Journal of Universal Computer Science, vol. 25, no. 8 (2019), 967-987
submitted: 26/12/18, accepted: 25/6/19, appeared: 28/9/19

The Role of Verification and Validation Techniques within Visual Analytics

Benjamin Weyers
(University of Trier, Germany
weyers@uni-trier.de)

Ekaterina Auer
(University of Applied Sciences Wismar, Germany
ekatarina.auer@hs-wismar.de)

Wolfram Luther
(University of Duisburg-Essen, Germany
wofram.luther@uni-due.de)

Abstract: We suggest to widen the focus of the scientific computations community from an isolated consideration of reliable numerical algorithms using standardized arithmetic to a broad user-centered system modeling and simulation approach relying on an appropriate verification and validation (V&V) design. Most V&V works rarely consider human-related issues specifically. However, modern applications generate and employ huge amounts of heterogeneous data and usually exhibit high complexity – challenges that are best tackled by augmenting human reasoning with automated techniques. That is, novel visual and collaborative approaches are needed to interpret the results, which has to be accounted for in the general V&V procedure. This should include an assessment of (meta-) data and code/outcome quality, selection of methods to propagate and bound uncertainty and, lastly, formally rigorous validation efforts. We present an approach to reliable visual analytics (i.e., analytics subjected to this V&V assessment), which can in turn contribute to the overall V&V procedure after that. Two use cases illustrate the potential of the introduced framework for reliable visual analytics.

Keywords: Verification and validation assessment, accurate modeling and simulation systems, reliable collaborative visual analytics, data quality assessment
Categories: D.2.4, H.1.2, H.5.1, I.3.6, I.6.4

1 Introduction

With the advance of ubiquitous computing, sensor-based systems with various interaction modes (*responsive systems* RS) gain more and more importance for supporting mobile users in all areas of their daily lives, helping them to interact and collaborate in a natural way. For example, people are supported in their surroundings by ambient assisted living systems [Munstermann 12] that rely on information gathered and used for inference by corresponding hardware devices or their networks. Nowadays, such environments deal with a wide spectrum of tasks, from those at peoples' homes (e.g., from the areas of elderly care or general healthcare), to those pertaining to commerce and business, up to museums and other leisure activities as well as to group decision-making [Sadri 11]. Larger systems (e.g., in smart city applications) employ technologies like Internet of Things that depend on reliable cloud computing methods to guarantee high performance, data integrity, privacy, network security and accuracy.

Apart from the obvious societal concerns about privacy loss, this implies a number of technical difficulties which must be overcome for RS to fulfill the requirements. To handle *large amounts of heterogeneous data* in these systems, techniques in the visual analytics (VA) domain offer indispensable tools allowing developers and users to exploit the information encoded in the input data and the system outcome. Additionally, novel concepts are needed for *collaborative human computer interaction* based on new task (meta-)models for post-WIMP (windows, icons, menus, pointer) interfaces, for solving arising user awareness, security and privacy issues or for assessing user satisfaction.

The RSs we focus on require a high level of reliability so that it is essential to apply a modern *verification and validation* (V&V) assessment to the process of their development. This includes appropriate user interaction and recommending services based on criteria which are adaptable to the addressed task, supplemented by a careful evaluation of these activities. The term validation concerns the process of determining the degree to which a computer-based model is an accurate representation of the real world and is appropriate for its purpose. Validation can be carried out using special metrics that help to establish the similarity degree to real life or to compare reconstructed objects and their behavior with the real world instances. Verification is necessary to ensure that the implementation of the developed model of a system or a process and its result are correct. If human-centered visualization environments are to be used as tools for developing reliable responsive systems and interfaces, they need a specialized evaluation methodology and V&V support [Kerren 07].

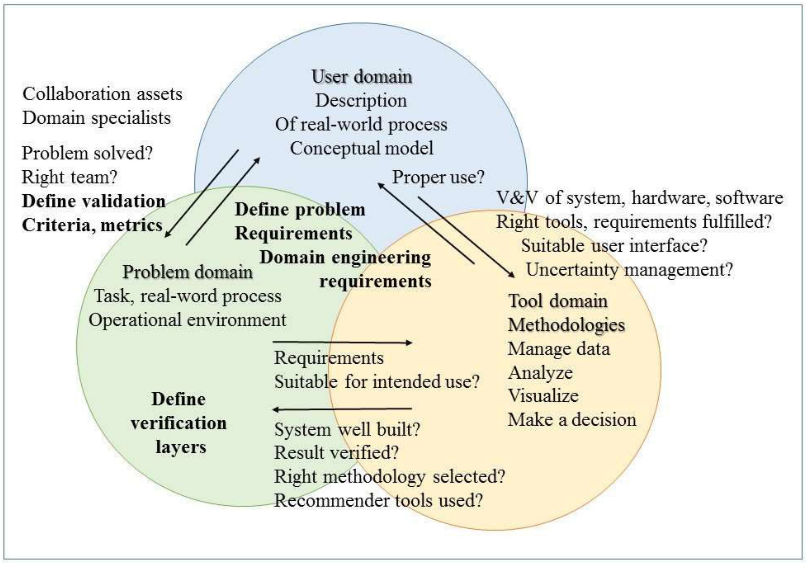

Figure 1: Stakeholder domains and key terms from the Introduction, displayed in a processing context based on a graphic from [Bingue 14] extended to user-centered aspects

Current V&V management does not yet completely reflect the requirements of modern computer-based systems (e.g., RS) and their interfaces, see Section 2. In particular, since these highly complex applications generate and employ large amounts of different kinds of data, it becomes necessary to augment automated techniques with human reasoning (and vice versa), which in turn calls for changes in the general V&V procedure. That is, new research questions have to be aimed at, on the one hand, verifying, validating, and evaluating novel visual and collaborative approaches employed in such systems. On the other hand, once their reliability is ensured, such approaches can be used inside the overall V&V scheme.

