> An intuitive theory may be thought of as a system of related concepts, together with a set of causal laws, structural constraints, or other explanatory principles, that guide inductive inference in a particular domain.

While it has long been recognized that the study of epidemiology is fundamental to effective statistical practice (Abramson, 2000), particularly biostatistics, epidemiology is not generally a part of the required curriculum for the training of statisticians and data scientists (DeMets, Stormo, Boehnke, Louis, Taylor, & Dixon, 2006). Rather, epidemiology is usually listed as an optional course or incorporated into other elective courses. There have been repeated calls for an integrated curriculum that "merges biostatistics with clinically relevant medical discussions, such as those that occur in many EBM curricula for epidemiological principles" (West & Ficalora, 2007, p. 939; Gouda & Powles, 2014) as "one way to enrich the applied learning of statistics" (Bennett, 2016, p. 4). Moreover, this approach is associated with higher levels of statistical competence (Stevenson, 2016), greater satisfaction, and more positive attitudes toward statistics and research (West & Ficalora, 2007, p. 939; Al-Zahrani, & Al-Khail, 2015; Lawrence, 2016).

Despite the noted benefits of an integrated (epidemiology and biostatistics) curriculum, the long-standing challenge of adequately preparing instructors and practitioners and developing curricular materials remains an open task (Simpson, 1995; Del Mar, Glasziou, & Mayer, 2004; Bellan, Pulliam, Scott, Dushoff, & the MMED Organizing Committee, 2012). Two epidemiological concepts or frameworks of core relevance in this regard, especially in the context of Big Data, where exploratory data analysis is predominant, are as follows.

- The epidemiological triad: A taxonomic concept that provides a basis for gaining insight into and modeling health issues, specifically allowing for the orientation and classification of variables. This model allows for investigating the interaction of the agent (risk factor or cause), the host (person at risk), and the environment (the context).
- The Bradford-Hill criteria: A set of factors for evaluating research outcomes, particularly non-experimental data, to determine if causality can be inferred from observed associations.

These two concepts are typical of an introductory epidemiology course; they provide a basis for inductive reasoning including classifying and organizing variables, and identifying causal relationships.

6.1 THE EPIDEMIOLOGICAL TRIAD

The epidemiological triad (see Figure 1) is a concept that can facilitate statistical thinking and reasoning in the clinical and public health contexts, including infectious and non-infectious diseases (Galea, Riddle, & Kaplan, 2009); this triad is particularly relevant to Big-Data analytics, which is largely exploratory in nature.

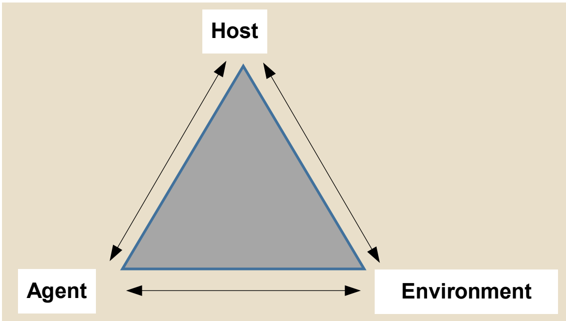

Figure 1. The Epidemiological Triad

This triangular representation depicts the three major factors that interact to cause or influence disease or other health outcomes (Bhopal, 2016). It is an empirical framework that provides a basis for conceptualizing and organizing health-related data, and generating plausible hypotheses, interpretations and explanations (Lipton & Ødegaard, 2005).

Agent The agent refers to the causal or risk factor(s) for the disease or outcome, recognizing that this could be a single factor, as in the medical and biological sciences where one organism (such as a virus) is the causal agent of an infectious disease. Multifactorial causation (or association) – generally referred to as a "web of causation" (Detels, Beaglehole, Lansang, & Gulliford, 2011) – is more applicable to the behavioral sciences, where it reflects the complexity of the underlying causal or explanatory mechanism. A common etiological model is the biopsychosocial concept of disease and wellness, which implicates biological, psychological, and sociological factors in multifactorial causality (Engel, 1980; Porta, 2014). In essence, the agent is what explains or brings about the outcome.

Host The host is generally the person who is at risk of developing or harboring the disease.

Environment The environment refers to the context where the agent can interact with the host to bring about the disease or other outcome. It encompasses the surroundings, conditions, and other influences external to the host.

Moreover, this triad reinforces the principle of "necessary but not sufficient" (Porta, 2014) in terms of modeling data to explain and predict outcomes, and for causal reasoning. That is, while each factor is required for development of the disease or outcome, it is the combination and interaction of these factors that determine if and the extent to which the outcome occurs. The epidemiological triad can be further detailed and expanded by noting that each of the three factors (agent, host, and environment) is itself influenced by other factors, which can alter or modify the level of susceptibility or risk of disease occurrence, as well as determine approaches to prevention and control. The interrelationships among these factors can be understood as follows.

