Abstract. NSIs are using frameworks to organise and set up their official statistical production, e.g. for more than one decade the GSBPM (“Generic Statistical Business Process Model”). As a sequential approach of statistical production, GSBPM has become a well-established standard using deductive reasoning as analytics’ paradigm. For example, the first GSBPM steps are entirely focused on deductive reasoning based on primary data collection and are not suited for inductive reasoning applied to (already existing) secondary data (e.g. big data resulting, for example, from smart ecosystems).

Taken into account the apparent potential of big data in the official statistical production, the GSBPM process needs to adapted to incorporate both complementary approaches of analytics (i.e. inductive and deductive reasoning) and, for example, through the usage of, for example, data-informed continuous evaluation at any GSBPM step.

This paper discusses the limitations of GSBPM with respect to the usage of big data (using inductive reasoning as analytics’ paradigm), and also with respect to “trusted smart statistics”. The authors give insights on how to augment and empower current statistical production processes by analytics and also by (trusted) smart statistics. In addition, it will also address which cultural changes (e.g. skills, organisational structure and agility) should be addressed by the senior management of NSIs to embrace this major paradigm shift.
1. Introduction

“Just as haute cuisine must incessantly reinvent itself in order to stay at the forefront of gastronomy, official statistics is also confronted with a rapidly changing context and needs.” This quote from Walter J. Radermacher [1], the former Director General of Eurostat, describes quite well the situation official statisticians are facing. Particularly, in the context of the digital transformation, they need maybe more than any others to reinvent themselves and maybe also to change the way they currently produce official statistics.

The digital revolution is built on data and is well underway: terms such as big data, cloud, internet of things, internet of everything, fourth industrial revolution, smart cities and data economy are no longer just words on everyone’s lips, but concepts that are changing the habits of consumers and businesses. This digital transformation started with the first wave of digitalisation: the technical digitalisation of converting analogue contents and services into digital ones – without fundamental changes to the underlying processes – that resulted in the (big) data revolution. However, big data are a data management infrastructure with underlying hardware, software, and architecture and should not be “taken for museum purposes” only. As early as 1942 the official statistician W. Edwards Deming noted that the “ultimate purpose of taking data is to provide a basis for action or a recommendation for action”. As such, a second wave of digitalisation is needed to enable learning from big data and to generate value from this (big) data revolution for society as a whole. The biggest challenge in this respect is the veracity of the “data pedigree”, i.e. the trustworthiness of the data, including the reliability, capability, validity and related quality of the data, as well as the transparency of the related production process. In a world of post-truth politics and fake news, data veracity is more important than ever.

2. Demystifying the “big data” hype

The term “big data” – coined in 1997 by two researchers at the NASA – has acquired the trappings of a religion. The term “big data” applies to an accumulation of data that can not be processed or handled using traditional data management processes or tools. As such, big data are a data management IT infrastructure which should ensure that the underlying hardware, software, and architecture have the ability to enable learning from data or making sense out of data, i.e. analytics; enabling data-driven decision making and data-informed policy making.

The following characteristics – “the five Vs” – provide a definition of “big data”: “Volume” (“data at rest”), “Variety” (“data in many forms”) and “Velocity” (“data in motion”) as essential characteristics of (big) data, and “Veracity” (“data in doubt” or “trust in data”) and “Value” (“usefulness of data”) as qualification for use characteristics of (big) data. As illustrated in Figure 1,
the veracity, including the reliability (quality over time), capability and validity of the data, and the related quality of the data, is key. Therefore, existing “small” data quality frameworks need to be extended, *i.e.* augmented, in order to cover this new (big) data ecosystem.

