Journal of Big Data

RESEARCH
Open Access

Addressing big data variety using an automated approach for data characterization

Georgios Vranopoulos*, Nathan Clarke and Shirley Atkinson

*Correspondence:
georgios.vranopoulos@plymouth.ac.uk
School of Engineering, Computing & Mathematics, University of Plymouth, Plymouth, UK

Abstract

The creation of new knowledge from manipulating and analysing existing knowledge is one of the primary objectives of any cognitive system. Most of the effort on Big Data research has been focussed upon *Volume* and *Velocity*, while *Variety*, "the ugly duckling" of Big Data, is often neglected and difficult to solve. A principal challenge with *Variety* is being able to understand and comprehend the data. This paper proposes and evaluates an automated approach for metadata identification and enrichment in describing Big Data. The paper focuses on the use of self-learning systems that will enable automatic compliance of data against regulatory requirements along with the capability of generating valuable and readily usable metadata towards data classification. Two experiments towards data confidentiality and data identification were conducted in evaluating the feasibility of the approach. The focus of the experiments was to confirm that repetitive manual tasks can be automated, thus reducing the focus of a Data Scientist on data identification and thereby providing more focus towards the extraction and analysis of the data itself. The origin of the datasets used were Private/Business and Public/Governmental and exhibited diverse characteristics in relation to the number of files and size of the files. The experimental work confirmed that: (a) the use of algorithmic techniques attributed to the substantial decrease in false positives regarding the identification of confidential information; (b) evidence that the use of a fraction of a data set along with statistical analysis and supervised learning is sufficient in identifying the structure of information within it. With this approach, the issues of understanding the nature of data can be mitigated, enabling a greater focus on meaningful interpretation of the heterogeneous data.

Keywords: Big Data, Variety, Data characterization, Data origination, Data Format, Data confidentiality, Delimiter determination, Metadata, Contextual integrity

Introduction

Laney, often referred to as the father of Big Data, had introduced three dimensions that characterise Big Data which have become the industry standard for defining Big Data [1, 2]. *Volume* is the first and refers to the amount of data created and stored in the digital universe [3]. The second is *Velocity*, which in Big Data environments refers to the speed at which data changes. Last but not least is *Variety*; this characteristic has to do

Springer Open

with the data itself and its manifestations. Sensors, Internet of Things (IoT), database records, video and audio have different formats and standards. Further to these dimensions which are mainly technical in nature, they have been subsequently augmented with some additional considerations. The lack of the required governance and homogeneity is identified by *Veracity*. *Validity* is concerned with correctness and accuracy and *Volatility* refers to the how long the data is valid for and for how long it should be stored for [3, 4]. Business is concerned with income and realization of competitive advantages; as a result, *Value* is another dimension that poses a significant challenge. These business dimensions have become essential since Big Data started gaining growing acceptance towards the data-driven decision (DDD) making approach [5].

Although there is arguably adequate literature and research on *Volume* and *Velocity*, research on *Variety* revealed a different trend [6–9]. One of the biggest challenges for data scientists is to cope with the heterogeneity of the data in identifying and understating the datasets [10]. There are no technological tools that specifically deal with the issue of *Variety* and the proliferation of data from many sources, internal and external, public and private and in numerous formats further increases the challenge [11]. The sheer complexity of data sources, accurate and inaccurate data mixed together with multiple formats and units of measurement pose a significant risk in handling data [12]. To mitigate the risks, companies spend 70%-80% of their time and effort understating the data structures and preparing the data rather than actually interacting with the data in generating meaningful insights [13, 14]. The task of contextual placement and definition are tasks assigned to humans; thus, *Variety* has remained resilient to software solutions [15]. Process automation could reduce the dependence on humans and their respective skills, which will, in turn, enable better business penetration and adoption [12, 16].

In Bloomberg's survey for the organizational challenges in adopting Analytics, it was identified that *Variety* had the leading position [17]. One of the factors impacting business adoption is the Total Cost of Ownership (TCO) where hardware, maintenance, energy, and software licences can be high [18]. Part of the TCO is also the cost to be paid on services in addressing *Variety,* which for any Big Data project will be the "key cost element" [12]. The arrival to erroneous conclusions due to data *Variety* is also a significant concern [19]. Organizations in realising Return on Investment (RoI) through Big Data [18] have understood the value of minimizing the effect, and in fact, 74% of businesses would like to harmonize their data [11]. Since *Variety* is "more difficult" to measure and quantify [20], no precise quantitative illustration of the lost RoI attributed to *Variety*, has been identified. Big Data projects RoI is approximately 55cents per dollar spent in contrast to any other project, which would reach $3–$4 [21]. This gap in RoI can designate the scale of the challenge from the lack of automation in a monetary manner. The challenge is further increased once the very high scale of profit realization from Big Data analytics implementation is taken into consideration, example is the case of Wallmart retailer with an additional 150 million of revenue post their analytics project [22].

This research aims to target the issue of heterogeneity and complexity of data by proposing and evaluating techniques to automate this identification process. The paper focusses upon two contributions, the automated identification of confidential data through the application of additional features, referred to as *Booster Metrics*, and a wider dataset identification approach through machine learning. Both contributions are supported through empirical experiments to investigate the impact and performance of these approaches. By doing so, the study will confirm the plausibility of an autoimated solution to *Variety* that can become an enabler for business adoption of Big Data. In doing so the techniques will minimise the level of human-analyst interaction towards data classification and private data identification [23].

Background research

In trying to identify the current state of the art for *Variety* (with respect to Big Data), a broad review of the prior art was undertaken acress a range of repositories (ScienceDirect, Springer Link, IEEE Xplore, and Google Scholar). The methodology used included multiple rounds of research, evaluation and classification of Big Data reference material. The keywords included in the search were "Big Data", "Volume", "Velocity", "Variety", "Technological Advancements", "Data Lake", "analytics", and combinations of them. The documents were initially evaluated based on the search engine results. The second and third screening extended to the abstract, introduction and conclusions along with a complete document review, respectively. Based on these iterations, a set of 414 articles were identified and categorized as shown in Table 1.

Notable within this analysis, whilst a good number of research studies and publications have focused on big data, a relatively small number have specifically focused upon *Variety*—even though the issues with respect to *Variety* have been well established. Unfortunately, an analysis of the thirteen papers specifically on *Variety* revealed that all were focused on confirming the challenge rather than proposing or validating solutions. In ensuring the review captured all relevant research, it was decided to undertake further searches with a broadening set of keywords—including related terms, "Heterogeneous Data", "Heterogeneity", "Data discrepancy", "NoSQL" and combinations thereof. From the new result set, one hundred and seventy (170) studies were identified as initially relevant and further reviewed. This process resulted in twelve papers being identified as relevant, focusing on challenges originating from automatic schema management, federated data sets, missing and erroneous data and the interdependence of heterogeneity to volume. Since most of the searched and reviewed documents of the new set, were primarily related to specific case studies on datasets pertaining to Biology (proteins), Genetics (Gene sequences), Medical (Cancer & C-Scan Recognition),

Table 1 Big data references classification

Subject	Number of resources
Big data in general	152
Business/corporate	37
RDBMS	40
Legal, social, ethical	23
Metadata	56
Productivity tools	42
Volume	23
Velocity	14
Variety	13
Heterogeneous data (in Big Data)	14

Disaster (floods, droughts, earthquakes), it was decided to further limit the search by applying filtering terms including "RDBMS", "Big Data" and "Database". From the new pool of references, a hundred (100) were reviewed, but only two met the criteria since most of the references were focusing on storage implications, NoSQL Databases, filtering and contrasting rather than the actual *Variety* challenges like integration and unification. All reviewed papers, the thirteen directly related to *Variety* and fourteen indirectly related by means of heterogeneity, were utilized in setting the bases of understanding the current art related to *Variety*.

Timeline analysis of the references denotes the researchers have mainly identified the challenges and offer limited solutions if any. Kumaridentified that the cost to be paid on services in addressing *Variety* in any Big Data project would be the "key cost element" and suggested that internal resources should be used in lowering the costs [12]. Kimura suggested 74% of businesses would like to harmonize their data [11]. Lennard, Shacklett and Brown all independently agree that specialized personnel have to be utilized and that the respective specialization is not present in many organizations [6, 8, 23]. These studies help to substantiate that human capital is a critical and scarce resource. *Variety* being difficult and costly is not the only challenge, errors and controversial results are also challenges [3], and organizations have understood the value of minimizing the effect.

Given the lack of relevant literature within *Variety*, it was deemed prudent to broaden the problem and search criteria. In investigating the issue of data heterogeneity, a wider body of literature was identified. The topic dates back to the middle 1970s [24]. Initially, methodologies like Common Object Request Broker Architecture (CORBA), Distributed Component Object Model (DCOM) and Electronic Data Interchange EDI were devised and employed, but they all were focused on defining a clear and rigid communication framework that in essence will prevent *Variety* by abolishing any diversification [24]. These approaches were arguably sufficient for the early years, but when the World Wide Web came into existence and was eventually extensively adopted, such methodologies had to be abandoned since it was impossible to enforce such rigid communication protocols. With the exponential increase of the generated data, it became apparent that standardization of data after it is generated and not beforehand, as prior methodologies did, was required. Towards this post-publication data integration approach related to diverse data structures, several frameworks, like ARTEMIS-MOMS, Cupid, "similarity-flooding", and iMAP, suggested to have reconciliation at the target schemas and have federated queries executed across all of them [14, 25, 26]. These methodologies/techniques utilize linguistic matching, translation into graphs and semantic matching. However, as Banek et al. highlight, all of these techniques will produce candidates that will have to be confirmed or will have to be "taught" using examples (e.g. Neural Network training) Target federated queries would require two levels of human intervention. First level would include importing the data thus understanding the structures whilst at the second level the schema reconciliation will have to be confirmed based on the automated identification.

Instead of federated queries, the concept of an aggregated model/meta-model can be used [26]. The model will contain the information in linking the metadata of the distinct models into an aggregate model, which will be used in the reconciliation or definition of the heterogeneous schemas/information storage. The Heterogeneous Data Quality Methodology (HDQM) presents an approach where the focal element is the Conceptual Entity, an abstraction of any single phenomenon of the real-life instantiation of an entity of interest [27]. If the concept is adapted to Big Data, it could address some of the *Variety* challenges. Automating the identification of such an aggregation or abstraction layer in the multiverse of Big Data could become a challenge by itself especially when *Velocity* is taken into consideration.

Wang explicitly refers to heterogeneous of big data, but the research is mainly confined to identifying/raising the challenges instead of proposing alternatives in overcoming them. Of interest is the new term *Data Swamp* which is used to identify ungoverned Data Lakes where data is dumped without any metadata that will lead to confusion and limited usage since the semantics and structure of the data will be unknown [28]. The data swamp can have devastating implications for an organization due to regulatory implications of data confidentiality. In addition, fuzzy data sets are of little use to data scientists since they will have to first format and break down the data before using them.

The prior art indicates the absence of extended research on *Variety*. The research was extended to the similar concept of heterogeneity in identifying means and methods of addressing the challenge. Most of the respective techniques are human driven or have little applicability on big data. The term *Data Swamp* is indicative of the *Variety* challenge since it will transform a corporate data lake to an unsalable conglomeration of data sets.

Addressing the challenge

The objective of this paper is to investigate the feasibility of transforming the repetitive work required to identify and ingest the data from a labour-intensive human effort to a software-driven automated process. The ingestion processes are depicted in a high-level diagram (see Fig. 1), based upon which the automation capabilities will be subsequently discussed. These tasks are prerequisites before a Data Scientist can start analysing the data to identify new knowledgeand obtain value from big data.

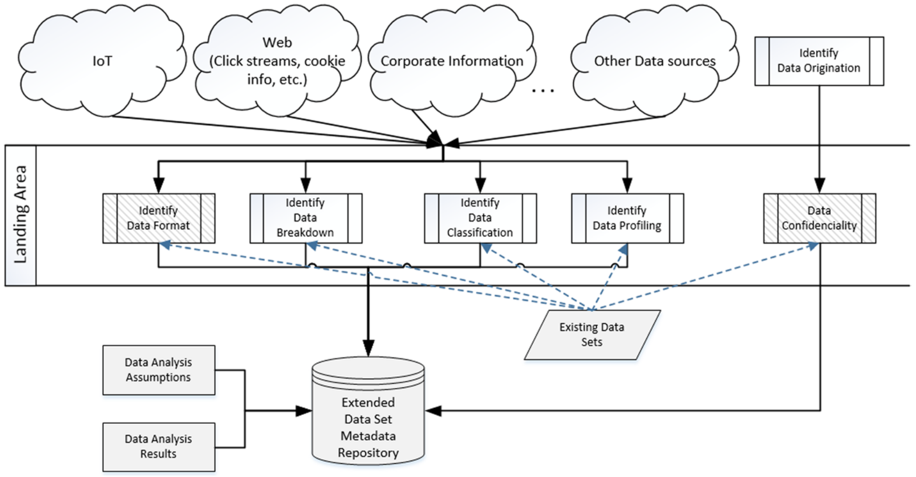


Fig. 1 Proposed model for big data variety

Data origination—confidentiality

Following the ingestion journey of data, the first step would be to understand where the data is coming from. This step, named *Data Origination*, has little room for automation since it is outside the realm of the actual ingestion but is essential since it will provide contextual information. Contextual information would include whether the data set origination is public or private, the originating industry or discipline, the collection methods, which in turn identifies the second milestone in ingestion, Data Confidentiality. Information is increasingly being regulated in multiple sectors with acts including the Personally Identifiable Information (PII), Public-Sector Information (PSI) Directive, General Data Protection Regulation (GDPR) law and Payment Card Industry (PCI) Security Standards along with anonymization standards like European Medicines Agency Policy 0070 or the Health Insurance Portability and Accountability Act. These regulatory acts can have an immediate financial impact in the form of fines that can reach 4% of the annual global turnover. An example is Facebook, with a confirmed $5 billion fine and another €56 million potential fines depending on the outcome of 11 ongoing GDPR investigations [29, 30]. Due to the Data Loss Prevention (DLP) risk, data confidentiality is a candidate for automation, but identifying content in Big Data can be a challenge due to *Volume*, in terms of source datasets but most importantly in relation to results which tend to be vast especially when false positives are present [31].

