

Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Bridging BigData and sampling methodology: What is big and where is the bridge?

by Fulvia Mecatti

Release date: June 30, 2025



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “Contact us” > “[Standards of service to the public](#).”

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2025

Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

An [HTML version](#) is also available.

Cette publication est aussi disponible en français.

Bridging BigData and sampling methodology: What is big and where is the bridge?

Fulvia Mecatti¹

Abstract

BigData users and the BigData research community are expanding rapidly, while statisticians at large are seemingly becoming divided between those who are enthusiastic and those who are concerned, if not downright hostile. Is BigData also a big step ahead, truly advancing our ability to extract meaningful information and actual knowledge from data? Is BigData underplaying traditional statistical inference as we know it, supplanting survey methodology as a low-cost futuristic option? In this paper I will attempt to unravel the multifaceted relationship bridging BigData to sampling methodology. Starting by reasoning why it should be interesting to look at BigData from a sampling statistician's perspective, I will delve deeper into the somewhat ambiguous definition of BigData and share some very personal considerations and views on the matter. In the process, several open questions will arise while discussing a personal selection of insights that are traceable through the vast body of statistical literature around BigData and sampling methodology. The discussion will take various angles explored across nine key points, and it will conclude with a forward-looking perspective on a main challenge for future research: addressing the strong assumptions needed to manage deviations from purely randomized data collection.

Key Words: Bayesian network models; Causal inference; Data quality; Digital data and sources; Non-probability samples; Observational data.

1. Introduction

I am grateful to Editor Jean-François Beaumont for his kind invitation to join the celebration of the 50th anniversary of Survey Methodology, and very glad to contribute to this special issue. Indeed this paper was born thanks to a number of kind invitations. The International Association of Survey Statistician (IASS) initiated this by inviting me to join their webinar series (Mecatti, 2022) and then to give the IASS President's invited paper at the 2023 World Statistics Congress of the International Statistical Institute. The IASS also suggested the intriguing subject: BigData and Sampling Methodology. While I would not consider myself an expert on the topic, I was intrigued, and compelled to dive deeper as a result, which led to the perspective I am presenting here.

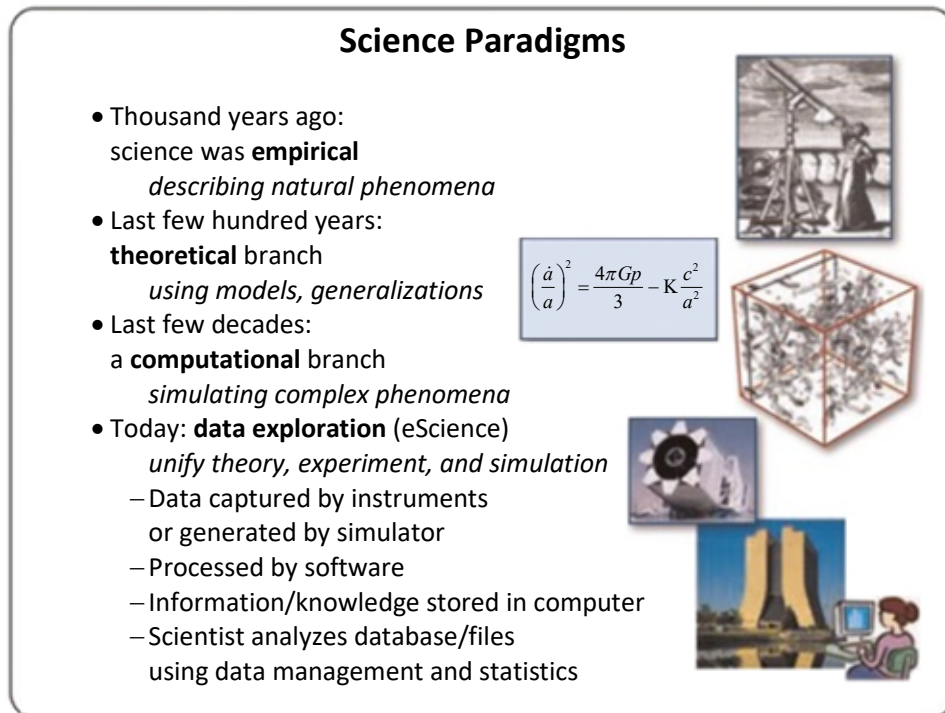
To discuss the *What* and the *Where* in the title, I will start with focusing on the *Why*. Why should it be interesting to elaborate on the bigness of BigData from a sampling statistician's perspective?

For some 80 years, sampling theory and methods have provided a successful, reliable standard for creating new knowledge from sample data. However, we are now experiencing a paradigm shift in data collection, data management, statistical analysis, and inference, driven by digitisation and the explosion of the Internet. Some have even referred to this shift as the rise of a new scientific paradigm. Figure 1.1 comes from a 2009 paper authored by Information Technology (IT) experts from Microsoft (Hansen, Johnson, Pascucci and Silva, 2009) where a new (fourth) science paradigm is theorized, named eScience, and characterised by data exploration as the new scientific method. The term Data Science was still to come at

1. Fulvia Mecatti, Professor of Statistics, University of Milano-Bicocca, Dpt of Sociology & Social research, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy. E-mail: fulvia.mecatti@unimib.it.

the time, but the new scientific method cutting across computer science and statistics seems to be conjured in the last row of Figure 1.1. The same figure also appears in the 2015 report by AAPOR Task Force on BigData (AAPOR Task Force on Big Data, 2015), while discussing the new (Big) data sources as drivers behind the paradigm shift. Survey researchers and sampling statisticians have been observing signs of this shift for some time, including a steady decline in survey response rates, a corresponding increase in survey costs, due in part to the need to meet accelerated information needs, and the Internet's ubiquitous presence in all aspects of our daily lives. In response to these early clues, web surveys have emerged as a first, natural solution, producing extensive literature and have rapidly penetrated the practice (see among others Biffignandi and Bethlehem (2021) and Pfeffermann (2015) Section 6 for a discussion on the use of web surveys and web opt-in panels for the production of official statistics).

Figure 1.1 A paradigm shift



Examples and sources usually related to BigData include:

- web data from all the WWW (World Wide Web) traffic, e.g. all our social network commenting and blog posting;
- digital data automatically created by IT systems and software, e.g. business transactions, shipments, and supplies, digitised health records and human genome databases, financial data streamed by stock exchanges, administrative electronic registries;

- data captured by sensors, e.g. temperature measures in industrial processes, customer access to retail facilities, and rainfall/pollution levels;
- videos, images and audios, e.g. from surveillance cameras all around our cities, postings on Instagram and YouTube;
- click streams, e.g. on e-commerce platforms and the IoT (Internet of things);
- log-files and search queries, e.g. all our Google-ing;
- satellite imagery and geo-localised data, e.g. traffic and weather data from smart meters, signals from our mobile phones;

and this is just to mention a few, since digital data are continuously generated by human activities, by devices and by person-device interactions. The BigData era is, in fact, a tempting abundance of datasets gathered passively, *found data* as they are also called to indicate that they are *already there* as opposed to actively (and expensively) collected, self-creating and self-updating at great speed and mostly free or at much lower costs than in a primary data collection, readily (or at least conveying the impression to be) accessible from our own computers and cell phones. Thus, my question is: why should we not want to use it not only to answer daily-life questions, to network and to connect but also to deal with issues that challenge sampling theory and practice, and to develop innovative statistical methods that can be applied to expanding and emerging domains. On these premises, in the rest of this paper I will share some insights I have gained from my personal journey through the wide-ranging literature on the multifaceted relationship between BigData and sampling methodology. My reasoning has taken various angles, which I will briefly explore in the following 9 discussion points, each with a subheading that conveys its unique perspective as outlined here below.

Is there a threat? My first point addresses the perceived threat to survey sampling that many researchers and practitioners might associate with BigData. I will provide possible reasons behind this fear, arguing that rather than a threat it should be seen as a call to action for sampling statisticians.

