

Using the Information and Communications Technology Data Deluge from a Semantic Perspective of a Dynamic Challenge: What to Learn and What to Ignore?

-Part 1-



Prof.Eng.Ph.D. Victor GREU

Abstract

The paper approaches the Data Deluge which is generated, at World scale, by the flows of data created by the complex proliferation and exponential development of Information and Communications Technologies (ICT), as main driving factor of the progress of the Information society (IS) toward Knowledge Based Society (KBS), but the paper analysis is focused on a systemic approach in order to observe the main premises and features of such complex processes, aiming the optimal efficiency of data generation and use, for humankind and Earth survival.

The main emergent (hype) technologies for ICT exponential development in 2019 are considered, as contributing to Data Deluge and including Artificial Intelligence (AI), Internet of Things (IoT), Cloud, Big Data, 3D Printing, Robotic Process Automation, Hardware Robotics, Blockchain, Augmented/Virtual Reality. ICT, but mainly all the above mentioned emergent (hype) technologies, contribute, by their complex processes that involve people, machines and devices in a planetary digital disruption, to the huge phenomenon called Data Deluge, but behind is still, mostly but not exclusively, the Internet (3.0 industry revolution) as a backbone.

For the paper entry in Data Deluge issue the CERN amazing "temple" of science and technology is chosen, but social media have also produced crucial changes in the way we live, including business models, although entertainment and other applications that use broadband mobile communications provided are also impressive in generating Data Deluge.

All these Data Deluge sources are illustrated with global figures that seem to be more impressive every year (over half a Yottabyte are generated), as the overwhelming park of connected devices and people are exponentially increasing, approximately following the Moore Law consequences as an invitation to Data Deluge.

*Although the cost of computing and communication is falling to Zero, this reality is not necessary a guarantee that the benefit in information/knowledge is automatically high, but on the contrary, **it is necessary to have higher expertise (means/methods) to extract information and eventually knowledge from the Data Deluge.***

*As prominent and relevant source of Data Deluge, CERN project is presented in detail (it includes 22 member states and a global community of 15,000 researchers). The CERN's mission is pointing international research, technology, education and collaboration. CERN will advance the frontiers of knowledge like the fundamental structure of the Universe, the generation of Universe by Big Bang, the kind of matter within the first moments of the Universe's existence, understanding the very first moments of Universe after the Big Bang or Dark Matter looking for Antimatter. In addition, CERN will develop new technologies for particle accelerators and detectors, but also for emergent fields like advanced ICT (including Quantum Computing), Web (the **World Wide Web was invented at CERN** in 1989 by British scientist Tim Berners-Lee) or the*

(computing) GRID. The medical diagnosis and therapy will also considerably benefit from the unprecedented advances achieved by the CERN. The CERN LHC is a machine of records, including: hottest spots in the galaxy; colder temperatures than outer space; the most sophisticated detectors ever built; the detectors are like gigantic digital cameras built in cathedral-sized caverns.

Concluding that CERN is just a tip of the iceberg that Data Deluge is or could become, other relevant examples could be given, but none could reach the CERN unique records (although, in the same class of Data Deluge “giants”, there are Facebook, Google, Amazon, Netflix etc.). A paper goal is **to try to analyse what and how, under these storm waves of Data Deluge, is going to melt these icebergs in order to extract and use the best of information/knowledge humankind and Earth need today and especially tomorrow.**

In the second section of the paper, a disproportional comparison of the antiquity Pyramids with CERN was deliberately chosen just to **emphasise the huge role of using the technology advances (mainly enabled by ICT) for generating Data Deluge and then extracting information and eventually knowledge**, even from sources (like the antiquity Pyramids) which almost have “run dry” before this new technological support. **It is mentioned the importance of the humankind evolution phase we are in, for the relevance of any analysis results**, as in every phase the technology advances push the data, information and eventually knowledge generation to a higher level (quite incredible before), which explains also the miracle in the case of Pyramids. The analysis also considered the deep and complex processes where data, information and eventually knowledge are linked with a multitude of goals the people could have when they expect the desired data and look, **from a semantic perspective**, for using them to fulfil these wishes. The difficulty and complexity of such analyses and optimization approaches are badly increased by the fact that all the mentioned processes premises are fast and nonlinearly changing, mainly because of ICT/IS/KBS exponential pace, this way generating **everywhere a dynamic challenge** of the mentioned **semantic perspective**, which in a simple expression could be: **What to learn and what to ignore?**

The paper also approached The Difference Between Data and Knowledge, observing that for leveraging knowledge refining it is necessary to timely think (have a thought) and create appropriate ICT tools because creating large amounts of data does not automatically generate lots of knowledge. Approaching both tools and thoughts related to the processes involved in knowledge creation in this epoque of Data Deluge, we pointed the diversity, complexity and difficulty of semantic context cases where we have to **select the optimal data (amount) leading to the desired information and eventually the knowledge benefic to wisdom**. As tools prominent example, the heart of **CERN computing infrastructure**, was given, as the Worldwide LHC Computing Grid (WLCG) includes: 170 computing centres in 42 countries; 1M CPU cores; 1EB of storage; 340 Gb/s transatlantic; 3PB moved per day.