In this work we focus on studying VA in the above context. Along the way, we define *reliability dimensions* (such as visual integrity or quality of user interfaces and interactions) together with their specific *quality criteria* and *metrics*. Special attention is given to the criteria Accuracy, Adequacy, and Efficiency. Further important aspects are recommendation of scientific tools that best meet the required quality criteria, specification and calibration of corresponding measures as well as guidelines for the evaluation of reliable VA environments. In accordance with these goals and based on various community evaluation efforts for multi-purpose VA frameworks [Isenberg 12], we introduce an enhanced V&V approach, which is the major contribution of this paper. The proposed approach can provide the needed level of V&V that would allow developers to use visual analytics as a reliable tool for RS development. Since our general topic is very broad, some of the issues (e.g., evaluation) need to be covered in separate publications. For example, in [Auer 19], we summarize and complement the results of the biennial BELIV workshop (https://beliv-workshop.github.io) focusing on the challenges of evaluation in visualization as well as those of VA Science and Technology Challenge (2006-2014) offering a choice of metrics along with their appropriate implementations, supplemented by datasets and evaluators [Scholtz 14].

The presented V&V procedure is shaped by the *IEEE Standard for System, Software, and Hardware Verification and Validation* [IEEE 17] that specifies how to assess systems and tasks using *quality criteria* and *metrics*. This includes analysis, evaluation, testing, and inspection of software, hardware, and their interfaces. It has a much wider focus than the approach to V&V management proposed by the authors in 2009 and refined in [Auer 14]. There, a four-tier numerical V&V taxonomy is suggested with the focus on result verification.

Human issues and usability analysis are addressed by the mentioned IEEE V&V standard only peripherally: “Verify that stakeholder needs and interests are considered during development, operation, and maintenance process activities. The analysis will assure that: human-centered design activities are performed; human factors and ergonomics considerations are incorporated into the design, potential adverse effects on human health and safety are addressed in the design; and user needs are satisfied in a manner that supports user effectiveness and efficiency” [IEEE 17, p. 216]. In Figure 1, the key concepts from the standard and their interconnections are illustrated. We include Tool Domain Methodologies because of our focus on visual analytics. However, to build trust in the outcome of modern responsive systems, the general V&V principles need to be complemented by a careful, well-specified and standardized study of human factors such as interactivity, collaboration, and visual analytics, which we attempt in this paper.

First, we describe the state-of-the-art approaches to general V&V assessment (Section 2). After that, we aim our attention at reliability dimensions, quality criteria and their measures in the area of visual analytics including various levels of interactivity and collaboration in Section 3. In Section 4, we give a brief overview of our work in the field of V&V assessment. Two use cases are presented in more detail: reliable VA in computational neuroscience for the visualization of the activity of simulated neurons over time and in steel quality rating helping to visually assess data on thousands of steel samples. This article is an updated and extended version of the contributions [Auer 18, Weyers 18] to the CODASSCA meeting on Collaborative Technologies and Data Science in Smart City Applications in Yerevan, Armenia, September 12–15, 2018.

2 Approaches to V&V Assessment

In this section, we describe modern standards and approaches to V&V assessment to point out their strengths and weaknesses. One of the most important recent documents on V&V is the already mentioned IEEE standard [IEEE 17]. It states that "the V&V effort should analyze the artifacts (e.g., plans, models, and architecture) of the domain engineering as part of the required V&V tasks. Significant analysis of the domain engineering products should occur during system requirements review, software requirements evaluation, interface analysis, software design evaluation, source code and source code documentation evaluation, and all test planning." [IEEE 17, p.202]. This standard is synthesized from the following series of new or revised standards for system and software quality:

- IEEE 12207-2017, which establishes a common framework for "defining, controlling, and improving" software life cycle processes.
- IEEE 730-2014, which establishes requirements pertaining to quality assurance processes.
- ISO 9001:2015, which is a globally approved standard for quality management.
- ISO/IEC 25000 series, which is a framework for the evaluation of software product quality.

The standard also defines "minimum software and hardware V&V tasks to address system issues. These tasks include hazard analysis, security analysis, risk analysis, migration assessment, and retirement assessment" [IEEE 17, p 11]. It doesn't mention the IEEE standard 1788-2015 for Interval Arithmetic [IEEE 15] which specifies basic interval arithmetic operations with the help of one of the most widely used mathematical interval models. As shown in [Auer 14], intervals complying with IEEE 1788-2015 are efficient for modeling bounded uncertainty and propagating the associated error bounds. IEEE 1788-2015 integrates the IEEE 754-2008 floating point format necessary for interval computations.

The next important standard is NASA-STD-7009, the permanent NASA Standard for Models and Simulations (M&S) issued by the NASA Chief Engineer M. G. Ryschkewitsch on July 11, 2008 and revised in 2016 by M. J. Steele et al. [NASA-STD]. In [Blattnig 13], an instructive summary is published. According to it, a primary goal is to "ensure that the credibility of the results from M&S is properly conveyed to those making critical decisions". This goal translates directly into assessment of the risk that arises from the use of the M&S. Risk means "the probability a program or project will experience an undesired event; and the consequences, impact, or severity of the undesired event if it were to occur" [de Weck 15]. All of the standard's 49 requirements

(39 in the revised version) concern programmatics, development and use of an application. Twelve of them deal directly with verification, validation and uncertainty quantification. The introduced credibility assessment scale includes five level definitions to assess the M&S results and key factors for the rigor of the processes used to produce them. These level definitions contain nested requirements concerning the key factors *verification, validation, pedigree of input and other data, uncertainty of results, robustness of results, use history, M&S management,* and *people qualifications*. Data pedigree is meant as "a record of traceability from the data's source through all aspects of its transmission, storage, and processing to its final form used in the development of an M&S". For working with quality criteria, evidence meaning relevance for a given factor or sub-factor must be given. In this case we can evaluate for quality criteria to what extent they are met on a specified level. Level 0 is characterized by insufficient evidence for all factors. To qualify for a higher level, a process must meet the criteria for all the lower levels.