What is Big in BigData? Building on that, in my second point, I wonder what is big in BigData, which first requires an attempt to define it. This leads me to discuss a popular and ever-growing list of 18 V-words (words beginning with the letter V) used to describe the different characteristics of BigData.

A personal, limited taxonomy. In my third point, I offer a personal and limited taxonomy of this abundance of Vs to define BigData. This taxonomy shifts the focus toward the many opportunities that BigData presents to sampling statisticians. I argue that these opportunities can be understood in two main directions. First, I consider how statistical reasoning and sampling principles can be applied to address the weaknesses and risks associated with BigData. This brings me to my fourth point.

Machine learning versus sampling methodology and practice. Here, I briefly discuss the evolution of machine learning techniques typically associated with BigData. The second direction focuses on how BigData can be leveraged to improve sampling methods and practices. This forms the focus of the latter part of the paper, where I explore these opportunities in 4 progressive steps.

Not just about size: balancing quantity and quality. First, I revisit the crucial need to balance quality and quantity, particularly important in the context of BigData, which I discuss in this fifth point.

The BigData paradox. Next, this balance is examined in more formal terms from a sampling statistics perspective in this sixth point, where I introduce what is known in the literature as the BigData paradox, i.e. a formal framework for understanding what can happen when quantity prevails over quality.

BigData versus non-probability samples. This sampling perspective on the quality of BigData leads me naturally to the current renewed attention within the sampling community on non-probability samples, which I address in this seventh point.

Propensity score versus participation probability. The practical relevance of propensity score-based methods to address biases inherent in non-probability samples triggers the reflection discussed in my eighth point. I argue around the similarities in the rising costs and other practical issues affecting both sample surveys and randomized biomedical studies, which in the context of BigData appear to align practice and research in causal inference from observational data and in valid inference from non-probability sampling.

Bayesian Network: A unifying and promising tool for inference and prediction from BigData. After navigating through these varied angles and ideas, I try to wrap up in my final point with a brief exploration of Bayesian Networks as a promising and unifying tool for both inference and prediction in the context of BigData.

The paper concludes by offering a look ahead, focusing on what I believe is one of the main challenges for future research in the fascinating area of BigData from a sampling statistician's perspective.

2. Is there a threat?

There was a time when, as statisticians, we might have felt threatened by the impact of data abundance and the seemingly easy access for everyone. We might have feared that our science and competence were endangered by this excitement. We could even see a potential menace for scientists in general, as illustrated by the following two, emblematic examples. In 2008 the popular technology and lifestyle magazine WIRED published an article titled “*The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*” (Anderson, 2008). The article apparently introduced the term *Data Deluge* for the first time, while it also prophesized the obsolescence of the scientific method as we knew it. The following quotes from that paper, though in need of being updated from the year of the paper's publication to the present day, make points that align closely with the objectives of the present paper:

60 years ago, digital computers made information readable. 20 years ago, the Internet made it reachable. 10 years ago, the first search engine crawlers made it a single database [...] measured in PetaByte [i.e. 1 million GigaByte] [...] stored in cloud. [...] faced with massive data, [the] approach to science - hypothesis, model, test - is becoming obsolete.

The second example is from a 2014 paper that discusses BigData as *A revolution that will transform how we live, work, think* (Mayer-Schönberger and Cukier, 2013). The authors are two data professionals who, at that time, were respectively a professor of Internet governance and regulation at Oxford University and a data editor at the UK magazine *The Economist*. The following is an illustrative quote from that paper:

Reaching for a random sample in the age of big data is like clutching at a horse whip in the era of the motor car.

This is, in fact, a longstanding process: at the dawn of the new millennium, a forerunner paper by Bellhouse (2000) was warning the statistical community:

Although theoretical developments in sampling theory have often run ahead of computational capabilities, it is now the case that survey statisticians are now followers of computing technology that has been motivated by others instead of acting as the catalyst that leads to technological change.

It has been on such grounds that sampling statisticians have been pondering whether BigData might overshadow survey sampling theory and methodology, emerging as a low-cost futuristic option. Several examples of similar questioning can be traced in the literature. Two main examples are the 2013 paper by Couper (2013) from the perspective of a survey researcher and the 2020 paper by Beaumont (2020) from the perspective of a statistician working in official statistics. Since the papers' title, the former wonders *Is the Sky Falling? New technology, Changing media, and the Future of surveys*, and the second inquires *Are probability surveys bound to disappear for the production of official statistics?* Couper's paper explores various limitations of BigData, including the potential for mischief. Nonetheless, he concludes by asserting that *Big data are here to stay*, and advocates for welcoming these new developments rather than fearing or opposing them. Beaumont's analysis starts by recognising five key factors that make the question posed in his title relevant for national statistical offices. Besides the decline in survey response rates and the increasing cost and burden of data collection, key factors include timeliness and the proliferation of non-probability data sources, such as BigData. According to Beaumont, these factors underscore why:

[...] a wind of change has been blowing over national statistical agencies, and other data sources [than standard probability surveys] are being increasingly explored.

Along this line of reasoning, a need to reposition seems to have emerged. Two further significant examples from the literature are worth recalling here. The 2017 IASS presidential invited paper by Rao and Fuller (2017) aimed to stimulate a discussion on the future. Here is a quote that makes a direct reference to BigData:

Future directions in research and methods will be influenced by [...] improved data collection devices, and availability of auxiliary data, some of which will come from Big Data. Survey taking will be impacted by changing cultural behaviour and by a changing physical-technical environment.

The second example is the (latest version at the time of writing this paper) 2022 paper *Positioning Household Surveys for the Next Decade* by the UN Statistical Division Inter-Secretariat Working Group on Household Survey (Carletto, Chen, Kilic and Perucci, 2022). It highlights that

household surveys are facing funding challenges and skepticism on their continued utility within the changing data landscape,

and asks

How can we establish sustainable household survey programs that are resilient and versatile to future shocks like COVID-19?

Indeed on one hand, for a large part of the world, surveys are crucial to strengthen and improve the digitisation of their national statistical offices (and health surveillance systems). The same paper reports that a mapping exercise found that approximately one-third of the global indicators in the UN Agenda 2030 of 17 Sustainable Development Goals, can be sourced from household surveys - specifically, 80 out of 232 indicators covering 13 different goals in the Agenda.

On the other hand, online tools for in-house surveys, e.g. SurveyMonkey and Qualtrics, seem to be flourishing, with millions of surveys completed every month and major companies in their portfolio. Related to this fact, the following quote from Marker (2017) sounds thought-provoking:

[...] the existence of big data has changed the expectation of timeliness, and national statistical offices will need to figure out how to carry out surveys and censuses quicker, or users will rely on available big data without understanding what they are losing.

This quick journey through the BigData scene leads to a first, personal conclusion. It might not suggest an imminent end to survey sampling, or the scientific method as we know them. Instead, it calls for action from statisticians, and from sampling statisticians especially. Here is a final example, from a general public use book titled *DATA-ISM: the revolution transforming decision making, consumer behaviour and almost everything else* authored by a technology reporter (at the time of its publishing in 2015) for *The New York Times* (Lohr, 2015). As statisticians, we may easily agree with the neologism, which effectively captures a world increasingly shaped by data, as well as its transformative effects. Still, a quick look at the book's "S" section of the subject index would reveal that some words are missing, and in particular words such as statistics, sampling, and survey. A surprising oversight for a book on data.

3. What is Big in BigData?

While the term BigData is certainly a household buzzword, what BigData actually is does not seem to be an easy question to answer, at least by a definition precise enough to satisfy a (sampling) statistician. Survey literature offers some alternative terms to BigData. Groves (2011) coined the term *organic* data to

describe data automatically created in the digital ecosystem, as opposed to *designed* data produced through quantitative studies and official statistics for a precise purpose and well-designed use. Another example is the previously mentioned *found* data, which emphasizes the benefit of not requiring actual collection efforts. Other sources highlight the multi-faceted nature of the term BigData and its pervasive use. According to Wikipedia (accessed October 2022), BigData is both a *blanket term* and *an obsession with entrepreneurs, consultants, scientists, and the media*. Official statisticians refer to it as *a collective term for the increasingly diverse range of data sources available through the web of everything* (Tam and Clarke, 2015). A 2019 post on the blog of the Berkeley School of Information (2019) (accessed May 2024) states:

Today, the concept of big data is not only less compelling but also potentially misleading.