The analysis of complex processes where data stream to information and eventually knowledge pointed the two main factors that could influence this (long) road, considering that environmental factors could be located among the Data Deluge sources (**where the Data Deluge comes from**) discussed mainly in the first section, while the cultural factors are referring to the intimate, diverse, complex and dynamic processes where data are analysed, interpreted or selected by humans or machines, usually (but not exclusively) by **semantic methods which naturally benefit from ICT prominent advances like AI/ML/CAIS**.

The final conclusion is that our **analysis needs to be further continued**, in order to get deeper (usually, but not always, scientific) insights when looking to **Data Deluge, which, in our days comes faster and everywhere**.

Keywords: Data Deluge, CERN Large Hadron Collider, semantic methods, Worldwide LHC Computing Grid (WLCG), World Wide Web, Particle Physics, Digital Disruption, Internet of Things, information society, knowledge based society, broadband mobile communications

JEL Classification: L63; L86; M15; O31; O33

*“Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?”*
— T.S. Eliot, *The Rock*

1. Where the Data Deluge comes from?

Often people say that in autumn they count the harvest results, thinking to the mature evolution the summer has generally enabled, but I guess that a similar thought, as a

metaphoric expression, could be experienced when reading about CERN (*Conseil Européen pour la Recherche Nucléaire*, a provisional body founded in 1952 with the mandate of establishing a world-class fundamental physics research organization in Europe) Open Days 2019 [2]:

“On 14 and 15 September 2019, CERN will open its doors to the public for two special days at the heart of one of the world’s largest particle-physics laboratories... With an estimated 30 000 to 40 000 visitors per day, we expect traffic disruption in the neighbouring area ... To the 75,000 visitors of the CERN Open Days: thank you for exploring the future with us and sharing our passion for science! ... The CERN Open Days have become a regular feature of the period that we call the “long shutdown” during which our accelerators stop for around two years, to benefit from upgrades and renovation work. We are now in Long Shutdown 2... Remember that CERN is open all year ...”

The comparison could be more relevant if we recall that for a football match, an audience of 40000 people is quite remarkable and further comments are not necessary, considering the impressing “harvest” that allows a stop of *around two years*, but one could still ask: Which is the point?

Perhaps from *exploring the future with us and sharing our passion for science* to a football much entertainment seems to be a long way, but our paper point is to analyse how data, information and knowledge are perceived in the digital (including Data Deluge) era we are living in, generally by **all people** on Earth.

Obviously, from here, the Information and Communications Technology (ICT) must be recalled, as the main driving engine of the Information Society (IS) toward Knowledge Based Society (KBS), because it is well known that its overwhelming influence on all humankind activity and beyond, due to the fact that ICT products and services succeeded to bring, everyday and everywhere, more and more benefits for humankind life and IS/KBS progress.

The main emergent (hype) technologies for ICT exponential development in 2019 are also largely recognized [5][6][3] as including Artificial Intelligence (AI), Internet of Things (IoT), Cloud, Big Data, 3D Printing, Robotic Process Automation, Hardware Robotics, Blockchain, Augmented/Virtual Reality etc.

Generally ICT, but mainly all the above mentioned emergent (hype) technologies, contribute, by their complex processes that involve people, machines and devices in a planetary digital disruption (DD), to the huge phenomenon called Data Deluge, but it is worth to notice that behind is still, mostly but not exclusively, the Internet (3.0 industry revolution) as a backbone.

The Data Deluge is generated by the flows of data created by the complex proliferation and exponential development of ICT at Earth scale.

As a matter of fact, the Data Deluge has a huge diversity of generation sources, but our analysis will focus on a systemic approach in order to observe the main premises and features of such complex processes, aiming the optimal efficiency of data generation and use, for humankind and ... Earth survival [7][10].

The scientific and economical world contain, naturally, the most of processes to generate data deluge [4]:

“What Causes Data Deluge?”

Global scientific output doubles every nine years (Noorden, 2014) and this combines with the recent estimate from IBM that over 90% of the world's data have been produced in the last four years (IBM, 2013). The Big Data explosion is well-documented and truly a part of modern scientific research."

Still, the wide picture of Data Deluge is far from being only about the scientific and economical World, as we already have only suggested by CERN – football comparison.

