AIChE Journal
DOI 10.1002/aic.16203

Assessing the Value of Information of Data-Centric Activities in the Chemical Processing Industry 4.0

Marco S. Reis[1,*] Ron Kenett[2]

(1) *CIEPQPF, Department of Chemical Engineering, University of Coimbra, Rua Sílvio Lima, 3030-790, Coimbra, Portugal, * marco@eq.uc.pt*

(2) *KPA Ltd. and Samuel Neaman Institute, Technion, Israel, ron@kpa-group.com*

Abstract

The quality of information generated in data-driven empirical studies is of central importance in Industry 4.0. However, despite the undeniable and widely accepted importance, not sufficient attention has been devoted to its rigorous assessment and analysis. Consequently, if information quality cannot be measured, it also cannot be improved, and therefore current efforts for extracting value from big data empirical studies and data collectors are exposed to the risk of generating limited findings and insights, leading to suboptimal solutions. In this article we describe and apply a framework for evaluating, analysing and improving the quality of information generated in empirical studies called InfoQ, in the context of the Chemical Processing Industry (CPI). This systematic framework can be used by anyone involved in data-driven activities, irrespectively of the context and specific goals. The application of InfoQ framework to several case studies is described in detail, in order to illustrate its practical relevance.

Keywords

InfoQ, information quality, industry 4.0, big data, chemical processing industry

Received: Nov 27, 2017; Revised: Mar 28, 2018; Accepted: Apr 30, 2018

Introduction

The first industrial revolution took place in the 18th century and consisted of the mechanization of manufacturing processes using water and steam power. The second revolution was powered by electricity, enabling mass production through assembly lines and work standardization, during the early 20th century. Computers and electronics elevated even more the degree of automation during the two last decades of the 20th century, leading to the third revolution (the digital revolution). Currently, in the down of the 21st century, industry is undergoing another expansion period, building over the success of the digital revolution, and taking advantage of the notable technological developments on cybernetics, distributed physical devices with built-in computing and communication capabilities (the internet of things, IoT), IT infrastructures (including cloud storage and computing), new sensor technology, new production processes (additive manufacturing), etc. When synergistically combined, these new technological ingredients create conditions for the development of what is now called "smart manufacturing" processes, where the units are able to communicate autonomously with each other and to self-adjust their operations in order to accommodate for changing disturbances, demands and constraints. These "smart" systems are able to collect, distribute and integrate information of diverse nature dispersed across the supply chain, and to use it for the sake of enhancing safety, productivity, efficiency (in the use of energy, as well as in human and material resources), environmental sustainability, product quality and economic performance. They constitute the basis of the 4th industrial revolution labelled "Industry 4.0".

More than ever, data abounds in Industry 4.0. Virginia Rometty (CEO of IBM) provides a clear signal that the critical role of data is finally being widely acknowledged by Industry stakeholders. She said: "*What steam was to the 18th century, electricity to the 19th and hydrocarbons to the 20th, data will be to the 21st century. That's why I call data a new natural resource*.". Therefore, the pressure is rapidly building up on enterprises around the world to identify how to take advantage of this resource, in order to turn it as a source of competitive advantage, process & quality improvement and economic growth.[1] This process is taking place not only in large companies, [2] but also in small and medium enterprises (SMEs) – in spite of their natural structural limitations; for more on this topic see ref. [3]. Consequently, the due importance currently given to data is being transferred and creates a sense of urgency in developing suitable analytical platforms able to handle their Volume, Variety, Velocity, Veracity and Value issues (the 5 V's of Big Data [1,4]). Briefly, Volume is the existence of large amounts of data, either stored or in transit, and Velocity is the increasing pace with which it is created and transferred. Data can be in a wide Variety of formats – structured (low or high-dimensional arrays of numbers) or unstructured (e.g., text or images) – and the assessment of their quality and accuracy often cannot be easily established. This corresponds to the Veracity issue of Big Data. Finally, the use of the Data resource must lead to some added Value for the companies and their stakeholders.

The final aspect referred to in the previous paragraph, Value, is, arguably, the most important of the five, [5] as it determines the value of the entire empirical effort. But how can practitioners, engineers, data analysts, managers, etc. –i.e., all those interested in collecting the benefits of empirical studies – assess their potential value in a given, well-defined, application context? This is a challenge needing urgent solutions. Data-centric activities are increasingly present in industrial processes, accumulating and circulating data at increasing volumes and rates, but surely not all of them have the same quality, value, or importance, for their specific problems. As companies invest more and more on technology and data analytics resources (see for instance the special issue of Chemical Engineering Progress, "Big Data Analytics", published in March 2016), including highly skilled personnel, software, computer power, IT infrastructures, etc., it becomes clear that something is still missing: a systematic framework for assessing the quality of information produced with a specific set of methods, in a given application context.

A framework for measuring the quality of information generated by empirical studies is therefore a critical asset for all those routinely involved in the analysis of a variety of industrial problems or challenges. This requires the definition of a data collection strategy, the required analytical tools and the reporting strategies. The InfoQ framework allows for conducting a preliminary risk assessment of each project, in order to identify gaps, improvement opportunities and establish priorities (expected cost-benefit analysis) before diving into them. It also enables for a posteriori diagnosis of existing or past activities. Most of us involved in data analysis have a story (or perhaps more) to tell about the time invested in analysing a problem or dataset, only to conclude, with a tone of resigned frustration, that nothing could be done or achieved with it, while other burning problems were waiting to be solved… Problems of this type, and limitations of Big Data initiatives become clear as more experience is acquired; see for instance the discussions in [1,6-8]. With a systematic framework available, not only situations such as these can be spotted in the early stages with higher probability, but also the weaknesses detected in the preliminary analysis will prompt effective actions to complete the project with missing elements and improve the quality of information it can potentially deliver.

This article is dedicated to the description and application of a systematic framework, in the context of the Chemical Processing Industry (CPI). The framework is based on the concept of Information Quality, InfoQ, originally proposed by Kenett and Shmueli. [9,10] InfoQ is defined as "the potential of a dataset to achieve a specific (scientific or practical) goal by using a given empirical analysis method". It is a general methodology applicable to empirical research in general, and to the practice of data science in industry, in particular. The goal of the assessment is to evaluate the value of the information likely to be generated in an empirical study and to devise actionable measures to improve it, maximizing InfoQ. The process of assessing and increasing InfoQ, requires the definition of a set of structuring components of any data-driven project, namely: the specific analysis goal, *g*; the available dataset or data collection protocol, *X*; the empirical analysis method to use, *f*; an utility measure, *U*. According to the definition of InfoQ, these elements are related with each other through the following analytical expression (in words, it is the level of Utility, *U*, achieved by applying the analytical method *f* to the dataset *X*, given the activity goal *g*):

$$InfoQ\left(f,X,g\right)=U\left\{f\left(X\mid g\right)\right\} \tag{1}$$

InfoQ is determined by the quality of its components (*g, U, X, f*), whose assessment comprises 8 partially independent dimensions (see the third section). The base implementation of the framework follows a Top-Down approach, starting with the definition of the goals, and then progressing towards the analysis of the specific dataset available and methods adopted. In this article we present an alternative procedure especially designed to the CPI context, namely for supporting the planning and development of Big data initiatives for Industry 4.0, but also applicable to any data-centric industrial activity.

The article proceeds as follows. The next section provides an overview of some data-centric activities in the CPI. In the third section, the InfoQ framework is introduced and its eight assessment dimensions described. The fourth section is dedicated to the presentation of case studies, where the application of framework is described in detail. The fifth section concludes the paper, with a summary of key aspects of the InfoQ framework and prospects for future activities.

Data-driven activities in the CPI

The scope of application of the assessment framework for the quality of information generated in data-centric activities is quite large and diverse. In this section, several activities whose InfoQ level is important to be high, are referred and briefly described. They can be classified, in a broader sense, as Type 1: Exploratory/descriptive studies; Type 2: Process monitoring & surveillance; Type 3: Predictive modelling (for control/optimization or for virtual metrology) and Type 4: Diagnosis or causal explanation. With the progress of Industry 4.0, new sources of data will be created, from all parts of the supply chain, as well as from the technology enabling their rapid and facilitated acquisition, storage, processing and reporting. It is therefore highly probable that other tasks will soon be added to the list presented below and become part of the companies' routines, acting as new drivers for competitive advantage.

Data visualization, fusion, selection, screening and reporting

Process management requires, at all levels, the access and display of data necessary to supervise, diagnose and report the current status of the process. Clearly, the most convenient way to convey the information extracted is through properly designed visual displays of data. [11] However, in order to be effective, information must be presented in a targeted way, only showing what is relevant to the user's function, and at the adequate time resolution. For instance, operators require data resolutions of minute-hour; for process managers, hour-day resolutions are normal; as to plant directors, day-week resolutions are usually adopted. For the administration board, monthly or coarser time resolutions are adequate. Furthermore, data is displayed in a hierarchical way, starting with key process indicators at a first level. More detailed information is accessed only if, and when, it is necessary, conditionally on the analysis outcome of the key indicators.

These activities belong to the class "Exploratory/descriptive studies" and Industry 4.0 offers new opportunities for increasing InfoQ at this level. An example is the creation of human-centric platforms for process supervision, where better interfaces are developed (involving a merging of both engineering and psychology), extending the capabilities of the report media (any digital portable device is illegible) and promoting the development of new graphical solutions with the emergence of infographics, dynamic plots, etc. Typically, only recent snapshots of process operation data are of interest for reporting activities.

Process monitoring

Since its introduction in the early 1930's by W.A. Shewhart [12], statistical process monitoring has become an important part of process operations and management in organizations. The main goal of this activity is to access whether process is stable in the present (if not, procedures are triggered to stabilize it) and whether it continues in this way in the future (when an out-of-control state is detected, the root cause, also called special or assignable cause, should be searched and fixed). The data of interest is usually the most recent one, with the proper resolution for taking fast decisions about the state of the process. Process monitoring methods should be able to capture and describe in a probabilistic compact way, the normal operation conditions behavior of the processes, which give rise to a variety of approaches dedicated to different typologies of processes and operations. In short, methods were developed for addressing static univariate, [12-14] multivariate (full-rank), [15,16] and high-dimensional continuous processes, [17-22] as well as for continuous dynamic processes, [23-27] non-stationary batch processes, [28-33] and for product and process profiles, [34-39] among others [40-45].