Hence, the personal impression is that BigData is (still) a blurry definition of a broad phenomenon that is not simply a matter of size. The vagueness of the term can create misunderstandings and hamper discussion. The unprecedented scale is just one of several characteristics that can be observed in many different fields of knowledge and tends to be context-specific.

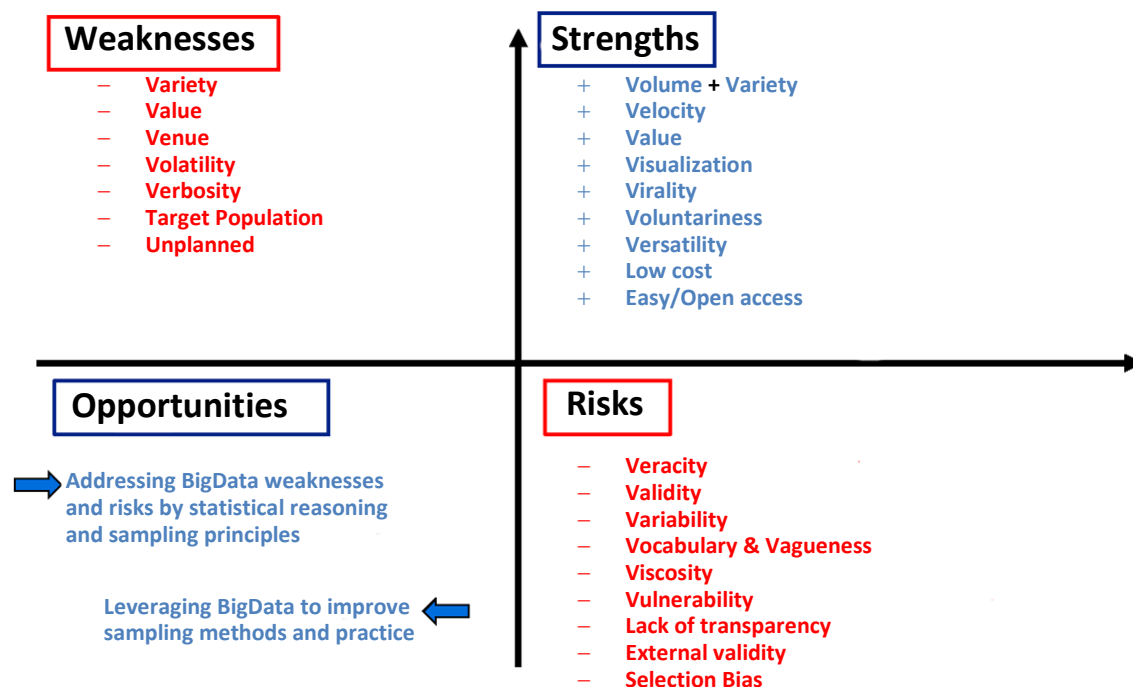
A popular way to define BigData is through the description of its inner characteristics, usually synthetically expressed by V-words. However, according to a personal limited web-scraping, it appears there is little agreement about how many V-words are needed to effectively describe what BigData is. It started in 2001, with the classical trio *Volume, Velocity, and Variety*, first introduced by the industry analyst D. Laney, that catches the amount, speed, and complexity of BigData. Quickly the trio became a 5 Vs-definition with the addition of *Value* and *Veracity*, allegedly by Oracle and SAS Institute. In 2014 a 10 Vs-definition appeared, due to K. Borne from Data Science Center, with the addition of *Validity, Variability, Venue, Vocabulary* and *Vagueness*, the last two being a direct reference to the domain-specific and the pluri-sidedness of the BigData notion. A 2017 paper by Panimalar, Shree and Kathrine (2017) reports a 14 Vs-definition that includes *Volatility, Visualization, Virality, and Viscosity*, and then stretches it up to a 17 Vs-definition by supplementing *Verbosity, Voluntariness, and Versatility*. Moreover, at some point in between, the addition of *Vulnerability* also happened (insideBIGDATA, 2019), which leads us to a grand total of 18 Vs in trying to inform an actual definition of BigData. With the purpose of making sense of such an abundance of Vs, a personal classification is offered in the next section. However, for possibly interested readers it is worth mentioning that the 18 Vs-definition I am considering here is neither comprehensive nor the end, for as much as 42 Vs can be tracked online by pairing BigData with Data Science (Shafer, 2017).

4. A personal, limited taxonomy

The personal taxonomy I am now proposing is limited to classifying the 18 Vs of BigData along 4 dimensions: 1) weaknesses, 2) strengths, 3) risks, and 4) opportunities. Of course other more complex organisational structures can be considered to map the 18 Vs, as for the *Framework for Data Quality* recently released by the Federal Committee on Statistical Methodology (FCSM-20-04, 2020). This framework

comprises 3 main domains (Utility, Objectivity, Integrity) and, within them, a total of 11 dimensions. It is comprehensive enough to incorporate the Total Error Framework for BigData, which extends the popular Total Survey Error approach (Amaya, Biemer and Kinyon, 2020), making it readily interpretable from the sampling statistician's perspective and certainly useful for identifying BigData's weaknesses, strengths, risks, and opportunities. However, this level of detail seems outside the purposes of this paper. For simplicity, the basic 4 dimensions have allowed me to compile the categories shown in Figure 4.1, enriched with some personal additions more directly related to statistics and sampling statistics, though this violates the V-wording rule.

Figure 4.1 A personal taxonomy of 18 Vs of BigData, and something more



Abundance, fast-generating flow, timeliness, and real-time availability seem to be the most relished features of BigData since the appearance of the term, thus *Volume* and *Velocity* quite naturally qualify as strengths. *Variety*, on the other hand, is typically understood as the diversity of BigData, which may not only be structured, such as Excel files from IT systems, but often as semi-structured and unstructured data. These may require pre-analysis steps for extracting information, especially when they come in a mixture of these formats. For instance, texts captured on social networks and e-commerce platforms usually demand implementing such preparatory steps to render BigData usable, which can be complex, costly, or impractical to implement, qualifying *Variety* as a weakness. Simultaneously, the *Variety* of online data, despite not being designed for statistical purposes, can be viewed as providing ample choices. Therefore, it can also be considered a strength of BigData. For instance, online data tends to offer greater detail, disaggregation, and richness in space and time compared to data collected through surveys and other traditional offline methods,