Empirical evidence suggests that the existing automated systems, which rely heavily on Regular Expressions (RegEx), will produce a high number of false positives [32, 33]. For example, if a pattern of a Credit Card is used, then any number of fifteen or sixteen digits will be identified as a possible candidate; samples are available in Table 2 [34–39].

The same would apply for PII governed elements such as Identification Card numbers, passport numbers or social security numbers, e-mail, physical address, internet protocol (IP) address, media access control (MAC) Address. For more generic data elements like account numbers, International Bank Account Number (IBAN) and

Table 2 Credit/debit card and personally identifiable information sample RegEx

Confidential data	Regular expression
mastercard	(?:5[1-5] [0-9]{2}|222[1-9]|22[3-9] [0-9]|2[3-6] [0-9]{2}|27[01][0-9]|2720)[0-9]{12}
VISA	4[0-9]{12}(?:[0-9]{3})?
AMERICAN EXPRESS	3[47][0-9]{13}
Diners Club	3(?:0[0-5]|[68][0-9])[0-9]{11}
Gulf Countries Civil ID	\d{1} (?!00)\d{2} (?!00)\d{2} (?!00)\d{2} (?!0000)\d{4}
Greek Civil ID	[A-Ω]{1,2}[0-9]{6}
International Passport	[A-Z0-9&lt;]{9}[0-9]{1}[A-Z]{3}[0-9]{7}[A-Z]{1}[0-9]{7}[A-Z0-9&lt;]{14}[0-9]{2}
IBAN	[a-zA-Z]{2}[0-9]{2}[a-zA-Z0-9]{4}[0-9]{7}([a-zA-Z0-9]?){0,16}
eMail	(?:[a-z0-9!#$%&amp;'*+/=?^_`{\|}~-]+(?:\.[a-z0-9!#$%&amp;'*+/=?^_`{|}~-]+)*|\"(?:[\×01-\×08\x0b\x0c\x0e-\x1f\×21\×23-\x5b\x5d-\x7f]|\\[\×01-\×09\x0b\x0c\x0e-\x7f])*\")@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?|[a-z0-9-]*[a-z0-9]:(?:[\×01-\×08\x0b\x0c\x0e-\x1f\×21-\x5a\×53-\x7f]|\\[\×01-\×09\x0b\x0c\x0e-\x7f])+)\])
MAC Address	([0-9A-Fa-f]{2}[:-]){5}([0-9A-Fa-f]{2})

Table 3 Classification to "booster metrics" association

Classification	Search Condition	Booster Metrics
Card	List of RegEx expressions to get the initial data match	Linguistic boundary characters, e.g. space, coma, quotation Luhn algorithm, for check digit verification institutional bank identification numbers (BINs)
Lists	List of RegEx expressions to get the initial data match	Monitor terms proximity. The distance of the occurrence with words like password, account, card, credit, id etc. is calculated
Absolute XML	–	List of specific xml tags e.g. < Civilld > ID123456 < /Civilld >
Relative XML	–	List of XML tags containing terms < *Passport* > where * indicates any number of any character

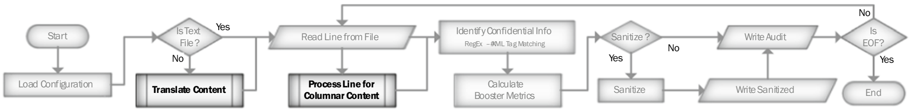


Fig. 2 Confidentiality process flow

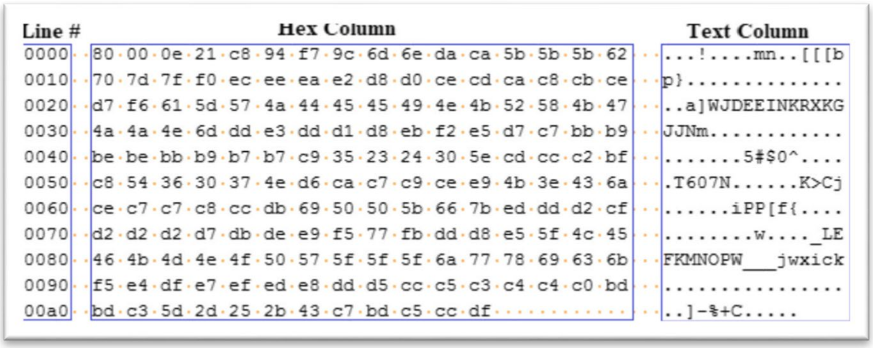
customer numbers which tend to be sequential numbers, will produce even more false positives since any number would qualify.

A confidence level is proposed in reducing the number of false positives and providing an accurate metric on the content of confidential data. In calculating the confidence level, the existing RegEx methodology will be extended with multiple other metrics, which will serve as “booster metrics” in increasing the confidence of the match. Metrics like Soundex, proximity and structural confirmation/check digit calculations will be used (as illustrated in Table 3).

In addressing the false positives impediment in a structured manner, the system utilises four main classifications to accommodate the different types of confidential information to be identified.

The system was extended with a pre-processor (Translate Content) to ensure the textual nature of the data fed into the system. Co-processors (Process Line for Columnar Content) were introduced to cater for different/multiple file structures. In Fig. 2, the pre-processor and a co-processor are shown as part of the entire system process.

For the pre-processor, the Apache Tika project was used to ensure that any nontext feed or file (e.g. pdf, MsWord, MsExcel) were converted into a text format before being sent into the matching subsystem [40]. In order to manage multiple file formats co-processors /alternative readers were introduced on top of the traditional line reader, where block reading was enabled since certain file exhibit a columnar structure, see Fig. 3. In this way, characters from line 1 from column 5 to 30 were merged with characters from line 2 from column 5 to 30 in making a logical sequence. This kind of formatting, shown in Fig. 3, is most common in trace files. There are two content columns in addition to the line number column, one with hex and one with text and in documents or pixel-perfect reports that utilize columnar writing styles. By incorporating different types of readers through parameterization, the system caters for the variety of file formats.



```
0000  80 00 0e 21 c8 94 f7 9c 6d 6e da ca 5b 5b 5b 62  ...!....mn..[[[b
0010  70 7d 7f f0 ec ee ea e2 d8 d0 ce cd ca c8 cb ce  p}..............
0020  d7 f6 61 5d 57 4a 44 45 45 49 4e 4b 52 58 4b 47  ..a]WJDEEINKRXKG
0030  4a 4a 4e 6d dd e3 dd d1 d8 eb f2 e5 d7 c7 bb b9  JJNm............
0040  be be bb b9 b7 b7 c9 35 23 24 30 5e cd cc c2 bf  .......5#$0^....
0050  c8 54 36 30 37 4e d6 ca c7 c9 ce e9 4b 3e 43 6a  .T607N......K>Cj
0060  ce c7 c7 c8 cc db 69 50 50 5b 66 7b ed dd d2 cf  ......iPP[f{....
0070  d2 d2 d2 d7 db de e9 f5 77 fb dd d8 e5 5f 4c 45  ........w...._LE
0080  46 4b 4d 4e 4f 50 57 5f 5f 5f 6a 77 78 69 63 6b  FKMNOPW___jwxick
0090  f5 e4 df e7 ef ed e8 dd d5 cc c5 c3 c4 c4 c0 bd  ................
00a0  bd c3 5d 2d 25 2b 43 c7 bd c5 cc df              ..]-%+C.....
```

Fig. 3 Columnar log file sample (3 columns)

It is anticipated that through the use of confidence levels and the inclusion of booster metricsit it will be possible to more reliable identify confidential data elements in an automated fashion.

Data format—delimiter determination

Having cleared the confidentiality barrier and being compliant, the data should be broken down and imported into a landing area. This step is the *Data Format Identification* followed by the *Data Breakdown* phase. In addressing the heterogeneity challenge, the proposed approach is focused upon viability and accuracy. As *Volume* can be an impediment when addressing *Variety*—as analysing complete data sets can often be impossible/infeasible, the proposed approach seeks to achieve format identification through an analysis of the delimiters present.

A file can be delimited with any character or sequence of characters. Common delimiters across the industry are comma, semicolon (comma-delimited files) or tab (tab-delimited files). This poses a challenge since these characters are also used as common characters in many digital forms, including documents, textual data, etc. In an attempt to incorporate this manifestation of *Variety* where the same character can have different contextual importance, the possible file delimiters were given different weights given their probable occurrence as punctuation marks for ordinary text. For instance, a comma would have a higher possibility of occurring as a punctuation mark than tilde. As a result, tilde would have a higher weight attributed to it. The rationale behind this is to increase the significance of a reoccurring rare character combination as a separator against common character separators. Detailed information on the delimiters and associated weights attributed are provided in Section 4.2.

The system used multiple independent components, as seen in Fig. 4, feeding information to each other. Pre-processors would be used in harmonizing the data and performing the initial set of classifications:

(a) Identify the nature of the file (Text Vs Binary) using MIME.
(b) Identify the files encoding in order for the system to be able to identify the content language and accordingly configure the subsequent input readers (e.g. an ANSI reader will not be able to load data for UTF encoded files correctly) for subsequent parsing

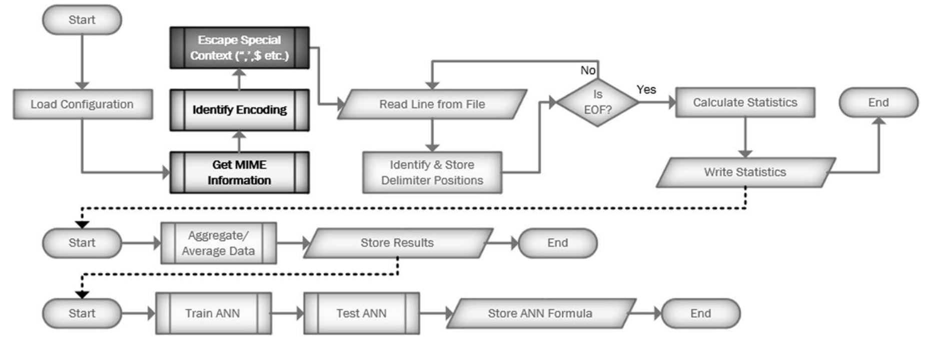


Fig. 4 Delimiter determination process flow

(c) Removal of special characters where quotations, JSON and XML notation characters were nullified since they tend to “break” the parsers,
(d) Adjust for specific set characteristics, like long paraphrases embedded as files, lists and other structures.

Based on preliminary results of the experiment, some inconsistency was observed, and after investigation, a further component was introduced into the framework. It became apparent that there was a difference in one of the experimental datasets that exhibited low accuracy in predicting the correct file delimiter. *Variety* was in play; something was different with this set that was not taken into consideration as an initial variable in the definition of the experiment. The pre-processor labelled “*Escape Special Context*” was incorporated as a compensating control for the following characteristics that attributed to the:

- The set utilised multiple delimiters for segmentation. Although the file, for instance, was delimited with a comma, one of the fields had in it multiple values delimited with semicolons. As a result, both delimiters exhibited high degrees of conformity.
- The records were extended into multiple lines while being enclosed in double-quotes.
- Many of the fields had long text, document sized, that had a very high degree of variance in length stretching from a couple of lines to hundreds of lines.

The following set of primary metrics were mapped and calculated to map the characteristics of the delimiters’ properties identified in each file, or each line read:

- The number of lines read from the file.
- The number of lines having the same number of columns.
- Min, Max and Mean of Standard Deviation for the position of the delimiter across all lines read per position.
- Min, Max and Mean of Coefficient of Variation for the delimiter position across all lines read per position.
- Min, Max and Mean of Standard Deviation for the relative position (distance) of the delimiter from the previous delimiter across all lines read per position.

- Min, Max and Mean of Coefficient of Variation for the relative position (distance) of the delimited from the previous delimiter across all lines read per position.

