A Survey on Big Data and Collective Intelligence

Ioannis Karydis*, Spyros Sioutas*, Markos Avlonitis*, Phivos Mylonas*, and Andreas Kanavos+

Ionian University, 49132 Kerkyra, Greece*
University of Patras, 26504 Patra, Greece+
{karydis,sioutas,avlon,fmylonas}@ionio.gr,kanavos@ceid.upatras.gr

Abstract. The creation and accumulation of Big Data is a fact for a plethora of scenarios nowadays. Sources such as the ever-increasing diversity sensors as well as the content created by humans have contributed to the Big Data's enormous size and unique characteristics. Making sense of these data has primarily rested upon Big Data analysis algorithms. Still, in one too many cases the effectiveness of these algorithms is hampered by the very nature of Big Data: analogue, noisy, implicit, and ambiguous. Enter Collective Intelligence: the capability of interconnected intelligences achieving ameliorated results in activities than each of the single intelligences creating the collective solely would. Accordingly, this work presents existing research on Big Data and Collective Intelligence. The work is concluded with the presentation of the challenges and perspectives of the common ground between the directions of Big Data and Collective Intelligence.

Keywords: Big Data; Collective intelligence; Cloud computing; NoSQL; Crowdsourcing; Distributed systems; Synergetic networks.

1 Introduction

The capability to create and store information nowadays is unparalleled. The gargantuan plethora of sources that leads to information of varying type, quality & consistency, large volume, creation rate per time unit has been identified as of 2001 [16]. Such data, also known affectively as Big Data, are currently at the order of tens of pebibytes [14] with increasing tendency, while their management and analysis are unsurprisingly a prominent research direction [13].

The continuously expansive ubiquitousness of cheap, mobile, network capable, multi-sensory, power-efficient processing capability is in part responsible for the aforementioned creation of Big Data (sets). In relation to the rate of evolution of the processing capability, storage capability of the produced information advanced somewhat less and thus the notion of streaming data with transient character. Another, less commonly identified factor for the emergence of Big Data, is the change of paradigm in content creation wherein willingness to contribute, domain expertise and access to content sharing methods became no longer a capability/privilege of few people but a widespread common ground, especially under the auspices of "web 2.0"-based social networks' practices.

The collection of Big Data and the requirement for their management and analysis can potentially provide valuable insight that only such volumes can. Most importantly through, the aforementioned abundance of Big Data information as well as the existence of synergetic networks have the potential to "*shift*

knowledge and power from the individual to the collective" [1] by catering for Collective Intelligence. This is achieved by addressing at least two of the three elements of the Collective Intelligence property as provided in the definition by Glenn [12]. In other words, Big Data address the element of information while the synergetic networks, that also supported Big Data, provide for Collective Intelligence's underlying interaction web for its intelligence units.

Despite the fact that Collective Intelligence is as old as humans are, its significance is by no means diminished. As Collective Intelligence refers to the capability of interconnected intelligences achieving ameliorated results in their activities than each of the single intelligences creating the collective solely would, its implications have been, are and most probably will be ever reaching. Some current well known examples of Collective Intelligence are collaboration projects such as Wikipedia[1] and the operating system Linux[2] as well as the knowledge extraction methods such as the PageRank algorithm of Google's search engine.

The advent of the domination of digitised information[14] that lead to Big Data has already had a profound effect on Collective Intelligence and is widely accepted that will continue to do so in an increasing manner[18]. The methods of analysis applied on Big Data refer to the very same behaviour that the "intelligent" part of Collective Intelligence aims to do. On the other hand, the activity of processing the Big Data by algorithms in distributed systems refers to one instance of the "collectivity" part of Collective Intelligence.

1.1 Motivation & Contribution

The identification of the overlapping segments in research for Big Data's analysis methods and Collective Intelligence's "intelligent behaviour" characteristics will provide a fertile ground for a synergy between, at least, these two domains that is necessary to be addressed. This necessity emerges not only from the *suis generis* interdisciplinary character of both Big Data's and Collective Intelligence's methods but also by the requirement to avoid reaching a state that Big Data accumulation and analysis does not lead to more intelligent insight than any individual actor.

In order to achieve the aforementioned aims, this work:

- presents concisely the state-of-the-art methods for Big Data analysis,
- provides a comprehensive account of the state-of-the-art methods for Collective Intelligence in various disciplines,
- identifies research challenges and methodological perspectives in the amalgamated domain of Big Data and Collective Intelligence.