It is well known that social media have produced crucial changes in the way we live, including business models, but entertainment and other applications that use broadband mobile communications provided are also impressive in generating Data Deluge [8]:

"According to a report generated by International data corporation (IDC), a research company informs that in between 2010 and 2020, the amount of data in this digital universe will grow by 35 trillion gigabytes. In the year of 2000s, the emergence of social media, cloud computing and other powerful platforms are the main cause of exponential growth in big data. As of December 2015, Facebook has an average of 1.04 billion daily active users, 934 million mobile daily active users, available in 70 languages, 125 billion friend connections, 205 billion photos uploaded every day 30 billion pieces of content, 2.7 billion likes, and comments are being posted and 130 average number of friends per Facebook user (Facebook, 2015)."

Similarly, after 1 year, the figures seem to be more impressive, as the overwhelming park of connected devices and people are exponentially increasing, following approximately the Moore Law consequences as and invitation to ... Data Deluge [9]:

"The internet of things ... 50 billion connected devices by 2020 Digital Disruption And what of the internet of things? It is estimated the number of "things connected to the internet" could be over 50 billion by 2020. This may well be true. It could even be higher. According to Gartner, there will be nearly 21 billion devices on the internet of things by 2020. ABI Research estimates that more than 30 billion devices will be wirelessly connected to the internet by 2020. 50 billion appears to be a reasonable assumption assuming our 11 billion digital connected devices is correct.

Beyond the exponential figures we have to observe the huge implications of this stormy evolution, mainly in the business sector but fast spreading in all domains of IS/KBS as we repeatedly mentioned [20][13][16], although here the point is deeper, in the cultural and humankind evolution, both aspects of Data Deluge being very well reflected by [9]:

<< "Get Smart" will take on a whole new meaning" ...

In 2015, we probably produced more every day since the birth of the printing press in the first 555 years. The rate of increase is exponential. Videos, photos, audio messages, blog posts are streaming onto the internet in ever increasing proportion. That's lots and lots of information and lots of need to store. According to the Cisco Cloud Global Index 2015, annual global data centre traffic will reach 10.4 zettabytes by the end of 2019, up from just over 3.4 zettabytes in 2014. Global cloud traffic will reach 8.6 Zettabytes up from 2.1 ZB over the same period. Data created by the internet of everything will exceed 500 ZettaBytes by 2020 up from 150 ZettaBytes in 2015. That's over half a Yottabyte each year with more to come. That's a lot of data. How big is a Zettabyte - think Kilobyte (1) Megabyte (2) - Gigabyte (3) - Terabyte (4) - Petabyte (5) - Exabyte (6) - Zetabyte (7) and

still to come yottabyte (8). Figures in brackets are the power factors from the humble Kilobyte. Consider the Encyclopaedia Britannica is approximately 24 GB in total! ... And what data there will be! Six years ago, Google's Eric Schmidt claimed we create as much information every two days as has been created since the dawn of civilisation and the year 2003>>

The new meaning of *get smart* is in fact leading to our paper title (only when smart one knows **what to learn and what to ignore**), as ICT Data Deluge use is enabling the extraction of information/knowledge with semantic methods that mainly benefit from the ICT advances and the consequences are more than fruitful, as it is further remarked:

"Storage on the other hand is becoming more accessible to new and fast growing businesses. Personal cloud usage has increased from zero in 2010 to over one billion users this year and will increased to 2.5 billion by 2020. The volume of data stored has increased from zero 2010 to 17 exabytes by 2015 and will increase to 45 exabytes by 2020. New industries have emerged to accommodate the growth of the digital era and personal / small business cloud storage. Box.com with a market share of over 50% in the USA, lays claim to a 30% share world wide. The company still lays claim to a market cap of \$1.3 billion US."

The support of all these figures, which are naturally changing every year (or even month), is provided, as we have above mentioned, by the two pillars of ITC [11]:

"The two key technological drivers of the IT revolution are Moore's Law – the exponential increase in computing power and solid-state memory - and the dramatic increase in communication bandwidth made possible by optical fibre networks using optical amplifiers and wave division multiplexing. In a very real sense, the actual cost of any given amount of computation and/or sending a given amount of data is falling to zero."

It is worth to observe that here is explained one of the prominent mechanisms of ICT and Data Deluge proliferation, as cost of computing and communication *is falling to zero*.

On the other hand, our paper point is to show that this reality is not necessary a guarantee that the benefit in information/knowledge is automatically high, but on the contrary, **it is necessary to have higher expertise (means/methods) to extract information and eventually knowledge from the Data Deluge**.

We still have to add that recently the communications pillar is clearly supported and unprecedentedly expanded by the **advances of broadband mobile communications**, starting with 4G and now rising with 5G on a level that could be a new revolution (industry 4.0) of IS/KBS exceeding the ICT area in conjunction with IoT [13]17[19][20].

Broadband mobile communications and IoT will probably become the main support contributors of the Data Deluge, but the mechanisms of increasing the data flows are more that the technical capacity, including the complex processes where the new applications enabled by the 5G low latency (1ms) and high speed (20 Gb/s) will leverage the people and devices to generate more and more ... data.