The primary set of metrics was averaged out in aggregating the data from the line level to the file level, taking into consideration each delimiter found. The derived set of metrics include:

- The number of identified delimiters in the file.
- Whether the number of columns is consistent across all lines read (Boolean metric).
- The average Standard Deviation for the absolute position of the delimiter.
- The average Coefficient of Variation for the absolute position of the delimiter.
- The average Standard Deviation for the relative position of the delimiter.
- The average Coefficient of Variation for the relative position of the delimiter.

These metrics are subsequently used as the input parameters into a supervised feed-forward multi-layer perception (MLP) neural network to classify the file format without any human input [41]. The neural network will be used to verify, based on the delimiter characteristics set by the calculated metrics, if the file was delimited with the respective delimiter. In this way, the identification process will be automated and shall require no human intervention or supervision.

Through automating the task, the required scope of Data Scientists services will be limited to the most important task of interpreting the data instead of identifying it. This synergy will result in multiple perspectives benefit (a) adaptability and responsiveness to changing business environment since efforts required in incorporating new data will be less [42], (b) profitability, by allocating internal or external resources towards data mining, instead of data identification, activities which will identify a comparative advantage [43], (c) lower TCO and in turn achieving higher business adoption rates [12].

Experimental methodology

A series of experiments have been conducted in order to evaluate the proposed approaches to data origination and data format identification:

- The main objective for Data Origination—Confidentiality was to confirm that the high number of false positives generated with conventional techniques like RegEx can be minimized, thus making the identification more accurate for further actions, e.g. masking. For the Proof of Concept (PoC), datasets from application logs, audit logs and network captures were used. The respective set was selected since there is a high probability of such files being shared by the organization with vendors and other external entities.
- Regarding Data Format—Delimiter Determination, the objective was to confirm the size of data, variables, and viability of a solution that will identify a file's delimiter and thereby enable identification of the data. In the second experiment, statistical data derived from an analysis of big data are used as an input in a neural network to enable identification.

In order to evaluate the approach, a variety of datasets were identified whose characteristics (origin, number of files and size) varied in order to investigate the impact this could have upon performance (as illustrated in Table 4).

A decision was taken to incorporate a proprietary dataset in addition to public since public data sets have typically been worked on and reviewed thus tend to be already more structured and sanitized. Proprietary datasets tend to be present unique characteristics since they are developed to serve a specific need rather than being disseminated. All proprietary dataset originating from the financial sector have been sourced for a nonproduction environment thus contain no sensitive information. In this way, the experiment will incorporate a broader spectrum of data anomalies and definitions. Volume and object size were introduced in incorporating the *Volume* and *Velocity* concepts where the behaviour could differ in case of a high number of small data elements in contrast to a low number of instantiations with a large data element. These dimensions will provide the processing time and computing resources requirements to be evaluated during the experiment.

In implementing and evaluating the experiments, a desktop PC with Inter® Core™ i7-6700 CPU @ 3.40 GHz and memory of 16 GB was utilized. The system and application software used would include Windows 7 (64-bit), Ms Office (2010, 2013, 2016), MatLab (2016b), Eclipse (Mars.2 – Java 1.8).

Data origination—confidentiality

In the Data Origination stage, we are interested in identifying the source and content of the data so that the accidental use of private, confidential and regulated, data is minimized if not abolished. Along with standard user-generated files like word processor files, spreadsheets, emails, logs were utilized to investigate and validate the proposed approach. Logs were included since they are widely used in the industry for security analysis, click-stream insights and other analytics insights [44, 45].

The four major classifications, along with lists and parameters (e.g. RegEx, bin numbers, associated "booster" add-on percentages, sanitization option), were stored in an XML configuration file, making the system highly flexible. In this manner, the system could quickly be adapted to any required changes or extensions regarding the required configurations. Part of the system initialization is to set the "booster metrics" configuration along with the location to be searched and the flag of whether the data should be sanitized. Sanitization depends on the classification configuration where the technique to be used is defined as mask, hash, encrypt or replace/truncate the identified

Table 4 Datasets volume & origin characteristics

Experiment	Dataset	Origin	Number of Files	Disk Size (GB)
Confidentiality	Mobile Banking logs	Proprietary	4	0.50
Confidentiality	Loan Origination System Logs	Proprietary	3	1.10
Confidentiality	Network Trace	Proprietary	43	3.98
Delimiter Identification	Banking Set (ODS)	Proprietary	8,605	15.60
Delimiter Identification	National Climatic Data Center (NCDC)	Public	14,030	9.40
Delimiter Identification	Center for Disease Control and Prevention (CDC)	Public	920	12.80

information. For each classification, the percentage contribution of each booster metric is configured, and the confidence level is set to a percentage greater or equal to the RegEx identification as shown in the examples of Table 5—for respective RegEx, refer to Table 2.

For ease of use, the system will recursively search in any path provided for all available files in any folder depth, thus allowing for one or multiple sets processing in a single execution. By adding the contribution of each metric, the system will arrive at a total confidence percentage per occurrence. If the respective percentage of the instantiation exceeds the defined high watermark for confidence level, the entry will be considered confidential. The confidence levels are parameterized in the system and should be higher than the initial contributor, in our case RegEx, and lower than the sum of contributors. In filtering out false positives, the level can be increased depending on how many contributors the analyst would like to be taken into consideration. Multiple post-processors are used depending on the configuration to (a) sanitize the data by applying the requested algorithm (b) remove the data for following classification occurrence so that multiple recordings of identical instantiations are not recorded, (e.g. Credit Card is not also identified as debit card) (c) record all identified values from identical occurrences of a tag value in the Relative XML class to be forwarded to an external AI system for identifying new RegEx patterns.

All the system results are recorded in a log file so that the lineage and detailed results of identified and qualified entries are available for review. If the parameterization of the system indicated the requirement to sanitize the file, a new file would be created so that the original file is preserved whilst the new one can be released for future usage. In this

Table 5 Sample metrics definition

Occurrence	Classification	Metrics definition	Value/add-on contribution to confidence level
Credit Cards	Card	RegEx Identified	40%
		Linguistic boundary	20%
		No linguistic boundary	10%
		Luhn algorithm	40%
		Exists in institutional BINs	5%
		Sanitization method	Masking (first six and last three chars)
		Confidence Level	60%
PII	Lists	RegEx Identified	40%
		Linguistic boundary	20%
		No linguistic boundary	10%
		Proximity	10%
		Sanitization method	Hash
		Confidence Level	50%
PII	Absolute XML	RegEx Identified e.g. (<CIVIL_ID>)	100%
		Sanitization method	Truncate
		Confidence Level	50%
PII	Relative XML	RegEx Identified e.g. (<*ID*>)	50%
		Sanitization method	Truncate
		Confidence Level	50%

way, Data Scientists can tune the system by altering the contribution percentage and fine-tune the high-water marks per dataset and containing entity.

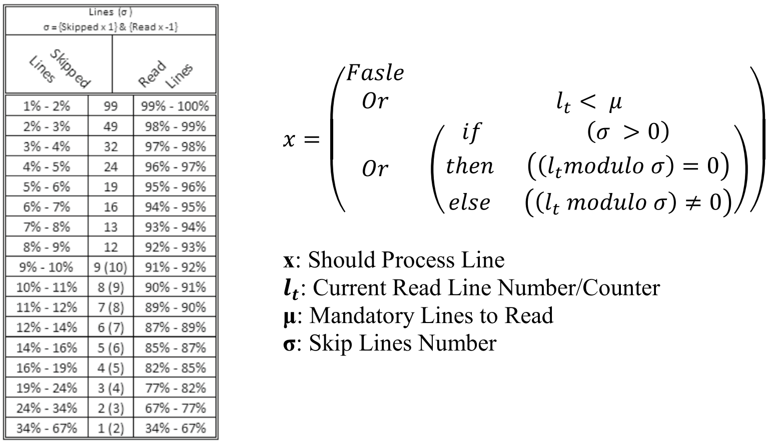
Data format—delimiter determination

This PoC was used to confirm that an automated solution for the data format is feasible. The experiment would have to identify the quantity of data that should be processed to attain reasonable confidence of the data se. In attaining a uniform read pattern across the data set, the system will have to process from all parts of the file instead of the common practice to process only the beginning of the file. Microsoft tools like MS Access, MS Excel and SSIS read a chunk, depending on the tool, that may vary from 100 to 100,000 lines, in presenting the user with a sample for import processing, which will arguably not suffice in Big Data sets. A failsafe was implemented where in the configuration, apart from the file percentage to be read; the Data Scientist can enforce a certain number of lines to be read from the beginning of the file. For the purpose of the experiment, the initial set of lines to be mandatorily read was set to 100 lines. In identifying whether a line should be read and considered in the analysis as input data, a skip line algorithm was utilized for 100 lines of input (as illustrated in Table 6).

The system should be able to accurately identify data formats irrespective of data set volume or data sets individual element size. To that end, the approach investigated data sets with different characteristics in terms of origin, file size and number of files in each set as shown in Fig. 5.

The selected datasets constitute a representative sample where all three combinations off the two-volume measures identified in addition to the origin factor: (a) Public, High number of files and Low volume (NCDC) (b) Private, Moderate number of files and Moderate volume (ODS) c) Public, Low number of files and High volume (CDC). Although the ratio of file numbers to their size indicates the number of lines per file, it is important to have a better understanding since processing many small files compared to a limited number of files with a high number of lines presents different challenges (e.g. memory constraints). The classification of lines number bands

Table 6 File lines %



Lines (σ) σ = {Skipped x 1} & {Read x -1}		
Skipped Lines		Read Lines
1% - 2%	99	99% - 100%
2% - 3%	49	98% - 99%
3% - 4%	32	97% - 98%
4% - 5%	24	96% - 97%
5% - 6%	19	95% - 96%
6% - 7%	16	94% - 95%
7% - 8%	13	93% - 94%
8% - 9%	12	92% - 93%
9% - 10%	9 (10)	91% - 92%
10% - 11%	8 (9)	90% - 91%
11% - 12%	7 (8)	89% - 90%
12% - 14%	6 (7)	87% - 89%
14% - 16%	5 (6)	85% - 87%
16% - 19%	4 (5)	82% - 85%
19% - 24%	3 (4)	77% - 82%
24% - 34%	2 (3)	67% - 77%
34% - 67%	1 (2)	34% - 67%

$$x = \left(\begin{array}{lll} Fasle & & \\ \quad Or & & l_t < \mu \\ & \left(\begin{array}{ll} if & (\sigma > 0) \\ then & \left((l_t modulo\ \sigma) = 0 \right) \\ else & \left((l_t\ modulo\ \sigma) \neq 0 \right) \end{array} \right. & \\ \quad Or & & \end{array} \right)$$

x: Should Process Line
$\boldsymbol{l_t}$: Current Read Line Number/Counter
μ: Mandatory Lines to Read
σ: Skip Lines Number

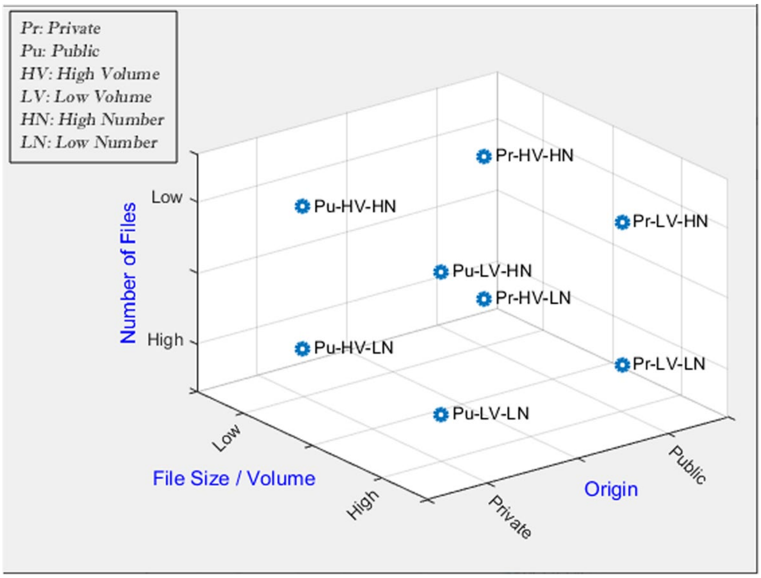


Fig. 5 Experimental data composition concerning Origin, Volume and Size

Table 7 Dataset lines no classification

Number of files	NCDC		CDC		ODS	
	# of files	% of Data set	# of files	% of Data set	# of files	% of Data set
0–100	2,742	95	84	15	3,737	63
101–500	129	4	102	18	330	
501–10,000	18	1	246	44	1,592	27
10,001–100,000			68	12	185	3
100,001–10,000,000			59	11	74	1

Table 8 Delimiters confidence level weights

Delimiter	Weight	Delimiter	Weight
Comma	1	Tilda	3
Semicolon	2	Tilda Pipe Tilda	4
Tab	2	Tilda Pipe Pipe Tilda	5

and respective counts for each dataset is available in Table 7. The presented count and percent is based on the text files identified in each set and exclude the binary files count, which eventually are excluded from the experiments.

In an attempt to auto-detect delimiters, the system was configured with different confidence levels for possible delimiters, as shown in Table 8. Weights were granted based upon the usage of the respective character(s) combination. Thus a comma "," which is extensively used, has the lowest weight compared to a " ~||~ " set which is rarely used in everyday communications.