The remaining of the work is organised as follows: Section 2 presents existing research on Big Data's key pillars, i.e. analysis methods, systems' architecture and databases solutions. Next, Section 3 provides a comprehensive account of the state-of-the-art methods for Collective Intelligence in prominent disciplines for varying type of actors. Finally, the work is concluded in Section 4 by a summary including the challenges and perspectives of the common ground between the directions of Big Data and Collective Intelligence.

[1] `https://en.wikipedia.org/wiki/Collective_intelligence`
[2] `http://www.linuxfoundation.org/`

2 Big Data

In an attempt to introduce Big Data based on their characteristic attributes, Laney's [16] focus rested on Volume, Velocity and Variety attributes, affectionately referred to as the "3 V's".

Volume The increase of aggregate volume of data due to lower cost of retention as well as the perception of information as tangible asset.

Velocity The pace at which information is created and thus requires attention (such as storage, caching, rooting, latency balancing, etc.).

Variety The nature of the data that most often than not are in varying data formats/structures/types in addition to inconsistent semantics.

Big Data is referred to by Snijders et al. [28] as "*a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software*". Continuing further with the attribute based definition, and as the field evolved further, more V's have been added to the initial "3 V's" in order to sufficiently address the challenges encountered, producing thus a more detailed nature of a total ten characteristic attributes (thus "10 V's"[3]) of Big Data [5].

Veracity The requirement of access to the appropriate and enough (for training, validation, testing) data in order to be able to verify hypotheses.

Validity The quality of the data originating from various sources based on varying schemata leading to the requirement of "cleaning processes".

Value The business value of the data, such as Return On Investment as well as their potential to transform an organisation.

Variability The non-static nature of data sources that lead to information that is dynamic and evolving.

Venue The multiplicity of sources (distributed) the data originate from that make for the heterogeneity of the data.

Vocabulary Inherent descriptors of the data such as schema, semantics and data models that depend on the content and/or the context of the data and refer to the data's structure, syntax and content.

Vagueness Lack of clear definition of the complex & evolving term "Big Data".

The collection of the aforementioned characteristics imposes a plethora of challenges to be addressed, as put by Borne [5]: "*the capture, cleaning, curation, integration, storage, processing, indexing, search, sharing, transfer, mining, analysis, and visualization of large volumes of fast-moving highly complex data*".

2.1 Analysis Methods

Efficient analysis methods, i.e. data-driven decision making methods, in the era of Big Data is a research direction receiving great attention [30]. The perpetual interest to efficient knowledge discovery methods is mainly supported by the nature of Big Data and the fact that in each instance, Big Data cannot be handled and processed to extract knowledge by most current information systems.

[3] Henceforth appearing with a capital first V in order to denote the specific meaning these have for Big Data.

The generic simplified knowledge discovery process pertains to three stages from raw data input to knowledge output. The first step is to gather from multiple sources the raw data input an perform tasks such as selection of appropriate to the process data-set, preprocessing in order to clean the data into a useful state as well as transformation to achieve common representation suitable for the next step. The second step is to perform analysis utilising methodologies of data mining, thus leading to information as an output. In the third step, information is conversed into knowledge by means of evaluation and interpretation.

Analysis methods focus on identifying hidden (due to the "10 V's") patterns, rules, associations, groupings and - in general terms - information that is new, valuable, non-obvious and very hard to get manually, from Big Data using methodologies such as artificial intelligence, machine learning and statistics.

The conversion of information into knowledge, the last step of the knowledge discovery process, is of great importance to the analysis methods of Big Data. It's the step allowing information that can only answer simplistic questions such as "who", "what", "when" and "where" to be enhanced into information that has been verified and augmented by contextual information, i.e. it's useful. Thus, knowledge allows to answer the much more complicated "how" questions [3].

Traditional data mining approaches face many challenges in Big Data [30]:

Lack of scalability Design and implementation does not address scalability issues necessary for the Volume, Variety and Velocity of Big Data.

Centralisation Execution is traditionally assumed to be taking place on single information systems while having all the data into the memory, a scenario that cannot be achieved for the Volume of Big Data, among other V's.

Non-dynamic attitude The design does not tackle data that require dynamic adjustment based on analysis of the input with the Velocity of Big Data.

Input structure uniformity The design and implementation does not cater for the large Variety of Big Data.