Predictive maintenance

Process monitoring is focused on assessing the "process health". Predictive maintenance and Condition Based Maintenance, [46] are focused on "Equipment Health". Equipment wearing and deterioration is expected, and is responsible for non-stationary components in process behaviour, which are usually overlooked by process engineers. They are however the focus of maintenance & reliability engineers, and the center of their activity. Clearly, several important opportunities lie ahead from a better integration of both domains, typically managed as separated silos. [40] Combining both perspectives, allows to expand process monitoring beyond its current static and reactive dominating view, and tackle the true nature of processes, which are non-stationary, evolving, and therefore, to some extent, predictable. Maintenance & reliability also benefit from the technological and data resources made available within Industry 4.0. [47]

Data for predictive maintenance involves the history of all interventions in the equipment and records of operation settings. Contrary to visualization & reporting and process monitoring applications, data for predictive maintenance requires the access to long records of past operation and equipment interventions.

Process diagnosis & troubleshooting (data-driven process improvement)

Process data bases accumulate large amounts of data. With the appearance of new measurement sources and the development of efficient collection, transmission and storage technology, much more data will be available in the future. This resource remains, to a large extent, unexplored in the CPI, as most current data-driven applications are on-line (e.g., process monitoring, process control, visualization and reporting of current operation, etc.). The off-line analysis of these immense data repositories can highlight and identify important improvement directions, such as for increasing productivity (reducing production cycle times), reducing chemicals consumption, improving process stability and product quality; reducing scrap and rework, etc. [19,31,48,49].

Of interest in this activity, is all available data related to the current configuration of the process. Operation data from different technological environments (e.g., when other main pieces of equipment were in use) is usually of limited value, as they reflect a past reality that does not exist anymore.

Predictive modelling (quality prediction and prognostic estimates)

Predictive modelling is one of the most common activities in industry. It consists in using a set of predictors or input variables, to estimate the value or state of a set of response variable of interest. Many methodologies have been developed for handling the diversity of problems, data structures and goals. For instance, and focusing only on regression approaches directed to large scale settings (those likely to be found in Industry 4.0 applications), one can identify several categories of fundamentally different approaches, such as variable selection methods, [50-52] penalized regression methods, [53-61] latent variables methods [62-66] and tree-based ensemble methods. [56,67-69]

With the increasing complexity of industrial processes, all information available should be used to optimize their performance. In this regard, not only the information extracted from process data is relevant, but also the one available *a priori* about the underlying process structure. In this regard, multi-block methods [70-76] and Bayesian networks, [77-79] offer good solutions to explore more in-depth in the future.

Predictive modelling it is not an end in itself, but a mean to accomplish a certain goal, such as process optimization, fast product release, stability through process control, variability reduction, etc. The data required to be implemented covers all operation records, including those periods where certain perturbations took place in the process, as they contain information about a larger region of the operational space. However, despite the large amount of observational data stored, they may not be of enough quality to derive good prediction models. This critical aspect is highly dependent on the goal of the analysis that defines the purpose of the predictive model. Existing process data may suffice for virtual metrology and soft sensor applications, but not for process control and optimization tasks. For these situations, the active collection of new data through properly designed statistical design of experiments (DOE), should be pursued.

Quality by design (QbD) and Product Development

With Quality by Design, the central aim is to build quality into the products and processes, through appropriate planning and following a systematic approach, rather than *a posteriori* testing and inspection. This is particularly relevant for the design and development of new products and processes in the Pharmaceutical industry, where QbD has gained notoriety and is today widely accepted (see for instance the ICH Q8 guideline issued by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceutical for Human Use). This is a science based approach that makes use of existent knowledge about the process and product, in order to guarantee compliance with the specifications required for their Critical to Quality Attributes. Most often, data involved in this process arises from experiments planned using DOE. However, a current trend is to fuse this approach with the existence of data collected over the years, in order to develop more efficient and informative designs. This approach is called retrospective QbD (rQbD) [80] and is particularly suited for companies with a long history of product development and production, that started quite often before the emergence of QbD initiatives in the field.

CPI has still a lot to gain from the adoption of QbD and rQbD principles. However, some data-driven initiatives for improving product design were already proposed and implemented in the CPI context. For instance, MacGregor and co-workers fitted latent variable models using historical data records, where process constraints and operating policies are already implicitly incorporated in the data correlation structure, and used them to address the development of new products, exhibiting a set of specified properties. [81,82] The necessary operating windows are derived from the definition of the desired quality specifications for the new product and an inversion in the latent variable model space (from the **Y** to the **X** space). The solution thus found, will not only comply with the required properties, but will also be compatible with past operating policies [81]. QbD studies tend to fall in the scope of “predictive modelling for control & optimization” or “diagnosis or causal explanation.”

The InfoQ framework

The activities referred in the previous section should lead to high quality of information – high InfoQ. In order to effectively manage these tasks, and improve the way they are carried out, it is fundamental to develop a suitable referential for measuring the quality of information generated – remembering the celebrated Lord Kelvin's citations: "If you can't measure it, you can't improve it" and "To measure is to know". In this regard, we present the InfoQ framework as an effective solution for measuring the value of information in the context CPI data-centric applications. The InfoQ framework will be presented here, with necessary adaptations to CPI.

As mentioned in the introduction, the InfoQ framework assesses the quality of the information generated by any empirical study, taking into account the following four building blocks linked through equation (1):

- *Analysis goal*, *g*. The purpose of the analysis, in statistical or analytic terms. Activities described in the previous section, have distinct analysis goals. Some include being able to make reliable predictions (of future values or about the state of the process or equipment, etc.), others are focused on diagnosis (causal explanation, for troubleshooting and process improvement) and others on description (for supervision and reporting). These are general types of goals commonly found in practice. The different applications bring different particularities to their fine definitions.
- *Dataset*, *X*. The dataset to be used for accomplishing the goal. Data can arise from different sources, such as observational industrial data, data collected from planned experiments, laboratory data, computer simulations, etc. Furthermore, they can have any structure: scalar quantities (i.e., zeroth order tensors, such as

industrial sensors and univariate quality parameters), first order tensors (i.e., 1-D profiles, such as spectra, chromatograms, particle size distributions), second-order tensors (e.g., grey-level and thermographic images), third-order tensors (e.g., hyper-spectral images, hyphenated data), etc. They can also contain unstructured data, namely text, such as tags from operation alarms and warnings, or from operator introduced comments.

- *Empirical analysis method*, *f*. The data analysis method adopted to process the dataset *X*, in order to achieve the goal, *g*. Methods can be of different types, such as parametric / semi-parametric / non-parametric, probabilistic / deterministic / algorithmic, linear / non-linear, single-block / multi-block, etc.
- *Utility*, *U*. A measure of the extent to which the analysis goal, *g*, is achieved. It usually consists of suitable performance metrics such as root mean square error of prediction (RMSEP) or R^2_{Pred} for predictive activities, measures of statistical power (e.g., p-values) for diagnosis, and goodness of fit or discrimination for descriptive goals. In many occasions however, practitioners need to access the potential of their activity, before implementing it, in order to establish priorities and make decisions where to invest the limited resources they manage. This happens, for example, during the Definition phase of process improvement and troubleshooting activities (e.g., in Six Sigma projects), and in the Design stages of new processes or products where alternatives are preliminarily screened. In these occasions, *U* needs to be estimated before full application of *f* to *X*, in order to make some decisions about the most favourable alternatives or on how to improve the base solution. In these cases, the utility can be evaluated through semi-quantitative assessment of *X* and *f*, *w.r.t. g*, as described in the next section. This is a new procedure, proposed for the first time in this article that is

particularly relevant to the scope of the CPI. It will avoid design errors, mitigate the consequences of bad practices and allow for improvement actions, before the activity is implemented, leading to gains in efficiency and quality of results.

The quality of the information generated in an empirical study is measured by its InfoQ level. The InfoQ level depends on the quality of the four components: quality of goal definition, data quality, analysis quality and quality of the utility measure, as well as the relationships between them.

The *quality of goal*, regards the adequacy of the translation of the practical goal, expressed in the words of the problem owner, to a goal stated in objective statistical or data analytic terms. Several goals may be posed, and for some, the dataset X may be suitable, for others, useless. For instance, observational datasets are adequate for descriptive studies, but may of limited value for some predictive or even diagnostic studies, where causality is required. In the words of Tukey:[83] "*Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise*".

Data quality regards the appropriateness of the dataset available, for achieving the intended purpose. How relevant is the data for answering the research question? In analytical terms, data quality reflects the potential utility of X in the universe of the analytical methods applicable to achieve the goal g, $\mathbf{\Psi}_g$:

$$Data\,Quality : E_{\forall f_i \in \mathbf{\Psi}_g}\left[U\left\{f_i\left(X \mid g\right)\right\}\right] \qquad (2)$$

where, $E[\cdot]$ stands for the expectation operator. In practice this assessment is usually made in qualitative terms and translated into some categorical scale. However, recent developments on the simultaneous testing and assessment of a wide variety of methods [84] can make this assessment more quantitative and precise.

The *quality of analysis* regards the adequacy of the analytical method to the purpose *g*, and to the dataset available *X*, as well as the way the methods are implemented (the analysis technique [85]). For instance, collinearity is usually a characteristic of data collected passively in Industrial 4.0 applications, given the existence of relationships between process variables, which arise from mass and energy conservation, control policies, redundant measurements, etc. Therefore, when addressing modelling problems, ordinary least squares (OLS) is of limited value (lower analysis quality), when compared with other alternatives such as variable selection methods, [50-52,86] penalized regression methods [53-61] or latent variables methods [62-66].

The *quality of the utility U*, depends on the goal *g*. It addresses the appropriateness of the performance metric to evaluate the utility of the empirical study with a given goal. For instance, a common mistake is the use of quality of fit metrics to assess the predictive performance or methods. This is not appropriate in general and easily leads to overfitting and a wrong assessment of prediction capability. Another example is the use of *p-values* for testing statistical hypothesis in large samples – given the higher power of the statistical tests in these conditions, any small difference will be signalled as statistically significant, even if the deviation is not relevant for all practical purposes.

