IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 16, Issue 4, Ser. I (Jul. – Aug. 2021), PP 33-53
www.iosrjournals.org

Big Data in the Economy of Things: The V's and The Anti-V's

Bel G. Raggad
Seidenberg School of CS & IS

Abstract
We introduce the economy of things and discuss the state of big data given the economy of things. We explain our disappointment in digitalization and present our reasons. Despite the championing of big data by the white literature, the refereed literature is still discontent of data science and big data analytics. In an effort to advance the big data technology in an economy of things, we propose a value-based taxonomy of big data and propose a framework to integrate big data engineering, data science, and decision engineering. We also debate current big data analytics problems and claim that the big data V's are at the origin of these problems and we propose anti-V's to remedy for added complications.

In this taxonomy, big data continues to produce non-data facts. While a lot of these non-data facts are datafied to create grey data, and a lot of them are datafied to create dark data, there will also be a lot of non-datafiable non-data facts that will be discarded in dark holes.

For a justifiable competitive advantage, businesses will process data that is born data, and grey data to produce a sufficient decision support power to achieve their strategic goals. While these companies may also, as needed, supplement their big data inductive analytics for testing or to add veracity, there are rare occasions where an aggressive strategy may need to process dark data to achieve a tactical interceptive position in the economy of things. The dark hole data cannot be datafied and unless stronger and feasible non-data analytics comes around, this type of non-data remains inaccessible. However, digging deeper in grey data is often a feasible activity to attempt an extensive search for decisional insights that can advance the organization's business value generation capabilities. In contrast, accessing dark data may be an expensive alternative that is only advised to supplement or test big data inductive analytics.

We also discuss the mostly non-Bayesian decision engineering activities in an economy of things and propose Smets' Transferred Belief Modeling, in Dempster and Shafer theory, that presents a mathematically sound approach to manage non-Bayesian uncertainty.

Keywords*: Big data, Economy of Things, Internet of Tings, Digitization, Digitalization, Decision engineering, Big data engineering, Datafication, Grey data, Dark data, Dark hole.*

Date of Submission: 25-08-2021 Date of Acceptance: 09-09-2021

I. A Big Data-Driven Economy

The internet of things has changed the world and imposed the economy of things. Only few years old and big data continues to pour quintillions of data in this new economy. Unfortunately, big data came with a great deal of complexity through the many V's that add velocity, volume, variety, and veracity complications. The literature reports multiple attempts to study the effect of digitization on business growth ([1], [2]) and is aware that business transformation must rely on the cloud, big data/analytics, social and business dynamics, and mobility. In this new economy, the volume of digital data has now exceeded the volume of analog data all around and members are sinking in very deep data and cannot find the big data analytics capable of giving them the decision support power they seek out. Members continue to extract and amass large volumes of costly data that change at high speeds while their ultimate goal should be instead to extract actionable decision support that lasts. The digitalization efforts attempted including statistics, data mining, expanded machine learning have all failed to generate the decision support capability that yields a stable economy, and one that survives the big data V's.

An economy is, by definition, made of events where members can trade with confidence and with a known advantage. This advantage promotes the rational member's decision support, as in a Simon's decision process. In a digital economy ([3], [4]), however, this decision support is the result of a member's processing of all accrued intelligence, given an abundance of big data reports. With all the V's branding these deep big data resources, a member's decision support remains an unsure quest, may be, of the type of garbage in garbage out. Imagine what those V's can do to any member in a digital economy: a diversity of noisy unstructured facts continue to crop up and overflow any costly storages a member can feasibly have, with a high speed, and with unknown veracities. Under these conditions, it is probably more suitable, as in [5], to adopt a redefinition of the conceptual resources forming in big data as four types: noise, data, information, and knowledge. Noise, as in

unstructured data, is termed as raw facts with an unknown code system. Data are, instead, raw facts with a known code system. Information is, however, defined as a network of data facts that generate apprehension (cognitive ability), and a Bayesian update (a surprise). Knowledge is inference that can be tested and validated as principles or rules of thumb.

Big data, in an economy of things, is populated by instances of those conceptual resources. We later in this article propose a new value-driven taxonomy, using those conceptual resources. Members who may be individuals, small businesses, or midsize or large companies, process those resources to produce the decision support power they need to collect the event's gain associated with any executed trade in this economy. In a digital environment, as in [6], we view the digital economy in terms of two components: digitization and digitalization. The white literature is creating great confusion in distinguishing between digitization and digitalization. As in [7], digitization is the converting of analog contents into a digital format. As in [8], digitalization is a step beyond digitization to process digitization outputs to generate new business value generation capabilities. Digitalization is concerned with the analysis of data poured by the internet of things in our economy of things. Organizations have the opportunity to acquire the necessary digitalization power to transform these volumes of data into business value.

Members in the economy of things continue investing in big data on a grand scale. They continue hadooping without any planning of sound data analytics ([8], [9]). The produced decision support, using digitalization technology, remains unproven. There is just too much change in the data generated by big data and what members extract now may not be the same a minute later.

Massive volumes are produced, to an unimaginable rate. Every minute, about 2 quadrillions of bytes of data are created; that is, more than two quintillions of bytes are created in one day. Despite these unbound volumes, members still desire to have more. A rational member of this economy can no longer continue without a finite plan devised to produce efficient decision support with a Bayesian update. The volume concept attached to big data continues to be a paradox until the member's objective of a suitable extraction can be achieved.

The same may be said for the velocity concept in big data. A member of the economy of things continues to invest in high speed data, yet none of the members knows how to coordinate their digitalization technology to achieve the decision support production rate needed in the economy. Despite a high velocity of more than 50,000 GB per second through the global internet traffic, members are still asking for more. This paradox will continue until the objective of a synchronized decision support rate can be achieved.

The third V in big data is the coerced variety concept which meant the diversity of big data facts in terms of noise, data, information, and knowledge. As much as 80% of big data facts come unstructured and without known code systems. Members may exchange tweets, photos, videos, documents, etc. More than 80% of data growth is in videos, images, and documents. Most of these facts are produced with an unknown code system and come in a high speed. Still, members desire to get more by investing more in big data. The continuous acquisition of big data despite the coerced variety concept remains a great paradox that can only be resolved if the big data is filtered at a matched speed to collect only those facts with known code systems. It will be very difficult to produce a Bayesian update from unknown code systems.

Big data is also tainted with a greater paradox associated with the veracity concept. It is estimated that, in this economy of things, more than 30% of business leaders do not trust the decisional information they rely on to make decisions [10]. It is also estimated that poor data costs the economy more than three trillion dollars per year. Despite all types of ambiguities, inconsistencies, uncertainties that hinder the veracity of big data, members of this economy are still seeking more at a big scale.

There may be other V's added to big data [10], but as long as a feasible Bayesian update cannot be achieved, none of these V's can come without a paradox. With the internet of things, the world is all connected and all digitalized and all the events making the digital economy pour their facts in big data logs and repositories. These continual currents of indigestible facts do not bring along any feasible quality analytics to overcome the speed and size of reception. Big data emerged just too much earlier than the tools needed to tackle it. Big data will remain rogue as long as members fail to transform its V's into lasting decision support capabilities ([10], [11], [12]). Later in this study, we propose anti-V's remedies to advance data science and the big data technology.

Disappointments in digitalization

The literature [13] acknowledges the absence of a generally agreed definition of digital economy. Most however accept that the digital economy is the availability of digitalization throughout all sectors of the economy. Others see the digital economy as online platforms, and activities that owe their existence to the internet of things and digital technology. In this economy, every interned-enabled member counts, including individuals, devices, businesses of any size, etc. Digitization will touch all members of this economy so fast and they all impact this economy through social and feasibility requirements. Leaving out any member of this economy can be a great source of risk of failure.

Despite all great advances in technology, there is still no adequate digitalization power that matches them. While the internet of things gave new opportunities to members of the economy to share a great deal of data resources, despite the massive availability of data, our economy is still suffering from the traditional economic informational problems: information asymmetry, lemon markets, inherent principal-agent, inherent moral hazard, and framing problems. This section will briefly discuss these problems and show that the internet of things alone cannot solve all of them. A powerful data science providing for efficient big data analytics is still very consequential.

A digital economy would be both digitized and digitalized. The digitalization property requires the adoption of digital technology, including digitization processes, big data, and data analytics. The digitization property requires that all noise, data, information, and knowledge resources be converted using digitization technology into digital formats.

In such a digital economy, with current digitalization capabilities, including big data and the internet of things, becomes available to all members of the economy. As a result, any trade subjects and objects in a digital economy would be all internet-enabled. With the internet of things, we are achieving an economy where big data, through digitalization and digitization, continues to transform the digital economy to an economy of things where all members of the economy are connected and where all decision support opportunities become available to all members.

Assuming an economy of things, all members of the economy are internet enabled and share their information feasibly throughout the internet of things. Feasibility applies with all its five components (technical, economic, social, operational, and legal/ethical) to all security policies for all members of the economy. Feasibility remains the glue connecting all members of the economy in all events planned in the economy.

In contrast to an economy of things, we are facing an information asymmetry where events fail to reach and stay in stable equilibria achieved through acting upon maximum decision support capabilities. Simply said, asymmetry arises when a member of the economy does not possess or cannot process, organize, or communicate event information. In the absence of an economy of things, asymmetry of information can lead to a market failure due to unstable equilibria provisionally reached in various events of the economy. For example, in the absence of an economy of things, in a trade event, stock sellers, brokers, and buyers cannot avoid information asymmetry and such event can only achieve an unstable equilibrium unless all needed decisional information becomes feasibly available to all event members.

This rogue big data technology produced many hypes and we are in a state where both the literature and the white literature continue blasting us with countless claims that are misleading members of the economy of things in a real time manner due to the strong connectivity of the internet of things. While the literature is attempting to educate members of the economy in a consistent manner, the white literature works with a faster speed to spread deceptive assertions about the business advantages and profit abilities generated by the big data technology ([14], [15]).

Before the economy of things, we got used to the traditional problems of information asymmetry, lemon markets, the inherent principal-agent problem, the inherent moral hazard, and framing. Unfortunately these problems are still around even in the economy of things. Nobody knows for sure whom to blame, the internet of things, big data, data science, or something else. One thing for sure is that an economy of things, with a rogue big data, will certainly lead to greater complications that can produce great mishaps, as explained below.

The feasibility condition is required in an economy of things to assure all information is feasibly available to all members of the economy. This does not mean that all members possess all information in the economy; they instead only possess the information judged feasible to them, technically, economically, operationally, socially, and legally/ethically.

Precision of information is essential for sound economic decisions by members, but these sound decisions will embed members' sound judgment in the allocation of their feasibility. A sound allocation of feasibility throughout the economy of things will assure a stable economy and prevent any market failure due to information asymmetry.

In an economy of things, the feasibility requirement will lead to efficiently manage all economy information by receiving, processing, organizing, and communicating feasible information as needed. Members will detect any faults in the economy, for example, in case of production overflows or shortages or misallocation of scarce resources which will lead to market shutdowns when too many members spend a lot more or a lot less and when members produce a lot more or a lot less.