For example, age, immunization status, genetic predisposition, and nutritional status can affect the susceptibility or risk of the host or person. Sanitary conditions, availability of public health services such as vector control (e.g., mosquitoes), and housing quality (such as crowding) can influence risk posed by the environment. And the virulence of a pathogenic or disease causing organism (the agent) or the potency of a drug (the agent) in the case of substance use disorders, will impact the degree or severity of the disease or outcome.

The epidemiological triad therefore provides a basis for modeling disease and other health outcomes, specifically allowing for the orientation and classification of variables as follows (Bhopal, 2016; Porta, 2014).

1. An exposure, cause, or risk factor, which is also known as the independent, predictor, or explanatory variable.
2. An outcome, response, or criterion variable, which is generally referred to as the dependent variable.
3. A mediator (intervening or intermediate) variable, which accounts for or explains the relationship between the exposure or risk factor and the outcome.
4. A moderator or effect modifier also known as an interaction variable, which can alter the effect of the risk factor (the independent variable).
5. A confounding factor, which is a variable that is associated with both the exposure (independent variable) and the outcome (dependent variable) resulting in biased or spurious associations.

Any inference resulting from statistical modeling, particularly exploratory data analysis and data from non-experimental research designs, requires strict evaluation to determine whether the relationships observed can imply causality or are just associative. The Bradford-Hill criteria (Hill, 1965; Fedak, Bernal, Capshaw, & Gross, 2015), establish a model of causal reasoning, which may be relevant in this regard.

6.2 THE BRADFORD-HILL CRITERIA

In the context of evidence-based practice, healthcare research emphasizes experimental methods (Greenhalgh, Howick, & Maskrey, 2014) that are hypothetico-deductive in nature. In other words, the researcher begins with a theoretical framework, based on the problem and context under investigation, and the analysis and interpretation of the data are informed or guided by the underlying theory. This approach however, does not capture the reality of Big-Data analytics, which is mostly focused on observational data and exploratory data analysis to generate insights and hypotheses largely through inductive reasoning. Therefore, in the Big-Data context, the quality and utility of the evidence, often referred to as veracity and evaluated in terms of reliability and validity, attract greater scrutiny, especially given the myriad of potential confounding factors (Brookhart, Stürmer, Glynn, Rassen, & Schneeweiss, 2010).

The "gold standard" of medical research is the randomized controlled trial or RCT (a true experimental study), which can determine whether an outcome can be attributed to the manipulated variable or intervention (Vogt, Gardner, & Haeffele, 2012). However, given that Big Data is predominantly observational, statistically significant associations must be carefully evaluated to determine whether causality can be implied. Statistical significance by itself must not be misconstrued as evidence of importance or causality.

Hill (1965) proposed a set of factors for consideration when evaluating statistically significant relationships or associations to determine whether causality can be inferred. These considerations are now commonly referred to as the Bradford-Hill criteria (Rothman & Greenland, 2005), a standard that is widely recognized in the field of epidemiology. The nine criteria are discussed below. They should be used as guidelines for assessing the quality of evidence and not as absolute and exhaustive criteria for determining causality.