The eight dimensions of InfoQ

The evaluation of InfoQ can be made directly upon the analysis of its four components. In this case, each component entails several aspects that should be considered and properly weighted during their individual assessment. This unspecified multidimensional assessment process raises questions of reproducibility and operationalization, which adversely affect the adoption by industrial practitioners. In order to make it well-defined and systematic, and to prevent overlooking important aspects to consider during the assessment of InfoQ, a set of eight dimensions were proposed that should be explicitly addressed during the assessment process. They cover different aspects that are necessary, in general, for determining the value of information in a data-driven empirical study. These dimensions, $\theta = \left[D_1 \quad D_2 \quad \cdots \quad D_8 \right]^T$, intervene in the quality of the four InfoQ components (*g, U, X, f*), in a way that may be different, depending on the component. Therefore, as an alternative to compute InfoQ by assessing directly the quality of the four components, one can do it indirectly, analysing the eight underlying dimensions that structure their quality, using the following function composition:

$$\begin{aligned} InfoQ\left(f, X, g\right) &= U\left\{ f\left(X \mid g \right) \right\} \\ &= U\left(\theta\right)\left\{ f\left(\theta\right)\left(X\left(\theta\right) \mid g\left(\theta\right) \right) \right\} \\ &= InfoQ\left(\theta\right) \end{aligned} \qquad (3)$$

The following dimensions are based on the original proposal of Kenett and Shmueli [9,10]. However, some adjustments were made in their definitions, in order to adapt the framework to the CPI context, and to facilitate the operationalization of the structured assessment scheme proposed in this article (also, a new procedure is put forward for estimating InfoQ from the eight dimensions; see third section).

Data Resolution (D_1)

In CPI context, resolution is usually connected to the aggregation level of data. One type of aggregation, regards data granularity. It often occurs that collected data may have different levels of granularity, meaning that their values regard the state of the process over different windows of time, during which measurements were collected and averaged, resulting eventually in a single aggregated value. This process results in recorded values representing averages of minutes, hours, days, weeks, shifts, production units (lots), etc. Another way of performing aggregation is through compound/composite sampling, i.e. through sampling schemes that combine a number of discrete samples collected from a process stream or amount of material into a single homogeneous sample for the purposes of experimental analysis. The existence of variables with multiple resolutions, i.e., containing different levels of aggregation, has been greatly overlooked in data science and statistics. In the CPI context, only a few multiresolution methodologies were developed for large-scale monitoring of industrial processes [87] and for soft sensor development. [88]

A distinct topic (but often confused with multiresolution), is multirate data. [89] Multirate regards the existence of multiple acquisition rates, usually from instantaneous (high resolution) measurements. For instance, quality variables tend to be available at much lower acquisition rates than process sensors.

In the scope of this InfoQ dimension, one considerer the appropriateness of both data granularity and acquisition rate for the purposes of analysis.

In case of process/quality monitoring applications, data aggregation is also related with the formation of rational subgroups. [90] The strategy for forming rational subgroups affects the process variability that is monitored and the type of faults that can be detected. Therefore, it is an important aspect to assess the quality of this InfoQ dimension for statistical process/quality control applications.

Data Structure (D_2)

Data structure refers to the types of data and their characteristics:

- Structure (arrays of numbers, cross-sectional, network data, time series) or unstructured (text, images, sound & vibration records);
- Tensor nature (0th-order, such as process sensors; 1st-order, such as spectra, etc.)
- Presence of noise, outliers, missing data, bad segments (plant shutdowns and transients);
- Single-block or multi-block (i.e., when a single or multiple natural groups of variables exist and their integrity should be maintained);
- Static or time-delayed structure (meaning a lagged-correlation pattern);
- Observational (i.e. “happenstance data”, using R.A. Fisher terminology) or Casual (namely collected following a DOE plan).

The way these aspects are considered during the InfoQ assessment, depends on the actual components being addressed (see third section). They will impact both the Data and Methods, InfoQ components. Data should have the right structure for achieving the analysis goal, but the methods must also have the capability to properly incorporate and deal with all their features. For instance, they should handle all variables and their dynamics and time-lagged behaviour; this is particularly challenging in large-scale contexts, where classic or VARMA/VARIMA time series models are no longer reliably applicable. Alternatives may pass by the use of dynamic latent variable methods [26,27,91-93] or subspace state-space models. [94-97]

Data Integration (D_3)

This dimension regards the existence of multiple sources of data that convey relevant information for achieving the project goal, if they could be properly integrated through *f*. Examples include the existence of several advanced instrumentation technologies, [72] data collected from different points in the supply chain (raw materials, process, quality laboratory, customer, service, etc.), additional meta data (process and alarm tags), etc.

Temporal Relevance (D_4)

The extraction of knowledge from data happens in a workflow, roughly composed by the following stages: i) planning; ii) data collection; iii) data analysis; iv) deployment. Dimension D_4 regards the impact of the duration of each stage, and the gaps in between, on InfoQ. For instance, the data collection time may increase or decrease InfoQ, depending on whether the study goal is longitudinal or cross-sectional. In predictive studies, the existence of a temporal gap between data collection and deployment usually decreases InfoQ, as a result of unmodelled non-stationary components present in CPI. This is connected with the timeliness concept: [98] "*solving the right problem too late.*"

The time for data analysis is also relevant (i.e., to implement *f*). Methods based on first principle mechanistic models can lead to more accurate and interpretable results, but require highly skilled personnel (often working in academia or consulting companies) and take much more time to be completed, which can raise timeliness issues. On the other hand, data-driven process improvement initiatives can be implemented with the plant personnel and take less time for achieving good results. The complexity of method *f* is another related aspect to consider (see also section on the dimension

"Operationalization"). More complex methods involve more parameters to tune and more time to run and analyse. Unless there is enough information to provide good settings for the tuning parameters, they tend to be less robust and prone to overfitting, despite their potential for being more accurate. [31]

As further example, typical from an Industry 4.0 context, let us consider the case where the goal is to perform a graphical descriptive analysis of a large structured dataset. In this case, conventional Principal Components Analysis (PCA) can lead to higher D_4-scores than the option of graphically analysing all the possible univariate and bivariate plots. PCA only requires the graphical analysis of a few score vectors and loadings, whereas the alternative (graphical analysis of all variables in a univariate or bivariate way) is usually impossible to accomplish within the time-frame available, and is much less informative.

Chronology of Data and Goals (D_5)

This dimension regards the variables selected and the temporal relationships between them, in the context of *g*. Much of the success of constructing models for process optimization and diagnosis goals rely on having access to measurements of critical variability drivers. This is fundamental for developing input-output models for process control & optimization or to perform troubleshooting activities, and not so critical for process monitoring and soft sensor applications. When some of the variation drivers are not observable, latent variable methods provide a way to incorporate their role in the analysis, either explicitly trough path modelling and structural equations, or implicitly through multivariate regression frameworks such as Principal Components Regression and Partial Least Squares. [64,65,99]

The chronology aspect is related to the retrospective vs. prospective scope of the study and to whether the goal is causal explanation, prediction or description. [10] For instance, for soft sensor development, inputs and outputs must be available at the same time and for the same samples, where as in causal explanation applications, the system architecture and operation determine the actual time structure connecting the different variables to be analysed.

In summary, this dimension is related with having the adequate set of variables for achieving the analysis goal, at the right chronological order.

Generalizability (D_6)

This InfoQ dimension regards the potential to generalize the analysis outcome to the desired universe, targeted by the empirical study. Observational data allows inferences regarding similar operation conditions. On the other hand, the active collection of data (through DOE) provides the capability for exploring operation modes beyond those used before, generalizing inferences to other conditions. Therefore, this dimension assesses the ability of X and f to be extended to the circumstances of interest (established in g), as well as the adequacy of U to capture this performance. For instance, more parsimonious models tend to be more stable and generalizable. [31,100] The use of dimensional analysis and the Buckingham π theorem, [101] can be used to extend experimental studies to wider domains. First principle mechanistic models are more generalizable than empirical data-driven models, though more complex (see D_7) and time consuming to develop (see D_4).

In CPI, one is more interested in the "engineering generalization", i.e., extending the results obtained to other conditions, processes, units, laboratories, etc., rather than "statistical generalization", where inferences are restricted to the population where data was collected. For example, over-parametrized data-driven methods are prone to overfitting, which compromises engineering generalizability, namely the ability to predict new observations in the future.

Operationalization (D_7)

This seventh dimension addresses the complexity in operationalizing the empirical study within the existent capabilities of the company. It regards the difficulties involved in data collection, analysis and deployment of solutions. Timeliness (D_4) addressed the aspect of time, but here the emphasis is in the complexity of carrying out the several stages involved and the access to the necessary resources to do it (other than time). Considering the example provided in the section regarding dimension D_4 – Timeliness, and recalling the use of PCA which led to a higher InfoQ score in the visual analysis of large datasets with descriptive purposes: it may happen that, in some contexts (especially in SME's), no employee is aware of how to conduct this analysis or there may be no software available for doing it. This would decrease the InfoQ score regarding operationalization in this particular setting. In general, some methods may offer good solutions to accomplish a given goal, but are outside the domain of skills or assets in the organizations, or their implementation is significantly more complex.

Communication (D_8)

This final dimension comprises the rigor, completeness and clarity, by which the following aspects are established and communicated:

- The goals of the project – to the project team;
- The results obtained – to the project stake holders.

Goals should obey to the SMART principles (Specific, Measurable, Achievable, Relevant and Timely), [102] where it should become clear whether they regard to a prediction, diagnosis (causal explanation) or descriptive analysis, or even the specific activity to be carried out (see list in the second section). The communication of results should be carefully planned, including the methods to adopt (e.g., visualization tools [11]) and organization issues. [103]

Operationalization of the InfoQ assessment in CPI activities

The eight dimensions described in the previous subsection (InfoQ-dimensions) need to be properly combined in order to compute an InfoQ-score.

We propose here a new InfoQ assessment strategy, that is based on the original decomposition of InfoQ into its 4 components, and then on the 8 dimensions that contribute to them: $\theta \to \gamma \to InfoQ$, where $\theta = \begin{bmatrix} D_1 & D_2 & \cdots & D_8 \end{bmatrix}^T$ represents the eight InfoQ-dimensions, and $\gamma = [g, U, X, f]^T$ stands for the 4-dimensional vector of InfoQ-components.