An important contribution of NASA-STD-7009 is the M&S criticality assessment matrix (cf. Figure 2). The cells of the matrix contain the risk type and the corresponding score for the likelihood of its occurrence. Risk types can have technical or programmatic sources. Examples are cost overruns, schedule slippages, safety mishaps, health problems, malicious activities, negative environmental impacts, or failure to achieve a needed success criterion [de Weck 15]. A color scheme rates the risks according to green: irrelevant, yellow: at project discretion, red: in the focus. Alternatively, a numeric risk consequence scoring can be given. Columns describe bad decision consequences (with failure effects from negligible to catastrophic). The rows show how M&S results influence process/project engineering decisions, from negligible to near certainty. For each category, there is a threshold level recommendation to avoid certain types of risk, for example, project/mission failure, further costs, or technical (component) failure. For reporting these "M&S credibility assessment and sufficiency thresholds" the bar graph and the spider (or radar) plot were proposed. This standard, like the previously described IEEE standard, addresses human factors only insufficiently. In this paper, we are interested in the factors interactivity, collaboration and visual analytics. In the following, we describe the newest developments in these areas.

The state of the art in building reliable interactive systems has been reported for the first time in [Wise 93] and recently in [Weyers 17]. Technological systems today are highly dependent on the interaction between humans and machines. The mentioned books present systematic research on how this interaction can be described, formalized, analyzed, and technically supported and how the reliability can be assessed. Interfaces and interaction have to meet various requirements with regard to human performance (or limitations) and their safety and fault tolerance.

Near certainty					
Significant					
Moderate					
Minor					
Negligible					
Relevance of M&S outcome	Negligible	Minor	Moderate	Critical	Catastrophic

Figure 2: Risk matrix "Scorecard" for project/process quality factors and descriptors

In Figure 3, we augment the traditional modeling/simulation and V&V cycle in engineering [Schlesinger 79] with features based on human-centered paradigms. In particular, such an additional feature is the use of advanced data and visual analytics for sense-making in big data environments. As a prerequisite for their employment inside the V&V scheme, they need to be assessed with regard to their quality, their data provenance and characteristics, and their safety, security and privacy. Cai and Zhu [Cai 15] describe a dynamic assessment process for big data based on a two-level quality standard. At the first level, there are five important assessment dimensions: availability, usability, reliability, relevance, and presentation quality. At the second level, reliability consists of accuracy, integrity, consistency, completeness and auditability. The last term means that it is possible for auditors to evaluate accuracy and integrity of data fairly. [Cai 15] is complemented by [Blytt 13], where the author emphasizes the frequently referred three Vs (volume, velocity, variety) as the biggest barrier to sense-making of big data. Veracity, validity, volatility and value are mentioned as further assessment factors. The value is postulated to comprise data discovery, integration and exploitation. Moreover, advantages and challenges of VA methods in the context of big data are described, which is also important for system evaluation. Finally, "interaction in the context of VA" and "the need for collaborative analysis" are emphasized, which highlights further human factors.

Obviously, not only (big) data, but also VA needs to be assessed. In [Keim 08], VA is described as combining automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision-making on the basis of a very large and complex data set. Although VA methods can be employed efficiently for a variety of applications, they sometimes exhibit a lack of attention to safety or reliability issues. However, accurate understanding of, for example, complex health data is necessary to make informed decisions about treatment of critically ill patients. For this reason, *reliable* VA methods should be introduced in decision-making sessions of various stakeholders. Only few publications explicitly address reliable VA, the main topics of discussion being data conversion to standard formats for visualization on various devices, accurate understanding of outcome using reliable mapping algorithms and standardized procedures to automatically select, analyze, refine and combine visual data. This makes developing clear guidelines for ensuring the reliability of VA an important research direction.

A further feature added to the cycle in Figure 3 is the option of collaborative outcome analytics done by various stakeholders with multiple expertise to cover the topics of system evaluation, effective data mining and problem solving as well as to organize follow-up actions. An example of using collaborative problem solving in combination with VA is given in [Jeong 15]. Here, developers design and evaluate a complete VA system and a collaborative touch-table application with two integrated components: a single-user desktop and an extended system suitable for a collaborative environment.

The final enhancement in Figure 3 is the possibility of evaluation, which aims at computing key indicators and parameters characterizing a computer-based system model based on existing standards and rules, requirements, and individual needs. Evaluation determines the extent to which goals are achieved and quality features are met, which somewhat overlaps with verification and validation.

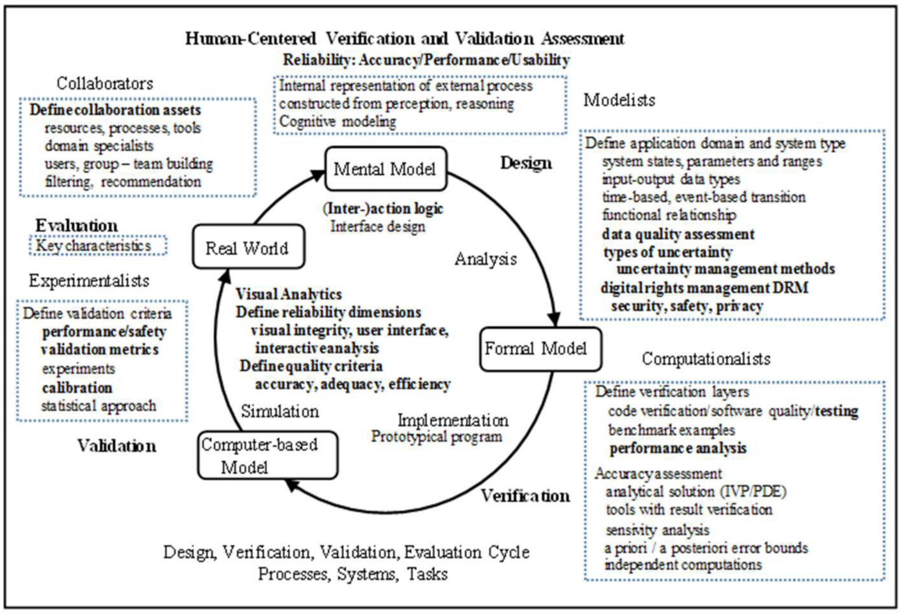

Figure 3: V&V – old and new assessment features. Features of special interest are pointed out in bold.