The above presented premises and features created a frame for Data Deluge and most important for what it could be for humankind and Earth in IS toward KBS, but a relevant example would clearly complete the actual image and add credible forecasts for its evolution.

By the next synthetic description, based on the comprehensive presentation of [1], it will be evident why for the paper entry in Data Deluge issue we have chosen the CERN amazing “temple” of science and technology.

CERN is the European Laboratory for Particle Physics, located at the Franco- Swiss border near Geneva. ***CERN has been pushing the boundaries of knowledge and technology for more than 60 years.***

The CERN project includes 22 member states and a global community of 15,000 researchers, but there are also 3 observers and 8 associate states (Romania being an associate member), having a general world-wide budget 1100 MCHF (2017). Experiments are run by collaborations of scientists from institutes all over the world.

The CERN’s mission is well tailored for such structure and budget, pointing international research, technology, education and collaboration. As a consequence, CERN will advance the frontiers of knowledge like the fundamental structure of the Universe, the generation of Universe by Big Bang, the kind of matter within the first moments of the Universe’s existence, understanding the very first moments of Universe after the Big Bang or Dark Matter looking for Antimatter.

In addition, CERN will develop new technologies for particle accelerators and detectors, but also for emergent fields like advanced ICT (including Quantum Computing), Web (the ***World Wide Web was invented at CERN*** in 1989 by British scientist Tim Berners-Lee) or the (computing) GRID. The medical diagnosis and therapy will also considerably benefit from the unprecedented advances achieved by the CERN.

For an accurate and astonishing image of the main features and performances of **CERN infrastructure for Particle Physics**, the next details are also extracted from [1]:

“The LHC (Large Hadron Collider) is the world’s largest and most powerful particle accelerator. It is built around 100 m underground and has a circumference of 27 km.

The LHC is a machine of records, including:

- *HOTTEST spots in the galaxy*
- *COLDER TEMPERATURES than outer space*
- *The FASTEST RACETRACK on the Planet*
- *The Most Powerful MAGNETS*
- *The Highest VACUUM*
- *The Most SOPHISTICATED DETECTORS ever built: CMS ALICE ATLAS LHCb*
- *The particles are accelerated to close to the speed of light*
- *The detectors are like gigantic digital cameras built in cathedral-sized caverns.*

Two general-purpose detectors cross-confirm discoveries, such as the Higgs boson:

- *ATLAS 46m long, 25m diameter weights 7’000 tonnes 100 million electronic channels , 3 000 km of cables*
- *CMS 22m long, 15m diameter weights 14’000 tonnes Most powerful superconducting solenoid ever built.”*

A remarkable feature is that CERN research is very mature planned, as the two *general-purpose detectors* (Atlas and CMS) are designed to *cross-confirm discoveries* (like the Higgs boson is a historical one), this way the results have a high degree of confidence and reliability, reflected also by the other two detectors:

“ALICE and LHCb (the “b” stands for beauty) experiments have detectors specialised on studying specific phenomena. LHCbALICE Studies the «Quark Gluon Plasma», state of matter which existed moments after the Big Bang. Studies the behaviour difference between the b quark and the anti-b quark to explain the matter-antimatter asymmetry in the Universe.

Collisions generate particles that decay in complex ways into even more particles. Up to about 1 billion particle collisions can take place every second. This can generate up to a petabyte of data per second. Filtering the data in real time, selecting potentially interesting events (trigger).”

After these impressing performances of **CERN infrastructure for Particle Physics**, is hard to add something, but it is worth to mention that, just like other prominent achievements of IA/KBS where ICT is pushing exponentially all the limits, CERN infrastructure has a dynamic development (i.e. the story goes on higher), which include 6 runs from 2010 to 2038.

Still, the fact that CERN experiments are run by collaborations of scientists from research organizations all over the World, although seem less spectacular, has also huge consequences either for the general scientific results and their impact on IS/KBS but especially for the impact on the CERN contribution to Data Deluge, by CERN computing infrastructure and especially by what at CERN is called GRID (an international ICT network which supports even the real time processing of CERN results).

Perhaps another astonishing aspect is that CERN is planning to extend LHC to a new one, about 100km long, connected with the actual one of 27km longness [12].

While these types of forecasts send us nearly to science fiction, the heart of **CERN computing infrastructure**, i.e. of the Data Deluge CERN contribution, will be further valued and better perceived in the next section (where we have to analyse **how** anything about Data Deluge could come **faster and ... everywhere**), because here (**where the Data Deluge comes from**) we have seen just the tip of the iceberg that Data Deluge is or could become, as this section might have been continued with other relevant examples, but we think that none could reach the CERN unique records (although, in the same class of Data Deluge “giants”, we could add Facebook, Google, Amazon, Netflix etc.).