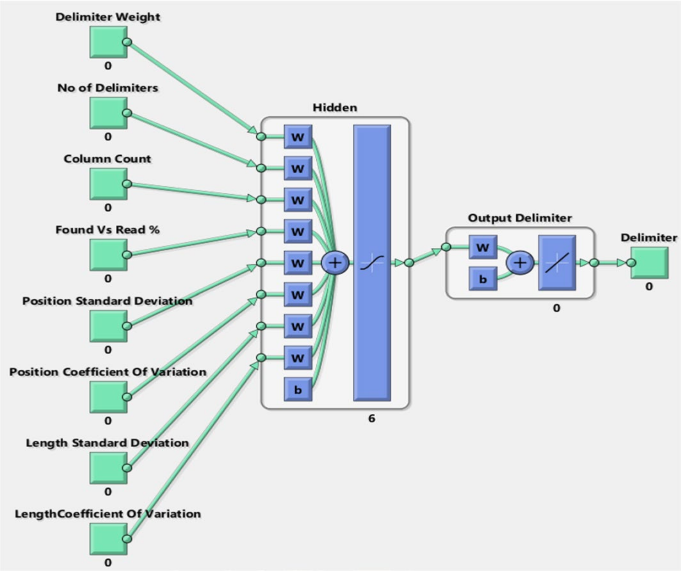

Delimiter Weight
No of Delimiters
Column Count
Found Vs Read %
Position Standard Deviation
Position Coefficient Of Variation
Length Standard Deviation
LengthCoefficient Of Variation
Hidden
W
b
6
Output Delimiter
0
Delimiter

Fig. 6 The experimental configuration of the ANN

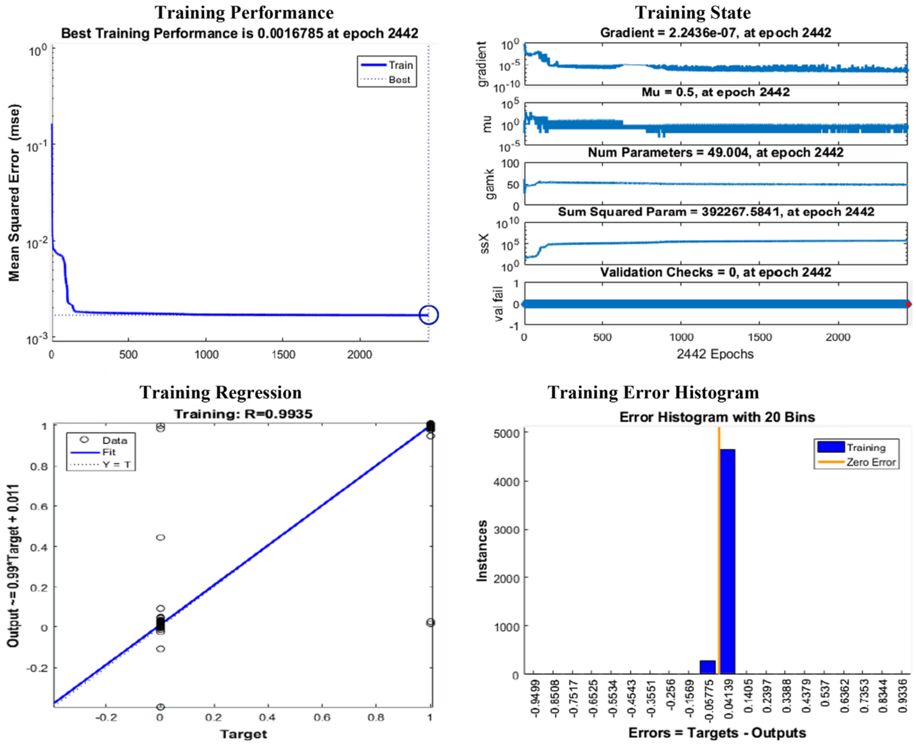


Fig. 7 Training performance graphs

Table 9 Metrics formulas

Metric	Formula
Mean (μ)	$\mu = \frac{\sum x_i}{n}$
Standard Deviation (σ)	$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$
Coefficient of Variation (C_v)	$C_v = \frac{\sigma}{\mu}$

Table 10 False positive percentages

Classification	Standard RegEx	"Boosted" RegEx confidence level ≤ 50%	"Boosted" RegEx confidence level > 50%
Cards	18,422	7,337 (39.83%)	11,085 (60.17%)
Lists	5,394,547	1,565,160 (29.01%)	3,829,387 (70.99%)
Total	5,412,969	1,572,1497	3,840,472

Table 11 Additional Hits using XML tags

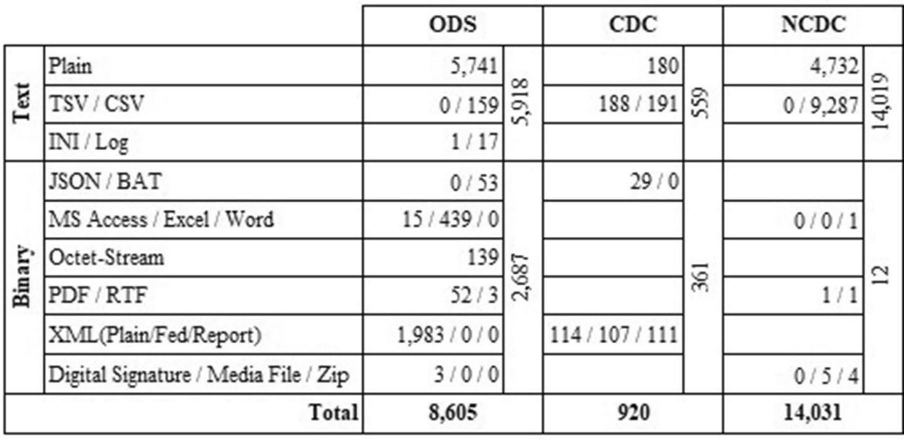
Classification	"Boosted" Absolute and Relative XML confidence level ≤ 50%	"Boosted" Absolute and Relative XML confidence level > 50%
Absolute XML		22,741
Relative XML	325,530	105,758

Once the unit testing of the system was concluded and several pre-processors were developed, multiple executions were conducted in incorporating different sampling sizes, as identified in Table 6 for each set. The formulas used to calculate the metrics are available in Table 9.

The primary derived metrics and features like data set name and delimiter name were input to the next component of the systems, a multi-layer perceptron (MLP) neural network, to approximate the relation between the delimiter characteristics and the delimiter [46]. This component was implemented in MatLab with a graphical illustration of the network shown in Fig. 6. In reducing model overfitting, the train-validate-test technique is used and the input data set was divided into three accordingly, with a ratio of 70%, 15% and 15% [47]. Although there is no clear ratio rule, most researchers tend to use this mix [48, 49]. The Bayesian Regularization Back Propagation (trainbr) training function was used since it presents several advantages and due to that, it is extensively used [50, 51]. The network utilized the Mean Squared Normalized Error (mse) performanceas is typical [52]. In respect to the executions, the limiting number of epochs was set to 10,000 and the hidden neurons were set to 6, using the "rule of thumb" $\frac{2\times(\#Input+\#Output)}{3}$ [53].

To train the neural network, for the training data was incorporated with target outputs using the values of zero (0) and one (1) were used. One (1) was used to represent valid combinations of file and delimiter whilst all other combinations, representing false

Table 12 Data stets file content classification



		ODS		CDC		NCDC	
Text	Plain	5,741	5,918	180	559	4,732	14,019
	TSV / CSV	0 / 159		188 / 191		0 / 9,287	
	INI / Log	1 / 17					
Binary	JSON / BAT	0 / 53	2,687	29 / 0	361		12
	MS Access / Excel / Word	15 / 439 / 0				0 / 0 / 1	
	Octet-Stream	139					
	PDF / RTF	52 / 3				1 / 1	
	XML(Plain/Fed/Report)	1,983 / 0 / 0		114 / 107 / 111			
	Digital Signature / Media File / Zip	3 / 0 / 0				0 / 5 / 4	
	Total	**8,605**		**920**		**14,031**	

Table 13 Files encoding classification

File Encoding	ODS	CDC	NCDC
ISO-8859–1	5,625	126	13,664
ISO-8859–8	1		
Windows-1252	35		253
KOI8-R	1		
MACCYRILLIC	7		
UTF-16LE	205		
UTF-32LE	1		
UTF-8	43	443	102
Total	5,918	559	14,019

positives, were assigned zero (0). The performance graphs of a sample training from one of the input datasets are presented in Fig. 7.

Post-training and the final calculation of the neural network confidence level to confirm that the data on whether the file was delimited with the specific delimiter, the neural network result / output was adjusted to the range of -1 and 2 shown in Fig. 8.

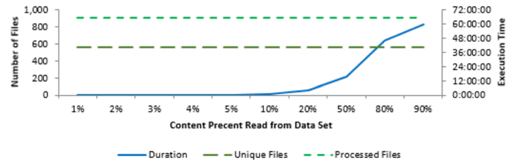
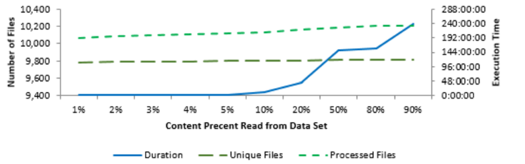
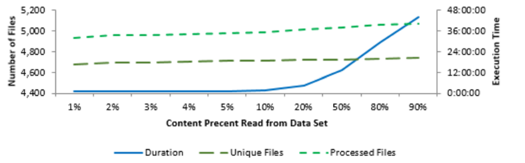
Multiple runs were executed with different sampling sizes. Based on each set of calculated metrics, a neural network was created, trained, and used to predict the delimiter's correctness. In this manner, the automation of the process would be achieved whilst its viability would be guaranteed by lowering the effect of Big Data *Volume* dimension with the use of small samples in correctly discovering the structure of a delimited file.

Results

Data origination—confidentiality

The purpose of the experiment was to confirm the capability of minimizing the number of false positives along with the capability of identifying more occurrences of confidential data with the use of extended pattern matching. In verifying the method used, the results were compared to a traditional simple RegEx matching. Files with a combination

Table 14 Datasets execution timing

		Unique Files	Processed Files	Duration	
NCDC	1%	9,782	10,068	1:28:00	
	2%	9,786	10,089	1:37:00	
	3%	9,792	10,099	1:56:00	
	4%	9,793	10,107	2:36:00	
	5%	9,798	10,120	3:07:00	
	10%	9,797	10,134	10:38:00	
	20%	9,804	10,166	44:56:37	
	50%	9,808	10,185	148:51:13	
	80%	9,810	10,203	155:51:08	
	90%	9,810	10,209	239:41:31	
CDC	1%	558	905	0:12:00	
	2%	558	905	0:13:00	
	3%	558	908	0:14:00	
	4%	558	906	0:17:00	
	5%	558	911	0:20:00	
	10%	558	912	0:49:00	
	20%	558	912	3:53:00	
	50%	558	912	15:24:50	
	80%	558	912	45:40:55	
	90%	558	912	59:59:45	
ODS	1%	4,674	4,928	1:14:40	
	2%	4,695	4,954	1:14:40	
	3%	4,691	4,954	1:15:42	
	4%	4,698	4,965	1:17:56	
	5%	4,712	4,977	1:20:36	
	10%	4,711	4,987	1:44:58	
	20%	4,718	5,010	4:14:59	
	50%	4,724	5,034	13:14:57	
	80%	4,733	5,054	29:12:43	
	90%	4,737	5,066	43:55:11	

of actual positives and false positives were utilized in identifying the value added by the proposal. In order to calculate the performance against the simple RegEx, the 50% confidence level was considered as the benchmark. Based on the parameters, standard RegEx would always yield values lower than 50%, whilst the "boosters" would elevate the respective to higher percentages. Any value below 50% should be considered a "false positive" since no "booster" was utilized to verify its validity. Utilizing the "un-boosted" RegEx matching methodology, 5.4 million occurrences were identified. Based on a sample verification, 3.8million occurrences were confirmed hits, which would coincide with the > 50% "boosted" result set. On average, it can be identified that there was an attainment of improvement of ≈35% in filtering out false positives, as shown in Table 10.

In calculating the system's performance for attaining a better matching by introducing Absolute and Relative XML, all 450 K additional matches are considered; see details in Table 11. This would constitute an increase in the hit ratio by ≈12%. Out of these, by utilizing the aforementioned 50% confidence rule, the immediate contribution would be ≈3% without false positives. The remaining 9%, which in essence is the 325 K pertaining to Relative XML without a "booster", can be further mined in identifying new RegEx expression or new "booster metrics".

Data format—delimiter determination

This experiment sought to explore the viability of identifying data using machine learning and to establish how much data needs to be processed to achieve this. The results of this experiment are presented in a structured manner since each result set contributes as an input into the next phase of the analysis.

The first result set consists of information about the identification of the files. For a source to be eligible for further processing, it should be identified as text. In Table 12, the classification text vs binary and the respective file MIME is presented. The files identified were: 559 CDC, 5,918 ODS and 14,019 NCDC; a total of 20,496 will participate in the subsequent analysis step. In Table 13, the encoding distribution that was identified for all the text files is presented.