While verification characterizes meeting requirements from a formal point of view, evaluation allows us to draw conclusions concerning less rigorously specified aspects such as feasibility, practicability or acceptability. It can be understood as a review and validation of existing requirements and quality criteria. Similarly to data analytics and VA features, evaluations including performance, efficiency and safety measures can be used for V&V purposes if they comply with the requirements of underlying V&V management [Wise 93]. Evaluation can be considered from the vantage points of stakeholders, designers, software engineers, the general public and researchers involved in the development and use of the model. In [Framework 99] and [Al-Hajj 17], examples of how such evaluations can be conducted in health sciences are given. In [Framework 99], the quality of evaluation activities is assessed using the following four groups of standards: utility, feasibility, propriety and accuracy. In [Al-Hajj 17], it is shown how group analytics can be used to evaluate VA problem-solving and to support multi-stakeholder decision-making sessions in the context of child injury prevention. Here, typical questions concern group building, filtering visual data due to space, access time or relevance, collection of stakeholders' statements, information fusion and reporting. Collected data include stakeholders' observations, audio and video recordings, questionnaires, and follow-up interviews. The group analytics sessions are analyzed using the joint action theory protocol analysis and pair analytics methods that prove the emergence of 'common ground', that is, mutual, common, or joint knowledge, beliefs and assumptions among stakeholders. This is a precondition to solving problems collaboratively using VA.

3 Visual Analytics - Visualization and Interaction Techniques

As a typical multidisciplinary research area, which integrates knowledge from data sciences, computer graphics and cognitive science, VA fosters constructive process model validation, revision and improvement. Moreover, it facilitates data processing and result evaluation by combining human and machine intelligence during interaction and collaboration. In this section, we describe key steps of a VA process as well as VA reliability dimensions, quality criteria and measures.

3.1 Key Steps of a VA Process

The key steps of the VA process are data representation and transformation, visual layout and mapping, model-based analysis, and interaction/collaboration options [Sun 13]. Additionally, modeling, propagation and visualization of uncertainty at each step play a crucial role in ensuring the reliability and trustworthiness of VA. Below, we point out in short important issues at each step.

For the data representation and mapping process, a user-controlled selection of data types and reduction of data dimensionality are essential as is shown in [Weyers 16, Potter 12].

At the next step, that is, actual visualization, it is important to properly take into account sources of uncertainty, even for crisp data. To visualize certain and uncertain scientific or information ensemble data, [Brodlie 12] introduces the E- resp. U-notation which uses a subscript to indicate the number of independent variables or parameters and a superscript to indicate the type of dependent variables, e.g., S for a scalar. Uncertainty is visualized as a geometric form, such as a bar, a rectangle or a thick surface, a truncated upper and lower PDF or the interval mean and the standard deviation of a PDF. Additionally, color maps, glyphs or isosurfaces with color depending on U_{xyz}^{S} can be used.

At the final step of the result interpretation and general sense-making, a modern VA system architecture and an appropriate task model should enable users to choose optimal devices and interaction for their current task or collaboration style. Whereas WIMP interfaces utilizing mouse and keyboard-based interaction on screens are well suited for presenting text and 2D content, post-WIMP interfaces introduce new interaction paradigms for users to navigate, manipulate and interact with objects within a 3D virtual reality environment (e.g., with virtual hands and arms). Interaction styles are related to application scenarios including virtual and augmented reality, ubiquitous and context-aware computing and computer-supported cooperative learning or work. Extended post-WIMP task models (i.e., collaborative VA or highly interactive systems like CAVEs [Weyers 16]) have to include profiles depending on the application type, adequacy of interaction elements, flexibility in partitioning the task among multiple actors.

3.2 Assessment of Reliability in VA Applications

VA tools can be seen as interactive computing systems necessarily including a human. [Norman 88] defines this human-in-the-loop paradigm through the human action cycle. A crucial part in Norman's work is his discussion of the *gulf of execution* and *the gulf of evaluation*. Both gulfs characterize a potential gap between the users' mental model (including their mental state and their knowledge) and the interactive system used in context of the users' task. The user needs to understand both the current system state and the available set of operations implemented in the system's internal functional core, the latter presented and accessible through the user interface. For example, users might not be able to achieve a certain goal because their mental model of the system's state is wrong due to the fact that the interpretation of the perceived system values presented by the user interface is wrong or incomplete (*gulf of evaluation*). An example for the *gulf of execution* is that a user wants the system to execute a certain operation but unfortunately presses the wrong button.

If these ideas are put into use for VA applications, the gulf of evaluation becomes critical because it leads to wrong interpretation and thus to erroneous data visualization, which is in turn inadmissible in the context of V&V. In the worst case, visualization breaks visual integrity as has been characterized by [Tufte 91] with the help of the so called *lie factor*. The lie factor relates the (normalized) size of a given effect shown in the visualization to the size of the same effect in data. To answer the question of visual integrity for mappings of data to more complex spatial geometries and structures, Tufte's definition can be applied by specifying the size of a complex object. For instance, the size of a convex polygon with vertices $(x_1,y_1),\ldots,(x_n,y_n),(x_1,y_1)$ can be defined by its area, which grows proportionally to a^2 if the vertices are scaled with the factor a. To avoid or minimize also the gulf of execution, all rules concerning usability and user experience need to be applied to VA environments. That is, VA should be supported by formal methods not only on system and data level but also on the level of user interaction if it is to be employed in V&V context.

This paper relies on the concept of *reliable visual analytics* (RVA), originally introduced in [Weyers 18], with the focus on the human-in-the-loop and the type of visualization method, that is, on the interaction level. The general goal of the resulting *RVA framework* is to offer a set of quality criteria (QC) that can be assessed empirically. Users of various levels of experience can rate the subjective reliability of the inspected VA application with the help of predefined metrics. Additionally, formal descriptions and models of the application parts add an objective component to the reliability analysis because they can be used to evaluate the compliance of a tool with predefined requirements.

Reliability in the VA interaction context can be defined as the degree of accuracy of the user's mental model representing the visualized data. Along with the aspect of *visual integrity* (VI), reliability depends on the quality of the *user interface* (UI) and the *interactive analysis process* (IP) implemented by this user interface following the user's task. This specifies the potential VA *exploration space* (defined as set of all possible interactive analysis processes) implemented by the user interface and thus by the set of operations offered by the VA environment and its dialog model, similarly to [Weyers 14]. These three reliability *dimensions* (VI, UI and IP) are embedded into the context of the user's task, experience, goals and organizational environment. Gathering subjective estimations of the QC for solving a representative analysis problem with a VA tool helps to map this information to the reliability dimensions and embed it into the working context and organization [Weyers 18], thus evaluating the tool's reliability from the point of view of interaction.