A stable economy is the result of a stable market when all members of the economy apply full feasibility in managing all the economy events facing them. In an economy of things, members of the economy cannot ignore certain properties of the trade events such as the quality of the object or the reliability of the subjects. They, instead, make assumptions based on other features of the event when an economic decision is made for the event. For example, a home buyer may assume a safe neighborhood when school taxes are high, or that the school district is highly ranked. This phenomenon is termed as the lemon problem which we thought

cannot be present in an economy of things unless big data owners failed to achieve the essential requirements discussed earlier [16].

In an economy of things, a member's economic decision will rely on decision support produced by big data analytics. Unfortunately, the complexity of big data analytics produced a great deal of confusion that may put more trust in the produced decision support than what it truly has. The variety in big data may render a member's decision very erroneous when certainty factors are inflated. The inflation of certainty factors associated with big data-driven decision support is similar to the inefficiency associated with the principal-agent mismatch of knowledge [17].

Brokers, for example, may employ data analytics tools that inflate principal's profitability based on which stock purchase decision is executed. The agent can be either informed of the inflation of results that lead to the purchase decision or the agent may be acting in a deceitful manner to achieve fraudulent commission gains.

Related to this phenomenon, the moral hazard problem can occur when the data analytical tools used to process big data in the economy of things is not carefully designed. The inefficiency of these tools may be due to the carelessness of the members of the economy or due to the overconfidence members have put in the big data. In either case, a great deal of adverse effects can happen. This is what is referred to as inherent moral hazard. The inherent moral hazard can also result from over-trusting the veracity property of EOT's big data.

Many members still apply a variety of statistical models to many extracted structured data subsets that produce results that misguidedly encourage certain managerial decisions that can produce poor consequences to members in the economy. For example, nutritionists may have advanced many herb supplements claiming they induce weight loss. These supplements have been sold in great quantities without leading to weight loss for buyers. Big data analytics can lead to a great deal of misleading information.

An essential question in an economy of things is, also, for example, for a member to inquire about what forms of diversity are adopted in extracting big data elements that can be processed through digitalization to produce decision support. The variety property in big data advances forms like videos, images, documents, and all types of unstructured noise resources. Members only prioritize diversity forms arbitrarily and have no basis in preferring other properties of big data. Is the veracity of the data forms an important criterion in selecting the big data forms to process? Can they adapt to the volume and velocity properties and achieve real-time decision support?

One may note the bias that may be present in selecting the big data forms to process. Members may think that a particular diversity form can produce all the decision support they needed, but it is possible that another form may be considered and may produce better results. Most often, everything seems to be identical, in difficulty or in easiness, in processing costs and availability, and members have to make a confusing choice. This is referred to as the framing bias.

With the great advances in big data and technology in general, it is very sad to still have the above problems and many others [10]. We thought that with economic intelligence, advances in artificial intelligence, advances in technology, and the internet of things, those problems are gone by now. The literature confirms they are still here ([18], [19]). We think that by improving and integrating big data, data science, and decision engineering, the above economy problems above and the new problems, we are listing below, associated with big data and data science will be significantly alleviated.

The state of big data

While digitization continues to grow beyond control, digitalization becomes more and more panting to further widen the gaps between the data generation stream and its process. Despite the expansion of big data and the variety of sources of facts, most owners, about 65%, are only interested in their internal data, not any other sources of facts, including the internet of things. And even, with those owners who undertake big data analytics for their internal data, most of them, about 83%, are only interested in the structured part of the big data. This may be interpreted as due to their perception of generated outcomes that can be lacking due to deficient data analytics employed in processing owners' data. The power of decision support that can be milked from unstructured data, grey and dark (see below), can be so consequential if the right data analytics power becomes available. Great business intelligence capabilities can stem from IoT sensors, social nets, smart grids, astronomic data, GIS, CCTV, medical data, gene information, and many other valuable sources.

The current trend in big data adoption, is to think of a set of attributes intended in a problem solving activity, for which data is sought and extracted from the big data. Unfortunately, this type of filtering drops away great business value that could have been embedded in thrown data. As a result, we have then declined a fully realistic view on things and we are left with incomplete and possibly faked results upon which false decisions may be made. Big data analytics should not allow for such a clipped data processing effort in any business decision support capability. Instead of such a clipping effort, big data analytics should, as needed, link to other data sources that may be useful in supporting, testing, or refining its generated inductive or exploratory

power. Big data is a great source of business value and a great decision support power; so why do we want to trim it and produce incomplete and artificial recommendations that often do us more harm than good?
Additional sources of data can be used to even further expand big data, whether it is coming from other servers, humans, data processing, or digitalization efforts. We can talk about the IoT and sensors integration, the social networks, audio sources, images or videos, cameras, geographic data, etc. This data expansion in big data can be seen more often in healthcare sources of big data where medical recommendations can only be possible if several big data resources are combined together. This is an example of a very complex and diversified data environment where accurate and a maximized big data analytics should be available ([13], [14]).
In contrast to big data, we however have the traditional data that is not big data, but data that is static, with a fixed format, and with measurable veracity. The literature [20] have sometimes referred to this type of data as 'small data,' and we will just refer as such throughout this article.

Table 1: Some differences between small data and big data

Small data	Big data
Fixed location	Anywhere on IoT servers
Well-structured data with known purpose	Mainly unstructured and will only have a purpose when datafied. Contents may or may not relate to known purposes.
One known owner	Many owners and a lot of it will remain with no owners until archived in grey data or lost in dark holes (see taxonomy below).
Fixed usage term. It is usually archived or discarded after use.	Has a life-cycle and continues to exist except for the dark hole data that is either discarded or archived (see taxonomy below).
Measured or sampled from data resources with known metrics.	More than 80% of big data has no known metrics until datafied. A lot of the big data will remain unmeasured even after datafication.
Often reproducible.	Big data comes as is. Despite a thorough process of datafication and cleansing, reproducibility is not guaranteed.
Associated with an accepted bound project risk.	While a risk management strategy is possible for big data, adventuring in big data adoption with unknown feasible purposes and resilience strategies can ruin the company.
Fully described in its structure or in its metadata tags.	Requires introspection techniques that may not be so accurate or feasible.
Treated using statistically-sound techniques.	Treated using inductive reasoning. Current data science only offers unproven data analytics for which veracity and soundness cannot be verifies due the many V's complicating big data streams.

While we fully know the owner, the location, the structure, the life cycle, and the purpose of small data, big data can reside anywhere on IoT servers and has no known owner, purpose, lifecycle, or structure. Moreover, while small data contains measureable attributes and reproducible data, and comes with known analytics and bound and acceptable risks, big data is irreproducible and has no known metrics or sound analytics, and brings great risks. Table 1 provides a clear distinction between small data and big data.

We are now in a state of big data where nobody knows what he/she is doing. We are in a mode of 'grab and go' where analysts grab some data, process it, produce artificial recommendations and claim big data championship. In an economy empowered by an internet of things, this state can ruin our new economy.
As discussed earlier, most of the existing big data analytics is just data analytics and not real big data analytics. We have identified, at least, the following problems:

1. Data myopicness
2. Unstructuredness intolerance
3. Venipuncture withdrawal
4. Analytical impotence
5. Stativity
6. Perishable inventory
7. Analytics fusionless
8. Infeasibility
9. Absence of Bayesian update-driven taxonomy
10. Absence of value-driven storage solutions
11. Indifference to security complications
12. Irreproducibility
13. Dependence on the big data V's

Decision support generators do not have to be any longer static but they are continuously reassessed and refined as big data changes along its V's. Only those models that stop changing when the big data V's change are declared admissible as big data analytics.

We keep producing data in different formats, at a high speed, everywhere. At every point in the Internet of Tings, data is accumulated, archived, or sent to a cloud or a big data plant. But if we fast-forward a little bit we are not sure we know how to feasibly create a decision support capability from it. Most often, however, data is stored or transmitted without any method of treatment for it. A digitalization process is taking place without any provision for a digitalization course. Overtime, we will amass an abundance of data without any digitalization power and a great deal of it will remain untreated or treated in unsound manners.

Big data suffers from data myopicness as current big data analytics aims at data and not at decision support. The main purpose of maintaining big data is not to amass data but to generate decision support and act upon it as needed. Any big data project has to have business value production as a result of the decision power obtained from big data analytics. Looking at big data as just data is a great mistake.

Big data also suffers from Unstructuredness intolerance as current big data analytics failed to conduct unstructured data analytics that concerns more than 80% of big data contents. Big data contains mostly unstructured facts, including tweets, images, videos, and symbols with no known language. Big data analytics does not provide sufficient analytics for the unstructured part of the big data. A great part of the big data remains untreated and may be hiding great potential business value that we still cannot reach using current big data analytics ([13], [14]).

Big data also behaves as in venipuncture withdrawal attempts as current big data analytics follow a blood work indicative approach, very much like the medical analysis a family doctor performs on a patient. As soon as he/she receives a patient's medical complaint, he/she formulates some possible medical conditions and tests for their presence. He/she orders a blood work which will show to the doctor whether or not the patient has the tested medical condition. This approach is a very sound medical test because the patient's blood is the same everywhere in his/her blood circulatory system. Unfortunately treating data in big data as the blood in a patient's body is wrong because the data changes and the blood does not. Then extracting some data and use it to test some new ideas is not statistically sound because new data will accumulate which may not give the same test results found in the initial data sets.

Most owners just treat big data as small data stores. There is no much data analytics that really applies to big data with its changing contents; and somehow, not many owners cared much about that and they seemed to be content with the models they inherited from small data. This analytical impotence makes big data lack the adequate analytical power needed to create actioned knowledge that is valid when new data comes by. Most current big data analytics still use the same machine learning techniques they apply to small data. These techniques only apply to the current data sets in question and not the data that is still to come with big data streams. Big data is a continuous source of data and findings from earlier extractions may not be still valid when tested on new data. Current big data analytics fails to produce results that stay valid when data changes as indicated by the big data V's.

The big data technology also suffers from stativity. Nothing that is recommended now by data analytics may work later when data changes. Current big data analytics follows a stative approach to elicit data support. Most big data analytics follows a stative approach and is designed to produce results based on past and current extractions.

With time, big data should become like a Sprague and Carlson's decision support generator [20] with a valid model base containing a rich inventory of data analytics. Current big data analytics maintains a perishable inventory. Over time, a model base for big data analytics gets populated. This model inventory has an expiration time beyond which the model is no longer valid because of changes in big data veracity, in its types, and its data structure.

At the same time, a big data model base should contain compatible inventory where veracity can be improved by fusing compatible models. Current big data analytics inventory is mostly unfusable and incompatible. Given the size of big data, there is no single model capable to process all the data in big data. A functional composition of the big data specifications is essential in designing big data analytics. Unfortunately, unless these models are compatible, their outputs cannot be combined to produce fused results. In addition to compatibility, the big data analytics model has to be fusable.