The execution time in respect to the results is the next metric that was evaluated. The execution time for each % of line reads was measured and compared with the number of delimiters found in the files. "Unique Files" stands for the actual number of files in the dataset, whilst "Processed Files" stands for the actual number of files processed. The difference is that files identified to have multiple delimiters will have to be processed by the system as many times as the identified delimiters in procuring the required metrics and statistics for each delimiter. As expected, the results showed a gradual increase in execution time whilst the read % was increasing. The interesting outcome was that the variation of the delimiter identification was minimal throughout all read percentages. Detailed information per dataset, in tabular and graphical format, can be seen in Table 14.

Based on the results, it can be concluded that the most interesting percentages are between 1 and 3, since they have low execution timings and relatively high accuracy. Nevertheless, it is imperative to validate other metrics like the efficiency of the location of delimiters and the file structure's actual breakdown in reaching a conclusion. For that reason, further statistical analyses were performed and were imported into an neural network as input parameters. The network was utilized in accurately identifying the delimiter of each file. In matching the results of the ANN with the actual data, a manual and semi-manual process was employed to verify the files delimiters and cross-reference with the ANN output. The initial series of tests, unit testing of the implementation, with one dataset whilst reading 1% of the file's content, was quite revealing and unexpected. The accuracy was so high that in certain cases where the ANN result was conflicting with the manually identified delimiter, it was confirmed that there was a human error in the manual classification and the ANN identified the correct delimiter. At that time, having confidence in the results from the concluded unit testing, the experiment was expanded in covering all content reading percentages (1%, 2%, 3%, 4%, %5, 10%, 20%, 50%, 80%, 90%) and all the datasets (ODS, CDC, NCDC) which resulted in a set of 30 files to be referred to as "% read" files. The training and final calculation of the ANN Function for each set, resulting in 30 neural; networks, was stored in independent files. Each derived ANN was then applied into the respective 30 files from the "% read" files", producing a set of 900 result files, one file per neural network per "% read" file. A cross-tabulation and the related heat map, was created for each of the 30 training networks and its associated 30 files. Fig. 9 illustrates the output for the ODS based network with respect to 1% lines read against all other sets and read percentages. The first cross-tabulation shows the number of unmatched files and the second cross-tabulation heat-map shows the match percentage. We have the input file being tested per data set, per percentage of lines read on the horizontal axis. On the vertical axis, we have an inverted percentage representing the percentile difference per file % that was exhibited when comparing the ANN result with the actual delimiter value of the file. This percentile labelled "Deviation between Actual and ANN results" indicates how much difference is exhibited between the actual value (0 or 1) and the network calculated value (between − 1 and 2 as shown in Fig. 8), which in essence is the neural network performance. The ANN formula is defined by training the ANN using the data collected from processing the ODS data set and reading 1% of each file's content. The formula identified is then used to predict the confidence level of the delimiter using the CDC and NCDC data sets and for all content read percentages (1%, 2%, 3%, 4%, %5, 10%, 20%, 50%, 80%, 90%). The data elements represented are counts and their respective percentile values.

- In the upper part of the figure cross-tabulation, the number of files that fall into that category. The highlighted data element (in purple) represents that 7,687 files exhibit a difference less or equal to 1% when comparing the ANN result with the actual delimiter. The ANN formula being tested is calculated based on the ODS dataset having read 1% of the file content and tested against the NCDC data set while reading 3% of the file content.
- In the heat-map, the percentage of files that fall into that category. The counts from the upper cross-tabulation are depicted as percentages (count/file number). Taking the same example, 76% (in purple) of the files exhibit a difference less or equal to 1% when comparing the ANN result with the actual delimiter.

The heat map is used in visualizing the progression of differentiation, exhibiting the margin of error. Based on the colour coding, the map indicates that although there is high differentiation when testing with NCDC (only 23–26% matches), the bulk number of erroneously classified files is between 1 and 4% since, after that, the number of files is dramatically reduced in contrast to CDC where although 40% is matched the deviation levels remain high thought-out.

For each of the aforementioned cross-tabulations created, one column would be invalid since the training and testing file would be the same. In the sample depicted in Fig. 9, the column ODS/01 with 94% match, highlighted in the red parallelogram, will be invalid since the training and testing data for the ANN is ODS dataset file having read 1% of the file content. It quickly became apparent that the difference mentioned above in CDC variation was exhibited in all 30 cross-tabulations. The exhibited difference ranged from 49% and reached a rate of 98%, depending on the data set and read percentage used to train the ANN, which indicated an error or a peculiarity with the specific dataset. Having confirmed the calculations had no error, it became apparent that *Variety* was in play and the "Escape Special Context" pre-processor as described in Sect. 0 was implemented in adjusting the model to cater to the CDC set's unique characteristics. As a result of the pre-processor, the effect was minimized, having values from 10 to 60%, but it was still apparent that the specific set was acting differently. This preliminary differentiation is vital since it was an indication that the model can auto-adjust itself and perform with high accuracy, even though an outlier dataset was introduced into the ecosystem.

A meta-heat map was created to understand the results further, aggregating the previous results from the independent cross-tabulations. Given an acceptable percentage for the probability of error, the heat map aggregated the data from the 30 individual maps by displaying the respective slice of each table. In Table 15, we can see the heat map for a mismatch of 10%. Just like in the original cross-tabulation, the cases where training and testing was performed on the same dataset should not be considered (the values in blue). Depending on the risk "appetite", the Data Analyst/Scientist can identify the percentages

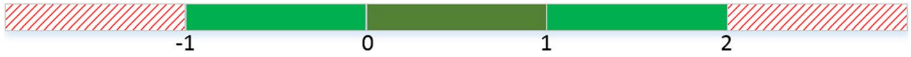


Fig. 8 Neural network result adjustment

that each ANN function would yield against each set and retrieve the respective heat map. Taking into consideration that a) an error level of 1% until 10% would be a viable risk for an analysis and b) the timings of reading percentages for 1% until 3% were the fastest; a subset of the heat map in Table 15 is created for each percentage, e.g. third heat map of Fig. 10. The comparative analysis confirmed that the best results yield from 2 to 3%; see details in the first and second heat map of Fig. 10. Since 2% has a lower execution time than 3% for all dataset processing, see Table 14, it is concluded that the optimal percent would be 2%.

Having identified the optimal read percentage to be 2%, the experimental work resumed in further investigating how the impediments presented by the CDC dataset could affect the framework's accuracy and adaptability. In introducing *Variety*, to the training of the ANN two new sets were devised that consisted of files from all three datasets where each set had an equal contribution of 33%. The first, which was used for training, was composed of 4,801 files, whilst the second used for testing had 4,839 files. Precautions were exercised so that there were no common files between the two sets. A new neural network was implemented using 2% reading of lines. Similarly, to prior tabulations ad heat maps, Fig. 11 presents the results from this investigation.

It can be observed, in contrast to CDC, that there is a high match when it comes to the NCDC set, 4250 out of 4839, and a relative high match when it comes to ODS, 1824 out of 4,839, where the result calculated from the ANN exactly matches the actual file delimiter. Confirming the prior experiment results, the network exhibits a high differentiation with the CDC dataset that is steeply phased out compared to the ODS-based network with lower differentiation but is phased out relatively slower. In contrast, the network based on NCDC and Aggregated set exhibits high accuracy and constant smooth phasing (see Fig. 12).

Based on these new results from the aggregated sets, it was proven that the Automated Approach for Data Characterization framework could accurately classify structures even when the data sets exhibit unique trends and characteristics. The results have shown that: (a) by incorporating multiple sets, the framework is capable of absorbing *Variety* and transforming it into an asset by auto-calibrating the neural network that caters for the new multi-sourced set (e.g. 94% accuracy on the aggregated set) (b) the neural network can efficiently and effectively classify other independent sets (e.g. classify NCDC based on ODS) or multiple sets (e.g. classify all three sets—aggregated set based on NCDC).

Discussion

The paper is contributing to the body of knowledge by exploring and empirically validating a content classification method and a structure identification method. The experiment towards content classification, introduced the use of "Booster Metrics" and measured their effectiveness in removing data noise. For structural identification, the experimental work included the processing of datasets with different quantitative and qualitative characteristics through a two-step procedure: (a) statistical data analysis of big data in calculating specific metrics (b) incorporation of these metrics into a neural network to detect the structure. The experiment was focused on the feasibility of the approach in respect to computation power, representative size of dataset sample to be analysed and time.

The experiments suggest that the accuracy of identifying confidential information can be increased and that there can be an automated process for identifying files structures based on a fragment of the dataset content. Although several enhancements can be utilised in further implementing improved accuracy to the automation, it has to be noted that automation was proven to be a viable proposition.