We identify the following three QC: accuracy (AC), adequacy (AD), and efficiency (EF). Their semantics depends on the interpretation in the context of each individual dimension VI, UI, IP. **AC** is the potential for error prevention the VA environment offers.

Examples of issues AC addresses in the context of each reliability dimension are correctness of the user's interpretation (VI), usability of the user interface considering the analysis task (UI), or robustness (IP). That is, VA environments should provide all necessary operations at the right moment while users execute an operation sequence so that interaction errors are prevented. **AD** is defined as the level of suitability of the VA environment for the general analysis question. Applied to the reliability dimensions, AD addresses the quality of the mapping of data to their visual representation (VI) or the suitability of provided operations (e.g., as a number) of the user interface (UI) considering the user's task (IP). **EF** is defined as the performance achieved using the VA environment for the general analysis question. EF addresses the ease with which the visualization can be perceived (VI), usage simplicity of the user interface (UI), and the speed with which an analysis workflow can be executed (IP).

3.3 Measures for Reliable Visual Analytics

A first draft of the set of measures with the help of which QC are evaluated in the context of VI, UI and IP is shown in Figure 4 and has been initially presented in [Weyers 18]. The measures can be computed using the previously outlined empirical approach for a given VA application v. Additionally, a formal approach complementing the empirical one can be chosen for assessment. A quality function q and the associated algorithm A are defined in dependence on the visualization tool v, several descriptors, for example, pertaining to data, users, and tasks as well as some side conditions. As proposed in [Behrisch 18], the algorithm A has to solve a multi-objective optimization problem in order to find a visualization instance v to maximize (minimize) q. The choice of parameters to optimize and their ranges depends on the task definition, requirements, valid standards or measurements/experiments for validation. This choice can be made by trial and error, searching or ranking, and may follow iterative processes similarly to software engineering.

The goal of both the empirical and formal approaches can be formulated as finding an optimal tool v for the task. In this case, automatic recommendation ranking the suitability of v might be a different technique to achieve it if an effective and efficient implementation A of the target function q and its computation can be provided. The function q is characterized by QCs and QMs defined in the context of the user group and its profile, the task and its model, the data, metadata and data types mapped via a tool v belonging to a predefined set of computer-based visualizations, the hardware and its interfaces. The QCs encompass performance of the task completion including effectiveness and efficiency, reliability criteria for the data that need to be mapped by v to the visual space accurately and efficiently as well as fidelity and usability of user interfaces, graphical presentation and interaction styles with the focus on perception, navigation and manipulation of the graphical objects. If the optimization problem can be solved automatically with the help of A, then the visualization v and its QM parameters fulfill the requirements and can be recommended to the user (group) with a certain ranking for drawing conclusions and making decisions. [Behrisch 18] analyzes and categorizes more than 300 references which address quality metrics and data categories established in the information visualization (for multi- and high-dimensional, relational, sequential, geospatial and text data) and requirements on quality criteria.

To describe measures from Figure 4, we use the abbreviations [RD|QC]QM, where *RD* is the reliability dimension, *QC* the quality criteria and *QM* the addressed measure.

[VI|AC] Accuracy regarding visual integrity is a combination of two measures: correctness as a constraint and mapping as a sufficient condition. Both (classes) of measures are characterized by potential algorithmic evaluation using either formal model checking or the comparison of rendered visualizations with a pre-defined reference.

Correctness: This measure addresses the degree of correctness of the visual representation of values in the considered data set (a constraint). For instance, representing data with value 4 using the visual representation for 5 means low correctness. This might be measured by an automatic comparison of the rendered visualization (pixel by pixel, e.g. by calculating the distance between pixel's colors) with a set of reference images, (i.e., correct visual representations). Moreover, correctness should reflect whether uncertainty in the data is visualized in such a way that users are able to interpret it correctly in the context of both the data point and the whole data set. The correctness of interpretation may be measured automatically or within an empirical user study challenging/urging users to identify certain (known) features in a given visualization and compare the results with the expected outcome.

Mapping: The quality of the mapping of a data set w.r.t. its visual representation can be measured (automatically) via a function preserving the object descriptors, for example, as defined by Tufte's lie factor (the sufficient condition).

[VI|AD] Adequacy regarding visual integrity considers the task the user addresses with the help of the given interactive visualization. It is characterized by mapping and layout, which both need to be measured in accordance with the user's task.

Mapping: This measure addresses the mapping of data values to their visual representations in a sense similar to [VI|AC]Mapping but with a focus on the user's task. A model-based description of the task might be necessary to evaluate for a giving interactive visualization whether the visual representation is helpful for answering a certain (sub-)question. These models need to define what mapping for a (sub-)task in the analysis is adequate or required. For instance, the VA application's engineer can check in a testing session, in which he tries out all operations offered by the user interface, whether all operations exist and can be executed in reasonable time.

Layout: This measure characterizes adequacy of the design of a visualization for a given task with a specific focus on the layout of the visualization. Thus, this measure needs to consider the task and might involve the analyst's workflow. Not only the task itself, but also specific dependencies between tasks and sub-tasks need to be assessed w.r.t. adequacy with which their relevant visual representations are provided. This measure can be assessed similarly to [V|AD]Mapping in the same test session.

[VI|EF] Visual integrity influences the efficiency that characterizes the potential of success in answering a research question using the visualization under consideration. We identified two major measures to assess efficiency: readability and layout.

Readability: This measure characterizes the absence of visual clutter. Additionally, the complexity due to unnecessary amount of represented information is considered (e.g., complex plots rendered in limited screen space).

Layout: Similarly to [VI|AD]Layout, this measure addresses the spatial organization of the elements of the visualization, also taking the task to be solved into account. Additionally, a visualization should be structured in such a way that the relevant content for the task is easily recognized. Similarly to [VI|AD], both measures may be assessed using an expert walk-through.

[UI|AC] Accuracy of the user interface focuses on the basic requirements addressing general usability. The three identified measures are understood as the bare minimum for an interactive VA tool. A general guideline is the fulfillment of ISO 9241-11, the compliance with which can be measured in an empirical user study with such instruments as the SUS questionnaire [Brooke 96].

