A Theory of Big Data

Jan de Meer[1]

Abstract: CLAIRE the 'initiative of a pan-EU confederation of AI Research Labs'[Cl18] anticipates a humane AI which is based on ethical and trustworthy tenets empowering citizens and society. Hence in order to achieve a human AI, new semantic categories of standards must be written enabling stakeholders to implement a responsible AI. Communication however generates lots of unstructured data sets to be classified and structured into data types. AI-based algorithms are suitable to derive - from data sets and data types - more sophisticated and implicitly given information that can further be enriched to knowledge about autonomous communicating processes or even autonomous behaving humans. The knowledge gained from applying data enrichment algorithms in turn may be used for reasoning and prediction purposes, thus improving applications a lot. The aim of this project is to find a common unified tool able to handle big data flows and types that allows the recognition of hidden information from both data flows and data types. The hidden information is sometimes also called meta-data, thus being implicitly existing but by tools becoming explicitly seen.

Keywords: Big Data Analysis, AI Standardization, Big Data Lake, Data Science, Data, Information and Knowledge, IACS.

1. Introduction

CLAIRE the 'initiative of a pan-EU confederation of AI Research Labs'[Cl18] anticipates a humane AI which is based on ethical and trustworthy tenets empowering citizens and society. Hence in order to achieve a human AI, new semantic categories of standards must be written enabling stakeholders to implement a responsible AI.

This kind of Responsible AI however deals with both, i.e. avoiding risks and biases and to achieve socio-economic impacts improving and making safe and secure our human lives, work and daily communications. Thus the new standards are catalysts for innovations and thus shall provide frameworks, guidelines, rules, evaluation and measurement methods and techniques on how to gain trust in and safety of socio-technical ecosystems in which citizens have to communicate with various infrastructures as their digital twins.

Although the notions of Big Data Flows & Types (BD F/T [Cc18]) in their intended semantics differ from Big Data Science (BDS), such that the two notions are used partially synonymously. Whereas 'Data Science' in the context of the ISO/IEC standard

[1] smartspacelab,eu GmbH, Berner Str.21B, 12205 Berlin, demeer@ACM.ORG

doi:10.18420/inf2019_ws30

FDIS 20546 'Big Data Vocabulary', it is referred to as the 'process of extracting actionable knowledge from data' (i.e. 42AA N0767 Vocabulary Clause 3.1.10) and 'Data Type' is defined as '*a set of data objects of a specified data structure and a set of permissible operations …*'

2. Data, Information and Knowledge

2.1. AI Standardization Framework

The ISO/IEC standardization committee JTC1/SC42 on AI recently has herself restructured into 4 working groups(WG) and several study or ad-hoc groups (SG/AhG) in order to serve for the advanced working programme on standardizing frameworks, concepts, methods, vocabularies and architectural requirements on AI:

1. WG1 is on foundational standards that cope with AI concepts and AI terminology necessary for the full AI life cycle;
2. WG2 is on big data that aims at vocabulary, framework and reference architecture for big data;
3. WG3 deals with requirements for trustworthy and bias-free AI systems that include assessment of robustness of Neural Networks;
4. WG4 is oriented towards applications and use cases to demonstrate feasibility on AI standards
5. SG1 investigates into computational approaches comprising Machine Learning (ML) algorithms, reasoning approaches, NLP etc.
6. SG2 investigates into aspects of trustworthiness and pitfalls, whereas the former aspects deal with system properties like transparency, verifiability, explainability, controllability and the latter aspects deal with robustness, safety, security, privacy system properties.
7. a new standardization project NWIP 24300 is planned to come and is related to the AI process management for Big Data Analysis (BDA) [Pm18].

The 'AI Technical Report (TR) on trustworthiness' [Ai18] and the IEC white paper on 'AI across Industries' [Bd18] present the following somehow diverging definitions of AI:

- AI is the 'capability of a functional unit (of a system) that are generally associated with human intelligence, such as reasoning and learning'.

- AI is a 'branch in computer science that simulates intelligent behavior in computers including problem solving, learning or pattern recognition'.

The AI definitions above are said to be diverging, because the subjects of reference are different. In the first case the AI capabilities of a functional unit are compared to human intelligence, i.e. cognitive capabilities - whereas in the second case the AI capabilities are restricted to machines, i.e. to computers obeying just computational capabilities.

2.2. AI Vocabularies

The aim of this project is to find a common unified tool able to handle big data flows and types that allows the recognition of hidden information from both data flows and data types. The hidden information is sometimes also called meta-data, thus being implicitly existing but by tools becoming explicitly seen.