Anyway, further we have **to try to analyse what and how, under these storm waves of Data Deluge, is going to melt these icebergs in order to extract and use the best of information/knowledge humankind and Earth need today and especially ... tomorrow.**

2. The perpetual dilemmas of selecting information now come faster and ... everywhere

After observing the CERN amazing “temple”, another huge “image” that could be recalled from ancestral World history is the antiquity Pyramids (Egyptian or Mesoamerican), as they still continue to generate new data for the archaeological researchers.

Of course, the CERN huge dimensionality are astonishing by its multilateralism, including the unique combination of dimensions like physical, electric and magnetic power, speed of light, high tech, number of scientists or ... data flows.

On the other hand, the Pyramids mainly impress by their physical, cultural or even high tech achievements, considering the realities of thousands years ago.

Still, for both “giants” we can observe, in our days, a common feature: they generate a lot of data, **but keeping the scale**, which is really a miracle in the case of Pyramids.

Here the point is to reveal the fact that the great humanity achievements have a special power to generate data, information and eventually knowledge, **but at a level which depends on humankind evolution phase we are in**.

With other words, we have to notice the incredible differences this recent evolution created, in the last decades, especially on high tech and ... data flows, i.e. the Data Deluge.

More than these, the disproportional comparison of the antiquity Pyramids with CERN was deliberately chose just to **emphasise the huge role of using the technology advances (mainly enabled by ICT) for generating Data Deluge and then extracting information and eventually knowledge**, even from sources (like the antiquity Pyramids) which almost have “run dry” before this new technological support.

Anyway, this section is generally focusing on the analysis of the deep and complex processes in which people generate or perceive data, information and eventually knowledge, aiming to optimize their efficiency for a sustainable development of IS/KBS, considering, as we have already presented [18][10], that the exponential pace of ICT and other progress factors could have side consequences, first including climate changes, Earth resources fading and humankind health on long term.

It is worth to notice from the start **the importance of the humankind evolution phase we are in, for the relevance of any analysis results**, as the above revealed that in every phase the technology advances push the data, information and eventually knowledge generation to a higher level (quite incredible before), which explains also the miracle in the case of Pyramids.

For clarity, here we just shortly mention the crucial interest for other cases, like DNA (human genome) and generally for biogenetic engineering, where long expected results would have to appear just in our days phase, when the incredible advances of ICT and other associated technologies (like nanotechnology) have the chance to push the results, i.e. data, at content and semantic levels that could further provide information and eventually knowledge with the desired practical and sustainable applications/implications and then benefits (sometimes considered impossible until recently).

This way we have entered in the deep and complex processes where data, information and eventually knowledge are linked with a multitude of goals the people could have when they expect the desired data and look, **from a semantic perspective**, for using them to fulfil these wishes.

The difficulty and complexity of such analyses and optimization approaches are badly increased by the fact that all the mentioned processes premises are fast and nonlinearly changing, mainly because of ICT/IS/KBS exponential pace, this way generating **...everywhere a dynamic challenge** of the mentioned **semantic perspective**, which in a simple expression could be: **What to learn and what to ignore?**

Surely this question is not simple to be answered, nor if we find for it another form, like **the perpetual dilemma of selecting information**, considered in our days, when ... it comes faster and ... everywhere.

For starting to understand the mentioned complex processes, it is important to approach *The Difference Between Data and Knowledge*, as it is also intended in [4]:

“Data is the lowest tier of the classical Data, Information, Knowledge, Wisdom (DIKW) pyramid (Ackoff, 1989) and forms the basis of our knowledge and wisdom. We interpret ‘data’ to generate a tangible insight that we then call ‘knowledge’ but this distinction is often missed, and the terms of data and knowledge are frequently used interchangeably. Simply having data does not create knowledge and creating large amounts of data does not automatically generate lots of knowledge. Instead, more thought and more tools are needed to help us to ascend this value chain.”

Here we find a clear confirmation of our continuous preoccupation [7], as for leveraging knowledge refining it is necessary to timely think (have a thought) and create appropriate ICT tools [3], because *creating large amounts of data does not automatically generate lots of knowledge*.

As a matter of fact, in this paper, we approach both *tools* and *thoughts* related to the processes involved in knowledge creation in this epoche of Data Deluge, but pointing the diversity, complexity and difficulty of semantic context cases where we have to **select the optimal data (amount) leading to the desired information and eventually the knowledge benefic to ... wisdom**.

Speaking of *tools*, now is the case to give as relevant example the heart of **CERN computing infrastructure**, i.e. of the Data Deluge CERN contribution [1]:

“The CERN data centre(s) processes hundreds of petabytes of data every year. CERN’s data centre in Meyrin is the heart of the laboratory’s computing infrastructure... The Wigner data centre in Budapest serves as an extension to the one in Meyrin. The two centres are connected by three 100 Gb/s fibre-optic links. MEYRIN CENTRE (CH): 300,000 processors cores 180 PB on disk 230 PB on tape ... WIGNER CENTRE (H): 100,000 processors cores 100 PB on disk...The two centres are connected by three 100 Gb/s fibre-optic links.”