Little attention is paid to feasibility concerns in big data analytics. Current big data analytics does not enforce feasibility. Because of the size of big data and its multiple sources, big data analytics will encounter many feasibility complications: economic feasibility, technical feasibility, operational feasibility, social feasibility, and legal/ethical feasibility.

Big data also lacks a taxonomy that stresses business value and a Bayesian update. There are not many viable taxonomies for big data. Most studies however divide big data into structured, semi-structured, and unstructured data. This taxonomy does not distinguish between data and non-data. This taxonomy does not say much about the language, the vocabulary, or the grammar of the facts, and does not relate to any semantics or

business value for the facts. We hence cannot imply any decision support interest or Bayesian update in the unstructured part of the facts.

Given the speed of data generation in big data, this technology still lacks value-driven storage solutions capable of managing the high volumes accumulated and making them available to data engineering. Big data produces data at a high speed and unless a value-based smart storage solution is adopted, great business value may be lost with the data that has not been stored for further processing. Any smart storage solution cannot be designed without a great emphasis on decision engineering. Moreover, with this data flooding from continuous streams, big data is still indifferent to security concerns. Big data analytics is not concerned with security complications: the larger the big data the larger are the threats. Current big data analytics does not assume any complications due to information leakage, information corruption, or denial of service. The risks are very high ([22], [23]).

There are many contributing factors affecting the big data V's. The data generated may contain great inconsistencies that make it impossible to reproduce results. Reproducibility is an essential feature in the scientific method and cannot be compromised. Despite a sound data engineering method, irreproducibility may be present due to invalid data analytic techniques applied to sound data sets.

More importantly however is the absent capability of generating recommendations that are independent of the big data V's. We do not want to act upon recommendations that will no longer be valid as soon as velocity, volume, variety, or veracity of data changes. Current big data analytics does not manage the V's risks. Each one of the V's brings with it great difficulties for the big data analytics. Current big data analytics does not provide solutions for the big data V's. Appropriate controls should be in place for the velocity of the data, its volume, its variety, and its veracity.

An easy review of the above big data analytics difficulties will conclude that the main source of problem is in fact the four V's. Once we put all those V's under tight control we will see that all those difficulties fade away. Let us discuss those V's and prescribe ways to manage them. We will propose some solutions to counter the V's that we call the anti-V's, throughout this article.

Salvaging the big data technology:

Most of the literature ([24], [25], [26]) is mistakenly championing data science, as is. Data science is in fact at the origin of many of the big data problems discussed above. We all talk about big data and data science. Most, however, simply meant the data they extract from big data and not the big data itself. The claims and success stories they tell stem from processing data obtained from big data and not necessarily from processing the entire big data they have. Most of big data users have treated big data analytics as a blood work activity where a doctor orders a lab work where blood is drawn from a patient's veins to test for some suspected medical conditions. While the blood work activity is scientifically sound because the same blood passes through the patient's veins, it is not the case for big data analytics because the data in big data is not the same over time due the changes in its speed, volume, variety, and veracity.

Big data analytics should leave the traditional deductive reasoning and move to more dynamic data analytics including inductive reasoning and explorative analysis. Instead of accepting a prescribed hypothesis based on partial data, as it is the case in deductive reasoning, a sequential mechanism will instead create new candidate hypotheses, and continuously reevaluate and validate all candidate hypotheses.

Our salvage attempt looks into three forces that need to work together to reclaim the big data technology: decision engineering, data science, and data engineering. Data science alone cannot do it, because data science will need adequate quality data to perform its analytics and this data comes from data engineering; and will need correct specifications of the decision process, which comes from decision engineering, to guide its analytic power towards producing the exact decision support owners need.

Figure 1 depicts a possible integrated protocol of cooperation between data science, data engineering, and decision engineering in order to achieve a feasible actioned decision process.

Decision engineering is the application of technical, economic, operational, social, and legal/ethical knowledge in understanding an existing system and exploring improvement options to secure sufficient data support for these options, and predict possible system conditions capable of permitting these options, review outcomes in terms of new data as it become available.

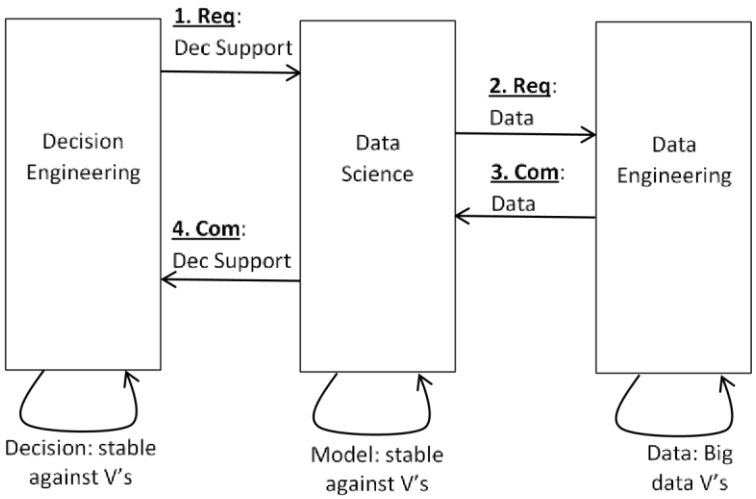

Figure 1: Integrated protocol for data engineering, data science and decision engineering

Big data engineering is an integrative approach to a Simon-based big data analytics. Big data is considered a supply source of facts that can be partially datafied and transformed into descriptive information, decisional information, and knowledge. Big data engineering can process facts in a sequential manner, datafy facts, create information and knowledge for the purpose of supporting decision making under unstructured uncertainties in dynamic environments.

Big data engineering:

Just too much confusion in big data. People say data and mean information, and people say information and mean data. Others say knowledge to mean information, and information to mean knowledge. Also, most say unstructured data and mean non-data without a structure. We just cannot build data science with all of this mix-up, or think of big data engineering in all this muddle. We have to redefine these big data concepts in order to advance big data engineering on sure grounds.

Let us redefine things.

Not all big data facts are data, but some are data, less than 20%, and some are non-data, more than 80%. So, big data contains facts. Those facts that have a known code system are called data [5]. The facts that have no code systems, or a vocabulary and a grammar that produce a cognizable meaning to owners, are called noise or non-data [5].

The data part of the big data is feasibly explorable using deductive, inductive, and abductive reasoning as needed. The noise part of the big data, however, has to be datafied before explored. Unfortunatly not all noise in big data is datafiable. We therefore divide non-data facts into three non-data types: 1) feasibly datafiable non-data facts that we call grey data; 2) infeasibly datafiable non-data facts that we call dark data; and 3) non datafiable non-data that we call dark hole non-data.

At a higher level, after datafication, we produce two types of information: descriptive information and decisional information. Figure 2 depicts this new big data taxonomy.

Descriptive information consists of a network of data facts with a vocabulary and a grammar that produces cognizable meaning to owners. Decisional information is descriptive information that generates a surprise to owners, which is referred to as a Bayesian update in decision theory.

At a higher level after information creation, there is knowledge formation. Knowledge is inferences from information that are tested, validated, and adopted as principles or rules of thumb.

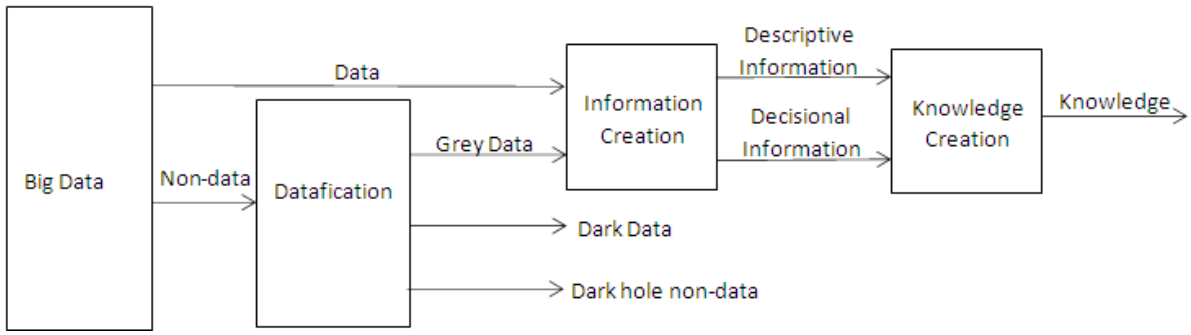

Figure 2: New big data taxonomy

[24] and [26] conducted an in-depth systematic review of IS literature on how owners achieve business value from big data and found that portability and interconnectivity, as in the internet of things, are essential socio-technical feature that are behind the business value capabilities companies gain from big data. The generation of distributed data and its availability to all members of the economy of things are required conditions for creating a competitive advantage. They also advised that realigning work practices, organizational models, and stakeholder interests will secure great business value from big data [27].

In supporting a sound data science, a big data study may be better conducted using an engineering management approach where we apply the practice of management to the practice of engineering. We proceed to a big data engineering task where we, as engineers, 1) recognize a well-defined need; 2) define the problem, the objectives and the constraints; 3) collect information and data, 4) generate alternative solutions, 5) evaluate the consequence of different solutions, and 6) decide and specify the final best solution.

A management approach to big data engineering, in this new economy (economy of things), can benefit a great deal from assuming a Simon's decision process, at least for midsize and large size companies. While most engineering methods implicitly include a review task when the best solution is reached, the Simon decision process explicitly imposes a review phase where the best solution is reviewed which may require revisits as needed to the earlier phases of the process. The revisits in a Simon process fit well the big data engineering task that has to revalidate data analytic outputs when data changes due to big data velocity, volume, variety, and veracity. Such a process aims at nothing but reaching the best course of actions given big data. In this new economy, we have plenty of data, from things like IoT sensors, social media, GIS, astronomy, CCTV, etc.

Simon follows 4 management phases: intelligence, design, choice, and review. As you can see, Simon aligns well with the big data concepts because intelligence is in abundance due to volume, and design is meant to impose an engineering method where the owner is fully aware of all possible feasible solutions to the problem. Design imposes an engineering requirement that the selection process among the feasible solution is a scientifically sound mechanism. The review phase aligns well with an engineering requirement for validation of the design and choice when data changes at the big data speed and the quality and format of data change due to the veracity and variety complications of big data.

Decision Engineering:

Decision engineering is the application of technical, economic, operational, social, and legal/ethical knowledge in understanding an existing system, exploring improvement options, secure sufficient data support for these options, predict possible system conditions capable of permitting these options, review outcomes in terms of new data as it become available.

This definition recommends inductive reasoning in a feasible Simon's decision approach. The five feasibility requirements of technical, economic, operational, social, and legal/ethical components are present in the definition which also includes the Simon's intelligence, design, choice, and review phases.

According to Bayesian decision theory [28], a choice situation is characterized by a set of possible alternatives, a probability distribution over the set of possible states of the world, the outcome of each alternative in each possible state, and a utility function over the outcome space. The optimal alternative is the one that maximizes expected utility. Alternatives can be operational activities, functional activities, complex plans, strategies, etc.