Table 15 Meta-heat map for 10% mismatch of the neural network results

Mismatch %

10%

			Testing File																													
			ODS										CDC										NCDC									
			01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90
Training File	ODS	01	0.41%	0.85%	1.07%	1.41%	1.27%	1.46%	1.42%	1.67%	1.82%	2.01%	47.62%	47.07%	47.03%	47.46%	46.76%	46.49%	47.59%	47.04%	48.25%	48.90%	1.70%	1.77%	1.85%	1.74%	1.97%	2.22%	2.26%	2.05%	1.84%	1.72%
		02	0.43%	0.12%	0.69%	1.05%	0.92%	1.22%	1.58%	2.24%	2.35%	2.61%	37.24%	38.78%	41.30%	40.84%	41.82%	45.83%	49.89%	53.18%	54.71%	54.93%	1.19%	1.21%	1.23%	1.22%	1.26%	1.30%	1.64%	1.95%	2.18%	2.23%
		03	1.64%	1.51%	1.19%	1.71%	1.65%	21.40%	25.41%	25.43%	25.35%	25.37%	36.02%	35.25%	37.11%	36.42%	35.46%	36.73%	39.69%	40.35%	39.14%	39.47%	2.52%	2.63%	2.64%	2.69%	2.65%	2.67%	2.81%	2.74%	2.97%	2.99%
		04	2.44%	2.28%	2.24%	1.99%	2.29%	25.87%	26.07%	26.52%	26.89%	26.81%	48.29%	46.85%	47.25%	45.81%	46.10%	47.92%	50.66%	52.52%	54.28%	54.93%	52.50%	52.64%	52.80%	52.87%	52.65%	52.73%	52.65%	52.34%	52.38%	52.35%
		05	1.89%	1.64%	1.78%	1.91%	1.08%	2.19%	25.49%	25.74%	26.02%	25.88%	34.48%	34.25%	35.35%	36.87%	36.55%	35.53%	35.09%	34.21%	34.32%	34.65%	15.25%	15.25%	15.45%	15.45%	15.39%	15.32%	15.04%	15.05%	14.92%	14.90%
		10	2.29%	2.24%	2.54%	2.62%	2.41%	1.76%	2.73%	2.92%	3.28%	3.12%	41.44%	44.31%	43.39%	44.04%	44.57%	47.04%	49.89%	50.55%	49.78%	50.33%	15.20%	15.27%	15.20%	15.42%	15.47%	15.68%	15.87%	16.10%	16.34%	16.30%
		20	1.72%	1.51%	1.55%	1.65%	1.57%	1.58%	1.02%	1.47%	1.84%	1.89%	25.41%	24.97%	25.44%	24.72%	24.26%	25.00%	22.81%	25.33%	25.66%	26.10%	1.81%	2.03%	2.09%	2.22%	2.41%	2.73%	2.73%	2.93%	2.78%	2.83%
		50	1.99%	2.20%	2.12%	2.32%	2.13%	2.09%	1.92%	1.47%	2.24%	2.25%	31.16%	32.82%	35.24%	35.87%	37.21%	38.82%	39.14%	40.79%	41.78%	41.78%	1.06%	1.16%	1.05%	1.09%	1.24%	1.78%	2.19%	2.53%	2.82%	2.95%
		80	2.41%	2.26%	2.22%	2.26%	2.01%	2.03%	1.86%	1.63%	1.27%	1.56%	21.44%	20.33%	20.48%	20.42%	20.64%	19.41%	24.45%	23.79%	26.10%	25.44%	1.98%	2.10%	2.20%	2.40%	2.56%	2.92%	3.10%	3.23%	3.45%	3.38%
		90	2.76%	25.72%	25.76%	25.78%	25.56%	25.35%	24.93%	1.95%	1.64%	1.14%	27.18%	27.96%	27.09%	27.26%	27.88%	30.26%	31.25%	32.46%	32.02%	32.02%	1.40%	1.58%	1.76%	2.00%	2.01%	2.51%	3.08%	3.58%	3.64%	3.83%
	CDC	01	4.65%	4.72%	4.80%	4.85%	4.84%	4.87%	4.89%	4.97%	5.66%	5.68%	7.40%	10.06%	10.68%	10.82%	11.53%	13.71%	13.38%	13.27%	12.83%	13.60%	1.97%	2.09%	2.09%	1.99%	2.03%	1.97%	2.03%	1.98%	2.10%	1.96%
		02	6.17%	6.18%	6.26%	6.38%	6.19%	6.24%	6.13%	6.26%	6.67%	6.65%	7.96%	6.63%	8.26%	8.39%	9.99%	10.75%	11.62%	12.61%	11.95%	11.95%	0.83%	0.79%	0.86%	0.95%	0.93%	0.97%	0.92%	0.88%	0.84%	0.86%
		03	4.02%	3.90%	3.86%	4.03%	3.94%	4.07%	3.93%	4.05%	4.06%	4.17%	7.96%	7.07%	6.50%	7.28%	7.68%	8.00%	8.33%	10.31%	12.72%	12.83%	0.27%	0.31%	0.40%	0.40%	0.44%	0.67%	0.90%	1.09%	1.40%	1.29%
		04	10.47%	10.72%	10.62%	10.78%	10.99%	11.01%	10.76%	10.73%	10.90%	10.80%	8.95%	9.50%	8.70%	7.62%	8.45%	10.53%	11.84%	9.76%	10.31%	10.86%	2.73%	2.69%	2.85%	2.91%	2.93%	3.06%	3.10%	3.37%	3.52%	3.54%
		05	9.60%	9.79%	9.79%	10.03%	10.03%	10.05%	9.78%	9.54%	9.62%	9.46%	8.40%	7.96%	8.04%	7.73%	7.46%	8.33%	8.55%	8.55%	8.33%	8.33%	2.13%	1.94%	2.03%	1.95%	1.82%	1.83%	1.81%	1.91%	2.00%	2.00%
		10	6.70%	6.54%	6.62%	6.75%	6.55%	6.72%	6.75%	6.75%	6.89%	6.83%	7.51%	8.40%	7.60%	7.51%	7.35%	6.69%	7.68%	8.33%	8.77%	9.10%	0.21%	0.17%	0.25%	0.20%	0.25%	0.24%	0.31%	0.35%	0.47%	0.47%
		20	4.44%	4.40%	4.50%	4.67%	4.34%	4.35%	4.23%	4.41%	4.47%	4.38%	8.51%	8.62%	8.70%	8.17%	8.89%	8.55%	7.13%	8.11%	9.10%	8.99%	1.30%	1.49%	1.71%	1.85%	2.08%	2.16%	2.13%	2.09%	2.14%	2.13%
		50	11.61%	11.85%	11.99%	12.21%	12.20%	12.09%	11.84%	11.76%	11.81%	11.71%	8.29%	8.40%	8.70%	9.05%	9.11%	9.54%	8.44%	6.14%	7.46%	7.24%	2.57%	2.55%	2.65%	2.55%	2.51%	2.30%	2.05%	1.81%	1.81%	1.77%
		80	9.98%	10.23%	10.38%	10.49%	10.47%	10.31%	9.90%	9.71%	9.89%	9.63%	8.62%	8.40%	8.15%	8.06%	8.12%	8.22%	7.57%	7.35%	6.69%	7.02%	1.59%	1.27%	1.27%	1.26%	1.27%	1.38%	1.53%	1.78%	1.75%	1.94%
		90	6.47%	5.95%	6.34%	6.49%	6.45%	6.30%	6.31%	6.26%	6.37%	6.45%	8.73%	8.95%	9.47%	10.26%	10.43%	9.87%	10.09%	9.54%	7.68%	6.58%	1.52%	1.77%	2.02%	2.20%	2.16%	2.30%	2.31%	2.46%	2.46%	2.54%
	NCDC	01	6.59%	6.58%	6.64%	6.71%	6.55%	6.68%	6.67%	6.67%	6.63%	6.65%	26.96%	26.19%	26.10%	25.83%	25.58%	24.45%	24.34%	23.90%	23.79%	23.79%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		02	6.03%	6.00%	6.06%	6.14%	6.01%	6.16%	6.13%	6.12%	6.11%	6.10%	27.96%	26.74%	26.76%	26.49%	26.34%	25.33%	25.22%	24.89%	24.67%	24.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%
		03	6.07%	6.06%	6.14%	6.22%	6.09%	6.22%	6.19%	6.18%	6.15%	6.14%	27.40%	26.74%	26.76%	26.60%	26.45%	25.33%	25.11%	25.00%	24.56%	24.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		04	5.68%	5.65%	5.75%	5.86%	5.69%	5.84%	5.83%	5.80%	5.78%	5.76%	26.96%	26.63%	26.21%	26.27%	26.23%	25.00%	24.56%	24.45%	24.12%	24.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		05	6.74%	6.70%	6.82%	6.95%	6.73%	6.86%	6.91%	6.89%	6.79%	6.77%	30.61%	30.39%	30.29%	29.69%	29.53%	28.62%	28.40%	27.74%	27.63%	27.63%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		10	5.54%	5.51%	5.57%	5.68%	5.53%	5.59%	5.63%	5.60%	5.58%	5.57%	25.75%	25.86%	25.66%	25.50%	25.91%	25.11%	24.56%	23.57%	23.46%	23.36%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		20	5.56%	5.53%	5.59%	5.68%	5.53%	5.63%	5.65%	5.64%	5.60%	5.61%	26.30%	26.30%	26.21%	26.38%	26.45%	26.21%	25.44%	24.89%	24.34%	23.79%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		50	5.91%	5.87%	5.93%	6.08%	5.89%	5.92%	5.93%	5.90%	5.82%	5.80%	43.87%	43.76%	43.50%	43.16%	43.36%	42.87%	42.54%	41.89%	42.32%	42.00%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%
		80	6.25%	6.22%	6.30%	6.49%	6.23%	6.30%	6.31%	6.30%	6.19%	6.18%	40.44%	40.44%	40.42%	40.29%	40.72%	40.79%	40.79%	41.23%	41.45%	41.45%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%
		90	5.74%	5.73%	5.81%	5.98%	5.75%	5.86%	5.87%	5.92%	5.82%	5.80%	40.33%	40.33%	40.42%	40.29%	40.72%	40.79%	40.68%	41.12%	41.45%	41.45%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%

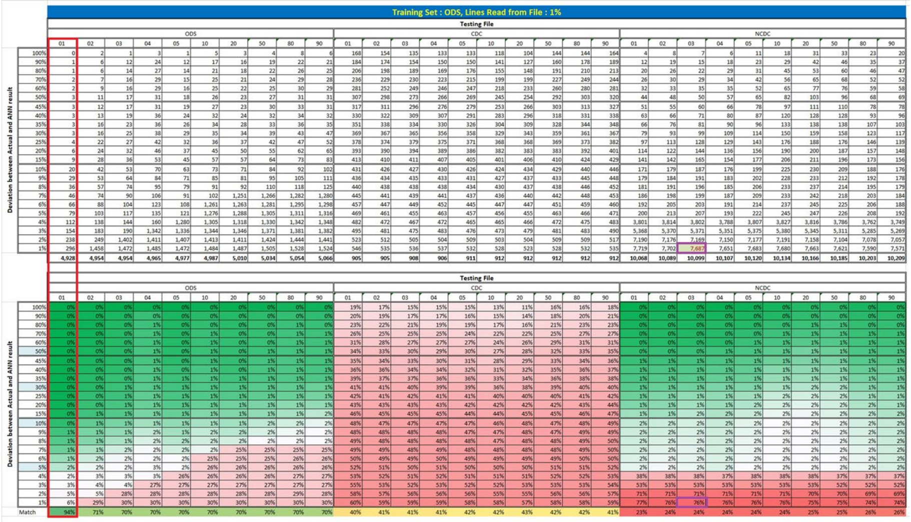
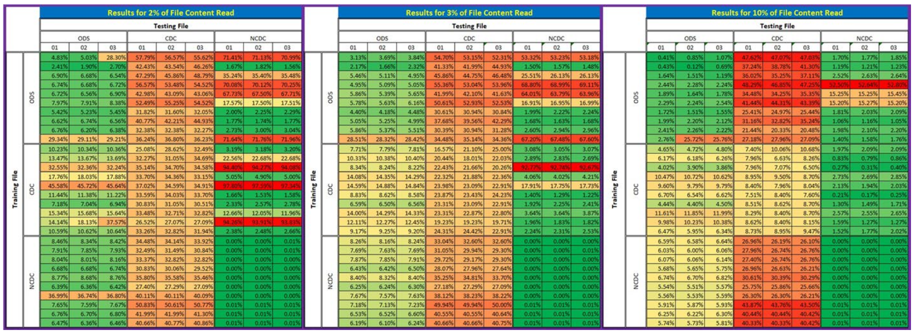

Training Set : ODS, Lines Read from File : 1%
Testing File
ODS
CDC
NCDC
Deviation between Actual and ANN result
Match

Fig. 9 Neural network results cross-tabulation heat map sample (1 out of 30)

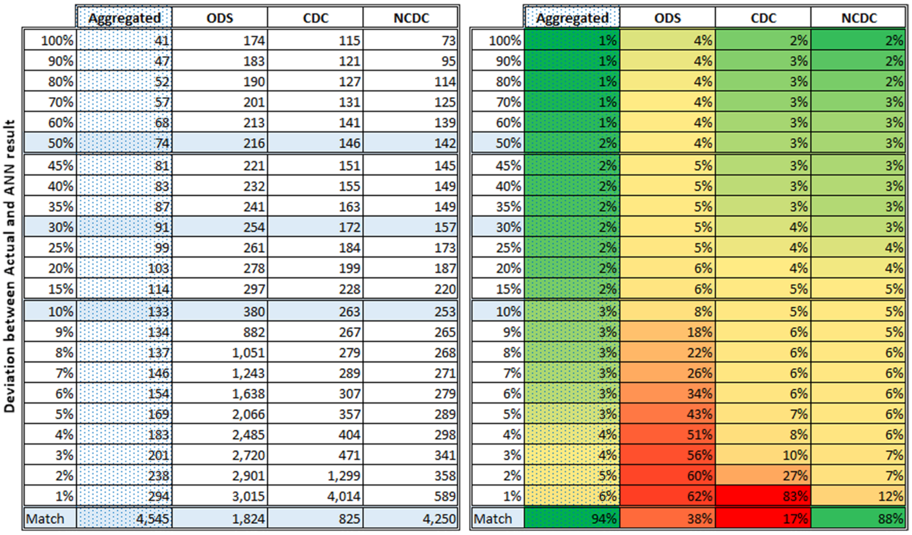


		Results for 2% of File Content Read								
		Testing File								
		ODS			CDC			NCDC		
		01	02	03	01	02	03	01	02	03
Training File	ODS	4.83%	5.03%	28.30%	57.79%	56.57%	55.62%	71.41%	71.13%	70.99%
		2.41%	1.90%	2.70%	42.43%	43.54%	46.26%	1.67%	1.82%	1.56%
		6.90%	6.68%	6.54%	47.29%	45.86%	48.79%	35.24%	35.40%	35.48%
		6.74%	6.68%	6.72%	56.57%	53.48%	54.52%	70.08%	70.12%	70.25%
		6.72%	6.56%	6.90%	42.98%	43.09%	43.06%	67.73%	67.50%	67.71%
		7.97%	7.91%	8.38%	52.49%	55.25%	54.52%	17.57%	17.50%	17.51%
		5.42%	5.23%	5.45%	31.82%	31.60%	32.05%	2.00%	2.25%	2.29%
		6.62%	6.74%	6.56%	40.77%	42.21%	44.93%	1.77%	1.74%	1.77%
		6.76%	6.20%	6.38%	32.38%	32.38%	32.27%	2.73%	3.00%	3.04%
		29.34%	29.11%	29.21%	36.24%	36.80%	36.23%	71.64%	71.76%	71.96%
	CDC	10.23%	10.34%	10.36%	25.08%	28.62%	32.49%	3.19%	3.18%	3.20%
		13.47%	13.67%	13.69%	32.27%	31.05%	34.69%	22.56%	22.68%	22.68%
		32.55%	32.36%	32.24%	35.14%	34.70%	34.58%	94.40%	94.27%	94.08%
		17.76%	18.03%	17.88%	33.70%	34.36%	33.15%	5.05%	4.90%	5.00%
		45.58%	45.72%	45.64%	37.02%	34.59%	34.91%	97.80%	97.59%	97.34%
		11.44%	11.38%	11.22%	33.59%	34.03%	33.70%	1.66%	1.53%	1.58%
		7.18%	7.04%	6.94%	30.83%	31.05%	30.51%	2.33%	2.57%	2.78%
		15.34%	15.68%	15.64%	33.48%	32.71%	32.82%	12.66%	12.05%	11.96%
		14.14%	18.13%	37.57%	26.52%	27.07%	27.09%	94.26%	93.91%	93.83%
		10.59%	10.62%	10.64%	33.26%	32.82%	31.94%	2.38%	2.48%	2.66%
	NCDC	8.46%	8.34%	8.42%	34.48%	34.14%	33.92%	0.00%	0.00%	0.01%
		7.91%	7.85%	7.93%	32.49%	31.49%	30.84%	0.00%	0.00%	0.01%
		8.04%	8.01%	8.16%	33.37%	32.82%	32.82%	0.00%	0.00%	0.00%
		6.68%	6.68%	6.74%	30.83%	30.06%	29.52%	0.00%	0.00%	0.00%
		8.77%	8.68%	8.76%	35.80%	35.58%	35.46%	0.00%	0.00%	0.00%
		6.39%	6.36%	6.42%	27.40%	27.29%	27.09%	0.00%	0.00%	0.00%
		36.99%	36.74%	36.80%	40.11%	40.11%	40.09%	0.00%	0.00%	0.00%
		7.65%	7.59%	7.67%	50.83%	50.61%	50.77%	0.01%	0.01%	0.01%
		6.76%	6.70%	6.80%	41.99%	41.99%	41.30%	0.01%	0.01%	0.01%
		6.47%	6.36%	6.46%	40.66%	40.77%	40.86%	0.01%	0.01%	0.01%