A similar double-faced classification holds for *Value*. When trying to answer the question *is there (always) value in (all) BigData?*, *Value* can be cast both in the weaknesses quadrant and in the strengths one. For instance, *BigData - Big Noise* is a popular sentence (e.g. Waldherr, Maier, Miltner and Günther (2017)) that points out the usually low signal-to-noise ratio. At the same time, a 2012 report from the World Economic Forum (2012) declared data *a new class of economic asset, like currency or gold*. *Veracity* appears as a downright risk, pointing to the truthfulness of data and the data-source, as efficiently summarised in the metaphor by Couper (2013), *Like good wine, the provenance of the data we analyse is important, as is quality*. None of the 5 more Vs that led to the 10 Vs-definition can add to the strengths quadrant. By referring to the tendency of BigData to be scattered across multiple platforms, *Venue* can be seen as a weakness, necessitating proper and efficient methods for data linkage and integration. The remaining 4 Vs all appear as risky characteristics, calling for increased knowledge and awareness, as well as statistical competencies. *Validity* relates to both the quality of data and to the appropriateness of the statistical analysis performed on it. *Variability* is a risk of BigData that reaches beyond statistical complexity, such as the presence of outliers, to include heterogeneity and inconsistencies across sources and time. Both *Vocabulary* and *Vagueness* pertain to the lack of a standard definition of BigData. The different context-specific interpretations of the concept entail the risk to create misunderstanding among users with different backgrounds and application interests. Going up to 17 Vs, *Volatility* and *Verbosity* both classify as weaknesses of BigData, the former related to the tendency of BigData to fall quickly out of date, the latter involving its widespread, multi-source nature that can cause overlaps, redundancy and ultimately a waste of computing resources. Four plain strengths follow. *Visualization* concerns the fast development of Infographic as a new technique prompted by the need to represent BigData. This is useful both in a first stage of the analysis to tackle the actual information it contains and its possible use, and in the final step to disseminate the knowledge extracted from it. *Virality* refers to the rapid dissemination of BigData over the Internet, *Voluntariness* to its availability to a range of different users, and *Versatility* to its flexibility across different contexts, domains and consumers. *Viscosity* relates to risks such as frictions in the data flow, for instance due to time lags that can lead to linkage errors. Finally, *Vulnerability* is a main and ever-growing risk inherent to BigData. It relates not only to the need to protect privacy and ensure confidentiality, but also to *a vision of technology that strengthens democracy* (Harris and Frueh, 2023). Indeed BigData is vulnerable to inappropriate use, both unintentional and intentional, which can lead to misleading conclusions, online and offline disinformation, manipulation, forgery and fakes - e.g. the 2010 Facebook-Cambridge Analytica scandal.

The remaining features displayed in Figure 4.1 are intended as either personal additives distinctly pertaining to sampling methods and practice, or re-interpretations, under this same approach, of some already mentioned V-words. Sampling statisticians would find weaknesses in that a foundational concept such as the target population does not always fit clearly into the BigData framework, both as the parent of data and as the destination of the inference. Similarly, the unplanned and essentially unintended BigData collection may poorly align with main research questions and variables of interest of our specific study. In contrast, the increase in low-cost and free BigData sources and its steady flow toward public platforms (e.g. UNData portal, GitHub) are plain strengths in the face of ever-growing issues of traditional surveys, i.e. tightening

budget, respondent burden and need to improve timeliness. Both these weaknesses and beneficial strengths reflect into risks that are evident and significant from the perspective of sampling statisticians. Despite its large to massive size, BigData still means partial coverage. BigData is in fact a sample, one for which the automated data generator or capturer mechanism, remains unknown and difficult to explain, leading to error rates that are widely or entirely out of control. This presents a constant potential for selectivity, introducing a selection bias that undermines the external validity of any generalisation derived from even the most highly sophisticated output generated by BigData. All sorts of non-sampling errors can also impact BigData, with under-coverage and measurement bias as the most discussed. A general lack of transparency is not limited to the phase of the BigData creation but also extends to many popular processing algorithms. Often, no context is provided to assess the so-called *black box effect*, where the internal decision-making processes of complex models remain opaque and difficult to interpret, as for some Machine Learning traditional techniques commonly associated with BigData. To conclude, here is a personal belief: it is in the last Opportunities quadrant in Figure 4.1 where, I believe, we should look to unfold what is big in BigData. All weaknesses and risks discussed above are in fact challenges that may require a fresh mindset for statisticians, who surely have a significant role to play in the new science paradigm. These opportunities may be seen as a two-way story. On one way, say W1, how can we address and improve upon all the BigData weaknesses and risks discussed above using statistical reasoning and sampling principles? On the other way, say W2, how can we enhance sampling methods and practices by leveraging the strengths of BigData highlighted above? In the remaining paragraphs I will present, with a personal angle, a selection of insights drawn from the extensive and rapidly expanding literature in both of these ways.

4.1 W1: Machine learning versus sampling methodology and practice

Data Science and Machine Learning are perhaps the most trendy terms associated with BigData, techniques naturally referred to for analysis and prediction upon it. Machine Learning methods and computational approaches to learning have been brewing for a while and have taken quite some time to develop into the techniques we utilise today and the meanings we currently associate with these terms. According to Wikipedia (accessed May 2024), the term Machine Learning first appeared back in 1959, coined by Arthur Samuel of IBM, who envisioned computers capable of self-teaching. Since then, the field has evolved beyond its initial pursuit of artificial intelligence, with a focus on machines that learn from data. The process began in the 1960s, primarily centered around pattern classification, gradually branching into pattern recognition in the 1970s, and then into probabilistic logic and reasoning in the 1980s, along with neural network training. The first International Workshop on Machine Learning convened in 1980, followed by the second in 1983, and the birth of the scientific journal *Machine Learning* by Springer in 1986. In the 1990s, Machine Learning intersected with statistics and the handling of large datasets, birthing what we used to call *data mining*. The 2000s brought a focus on *statistical learning* with prediction as the primary objective, marked by the publication of the first (2001) and second (2009) editions of the pivotal text by Hastie, Tibshirani and Friedman (2009). At this juncture between Machine Learning and statistical inferential goals, in my view, a dynamic emerged: a sort of friction between a statistical approach and a

purely computational approach. Yet, Machine Learning enthusiasm within the statistical community surged with the Data Deluge (see the previous section concerning the appearance of the term in 2008). While Wikipedia frames this evolution as a cultural shift, in my perception it remains rooted in statistical reasoning and inference. However, it certainly offers a new lens and enhanced computational support, paving the way for the line of research I have dubbed W1, which is well worth investigating. Breidt and Opsomer (2017) (and the references therein), and the recent contributions by Goga and Haziza, and Rueda and coauthors provide remarkable overviews in this area (Ferri-García and Rueda (2020); Castro-Martín, Rueda and Ferri-García (2020); Dagdoug, Goga and Haziza (2023b); Dagdoug, Goga and Haziza (2023a)).

4.2 W2: Not just about size: balancing quantity and quality

In the opposite way, to take advantage of BigData opportunities as discussed at the end of the taxonomy above, another significant driver is the tension between quality and quantity inherent in the inferential use of BigData. The statistical community has long known that *bad sampling cannot be cured by merely increasing the sample size* (Conti, 2022), perhaps since the iconic Literary Digest failure in the USA Presidential Elections in 1936, as well as in the many other examples of poll failures later on (see among other Elliot and Valliant (2017)). Nevertheless, in commercial and common use of BigData there is too often a naïve appeal to its sheer magnitude, a mis-interpreted law of large numbers that supposedly corrects all biases and erases all risks discussed above, which is unlikely to be justified. Instead, the selectivity of BigData sources is identified as a main issue in exploring the potential of BigData for the production of official statistics (see the 24th Morris Hansen Lecture by Pfeffermann (2015) and Comments, and Beresewicz, Lehtonen, Reis, Consiglio and Karlberg (2018)). Coping with sampling biases and validating inferences based on BigData are prominent challenges, in urgent need of innovative approaches to support the BigData era. The common misconception that bigger is always better oversimplifies the complex, multi-layered, and multi-purpose nature of BigData, and disregards the fundamental statistical principle that large sizes alone cannot correct selection bias and poor data quality (see among others Tam and Holmberg (2020)). However, it is apparent that the boom in new digital data sources has sparked, and is fuelling, an unstoppable reliance upon them by a wide variety of users, despite the bias potential and quality issues perhaps visible to only well-trained statistical eyes. Simultaneously, this boom fosters the fallacy that quantity outweighs quality in BigData, that it can correct any pitfalls and diminish the importance of quality. This underscores the classical imperative to balance quality and quantity and makes it even more crucial to fully harness the opportunities around BigData.

5. The BigData paradox

To strike a balance between BigData quantity and quality, we certainly need to innovate traditional statistical methods, but we also need to remain practical. Grounded in the power of sampling methodology, Meng (2018) offers both a fresh perspective and a practical framework toward the balancing goal. In fact, Meng sees more than opportunities created by BigData whilst warning us:

I see a paradise, or even paradises, gained if there is a sufficient number of us who can engage in what we have advertised to be the hallmark of our discipline, that is, principled thinking and methodology development for dealing with uncertainty.