Contrary to the standard assessment protocol, where eight dimensions are combined to compute directly the final Info-Q score, in the new proposed scheme, each InfoQ-dimension, θ_i, is assessed *w.r.t.* a given InfoQ-component, γ_j, with which it is related. In fact, the InfoQ-dimensions are not necessarily related with all InfoQ-components (as will become clear below), and the way each InfoQ-dimension is assessed, depends on the particular component being considered. For instance, one perspective is to analyse *D_1 – data resolution* in the scope of the component, *X – dataset*: “Do collected data have the adequate resolution to address the project goal?”, and another is to assess its quality in the scope of component, *f – method*: "Does the method *f* has the capability to deal with the resolution or multiple resolutions existent in data?" (note that similar questions apply to dimension D_2 – *data structure*, by replacing the term "resolution" with "structure", etc.). These are clearly different problems, which were analysed conjointly in previous InfoQ assessment schemes. The proposed approach has the advantage of making it clear in which scope should each dimension be regarded, making the process more systematic, streamlined and reproducible for practitioners. Furthermore, it safeguards against the possibility that some important aspects are overlooked during the assessment, because their analysis is now explicitly solicited. The new assessment strategy is described in the following sections.

InfoQ assessment structure and workflow

Figure 1 presents the proposed InfoQ decomposition into its 4 InfoQ-components and finally into the 8 InfoQ-dimensions. This scheme provides structure to the quantitative assessment as presented next.

[Insert Figure 1 approximately here]

Figure 1. The decomposition of InfoQ into its components (*X, f, g, U*) and then on the 8 dimensions that determine their quality. Also shown, is the connection between dimensions and the components in which assessment they take part.

The relationships between the InfoQ-dimensions and the components are depicted in Figure 1. As their identification may not be easy, Table 1 provides the respective connectivity matrix.

[Insert Table 1 approximately here]

Table 1. Summary Table of the InfoQ-Dimensions Affecting the Four Components (*X,f,g,U*)

The computation of InfoQ comprises the following three stages to be carried out:

- **Stage 1.** For each component, C_j, assess each dimension, D_i, connected to it (see Figure 1 and Table 1) and compute the associated scores: $score-\mathbf{d}_i^j$
- **Stage 2.** Combine the scores obtained from stage 1 for each component and compute the scores for the quality of each component: $score-\mathbf{d}_i^j \to score-\mathbf{c}_j^{InfoQ}$
- **Stage 3.** Combine the component scores, and obtain InfoQ: $score-\mathbf{c}_j^{InfoQ} \to InfoQ$

In the first stage, the user assesses each dimension, *w.r.t.* a given component. Stage 2 and 3 are computational stages, where the quality of the components and InfoQ are successively obtained. These two stages can easily be implemented with resort to a spreadsheet application (one such application can be made available upon request to the corresponding author).

By decomposing the assessment in a hierarchical way (Figure 1), the assessment is made more focused, objective and reproducible. As referred above, it constitutes a fundamentally different problem assessing the quality of the dimension “Data Structure” *w.r.t.* to the component “Dataset” or to the component “Method”: in the first case, on assess the information content in the dataset for achieving the analysis goal, arising from the underlying structure, whereas in the second case one evaluates the method’s capability to process data with a that structure and produce the desired outcome. Furthermore, it becomes possible to explore different levels of aggregation in the analysis of framework outcomes. These aspects were not so conveniently handled in the conventional InfoQ procedure. To support the assessment to be performed during Stage 1, we present in Appendix A a list of questions that may provide a useful template for users to adopt in the future, which can easily be adapted to new applications.

With the proposed structured assessment scheme supported by focused and precisely made questions, the assessment subjectivity of the original approach is greatly reduced. Assessment subjectivity is just an informal way of referring to measurement uncertainty [104-107] in the current scenario. But measurement uncertainty is a common trace of any evaluation system, and therefore ours must necessarily be subjected to it to some extent.

Methodological and computational details

This section provides more detailed information on how to execute the operations and computations associated with each stage of the proposed procedure for assessing the value of information in data-driven studies, i.e., to compute their InfoQ.

Stage 1

The assessment of each dimension, D_i, *w.r.t.* a given component, C_j, is made by answering the questions in Appendix A, with resort to a Likert scale. For instance, a 5 level Likert scale can be adopted, [1–5], with "1" indicating low achievement in that dimension and "5" indicating high achievement. These ratings, $\left\{\mathbf{d}_i^j\right\}_{i=1:8}$, are filled by the user, and are then normalized using a desirability function approach [108,109] into a scale [0–1], leading to the normalized assessment scores, represented by $score-\mathbf{d}_i^j$, through following mapping: $score-\mathbf{d}_i^j = desirability\left(\mathbf{d}_i^j\right)$, with:

$$desirability\left(x\right)=\begin{cases}0\Leftarrow x=1\\0.25\Leftarrow x=2\\0.5\Leftarrow x=3\\0.75\Leftarrow x=4\\1\Leftarrow x=5\end{cases}\qquad(4)$$

This assessment can be implemented directly by the knowledgeable user, or following the Delphi consensus approach, in case the team counts with several experts on industrial data analytics. [110,111]

Stage 2

In stage 2, the scores obtained from the assessment of the expert are combined in order to obtain the quality of each InfoQ component ($score-\mathbf{d}_i^j\rightarrow score-\mathbf{c}_j^{InfoQ}$). The fusion is made through the weighted geometric mean of the individual desirabilities that are relevant to a given component. Contrary to the original approach, weights are now introduced, to reflect the different focus and priorities associated with the different analysis goals (more details on the selection of the weights are provided further ahead in the text). The computation details are provided below.

- Let $\mathbf{D}^j$ be the set containing the relevant dimensions necessary for establishing the quality of the InfoQ-component, C_j.
- Let $\mathbf{I}^j$ be the set of indices for the dimensions in $\mathbf{D}^j$.

The scores for InfoQ-components, $score-\mathbf{c}_j^{InfoQ}$, are obtained by the following weighted geometric expression:

$$score - \mathbf{c}_j^{InfoQ} = \left\{ \prod_{k \in \mathbf{I}^j} \left(score - \mathbf{d}_k^j \right)^{w_k^j} \right\}^{\frac{1}{\sum_{k \in \mathbf{I}^j} w_k^j}} \tag{5}$$

The weights used in equation (5) depend on the type of problem under analysis. In order to facilitate the InfoQ assessment scheme we have considered 5 distinct categories of problems, and defined typical weighting profiles for each one of them. These categories are: exploratory/descriptive analysis; process monitoring; diagnosis and causal explanation; predictive modelling for virtual metrology; predictive modelling for control & optimization. The tables of weights associated with each category are presented in Appendix B. According to the type of problem, the user just needs to select the corresponding table of weights. These weighting profiles were defined based on the authors' accumulated experience in driving multiple empirical studies and on the hierarchy of importance attributed to the different dimensions w.r.t. to the InfoQ components. Therefore, they reflect, in general, the expert's knowledge and experience, and constitute a way to state his assumptions explicit and defined in a clear and transparent way, making the analysis reproducible and shareable. These weights should be interpreted as the analysis defaults, and the user can always change them to better reflect his priorities or preferences (this is straightforward to do in the evaluation spreadsheet).

Stage 3

In this final stage, the InfoQ-component scores are combined, and the quality of information generated in the data-driven activity obtained. In this work, we did not consider different weights for the different components, which amounts to assume that they are all equally relevant for establishing InfoQ (this option can easily be changed in the future, in case a differentiation reflecting the importance of the different components is justifiable).

$$InfoQ = \left\{ \prod_{j=1:4} \left(score - \mathbf{c}_j^{InfoQ} \right) \right\}^{\frac{1}{4}} \qquad (6)$$

Pre- vs Post- Assessment

Pre-assessment regards the analysis and evaluation of the study design (strategy, technical components, workflow, and restrictions). It is made before applying method *f* to the dataset *X*, but some quick analysis over *X* may be carried out in case (some) data is already available. The main focus is to evaluate the appropriateness of all InfoQ components *w.r.t.* to the goal to be achieved (*g*). We will also call it a Type B assessment ("B", from "Before"). Post-assessment refers to the analysis and evaluation of the way the study was actually carried out, conclusions were made and communicated, as well as the quality of the final results achieved and how they were assessed. It will be also referred as Type A assessment ("A", from "After").

The difference between pre- and pos- assessment is important, as they have different purposes and imply distinct mindsets. Type B assessment addresses the planning, cost-benefit and risk assessment stages of a project. The decisions may be to improve it (using the assessment outcomes) and to go ahead, or to decide not to implement the study, due to adverse cost-benefit considerations (high cost of data collection) or intrinsic limitations (no data available under certain conditions, or with a given structure) in the design. On the other hand, Type A is more directed to the critic analysis of the study, making sure that the results are solid and meaningful, in order to use them as factual support for decision making. For instance, when addressing a prediction problem, the utility, *U*, may be the value of R^2_{Pred} actually achieved in the empirical study for a Type A assessment, whereas for a Type B it is the appropriateness of using this metric *w.r.t.* to the goal that is accessed.

Case Studies

In this section, several case studies regarding different industrial data-centric activities are presented, and the quality of information generated assessed. The impact of options followed at the level of the methods adopted, *f*, or regarding features present in the dataset, *X*, are also brought to the analysis and discussed.

InfoQ assessment of an industrial semiconductor process

Description

This case study is based on a project conducted in collaboration with a semiconductor company (companies' name cannot be disclosed), whose purpose was to derive an inferential model (virtual metrology) to be used in the future for purposes of fast release of wafer batches and maybe for process control (run-to-run control). FDC data was provided by the semiconductor manufacturer (FDC means Fault Detection and Classification, and consists mostly of process operation variables, such as flows, pressures, temperatures, etc.), together with Metrology data for the key dimensions of the wafer. The FDC data contains information about almost 1000 wafer batches, but the Metrology data was collected for only approximately 50 batches, which furthermore do not always coincide with those in the FDC dataset.

The analytics team decided to fuse the two datasets (FDC and Metrology) using the wafer lot code as reference for performing the merging operation, and developed inferential models using several predictive modelling approaches, such as least squares regression with variable selection (forward stepwise variable selection), penalized regression (LASSO) and partial least squares (PLS). The methods' performance was assessed on the basis of the prediction cross-validation errors (internal validation). Good fitting and predictive scores were obtained for the least squares variable selection methodology.