Quality Criteria / Reliability	Accuracy [AC]	Adequacy [AD]	Efficiency [EF]
Visual Integrity [VI]	- Correctness - Mapping	- Mapping - Layout	- Readability - Layout
User Interface [UI]	- Readability - Intuitiveness - Ergonomics	- Perceivability - Usability	- Usability - UX
Interaction Process/Dialog [IP]	- Intuitiveness - Learnability - Robustness	- Complexity - Structure - Mapping	- Seq. Length - Usability - UX

Figure 4: A measure matrix to characterize the reliability of VA by means of three reliability dimensions and quality criteria

Readability: This measure addresses the quality of the user interface in terms of clutter. All operations and interaction elements needed for a task or research question to be answered should be well visible and easy to find. Thus, this measure is very similar to that for visual integrity but in this case with the focus on the user interface. A potential operationalization of this measure is a performance test, where better performance (that is, better results the user is able to gather) indicates better readability.

Intuitiveness: This measure characterizes the level of the intuitiveness of the user interface, for example, how easy it is to identify the right operation/widget for a specific subtask. For quantification, this measure needs well designed user studies considering the addressed task(s) and user group(s). A potential scale is the SUS questionnaire.

Ergonomics: Similarly to intuitiveness, this measure characterizes the quality of the user interface design for a given task, for instance, ease of access to widgets or operations. Quality of the used input and output devices (e.g., how heavy a hand-held input device or how intrusive a display is) can be also addressed. In both cases, user studies need to be carried out using certain scales (such as SUS).

[UI|AD] Adequacy regarding the user interface concentrates on higher level qualitative measures such as perceivability and usability with an additional focus on the considered task.

Perceivability: This measures how simple it is to find a certain widget or element the user is looking for depending on the task. It needs to be characterized in a user study very specifically focusing on a task or a class of tasks. Compared to [UI|AC]Ergonomics, this measure characterizes how well a certain widget can be perceived in reference to the point of view of task performance. Here, error prevention plays a lesser role, although both can be measured by a task-dependent performance test involving the user.

Usability: This measure reflects compliance with ISO 9241-11 on usability with the focus on the task.

[UI|EF] Here, compliance with ISO 9241-11 is assumed as a prerequisite. That means that a user interface can only make work efficient if usability is high. A high level of user experience enhances interaction further.

Usability: Fulfillment of ISO 9241-11 with the focus on efficiency of completing a certain task, measured with similar tools as discussed above.

UserExperience: Fulfillment of user experience requirements and measures. For instance, the UEQ questionnaire [Laugwitz 08] can be used to measure user experience for a given user interface, user and task.

[IP|AC] The interaction process refers to a dialog model of a user interface. The dialog model describes what operations offered by the user interface (through widgets) can be executed and under what conditions. The previously described measures consider global qualitative characteristics of a user interface, such as usability. The dimension IP considers instead the execution of specific (task dependent) operation sequences, which might differ from task or sub-task sequence due to their atomicity. Thus, operations cannot be split further but task and sub-tasks are still more abstract and need to be mapped to operations. In the context of accuracy, the following measure can be applied.

Intuitiveness: Similarly to [UI|AC]Intuitiveness, intuitiveness in the context of [IP|AC] refers to the level of simplicity needed to find and execute not only an operation but a sequence or complex combination of (correct) operations in the context of the addressed analysis task. Intuitiveness can be measured in a user study in which the user has to execute a complex analysis processes and rate the system accordingly. An operationalization can take place using such performance measures as number of errors or quality of outcome/result of the applied process.

Learnability: This measure characterizes the effort needed to learn to work effectively with the VA tool focusing on the addressed analysis process. Thus, this measure is related to [IP|AC]Intuitiveness as high intuitiveness consequently reduces the needed learning effort. A classic operationalization of learnability is to apply pre- and post-knowledge tests before and after using the VA application. Questionnaires might help to estimate the perceived learning effort.

Robustness: A high level of intuitiveness and learnability might result in a high level of robustness of the tool when wrong input or similar errors happen. For instance, if the user changes the camera position inappropriately, the tool should offer the possibility to undo this change. That is, robustness also refers to basic usability measures and may be included into a user study for measuring the usability level of a given VA tool within an interaction process.

[IP|AD] Adequacy in the context of interaction processes addresses (in general) how adequate the interactive analysis tools support a given analysis workflow or process.

Complexity reflects the relation of complexity of the potential interaction sequences to the complexity required by the needs of the user's task and the user's experience. Too high complexity might lead to frustration and errors, too low complexity to a lower level of concentration resulting in missing operations and not completing the task. This measure can be quantified in an user study covering everyday work situations and can also be accompanied by formal methods to compare interaction sequences (from task models) with an interaction operation space, for example, extracted from a formal process-based description (e.g., Petri nets) [Weyers 17].

Structure: This measure is similar to [IP|AD]Complexity, but addresses the overall structure of the interaction space defined by the tool's interaction process/dialog. It can be operationalized similarly to measures described for [IP|AD]Complexity.

Mapping: This measure quantifies how adequately a VA tool maps a user's task to analysis processes. [IP|AD]Mapping is closely related to [IP|AD]Complexity and Structure since both can be affected by a given mapping. For instance, a user interface offering a wizard-like guidance through the analysis process reduces the complexity, simplifies the structure but simultaneously restricts the UI to only one or a set of processes, which might or might not be a good mapping to the user's analysis task. Along with a subjective characterization via a user study, a formal description of both process and user interface can be employed (as it is the case for [IP|AC]).

[IP|EF] Compared with [IP|AC] and [IP|AD], [IP|EF] is characterized mainly by subjective measures from user surveys accompanied by further empirical methods.

SequenceLength: This measure characterizes the minimum length of the needed number of operations to finish a given task/interaction process. A specific measure could be the needed number of mouse clicks or the number and complexity of gestures (characterized, e.g., in time – see also Fitt's law [MacKenzie 92]).

Usability: Similar to [UI|AC]Usability, but additionally addressing the analysis process. Usability should be measured empirically (as said above) in a user study. For assessing the interaction process, the study needs also to consider more complex analysis processes. The similar is true for user experience described below.