1. *Strength*: This is generally determined by the magnitude of the relationship or association using statistical measures. While there is the tendency to regard strong relationships as being potentially causal, the role of confounding must always be considered. As well, a weak relationship should also be carefully assessed as it could be implicated in multifactorial causality rather than a one-to-one or single-cause relationship.
2. *Consistency*: This refers to the reliability, replicability, reproducibility, or robustness of the findings. That is, was the finding repeatedly observed by different researchers, in different settings, using different methods?
3. *Specificity*: This was meant to apply to direct one-to-one relationships. That is, if a single specific exposure is linked to a particular disease or outcome, it is strong evidence of a causal effect. However, this is a narrow view of causation, as a particular factor or exposure can be causally implicated in multiple diseases or outcomes, hence a lack of specificity, but still causal. As well, a set of factors can bring about a particular disease or outcome, as in multifactorial causality.
4. *Temporality*: This is the "chicken and the egg" conundrum and refers to the sequence of the factors in the relationship. Which comes first? In order to establish causality, the hypothetical cause (or exposure) must precede or come before the disease or outcome (the hypothetical effect). Good logic dictates that this is an "inarguable" necessary criterion for causation (Rothman & Greenland, 2005, p. 148).
5. *Biological gradient*: This refers to a dose-response relationship. Simply put, if an exposure causes a disease or outcome, then an increase in the exposure (a higher dose) should increase the risk of occurrence of the disease or outcome, suggesting a linear or at least monotonic relationship. However, causality is often complex and nonlinear, and hence while the concept of biological gradient may be informative for data exploration, it might not be helpful or relevant to assessing causation. Also, assessment for confounding is critical in this context.
6. *Biological plausibility*: This means that the observed association is consistent with the established body of scientific knowledge and understanding. In other words, the association makes scientific sense as it has a rational and theoretical basis. Indeed, caution is required here, as assessment for plausibility is limited to the existing body of knowledge, and as Hill (1965, p. 10) reminds us, "the association we observe may be one new to science or medicine and we must not dismiss it too light-heartedly as just too odd." In other words, we must be open to discovering the unexpected. Also, plausibility is not always purely biological in the medical and health context. For example, the literature also refers to psychological and sociological plausibility.
7. *Coherence*: Hill (1965, p. 10) characterized this criterion by noting that "the cause-and-effect interpretation of our data should not seriously conflict with the generally known facts of the natural history and biology of the disease." The "natural history of disease refers to the progression of a disease process in an individual over time, in the absence of treatment" (CDC, 2006, p. 59), whereas biology relates to pathogenesis or the underlying disease mechanism. While this pertains to the clinical setting, the concept of coherence can be extended to other fields of research and contexts. Coherence is how consistent the observed association is with the wider body of direct and indirect scientific evidence, and is therefore akin to plausibility, albeit broader. In assessing the evidence, researchers should be mindful of potential misinterpretation and misunderstanding, as well as the limitations of the extant body of knowledge, and therefore, not readily "nullify" observed associations that cannot be substantiated.
8. *Experiment*: Experimental research provides the strongest evidence in support of causal inference, depending on the quality and complexity of the methodology. This design allows for examining the effectiveness of treatments or interventions, which are manipulated to test a predetermined hypothesis. While a core requirement of the experimental design, particularly a true experiment, such as a randomized controlled

trial, is control for confounding factors, bias must always be considered, and alternative explanations for the outcomes explored.

9. *Analogy*: This results when information from external sources is used to explain or clarify the observed association because of similarity in properties (Bartha, 2013). This criterion is directly related to plausibility and coherence, and is the basis of inductive reasoning (Landes, Osimani, & Poellinger, 2018), a cognitive process that is necessary for effective Big-Data analytics (McAbee, Landis, & Burke, 2017). The "absence of such analogies only reflects lack of imagination or experience, not falsity of the hypothesis" (Rothman & Greenland, 2005, p. 149).

7. SUMMARY AND IMPLICATIONS

The premise of this paper is that training programs for statisticians and data scientists in healthcare should give greater importance to fostering inductive reasoning toward developing a mindset for optimizing Big Data, and supporting evidence-based practice. Indeed, this is at variance with the current predominant focus on the hypothetico-deductive reasoning model which emphasizes algorithmic and mathematical procedures. In addition to augmenting our understanding of Big Data, incorporating strategies for inductive reasoning into the instructional repertoire can help to meet the diverse cognitive preferences of learners. As well, the ubiquitous perception of statistics as mathematics, and the associated high levels of anxiety and fear among some students and prospective candidates can be a major barrier to pursuing a career as a statistician or data scientist. Therefore, outreach communication should address the important role of inductive reasoning or creative thinking in Big-Data analytics, as this can generate greater interest in the discipline from diverse groups, and help to mitigate the dearth of qualified personnel.

A key assumption underlying this paper is that inductive and deductive reasoning are not orthogonal or mutually exclusive processes or constructs but occur on a cognitive continuum. Inductive reasoning is a more open system of logic; the ability to be creative, and generate novel and meaningful insights and hypotheses from the data rather than from theory. Of note is that while statistical inference associated with hypothesis testing is inductive; it is not inductive reasoning; rather statistical inference follows a pre-assigned rule.

The constructivist philosophy of learning and Gestalt theory constitute a theoretical framework that supports inductive reasoning as a necessary cognitive tool for effective learning and practice of Big-Data analytics. Pedagogical strategies such as problem-based learning (PBL) and cooperative learning can be effective in this regard. Constructivism posits that knowledge is constructed in the mind of the learner, and learning is a meaning-making experience. Whereas, the Gestalt principles explain how humans are inclined to perceive objects and images. Gestalt psychology reminds us that we have an innate tendency to "constellate" elements that seem similar, and being mindful of these automatic perceptions, can reduce the likelihood of bias. Moreover, there seems to be a natural human preference for visual representations.