Decision engineering applies complete and applied methods that offer decision makers a normative approach capable of reasoning and acting under conditions of uncertainty. This applied approach obeys the assumption that if the decision maker's preferences follow a set of intuitively appealing constraints, then there exists a probability function and a utility function, such that the most preferred alternative is achieved by maximizing expected utility.

In this manner, Bayesian theory remains the language for reasoning about uncertainty. Given limited assumptions about rational beliefs, a Bayesian update is demonstrated to be the optimal way to update prior beliefs with new information [29]. Unfortunately, most often, there is a great deal of unstructured uncertainties, and Bayesian ceases to work. A decision engineer, however, does not stop and do nothing when assumptions don't hold, computations stall, simulations don't produce results, or unstructuredness come by. In big data engineering, while fusion of probabilistic models is possible, the existence of these models cannot be assured, and even if assured, the quantification of uncertainties and the performance of model prediction remain unsure due to the unstructuredness of these uncertainties.

Some of the literature argues that the big data challenge is not the technical problem of moving the maximum amount of bits in least time, but the scientific challenge of representing the complex systems governing the digital world [25]. The current state of big data analytics whether it is computational analytics that produces statistics from extracted large data sets, data mining that produces hidden patterns and relationships in data for new knowledge discovery, or visualization techniques that depict creative maps, color-coded images, and coaching videos to aid understanding and guide decision making, is still lagging behind.

In decision theory [32], when objective probabilities do not exist we are allowed to seek subjective probabilities that should guide our decisions. The Bayesian theory also allows us to construct probability measures representing theses subjective probabilities to maximize our expected utility. Unfortunately, there is just too much complexity, ambiguity, and inconsistency in big data for most of the existing digitalization and statistical technology to work. We are left with machine learning attempts that will only work for the extracted data and as soon as extracted there is risk that it no longer represents the big data in question. We are then faced with a great deal of uncertainties where probabilities are neither defined, nor computable. The Bayesian reasoning is hence not of any immediate use, and we have to revert to a

Dempster and Shafer theory (DST) approach [33] where a Smets' transferrable belief model is applicable. ([34], [35]). And even with DST, we are still faced with unstructured data that we cannot include in the process without a great deal of mathematical acrobacy.

The TBM consists of two stages: the credal model stage and the pinistic model stage. One may alternatively opt for Shafer's plausibility functions as a substitute to Smets' pignistic probabilities, as both techniques stem from the same belief structure and both add greater interpretability to the TBM. [34].

The design of the creedal stage may be set to fully asserted evidence based on the selected subsets without accounting for any managerial judgment that may be sometimes relevant in some decision problems and without accounting for any certainty factors or discount factors associated with the evidence on hand. This means that the basic belief assignments expressing the uncertainty associated with the data subsets' evidence remain fully asserted. The overall evidence on hand on our decision parameters has been accepted and expressed as a single belief structure. As mentioned earlier, even though we here demonstrate the pignistic model, another way may alternatively choose to compute Shafer's plausibility functions as a substitute to the pignistic probabilities. Smets' pignistic probabilities may be induced from the obtained belief structure.

Data Science:

Data science is the intellectual and practical activity encompassing the methodical study of data towards producing knowledge, insights, and decision support capabilities through deductive reasoning, inductive reasoning, or other mathematically and statistically sound data analytics.

Data science includes both small data analytics and big data analytics. Big data analytics is a set of data analytic methods that process data and non-data facts generated by big data to produce decision support or actioned models that are stable with regard to the V's features of the big data.

That is, a data analytic model that ignores or cannot process the non-data facts generated by big data is not big data analytics. A data analytic model that produces outputs that change when the speed of fact generation, the volume, the veracity, or the variety of types of facts change. Big data analysts aim at devising techniques to capture and recapture big data facts, including data and non-data, and process and reprocess those structured and unstructured facts, as big data speed, volume, variety, and veracity change.

The decision support produced and the knowledge created, and the actionable insights have to be independent from the big data V's.

Given this new definition, most existing data analytic models are not big data analytics. They may however be very strong models for the creation of unobservable knowledge from observable data extracted from big data. Because the extracted data is not representative of the big data these models are not acceptable as is and if they are to be salvaged then they have to be validated against the big data V's.

Major data analytics approaches

The failure of big data analytics is not necessarily the data analytics methods they use, but mainly the way they use them. These data analytics methods work well for small data analytics but failed to take into account of big data features that can impact their outputs due to the big data V's. We will later propose the additions of big data analytics capabilities to empower data science methodologies. In this section, we however, present the current data analytics that data science proposes for big data processing. All these methods are valid to use for data generated by big data but they need to be later validated against the big data V's for which we will also propose anti-V's remedies.

As shown in Figure 3, big data analytics may be organized into four classes of data analytics: 1) Explorative analytics; 2) Adductive reasoning; 3) Deductive reasoning; 4) Inductive reasoning; 5) Deep learning; and 6) Perspective reasoning.

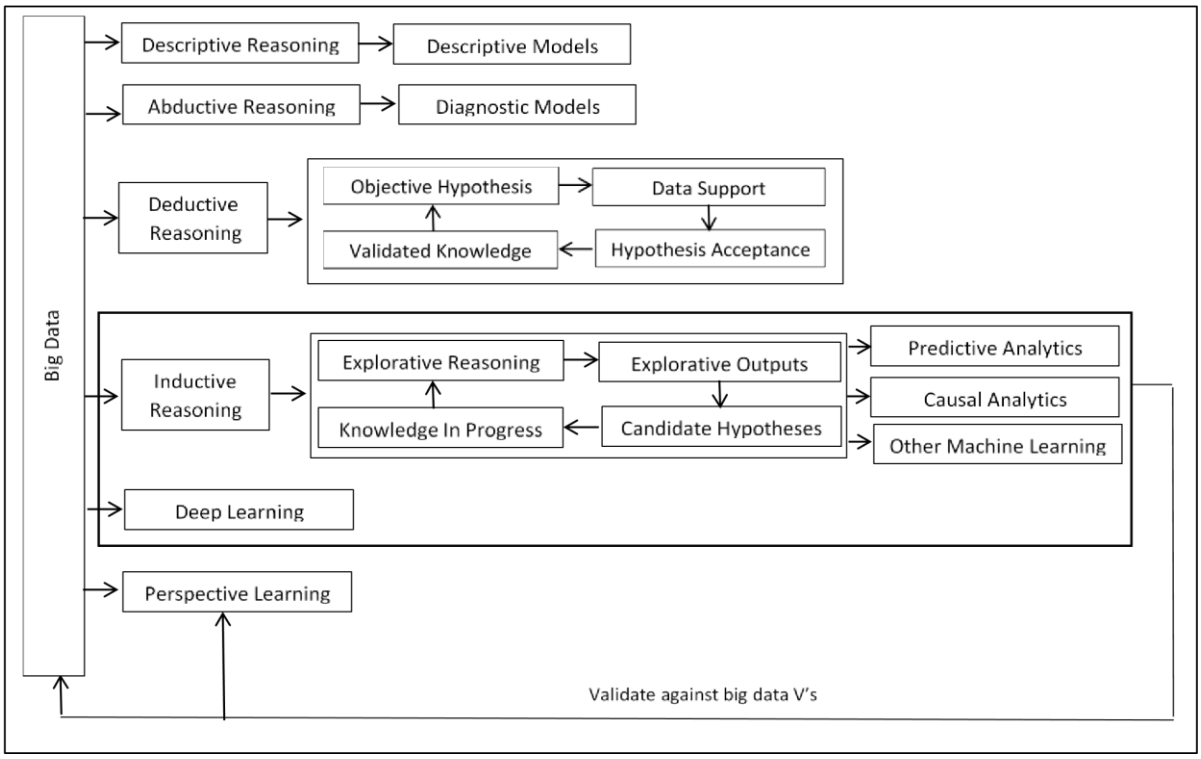

Figure 3: Big data analytics

Descriptive analytics uses descriptive models to understand the current system and obtain insights on how to plan further detailed studies to identify trends, and define relationships or patterns and discover new applied knowledge. This is also useful in planning deeper searches in the massive data repositories to uncover new knowledge that guides more studies or leads to more accurate paths for decision support. Abductive reasoning is the application of a diagnostic model where the consequences are given and the data analytic model is used to identify possible conditions that produce the consequences. The diagnostic models will help owners search for the presence of conditions that produce the desired or the undesired consequences.

Deductive reasoning is often needed to establish steady regulations and policies, and standard knowledgebased capabilities. Often, the owner defines a useful hypothesis to be tested and accepted as proven knowledge. Data is sampled and analyzed in a statistically sound manner to accept the right hypothesis at a prescribed acceptable significance level. Inductive reasoning goes through an explorative effort to generate useful ideas that will lead to candidate hypotheses. Each hypothesis will be tested using available data and a decision is made to accept it or reject it. The produced knowledge is only partial and temporary and as soon as new relevant data becomes available, more testing is needed to validate or reject the candidate hypotheses. Predictive reasoning can also be used to prepare observed data and model it to predict possible states of nature. The owner will be informed with possible scenarios that can be seen in the decision process. The analyst has to assemble sufficient data that will be processed in a statisticallysound manner to recommend the most likely outcomes. Approximate intelligent analytical models based on machine learning can also be used to make predictions.

The big data anti-V's to empower data science:

Even though there are many other V's added to big data, we are only considering the first four V's: velocity, volume, variety, and veracity. Velocity is the speed of data generation in big data. This concept will have a great impact on the validity of the decision support capability of the models adopted. The outputs of the data analytic models have to be stable when new data is generated. The volume concept indicates the big size of big data. I guess it is called big data because of the large volumes of data that populate it. You however never know what types of facts you can see in big data. Less than 20% will be data you know, as structured. The rest can be types with no known code system or grammar. There is not much what you can fit in a relational database. Object-oriented databases and object definition and manipulation languages will be needed. Variety is probably the most consequential feature that can be useful in planning big data analytics. Unstructured data (grey and dark data) will be a fundamental concept in big data for which no standardized processing methodology can exist. Veracity is also an important concept in big data analytics. The quality of decision

support outputs will depend on how much we trust the data. Let us propose some anti-V's to countermeasure the V's of big data.

The velocity anti-V: The objective of an anti-V for velocity V is to ease the velocity of data generation in the big data. Decision engineers will define the content of data needed to execute the decision process. The data engineering effort will define the data structure suitable for the decision process. Realtime private blockchaining will allow data engineering to create a block, hash it, and place it in the blockchain. The rest of the irrelevant facts, can be either data that was not blockchained or non-data.

A blockchain is an expanded list of blocks linked using hashing. Each block contains a cryptographic hash of the previous block, a timestamp, and content data. By design, a blockchain is resistant to modification of the data [35]. A blockchain is a time-stamped series of encrypted records of data that can be managed by co-owners over the internet of things. Even though the blockchain distributed network has no central authority, this can be redesigned to satisfy any requirements needed to plan the velocity management. Even though a blochchain is by definition a transparent medium and is shared by all members of the economy of things who see interest, we propose a virtual private blockchain that is only available to members of the decision engineering path; those who are authorized to add data to the decisional path, as in Figure 4. We propose that the virtual private blockchain be a fully automatic organizer capable of classifying received data of the big data in decisional paths registered as decisional paths by decision engineering. Figure 4 depicts a possible layout of a virtual private blockchain.