Vocabularies and semantics shall aim at formal description of specific characteristics of industrial standards like the 'Industrial Automation and Control systems (IACS)' series that describe Cyber Security Requirements of Industrial Systems and Components in:

1. IEC 62443-4-2 (IEC 65/735/FDIS) IACS - part 4.2 Technical Security Requirements for IACS Components:2018 [Ts18];

The compound bracket of the aforementioned three documents on big data is in-fact a common vocabulary and semantic notions that share the issue of Big Data generated, managed and applied in technical systems such as IACS. From the standard's point of view a common normalizing theory on knowledge gaining from data is mandatory to avoid diverging semantic understandings in shared domains.

A triggering point of thinking about a combined Big Data Theory comprising Flows and Types are the current standardization projects of European and International Standardization Organizations that are related to the issue of Big Data Analysis in Industrial System Environments and thus in the forthcoming Digital Society, e.g:

1. ISO/IEC 20547-3 Big Data Reference Architecture (BDRA) - part 3:DIS 10-2018 [RA18], [Cc18]
2. ISO/IEC 20546 Big Data Overview and Vocabulary (BDOV):FDIS 01-2019 [Bd18]
3. CEN/CLC/JTC13 Work Programme 2019 to achieve the Digital Society [Rm19].

In ISO 20547-3 (DIS:2018)[Ra18] 'Big Data' is defined as 'extensive data sets' (respectively data types) obeying the so-called 4V-Characteristics of Volume, Variety, Velocity

and Variability requiring a scalable architecture for efficient storage, data manipulations and analysis.

Whereas the characteristic of Volume stands for the extensive amount of data; the characteristic Variety means a high number of different data types respectively data domains; the characteristic Velocity is defined by the rate of data flow and the characteristic Variability means the variability of data rates, formats, structures or qualities.

Furthermore Big Data requires constraints on Data Governance, Data Quality and Data Veracity. Governance means to develop a strategic plan for the Data Asset Management, Quality means implied data needs to be satisfied under specific conditions, and veracity means to have a notion of Completeness referring to trustworthiness with respect to applicability, noise, bias, abnormality and other quality constraints.

The part 3 - the definition of the BD reference architecture - of the document DIS 20547 series comprises just two views, the user view and the functional view to define the concepts of the BDRA. This is not enough since Big Data must deal with efficient data analytics. Hence at least a third view - an information point of view is definitely needed yet not given in DIS20547. At the other side, the functional viewpoint of the current document is overloaded with the missing notions from the information viewpoint, although there are additional roles and activities defined. (Notice the cross-cutting aspects of 20547-3 contains under the issue of data governance the definition of a governance and management view)

The second single document FDIS 20546 defines the anticipated vocabulary partially in the document sections of ‚terms and definitions' and partially in the sections ‚Key Characteristics' and ‚Cross-cutting Concepts'.

3. Communication Semantics

3.1. Use Case ‚Smart Grid Solar Panels'

Information and data must be distinguished since data is so-to-say the carrier of information that must be made explicit by suitable analytic methods in order to be applied correctly (with respect to safety and security criteria) in applications. Knowledge and information must be distinguished since information describes the context of the related flow of data [Cc18] generated somewhere and from something (e.g. sensors); however knowledge describe the circumstances respectively the constraints that happened and on that can be reasoned on a certain given data value. Hence knowledge of a certain context can be considered to be equivalent with the semantics of data because of the obeyed reasoning capability.

For example: assume a temperature sensor in the field that measures in a certain rate the temperature. Before application it has been agreed that the measurement is represented

as a data<n-tuple> comprising several elements (n=5) that describe the possible event of a field temperature measurement of a) the sensor ID, b) the time of measurement (date, UTC), c) the measured value of temperature (integer, float), d) the assigned metric of temperature (^{o}C, ^{o}F) and in case f) the intended rate of measurements (sec^{-1}). It must be noted that the digital coding of this data<n-tuple> is from a semantical point of view out of interest and can be chosen anyway arbitrarily but luckily enough it is appropriately standardized (e.g. ANSI coding) for most cases of applications.

The information that can be gained from a stream of such type of data (<n-tuples>) is taken from an application point of view and is thus not objective. Information is subjective because the application defines the constraints respectively the invariants under which the temperature measurement is meaningful, e.g. when the temperature is measured on the surface of solar panels, i.e. the temperature must not overrun a certain threshold during operation.

Hence the ‚semantics of solar panel surfaces' of the given example is defined by its characteristic invariants, i.e. the surface temperature evaluated in real-time, its limits to avoid destructions and the frequency of measurements to recognize trends etc. With this information the measurement stream is constrained and thus filtered since not all measurements are so interesting to be kept forever.