Notice that **CERN computing infrastructure** is more than a computing core, i.e. a network which leverages, by international collaboration, the knowledge refining with a fast pace, close to ICT development pace.

The CERN computing infrastructure is further used along with an ICT huge network, called Worldwide LHC Computing Grid (WLCG) by the CERN researchers. WLCG is used sometimes in real time, for leveraging the best results [1]:

“Physicists must sift through the 30-50 PBs produced annually by the LHC experiments...The WLCG gives thousands of physicists across the globe near real-time access. Tier-0 (CERN and Hungary): data recording, reconstruction and distribution. Tier-1: permanent storage, re-processing, analysis. Tier-2: Simulation, end-user analysis...The Worldwide LHC Computing Grid integrates computer centres worldwide to combine computing and storage resources into a single infrastructure accessible by all LHC physicists...

With 170 computing centres in 42 countries, the WLCG is the grid that never sleeps!

The size of WLCG...

~1M CPU cores;

~ > 2 million jobs/day;

~1EB of storage;
~ 10-100 Gb/s links;
~ 340 Gb/s transatlantic;
~3PB moved per day.”

Generally speaking and leaving the CERN case, the analysis of complex processes where data stream to information and eventually knowledge, should continue pointing the main factors that could influence this (long) road, as it is also agreed by [4]:

“On a practical level, most scientific instrumentation can produce many channels of data for processing and further analyses. However, scientific knowledge generation from these data sources can be hampered by various limiting factors, each of which becomes more impactful as the volume and diversity of data increase. Each of these factors can work alone or in concert. If not controlled, however, these factors can combine to reduce the ability to produce information, and therefore scientific knowledge.

Looking at this in detail, we can quickly generate a shortlist of limiting factors that fall into two broad categories, classified as environmental and cultural factors.

Environmental factors can be defined as attributes that relate to the process of generating knowledge from data.

Conversely, cultural factors, when generating scientific insight, can arise as a consequence of how our science training directs us to look at data and are oftentimes institutionally/ historically-based in their origin.”

We appreciate that this processes image, although general and perhaps too abstract, is very comprehensive in synthetically covering such complex diversity.

In reality, both classes of factors are very difficult to be completely separated, because they are related with the same complex processes, where data are generated, either by humans and machines, while the most critical phases, the information acquiring and then knowledge generation are usually performed by humans, but lately AI by machine learning (ML) and soon by cognitive AI systems (CAIS) tend to approach just this high level of human intelligence (knowledge generation).

As a matter of fact, recent advances of AI are more and more efficient for extracting information (and even knowledge) from data, but both humans and AI machines use mainly semantic methods for selecting that information.

The main problem, as a dynamic challenge, remain the level of the desired knowledge, or, in a simpler expression, the ability of humans or machines to refine knowledge, because in fact, in any domain we could consider **a progress only when the old information/knowledge is improved/replaced by the new one (a refined one) with a higher efficiency/capacity to be useful and this way we have just arrived to the obsessive question: What to learn and what to ignore?**

We have to observe that, in spite of all our efforts to detail and then optimize these processes where data stream to information and eventually knowledge, every step, no matter how deeply described, contains relative and difficult to quantify elements, which in addition are time sensitive, i.e. **we face a dynamic challenge** and we have already compared such situations with *Morgan le Fay* (*Morgana mirage*) [3] [14].

The inherent causes of such a dramatic reality are linked with the ever changing human perceiving of the World (human personality subjectivity), which is amplified by the

exponential advances of ICT (which enable access to a large diversity of data/options from where people could choose/select everyday), because these two features are in a complex mutual dependence when generating the fundamental (above mentioned) *tools* and *thoughts*.

In the actual World, we would consider that *environmental factors* could be located among the Data Deluge sources (**where the Data Deluge comes from**) we have discussed mainly in the first section, while the *cultural factors* are referring to the intimate, diverse, complex and dynamic processes where data are analysed, interpreted or selected by humans or ... machines, usually (but not exclusively) by **semantic methods which naturally benefit from ICT prominent advances like AI/ML/CAIS**.

Arriving at this point, it is clear that our **analysis needs to be further continued**, in order to get deeper (usually, but not always, scientific) insights when looking to **Data Deluge, which, in our days comes faster and ... everywhere**.

3. Conclusions

The Data Deluge is generated by the flows of data created by the complex proliferation and exponential development of ICT at Earth scale and has a huge diversity of generation sources, but the paper analysis is focused on a systemic approach in order to observe the main premises and features of such complex processes, aiming the optimal efficiency of data generation and use, for humankind and Earth survival.