In order to justify the naming of the virtual private blockchain, you need to look at the interface between decision engineering and big data engineering as a set of tunnels where data goes in a tunnel from the big data either directly to the relevant decision process, or after datafication. This is a very realistic setting because the sources of big data should be known and the utility of the data should be acknowledged and the integrity of the data should be secured. Given those three conditions, only the relevant sources are authorized members of the permissioned private blockchain. The integrity of the data is guaranteed through hashing ([36], [37]).

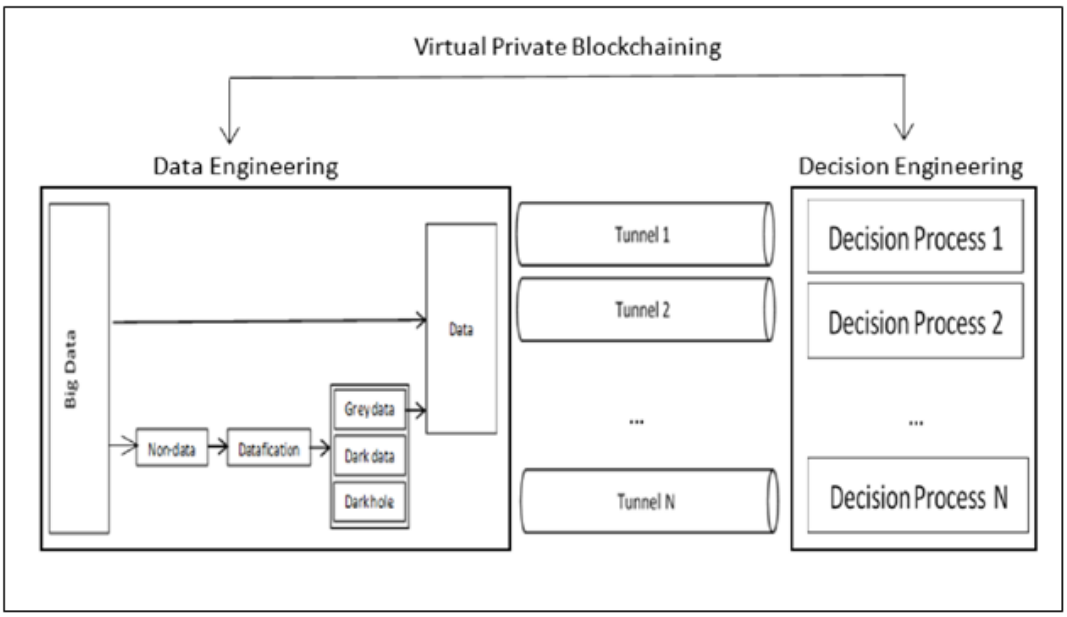

Figure 4: A possible layout of a virtual private blockchain

The feasibility of the permissioned virtual private blockchain (VPPB) method is assured because of the fusability requirement. Because data analytics model outputs are fusable then the data may be archived immediately after processing and included in dark data. All transferred belief models are fusable using Dempster rules of combination of evidence and data may not be needed after it is used in constructing the belief function structures needed for estimating the pignistic probabilities.

Of course, the Dempster and Shafer theory (DST) has been advanced and expanded [30] in diverse ways that are beyond the purpose of this study. We just wanted to show that even simple DST models can really solve many of the problems of big data and data science discussed in this paper. The development of the idea of adopting a permissioned virtual private blockchain involving the big data sources that are relevant to the existing decision processes. Figure 5 depicts one way to model the anti-V's complications in big data analytics.

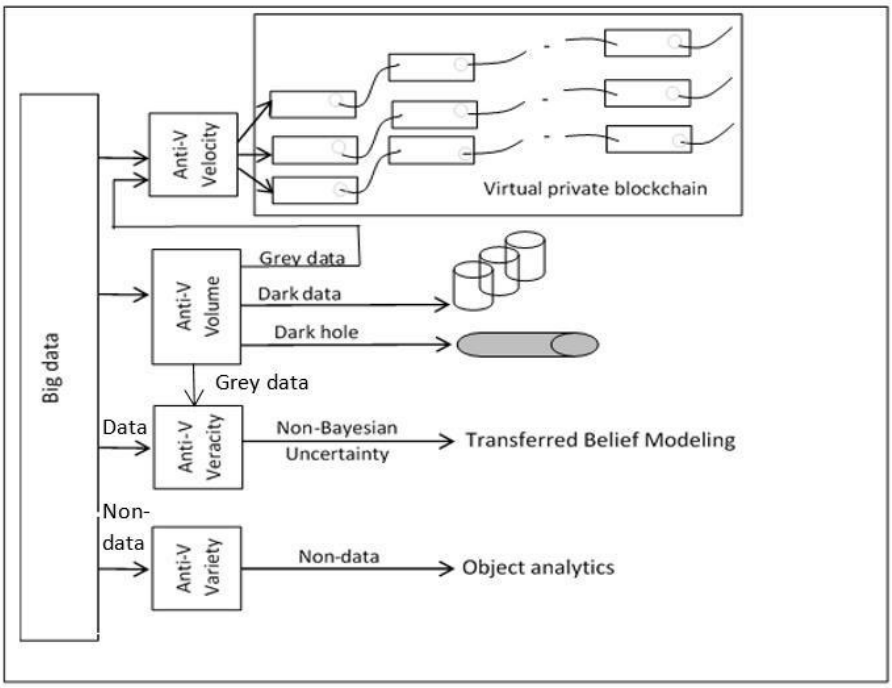

Figure 5: Anti-V's for big data V's

The blockchain is a simple yet ingenious way of passing information from one member to another in a fully automated and safe manner. One member initiates the process by creating a block that can be verified by the rest of the members including all computers distributed around the net. The verified block is added to a chain, creating a unique record with a unique history. Altering a single record would mean forging the entire chain in all the instances on the net. This is practically impossible.

The literature reported on few successful implementations of permissioned private blockchain platforms [24]. The literature is also noting an increase in business adoption in blockchain, in general, and an important advance in the development of permissioned private blockchain technology.

In a VPPB blockchain, the owner sends an invitation to sources relevant to decisional paths. These sources will need to register to be a member of the VPPB network. This can be achieved either manually by the big data owner or automatically by evoking the VPPB security policy.

Companies may choose to implement a VPPB network to impose restrictions on who is permitted to register in the permissioned network. The VPPB security policy will also define the roles in data contributions, and in maintaining the blockchain in a decentralized manner. In fact, with the adoption of blockchaining, as depiected in Figure 9, we have hit two birds with one stone: We solved the speed and volume problems, and the big data security problem. Security is a great feature in blockchaining ([36], [37]). A strong cryptographic mechanism using public and private keys make it feasibly impossible to steal owners' identities and their data. The data integrity of blockchain data is a sure feature given the hashing techniques used [38] [39. The literature [16] provides great discussion of blockchaining and big data security.

The volume anti-V: The objective of an anti-V for the volume V is to adopt a smart storage approach that is capable of datafying the non-data and dispatch it according to our proposed value-driven taxonomy. The grey data goes to direct access stores, the dark data to data warehouses, and the non-datafiable nondata to dark holes.

The variety anti-V: The objective of an anti-V for the variety V is to process as much of the non-data as possible. An object analytics is the only natural approach that works for non-data analytics. Objects are in fact data subsets that can be blockchained and processed using non-data analytics ([40], [41]).

The veracity anti-V: The objective of an anti-V for the veracity V is to adopt a mathematically sound approach to manage the non-Bayesian uncertainty associated with non-data. Smets' transferred belief modeling is a very appropriate and mathematically-sound approach that applies credal and pignistic modeling to manage uncertainty.

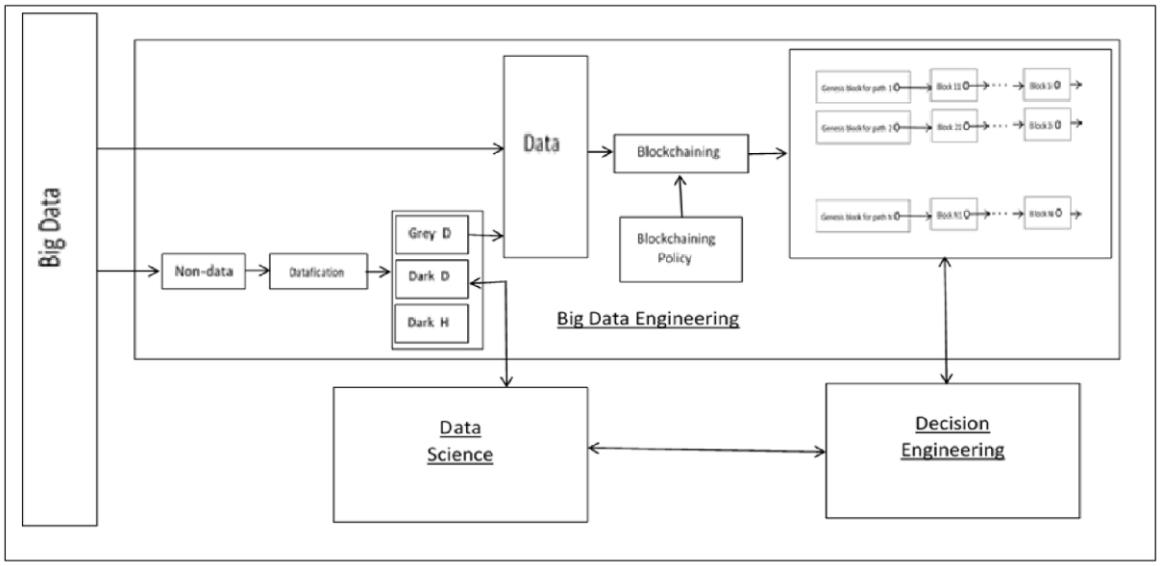

Figure 9: Virtual private blockchaining for big data

Data science unlimited

Data science is here to stay. It will also get stronger as long as big data continues to pour non-data facts requiring intelligent data analytics. Any lack in data science will lead to greater untreated volumes of dark data. In theory, dark data is only called upon for testing data analytics for planning an aggressive competitive power, but strategic matters may require looking everywhere there is business value, including dark data warehouses. Dark data, even though originated in big data, is still not big data because owners decide what to store in it which is usually the non-data scraps that data science could not tackle. Dark data may be treated instead as data warehousing where data mining and machine learning may still apply in testing activities or when more aggressive campaigns are planned in business gaming. The economy of things has however allowed for a more global use of big data and gave more computational freedom to all as more individuals and internet-enabled devices can now take advantage of big data in their own cost-effective ways. Small businesses also, especially in developing countries, may behave as individuals when big data is processed.