Big-Data analytics is primarily exploratory in nature, aimed at discovery and innovation or producing the unexpected and this requires fluid reasoning akin to inductive reasoning, which draws on systems of rich conceptual knowledge. Prior knowledge base, which can serve as a meaningful foundation for inductive reasoning includes taxonomic (classification) and causal reasoning models. These conceptual models function as intuitive theories but are not typically addressed in the core statistics curriculum. However, given the uncertainty inherent in Big Data, not to mention its largely observational or non-experimental nature, such conceptual models, which can be adopted from the discipline of epidemiology, can be particularly useful for deriving understanding and meaning. This calls for a more integrated curricular approach. Toward this end, this paper proposes that two core epidemiological frameworks, the epidemiological triad (a taxonomic concept) and the Bradford-Hill criteria (a causal reasoning model) have the potential to facilitate inductive reasoning toward more efficient and effective Big-Data analytics in healthcare.

While the charm of Big Data is centered on its size, its promise is based on assumptions about the intelligence that is nested in its complex and high-dimensional structure. The major challenge for Big-Data analytics practitioners is to separate the signal or useful information from the noise or confounding, and transforming Big Data into smart data. And as fascinating as artificial intelligence and machine learning are, this is just technology, and what can be produced is dependent on human capacity for creative thinking and conceivability.

Finally, empirical research is required to ascertain instructors' and practitioners' perceptions about the role of inductive reasoning in Big-Data analytics, as well as what constitutes effective pedagogy. The epidemiological triad and the Bradford-Hill criteria should be further explored, in this regard, to determine how the knowledge from each model or concept interacts with the evidence from Big Data to produce outcomes. Together, these can support the development of guidelines for an effective integrated curriculum for the training of statisticians and data scientists who understand how to maximize the value of Big Data.

ACKNOWLEDGEMENTS

I am grateful to the editors and reviewers for their superb feedback and guidance. Special thanks to my students, Mercy College, and the team from Canterbury Christ Church University, UK.

REFERENCES

Abramson, J. H. (2000, August). Teaching statistics for use in epidemiology. Paper presented at tha Joint Statistical Meetings 2000, Indianapolis.

Ahrens, W., Krickeberg, K., & Pigeot, I. (2005). An introduction to epidemiology. In W. Ahrens & I. Pigeot (Eds.), *Handbook of epidemiology* (pp. 3–42). Berlin: Springer.

Alzahrani, S. H., & Aba Al-Khail, B. A. (2015). Resident physician's knowledge and attitudes toward biostatistics and research methods concepts. *Saudi Medical Journal*, *36*(10), 1236–1240.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *The American Economic Review*, *84*(2), 406–411.

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., ... & Zanetti, G. (2016). Making sense of big data in health research: Towards an EU action plan. *Genome Medicine*, *8*: 71. [Online: doi.org/10.1186/s13073-016-0323-y]

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistics Education Research Journal*, *7*(2), 130–145. [Online: iase-web.org/Publications.php?p=SERJ_issues]

Bamparopoulos, G., Konstantinidis, E., Bratsas, C., & Bamidis, P. D. (2016). Towards exergaming commons: composing the exergame ontology for publishing open game data. *Journal of Biomedical Semantics*, *7*: 4. [Online: doi.org/10.1186/s13326-016-0046-4]

Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., & Viboud, C. (2016). Big data for infectious disease surveillance and modeling. *The Journal of Infectious Diseases*, *214*(S4), S375–S379. [Online: doi.org/10.1093/infdis/jiw400]

Bartha, P. (2013). Analogy and analogical reasoning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. [Online: plato.stanford.edu/entries/reasoning-analogy/]

Beatty, B. (2009). Transitory connections: The reception and rejection of Jean Piaget's psychology in the Nursery School Movement in the 1920s and 1930s. *History of Education Quarterly*, *49*(4), 442–464.

Bellan, S. E., Pulliam, J. R., Scott, J. C., Dushoff, J., & the MMED Organizing Committee. (2012). How to make epidemiological training infectious. *PLOS Biology*, *10*(4): e1001295. [Online: doi.org/10.1371/journal.pbio.1001295]

Bellazzi, R. (2014). Big data and biomedical informatics: A challenging opportunity. *Yearbook of Medical Informatics*, *9*, 8–13.

Bennett, C. (2016). The pluses and minus of teaching biostatistics and quantitative epidemiology. *Australasian Epidemiologist*, *23*(1), 4.

Bhopal, R. S. (2016). *Concepts of epidemiology: Integrating the ideas, theories, principles, and methods of epidemiology*. Oxford: Oxford University Press.