3.2. Graph-theoretical representation of Semantics

In a theory of Big Data semantics is represented by Graphs comprising edges and vertices (Notice: whereas a ‚vertex' means a node of a graph, and a ‚node' means a component from a certain system architecture, in order to make the difference between a mathematical graph and a technical system - which are related to each other by a homomorphism).

Up to now we have a data<n-tuple> stream generated by a sensor network and - with respect to the solar-panel application - a collection of specific invariants, constraining the stream and thus generating conditional events (e.g. IF the sun shines in the expected way THEN no preventive action must take place (i) - or vice versa: IF the sun irradiation gets too high AND thus the temperature overruns the safe limits THEN the panels must be protected from destruction (ii)).

Conditional events in graph theory are modeled by graph edges. A graph edge however is a directed or undirected pair of vertices, whereas an undirected edge models a relation between two nodes of a system (e.g. a network of solar panels under ‚normal' conditions and a network of solar panels under exceptional conditions) - a directed edge models the change of system state form normal to exceptional conditions. The difference between an event and a relation is that a relation has the potential to perform a system state it models so-to-say the static property whereas events model dynamic properties. Since edges defined as pairs if vertices that model potential events, the initial vertex is called ‚event head' and the resultant vertex is called ‚event tail'. In the example above we have

indicated two edges (i) and (ii) of possible state changes where the nodes contain all system variables describing the invariant conditions.

4. Characteristics of Devices

Having above invented the notion of semantics for the purpose of BD Analytics, the characteristics of BD can be stated as below. In order to express these characteristics in terms of typing, quantification, safety, privacy and security properties an adequate language based on formal semantics respectively a vocabulary denoting to semantic notions is required. Thus a suitable Big Data Vocabulary is able to express necessary characteristics but is not restricted to:

1. parallel evaluation of BD information [cp. 20547-3 Functional View - Graph Storage;

2. system scaling with respect to system resources (vertical scaling) and with respect to BD properties (horizontal scaling). Vertical scaling measures comprise system nodes like processors, memories, storages, connectivity - horizontal scaling measures comprise data descriptive notions like volume, velocity, variety, and variability.

3. metrics for the data measurement transparency with respect to data volume, velocity, variety, variability e.g. bit, pixels, bit sec^{-1}, security level, QoS level etc.

4. Formal Semantics by means of Graph and Set Theories. Graph theory is good for the analysis of the dynamic characteristics (i.e. processes) of Big Data manipulation and analytics; Set Theory is good for data typing, data type coercion, and equational logics (equations) of data type properties.

5. BD Analytics can be achieved using statistical methods (i.e. Platform Analysis), SCADA-like simulations, or AI/ML methods based on Neural Networks. The latter can easily be represented by subgraphs performing pattern recognition upon learned ‚similarity matrices' that allow, for example the safety properties of the solar panels above to be learned from sufficient examples. So all measurements that do not fit with the learned matrices may violate the ML-implemented safety, privacy or security invariants.