Meng's reasoning starts by introducing the notion of an R -mechanism generating a data set, however big. The R -mechanism extends beyond the traditional conceptual framework where R represents Random (or *iid* (independent and identically distributed) or probabilistic with ignorable non-response). It also encompasses all the ways, whether innovative and associated with Big Data or simply more actionable, in which a data set may come available for practical use. Straightforward examples are: 1) all designed samples actually informed by those who chose to Respond, e.g. probabilistic with non ignorable non-response; 2) all self-Reported or self-selected samples, e.g. surveys launched over the Internet, volunteer opt-in panels; and 3) self-Recorded datasets, e.g. from sensors, from all our registering into webpages and posting on social networks. The R -mechanism can be easily formalised in a vector \mathbf{R} of membership indicators for every unit i included in the population that is the interest of our inferences,

$$\mathbf{R} = \left[R_i, i = 1, \dots, N; \sum_{i=1}^N R_i = n \right] \quad (5.1)$$

where, with usual notation, N and n respectively denote the population and the sample size, both however big, huge or massive. When R stands for Random, the vector \mathbf{R} has a known and well specified probability distribution, i.e. the sample design of a probability sample. Meng's paper offers a practical framework for assessing the uncertainty of inference using data generated under any R -mechanism. This framework is easily illustrated by focusing on the average \bar{y}_n of the (Big) data set as the familiar estimate of the mean (proportion) of a variable y of interest in a larger population (Beaumont and Rao, 2021)

$$\bar{y}_n = \frac{\sum_{i=1}^N R_i y_i}{\sum_{i=1}^N R_i}. \quad (5.2)$$

Meng's approach, in fact an updated version of the approach in Hartley and Ross (1954), considers usual uncertainty metrics, such as Bias and Mean Squared Error, as associated to the critical role of the BigData bias versus the BigData variance. The bias dominates the MSE. The Bias of the estimator \bar{y}_n can be expressed simply as a three-factors product

$$\text{Bias}(\bar{y}_n) = E_R(\rho_{Ry}) \times \sqrt{\frac{1-f}{f}} \times \sigma_y \quad (5.3)$$

with $f = n / N$. Alternatively, according to Meng's terminology

$$\text{Estimation (in)accuracy} = \text{data quality} \times \text{data quantity} \times \text{problem difficulty}.$$

The quite uncustomary factor in Meng's expressions above is the first factor, that catches the data quality and shows it depends on the correlation between the data generating mechanism and the study variable (a

version of the notion of non-informativeness of a sample design that will be discussed more in the following sections). In Meng's language ρ_{Ry} is dubbed data defect indicator, which equals 0 if the R -mechanism is Random, otherwise it gives the sign and the degree of the selection bias in the (Big) data set. Both the remaining middle and the right factors in (5.3) are the usual error components in finite population estimation. In Meng's language, the right factor σ_y catches the problem difficulty, while the middle factor $\sqrt{(1-f)/f} = \sqrt{N/n-1}$, also known as dropout odds, catches the dependence of estimation accuracy on the data quantity. On this basis, it is easy to prove (see also Tam and Kim (2018)) that in order not to incur a dramatically large selection bias, which can reduce the effective sample size of a massive dataset to the level of a relatively small probabilistic sample, the BigData should guarantee a population coverage unrealistically close to 100%. Meng's paper states:

Once we lose control of probabilistic sampling, then the driving force behind the estimation error is no longer the sample size n , but rather the population size N .

This leads Meng to the conclusion that compensating for quality with quantity is a doomed game, and ultimately to what Meng calls the *BigData Paradox: the bigger the data, the surer we fool ourselves*.

6. BigData versus non-probability samples

The potential of BigData to act as a booster of biases and an amplifier of quality issues is a primary reason behind the accelerated and intensified research on non-probability samples (NPS). These are not new for sampling statisticians, but thanks to the BigData phenomenon, the topic is having a renewed consideration. As Couper (2013) reminds us:

Non-probability surveys have been around for a long time [...] but the recent attention that has been paid to such methods can be attributed to the rise of Internet surveys.

A more recent example (Yang, Kim and Hwang, 2021) points out:

While [BigData sources] provide timely data for a large number of variables and population elements, they are non-probability samples and often fail to represent the target population of interest because of inherent selection biases.

The qualification *non-probability* essentially refers to the unplanned, undesigned and unknown selection process providing a NPS, namely the R -mechanism in Meng's framework, which implies uncontrollable bias and possibly low data quality. For example, according to Nandram and Rao (2023) non-probability samples lack the probabilistic structure and hence can be grossly inaccurate. The key issue posed by the biased nature of NPS is that it cannot be corrected by using the sample itself. The state-of-the-art methodological proposals to improve NPS biases have mainly developed along two classical sampling approaches: the design-based *propensity score* approach, and the model-based superpopulation approach. Extensive reviews are offered in Wu (2022) and, with more disseminative intentions, in Beaumont and Rao (2021).

Both approaches rely on the availability of extra data and auxiliary population information, whether design variables known in advance for all population units (e.g. cluster and stratification indicators) or covariates and paradata measured on the sample alongside the study variable(s). Furthermore, the ability of both approaches to actually reduce the detrimental effects on the validity of inference of an uncontrolled R -mechanism, relies on assumptions that are hardly met in practice, as well as irremissible and non-testable in principle. In particular, the propensity score approach applies provided that a probabilistic sample, with known sample design and negligible non-response, is available to be used as a reference. In the practical case where the NPS and the reference probabilistic sample share a set of $p \geq 1$ measured covariates $\mathbf{x} = [X_1, \dots, X_j, \dots, X_p]$, though not the study variable y , the majority of adjusting methods are based on the modelling and estimation of the so called propensity scores into the NPS, also more appropriately called *participation probabilities*. These models are based on the common covariates as well as the known sample design of the reference sample. Using Meng's framework and the vector of sample membership indicator \mathbf{R} in equation (5.1), for all units in the NPS the participation probabilities are defined as $\pi(\mathbf{x}_i) = P(R_i = 1 | \mathbf{x}_i)$, and they are estimated by fitting some kind of parametric model, usually a logistic one. An unresolved limitation of any method based on estimated participation probabilities is that its effectiveness in adjusting for biases in the NPS relies on the (often implicit) assumption that the unknown R -mechanism providing the NPS is non-informative, i.e. independent on the study variable y , given the set of shared covariates: $P_R(R_i = 1 | \mathbf{x}_i, y_i) = P_R(R_i = 1 | \mathbf{x}_i)$ (Pfeffermann, 1983). Noticeably, this is also equivalent to the MAR (Missing At Random) assumption in the missing values literature (Rubin, 1976). Indeed, there is an inherent missingness in NPSs due to the lack of actual design variables, and typically, the reference probabilistic sample lacks the study variable y . The excellent review by Rao (2021) states:

Models that have been used for participation probabilities and for the study variable are based on strong assumptions. Understanding those assumptions and validating them is a big challenge in making reliable inferences from a non-probability sample alone.