Binani S., Gutti A., Upadhyay S. (2016) SQL vs. NoSQL vs. NewSQL – A Comparative Study. *Communications on Applied Electronics, 6*(1), 43–46. [Online: goo.gl/gSyvWz]

Bodner, G. M. (1986). Constructivism: A theory of knowledge. *Journal of Chemical Education*, *63*(10), 873–878. [Online: pubs.acs.org/doi/pdf/10.1021/ed063p873]

Bonell, C., Moore, G., Warren, E., & Moore, L. (2018). Are randomised controlled trials positivist? Reviewing the social science and philosophy literature to assess positivist tendencies of trials of social interventions in public health and health services. *Trials*, *19*:, 238.

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., & Schneeweiss, S. (2010). Confounding control in healthcare database research: Challenges and potential approaches. *Medical Care*, *48*(6S), S114–S120.

Butler, D. (2013). When Google got flu wrong. *Nature*, *494*(7436), 155–156. [Online: doi.org/10.1038/494155a]

Cavanagh, M., & Lane, D. (2012). Coaching psychology coming of age: The challenges we face in the messy world of complexity. *International Coaching Psychology Review*, *7*(1), 75–90.

CDC (2006). Principles of epidemiology in public health practice: An introduction to applied epidemiology and biostatistics. Centers for Disease Control and Prevention. [Online: www.cdc.gov/csels/dsepd/ss1978/index.html]

Chang, A. C. (2016). Big data in medicine: The upcoming artificial intelligence. *Progress in Pediatric Cardiology*, *43*, 91–94.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Chauhan, R., & Kaur, H. (2015). A spectrum of big data applications for data analytics. In D. P. Acharjya, S. Dehuri, & S. Sanyal (Eds.), *Computational Intelligence for Big Data Analysis* (pp. 165–179). Cham, Switzerland: Springer International.

Chen, Y., Argentinis, J. E., & Weber, G. (2016). IBM Watson: How cognitive computing can be applied to Big Data challenges in life sciences research. *Clinical Therapeutics*, *38*(4), 688–701.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from Big Data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Christensen, M. L., & Davis, R. L. (2018). Identifying the "Blip on the radar screen": Leveraging Big Data in defining drug safety and efficacy in pediatric practice. *The Journal of Clinical Pharmacology*, *58*(S10), S86–S93.

Church, A. H., & Dutta, S. (2013). The promise of Big Data for OD: Old wine in new bottles or the next generation of data-driven methods for change. *OD Practitioner*, *45*(4), 23–31.

Cobb, P. (1996). Where is the mind? A coordination of sociocultural and cognitive constructivist perspectives. In C. T. Fosnot (Ed.), Constructivism: Theory, perspectives, and practice (pp. 34–53). New York: Teachers College Press.

de Koning, E., Sijtsma, K., & Hamers, J. H. (2003). Construction and validation of test for inductive reasoning. *European Journal of Psychological Assessment*, *19*(1), 24–39.

Del Mar, C., Glasziou, P., & Mayer, D. (2004). Teaching evidence based medicine: Should be integrated into current clinical scenarios. *BMJ: British Medical Journal*, *329*(7473), 989–990.

DeMets, D. L., Stormo, G., Boehnke, M., Louis, T. A., Taylor, J., & Dixon, D. (2006). Training of the next generation of biostatisticians: A call to action in the U.S. *Statistics in Medicine*, *25*(20), 3415–3429.

Dennick, R. (2016). Constructivism: Reflections on twenty five years teaching the constructivist approach in medical education. *International Journal of Medical Education*, *7*, 200–205.

Detels, R., Beaglehole, R., Lansang, M. A., & Gulliford, M. (2011). *Oxford textbook of public health*. Oxford: Oxford University Press.

Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience*, *5*(1):12. [Online: doi.org/10.1186/s13742-016-0117-6]

Dolley, S. (2018). Big data's role in precision public health. *Frontiers in public health, 6*, 68. [Online: doi.org/10.3389/fpubh.2018.00068]

Ehrenstein, V., Nielsen, H., Pedersen, A. B., Johnsen, S. P., & Pedersen, L. (2017). Clinical epidemiology in the era of big data: New opportunities, familiar challenges. *Clinical Epidemiology*, *9*, 245–250.

Evans, J. (1993). The Cognitive Psychology of Reasoning: An Introduction. *The Quarterly Journal of Experimental Psychology Section A*, *46*(4), 561 567.
[Online: doi.org/10.1080/14640749308401027]

Fast, L., & Waugaman, A. (2016). Fighting Ebola with information: learning from data and information flows in the West Africa Ebola response. *USAID Report. Washington, DC.* [Online: www.usaid.gov/sites/default/files/documents/15396/FightingEbolaWithInformation.pdf]

Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: How data integration has changed causal inference in molecular epidemiology. *Emerging Themes in Epidemiology*, *12*: 14.
[Online: doi.org/10.1186/s12982-015-0037-4]

Feeney, A., & Heit, E. (Eds.). (2007). *Inductive reasoning: Experimental, developmental, and computational approaches*. New York: Cambridge University Press.