UserExperience: Equivalent to [UI|AC] but addressing the analysis process. Similarly to [IP|EF]Usability, this measure should be assessed in an empirical user study using standardized questionnaires (e.g., UEQ).

In summary, the reliability assessment outlined above relies on a three step engineering process: First, the relevant dimension and quality criteria need to be selected for a given analysis tool and task. Based on this selection, the developed user interface and analysis tool need to be investigated using a (partially) formalized interaction model by the engineers in a second step. Finally, in a third step, user studies need to be planned and conducted assessing the mentioned measures from a subjective, empirical point of view. Eventually, the gathered insight should be used in an iterative development process, for example, identified flaws in the design of a VA tool improved. The adaptions can be further investigated in a next iteration of V&V assessment as outlined above.

The presented framework is a first step towards a generalized set of qualitative and quantitative measures and criteria. Generalization, extension and validation of the presented measures is a challenging topic of (our) ongoing investigations. Nevertheless, the presented framework introduces an overarching strategy for addressing V&V for VA not only from the point of view of formal methods but also including empirical measures with the human in the loop as major recipient of the generated visualizations.

4 Use Cases

In recent years, the three authors and their collaborators have contributed reliable software tools and frameworks in various application areas, for example, analysis of (bio-)mechanical systems, solid oxide fuel cell systems [Auer 14] and steel technologies [Thurau 14, 14A]. We demonstrated how to enhance V&V assessment with the reasoning capabilities of VA through interactive visual interfaces, adding a dimension of collaboration to creation and evaluation of complex and responsive systems. We implemented virtual museums and laboratories, which showed, in particular, how involving people in collaborative and crowdsourcing activities could transform museums/labs into participatory spaces [Biella 16, Sacher 17].

Code verification and model validation techniques convinced in further projects: Traffic simulation systems using MAUDE [Brügmann 14, MAUDE]; a framework for development, assessment and interoperable use of verified methods with applications in distance computation, global optimization, and comparison systematics [Kiel 14]; a human autonomy assessment system [Munstermann 12] as well as uncertainty quantification via verified stochastic methods in geographic information system (GIS) applications [Rebner 15]. Moreover, a hardware and software architecture to assess and visualize important risk parameters and environment entities under uncertainty was described in [Weyers 16] with the aim to process, format and present relevant everyday threats from several risk classes that appear in specific contexts with typical parameters in a virtual house of risk. The layered architecture approach provided risk data and metadata in a virtual reality environment allowing for appropriate 2D and 3D visualization and various interaction styles, cooperation and co-creation of new content, and an individual degree of immersion into situations where threats typically happened.

To make modeling and assessment center on humans, a reference nets approach for modeling and reconfiguring user interfaces and their (inter)action logic can be employed as shown in [Weyers 12]. It is described how two student groups working collaboratively and on their own constructed their specific roles (Alice, Bob, Trusted Third Party, etc.) in a standardized cryptographic protocol step by step. The system automatically generated a reference net (a special type of the colored Petri net), which was matched against the existing protocol logic and supported by the robust implementation framework.

The examples mentioned above constitute relevant parts of the new human-centered V&V assessment introduced in Section 3.3 which includes not only traditional code and result verification, uncertainty management, validation and evaluation, but also evaluated user interaction, recommender techniques and reliable VA. In the following subsections, we will highlight two use cases, which address relevant parts of an enhanced V&V assessment.

4.1 Visual Analysis in Computational Neuroscience (CN)

The goal in CN is to apply simulation methods to investigate the dependency between structure and function of neural networks with the ultimate goal to understand the human brain. Mathematical models of neurons and data structures describing the connectivity of a neural network generate various types of data as output of a simulation run, such as spike events of neurons or membrane potentials. In the CN community, various types of simulators are being developed, for example, Neuron [Hines 97], TVB [Jirsa 10] or NEST [Gewaltig 07]. These simulators address various levels of abstraction, either focusing on very detailed neuron models (NEURON), the simulation of simplified point neurons (NEST), or the investigation of models on the level of neural areas (TVB). Here, visual analysis plays a central role in the investigation of simulation outcome [Senk 18]. In most cases, visualization is used in an interactive fashion in very early steps of a scientific workflow, for example, to validate the network's characteristics and behavior. Therefore, various visualization tools have been developed, such as VisNEST [Nowke 13], an interactive VR-ready tool visualizing multi-area models simulated with NEST. In the following, we characterize the so called "control view" (as shown in Figure 5, rendered in the aixCAVE [aixCAVE 18]) of VisNEST using the measures from Section 3.3. The view is composed of three different visualizations presenting activity data that results from the simulation of the visual cortex of a Macaque monkey. The visualizations are (a) a dot plot showing the activity of each neuron over time (also called spike plot), (b) a function plot showing the accumulated activity in an area per time step and over the whole simulation time, and (c) the accumulated activity per area mapped to a 3D rendered representation of the visual cortex' areas (right in Fig. 5). The measures from Section 3.3 are exemplified with a focus on the 3D rendering (c).

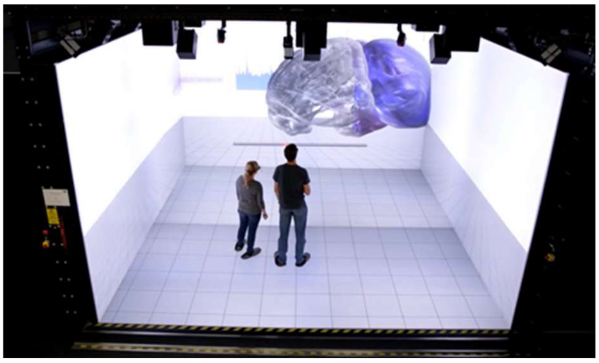

Figure 5: A VA tool for the investigation of NEST-based simulation of neural networks

[VI|AC]: For the quantification of correctness and mapping, the transfer function that maps the activity value to a color needs to be checked formally. In VisNEST, it is necessary that a certain distance between two values is perceived similarly to the perceived distance between the colors chosen for representing these values. This requirement is fulfilled by the use of the color map NASA MRO ice_freq (blue to red to yellow) [NASA GISS].