![Figure 1. The first four Vs of big data with “Veracity” as key. [2]](image)

### 3. Demystifying the “Internet of things” hype

The term “Internet of Things” (IoT) – coined in 1999 by the technologist Kevin Ashton – starts acquiring the trappings of a *new religion*. According to IEEE [3], the IoT covers many areas ranging from enabling technologies and components to several mechanisms to effectively integrate these low-level components. Software is then a discriminant factor for IoT systems. IoT operating systems are designed to run on small-scale components in the most efficient way possible, while at the same time providing basic functionalities to simplify and support the global IoT system in its objectives and purposes. Middleware, programmability – in terms of “Application Programming Interfaces” (APIs) – and data management seem to be key factors for building a successful system in the IoT realm. Management capabilities are needed in order to properly handle systems that can potentially grow up to millions of different components. In this context, self-management and self-optimisation of each individual component and/or subsystem maybe strong requirements. In other words, autonomous behaviours could become the norm in large and complex IoT systems. Data security and privacy will play an important role in IoT deployments. Because IoT systems will produce and deal with personally identifiable information, data security and privacy will be critical from the very beginning. Services and applications will be built on top of this powerful and secure platform to satisfy business needs. So many applications are envisioned as well as generic and reusable services. This outcome will require new, viable business models for IoT and its related ecosystems of stakeholders. Finally, IoT can have an impact on people and the society they live in, and so it must be conceived and conducted within the constraints and regulations of each country. However, *IoT is about data and not things*. As such, if we try to map the IoT concept to Figure 1, we can see that from a *data point of view* we can easily replace the concept of big data with IoT data. That means that from a data
perspective we are facing the same challenges; see Figure 2. Therefore and similarly to big data, if we want to use IoT (data) for statistical purposes, we need also to extend, *i.e.* augment, existing “small” data quality frameworks.

![Figure 2. The first four Vs of IoT (data) with “Veracity” as key. [2]](image)

The demystifications made so far clearly underline the point that *data are the fuel and analytics, *i.e.* learning from data or making sense out of data, is the engine of the digital transformation and the related data revolution.*

4. **“Analytics of things” and “smart (official) statistics”**

The “Analytics of Things” (AoT) corresponds to the “analytics layer” that occurs with the IoT devices and their generated data. Given the distributed nature of the connected devices and the explosive growth of IoT infrastructures and technologies, it becomes key to execute analytics and related data quality processes on the data-gathering devices themselves, *i.e.* at the edge (or at the “endpoint”), or as close to the originating data source as possible. This paradigm is referred to *analytics at the edge* or *edge analytics* (based on a distributed “IT architecture layer” called “edge computing”). For example, in practice, the most efficient way to control data quality is to do it at the point where the data are created, as cleaning up data downstream (and hence centralised) is expensive and not scalable. As such, it is about *moving the analytics and the data quality frameworks to the data* and not the data to the (centralised) analytics and (centralised) data quality frameworks. To do so, a centralised management of analytics will be needed; consisting, for example, of transparent central analytics model and rule development and maintenance, a common repository for all analytics models, *i.e.* “algorithms”, and a related analytics model version management. Additional concerns are security (*e.g.* will be improved by reducing complexity), privacy (*e.g.* sensitive data will be retained at the edge), analytics governance (*e.g.* no strong governance needed as the algorithms are decentralised and publicly available), reliability and scalability of the edge devices, and (public) trust; see also [2]. The usage of AoT within official statistics clearly illustrates another paradigm shift,
along with related considerations of transparency and glocalisation, i.e. producing official statistics according to both local and global considerations. Current key challenges (see [4]) include the lack of (glocalised) standards of both IoT data and analytics, and also of AoT approaches, e.g. using edge analytics. Saying this, standardisation efforts are urgently needed. Could official statistics play a key role in their development and governance?

Eurostat ([5] and [6]) defines “smart (official) statistics” as being the future system of producing official statistics where essentially, data capturing and data processing capabilities coupled with analytical and statistical capabilities will be embedded in the smart systems themselves, i.e. a “data layer for statistics” within smart systems. Therein, the traditional model of pulling data in – from data sources to NSIs – will not fit in the new scenario. Instead, Eurostat envisions a model based on pushing computation out – from NSIs to the data acquisition systems. This shift of focus from sources to systems lies at the core of what Eurostat call smart statistics, and corresponds to what we discussed above, i.e. AoT using edge analytics, in the context of official statistics.