		Results for 3% of File Content Read								
		Testing File								
		ODS			CDC			NCDC		
		01	02	03	01	02	03	01	02	03
Training File	ODS	3.13%	3.69%	3.84%	54.70%	53.15%	52.31%	53.32%	53.23%	53.18%
		2.17%	1.66%	2.32%	41.33%	41.99%	44.93%	1.50%	1.57%	1.48%
		5.46%	5.11%	4.95%	45.86%	44.75%	46.48%	25.51%	26.13%	26.13%
		4.95%	5.09%	5.05%	55.36%	53.04%	53.96%	68.80%	68.99%	69.11%
		5.86%	5.39%	5.65%	41.99%	42.10%	41.63%	64.01%	63.79%	63.96%
		5.78%	5.63%	6.16%	50.61%	52.93%	52.53%	16.91%	16.95%	16.99%
		4.40%	4.18%	4.48%	30.61%	30.94%	30.84%	1.99%	2.22%	2.24%
		5.05%	5.25%	4.99%	37.68%	39.56%	42.29%	1.68%	1.63%	1.68%
		5.86%	5.37%	5.51%	30.39%	30.94%	31.28%	2.60%	2.94%	2.96%
		28.51%	28.32%	28.42%	34.48%	35.14%	34.36%	67.20%	67.48%	67.60%
	CDC	7.71%	7.79%	7.81%	16.57%	21.10%	25.00%	3.08%	3.05%	3.07%
		10.33%	10.38%	10.40%	20.44%	18.01%	22.03%	2.89%	2.83%	2.69%
		8.34%	8.24%	8.22%	22.43%	21.66%	20.26%	92.77%	92.74%	92.67%
		14.08%	14.35%	14.29%	22.32%	21.88%	22.36%	4.06%	4.02%	4.21%
		14.59%	14.88%	14.84%	23.98%	23.09%	22.91%	17.91%	17.75%	17.73%
		8.83%	8.62%	8.58%	23.87%	23.43%	24.23%	1.40%	1.29%	1.22%
		6.59%	6.50%	6.56%	23.31%	23.09%	22.91%	1.92%	2.25%	2.41%
		14.00%	14.29%	14.33%	23.31%	22.87%	22.80%	3.64%	3.64%	3.87%
		12.11%	12.27%	12.45%	19.23%	19.23%	19.71%	1.96%	1.83%	1.82%
		9.17%	9.25%	9.20%	24.31%	24.42%	22.91%	2.24%	2.31%	2.53%
	NCDC	8.26%	8.16%	8.24%	33.04%	32.60%	32.60%	0.00%	0.00%	0.00%
		7.69%	7.63%	7.69%	31.05%	29.94%	29.30%	0.00%	0.00%	0.01%
		7.87%	7.85%	7.91%	29.72%	29.17%	29.30%	0.00%	0.00%	0.00%
		6.43%	6.42%	6.50%	28.07%	27.96%	27.64%	0.00%	0.00%	0.00%
		8.40%	8.32%	8.40%	35.25%	34.81%	33.70%	0.00%	0.00%	0.00%
		6.25%	6.24%	6.30%	27.18%	27.29%	27.09%	0.00%	0.00%	0.00%
		7.67%	7.57%	7.63%	38.12%	38.23%	38.22%	0.00%	0.00%	0.00%
		7.18%	7.13%	7.23%	49.94%	49.94%	50.00%	0.01%	0.01%	0.01%
		6.53%	6.52%	6.60%	40.55%	40.55%	40.64%	0.01%	0.01%	0.01%
		6.19%	6.10%	6.24%	40.66%	40.66%	40.75%	0.01%	0.01%	0.01%

		Results for 10% of File Content Read								
		Testing File								
		ODS			CDC			NCDC		
		01	02	03	01	02	03	01	02	03
Training File	ODS	0.41%	0.85%	1.07%	47.62%	47.07%	47.03%	1.70%	1.77%	1.85%
		0.43%	0.12%	0.69%	37.24%	38.78%	41.30%	1.19%	1.21%	1.23%
		1.64%	1.51%	1.19%	36.02%	35.25%	37.11%	2.52%	2.63%	2.64%
		2.44%	2.28%	2.24%	48.29%	46.85%	47.25%	52.50%	52.64%	52.80%
		1.89%	1.64%	1.78%	34.48%	34.25%	35.35%	15.25%	15.25%	15.45%
		2.29%	2.24%	2.54%	41.44%	44.31%	43.39%	15.20%	15.27%	15.20%
		1.72%	1.51%	1.55%	25.41%	24.97%	25.44%	1.81%	2.03%	2.09%
		1.99%	2.20%	2.12%	31.16%	32.82%	35.24%	1.06%	1.16%	1.05%
		2.41%	2.26%	2.22%	21.44%	20.33%	20.48%	1.98%	2.10%	2.20%
		2.76%	25.72%	25.76%	27.18%	27.96%	27.09%	1.40%	1.58%	1.76%
	CDC	4.65%	4.72%	4.80%	7.40%	10.06%	10.68%	1.97%	2.09%	2.09%
		6.17%	6.18%	6.26%	7.96%	6.63%	8.26%	0.83%	0.79%	0.86%
		4.02%	3.90%	3.86%	7.96%	7.07%	6.50%	0.27%	0.31%	0.40%
		10.47%	10.72%	10.62%	8.95%	9.50%	8.70%	2.73%	2.69%	2.85%
		9.60%	9.79%	9.79%	8.40%	7.96%	8.04%	2.13%	1.94%	2.03%
		6.70%	6.54%	6.62%	7.51%	8.40%	7.60%	0.21%	0.17%	0.25%
		4.44%	4.40%	4.50%	8.51%	8.62%	8.70%	1.30%	1.49%	1.71%
		11.61%	11.85%	11.99%	8.29%	8.40%	8.70%	2.57%	2.55%	2.65%
		9.98%	10.23%	10.38%	8.62%	8.40%	8.15%	1.59%	1.27%	1.27%
		6.47%	5.95%	6.34%	8.73%	8.95%	9.47%	1.52%	1.77%	2.02%
	NCDC	6.59%	6.58%	6.64%	26.96%	26.19%	26.10%	0.00%	0.00%	0.00%
		6.03%	6.00%	6.06%	27.96%	26.74%	26.76%	0.00%	0.00%	0.00%
		6.07%	6.06%	6.14%	27.40%	26.74%	26.76%	0.00%	0.00%	0.00%
		5.68%	5.65%	5.75%	26.96%	26.63%	26.21%	0.00%	0.00%	0.00%
		6.74%	6.70%	6.82%	30.61%	30.39%	30.29%	0.00%	0.00%	0.00%
		5.54%	5.51%	5.57%	25.75%	25.86%	25.66%	0.00%	0.00%	0.00%
		5.56%	5.53%	5.59%	26.30%	26.30%	26.21%	0.00%	0.00%	0.00%
		5.91%	5.87%	5.93%	43.87%	43.76%	43.50%	0.01%	0.01%	0.01%
		6.25%	6.22%	6.30%	40.44%	40.44%	40.42%	0.01%	0.01%	0.01%
		5.74%	5.73%	5.81%	40.33%	40.33%	40.42%	0.01%	0.01%	0.01%

Fig. 10 Comparative analysis of 1–10% of error for 1–3% of Lines Read



Deviation between Actual and ANN result		Aggregated	ODS	CDC	NCDC
	100%	41	174	115	73
	90%	47	183	121	95
	80%	52	190	127	114
	70%	57	201	131	125
	60%	68	213	141	139
	50%	74	216	146	142
	45%	81	221	151	145
	40%	83	232	155	149
	35%	87	241	163	149
	30%	91	254	172	157
	25%	99	261	184	173
	20%	103	278	199	187
	15%	114	297	228	220
	10%	133	380	263	253
	9%	134	882	267	265
	8%	137	1,051	279	268
	7%	146	1,243	289	271
	6%	154	1,638	307	279
	5%	169	2,066	357	289
	4%	183	2,485	404	298
	3%	201	2,720	471	341
	2%	238	2,901	1,299	358
	1%	294	3,015	4,014	589
	Match	4,545	1,824	825	4,250

	Aggregated	ODS	CDC	NCDC
100%	1%	4%	2%	2%
90%	1%	4%	3%	2%
80%	1%	4%	3%	2%
70%	1%	4%	3%	3%
60%	1%	4%	3%	3%
50%	2%	4%	3%	3%
45%	2%	5%	3%	3%
40%	2%	5%	3%	3%
35%	2%	5%	3%	3%
30%	2%	5%	4%	3%
25%	2%	5%	4%	4%
20%	2%	6%	4%	4%
15%	2%	6%	5%	5%
10%	3%	8%	5%	5%
9%	3%	18%	6%	5%
8%	3%	22%	6%	6%
7%	3%	26%	6%	6%
6%	3%	34%	6%	6%
5%	3%	43%	7%	6%
4%	4%	51%	8%	6%
3%	4%	56%	10%	7%
2%	5%	60%	27%	7%
1%	6%	62%	83%	12%
Match	94%	38%	17%	88%

Fig. 11 Aggregated set results heat map

The experimental work revealed challenges in respect to Big Data dimensions and system expandability/extensibility/maintainability. *Volume* manifested in relation memory (RAM) limitations while processing large files. *Velocity* limited the processing window which in turned drove the sampling initiatives. *Variety* was confirmed in both experiments and additional sub-systems had or will have to be introduced. Questions in relation to expandability, extensibility and maintainability included:

- How can the system be extended without human interaction in respect to new RegEx patterns?
- How can binary or HEX content be treated?
- Is there a way to have more contextual accuracy?
- Is there a way to retrofit prior structural identified information into the system?
- Would such a framework be applicable in a business implementation towards DLP?

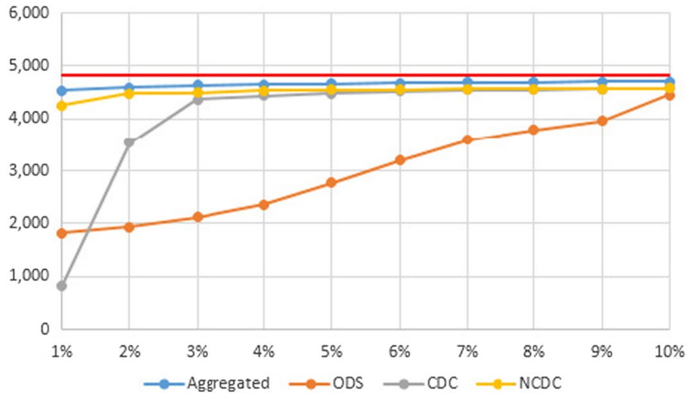


Fig. 12 ANN functions differentiation phasing out plot

Data Organisation—Confidentiality phase could benefit from enhancements in increasing the accuracy or broadening the identification spectrum. Techniques and technologies including (a) the use of Natural Language Processing in identifying information, (b) Hexadecimal parsing that would "reveal" information that is not visible before translated into text and (c) exploding of container/cabinet files (zip, cab, jar etc.) that would give the capability of investigating a broader number of sources. Multiple other pre-processors could be incorporated in transforming the digital data like video, picture and sound in becoming inputs for the respective module. A Soundex extension can be introduced in aggregating the information from the "Relative XML" classification recorded and identify new RegEx expressions that can be auto-retrofitted into the system in extending the system.

Data Format—Delimiter Detection, can be extended with enhancements in performance and accuracy. A suggestion could be implementing an "early warning system" where the entire file will not have to be read in calculating the indices required to identify the structure. The "progressive analysis" will be based on prior experience built from other data sets and provide quick indicators about the file structures. Delimiter patterns can be any number of character combinations. Based on that, a possible enhancement could include a pre-processor that would identify possible character sets that could serve as delimiters and incorporate them for the primary system to analyse. Grammatical analysis and Natural Language Processing could also enhance the performance since, in many cases, identified delimiters could be easily excluded, thus reducing the execution times and increasing accuracy. Double quotes, carriage returns and line feeds could also pause some threat to the experimented upon algorithms since the definition of a "line" is the basis for analysing a record into fields.