7. Propensity score versus participation probability

The notion of Propensity Score and related validity assumptions have been developed in the context of observational studies for causal effects, i.e. empirical studies to estimate the effect of treatments, agents, interventions or exposures (Rosenbaum and Rubin, 1983). Broadly speaking, data are *observational* as opposed to *experimental* in that the former are collected under uncontrolled non-experimental contexts. Experimental data, produced under completely randomised experiments and clinical trials, readily allows causal inference. By contrast, observational data, often termed real world data, lack a randomised structure and therefore suffer from inherent biases, usually referred to as external validity issues or failure of generalisability. In this sense, observational data versus experimental data pose inferential challenges that are parallel to NPSs versus probabilistic samples. Indeed the gold standard for causal inference, still represented by the experimental approach, is undergoing a crisis similar to the current probabilistic survey crisis,

basically for the same reasons and leading to the same reaction. On one hand, there are expanding costs to implement randomized trials, extremely long times required to reach conclusive answers to causal research questions, and increasing regulatory hurdles, such as ethics and privacy concerns. On the other hand, there is an escalating desire to improve timeliness and reduce data collection efforts by harnessing the growing availability of observational data from both traditional and new sources, such as follow-up longitudinal datasets, digitized hospital records, and AI-powered (Artificial Intelligence) portable devices and diagnostic tools. Observational datasets are often large-scale and significantly easier to attain than experimental data. In a sense, they are the new Big in clinical and epidemiological research (as well as in psychology and econometrics, see among others Athey and Imbens (2017), Imbens and Wooldridge (2009) and Hirano, Imbens and Ridder (2003)). A similar alignment between causal inference from observational data and valid inference from NPSs can be tracked under more principled, conceptual respects, including perfect equality in the formal definition of propensity score and participation probability (see previous section). Potential outcomes are (still) the mainstream approach for defining causal effects, whose roots date back to 1923 in J. Neyman's PhD dissertation in the context of completely randomised experiments (Neyman, 1923), and extensively researched since the 1970s in more general settings of observational studies (Rubin, 2005). In its basic formulation, the term potential outcomes refers to the two possibly different responses y_{i1} and y_{i0} that can be potentially observed on any unit $i = 1, \dots, n$ would it be *treated* or *not treated*, respectively, in the causal inference study. Hence, in principle, assessing casual effects implies comparing, at unit level, the potential outcomes. Clearly, exclusively one outcome y_i for each unit i can be actually observed in practice, a fact that is referred to as *the fundamental problem of causal inference* (Holland, 1986), i.e.

$$y_i = y_{i1}T_i + y_{i0}(1 - T_i)$$

where T_i denotes the assignment-to-treatment indicator, which is 1 if unit i was treated in the causal inference study and 0 otherwise. In the potential outcome framework, T is in fact the R -mechanism in Meng's framework: it is Random when it has a perfectly controlled probability distribution, e.g. $T_i = R_i \sim \text{Bernoulli}(0.5)$ in a completely randomised experiment, while it is (totally or partially) uncontrolled and unbalanced in an observational study, which aligns this latter to a NPS of treated (untreated) units. In standard causal inference studies, a vector $\mathbf{x} = [X_1, \dots, X_j, \dots, X_p]$ of $p \geq 1$ (baseline) covariates is measured on every unit under study, before the treatment assignment, e.g. demographics and medical history. The propensity score is defined equal to the participation probability into a NPS: $p(\mathbf{x}_i) = P(T_i = 1 | \mathbf{x}_i)$ and, once estimated, is used for rebalancing observational data for causal inference purposes. There are three main propensity score-based balancing methods: (i) statistical matching; (ii) stratification; and (iii) inverse probability weighting (IPW) (see for instance Imbens and Rubin (2015)). Noticeably, the latter is equivalent to the familiar Horvitz-Thompson (seldom Hajék) estimator in survey sampling methodology. A further connection between causal inference from observational data and valid inference from NPSs can be highlighted in the assumptions that must hold for propensity score methods. These assumptions allow both valid inference on population parameters from NPS data and causal inference on treatment effects from observational data within the potential outcome framework, referred to as POobs in

the sequel. For this latter, two main assumptions, regarding the assignment-to-treatment indicator and/or the propensity score, are:

A1-POobs Unconfoundedness or Ignorability: $T_i \perp (y_{i1}, y_{i0} | \mathbf{x}_i)$, i.e., the assignment-to-treatment indicator is independent of the potential outcomes given the covariates;

A2-POobs Common support or Positiveness: there exists a positive real δ such that $\delta \leq p(\mathbf{x}_i) \leq 1 - \delta$ for all covariates and all units in the study.

The practical meaning of A1-POobs is that no relevant covariate affecting the treatment and/or the response is missed in the causal inference study, i.e. no missing confounder. More conceptually, assumption A1-POobs is also known as strong ignorability of the treatment assignment mechanism, and can be equivalently stated as $P(T_i = 1 | \mathbf{x}_i, y_{i1}, y_{i0}) = P(T_i = 1 | \mathbf{x}_i)$. While both A1 and A2 in POobs context (can be managed to) hold true in a randomised experiment, where T is a random indicator with perfectly controlled probability distribution, they are both strong requirements under uncontrolled treatment assignment, as is the case in observational studies.

For NPS, two main assumptions regarding the R -mechanism and/or the participation probabilities are:

A1-NPS Non-informativeness: $P_R(R_i = 1 | \mathbf{x}_i, y_i) = P_R(R_i = 1 | \mathbf{x}_i)$;

A2-NPS Positiveness: $\pi(\mathbf{x}_i) = P_R(R_i = 1 | \mathbf{x}_i) > 0$ for all covariates and all units.

The first assumption parallels A1-POobs both conceptually and in its practical interpretation. The practical meaning of A1-NPS is that all design variables must be observed in the pool of shared covariates, that is all variables characterising the R -mechanism (also the participation behaviour) are included into both the reference probability sample and the NPS, no relevant covariate is missed. Furthermore, as already remarked in the previous section, assumption A1-NPS implies that the conditional distribution of the study variable given the covariates in the NPS equals the same conditional distribution in the population, i.e. the R -mechanism is *ignorable* since \mathbf{x} includes all relevant predictors of y (Pfeffermann, 1983). However, it is important to notice that despite the formal equivalence in the definition of propensity scores and participation probabilities, there is a relevant difference in how they function in practice. While the re-balance effectiveness of the estimated propensity scores is rooted in (observational) data observed on both treated and untreated units under the same, albeit unknown, R -mechanism, the effectiveness of the estimated participation probabilities heavily depends on the capacity of the known sample design of the reference probabilistic sample to “grasp and explain” the unknown R -mechanism generating the NPS. Finally, the parallelism is apparent for assumption A2 in both contexts. A2-NPS is the basic sampling methodology requirement that all units in the population possess a non null probability to be selected in the sample, which implies in practice, the complete coverage of the sampling frames. For a NPS the notion of sampling frames makes less sense or even none, so that its validation in practice is a challenge (see Wu (2022) for an extensive discussion of assumptions for valid inference from NPS data and their practical implications).

In both contexts, NPS and POobs, assumptions A1 and A2 are unlikely to be met and certainly hard to validate in practice. These remain a big challenge of our BigData era, in urgent need of further research and innovative perspectives.

8. Bayesian Network: A unifying and promising tool for inference and prediction from BigData

Trying to wrap up the discussion so far, around BigData opportunities, quality versus quantity tensions, parallel inferential issues in NPS and Observational data (POobs), and practical challenges, I reckon Bayesian Network (BN) models appear to be unifying and promising tools in all these respects.

In a nutshell, BNs are multivariate statistical models, which satisfy sets of conditional (in)dependence relationships as encoded in a Directed Acyclic Graph (DAG) (Pearl, 1988). A DAG is a collection of nodes (or vertices) and directed arcs (or edges) that connect pairs of nodes. The nodes represent variables, and missing arcs between nodes imply conditional independence between the corresponding variables. Figure 8.1 is a basic illustration of a DAG with $p = 5$ variables ($X_j, j = 1, \dots, p$). Notice that the graph is both directed, with oriented arcs depicted as arrows, and acyclic, meaning it does not allow starting from a node and, following arc directions, returning to the same starting node. A BN is a DAG equipped with a joint probability distribution that satisfies the Markov properties (Lauritzen, 1996). In practical terms, the DAG, i.e. the graphical part of the BN model, shows the independence structure among the set of p variables, in an easy-to-read, two-dimensional but fully multivariate and qualitative mapping. Each node (variable) is associated with the conditional probability distribution of that variable given its parents, which are all nodes/variables graphically linked and pointing to that variable (or the marginal distribution if the variable had no parents). This leads to a factorised expression for the joint probability distribution over the entire graph, known as chain rule:

$$P(X_1, \dots, X_j, \dots, X_p) = \prod_j P(X_j | X_{pa(j)})$$

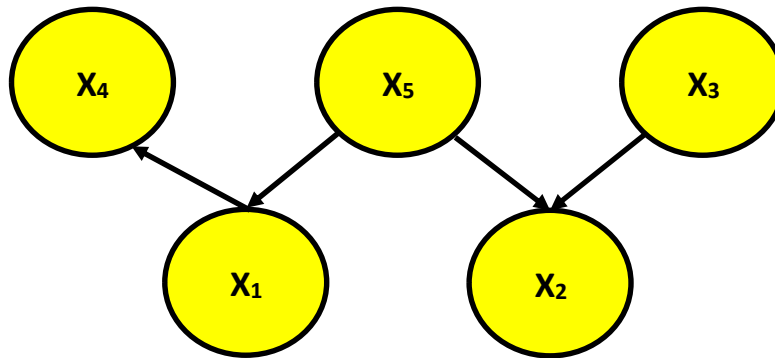
where $X_{pa(j)}$ is the set of parents of X_j , e.g. in the example in Figure 8.1: $X_{pa(2)} = \{X_3, X_5\}$.