Small business daily operations and individual activities will still be continuously feeding big data through the internet of things. They are greatly impacted by big data, but in an indirect manner. Competitive advantage may be achieved indirectly through big data analytics. Individual will mostly benefit, through informational and decisional conjecturing, from big data in a general manner without the need of sound big data analytics.

The unstructuredness in uncertainties should not stop the decision engineers from approaching the decisional framework and seek new probabilistic techniques capable of recovering the initial Bayesian decision theoretic model and solve it for the best actions that maximize expected utility. We will in a later section discuss the Smets' transferred belief model as a sound approximation of the Bayesian decision process.

Unstructured facts can include useful information and knowledge, in addition to data and noise. All dark data in big data that is not used in analytics or that its security cannot be verified is stored in black or dark holes. Noise contents of big data can be rich in business intelligence if only we have the intelligent data analytics power that can produce the needed Bayesian update, the surprising decisional information that owners needed but had no way to anticipate. Unless a Bayesian update is produced, where this surprise decisional information is made available for owners to act upon, the data analytics remains inadequate.

It is often believed in the literature that deep learning can squeeze decisional information of great value to owners; this faith in deep learning as it is currently implemented through neural computing is misleading for the least. For there will be always new and huge dark data that cannot be included in the training of deep learning models and there may be always new and huge dark hole cases that cannot be classified. The decision risks when deep learning output messages are actioned may be too high to bear.

The big data literature [42] confirms that unstructured data, both grey and dark, add great decision power to big data owners. Microsoft, IBM, Google, and others have attempted processing noise facts in structured data in text files, audio files, video files to produce decision support power. They are so far advancing at a faster speed with processing text files but slower on other types of non-data types. The datafication of noise facts with unstructured formats, in images and videos, is still slowing the digitalization process of big data. The existing image recognition tools adopted by most companies in digitalization are only capable to analyze informational morsels in single images and videos. They can perform local cognizance but

cannot provide for training, testing, or studying the effects of loads of images and videos to satisfy the need to prove a concept and validate learned knowledge. Deep learning outputs may produce meaningful messages, but many problems can be present: 1) these messages are of limited representations of useful concepts, most often in the form of unfound associations; 2) these messages, even in their most acceptable forms, remain unfound unless the rest of the dark data is used in validating them; 3) these messages, even if they are well formed and well tested using dark data, they will still suffer from stativity because the new big data contents may tell a different story; and 4) these messages, have to bring decisional information with a Bayesian update.

We then have discussed several concepts related to the deficiencies of big data analytics, by detecting deficiencies, proposing possible validations, and possible corrections that will work to salvage existing big data analytics. Those data analytic models are however an important part of data science and will still work for small data. In fact, many owners, like individuals or small businesses, may still elect to use small data sets they extract from big data streams, given the sizes and types of decisions they process. Table 2 depicts big data owners and indicate their sizes and types of decision problems. In real life, big data will be very costly for individual and small businesses to adopt in their decision processes. We are proposing some recommendations on how to feasibly employ big data analytics to secure their planned competitive advantages, in terms of business size and decision type. We are also proposing big datadriven decision paths that include possible sequential activities that explain how to navigate decision engineering, data science, and data engineering, for individuals, small businesses, and midsize and large business, as proposed in the integrated protocol depicted in Figure 1.

Table 2: Big data needs for business types

Type	Freq	Var	Str	Management	Type of facts	Reasoning	K
I	L	N	N	Operational	Data (Structured)	Still Inductive + Structured decisions	None
S	A	N	N	Operational	Data (Structured)	Serial Inductive + Structured decisions	U
M	A	Y	Y	Functional	Data + Some Noise (Semi-structured)	Deductive + Serial Inductive + Semistructured decisions	K
L	H	Y	Y	Strategic	Data + Non-data (Noise, Unstructured)	Deductive + Continual Inductive + Perspective + Unstructured decisions	Deep K

I: Individual; S: Small business; M: Midsize business; L: Large business; L: Low; A: Average; H: High; Freq: Frequency; Var: Variety; Str: Structure; K: Knowledge.

The individual big data-driven decision path:

As shown in Table 2, individuals will encounter problems for which data is needed with a low frequency; and these problems are often interrelated; the same data sets are often used to model them. Individuals are not owners of big data but are simply users. A data engineering approach will require that an individual passes through several sequential steps before achieving his/her decisional objective: definition effort, descriptive effort, explorative effort, inferential effort, predictive effort, causal effort, and an occasional mechanistic effort.

Big data engineering, for the individual decision path, is only active in an objective or problem-driven approach. Given a well-defined objective, the individual will look into all available data, including the continuous big data stream of data, and initiate relevant descriptive techniques on the existing working system, to see how and which data relate to the problem-driven objective. At this point, all data sets, stative and live, that are relevant to solving the problem in question have been identified. Remember, individuals are not owners of the big data. Given this well-defined scope of data, the individual can now apply inference techniques to find meaningful trends, patterns, and relationships that will guide the modeling of his/her decision problem. A decision model is now ready to use an inferential effort, where the individual searches within the very scope of data, for sufficient data support for the decision model in question. At this point, the individual has identified relevant data, and generated some hypotheses in terms of trends, patterns, and relationships uncovered through

explorative learning. He/she is then ready to initiate a review process where predictive techniques may be used to search for conditions favorable to the accepted hypotheses. After predictive reasoning, occasional causal reasoning and mechanistic learning may be needed. The causal technique is sometimes sought to assert an interesting causal-effect relationship that the individual can adopt as rule of thumb or principle when similar decision problems come around. The mechanistic effort, if available, will propose a deterministic mechanism capable of reproducing the same results in similar problem conditions. Figure 6 depicts the individual decision path for big data analytics.

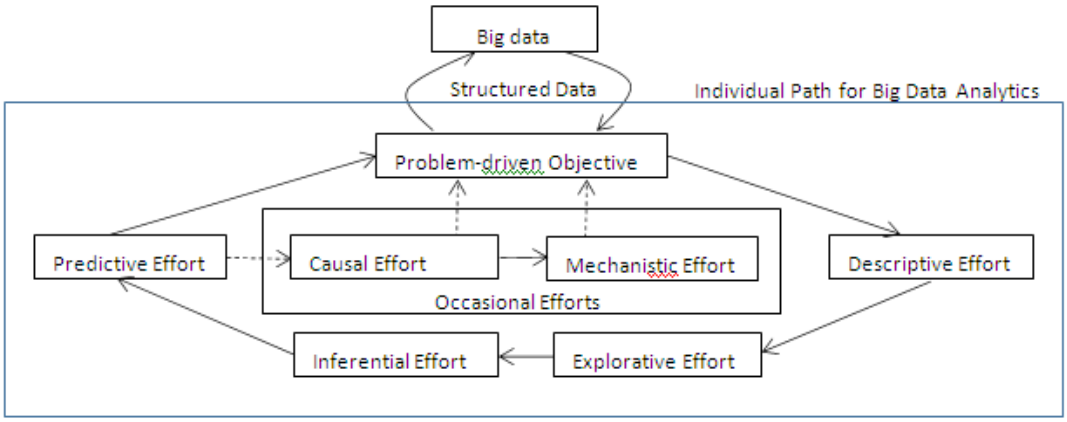

Figure 6: Individual big data-driven decision path

The small business big data-driven decision path:

A small business is often concerned with operational decisions. In real life, functional and strategic planning may implicitly take place, but it is only performed to define and model operational activities. The frequency of decision support is usually small and repetitive. There is no formal structure governing operational business decisions and most decisions are structured and achieved through inductive reasoning that can be validated using newer data subsets extracted from the big data stream. As depicted in Figure 7, a small business path for big data analytics is very similar to the individual path except for the serial validation of inductive models as needed.

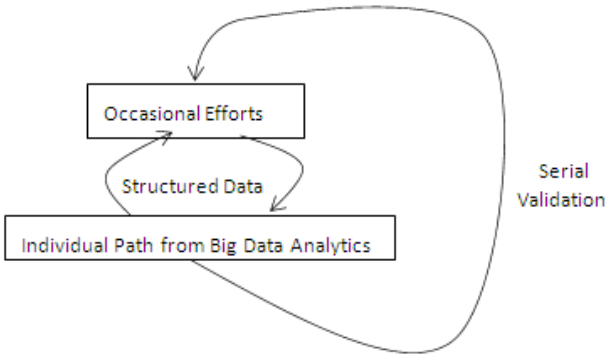

Figure 7: Small business big data-driven decision path

Midsize business big data-driven decision path:

A midsize company is often more concerned with functional decisions, even though strategic planning is a very important part of its management activities. Its functional decision support power is usually based on semi-structured decision models and are activated as often as needed with a moderate frequency. While deductive reasoning is often needed to establish steady regulations and policies, standard knowledge-based capabilities may be created to maintain a lasting competitive advantage. Midsize companies can play a more aggressive business presence by applying sequential inductive reasoning that is continuously refined based on new data brought by big data streams. While midsize companies can benefit a great deal of feasibility when a Simon's decision approach is integrated with big data analytics, they can always achieve a great decision support power, without Simon, in a heuristic manner. A midsize business path for big data analytics can tend to

a small business path for a smaller size midsize company, or to a large size path for a larger size midsize company.

Large size business big data-driven decision path:

A large size company is often more concerned with strategic decisions, while assuring that functional decisions be aligned with company's strategic plan, mission, and vision. Its strategic decision support power is usually based on unstructured decision models that are activated and validated with a high frequency. The velocity and variety of big data have to be matched by big data analytics. Deductive reasoning is often needed to establish new regulations and new policies, and to create knowledge that can be adopted as principles and rules of thumb.

A large size company treats a great variety of facts, including data and non-data, and processes a great variety of unstructured decisions. A sound decision support power in a large size company has to be very well structured to achieve a lasting business value generation capability. We require that a sound Simon's decision process should be integrated with any big data analytics, as shown in Figure 8. Large size companies can play a more aggressive business presence by applying continual inductive reasoning that is continuously refined based on new data brought by big data streams. Real business advantages can only be obtained if adopted data analytics satisfies the big data V's.