First InfoQ Assessment

Implementing the workflow for InfoQ assessment (Stage 1), each component was evaluated using the dimensions that are relevant for its quality (see Table 1). The following paragraphs contain some observations of the ratings given to each dimension *w.r.t.* to a given component (*g, U, X, f*). The weights profiles were selected for the problem category: predictive modelling for virtual metrology (Table B4).

- *Assessing InfoQ-X.* Several datasets are available, namely FDC and Metrology data, but their integration is limited because the overlap of records from both sources is low (small number of records for the same wafers). Therefore, the collection protocol could have been better designed from the standpoint of potentiating better integration capabilities (*D3*). Therefore, the rating given to the dimension 3 (data integration) *w.r.t.* X (dataset), is $D_3^X = 3$ (scale [1-5]). The low superposition between datasets also causes many records to be discarded, leading to low resolution data, $D_1^X = 3$. On the other hand, the dataset took considerable time to be collected and to be made available to the analytics team, and the collection process was very complex ($D_7^X = 4$) – by the time it was analysed, the process may have suffered changes, which can limit the deployment of results ($D_4^X = 3$). The data structure correspond to a 2-way table composed by observational or passively collected data ($D_2^X = 4$), and the main process variables were included in the analysis ($D_5^X = 5$), which are both positive aspects for developing a Virtual Metrology predictive model for this process. However, the generalization to other products and tools is limited, as the inference basis provided by the collected data is restricted ($D_6^X = 3$).

- *Assessing InfoQ-f*. The methods adopted are in general capable to deal with the features present in the dataset, such as multicollinearity, sparsity and noise ($D_1^f = 5$, $D_2^f = 5$ $D_3^f = 4$), and can be implemented in useful time and within the resources available in the team ($D_4^f = 5$, $D_7^f = 5$). The methods also have built-in features for selecting the relevant variables ($D_5^f = 5$) and for generalization to the process of interest (namely parsimony and parameter estimation stability, $D_6^f = 4$). The results were properly communicated using a summary table containing the relevant performance indicators, supported by graphs ($D_8^f = 5$).
- *Assessing InfoQ-g*. It is not clear from the goal statement whether the objective is to develop a predictive model for Virtual Metrology or for Control/Optimization. A better goal definition is therefore needed ($D_8^g = 3$), as the nature of the models required for these two goals can be quite different.

- *Assessing InfoQ-U*. The performance of the predictive model was evaluated using cross-validation, which is a sound approach for assessing the predictive capabilities of the model, under situations where data is not so abundant. However an independent test set would be a preferable solution in the future, especially if the purpose is to conduct process control ($D_6^U = 4$).

The assessment of the initial study resulted in the combined scores for the components and in the overall InfoQ, presented in [Insert Figure 2 approximately here]

Figure 2. From the analysis of these results, one can verify that the overall quality of information is not high (0,68), and the main concerns are in the InfoQ-components: dataset (*X*) and goal (*g*). Therefore these components should be carefully analyzed (focusing on the dimensions that contribute to them) and solutions devised for their improvement, in order to increase the value of information generated in the study.

[Insert Figure 2 approximately here]

Figure 2. Semiconductor case study. Decomposition of the InfoQ assessment: analysis of the initial study.

Final InfoQ Assessment

After a careful analysis of the elements of the initial study and the InfoQ assessment performed, several improvement opportunities were identified, namely:

- The decision of the data to be collected should result from a consensus discussion between the process team and the analytics team and not only a decision of the process team. With this, better integration capabilities ($D_3^X = 5$)

can be expected and the resolution of data will also be improved ($D_1^X = 5$), as well as their structure ($D_2^X = 4$).

- The goal definition must also be clearly defined, namely if it regards the development of a virtual metrology model, or if the purpose is to derive an input-output model for process control and optimization. This can make a significant difference on the type of models and the data structure required for analysis. For instance, input-output models for process control involve the realization of system identification experiments, which were not contemplated in the original data collection plan ($D_8^g = 5$).
- An independent test set should be collected, especially if the purpose is to conduct process control ($D_6^U = 5$).

With these changes implemented in the future, the quality of information generated by the study can improve from the initial level of 0.68 to 0.92, indicating a significantly higher level of achievement of the project goals (Figure 3).

[Insert Figure 3 approximately here]

Figure 3. Semiconductor case study. Decomposed InfoQ assessment: follow-up analysis.

InfoQ assessment of an industrial crystallization process

Description

A dataset was collected from an industrial crystallization operating in a batch mode [31], i.e., where process variables can present different types of profiles (usually, non-stationary), that repeat themselves from batch to batch. In the process under analysis, there are two driers operating in parallel, which lately have been found to be producing products with different levels of impurity. This happens even though they both share the same recipes and follow the same stages, a situation that motivated a more detailed analysis of their operation, looking for possible sources for the different behavior. The goal of the study was to conduct an exploratory analysis over the collected dataset, in order to identify relevant patterns of variation in the driers' operation and pinpoint possible sources of systematic or unstructured variability justifying the different levels of impurities. An additional goal was also set by the industrial company, which consisted in developing a model that could reasonably predict the amount of impurities.

Pre-assessment of the exploratory study (Type B)

The InfoQ assessment for each component, performed before conducting the exploratory analysis, can be summarized as follows.

- *Assessing InfoQ-X*. Data was collected at a sufficiently high sampling rate, using the IT industrial infrastructure and the local DCS ($D_1^X = 5$). The data collection process if facilitated by the existence of a query system that speeds up data retrieving from the process units, requiring some curation (for instance, not all variables are relevant in all stages) and cleaning. Given its complex structure (see below), data is not easy to handle or manipulate ($D_7^X = 4$), which also introduces some delay in the process ($D_4^X = 4$). Process data was fused with the available quality data (impurity measurements), leading to the final merged dataset for analysis ($D_3^X = 5$). Being a batch process, the data structure is very complex. It consists of time-profiles for the process variables, for the different stages and for all batches. These profiles do not have the same duration, nor are synchronized/aligned, as required by many batch data analysis tools [112-115]. Even though this operationalization issues, data potentially contains all the necessary information for conducting the exploratory study ($D_2^X = 5$), as all relevant process variables were successfully collected (overall, 20 process variable were analyzed, in the different batch stages; $D_5^X = 5$). The dataset is confined to a process unit and limited operation, and any generalization is therefore limited ($D_6^X = 3$).

- *Assessing InfoQ-f*. The complexity of multi-stage, unsynchronized batch data is very high, and there is a current lack of methodological approaches to handle its resolution and structure, in an efficient way ($D_1^f = 3$, $D_2^f = 2$). Combining all variable profiles from several stages is also difficult, given the variability in their durations and misalignment, which raise relevant integration and operationalization problems ($D_3^f = 2$, $D_7^f = 3$) to the current 2-way [28,29,116,117] and 3-way [118-120] batch data analysis approaches. These methods are not particularly capable of performing variable selection and are often overparameterized, hence potentially unstable (see ref. [31] for a discussion); therefore $D_5^f = 3$ and $D_6^f = 3$. This obstacles increase the analysis time and delay the deployment of results ($D_4^f = 3$). Finally, the results obtained by classical methods are not easy to communicate, as they involve the time-resolved analysis of many variables, in multiple stages ($D_8^f = 3$).
- *Assessing InfoQ-g*. The goal statement is clear regarding the purpose of the exploratory study. However, the secondary objective is not so clear, and the

predictive goal should be better defined. It may be the case that the purpose is not even predictive, but of conducting a diagnosis of potential root-causes using an empirical modelling approach ($D_8^g = 4$). For this reason, this goal was not further pursued.

- *Assessing InfoQ-U*. The performance of the exploratory study is mainly evaluated by the clarity, interpretability and usefulness of the graphical outcomes produced by the analysis. Contrary to other descriptive studies, where goodness-of-fit measures are recommended, here the goal is essentially process data visualization, regarding which more research is needed to define clear figures of merit to evaluate the associated performance. The evaluation is therefore qualitative and based on the knowledge of the spectrum of solutions available ($D_6^U = 4$).

With the evaluation performed during Stage 1 as detailed above, the computations of Stage 2 were performed using the table or weights for problems of the type "exploratory/descriptive analysis" (Table B1). The overall InfoQ was then computed in Stage 3, using equation (6). Figure 4 presents the pre-assessment of the InfoQ for the exploratory empirical study, indicating a rather moderate potential for generating information of quality. The main reasons lie in methodological limitations (component *f*) to cope with the complexity of the available dataset.

[Insert Figure 4 approximately here]

Figure 4. Crystallization case study. Decomposed InfoQ assessment: pre-assessment of the exploratory analysis study (Type B).

Post-assessment of the exploratory study (Type A)

In order to circumvent the main limitations detected during the pre-assessment stage, a new framework was developed to handle complex batch process data. This framework, called FOBA (Feature Oriented Batch Analytics), provides an alternative solution to handle batch data, by converting profiles into features that characterize the fundamental aspects of their trends. Features are profile-specific, and therefore the individuality of the time-profiles is retained and preserved. In this approach, there is no need to perform complex synchronization and alignment operations (which are simply transferred to normal variation in the features domain) and the batch runs can have different durations. In the feature-oriented representation, the three-dimensional array is converted into a structured 2-way array, where most of the current well-developed analytical approaches can be applied. The implementation of FOBA to this dataset, easily identified a subset of variables, and their respective stages, that exhibit distinct trajectories in the two driers (a difference that was also found to be statistically significant), namely those regarding product feed and water washing stages. With this approach available and implemented, a post-evaluation was made, as described next.

- *Assessing InfoQ-X.* This component was left unchanged, as no major concerns were found at the level of the dataset. Therefore, using the Pareto action principle, the focus is directed primarily to the few aspects having a bigger impact in information quality, which lie in component *f*.
- *Assessing InfoQ-f.* By design, FOBA is able to handle batch data with arbitrary complexity regarding the resolution, structure and integration dimensions ($D_1^f = 5$, $D_2^f = 5$, $D_3^f = 5$). The analysis becomes much simpler ($D_7^f = 5$) and faster ($D_4^f = 5$), which also means that more people will be able to apply it in

the shop floor and benefit from the analysis of data from batch processes for process improvement. The analysis models constructed from it are more parsimonious, which makes them more stable and generalizable (see examples in [31]; $D_6^f = 4$). Efficient approaches for variable selection can be implemented, allowing the identification of relevant variables/stages ($D_5^f = 5$). All results can be easily communicated using conventional plots, because the problem was converted into the feature space $D_8^f = 5$.