The main emergent (hype) technologies for ICT exponential development in 2019 are considered, as contributing to Data Deluge and including Artificial Intelligence (AI), Internet of Things (IoT), Cloud, Big Data, 3D Printing, Robotic Process Automation, Hardware Robotics, Blockchain, Augmented/Virtual Reality.

Generally ICT, but mainly all the above mentioned emergent (hype) technologies, contribute, by their complex processes that involve people, machines and devices in a planetary digital disruption, to the huge phenomenon called Data Deluge, but it is worth to notice that behind is still, mostly but not exclusively, the Internet (3.0 industry revolution) as a backbone.

Because the scientific and economical World contain, naturally, the most of processes to generate Data Deluge, for the paper entry in Data Deluge issue we have chosen the CERN amazing “temple” of science and technology. Also, social media have produced crucial changes in the way we live, including business models, but entertainment and other applications that use broadband mobile communications provided are also impressive in generating Data Deluge.

All these Data Deluge sources are illustrated with global figures that seem to be more impressive every year (*over half a Yottabyte* are generated), as the overwhelming park of connected devices and people are exponentially increasing, approximately following the Moore Law consequences as an invitation to Data Deluge.

Although the cost of computing and communication *is falling to Zero*, this reality is not necessary a guarantee that the benefit in information/knowledge is automatically high, but on the contrary, **it is necessary to have higher expertise (means/methods) to extract information and eventually knowledge from the Data Deluge**. This way, broadband mobile communications and IoT will probably become the main support contributors of the

Data Deluge, but the mechanisms of increasing the data flows are more than the technical capacity, including the complex processes where the new applications enabled by the 5G low latency (1ms) and high speed (20 Gb/s) will leverage the people and devices to generate more and more data.

As prominent and relevant source of Data Deluge, CERN project is presented (it includes 22 member states and a global community of 15,000 researchers). The CERN's mission is pointing international research, technology, education and collaboration. CERN will advance the frontiers of knowledge like the fundamental structure of the Universe, the generation of Universe by Big Bang, the kind of matter within the first moments of the Universe's existence, understanding the very first moments of Universe after the Big Bang or Dark Matter looking for Antimatter. In addition, CERN will develop new technologies for particle accelerators and detectors, but also for emergent fields like advanced ICT (including Quantum Computing), Web (the **World Wide Web was invented at CERN** in 1989 by British scientist Tim Berners-Lee) or the (computing) GRID. The medical diagnosis and therapy will also considerably benefit from the unprecedented advances achieved by the CERN. The CERN LHC is a machine of records, including: hottest spots in the galaxy; colder temperatures than outer space; the most sophisticated detectors ever built; the detectors are like gigantic digital cameras built in cathedral-sized caverns.

Concluding that CERN is just a tip of the iceberg that Data Deluge is or could become, other relevant examples could be given, but we think that none could reach the CERN unique records (although, in the same class of Data Deluge "giants", we could add Facebook, Google, Amazon, Netflix etc.). Anyway, further we have **to try to analyse what and how, under these storm waves of Data Deluge, is going to melt these icebergs in order to extract and use the best of information/knowledge humankind and Earth need today and especially tomorrow.**

In the second section of the paper, a disproportional comparison of the antiquity Pyramids with CERN was deliberately chosen just to **emphasise the huge role of using the technology advances (mainly enabled by ICT) for generating Data Deluge and then extracting information and eventually knowledge**, even from sources (like the antiquity Pyramids) which almost have "run dry" before this new technological support. **Notice the importance of the humankind evolution phase we are in, for the relevance of any analysis results**, as the above revealed that in every phase the technology advances push the data, information and eventually knowledge generation to a higher level (quite incredible before), which explains also the miracle in the case of Pyramids. The analysis also considered the deep and complex processes where data, information and eventually knowledge are linked with a multitude of goals the people could have when they expect the desired data and look, **from a semantic perspective**, for using them to fulfil these wishes.

The difficulty and complexity of such analyses and optimization approaches are badly increased by the fact that all the mentioned processes premises are fast and nonlinearly changing, mainly because of ICT/IS/KBS exponential pace, this way generating **everywhere a dynamic challenge** of the mentioned **semantic perspective**, which in a simple expression could be: **What to learn and what to ignore?**

The paper approached *The Difference Between Data and Knowledge*, observing that for leveraging knowledge refining it is necessary to timely think (have a *thought*) and create

appropriate ICT *tools* because *creating large amounts of data does not automatically generate lots of knowledge*. Approaching both *tools* and *thoughts* related to the processes involved in knowledge creation in this epoque of Data Deluge, we pointed the diversity, complexity and difficulty of semantic context cases where we have to **select the optimal data (amount) leading to the desired information and eventually the knowledge benefic to ... wisdom.**

As *tools* prominent example, the heart of **CERN computing infrastructure**, was given, as the Worldwide LHC Computing Grid (WLCG) includes: *170 computing centres in 42 countries; 1M CPU cores; 1EB of storage; 340 Gb/s transatlantic; 3PB moved per day.*