Ferrer, E., O'Hare, E. D., & Bunge, S. A. (2009). Fluid reasoning and the developing brain. *Frontiers in Neuroscience*, *3*(1), 46–51.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*(1), 39–82.

Fosnot, C. T. (2013). Constructivism: Theory, perspectives, and practice. Teachers College Press.

Galea, S., Riddle, M., & Kaplan, G. A. (2009). Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology*, *39*(1), 97–106.

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, *2*(43), 67–78.

Engel, G. L. (1980). The clinical application of the biopsychosocial model. *The American Journal of Psychiatry, 137*(5), 535–544.

Gouda, H. N., & Powles, J. W. (2014). The science of epidemiology and the methods needed for public health assessments: A review of epidemiology textbooks. *BMC Public Health*, *14*(1): 139.

Green, M. (1998). Toward a perceptual science of multidimensional data visualization: Bertin and beyond. Unpublished paper.

Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: A movement in crisis? *The BMJ, 348*: g3725.

Hammond, K. R. (1981). *Principles of organization in intuitive and analytical cognition* (No. CRJP-231). Colorado University at Boulder: Center for Research on Judgment and Policy.

Hammond, K. R. (2000). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York; Oxford University Press.

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19.
[Online: rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2014.00778.x]

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *WIREs Cognitive Science*, *9*(3): e1459. [Online: doi.org/10.1002/wcs.1459]

Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *WIREs Cognitive Science*, *1*(2), 278–292. [Online: doi.org/10.1002/wcs.44]

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition, 37*(6), 730–743.

Hayes, B. K., Ruthven, C., & Newell, B. R. (2007). Inferring properties when categorization is uncertain: A feature-conjunction account. In *Proceedings of the Annual Meeting of the Cognitive Science Society,* 29 (pp. 1073–1078). Cognitive Science Society. [Online: escholarship.org/uc/cognitivesciencesociety/29/29]

Hayes, B. K., Stephens, R. G., Ngo, J., & Dunn, J. C. (2018). The dimensionality of reasoning: Inductive and deductive inference can be explained by a single process. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(9), 1333–1351.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*(5), 295–300. [Online: www.ncbi.nlm.nih.gov/pmc/issues/145128/]

Holmes, D. E., & Jain, L. C. (2018). Advances in Biomedical Informatics: An Introduction. In D. E. Holmes, E. Dawn, & L. C. Jain (Eds.), *Advances in biomedical informatics* (pp. 1–4). Cham, Switzerland: Springer International.

Imenda, S. (2014). Is there a conceptual difference between theoretical and conceptual frameworks? *Journal of Social Sciences*, *38*(2), 185–195.

Jain, V. K. (2017). *Big Data and Hadoop*. New Delhi: Khanna Book Publishing.

Jenkins, T. (2013). Don't count on big data for answers. *The Scotsman*, *12 February*. [Online; www.scotsman.com/news/opinion/tiffany-jenkins-don-t-count-on-big-data-for-answers-1-2785890]

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology*, *89*, 39–70.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*(43), 18243–18250.

Jones, L. V. (Ed.) (1987). *The collected works of John W. Tukey. Philosophy and principles of data analysis 1965–1986, Vol. 4*. Monterey, CA: Wadsworth & Brooks/Cole.

Jordan, L. (2015). The problem with Big Data in translational medicine. A review of where we've been and the possibilities ahead. *Applied & Translational Genomics*, *6*, 3–6.

Karagiorgi, Y., & Symeou, L. (2005). Translating constructivism into instructional design: Potential and limitations. *Educational Technology & Society*, *8*(1), 17–27.

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education, 3*: 11.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, *59*(7), 613–620.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1). [Online: doi.org/10.1177/2053951714528481]

Klauer, K. J. (1996). Teaching inductive reasoning: Some theory and three experimental studies. *Learning and Instruction*, *6*(1), 37–57.

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, *78*(1), 85–123.

Koutkias, V., & Thiessard, F. (2014). Big data-smart health strategies. Findings from the yearbook 2014 special theme. *Yearbook of Medical Informatics*, *9*(1), 48–51.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*, *33*(7), 1163–1170.