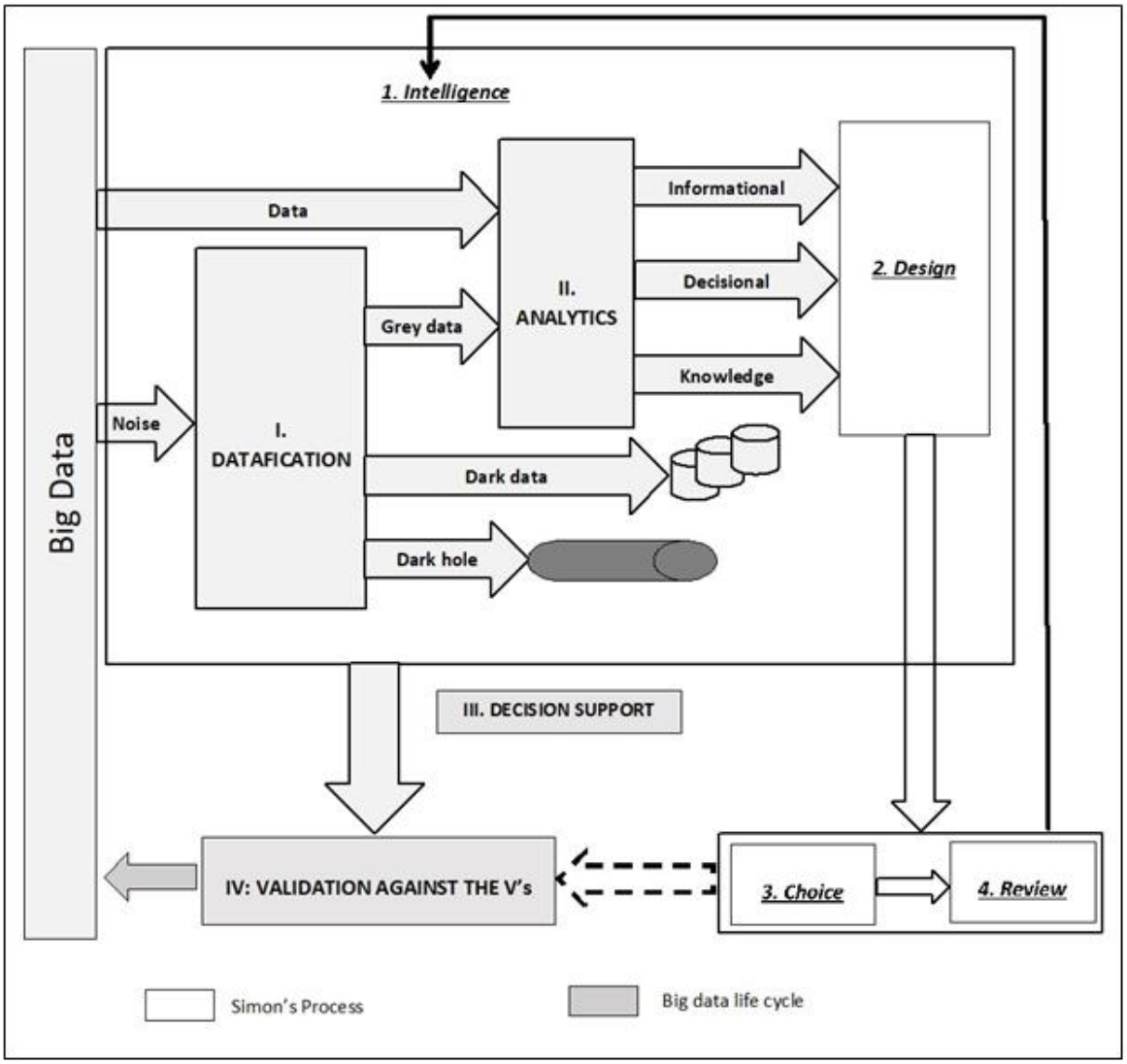

Figure 8: Simon's decision process integrated with any big data

Validating big data analytics

Most of the data analytics we discussed above only process data sets extracted from big data at a given time period. This is very much like treating small data, because the decision support produced may not be valid when a future data set is extracted and the current data set is not representative for the entire big data. That

is, in order to salvage the entire inventory of data analytics models we discussed above, we have to apply some validation techniques to accredit some of them for use in big data analytics.

We propose two simple ways to validate existing data analytics in order to accredit those models for use in big data analytics: a practical accreditation technique, and a congeneric accreditation technique.

In the practical technique, in order for a data analytics model to be accredited for use in big data analytics, this model has to be validated against the four big data V's. The produced decision support output should be stable when the big data V's change. This means that the model is tested at a high speed of data generation, when data comes in big batches, when data comes with a great deal of format diversity, and when data comes with an unsure veracity. The tested data analytics model should give the same decision support output, with an accepted prescribed small error term. Proposing methods of validation is beyond the scope of this paper, but more research is recommended to validate current data analytic models against big data V's.

In the congeneric method, we have two different cases: 1) Case of fusible decision support outputs; and 2) Case of infusible outputs.

If the outputs are fusible, then the accreditation method proceeds by applying the tested model to a large N randomly selected data sets and fuse the outputs to produce a single output. If any other data set from dark data produces the same output, given an acceptable prescribed error term, then the model will be accredited for use in big data analytics.

In the second case, the outputs are not fusible. We randomly select N data sets to which we apply the data analytics model. If the N outputs are similar, given an accepted prescribe error term, then the model is accredited for use in big data analytics.

Of course, in both methods, we cannot mathematically prove that future data from big data will produce the same decision support outputs. This mathematical proof is beyond the scope of this paper, but this proof is possible for large N when normal distributions are assumed. Testing the data analytics models on more data sets that are randomly selected from the big data can only provide more assurance that the tested models are more valid to use in big data.

Modeling non-Bayesian uncertainty in data science

We have redefined and discussed data science and related it to decision engineering and data engineering. We thought that data science will remain lacking when used in big data analytics if it does not add at least these three big data analytics capabilities: 1) the capability to treat non Bayesian decision processes, 2) the capability to treat data with unknown domains, and 3) the capability to process objects holding contents without a known code system.

Unfortunately, Bayesian reasoning may cease to work when too much ambiguity is characterizing the available data sources, like in big data. Under these conditions, data should be processed out of a Bayesian framework and sound measures should be produced to assimilate the probability distributions adopted in a Bayesian decision process. Alternatively, the Smets' Transferred Belief Model may be applied to produce pignistic probabilities that we can employ as Bayesian probabilities in our decision process. We later reserve a special section that demonstrates the working a Smets' model. We will show that these three capabilities can be achieved in a mathematically-sound way using Dempster and Shafer theory. These three capabilities will share the same mathematical setting as discussed in this section.

Our decision engineering approach requires that a detailed decision framework should be devised. Let us produce a simple framework as follows:

We have a decision maker, either an individual, a small business owner, or a manager in midsize or large organization. This decision maker has to choose among |A| actions from a set $A=\{A_1, \ldots, A_{|A|}\}$. There are N events from a frame of discernment $E=\{X_1, \ldots, X_N\}$. Given the great ambiguity associated with E and the lack of experimentation possibilities, Bayesian reasoning is hence inappropriate. We however have big data available to our decision maker to support his/her decision process.

According to our big data taxonomy above, big data will provide our decision maker with three types of facts:

1. Data (about 20%)
2. Non-data (about 80%)

The non-data part will, through datafication, produce grey data, dark data, and dark hole data.

The datafication process will organize the non-data into two types of non-data: lingual non-data and object non-data. The lingual non-data will require text mining and text-based analytics that may be used to produce decision support to the decision maker's problem. This aspect of non-data analytics is beyond the decision engineering approach in discussion. We are then left with the object non-data. An object can always be represented using a data set. For example, an image may be represented by a single data set and a video by multiple data sets.

The dark data is the same as grey data except that the former is, by definition, infeasible to process by the current decision maker. The dark hole data will be archived forever and is of no use to our decision maker.

That said, we now have data that we can use in our decision process: data that is born data as in 1 above, and data that was obtained as a result of datafication after all objects were translated into data subsets [42].

At this point, all data will be treated the same way whether data is born data or is representing objects. The origin of the data is not relevant in this proposed data analytics.

Let us assume a data subset $D=\{d_{ij}\}_{i=1,N;\ j=1,M}$. where the d_{ij}, j=1,M belong to X_i, the domain of attribute X_i, i=1,N. Without any confusion, we also use X_i to denote $dom(X_i)$.

Remember, we earlier said that we will address three big data analytics capabilities which are 1) the capability to treat non-Bayesian decision processes, 2) the capability to treat data with unknown domains, and 3) the capability to process objects holding contents without a code system.

We earlier announced three big data analytics capabilities that should empower data science. While the first and third capabilities are plainly discussed in this section, you can see that we did not make any assumptions on the definitions of the data attribute domains. They can be finite or unbounded, and nominal, ordinal, or intervals. That is, even the third capability that required some blind data analytics is implementable using this proposed framework because we did not impose the full knowledge of domains as in statistical analysis. Uncertainty management will be processed on Ω the Cartesian product of the power sets of the attribute domains X_i, i=1,N, as follows:

	2X1	- - -	- - -	2XN
e_1	{…}	- - -	- - -	{…}
- - -	- - -	- - -	- - -	- - -
e_K	{…}	- - -	- - -	{…}
$K = \|2^{X1}\| x \dots x \|2^{XN}\|$				

The universe Ω is in fact a matrix of hypertuples where cells contains subsets instead of single values. Consider then a hypertuple e, $e=\{e^1, \dots, e^N\}$ and where e^k= is a subset of X_k. Also let Δ_α be a partial order relation on all the data sets on hand. If x and y are elements of a set E, we say that $x\Delta_\alpha y$ if and only if $|x\cap y|/|x| \geq \alpha$. The intersection defines the amount of support x provides to y, or alternatively, the amount of α-compatibility between x and y (i.e., a compatibility with level α).

We define the evidence support $s_G^{\alpha}(e)$ of x in D as the set of y in D such that $y\Delta_\alpha e$. That is, $s_D(e) = \{y\varepsilon D$, such that $y\Delta_\alpha e\}$. The subset D is a poset with respect to the partial order relation Δ_α and it may hence have elements that are related to e (α-compatible with) and others that are not related to e (not α-compatible). Only the compatible elements y in D such that $y\Delta_\alpha$ e are accepted to support e.

Let us now construct the belief structure on hypertuples.

Let Ω, defined above, be our frame of discernment. The belief structure for D in Ω is defined as follows:

$m_D^{\alpha}: \Omega \rightarrow [0\ 1]$

$m_D^{\alpha}(e) = |s_D^{\alpha}(e)|/|s_D^{\alpha}(\Omega)|$ where $s_D^{\alpha}(\Omega) = \{\{y\varepsilon D$ such that $y\Delta_\alpha e\}, e\varepsilon\ \Omega\}$

Let us simplify our notations as follows:

$|s_D^{\alpha}(e)| = |x\Delta_\alpha D|$ = Cardinal of $\{y\varepsilon D$, such that $y\Delta_\alpha e\}$.

We then have the following:

$m_D^{\alpha}(e) = |e\Delta_\alpha D|/|\Omega\ \Delta_\alpha D|$, $m_D^{\alpha}(\Omega) = |\Omega\ \Delta_\alpha D|/|\Omega\ \Delta_\alpha D| = 1$.

At this point, we already have some decision support information we can act upon. There are two cases here: 1) Decisions are made based on the most probable events; or 2) we have decisions rules that may be applied based on the uncertainties associated with events.

In the first case, we need to compute the Belief value of all actions that the user will undertake based on the decision support we just provided in terms of belief functions. Let this action be denoted as A, we then have the following:

$$A = \text{argmax}_{[e_i \varepsilon \Omega]} Bel(e_i).$$

In the second option, the pignistic probabilities may be used be replace the Bayesian probabilities in computing the expected values of actions and the one that has the highest expected value should be chosen.