Bearing in mind the inefficiency of traditional data mining approaches in the face of the "10 V's", a complete redesign is required in order for these approaches to be useful for Big Data. In addition, state-of-the-art Big Data analysis focuses [8] on: (a) Inference capability based on large-scale reasoning, benchmarking and machine learning due to the Volume and Variety of Big Data, (b) Stream data processing in order to cater for the Volume and Velocity of Big Data, and (c) Use of linked data and semantic approaches in order to address challenges such as efficient indexing as well as entities' extraction and classification.

2.2 Processing Resources

The advent of Big Data with all the requirements for new analytics methods as described in Section 2.1, call for a new computing paradigm. A paradigm that will allow computations not to require prohibitive long time to execute while featuring disk arrays in order to hold the volume of data. Enter the Cloud and the process of computing onto the Cloud, thus Cloud Computing. The cloud metaphor refers to an obscured to the end user set of adaptable processing

resources and networking that are accessible from anywhere, just like a cloud would do to engulfed items.

Current experience with Cloud Computing applied to Big Data usually revolves around the following sequence: preparation for a processing job, submission of the job, wait a usually unknown amount of time for results, receive none-to-little feedback as to the internal processing events and finally receive results, a paradigm that resembles awfully a lot the mainframe computing age [10] of thin clients and heavy back-ends. In contrast to the common processing scenario of much less "10 V"-complexity than Big Data - where one would enjoy direct manipulation, realtime interactivity and within seconds response/results - Big Data do not allow for any of these. In other words, current systems for Big Data processing offer a computing workflow that indeed reminds the 1960s era.

Mell & Grance [21] define Cloud Computing as "*a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*" wherein the resources can be of physical or virtual.

Cloud Computing features a number of characteristics that make it the current choice for Big Data processing:

On-demand self-service Resources and services are available for deployment to consumers without the need for human intervention on the provider's side.

Broad network access Consumers can access their allotted resources and services through a variety of client platforms almost irrespectively of the clients' mobility or processing capabilities.

Resource pooling Resources and services are pooled in a dynamic fashion in order to service all provider's customers demands, while the locality of the resource is mostly of none-to-little interest.

Rapid elasticity The provision of resources and services is furnished and released in an elastic manner in order to scale with demand in both directions of the provider. At the same time, the consumer perceives the resources and services as limitless.

Measured service Resources and services are provided as measurable with an appropriate model in order to manage and charge the consumer.

Moreover, Cloud Computing is provided with three key service models (Software as a Service, Platform as a Service and Infrastructure as a Service) while the deployment methods include private clouds, community, public and hybrid clouds[4].

As far as the actual software that manages the distributed processing and storage is concerned, Apache Hadoop[5] is one such popular nowadays solution. In fact, Apache Hadoop is a software framework (an ecosystem of modules) that manages clusters of (commodity) hardware, be these on the Cloud or in-premises. Moreover, in order to fully take advantage of processing and storage resources for

[4] An extensive presentation of the service and deployment models is outside the scope of this work. Interested readers are referred to [21].

[5] `http://hadoop.apache.org/`

parallelisable problems, programming models exist that profit from the locality of data, aiming at processing as close as possible to the storage thus taking advantage of reduced transit time and bandwidth use. A popular such model is MapReduce [7] that is also implemented as a module of Apache Hadoop.

Based on the aforementioned broad definition of Cloud Computing, its architecture cannot strictly be defined. The delivery of Cloud Computing resources and services usually involves numerous components that interact and feature using an intelligent interdependence methodology in order to provide for the elasticity characteristic. One such common component that is of great importance to the theme of this work is the storage/database component.

2.3 Storage

With the Volume of Big Data as well as the distributed nature of the processing resources utilised for Big data, it is no surprise that storage of such information is achieved by using mostly distributed approaches [29]. Accordingly, storage solutions for Big Data mostly refer to distributed file systems, Cloud storage, NoSQL databases as well as NewSQL databases. Traditional relational databases can indeed, in some occasions, address some of the "10 V" requirements but have been shown to be less efficient and thus more expensive [19]. This distributed character addresses both the need for Volume as well as solutions' scalability.