All prior information about the association structure, such as subject-matter knowledge, is readily accommodated in the DAG in the form of either direct or missing arcs. When only partially compiled, as is often the case in practice, the BN model is estimated from data in a structural learning phase. One of the most popular methods for this is the PC constraint-based algorithm described in Spirtes, Glymour and Scheines (1993) (and named after its authors Peter and Clark). As a learning-from-data phase, the PC algorithm needs sufficiently large data sets, making BNs suitable for BigData settings. The actual functioning of the PC algorithm is straightforward. Starting from a complete undirected graph, the PC algorithm performs a sequence of conditional independence tests that recursively delete (1st phase) and orient (2nd phase) arcs between nodes. Eventually, the output provided allows the selection of an estimated BN, composed by both the DAG and the estimated joint probability distribution. In this sense, BNs are

transparent, fully statistical, and explainable learning-from-data models. The standard PC algorithm relies on the assumption of independent and identically distributed (iid) data. A modified PCcomplex algorithm that accounts for (known) complex sample designs is developed in Marella and Vicard (2022). BN models that include a specific assignment-to-treatment T -node, have been applied to improve the estimation of the propensity scores in POobs settings: Cugnata, Rancoita, Conti, Briganti, Di Serio, Mecatti and Vicard (2021) for causal inference from observational data with discrete covariates and a binary output, and Conti, Cugnata, Di Serio, Mecatti, Rancoita and Vicard (2023) for the more general case of a treatment with multiple different levels and a discrete outcome. The BN approach to estimate the propensity scores offers both appealing theoretical properties and practical advantages over traditional estimation methods and machine learning techniques. In POobs applications, propensity score estimation is routinely carried out by fitting a logistic regression model, with or without interactions. The BN approach is naturally parsimonious for not requiring the specification of a parametric model relating the response y to the $p \geq 1$ covariates, as it estimates the full multivariate association structure in the learning-from-data phase. Moreover, for all discrete covariates, BNs allow for maximum likelihood (ML) estimation (Cowell, Dawid, Lauritzen and Spiegelhalter, 1999). Therefore, the BN approach provides genuine ML propensity score estimators that are consistent, asymptotically unbiased and efficient, i.e. equally or more efficient than any other estimation method applied on big datasets, and asymptotically Gaussian distributed (Vicard, Rancoita, Cugnata, Briganti, Mecatti, Di Serio and Conti, 2023). Notice that collecting data and making decisions based on discrete or categorised variables is common practice in medical clinic and bio-medical studies, in addition to assessing the causal effects via point estimation of some kind of average effect, e.g., the popular average treatment effect (ATE) defined as the mean difference of the potential outcomes. The BN approach, by providing ML estimators of all probability distributions over the DAG (i.e., joint, conditionals, and marginals) allows for going beyond average point estimation to build tests for the presence/absence of treatment effect and confidence intervals. Incidentally, BN models can also act as powerful prediction engines for simulating scenarios and for what-if analysis (see, for instance, Mecatti, Vicard, Musella and Giammei (2022)).

BN models have been also applied for integrating data from different sources, a current research priority for Official Statistics. In Conti, Marella, Vicard and Vitale (2021) BNs have been considered to integrate data from two independent samples via statistical matching of multivariate categorical variables. The proposed method accounts for both the dependence structure between variables and the uncertainty around the estimated graphical model, along with, though separately, the uncertainty around the estimates of population quantities and super-population parameters. Moreover, the BN approach shows potential in extending to NPS settings. In these contexts, enhancing the estimation of the participation probabilities becomes a crucial prerequisite for implementing various proposals in the literature, including the particularly promising pseudo-empirical likelihood inference for NPS (Chen, Li, Rao and Wu, 2022), which preserves the desirable statistical properties of the empirical likelihood inference for an iid sample (Wu and Thompson, 2020).

Figure 8.1 Simple example of a DAG with 5 nodes (variables) and 4 directed arcs



9. Concluding remarks and THE challenge ahead

In this paper I have tried to articulate what is big in BigData, starting from its (still) blurry definition, delving into its multifaceted relationship with sampling methodology and practice, and passing through different kinds of data, application contexts, and goals. In the process, I have offered more open questions than conclusive answers and a few very personal views, including my own (scientific) keenness on the alluring potential of BNs. As a final personal consideration, THE challenge that I see as most urgent to sampling statisticians relates to what I described above as assumption A1. Indeed, the applications of the BN approach mentioned and envisioned in the previous section strongly rely on what is often referred to as the *usual assumptions* and, in particular, on some sort of ignorability, MAR, or unconfoundedness assumption holding true. Measures of bias have been proposed in the literature that can help assess this requirement for propensity score-based inference in both POobs and NPS setting. Examples include Andridge, West, Little, Boonstra and Alvarado-Leiton (2019) and Little, Boonstra and Hu (2020). However, the testability of these assumptions remains difficult, and the extent of their failure to hold in practice is largely unknown. Again, taking on the BN approach appears promising to address THE challenge of relaxing un-testable assumptions and trying to deal with selection bias under *any* R -mechanism by considering a proper R -node. This is, in fact, not a brand new challenge. Deep discussion and rigorous theoretical conditions can be tracked to the 1980s (Sugden and Smith, 1984). Again, the interest around BigData appears to have fueled a renewed, recent, and fast-growing literature. However, results and progress on this line of research are not easy to track as they lie on the borderline between scientific areas, e.g., Computer Science, Causal Inference and Sampling Theory and Methods, which traditionally aim at different goals and use different, specialised jargon. Recent literature, for instance Bareinboim and Pearl (2016) and Li, Irimata, He and Parker (2022), highlights a gap between theory and practice, making THE challenge a matter of stepping ahead from theoretical discussion toward practical solutions.

References

- AAPOR Task Force on Big Data (2015). *Report on Big Data*. AAPOR Council.
- Amaya, A., Biemer, P. and Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology*, 8(1, SI), 89-119.
- Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*.
- Andridge, R., West, B., Little, R., Boonstra, P. and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 68(5), 1465-1483.
- Athey, S., and Imbens, G. (2017). The state of applied econometrics: Causality and policy evaluation. *J. Econ. Persp.*, 31(2), 3-32.
- Bareinboim, E., and Pearl, J. (2016). Causal inference and the data-fusion problem. *PNAS*, 113(27), 7345-7352.
- Beaumont, J.-F. (2020). [Are probability surveys bound to disappear for the production of official statistics?](https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf) *Survey Methodology*, 46(1), 1-28. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf>.
- Beaumont, J.-F., and Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, 11-22.
- Bellhouse, D.R. (2000). [Survey sampling theory over the twentieth century and its relation to computing technology](https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000001/article/5174-eng.pdf). *Survey Methodology*, 26(1), 11-20. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000001/article/5174-eng.pdf>.
- Beresewicz, M., Lehtonen, R., Reis, F., Consiglio, L.D. and Karlberg, M. (2018). *An Overview of Methods for Treating Selectivity in Big Data Sources*. Statistical Working Paper EUROSTAT.
- Berkeley School of Information (2019). *Big Data Isn't a Concept - It's a Problem to Solve* - <https://ischoolonline.berkeley.edu/blog/what-is-big-data>.
- Biffignandi, S., and Bethlehem, J. (2021). *Handbook of Web Surveys, 2nd edition*. Wiley.
- Breidt, F., and Opsomer, J. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 3(2), 190-205.
- Carletto, C., Chen, H., Kilic, T. and Perucci, F. (2022). Positioning household surveys for the next decade. *Statistical Journal of the IAOS*, 38(3), 923-946.