5. Demystifying the two approaches of analytics

Statistics traditionally is concerned with analysing primary (e.g. experimental or made or designed) data that have been collected (and designed) for statistical purposes to explain and check the validity of specific existing “ideas” (“hypotheses”), i.e. through the operationalisation of theoretical concepts. The resulting primary analytics or top-down (i.e. explanatory and confirmatory) analytics aims to evaluate or test an idea (hypothesis). The corresponding analytics paradigm is deductive reasoning as “idea (theory) first”.

Data science – a rebranding of data mining and as a term coined in 1997 by a statistician – is typically concerned with analysing secondary (e.g. observational or found or organic or convenience) data that have been collected (and designed) for other reasons (and often not “under control” or without supervision of the investigator) to create new ideas (hypotheses or theories). The resulting secondary analytics or bottom-up (i.e. exploratory and predictive) analytics aims to generate an idea (hypothesis). The corresponding analytics paradigm is inductive reasoning as “data first”. Examples of inductive learning from data include machine learning algorithms, e.g. resulting in “supervised (machine) learning systems”; often referred to as artificial (or machine) intelligence (algorithms). However, note that machine learning is only one part of a larger artificial intelligence framework.

The two approaches of analytics, i.e. inductive and deductive reasoning, are complementary and should proceed iteratively and side by side in order to enable data-driven decision making, data-
informed policy making, and proper continuous improvement. The corresponding inductive-
deductive reasoning cycle is illustrated in Figure 3.

![Figure 3. The inductive-deductive reasoning cycle. [7]](image)

As already the statistician John W. Tukey told us in 1980: “Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.”; underlying again the need to use both approaches in order to achieve proper continuous improvement.

6. Process models for continuous improvement

According to W. Edwards Deming: “If you can not describe what you are doing as a process, you do not know what you are doing.”

Let us start with the well-known “Plan–Do–Check–Act” (PDCA) cycle below which is often referred to as the Deming cycle, Deming wheel or the Shewhart cycle. Walter A. Shewhart proposed this approach in the field of quality control in the 1920s, and W. Edwards Deming later popularised PDCA as a general management approach based on the scientific method.

![PDCA cycle](image)

However, we need to go further and differentiate the current process model for the production of official statistic from a general process model for analytics. Until we do not understand the fundamental differences between both process models and their links with respect to the usage of both analytics paradigms, i.e. inductive and deductive reasoning, we will not be able to understand what we need to care about.
First, a well-established process model for the production of official statistics is the “Generic Statistical Business Process Model” (GSBPM) which is coordinated through the “United Nations Economic Commission for Europe” (Version 5.0 as of December 2013). GSBPM is consistent with the PDCA cycle; see left panel of Figure 5. Moreover, GSBPM is a key conceptual framework for the modernisation (and standardisation of the production) of official statistics, and is nowadays (unfortunately) often shown as in the right panel of Figure 5 by removing (or hiding) the continuous (quality) improvement cycle.

Figure 5. The “Generic Statistical Business Process Model” (GSBPM). [8]

David A. Marker [9] also believes that “the reason [operational] cost [of different parts of the statistical business process] has not been a central focus is a difference between NSIs focus on measuring quality of their products and services, rather than continuous improvement of quality.” As such, the emphasis should clearly move from measuring quality to improving quality. Moreover, the GSBPM is a deductive reasoning and a sequential approach. For example, the first GSBPM steps are entirely focused on deductive reasoning for primary data collection and are not suited for inductive reasoning applied to (already existing) secondary data. In addition, the evaluation (“Evaluate” step) is only performed at the end. As such, the authors think that this process model needs to be adapted in order to incorporate both approaches of analytics (i.e. inductive and deductive reasoning) and through the usage of, for example, data-informed continuous evaluation at any GSBPM step.