Cloud storage. The popularity of Cloud Computing has inevitably lead to the development of Cloud storage [32]. Cloud storage solutions usually aim at achieving as many and as high as possible of availability, reliability, performance, replication and data consistency. The service refers to both end-users as well as enterprises, while access is achieved through the internet with a variety of devices, as per the general characteristic of Cloud technologies. End-users usually store therein their personal data and backups, while enterprises' needs for large volume of information are supported with scalable, effective to capacity change and cheap means. Cloud storage solutions usually provide for reach interfaces that cater for stored content's dissemination as well as sharing between accredited collaborators.

NoSQL databases. Currently, the predominant solution to Big Data storage, NoSQL databases primarily focus on availability, partition tolerance and speed, usually at the cost of consistency. When compared to relational databases NoSQL solutions utilise low-level query languages, lack standardised query interfaces and offer no true ACID transactions, but for few exceptions. On the other hand NoSQL solutions are of simpler design, do not require binding fixed table schemas, offer "horizontal" scaling to clustered hardware and provide fine-grained control over availability [17]. According to the data model used, NoSQL databases are categorised in key-value stores, columnar stores, document databases and graph databases.

NewSQL databases. NewSQL databases [4] are a hybrid solution of databases between NoSQL and relational databases, that features the advantages of both origins. Thus, NewSQL databases offer the transactional guarantees of relational databases in addition to the scalability of NoSQL databases. NewSQL databases

have five main characteristics [31]: SQL query interface, support for ACID, non-locking concurrency control mechanism, significantly increased per-node performance in comparison to relational databases and bottleneck resistant scale-out, shared-nothing architecture when executing in many nodes. NewSQL databases are expected to be "*50 times faster than traditional OLTP RDBMS*" [29].

Distributed File Systems. Distributed File Systems (DFSs) address the management of storage in a network of systems. A wide variety of DFSs exists, though Hadoop File System (HDFS) [27] and Google File System (GFS) [11] have recently received most attention. As HDFS is part of ubiquitous Apache Hadoop, it has become the *de facto* DFS of choice offering the capability to store large amounts of unstructured data in a reliable way on (commodity) hardware while providing very high aggregate bandwidth across the cluster.

Big Data query platforms. A number of solutions exist that are in fact an interface for Big Data storage querying platforms. Despite these solutions differ as far as their underlying technologies, most provide SQL type query interfaces aiming at integrating with existing SQL-based applications. Hive, Impala, Spark, Drill, to name a few, are some of these "*SQL-on-Hadoop systems over HDFS and NoSQL data sources, using architectures that include computational or storage engines compatible with Apache Hadoop*" [2].

3 Collective Intelligence

Collective Intelligence features a plethora of definitions, the most notable of which being the one of Malone & Bernstein [18] as "*groups of individuals acting collectively in ways that seem intelligent*". This definition presents a number of interesting characteristics, such as:

- the lack of explicit definition of the, elusive and complex, term "intelligence" catering for greater adaptability,
- the requirement of the notion of activity by the individuals, thus emphasising on the process and not the result,
- the vague definition of the individuals' grouping that allows for varying-grain individuals and group depending on the observing level,
- the requirement of collective activity by the individuals that is agnostic to their aims as long as their activities exhibit some interdependency,
- the use of the term "seems" that allows for subjective evaluation of the manifested intelligence by the observer suiting each case.

It is of interest to note that most definitions of Collective Intelligence neglect the element of the collective's orchestration. Indeed, numerous cases exist wherein the collective is formed *ad hoc* and without some for of centralised organisation, though equally in a plethora of occasions there indeed is some entity that organises the collective so as to achieve a desired or maximise an effect. In the latter case, the incentive and purpose for using resources by the organising entity is evidently focused on the amelioration of the collective's result for the profit of the organiser and the individuals. In the former case, individual entities attempt to balance the increased resources spent in order to exhibit Collective Intelligence as peer-organisation with the ameliorated, in comparison to solitary activity,

output that provides advantages directly to the individuals. A third case organisation could also be assumed based on the lowest level of interdependency of the individuals' activities. In this coordination-free scenario, the seemingly unrelated individual actors' activity output are aggregated and curated [15] by a third party that selects the individuals based on their activities' collective focus.