- *Assessing InfoQ-g* and *Assessing InfoQ-U*. The evaluation of these components was maintained, since the project goal were not revised and no conceptual evolution took place during the project duration, regarding new and better figures of merit for exploratory process visualization studies.

With the introduction of FOBA in the analysis of the industrial dataset, a significant evolution in the quality of component *f* was accomplished, which resulted in an increase of InfoQ from 0.62 to 0.85 (Figure 5), which is an interesting score for an exploratory study.

[Insert Figure 5 approximately here]

Figure 5. Crystallization case study. Decomposed InfoQ assessment: post-assessment of the exploratory analysis study (Type A).

Post-assessment assessment of the predictive study

A clarification of the objective regarding the secondary aspect, led to an improvement of the accuracy in the goal statement. The purpose should not be to develop a predictive model of the impurities level in the final product ($D_8^g = 5$) with the purpose of process tuning and optimization. The dataset does not present the right structure for that, nor exhibit enough variability to embrace this ambitious goal.

Conclusions

In this article we present a framework aimed at assessing the value of information generated in data-centric activities in the context of Chemical Processing Industry. The proposed framework decomposes the assessment of information quality (InfoQ) into 4 components (analysis goal, data set, methods and utility), which are related through equation (1) and can be evaluated using 8 dimensions. These dimensions are evaluated by the user using a template of questions (Appendix A) leading to scores expressed in a Linkert scale. The assessment scores are then processed in 2 stages, leading to normalised dimension scores, component scores, and finally to InfoQ. The decomposition approach followed provides a clear structure to the assessment scheme (previously, the assessment was made in a single step), making it more systematic, objective, reproducible and informative. Weights are now introduced, reflecting the different focus and priorities associated with the different analysis goals. Several weightings profiles are suggested for different online or off-line data analysis activities (process monitoring, predictive modelling, diagnosis or causal explanation, etc.), which can easily be adjusted by the user to reflect personal experience or preferences (Appendix B).

The proposed approach fills an existing gap in the assessment of the merits of data-centric activities. It is therefore an important asset providing support to the management of these activities in Industry 4.0 scenarios, where their importance will tend to grow and the quality of information generated will become a central issue.

The InfoQ framework can be applied, for instance, in the following contexts:

- Planning and optimization of data-driven activities in Industry 4.0.
- Risk assessment of data-driven empirical studies (InfoQ-RISK analysis).
- Tool for supporting decision making on how to improve the design of data-driven activities, maximizing InfoQ.
- A posteriori diagnosis and reporting of strengths and weaknesses of any data analysis activities (InfoQ-SWOT analysis).

Future work will contemplate the reporting and analysis of more applications of this methodology, with the purpose supporting practitioners in developing their data-centric projects in the era of Industry 4.0.

Acknowledgments

Marco Reis acknowledges financial support through project 016658 (references PTDC/QEQ-EPS/1323/2014, POCI-01-0145-FEDER-016658) financed by Project 3599-PPCDT (Promover a Produção Científica e Desenvolvimento Tecnológico e a Constituição de Redes Temáticas) and co-financed by the European Union’s FEDER.

Literature Cited

1. Reis MS, Braatz RD, Chiang LH. Big Data - Challenges and Future Research Directions. *Chemical Engineering Progress.* 2016;Special Issue on Big Data(March):46-50.

2. Colegrove LF, Seasholtz MB, Khare C. Big Data - Getting Started on the Journey. *Chemical Engineering Progress.* 2016;Special Issue on Big Data(March):41-45.

3. Coleman S, Goeb R, Manco G, Pievatolo A, Tort-Martorell X, Reis MS. How can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International.* 2016;32(6):2151-2164.

4. Qin SJ. Process data Analytics in the Era of Big Data. *AIChE Journal.* 2014;60(9):3092-3100.

5. Marr B. *Big Data - Using Smart Big Data Analytics to Make Better Decisions and Improve Performance*. Chichester: Wiley; 2015.

6. Hoerl RW, Snee RD. Post-financial meltdown: What do the services industries need from us now? (with discussion). *Applied Stochastic Models in Business and Industry.* 2009;25:509-521.

7. Hoerl RW, Snee RD, De Veaux RD. Applying statistical thinking to ‘Big Data’ problems. *WIREs Comput Stat.* 2014;6(222-232).

8. Harford T. Big data: are we making a big mistake. *Significance.* 2014;December:14-19.

9. Kenett RS, Shmueli G. *Information Quality: The Potential of Data and Analytics to Generate Knowledge*: Wiley; 2016.

10. Kenett RS, Shmueli G. On Information Quality. *Journal of the Royal Statistical Society A.* 2014;177(1):3-38.

11. Tufte ER. *The Visual Display of Quantitative Information*. 2nd. ed. ed. Cheshire, Connecticut: Graphics Press; 2001.

12. Shewhart WA. *Economic Control of Quality of Manufactured Product.* Vol Republished in 1980 as a 50th Anniversary Commemorative Reissue by ASQC Quality Press. New York: D. Van Nostrand Company, Inc.; 1931.

13. Roberts SW. Control Charts Tests Based on Geometric Moving Averages. *Technometrics.* 1959;1(3):239-250.

14. Page ES. Continuous Inspection Schemes. *Biometrics.* 1954;41(1-2):100-115.

15. Hotelling H. Multivariate quality control, illustrated by the air testing of sample bombsights. In: Eisenhart C, Hastay MW, Wallis WA, eds. *Selected Techniques of Statistical Analysis*. New-York: McGraw-Hill; 1947.

16. Lowry CA, Woodall WH, Champ CW, Rigdon SE. A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics.* 1992;34(1):46-53.

17. Jackson JE. Quality Control Methods for Several Related Variables. *Technometrics.* 1959;1(4):359-377.

18. Jackson JE, Mudholkar GS. Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics.* 1979;21(3):341-349.

19. Kourti T, MacGregor JF. Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods. *Chemometrics and Intelligent Laboratory Systems.* 1995;28:3-21.

20. Kresta JV, MacGregor JF, Marlin TE. Multivariate Statistical Monitoring of Process Operating Performance. *The Canadian Journal of Chemical Engineering.* 1991;69:35-47.

21. MacGregor JF, Jaeckle C, Kiparissides C, Koutoudi M. Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE Journal.* 826-838 1994;40(5).

22. MacGregor JF, Kourti T. Statistical Process Control of Multivariate Processes. *Control Engineering Practice.* 1995;3(3):403-414.

23. Montgomery DC, Mastrangelo CM. Some Statistical Process Control Methods for Autocorrelated Data. *Journal of Quality Technology.* 1991;23(3):179-193.

24. Harris TJ, Ross WH. Statistical Process Control Procedures for Correlated Observations. *The Canadian Journal of Chemical Engineering.* 1991;69:48-57.

25. Ku W, Storer RH, Georgakis C. Disturbance Detection and Isolation by Dynamic Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems.* 1995;30: 179-196.

26. Rato TJ, Reis MS. Fault detection in the Tennessee Eastman process using dynamic principal components analysis with decorrelated residuals (DPCA-DR). *Chemometrics and Intelligent Laboratory Systems.* 2013;125:101-108.

27. Rato TJ, Reis MS. Advantage of Using Decorrelated Residuals in Dynamic Principal Component Analysis for Monitoring Large-Scale Systems. *Industrial & Engineering Chemistry Research.* 2013;52(38):13685-13698.

28. Nomikos P, MacGregor JF. Monitoring Batch Processes Using Multiway Principal Component Analysis. *AIChE Journal.* 1994;40(8):1361-1375.

29. Nomikos P, MacGregor JF. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics.* 1995;37(1):41-59.

30. Rato TJ, Blue J, Pinaton J, Reis MS. Translation Invariant Multiscale Energy-based PCA (TIME-PCA) for Monitoring Batch Processes in Semiconductor Manufacturing. *IEEE Transactions on Automation Science and Engineering.* 2017;14(2):894-904.

31. Rendall R, Lu B, Castillo I, Chin S-T, Chiang LH, Reis MS. A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes. *Industrial & Engineering Chemistry Research.* 2017;56(30):8590-8605.

32. Van Sprang ENM, Ramaker H-J, Westerhuis JA, Smilde AK. Statistical Batch Process Monitoring using Gray Models. *AIChE Journal.* 2005;51(3):931-945.

33. Westerhuis JA, Kourti T, MacGregor JF. Comparing Alternative Approaches for Multivariate Statistical Analysis of Batch Process Data. *Journal of Chemometrics.* 1999;13:397-413.

34. Kang L, Albin SL. On-Line Monitoring When the Process Yields a Linear Profile. *Journal of Quality Technology.* 2000;32(4):418-426.

35. Kim K, Mahmoud MA, Woodall WH. On the Monitoring of Linear Profiles. *Journal of Quality Technology.* 2003;35(3):317-328.

36. Reis MS, Saraiva PM. Multiscale Statistical Process Control of Paper Surface Profiles. *Quality Technology and Quantitative Management.* 2006;3(3):263-282.

37. Woodall WH, Spitzner DJ, Montgomery DC, Gupta S. Using Control Charts to Monitor Process and Product Quality Profiles. *Journal of Quality Technology.* 2004;36(3):309-320.

38. Pereira AC, Reis MS, Saraiva PM. Quality control of food products using image analysis and multivariate statistical tools. *Industrial & Engineering Chemistry Research.* 2009;48(2):988-998.

39. Reis MS, Bauer A. Wavelet texture analysis of on-line acquired images for paper formation assessment and monitoring. *Chemometrics and Intelligent Laboratory Systems.* 15 February 2009 2009;95(2):129-137.

40. Reis MS, Gins G. Industrial Process Monitoring in the Big Data/Industry 4.0 Era: From Detection, to Diagnosis, to Prognosis. *Processes.* 2017;5,35:1-16.

41. Rato TJ, Reis MS. Statistical Monitoring of Control Loops Performance: An Improved Historical-data Benchmark Index. *Quality and Reliability Engineering International.* 2010;26(8):831-844.