This section showed how to add simple big data analytics capabilities, in processing data and grey data, but stronger blind analytics and object-oriented deep learning may be made feasible, for darker data, to model harder unstructured data analytics. It is recommended inhere to study objects as features on which belief structures can be constructed and processed. Deep learning can then be applied to basic belief assignments as objects representing dark data. This approach may be very promising due to the availability of evidence combination rules ([31], [41], [43]) capable of fusing terribly unstructured data.

Also, in blind analytics, there will be cases where data scope is undefined and the attribute domains keep changing. Some examples where domains change may be global weather data, stock markets, spatial data, etc. In these cases, Bayesian reasoning is not applicable to model and manage uncertainty. This is then a good situation where Transferred Belief Modeling is appropriate.

II. Conclusion

This paper discussed the state of big data given the economy of things. We discussed our disappointment in digitalization and explained the reasons. Despite the championing of big data by white literature, the refereed literature is still discontent of data science and big data analytics. In an effort to advance the big data technology in an economy of things, we proposed a value-based taxonomy of big data and presented a framework to integrate big data engineering, data science, and decision engineering. We discussed current big data analytics problems and we thought that the big data V's are at the origin of the trouble and we proposed anti-V's to remedy for added complications.

In this taxonomy, big data will continue to produce non-data facts. While a lot of these non-data facts will be datafied to create grey data, and a lot will be datafied to create dark data, there will also be a lot of non-datafiable non-data that will be discarded in dark holes.

For a justifiable competitive advantage, businesses will process data and grey data to produce sufficient decision support power to achieve their strategic goals. While these companies may also, as needed, supplement their big data inductive analytics for testing or to add veracity, there are rare occasions where an aggressive strategy may need to process dark data to achieve a tactical interceptive position in the economy of things. The dark hole data cannot be datafied and unless stronger and feasible non-data analytics comes around, this type of non-data remain inaccessible. However, digging deeper in grey data is often a feasible activity to attempt an extensive search for decisional insights that can advance the organization's business value generation capabilities. In contrast, accessing dark data may be an expensive alternative that is only advised to supplement or test inductive analytics.

We also discussed the mostly non-Bayesian decision engineering in an economy of things and proposed Transferred Belief Modeling, in Dempster and Shafer theory, that presents a mathematically sound approach to manage non-Bayesian uncertainty.

To expand on this study, we recommend that more sound research is needed to address essential issues related to big data. The following recommendations call for further research:

1. Re-coordinate work between data engineering, data science, and decision engineering
2. Understand the sources of big data
3. Manage the big data V's
4. Solve the irreproducibility problem
5. Manage big data security
6. Blockchain big data decision paths
7. Devise object-oriented analytics
8. Where Bayesian reasoning does not apply, move to more intelligent analytics, like, Dempster and Shafer theory, deep learning, and intelligent inductive reasoning. All should be tested using dark data.
9. Justify big data analytics by studying technical, economic, social, operational, and legal/ethical feasibilities.
10. Apply value-driven smart storaging
11. Achieve aggressive competitive advantage by dark data analytics
12. Apply blind learning for data without known domains

References

[1] V. Parida, Sjödin and W. ReimReviewing, "Literature on Digitalization, Business Model Innovation, and Sustainable Industry: Past Achievements and Future Promises," Sustainability, vol. 11, no. 391, p. 1., 2019.

[2] M.J. Davenport, and S. Kudyba, "Designing and dveloping analytics-based data products," MIT Sloan Management Review, vol. 58, no.1, 82-89, 2016.

[3] E. Brynjolfsson, and A., McAfee, "The Digitization of Just About Everything, Case Study," Harvard Business Review, https://hbr.org/product/the-digitization-of-just-about-everything/ROT275-PDF-ENG, 2015.

[4] D. Icholas and N.D. Evans, "Managing Innovation & Disruptive Technology," CIO, OCTOBER 01, 2015., https://www.cio.com/article/2988012/6-steps-for-digital-transformation.html

[5] B.G. Raggad, Information Security Management: Concepts and Practice, CRC Press, New York, 2010.

[6] R. Bukht, and R. Heeks, "Defining, Conceptualizing and Measuring the Digital Economy," Development Informatics, Working Series, No. 68, 2019.

[7] R. Rikowski, "Digitisation Perspectives," In Educational Futures Rethinking Theory And Practice, Sense Publishers , Volume 46, 2011, https://www.sensepublishers.com/media/263-digitisationperspectives.pdf

[8] Gartner, Insights From the 2017 CIO Agenda Report:Seize the Digital Ecosystem Opportunity, Gartner, 2017, https://www.gartner.com/imagesrv/cio/pdf/Gartner_CIO_Agenda_2017.pdf.

[9] C. K. Davis, "Beyond data and analysis," Communations of the ACM, vol. 57, no, 6, 2014, pp 39-41.

[10] Ekbia, et al., "Big data, bigger dilemmas: a critical review," Journal of teh Association of Information Science Technology, vol. 66, no. 8, pp 1523-1545, 2018.
[11] Abbasi et al., "Big data research in information systems: toward an inclusive research agenda,", Journal of Association of Information Systems, vol. 17, no. 2, 2016, pp i-xxxii.
[12] S. Vijayarani and S Sharmila, "RESEARCH IN BIG DATA – AN OVERVIEW" Informatics Engineering, an International Journal, vol. 4, no. 3, September 2016.
[13] IMF, Measuring The Digital Economy, Feb 2018, https://www.imf.org/~/media/Files/Publications/PP/2018/022818MeasuringDigitalEconomy.ashx.
[14] Seddon, et al., "How does business analytics contribute to business value?" Information Systems Journal, vol. 27, no. 3, pp 237-269, 2017.
[15] C. Dede, Christopher, "Next steps for "Big Data" in education: Utilizing data-intensive research," Educational Technology, vol. 56, no. 2, pp. 37-42. 2016, http://nrs.harvard.edu/urn-
[16] 3:HUL.InstRepos:28265473
[17] I-Scoop, Digital transformation: online guide to digital business transformation, https://www.iscoop.eu/digital-transformation/, 2018.
[18] G.J. Miller, "The Political Evolution Of Principal-Agent Models," Annual Revue of Political Science, vol. 8, pp. 203–225, March 2005.
[19] B. Gärtner, and M. Hiebl, Issues with Big Data. Chapter 13 in The Routledge Companion to Accounting Information Systems, 2017.
[20] EESC, The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context, European Economic and Social Committee, 2016, https://www.eesc.europa.eu/resources/docs/qe-02-17-159-en-n.pdf.
[21] M. Vozábal, Tools and Methods for Big Data Analysis, Master Thesis, Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, June 2016.
[22] R.H. Sprague and E.D. Carlson, Building Efective Decision Support Systems. Englewood Clifs, NJ: Prentice-Hall, 1982.
[23] Matturdi et al., "Big Data security and privacy: A review," China Communications, vol. 11, no. 14, pp. 135-145, 2014.
[24] Kaur et al., "Review: An evaluation of major threats in cloud computing associated with big data", Big Data Analysis (ICBDA) 2017 IEEE 2nd International Conference on, pp. 368-372, 2017.
[25] X. Wang, T. Laurence T. Yang, H. Liu, and M. Jamal Deen, "A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives", Big Data IEEE Transactions on, vol. 4, no. 3, pp. 325-340, 2018.
[26] A. Rahaman, S. Rajesh, and G. Rani, "Challenging tools on Research Issues in Big Data Analytics," International Journal of Engineering Development and Research, vol. 6, no. 1, 2018, pp. 637-644.
[27] K. Singhal, "Introduction to the Special Issue on Perspectives on Big Data," Production and Operations Management, vol. 27, no. 9, September 2018, pp. 1631.
[28] W.A. Gunther, M.H.R. Mehrizi, M. Huysman, and F. Feldberg, Debating big data: A literature review on realizing value from big data, The J of Strategic Information Systems, Vol 26, Issue 3, Sep 2017.
[29] L.J. Savage, The Foundations of Statistics, Dover Publications, New York, 1972.
[30] B. PUZA, BAYESIAN METHODS for Statistical Analysis, ANU eView, The Australian National University, Australia, 2015.
[31] D. Dubois and H. Prade, "Representation and combination of uncertainty with belief functions and possibility measures," Computational Intelligence, vol. 4, pp. 244-264, 1988.
[32] G. Shafer, "Dempster's rule of combination," International Journal of Approximate Reasoning, vol.
[33] 79, pp
[34] D.W. North, "A Tutorial Introduction to Decision Theory" IEEE Transactions on Systems Science And Cybernetics, vol. ssc-4, no. 3, September 1968.
[35] G. Shafer, A Mathematical Theory of Evidence. Princeton University Press, 1976.
[36] G. Shafer, Glenn (1990). Perspectives on the theory and practice of belief functions. International Journal of Approximate Reasoning, vol. 3 pp. 1-40, 1990.
[37] P. Smets, and R. Kennes,The transferable belief model. Artificial Intelligence, vol. 66, pp. 191–234, 1994.
[38] K. Sultan1, U. Ruhi, and R. Lakhani, Conceptualizing Blockchains: Characteristics & Applications, 11th IADIS International Conference Information Systems 2018, pp. 49-57.
[39] Gervais et al., "On the security and performance of proof of work blockchains," In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016,
[40] G. Karame, and S. Capkun, "Blockchain Security and Privacy," Article in IEEE Security and Privacy Magazine, vol. 16, no. 4, July 2018,
[41] M. Singh, M. N. Halgamuge, G. Ekici, and C. S. Jayasekara, "A Review on Security and Privacy
[42] Challenges of Big Data", Lecture Notes on Data Engineering and Communications Technologies
[43] Cognitive Computing for Big Data Systems Over IoT, Frameworks, Tools and Applications, Springer Volume 14, Chapter 8, ISBN: 978-3-319-70687-0, pp 175-200, January 201
[44] S. Kalid , A. Syed, A. Mohammad, and M. N. Halgamuge, "Big-Data NoSQL Databases:
[45] Comparison and Analysis of "Big-Table", "DynamoDB", and "Cassandra", IEEE 2nd International Conference on Big Data Analysis (ICBDA'17), Beijing, China, pp 89-93, 10-12 March 2017
[46] V. Vargas and M. N. Halgamuge, "Performance Evaluation of Big Data Business Intelligence Open
[47] Source Tools: Pentaho and Jaspersoft", Internet of Things and Big Data Analytics toward Next
[48] Generation Intelligence, Springer, ISBN: 978-3-319-60434-3, Chapter 6, pp 147-176, February 201
[49] Deloittes, Modern Business Intelligence: The Path to Big Data Analytics April 2018, https://www2.deloitte.com/content/dam/Deloitte/tr/Documents/deloitteanalytics/Modern%20Business%20Intelligence.pdf
[50] T. N. Hewage, M. N. Halgamuge, A. Syed, and C. Bellamy, "Review: Big data techniques of
[51] Google, Amazon, Facebook and Twitter", Journal of Communications, Volume 13, No 2, pp 94-100, Feb 201