Kulikowski, C. A., Shortliffe, E. H., Currie, L. M., Elkin, P. L., Hunter, L. E., Johnson, T. R., ... & Smith, J. W. (2012). AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *Journal of the American Medical Informatics Association*, *19*(6), 931–938.

Kyriacou, D. N. (2004). Evidence-based medical decision making: Deductive versus inductive logical thinking. *Academic Emergency Medicine*, *11*(6), 670–671.

Lamontagne, C. (2003). In search of M.C. Escher's metaphysical unconscious. In D. Schattschneider, & M. Emmer (Eds.), *M.C. Escher's Legacy* (pp. 69–82). Berlin, Heidelberg: Springer.

Landes, J., Osimani, B., & Poellinger, R. (2018). Epistemology of causal inference in pharmacology. *European Journal for Philosophy of Science*, *8*(1), 3–49.

Lane, D. A., & Corrie, S. (2006). *The modern scientist-practitioner: A guide to practice in psychology*. New York: Routledge.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *META Group Research Note*, *6*(70), 1.

Last, J. M. (Ed.). (2001). *A dictionary of epidemiology, 4th ed.*. New York: Oxford University Press.

Lawrence, G. (2016). An integrated approach to teaching introductory epidemiology and biostatistics to public health students. *Australasian Epidemiologist*, *23*(1), 20.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in Big Data analysis. *Science*, *343*(6176), 1203–1205.

Lee, C. H., & Yoon, H. J. (2017). Medical big data: Promise and challenges. *Kidney Research and Clinical Practice*, *36*(1), 3–11.

Lefèvre, T. (2018). Big data in forensic science and medicine. *Journal of Forensic and Legal Medicine, 57*, 1–6.

Lehane, E., Leahy-Warren, P., O'Riordan, C., Savage, E., Drennan, J., O'Tuathaigh, C., ... & Hegarty, J. (2019). Evidence-based practice education for healthcare professions: An expert view. *BMJ Evidence-Based Medicine, 24*(3), 103–108.
[Online: dx.doi.org/10.1136/bmjebm-2018-111019]

Lipton, R., & Ødegaard, T. (2005). Causal thinking and causal language in epidemiology: It's in the details. *Epidemiologic Perspectives & Innovations*, *2*(1): 8.

Luttenberger, S., Wimmer, S., & Paechter, M. (2018). Spotlight on math anxiety. *Psychology research and behavior management, 11*, 311. [Online: doi.org/10.2147/PRBM.S141421]

Malle, J. P. (2013). Big data: Farewell to Cartesian thinking. *Paris Innovation Review*, *15*.
[Online: parisinnovationreview.com/articles-en/big-data-farewell-to-cartesian-thinking]

Mayer, M. A. (2015). Expectations and pitfalls of big data in biomedicine. *Review of Public Administration and Management*, *3*(1).

McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review*, *27*(2), 277–290.

McAfee, A., & Brynjolfsson, E. (2012): Big data: The management revolution. *Harvard Business Review, 90*(10), 60–68. [Online: hbr.org/2012/09/big-datas-management-revolutio]

McCue, M. E., & McCoy, A. M. (2017). The scope of big data in one medicine: Unprecedented opportunities and challenges. *Frontiers in Veterinary Science*, *4*, 194–%%.

Meza, J. P., & Yee, N. (2018). Clinical practice can save evidence-based medicine from itself. *Clinical Research in Practice: The Journal of Team Hippocrates*, *4*(2): eP1833.

Mitchell, T. M. (1997). *Machine learning*. Boston: McGraw-Hill.

Monteith, S., & Glenn, T. (2016). Automated decision-making and big data: Concerns for people with mental illness. *Current Psychiatry Reports*, *18*: 112.

Nesbitt, K. V., & Friedrich, C. (2002). Applying Gestalt principles to animated visualizations of network data. In *Proceedings of the Sixth International Conference on Information Visualisation* (pp. 737–743). IEEE Digital Library.
[Online: doi.org/10.1109/IV.2002.1028859]

NIH (2018). *The NIH Catalyst. National Institutes of Health* (NIH). [Online: irp.nih.gov/catalyst/v27i6]

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.

Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: Challenges and research agenda. *Journal of Big Data*, *2*: 22.

Panda, B. (2016). *Big data for a big country: Mind the gap*. [Online: doi.org/10.1038/nindia.2016.71]

Phillips, C. V., & Goodman, K. J. (2004). The missed lessons of Sir Austin Bradford Hill. *Epidemiologic Perspectives & Innovations*, *1*: 3. [Online: www.ncbi.nlm.nih.gov/pmc/articles/PMC524370/]

Pinna, B. (2010). New Gestalt principles of perceptual organization: An extension from grouping to shape and meaning. *Gestalt Theory*, *32*(1), 11–78.