42. Rato TJ, Reis MS. On-line process monitoring using local measures of association. Part I: Detection performance. *Chemometrics and Intelligent Laboratory Systems.* 2015;142:255-264.

43. Rato TJ, Reis MS. On-line process monitoring using local measures of association. Part II: Design issues and fault diagnosis. *Chemometrics and Intelligent Laboratory Systems.* 2015;142:265-275.

44. Rato TJ, Reis MS. Multiscale and Megavariate Monitoring of the Process Networked Structure: M2NET. *Journal of Chemometrics.* 2015;29(5):309-322.

45. Rato TJ, Reis MS. Markovian and Non-Markovian Sensitivity Enhancing Transformations for Process Monitoring. *Chemical Engineering Science.* 2017;163:223-233.

46. Gruber A, Yanovski S, Ben-Gal I. Condition-Based Maintenance via Simulation and a Targeted Bayesian Network Metamodel. *Quality Engineering.* 2013;25:370-384.

47. Meeker WQ, Hong Y. Reliability Meets Big Data: Opportunities and Challenges. *Quality Engineering.* 2014;26(1):102-116.

48. MacGregor JF, Kourti T. Multivariate Statistical Treatment of Historical Data for Productivity and Quality Improvements. Paper presented at: Foundation of Computer Aided Process Operations - FOCAPO 981998.

49. Reis MS, Saraiva PM. Multivariate and Multiscale Data Analysis. In: Coleman S, Greenfield T, Stewardson D, Montgomery DC, eds. *Statistical Practice in Business and Industry*. Chichester: Wiley; 2008:337-370.

50. Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. Genetic Algorithms as a Method for Variable Selection in Multiple Linear Regression and Partial Least Squares Regression, with Applications to Pyrolysis Mass Spectrometry. *Analytica Chimica Acta.* 1997;348:71-86.

51. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems.* 2012;118:62-69.

52. Andersen CM, Bro R. Variable selection in regression—a tutorial. *Journal of Chemometrics.* 2010;24(11-12):728-737.

53. Draper NR, Smith H. *Applied Regression Analysis*. 3rd ed. New York: Wiley; 1998.

54. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological).* 1996;58(1):267-288.

55. Rasmussen MA, Bro R. A tutorial on the Lasso approach to sparse modeling. *Chemometrics and Intelligent Laboratory Systems.* 10/1/ 2012;119:21-31.

56. Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning.* Vol 1: Springer series in statistics Springer, Berlin; 2001.

57. Hesterberg T, Choi NH, Meier L, Fraley C. Least angle and ℓ1 penalized regression: A review. *Statistics Surveys.* 2008;2:61-93.

58. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2005;67(2):301-320.

59. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Statistics and computing.* 2004;14(3):199-222.

60. Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews.* 2010;29(5-6):594-621.

61. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. NY: Springer; 2001.

62. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems.* 2001;58(2):109-130.

63. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Analytica Chimica Acta.* // 1986;185:1-17.

64. Martens H, Naes T. *Multivariate Calibration*. Chichester: Wiley; 1989.

65. Jackson JE. *A User's Guide to Principal Components*. New York: Wiley; 1991.

66. Kourti T. Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing.* 2005;19:213-246.

67. Dietterich TG. Ensemble methods in machine learning. *Multiple classifier systems*: Springer; 2000:1-15.

68. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods.* 2009;14(4):323.

69. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and regression trees*: CRC press; 1984.

70. Reis MS. Network-Induced Supervised Learning: Network-Induced Classification (NI-C) and Network-Induced Regression (NI-R). *AIChE Journal.* 2013;59(5):1570-1587.

71. Reis MS. Applications of a new empirical modelling framework for balancing model interpretation and prediction accuracy through the incorporation of clusters of functionally related variables. *Chemometrics and Intelligent Laboratory Systems.* 2013;127:7-16.

72. Campos MP, Sousa R, Pereira AC, Reis MS. Advanced predictive methods for wine age prediction: Part II - a comparison study of multiblock regression approaches. *Talanta.* 2017;171:121-142.

73. Kourti T, Nomikos P, MacGregor JF. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control.* 1995;5:277-284.

74. Naes T, Tomic O, Afseth NK, Segtnan V, Måge I. Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemometrics and Intelligent Laboratory Systems.* 2013;124:32-42.

75. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Reserach.* 2014;238:391-403.

76. Westerhuis JA, Kourti T, MacGregor JF. Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics.* 1998;12:301-321.

77. Kenett RS. Bayesian networks: Theory, applications and sensitivity issues. *Encyclopedia with Semantic Computing and Robotic Intelligence.* 2017;1(1).

78. Whittaker J. *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley; 1990.

79. Pourret O, Naïm P, Marcot B. *Bayesian Networks - A Practical Guide to Applications*. Chichester: Wiley; 2008.

80. Silva BMA, Vicente S, Cunha S, et al. Retrospective Quality by Design (rQbD) applied to the Optimization of Orodispersible Films. *International Journal of Pharmaceutics.* 2017;528(1-2):655-663.

81. Jaeckle C, MacGregor JF. Product Design through Multivariate Statistical Analysis of Process Data. *AIChE Journal.* 1998;44(5):1105-1118.

82. Jaeckle C, MacGregor JF. Product Transfer Between Plants Using Historical Process Data. *AIChE Journal.* 2000;46(10):1989-1997.

83. Tukey JW. The future of data analysis. *Annals of Mathematical Statistics.* 1962;33(1):1-67.

84. Rendall R, Pereira AC, Reis MS. Advanced predictive methods for wine age prediction: Part I - a comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods. *Talanta.* 2017;171:341-350.

85. Godfrey AB. Eye on Data Quality. *Six Sigma Forum Magazine.* 2008;8:5-6.

86. Montgomery DC, Runger GC. *Applied statistics and probability for engineers*: John Wiley & Sons; 2010.

87. Reis MS, Saraiva PM. Multiscale Statistical Process Control with Multiresolution Data. *AIChE Journal.* 2006;52(6):2107-2119.

88. Rato TJ, Reis MS. Multiresolution Soft Sensors (MR-SS): A New Class of Model Structures for Handling Multiresolution Data. *Industrial & Engineering Chemistry Research.* 2017;56(13):3640-3654.

89. Lu N, Yang Y, Gao F, Wang F. Multirate Dynamic Inferential Modeling for Multivariable Processes. *Chemical Engineering Science.* 2004;59:855-864.

90. Kenett RS, Zacks S. *Modern Industrial Statistics - Design and Control of Quality and Reliability*. Pacific Grove: Duxbury Press; 1998.

91. Rato TJ, Reis MS. Defining the structure of DPCA models and its impact on process monitoring and prediction activities. *Chemometrics and Intelligent Laboratory Systems.* 2013;125:74-86.

92. Kaspar MH, Ray WH. Dynamic PLS Modelling for Process Control. *Chemical Engineering Science.* 1993;48(20): 3447-3461.

93. Lakshminarayanan S, Shah SL, Nandakumar K. Modeling and Control of Multivariable Processes: Dynamic PLS Approach. *AIChE Journal.* 1997;43(9): 2307-2322.

94. Shi R, MacGregor JF. Modeling of Dynamic Systems Using Latent Variable and Subspace Methods. *Journal of Chemometrics.* 2000;14:423-439.

95. Juricek BC, Seborg DE, Larimore WE. Identification of Multivariable, Linear, Dynamic Models: Comparing Regression and Subspace Techniques. *Industrial & Engineering Chemistry Research.* 2002;41:2185-2203.

96. Jiang B, Zhu X, Huang D, Paulson JA, Braatz RD. A Combined Canonical Variate Analysis and Fisher Discriminant Analysis (CVA-FDA) Approach for Fault Diagnosis. *Computers & Chemical Engineering.* 2015;77:1-9.

97. Jiang B, Zhu X, Huang D, Braatz RD. Canonical Variate Analysis-based Monitoring of Process Correlation Structure using Causal Feature Representation. *Journal of Process Control.* 2015;32(109-116).

98. Raiffa H. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*: Addison-Wesley; 1970.

99. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*. 5th ed. Upper Sadle River, NJ: Prentice Hall; 2002.

100. Nikolaou M, Vuthandam P. FIR Model Identification: Parsimony Through Kernel Compression with Wavelets. *AIChE Journal.* 1998;44(1):141-150.

101. Shen W, Davis T, Lin DKJ, Nachtsheim CJ. Dimensional Analysis and its Applications in Statistics. *Journal of Quality Technology.* 2014;46(3):185-198.

102. Kubiak TM, Benbow DW. *The Certified Six SIgma Black Belt Handbook*. 2nd ed. Milwaukee, Wisconsin: ASQ Quality Press; 2009.

103. Doumont J-L. *Trees, Maps, and Theorems - Effective communication for ratinal minds*. Kraainem, Belgium: Principiæ; 2009.

104. BIPM, IEC, IFCC, et al. *Guide to the Expression of Uncertainty*. Geneva, Switzerland: ISO; 1993.

105. Kimothi SK. *The Uncertainty of Measurements*. Milwaukee: ASQ; 2002.

106. Reis MS, Rendall R, Chin S-T, Chiang LH. Challenges in the specification and integration of measurement uncertainty in the development of data-driven models for the chemical processing industry. *Industrial & Engineering Chemistry Research.* 2015;54:9159-9177.

107. Reis MS, Saraiva PM. Integration of Data Uncertainty in Linear Regression and Process Optimization. *AIChE Journal.* 2005;51(11):3007-3019.

108. Figini S, Kenett RS, Salini S. Integrating operational and financial risk assessments. *Quality and Reliability Engineering International.* 2010;26:887-897.

109. Derringer G, Suich R. Simultaneous optimization of several response variables. *Journal of Quality Technology.* 1980;12:214-219.

110. Giannarou L, Zervas E. Using Deplhi technique to build consensus in practice. *Journal of Business Science and Applied Management.* 2014;9(2):65-82.

111. Dalkey N. An experimental study of group opinion: The Delphi method. *Futures.* 1969;1(5):408-426.

112. González-Martínez JM, Ferrer A, Westerhuis JA. Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping. *Chemometrics and Intelligent Laboratory Systems.* 2011;105(2):195-206.