Given the aforementioned definition's vagueness as well as broad spectrum of application, a number of domains relate Collective Intelligence [18]. Computer Science is a natural candidate as the notion of cooperating actors can easily be extended to include virtual entities (e.g. computational agents) as well. Cognitive Science and Biology are also related as the former focuses on the mind's functions that lead to intelligent behaviour while the latter on intelligent group behaviour and thus both provide for group intelligence matters. Social Sciences (such as sociology, political science, economics, social psychology, anthropology, organization theory, law, etc.) aim at behavioural characteristics of intelligent groups and thus relate with Collective Intelligence on matters of intelligent collective behaviour of complex entities with individual actors. Finally, Network Science is related with Collective Intelligence on matters where group intelligent activity is viewed on the context of networking between the group entities.

One of the key implications of Computer Science's relation to Collective Intelligence is the extension of intelligent behaviour to virtual entities. Thus, the range of individual actors is extended to human and virtual entities[6] while the collective to all combinations of these. Wikipedia is an example wherein, as far as Collective Intelligence is concerned, individual actors are in their overwhelming entirety contributing humans. On the other hand, weather forecasting is an example of almost solely virtual intelligences' cooperation on producing conclusions based on sensors' input. In between these extremes, Google's search engine is an example of human and virtual intelligences' cooperation given the interaction of human intelligences' created references to websites that are intelligently processed by virtual entities to furnish the search engine's functionality [25].

Focusing on the interrelation of Collective Intelligence with Big Data, a number of research directions exhibit common research challenges and methodological perspectives. Artificial Intelligence, crowdsourcing and human computer interaction, to name a few, are of interest to the theme of this work and thus presented in the sequel.

3.1 Collective Intelligence & Artificial Intelligence

Artificial Intelligence is, in general terms, the field focusing on the intelligence exhibited by machines. The first reference appears in McCarthy et al. [20] as a "*conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*". Russell & Norvig [23] organise the various definitions of Artificial Intelligence in four axes: systems that are concerned with the (a) thought process & reasoning or (b) behaviour while are evaluated based on (c) human performance or (d) an

[6] Collective behaviour in animals displaying intelligence attributes is established but outside the scope of this work. Interested readers are referred to Chapter 4 of [18].

ideal concept of intelligence. The most common methodology for the deployment of Artificial Intelligence is by means of autonomous software modules (a.k.a. agents). Agents usually exhibit characteristics such as planning for future events, assuming activity based on predefined mission criteria as well as redefining said criteria based on learning from their environment.

It is thus evident that based on the aforementioned range of individual actors of Collective Intelligence, Artificial Intelligence's agents are fulfilling almost all the breadth apart from the "solely human collective" edge. Moreover, the case of amalgamated collective intelligences of both virtual and human entities is of special interest to Artificial Intelligence as well. In any case, given the numerous Collective Intelligence alternative individual actor combinations that indeed include virtual entities, the role of Artificial Intelligence to Collective Intelligence is central in one too many ways. This is also true even in cases that Artificial Intelligence's capabilities are not used as another type of intelligence in the collective activity but solely to provide for (pre)processing of data that will independently support the activities of humans.

Machine learning, the ability of algorithms to improve performance through experience with data and predict future values of data [22], has been central to Artificial Intelligence since inception and a common requirement in Collective Intelligence. Email spam filtering, recommender systems, prediction markets, machine vision, fraud detection and biotechnology, to name a few, are some of the areas that utilise Collective Intelligence and machine learning methods in order to cope with data volume and the requirement of pattern detection.

3.2 Collective Intelligence & Human Computer Interaction

The field of research for Human-Computer Interaction (HCI) deals with the design and use of interfaces that allow the bidirectional interaction between ICT and humans. Bigham et al. in [18] refer to HCI as the study of "*the links between people and technology through the interactive systems they use*" while additionally emphasise the field's interest in both the human-to-human interaction using ICT as well as their interactions online.

HCI's relate with Collective Intelligence is clearly oriented towards Collective Intelligence's individual actors range that includes humans, given the aforementioned definition of HCI. To that end, HCI's contribution to Collective Intelligence is mainly focused on allowing the collective's members to interact in meaningful and user-friendly manner with other members, both virtual and human, in order to achieve their interdependent aim. In that sense, HCI's contribution to Collective Intelligence addresses only the point of view of the individual actors. Having described the possibility of an orchestrating entity for the activity of the collective in Section 3, HCI can additionally support Collective Intelligence by providing ICT tools that provide incentives to the individual actors while at the same time offering methodologies for the coordination and extraction of meaning and value from the collective's activity output.