Second, a well-known general process model for analytics is the CRISP-DM (“CRoss Industry Standard Process for Data Mining”) process, initially conceived in 1996, which is also consistent with the PDCA cycle; see Figure 6. By breaking down the life cycle of an analytics project into interrelated phases, CRISP-DM places a structure on the problem, allowing reasonable consistency, repeatability and objectiveness. The sequence of the phases is not rigid. Moving back and forth between different phases is always required.
However, as already indicated in the introduction using the words of W. Edwards Deming, analytics has value only when it is actionable. Embedding the developed analytics models, i.e. algorithms, inside the appropriate operational processes delivers that value. It is important to make sure that those models, i.e. algorithms, can be legally traced, monitored and secured, and this requires discipline and management capabilities. As such, deploying “analytical assets” within operational processes in a repeatable, manageable, secure and traceable manner requires more than a set of APIs and a cloud service; a model that has been scored (executed) has not necessarily been managed. It is all about “industrial deployment of analytics” as releasing a model to production is not enough; that “release moment” is followed by many steps that make the deployment cycle as important as the development cycle. Figure 7 illustrates the complementary cycles of developing (using CRISP-DM with slightly different names for the phases) and deploying “analytical assets”.

Figure 6. The “CRoss Industry Standard Process for Data Mining” (CRISP-DM) process. [10]

Figure 7. The complementary cycles of developing and deploying “analytical assets”. [11]
Following Erick Brethenoux [11] the foundation of both cycles should provide: a centralised way to store and secure analytical assets to make it easier for analysts to collaborate, allow them to reuse models or other assets as needed; the possibility to house the repository in an existing database management system (regardless of its provisioning) and establish security credentials through integration with existing security providers, while accommodating models from any analytics tool; protocols to ensure adherence to all internal and external procedures and regulations, not just for compliance reasons but, as an increasing amount of data gets aggregated, to address potential privacy issues; automated versioning, fine-grained traceable model scoring and change management capabilities (including champion/challenging models and/or “features” ) to closely monitor and audit analytical asset lifecycles; and bridges that eventually link both the development and deployment cycles to external information life cycle management systems (to optimise a wider and contextual reuse of assets) and enhanced collaboration capabilities (to share experiences and information assets beyond the four walls of the organisation).

The complementary cycles of developing and deploying “analytical assets” and their foundation is clearly in line with what the authors mentioned above when discussing AoT or edge analytics. From a data perspective, it will become key to embed into these cycles a “data veracity layer” by continuously applying and monitoring “augmented data quality frameworks” to the “data pedigree” before, during and after the execution of the models, *i.e.* algorithms. All this bends the well-established system of trust that lies at the basis of the official statistics production until today. Eurostat [6] sees the related opportunity official statistics faces in the design of a coherent framework for what they call “trusted smart statistics”.

7. **Conclusion**

Following the above the authors clearly think that current frameworks in the official statistical production (like GSPBM) are not appropriate for inductive reasoning applied to (already existing) secondary data (*e.g.* big data resulting, for example, from smart ecosystems). We tried to give some insights on how to augment and empower current statistical production processes by both approaches of analytics and also by trusted smart statistics, and even AoT using edge analytics. We do not have the solution on how to exactly link the process model for analytics (including the development and deployment) to the current GSPBM process. However, we think that a single process model will not be sufficient (especially with respect to trusted smart statistics and AoT), but that the “Embed Insights” phase in Figure 7 could be used as a link to the GSPBM process; see the Figure 8. The authors strongly encourage the European Statistical Systems (ESS) to address this topic, *e.g.* in the
upcoming ESSnet on Big Data. The necessity to ensure that the GSBPM process evolves consistent with the above mentioned challenges is given. Otherwise, every NSI will develop its own practice which is time consuming, risky and not efficient. In addition, cultural changes (e.g. skills, organisational structure and agility) should be clearly addressed as well by the senior management of NSIs to embrace all the paradigm shifts illustrated in this paper.

Figure 8. Current statistical production versus analytics process model. [4]

References