- Castro-Martín, L., Rueda, M.D.M. and Ferri-García, R. (2020). Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. *Mathematics*, 8(6:879).
- Chen, Y., Li, P., Rao, J.N.K. and Wu, C. (2022). Pseudo empirical likelihood inference for nonprobability survey samples. *Canadian Journal of Statistics*, 50(4 - Special Issue: 50th anniversary of CJS), 1166-1185.
- Conti, P.L. (2022). *Non-Probability Samples and Big Data: How to Use them?* (Eds., A. Balzanella, M. Bini, C. Cavicchia and R. Verde), Book of Short Papers SIS2022, Pearson.
- Conti, P.L., Cugnata, F., Di Serio, C., Mecatti, F., Rancoita, P. and Vicard, P. (2023). *Treatment Effect Assessment in Observational Studies with Multi-Level Treatment and Outcome*, (Eds., F.M. Chelli, M. Ciommi, S. Ingrassia, F. Mariani and M.C. Recchioni), Book of Short Papers SIS2023, Springer.
- Conti, P.L., Marella, D., Vicard, P. and Vitale, V. (2021). Multivariate statistical matching using graphical modeling. *Journal of Approximate Reasoning*, 130, 150-169.
- Couper, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3), 145-156.
- Cowell, R., Dawid, A., Lauritzen, S. and Spiegelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. New York: Springer.
- Cugnata, F., Rancoita, P., Conti, P.L., Briganti, A., Di Serio, C., Mecatti, F. and Vicard, P. (2021). A propensity score approach for treatment evaluation based on Bayesian Networks, (Eds., C. Perna, N. Salvati and F. Schirripa Spagnolo), Book of Short Papers SIS2021, Pearson.
- Dagdoug, M., Goga, C. and Haziza, D. (2023a). Imputation procedures in surveys using nonparametric and machine learning methods: An empirical comparison. *Journal of Survey Statistics and Methodology*, 11(1), 141-188.
- Dagdoug, M., Goga, C. and Haziza, D. (2023b). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542), 1234-1251.
- Elliot, M., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(52), 249-264.
- FCSM-20-04 (2020). *A Framework for Data Quality*. Federal Committee on Statistical Methodology.
- Ferri-García, R., and Rueda, M.D.M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLOS One*, 15(4:e0231500).

- Groves, R. (2011). Three eras of survey research. *Public Opinion Quarterly*, (75th Anniversary Issue), 75(5), 861-871.
- Hansen, C., Johnson, C.R., Pascucci, V. and Silva, C. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*, (Eds., T. Hey, S. Tansley and K.M. Tolle) - Microsoft Research.
- Harris, T., and Frueh, S. (2023). The Complexity of Technology's Consequences Is Going Up Exponentially, But Our Wisdom and Awareness Are Not. *Issues in Science and Technology*, <https://doi.org/10.58875/tqkw5953>.
- Hartley, H., and Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174, 2170-271.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics.
- Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using estimated propensity scores. *Econometrica*, 71(4), 1161-1189.
- Holland, P. (1986). Statistics and causal inference (with discussion). *Journal American Statistical Association*, 81(396), 945-960.
- Imbens, G., and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Imbens, G., and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *J. Econ. Lit.*, 47(1), 5-86.
- insideBIGDATA (2019). 5 Big Data Vulnerabilities You Could Be Overlooking - <https://insidebigdata.com/2019/07/12/5-big-data-vulnerabilities-you-could-be-overlooking/>.
- Lauritzen, S. (1996). *Graphical Models*. Clarendon Press Publication - Oxford Statistical Science Series.
- Li, Y., Irimata, K.E., He, Y. and Parker, J. (2022). Variable inclusion strategies through directed acyclic graphs to adjust health surveys subject to selection bias for producing national estimates. *Journal of Official Statistics*, 38(3), 875-900.
- Little, R., Boonstra, P. and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8(5), 932-964.
- Lohr, S. (2015). *Data-ism: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything*. New York: HarperCollins.
- Marella, D., and Vicard, P. (2022). Bayesian network structural learning from complex survey data: A resampling-based approach. *Statistical Methods and Applications*, 31, 981-1013.

- Marker, D. (2017). How have national statistical institutes improved quality in the last 25 years? *Statistical Journal of the IAOS*, 33, 951-961.
- Mayer-Schönberger, V., and Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.
- Mecatti, F. (2022). IASS webinar 23, <https://isi-web.org/webinar/iass-webinar-23-bridging-big-data-and-sampling-methodology-what-big-and-where-bridge>.
- Mecatti, F., Vicard, P., Musella, F. and Giammei, L. (2022). Bayesian networks versus gender bias. *Significance*, 19, 16-20.
- Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox and the 2016 US presidential election. *Annals of Applied Statistics*, 12(2), 685-726.
- Nandram, B., and Rao, J.N.K. (2023). Bayesian predictive inference when integrating a non-probability sample and a probability sample. Pre-print. <https://arxiv.org/abs/2305.08997v1>.
- Neyman, J. (1923). *On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9*. Roczniki Nauk Rolniczych Tom X [in Polish]; translated in *Statistical Science*, (1990), 5, 465-480.
- Panimalar, A., Shree, V. and Kathrine, V. (2017). The 17 V's of big data. *International Research Journal of Engineering and Technology* - <https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf>, 04(09).
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc.
- Pfeffermann, D. (1983). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2), 317-337.
- Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *Journal of Survey Statistics and Methodology*, 3(4), 425-483.
- Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā*, 83-B(1), 242-272.
- Rao, J.N.K., and Fuller, W.A. (2017). [Sample survey theory and methods: Past, present, and future directions](#). *Survey Methodology*, 43(2), 145-160. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017002/article/54888-eng.pdf>.
- Rosenbaum, P., and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

- Rubin, D. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
- Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decision. *Journal American Statistical Association*, 100(469), 322-331.
- Shafer, T. (2017). The 42 V's of big data and data science - <https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/>.
- Spirtes, P., Glymour, G. and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag.
- Sugden, R., and Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3), 495-506.
- Tam, S.-M., and Clarke, F. (2015). Big Data, Official Statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review*, 83(3), 436-448.
- Tam, S.-M., and Holmberg, A. (2020). New data sources for official statistics - A game changer for survey statisticians? *The Survey Statistician*, 81, 21-35.
- Tam, S.-M., and Kim, J.-K. (2018). Big data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, 34, 577-588.
- Vicard, P., Rancoita, P., Cugnata, F., Briganti, A., Mecatti, F., Di Serio, C. and Conti, P.L. (2023). Testing for causal effect for binary data when propensity scores are estimated through Bayesian Networks. Preprint. <https://arxiv.org/abs/2302.07663>.
- Waldherr, A., Maier, D., Miltner, P. and Günther, E. (2017). Big Data, Big Noise: The Challenge of Finding Issue Networks on the Web. *Social Science Computer Review*, 4(9), 427-443.
- World Economic Forum (2012). *Big Data, Big Impact: New Possibilities for International Development*. World Economic Forum, Cologny-Geneva.
- Wu, C. (2022). [Statistical inference with non-probability survey samples](#). *Survey Methodology*, 48(2), 283-311. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf>.
- Wu, C., and Thompson, M.E. (2020). *Sampling Theory and Practices*. Book Series in Statistics. Springer.
- Yang, S., Kim, J.K. and Hwang, Y. (2021). [Integration of data from probability surveys and big found data for finite population inference using mass imputation](#). *Survey Methodology*, 47(1), 29-58. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf>.