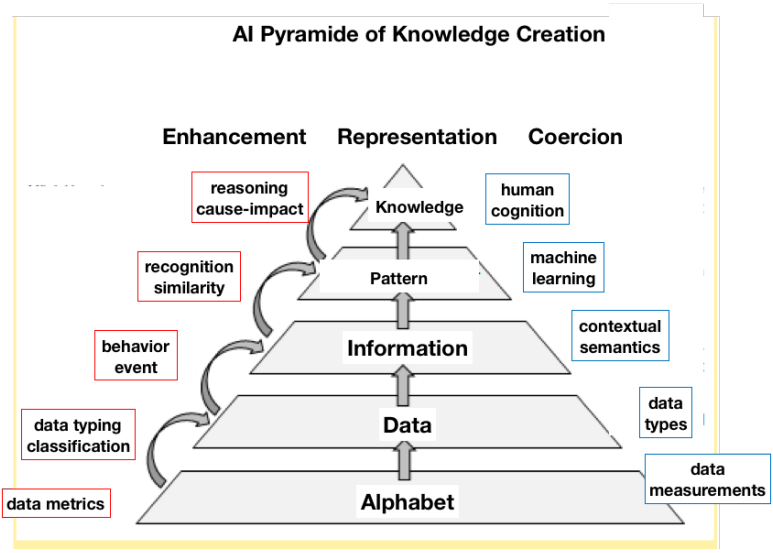

Figure 1: AI Pyramide of Knowledge Creation/Data Enrichment

5. Conclusions

In figure 1 the AI pyramid of Knowledge Creation is seen. It comprises three trajectories, i.e. the representation of data: starting with the alphabet of measurements, to typed data, to observable information, to recognition of patterns and finally to knowledge-based cognition. The data representation trajectory coincides with data enrichment steps on the lefthand side of figure 1. More sophisticated is the data coercion trajectory on the right hand side of figure 1: Since data is enriched from simple data measurements to data types to contextual semantics interpretation, to machine learning, to human cognition, the semantics transformed from simple to complex types must accordingly be coerced. This kind of transformation sometimes is also called 'semantics change' because a measurement is transformed stepwise into sophisticated knowledge, e.g. to reason why the measurement curve of the surface temperature of a solar panel shows the tendency to critical behavior.

Although by this report a certainly incomplete picture of the effort of international AI standardization organizations has been drawn on concentrating on Big Data Management[Pm18] respectively ETSI's 'Big Data Lake' whereas the latter platform approach is presented by the ISI Industrial Specification Group in the Group Specification of [Is19]. Nevertheless the urgent needs have been made plausible by discussing use cases of big data management standards based on AI and ML methods and techniques yet from such diverging viewpoints of society in the realm of CLAIRE 01 and of industry in the realm of I4.0 [Ts18].

German Institute of Standardization DIN, German Association of Electro-technique DKE, CEN/CLC[Rm19], ISO/IEC and many other standardization organizations have recently founded working and focus groups on AI in order to provide time schedules, reference models, architectures and techniques of AI techniques and new developments. So, by [Rm19] the CEN/CLC roadmap for AI standardization in support of ISO/IEC JTC1 SC42[Ai18] has been presented: In 10/19 it is planned to take stock of international standardization activities outcomes and in 03/20 the final AI standardization roadmap will be presented to international organizations, EU Technical Committees and to European policy makers.

Finally a good example from research of what is meant by AI-based data science is the 'LMU Curriculum of the Data Science Certificate Program'[Ds19] industries and companies shall absolve comprising the following items including but not limited to: data management and analysis, statistics, causality and causation visualization, prediction modeling, deep learning methods, tools and concepts for large data sets and last not least data privacy, safety and security aspects compared to JTC1 SC27 [St19].

In several standardization organizations - according their roadmaps - it is planned to propose several projects on so-called ‚semantic standards' aiming at standardized guidelines for using the 2 levels of semantics, i.e. context depending foreground ontologies and context independent background graph semantics.

Literature

[Cl18] CLAIRE: http://www.humane-ai.eu/index.html

[Ai18] ISO/IEC JTC1/SC42 WG3 N007v0_2018-12-06 Technical Report on Trustworthiness of AI

[Bd18] ISO/IEC JTC1/SC42 N199 FDIS 20546-1:2018-10-12 IT Big Data - part 1 Overview and Vocabulary

[Ra18] ISO/IEC JTC1/SC42 DIS 20546-3 IT Big Data - part 3 Reference Architecture

[Ts18] IEC 62443-4-2 (IEC 65/735/FDIS) IACS - part 4.2 Technical Security Requirements for IACS Components:2018

[Pm18] ISO/IEC JTC1/SC42 WG2 N1504 NWIP 24300 IT AI Process Management Framework for Big Data Analysis

[Is19] ETSI GS 006 Information Security Indicators (ISI) - An ISI-driven Measurement and Event Management Architecture (IMA) and CSlang, a Common ISI Semantics Specification Language: https://www.etsi.org/deliver/etsi_gs/ISI/001_099/006/01.01.01_60/gs_isi006v010101p.pdf

[Rm19] BT N11454-BT 161/DG 11173/DC CEN-CLC Focus Group on AI: Proposed Time Schedule towards the CEN-CLC technical roadmap

[Ds19] LMU Faculty of Math, Informatics and Statistics 'Data Science Certificate Program': https://www.m-datascience.mathematik-informatik-statistik.uni-muenchen.de/application/index.html; LMU 'Data Science Lab': https://dsl.ifi.lmu.de/cms/

[St19] ISO/IEC JTC1/SC27 IT Security Techniques with Working Groups:

WG1 ISMS - Information Security Management

WG2 CSM - Cryptography and Security Mechanisms

WG3 SETS - Security Evaluation Testing and Specification

WG4 SCS - Security Controls and Services

WG5 IMPT - Identity Management and Privacy Technologies

[Cc18] ISO/IEC SC38/WG5 ‚19944' IT Cloud Computing - Cloud Services and Devices: Data Flow, Data Categories and Data Use