An example of Collective Intelligence assisted by special HCI characteristics is the process of the musical creation industry from idea conception to the production that features understanding and support of the musical co-creative

processes dynamics. Such interfaces allow for implementation of collaboration methods and interfaces for the enhancement of the process of musical co-creation from the composition procedures up to performance and post-production levels, providing for connectivity / time / space separation of the collaborators and the artistic nature of music as well as its multifaceted creative contexts.

3.3 The Case of Crowdsourcing

Crowdsourcing, or the process of harnessing the crowd's potential in order to solve a problem, is a case of Collective Intelligence that requires special mention as it features a number of interesting, to the theme of this work, characteristics.

The original term of crowdsourcing was coined by Howe & Robinson [24] as "*outsourcing to the crowd*" but a more formal definition was provided by Estells-Arolas & Gonzlez-Ladrn-de-Guevara [9] that was the product of existing definitions' analysis, common elements extraction and establishment of basic characteristics. Their integrated solution defines crowdsourcing as "*a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken.*".

Accordingly, the parties involved in crowdsourcing fall within the classes of "requestors" that is external to the group of individuals (crowd) and have an objective and the "workers" that form the group of individuals called by the requestor to address the objective.

Apart from the obvious relation of crowdsourcing with Collective Intelligence, wherein human intelligences are cooperating in order to perform a task, crowdsourcing may also be associated, even though as a special case, with the notion of the wisdom of the crowd, wherein intelligent aggregation of insight by large number of humans (crowd) is shown to be more accurate than the majority of solutions of single members' of the crowd [33].

Moreover, crowdsourcing is utilising Artificial Intelligence and HCI advances in order to achieve its function. Artificial Intelligence is used to manage both (a) the large volume of the participants (crowd), that feature a multiplicity of capabilities and skills, and (b) the task set to be allocated to participants. Accordingly, manually addressing such issues would certainly endanger efficient task-allocation management as well as quality control. As far as HCI is concerned, both classes of workers and requestors will be engaged in a human-to-human interaction that need to be addressed. Workers' contribution is greatly affected by the appropriate incentives furnished through interaction interfaces that are user-friendly and self-explanatory, while requestors necessitate appropriate interfaces in order to coordinate and harness the output of workers.

4 Challenges & Perspectives

Having described the key pillars of both Big Data and Collective Intelligence in Sections 2 & 3 respectively, it is now necessary to address the challenges and perspectives of the shared domain of these two interrelated disciplines.

Making the most of Big Data is notoriously hard a problem, as it is common for practice for analytics to be applied with methods or on segments of data that are mostly expected to deliver results, some times in vain [26]. Bearing in mind the "10 V's" characteristics of Big Data, as described in Section 2, one must consider that not all "V's" need apply simultaneously in order for the data to be considered as Big Data. Accordingly, Volume and Velocity are enough criteria to label digital, clean, explicit and unambiguous data (collectively referred to as structured data) as Big Data. On the other hand, adding any of Variety, Variability, Vocabulary, Venue lead to analog, noisy, implicit or ambiguous data (collectively referred to as unstructured data), also labelled Big Data.

The human mind evolved to be able to process and make sense out of unstructured data by bringing to bear, seemingly unconsciously, an enormous amount of contextual knowledge [6]. On the other hand, computers excel at processing structured data [26]. Still, it is also a matter of processing power, not only type of data: increased Volume and Velocity of the data require more computing power of any sort. Accordingly, selection of the computing power to use on Big Data must be made based on the type of data to be processed: Collective Intelligence for unstructured Big Data, while Big Data efficient analytics for structured Big Data. A common such example is the statistical prediction task[7]: predictions will be made based on past information; virtual intelligence predictions in-line with past data are accurate; their accuracy drops for disruptive future information/events; human intelligences are surprising accurate, especially when Collective Intelligences.