Another possible enhancement based on the work mentioned earlier could be adopting a framework for corporate data compliance. Based on the experiments, the Data Scientist, before analysing the data, is automatically presented with indications, if not a confirmation depending on his/her risk appetite, about (a) the existence and nature of confidential information in a set and (b) the structure of the information in the set. Having empowered the analyst to proceed with mining the data, the results and possibly the sets themselves will be shared. In certain cases, new sets will be created from corporate data to be shared with other corporates or the public. During this process, it is imperative to comply with any industry rules for data confidentiality. Towards this regulatory requirement, a rule-based system that will enforce corporate rules in addition to the Data Scientist's explicit anonymization and depersonalization decisions can prove beneficial. The system could exploit the possibility of (a) defining different levels of rules, e.g. corporate or departmental, (b) identifying—preferably real-time and visually- compliance to rules, (c) what-if scenario analysis and calibration based on public compliance verification engines.

Although applicable to Big Data, these experiments, and enhancements towards minimizing *Variety* and empowering the work of Data Scientists could also prove beneficial to control functions and other business units like Group Information Office, Chief Data Office, Operational Risk, and Compliance. By increasing the usage of actionable insight derived from Big Data mining and processes, the promotion and adoption of Big Data can be further enhanced throughout multiple business entities.

Conclusion

The results have shown that Data Scientists will be assisted in their data classification and identification in a automated manner using the proposed approaches. Even though the experiments showed that human interaction is recommended as analysts will be required to understand and interpret the result, the work itself can be automated. The approach would remove much of the labour intensive part of the data origination and data format tasks, whilst only a residual critical decision remains with the Data Scientists. In this way, the initial implications of *Variety* in staging and ingesting the data can be minimized utilizing an algorithmic approach. With the use of the automated processes to identify any confidential information, any regulatory risk of noncompliance and possible data loss is also minimized. This will enable the use of software and minimize human-intensive tasks, which in turn will allow for higher adoption. In this way, the limitations identified earlier in the text by multiple scholars can be tackled. The approach confirmed its viability in a Big Data environment by utilizing 2% of the dataset data whilst producing high-quality results that can address the challenge at hand.

In both experiments, it is evident that a substantial number of executions on a large sample of diverse datasets have to happen in order to be able to retrofit the data onto the system and further validate and enhance it. The risk of not being able to acquire a sufficient number of sets and data could have an impact in fine-tuning the parameterization of percentages in the first experiment, whilst for the second, the effect would be to retrain the newral network whenever a new set is introduced. The suggested systems, although employing many automated techniques, lead to an automated process since the

Data Scientists interactions are limited in fine-tuning the models and understanding the exceptions.

Acknowledgements
Not applicable

Authors' contributions
GV was responsible for authoring and NC along with SA were responsible for guidance and editing. All authors read and approved the final manuscript

Funding
Not applicable.

Availability of data and materials
The datasets used and/or analyzed during the current study are available in the public domain for (DCD—https://data.cdc.gov/ and NCDC—https://www.ncdc.noaa.gov/cdo-web/datasets), ODS is not publicly available due to the proprietary nature of the origin. Results sets are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 August 2021 Accepted: 19 December 2021
Published online: 10 January 2022

References

1. Laney D. 3D data management: controlling data volume, velocity, and variety. Appl Deliv Strateg. 2001;949:4.
2. Supriya M, Chattu VK. A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health. Big Data Cogn Comput. 2021. https://doi.org/10.3390/bdcc5030041.
3. Khan MAUD, Uddin MF, Gupta N. "Seven V's of big data understanding big data to extract value. In: Proc. 2014 Zo. 1 Conf. Am. Soc. Eng. Educ.—"Engineering Educ. Ind. Involv. Interdiscip. Trends, ASEE Zo. 1 2014. 2014. doi: https://doi.org/10.1109/ASEEZone1.2014.6820689
4. Zhang L. A framework to model big data driven complex cyber physical control systems. In: 20th International conferenceon automation & computing. 2014. p. 12–13.
5. Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. Big Data. 2013;1(1):51–9. https://doi.org/10.1089/big.2013.1508.
6. Lennard, "Data variety—the ugly duckling of big data." DataShaka. 2014, [Online]. http://www.datashaka.com/blog/non-techie/2014/01/data-variety-ugly-duckling-big-data.
7. Mao R, Xu H, Wu W, Li J, Li Y, Lu M. Overcoming the challenge of variety : big data abstraction , the next evolution of data management for AAL communication systems, no. January. 2015. p. 42–47
8. Brown E. Big data problems—variety not volume|big data forum, Big-Data Forum. 2014. http://www.big-dataforum.com/605/big-data-problems-variety-not-volume. Accessed 02 Sep 2015.
9. Vranopoulos GE, Triantafyllidis AA, Lefteriotis K. Big data variety, 'where do we stand', an overview of big data and the variety challenge. Int J Manag Appl Sci 2020; 6(3).
10. McElhenny J. Leaving data on the table : new survey shows variety , not volume , is the bigger challenge of analyzing big data paradigm4 survey also shows half of data scientists fed up with traditional databases. InkHouse (for Paradig). 2014. p. 3–4.
11. Kimura C. Beyond the 'Big': solving for data variety requires new thinking—ClearStory data. CrearStory Data. 2014. http://www.clearstorydata.com/2014/12/beyond-big-solving-data-variety-requires-new-thinking/. Accessed 02 Sep 2015.
12. Kumar M. Variety is the unsolved problem in big data|SmartData collective. Smart Data Collective. 2013. http://www.smartdatacollective.com/maheshkumar1/156256/why-variety-unsolved-problem-big-data. Accessed 02 Sep 2015.
13. Baker P. Variety, not volume, biggest big data challenge in 2015—FierceBigData. FireceBigData 2015. http://www.fiercebigdata.com/story/variety-not-volume-biggest-big-data-challenge-2015/2015-01-14. Accessed 02 Sep 2015.

14. Labrinidis A, Jagadish H. Challenges and opportunities with big data. Proc VLDB Endow. 2012;5(12):2032–3. https://doi.org/10.14778/2367502.2367572.
15. Trader T. Big data future hinges on variety. 2014. http://www.datanami.com/2014/02/24/big_data_future_hinges_on_variety/. Accessed 05 Oct 2015.
16. Assunção MD, Calheiros RN, Bianchi S, Netto MA, Buyya R. Big data computing and clouds: trends and future directions. J Parallel Distrib Comput. 2014. https://doi.org/10.1016/j.jpdc.2014.08.003.
17. Bloomberg. The current state of business analytics: where do we go from here? Bloom. Businessweek Res. Serv. 2011.
18. Neves PC, Schmerl B, Cámara J, Bernardino J. Big data in cloud computing: features and issues. In: IoTBD 2016—proceedings of the international conference on internet of things and big data. 2016. p. 307–314. doi: https://doi.org/10.5220/0005846303070314.
19. Vranopoulos G, Trianatafylidis A, Yiannopoulos M. Putting ATM cash requirements into context, ANN relation to socioeconomic events and variety. In: BEFB 2016—international congress on banking, finance and business 2016. p. 526–545. doi: 212-4044
20. Dautov R, Distefano S. Quantifying volume, velocity, and variety to support (Big) data-intensive application development. 2017. Doi: https://doi.org/10.1109/BigData.2017.8258252.
21. Singh G, Gupta S. Big data: a 'big' leap towards profitability. Analytics India Magazine Pvt Ltd, 2014. https://analyticsindiamag.com/big-data-a-big-leap-towards-profitability/. Accessed 21 Nov 2021.
22. Romanov A. Putting a dollar value on big data insights. WIRED. https://www.wired.com/insights/2013/07/putting-a-dollar-value-on-big-data-insights/. Accessed 21 Nov 2021.
23. Shacklett M. How to cope with the big data variety problem—TechRepublic. TechRepublic. 2014. http://www.techrepublic.com/article/how-to-cope-with-the-big-data-variety-problem/. Accessed 02 Sep 2015.
24. Luo J, Dang A-R, Mao Q-Z. The study of integration of multi-sources heterogeneous data based on the ontology. 2008.
25. Kwadwo Antwi S, Hamza K. Qualitative and quantitative research paradigms in business research: a philosophical reflection. Online, 2015. [Online]. www.iiste.org.
26. Banek M, Vrdoljak B, Tjoa AM, Skočir Z. Automating the schema matching process for heterogeneous data warehouses. 2007.
27. Carlo B, Daniele B, Federico C, Simone G. A data quality methodology for heterogeneous data. Int J Database Manag Syst. 2011. https://doi.org/10.5121/ijdms.2011.3105.
28. Wang L. Heterogeneous data and big data analytics. Autom Control Inf Sci. 2017;3(1):8–15.
29. Abellan Matamoros C. Facebook to pay record $5 billion fine over privacy violations, but are they getting off lightly? Euronews, 2019. https://www.euronews.com/2019/07/24/facebook-to-pay-record-5-billion-fine-over-privacy-violations-but-are-they-getting-off-lig.
30. Lovejoy B. GDPR fines total €56M in first year as Facebook faces 11 investigations. 9To5Mac. 2019. https://9to5mac.com/2019/05/28/gdpr-fines/.
31. Wang J, Liu W, Kumar S, Chang S-F. Learning to hash for indexing big data—a survey. Sep. 2015, [Online]. http://arxiv.org/abs/1509.05472.
32. Koenig S. Can DLP protect credit card numbers without burying you in false positives? ZScaler 2019.
33. Solbers R. Data classification tips: finding credit card numbers. Varonis 2020.
34. regex—Algorithms for detecting Credit Card Number reducing false positives/negatives—Stack Overflow. https://stackoverflow.com/questions/18842081/algorithms-for-detecting-credit-card-number-reducing-false-positives-negatives. Accessed 03 Jan 2021.
35. Finding or verifying credit card numbers. https://www.regular-expressions.info/creditcard.html. Accessed 03 Jan 2021.
36. Java regex matching IP Address—yet another programming solutions log. https://farenda.com/java/java-regex-matching-ip-address/. Accessed 03 Jan 2021.
37. IBAN Regex design—stack overflow. https://stackoverflow.com/questions/44656264/iban-regex-design. Accessed 03 Jan 2021.
38. "Regular Expression Library." https://regexlib.com/(X(1)A(1No-5dalAoSIC8mpU-0wtp7B9cY7gTSTHVOaIrtzqHDEK9roERzQ2Qro29iJ7wrPJbrhL8nohAyR_1PpvX2SoTm4pa4WPt95rUUSi-P_9scEQpMFdfhWIgH_Rad6LjTZ9_gaGMBlTOllz7DwBix9fkQTZHNPpix3sG1xigIlVPjkvdyXSRSwkAvKCxN87A5j0))/REDetails.aspx?regexp_id=993&AspxAutoDetectCookieSupport=1. Accessed 03 Jan 2021.
39. SearchCode RegExp Library Formats.
40. Apache Tika—Apache Tika. https://tika.apache.org/. Accessed 05 Dec 2021.
41. Brown KA, Brittman S, Maccaferri N, Jariwala D, Celano U. Machine learning in nanoscience: big data at small scales. Nano Lett. 2020;20(1):2–10. https://doi.org/10.1021/acs.nanolett.9b04090.
42. Greenbaum J. Adding business value to database consolidation. Enterprise applications consulting 2008.
43. Gregory M. Strategies for implementing big data analytics. 2013.
44. Fan S, Lau RYK, Zhao JL. Demystifying big data analytics for business intelligence through the lens of marketing mix. Big Data Res. 2015. https://doi.org/10.1016/j.bdr.2015.02.006.
45. Mcdaniel P, Cárdenas AA, Rajan SP. SYSTEMS SECURITY big data analytics for security. [Online]. www.computer.org/security.
46. Pijanowski BC, Tayyebi A, Doucette J, Pekin BK, Braun D, Plourde J. A big data urban growth simulation at a national scale: configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment. Environ Model Softw. 2014;51:250–68. https://doi.org/10.1016/j.envsoft.2013.09.015.
47. NcCaffrey J. Neural network train-validate-test stopping—visual studio magazine. Visual Studio Magazine. 2015. https://visualstudiomagazine.com/articles/2015/05/01/train-validate-test-stopping.aspx. Accessed 30 Jan 2021.
48. Train test validation split python. IIIT-Delhi Blog. 2020. https://blog.iiitd.ac.in/wp-content/uploads/usborne-train-vihc/train-test-validation-split-python-943e13. Accessed 30 Jan 2021.

49. Rachel Lea Ballantyne Draelos. Best use of train/val/test splits, with tips for medical data—glass box. Glassbox Medicine. https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/. Accessed 30 Jan 2021.
50. Yue Z, Songzheng Z, Tianshi L. Bayesian regularization BP Neural Network model for predicting oil-gas drilling cost. In: BMEI 2011—proceedings 2011 international conference on business management and electronic information, vol. 2. 2011. p. 483–487. Doi: https://doi.org/10.1109/ICBMEI.2011.5917952.
51. Livingstone D. Artificial neural networks. Humana Press; 2008.
52. Christiansen NH, Erlend P, Voie T, Winther O, Høgsberg J. Comparison of neural network error measures for simulation of slender marine structures. J Appl Math. 2014. https://doi.org/10.1155/2014/759834.
53. Y'barbo D. Machine learning - multi-layer perceptron (MLP) architecture: criteria for choosing number of hidden layers and size of the hidden layer?. Stack overflow. 2012. http://stackoverflow.com/questions/10565868/multi-layer-perceptron-mlp-architecture-criteria-for-choosing-number-of-hidde, Accessed 05 Jan 2016.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.