The analysis of complex processes where data stream to information and eventually knowledge pointed the two main factors that could influence this (long) road, considering that *environmental factors* could be located among the Data Deluge sources (**where the Data Deluge comes from**) we have discussed mainly in the first section, while the *cultural factors* are referring to the intimate, diverse, complex and dynamic processes where data are analysed, interpreted or selected by humans or ... machines, usually (but not exclusively) by **semantic methods which naturally benefit from ICT prominent advances like AI/ML/CAIS.**

The final conclusion is that our **analysis needs to be further continued**, in order to get deeper (usually, but not always, scientific) insights when looking to **Data Deluge, which, in our days comes faster and ... everywhere.**

REFERENCES

- [1] Maria Girone, *Tackling Tomorrow's Computing Challenges Today at CERN*, ISC 2018, Published in: Technology, Aug 1, 2018
- [2] ***, *CERN Open Days 2019 | 14–15 September*, 2019, <https://opendays.cern/>.
- [3] Victor Greu, *The information and communications technology is driving artificial intelligence to leverage refined knowledge for the World sustainable development– (Part 2)*, Romanian Distribution Committee Magazine, Volume 10, Issue 1, Year 2019.
- [4] Kevin Teburi, *How to Stem the Tide of Data Deluge to Produce Better Science in Biopharma R&D* ibi teburi pdf how to stem the tide of data deluge- international biopharmaceutical industry summer/autumn 2018 volume 1 issue 2.
- [5] ***, *Emerging Technologies: Changing how we live, work and play*, EY-Mint Emerging Technologies Report 2019, <https://www.ey.com/ey-mint-emerging-technologies-report-2019>
- [6] ***, *Tech Trends 2019 - Deloitte*, https://www2.deloitte.com/content/dam/Deloitte/fi/Documents/technology/DI_TechTrends2019.
- [7] Victor Greu, *Extending information and communications technologies' impact on knowledge based society through artificial and collective intelligence –(Part 3)*, Romanian Distribution Committee Magazine, Volume 9, Issue 3, Year 2018.
- [8] Mudassir Khan, *Big data analytics emerging trends, technology and innovations for the future business in the global market*, International Journal of Scientific Research and Review, Volume 8, Issue 2, 2019

[9] John Ashcroft, *Six Things Everyone Should Know About Digital Disruption*, The Saturday Economist Published in: Business, Jul 6, 2016

[10] Victor Greu, *Searching the right tracks of new technologies in the earth race for a balance between progress and survival*, Romanian Distribution Committee Magazine, Volume 3, Issue1, Year 2012.

[11] Tony Hey, Anne Trefethen, *The Data Deluge: An e-Science Perspective*, <http://www.computing.surrey.ac.uk/courses/csm23/Papers/DataDeluge.pdf>], Wiley, 2003.

[12] Harry Cliff, *CERN: Large Hadron Collider replacement plans unveiled – here's what it could discover*, January 18, 2019, <https://theconversation.com/cern-large-hadron-collider-replacement-plans-unveiled-heres-what-it-could-discover-109983>.

[13] Victor Greu et al, *Human and artificial intelligence driven incentive-operation model and algorithms for a multi-purpose integrated crowdsensing-crowdsourcing scalable system* - paper submitted to International Conference Communications 2018 (Politehnica University of Bucharest, Romania Military Technical Academy, IEEE Romania), June 2018.

[14] Seth Earley, *There Is No AI Without IA*, 2016, IEEE IT Professional (Volume: 18, Issue:3, May-June 2016)

[15] Niko Mohr, Holger Hürtgen, *Achieving business impact with data - A comprehensive perspective on the insights value chain*, Digital McKinsey 04.2018 Copyright © McKinsey & Company www.mckinsey.com

[16] Victor Greu, *The Exponential Development of the Information and Communications Technologies – A Complex Process Which is Generating Progress Knowledge from People to People*, Romanian Distribution Committee Magazine, Volume 4, Issue2, Year 2013

[17] Mike Sirius, *How artificial intelligence and innovation will interact*, June 5, 2017, Idea Drop Ltd, <http://ideadrop.co/artificial-intelligence-innovation/>

[18] Victor Greu, *Tomorrow's paradox: refining knowledge by smarter information and communications technologies while humans tend to become a limited factor of performance*, Romanian Distribution Committee Magazine, Volume 7, Issue1, Year 2016.

[19] Bhupinder Kour, *The Rise of Machine Learning and AI is Improving Lives in 2018*, <https://www.smartdatacollective.com/rise-of-machine-learning-ai-improving-lives/>

[20] Victor Greu, *Information and communications technologies drive digital disruption from business to life on earth -(Part 1)*, Romanian Distribution Committee Magazine, Volume 8, Issue 2, Year 2017.