References

1. Collective intelligence. https://en.wikipedia.org/wiki/Collective_intelligence, Accessed: July 2, 2016
2. Abadi, D., Babu, S., Özcan, F., Pandis, I.: Sql-on-hadoop systems: Tutorial. Proceedings of VLDB Endowment 8(12), 2050–2051 (2015)
3. Ackoff, R.L.: From data to wisdom. J. of applied systems analysis 16(1), 3–9 (1989)
4. Aslett, M.: Nosql, newsql and beyond. https://451research.com/report-long?icid=1651 (2011)
5. Borne, K.: Top 10 big data challenges a serious look at 10 big data vs. https://www.mapr.com/blog/top-10-big-data-challenges-%E2%80%93-serious-look-10-big-data-v%E2%80%99s (2014), Accessed: July 2, 2016
6. Byrd, D.: Organization and searching of musical information. http://homes.soic.indiana.edu/donbyrd/Teach/I545Site-Spring08/SyllabusI545.html (2008)
7. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM 51(1), 107–113 (2008)
8. Domingue, J., Lasierra, N., Fensel, A., van Kasteren, T., Strohbach, M., Thalhammer, A.: Big Data Analysis, pp. 63–86. Springer International Publishing (2016)

[7] Interested readers are referred to Chapter 5 of [18] for an extensive set of Collective Intelligence forecasting examples

9. Estells-Arolas, E., Gonzlez-Ladrn-de Guevara, F.: Towards an integrated crowdsourcing definition. Journal of Information Science 38(2), 189–200 (2012)
10. Fisher, D., DeLine, R., Czerwinski, M., Drucker, S.: Interactions with big data analytics. Interactions 19(3), 50–59 (2012)
11. Ghemawat, S., Gobioff, H., Leung, S.T.: The google file system. In: ACM SIGOPS operating systems review. vol. 37, pp. 29–43 (2003)
12. Glenn, J.C.: Collective intelligence: one of the next big things. Futura 4 (2009)
13. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of big data on cloud computing: Review and open research issues. Information Systems 47, 98 – 115 (2015)
14. Hilbert, M., López, P.: The world's technological capacity to store, communicate, and compute information. Science 332(6025), 60–65 (2011)
15. Karydi, D., Karydis, I.: Legal issues of aggregating and curating information flows: The case of rss protocol. In: International Conference on Information Law (2014)
16. Laney, D.: 3D data management: Controlling data volume, velocity, and variety. Tech. rep., META Group (2001)
17. Leavitt, N.: Will nosql databases live up to their promise? Computer 43(2)
18. Malone, T., Bernstein, M.: Handbook of collective intelligence. MIT Press (2015)
19. Marz, N., Warren, J.: Big Data: Principles and best practices of scalable realtime data systems. Manning Publications Co. (2015)
20. McCarthy, J., Minsky, M., Rochester, N., Shannon, C.: A proposal for the dartmouth summer research project on artificial intelligence. `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html` (1955), Accessed: July 2, 2016
21. Mell, P.M., Grance, T.: Sp 800-145. the nist definition of cloud computing. Tech. rep., Gaithersburg, MD, United States (2011)
22. Provost, F., Kohavi, R.: Guest editors' introduction: On applied research in machine learning. Machine Learning 30(2-3), 127–132 (1998)
23. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson (2009)
24. Safire, W.: On language. `http://www.nytimes.com/2009/02/08/magazine/08wwln-safire-t.html` (2009), Accessed: July 2, 2016
25. Segaran, T.: Programming Collective Intelligence: Building Smart Web 2.0 Applications. O'Reilly Media (2007)
26. Servan-Schreiber, E.: Why you need collective intelligence in the age of big data. `https://blog.hypermind.com/2015/01/28/the-role-of-collective-intelligence-in-the-age-of-big-data/` (2015), Accessed: July 2, 2016
27. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: IEEE Symposium on mass storage systems and technologies
28. Snijders, C., Matzat, U., Reips, U.D.: "big data" : Big gaps of knowledge in the field of internet science. International Journal of Internet Science 7(1), 1–5 (2012)
29. Strohbach, M., Daubert, J., Ravkin, H., Lischka, M.: Big Data Storage, pp. 119–141. Springer International Publishing (2016)
30. Tsai, C.W., Lai, C.F., Chao, H.C., Vasilakos, A.V.: Big data analytics: a survey. Journal of Big Data 2(1), 1–32 (2015)
31. Venkatesh, P.: Newsql the new way to handle big data. `http://opensourceforu.com/2012/01/newsql-handle-big-data/` (2012), Accessed: July 2, 2016
32. Wu, J., Ping, L., Ge, X., Wang, Y., Fu, J.: Cloud storage as the infrastructure of cloud computing. In: Intelligent Computing and Cognitive Informatics
33. Yi, S.K.M., Steyvers, M., Lee, M.D., Dry, M.J.: The wisdom of the crowd in combinatorial problems. Cognitive Science 36(3), 452–470 (2012)