113. Ündey C, Williams BA, Çinar A. Monitoring of Batch Pharmaceutical Frementations: Data Synchronization, Landmark Alignment, and Real-Time Monitoring Paper presented at: 15th IFAC World Congress of Automatic Control2002; Barcelona, Spain.

114. Gins G, Van den Kerkhof P, Van Impe JFM. Hybrid Derivative Dynamic Time Warping for Online Industrial Batch-End Quality Estimation. *Industrial & Engineering Chemistry Research.* 2012;51(17):6071-6084.

115. Kassidas A, MacGregor JF, Taylor PA. Synchronization of batch trajectories using dynamic time warping. *AIChE Journal.* 1998;44:864-875.

116. Wold S, Kettaneh N, Friden H, Holmberg A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics and Intelligent Laboratory Systems.* 1998;44:331-340.

117. González-Martínez JM, Vitale R, De Noord OE, Ferrer A. Effect of Synchronization on Bilinear Batch Process Modeling. *Industrial & Engineering Chemistry Research.* 2014;53:4339-4351.

118. Yoo CK, Lee J-M, Vanrolleghem PA, Lee I-B. On-line monitoring of batch processes using multiway independent component analysis. *Chemometrics and Intelligent Laboratory Systems.* 2004;71(2):151-163.

119. Smilde AK, Bro R, Geladi P. *Multi-way Analysis with Applications in the Chemical Sciences*. Chichester, UK: Wiley; 2004.

120. Louwerse DJ, Smilde AK. Multivariate statistical process control of batch processes based on three-way models. *Chemical Engineering Science.* 2000;55(7):1225-1235.

List of Figures Captions

Figure 1. The decomposition of InfoQ into its components (*X, f, g, U*) and then on the 8 dimensions that determine their quality. Also shown, is the connection between dimensions and the components in which assessment they take part.

Figure 2. Semiconductor case study. Decomposition of the InfoQ assessment: analysis of the initial study.

Figure 3. Semiconductor case study. Decomposed InfoQ assessment: follow-up analysis.

Figure 4. Crystallization case study. Decomposed InfoQ assessment: pre-assessment of the exploratory analysis study (Type B).

Figure 5. Crystallization case study. Decomposed InfoQ assessment: post-assessment of the exploratory analysis study (Type A).

Figures

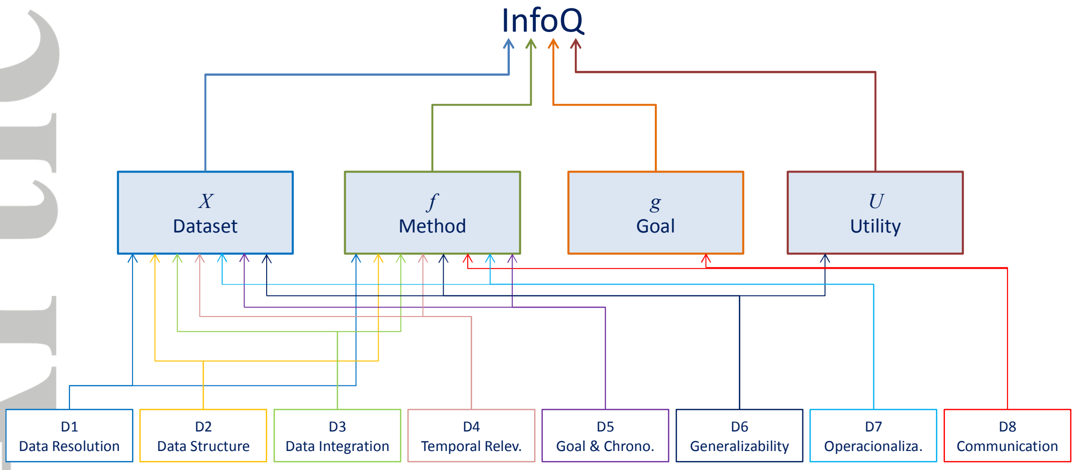

Figure 1

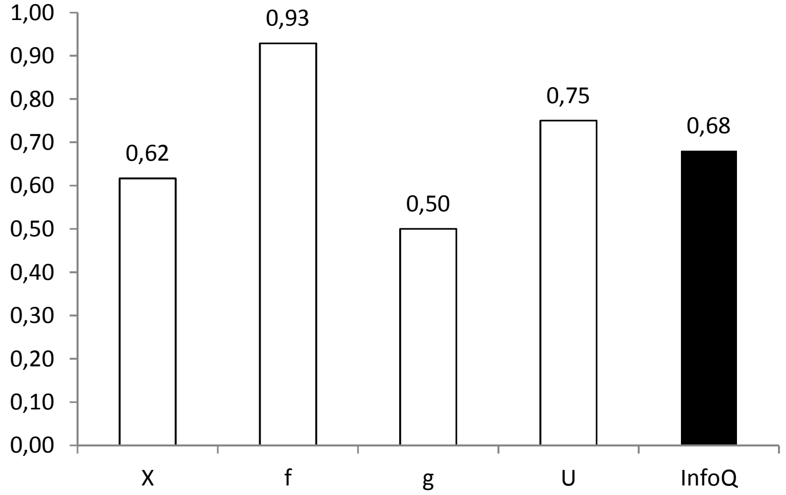
1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
0,62
0,93
0,50
0,75
0,68
X
f
g
U
InfoQ

Figure 2

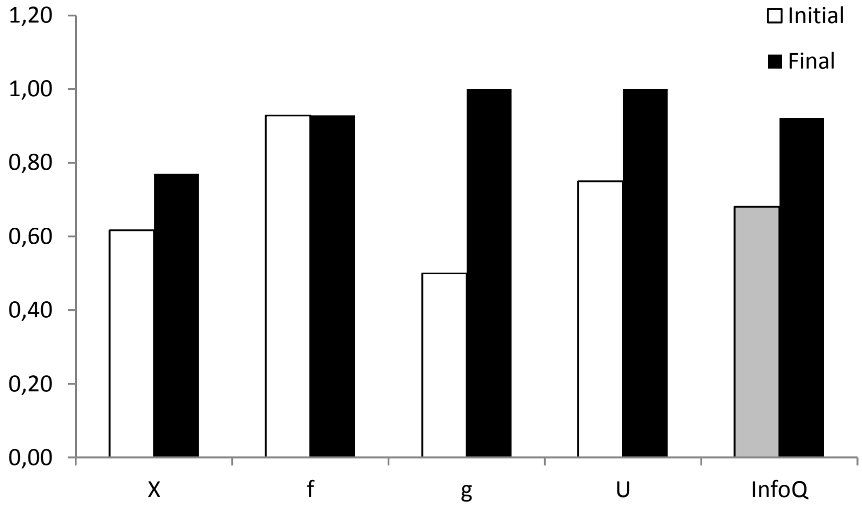

Figure 3

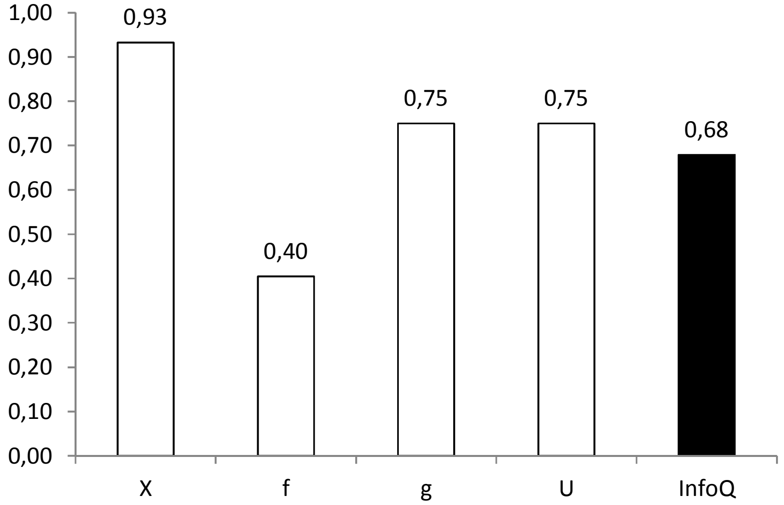
1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
0,93
0,40
0,75
0,75
0,68
X
f
g
U
InfoQ

Figure 4

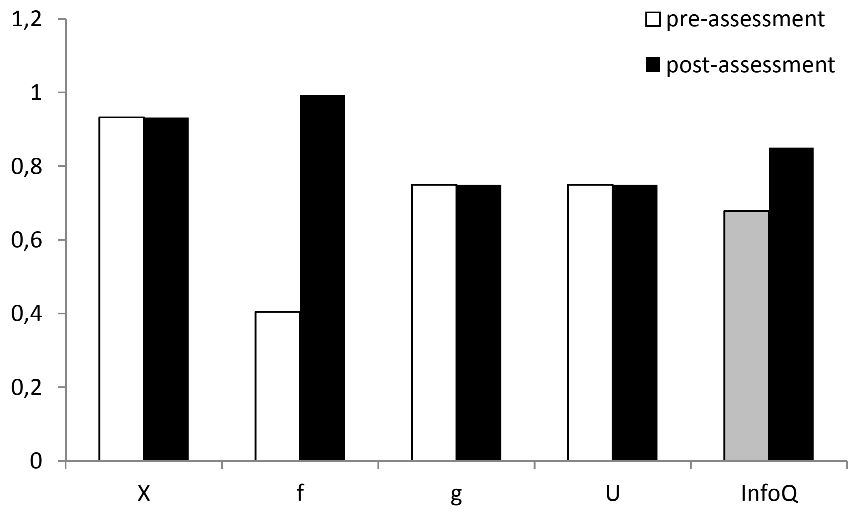

Figure 5

Tables

Table 1. Summary Table of the InfoQ-Dimensions Affecting the Four Components (X,f,g,U)

InfoQ-dimens.(↓) / InfoQ-compon.(→)	***X***	***f***	***g***	***U***
Data Resolution (D_1)	✓	✓		
Data Structure (D_2)	✓	✓		
Data Integration (D_3)	✓	✓		
Temporal Relevance (D_4)	✓	✓		
Chronology of Data and Goals (D_5)	✓	✓		
Generalizability (D_6)	✓	✓		✓
Operationalisation (D_7)	✓	✓		
Communication (D_8)		✓	✓	