Porta, M. (Ed.). (2014). *A dictionary of epidemiology, 6th ed*. New York: Oxford University Press.

Prince, M. J., & Felder, R. M. (2006). Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education*, *95*(2), 123–138.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review, 84*(3), 528–549. [Online: doi.org/10.1111/insr.12110]

Rock, I., & Palmer, S. (1990). The legacy of Gestalt psychology. *Scientific American*, *263*(6), 84–90.

Rothman, K. J., & Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health*, *95*(S1), S144–S150.

Russom, P. (2011). Big data analytics. *TDWI Best Practices Report*, *19*(4), 1–34.

Sackett, D. L. (1997, February). Evidence-based medicine. *Seminars in Perinatology,* 21(1), 3–5).

Sacristán, J. A., & Dilla, T. (2015). No big data without small data: Learning health care systems begin and end with the individual patient. *Journal of Evaluation in Clinical Practice*, *21*(6), 1014–1017.

Savery, J. R., & Duffy, T. M. (1996). Problem-based learning: An instructional model and its constructivist framework. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies – Instructional design* (pp. 135–150). New Jersey: Educational Technology Publications.

Simpson, J. M. (1995). Teaching statistics to non-specialists. *Statistics in Medicine*, *14*(2), 199–208.

Singh, R., Yang, H., Dalziel, B., Asarnow, D., Murad, W., Foote, D., … Fisher, S. (2013). Towards human-computer synergetic analysis of large-scale biological data. *BMC Bioinformatics, 14*: S10. [Online: doi.org/10.1186/1471-2105-14-S14-S10]

Sjøberg S. (2010). Constructivism and learning. In P. Peterson, E. Baker, & B. McGaw, (Eds.), *International encyclopedia of education* (pp. 485–490). Oxford: Elsevier.

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, *112*(1), 80–116.

Stevenson, C. (2016). Reshaping biostatistics 1 – Moving from the 20th to the 21st century: A case study in teaching biostatistics in the Deakin master of public health. *Australasian Epidemiologist*, *23*(1), 8–15.

Stricker B. H. (2017). Epidemiology and ‘big data’. *European Journal of Epidemiology*, *32*(7), 535–536.

Stroup, D. F., Goodman, R. A., Cordell, R., & Scheaffer, R. (2004). Teaching statistical principles using epidemiology: Measuring the health of populations. *The American Statistician, 58*(1), 77–84.

Tanenbaum, S. J. (2005). Evidence-based practice as mental health policy: Three controversies and a caveat. *Health Affairs*, *24*(1), 163–173.

Taylor, L., & Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. *Geo Journal*, *80*(4), 503–518.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney, & E. Heit (Eds.), *Inductive Reasoning. Experimental, Developmental, and Computational Approaches*. New York: Cambridge University Press.

Tobias, S. (2010). Generative learning theory, paradigm shifts, and constructivism in educational psychology: A tribute to Merl Wittrock. *Educational Psychologist*, *45*(1), 51–54.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525–547.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. In: Wendt, D., & Vlek, C. (Eds.), *Utility, probability, and human decision making* (pp. 141–162). Dordrecht: Reidel.

Vogt, W. P., Gardner, D. C., & Haeffele, L. M. (2012). *When to use what research design*. New York: Guilford Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Ed. by M. Cole, V. John-Steiner, S. Scribner, E. Souberman). Cambridge, MA: Harvard University Presse.

Walesby, K. E., Harrison, J. K., & Russ, T. C. (2017). What big data could achieve in Scotland. *Journal of the Royal College of Physicians of Edinburgh*, *47*(2), 114–119.

Weissmann, G. (2010). Pattern recognition and Gestalt psychology: The Day Nüsslein-Volhard Shouted “Toll!”. *The FASEB Journal*, *24*(7), 2137–2141.

West, C. P., & Ficalora, R. D. (2007). Clinician attitudes toward biostatistics. *Mayo Clinic Proceedings, 82*(8), 939–943).

Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of Big Data. *Journal of Learning Analytics, 2*(2), 5–13.

Wittrock, M. C. (1992). Generative learning processes of the brain. *Educational Psychologist*, *27*(4), 531–541.

Zou, X., Zhu, W., Yang, L., & Shu, Y. (2015). Google Flu Trends – The initial application of big data in public health. *Zhonghua Yu Fang Yi Xue Za Zhi [Chinese Journal of Preventive Medicine]*, *49*(6), 581–584. (Article in Chinese).

ROSSI A. HASSAD
Mercy College
School of Social & Behavioral Sciences
555 Broadway, Dobbs Ferry
New York, USA