[VI|AD] The design is very simplistic relating to the measures Mapping and Layout. It uses the underlying geometrical shape of the simulated brain areas. The geometries are gathered from a scalable brain atlas, which offers anatomically correct mappings of experimentally (valid) data of the represented brain to geometrical models [Bakker 12].

[VI/UI/IP|EF] concerns efficiency along the three reliability dimensions. VisNEST has been investigated empirically by various experts who also took into account their workflows to investigate the addressed data without VisNEST. In [Nowke 13], the original authors present their findings, which were overall positive. The tool has been developed in very close collaboration with the later users (the domain scientists), which implies validation of the tool by design and through the design process. This is a benefit gained by the *user centric design* methodology, which ensures the creation of a successful user interface design that takes care of [UI|AC/AD] as a by-product. Additionally, the authors focus on the design of the visualization and tools not only in reference to the user's requests but also with respect to the original workflow described in [Nowke 13]. This makes it possible to address relevant measures for [IP|AC/AD] too.

Although a user centric design approach ensured the success for VisNEST, not all aspects from Section 3.3 have been assessed for the tool, in particular, the empirical measures usability and user experience have not been considered. This kind of characterization needs well designed user studies with a statistically relevant number of participants. Note that since VisNEST is a tool for a limited circle of experts, it is challenging to acquire enough participants with a good understanding of both the analysis workflow and of the simulation for a meaningful user study.

4.2 Analysis of Steel Samples SILENOS©

In this subsection, we illustrate how to visualize big data influenced by uncertainty, an aspect pertaining to data quality assessment, uncertainty management, performance analysis, and reliability of VA, using the steel inclusion processing framework, IPFViewer. It has been designed by the first author of [Thurau 14,14A] in cooperation with a large German steel production facility to analyze collected data about nonmetallic inclusions and other defects in steel samples using the steel inclusion level evaluation by the numerical optical system SILENOS© patent-protected by Hüttenwerke Krupp Mannesmann GmbH. The tool can analyze the ensemble data set in various ways, for example, perform outlier detection to identify samples and defects that differ from others by position, size, type and number. To rate steel quality, it can carry out trend analysis to study the influence of different process parameters on the steel samples and their meta-data and variance analysis to examine natural fluctuations within the samples and desired variations that result from process parameters.

After the three-dimensional shape of inclusions and other defects is reconstructed, the nonmetallic inclusions are classified into globular defects, crack-like defects and artifacts near the border of the sample volume [Buck 16]. If a chemical analysis is desired, it is possible to ask the system for a map with interesting inclusions. This allows the engineers to analyze the spatial distribution of inclusions and defects in general, classify them according to their three-dimensional shape and visualize the items using real-time rendering methods. The development of two subsequent IPFViewer versions involved a dozen university researchers and employees of the industrial partner. An expert survey concerning usability [UI|AD], performance and visualization concepts is ongoing work. In the following, IPFViewer is described using some of the measures from Section 3.3.

[VI|AD] Based on a new data tree and visualization model for the analysis of hierarchical ensemble data sets, IPFViewer utilizes multiple view techniques, data grouping and aggregation, data mining, and a reference data visualization (Mapping). Due to the immense size of the ensemble data set, which can contain up to a hundred thousand steel samples and ten billion defects, the main challenge is to present the relevant data in a way that enables the users to perform the evaluation quickly.

[VI|AC] Uncertainty visualization is used to analyze variations in the data: Not only small natural fluctuations from the production process can be visualized but also larger variances due to variation in the process parameters (e.g., temperature in the smelting furnace). The tool has a standardized and customizable reporting functionality and can be used with ensemble data sets generated using other application tools as well.

[IP|EF] When required, IPFViewer relies on new incremental, approximate analysis techniques to ensure the responsiveness of the application while sufficient precision is guaranteed for queries with fast response times [Thurau 14A].

[IP|AD] The layout allows workers at the steel production facility to quickly and interactively analyze data with millions of data rows. The resulting data tree is visualized as a huge grid in a scrollable area. Each grid cell incorporates a multiple view system with such standard visualization techniques such as scatter plots, bar charts and trend graphs [VI|AD]. Steel experts examine the histogram about defect diameters and the largest found defects to evaluate a sample quickly without having to analyze each defect manually. They can also scroll through all the samples and compare them, create and save various layouts that visualize different aspects.

5 Conclusions

This paper introduced human-centered paradigms into a formal V&V assessment within a workflow for designing, modeling, and implementing various real life processes. We showed that numerical result verification was only one of the important issues. Big data quality assessment, accurate task and process modeling, user interaction modeling and visual analytics should also be carried out in a reliable way. For reliable VA, trust in its outcome should be established, which we proposed to do by following the (meta-)design principles of a human-centered V&V assessment and also in dependence on users' task models and interaction styles, since the possibility to work with the visualization interactively was an integral part of VA. As use cases, we described toolboxes supporting reliable VA for visualizing steel artifacts and activity in simulated biological neural networks V&V assessment and evaluation of these systems was conducted from various vantage points concerning design, code or numerical result verification, software testing, visualization and usability. The subject of VA evaluation, which is not covered in this paper but is highly interconnected with the research questions posed by verifying and validating VA environments, is the topic of a further paper [Auer 19].

Not every issue raised by the considered research area has been examined yet. For example, the introduced list of reliability dimensions relevant for analyzing and understanding visual artifacts is not complete. Nevertheless, the paper delivers a description of an extended V&V assessment, an insight into the term reliable visual analytics and its importance for result evaluation in engineering (possibly under uncertainty) as well as a point of reference for the implementation of the specified requirements (provided by our applications from two different areas of engineering).

References

[aixCAVE 18] http://www.itc.rwth-aachen.de/cms/IT-Center/Forschung-Projekte/Virtuelle-Realitaet/Infrastruktur/~fgqa/aixCAVE/?lidx=1

[Al-Hajj 17] Al-Hajj, S., Fisher, B., Smith, J., Pike, I.: Collaborative Visual Analytics: A Health Analytics Approach to Injury Prevention. Int. J. Environ. Res. Public Health, 14 (2017) 1056

[Arias 17] Arias-Hernandez, R., Kaastra, L. T.: Joint Action Theory and Pair Analytics: In-vivo Studies of Cognition and Social Interaction in Collaborative Visual Analytics. Proc. 33rd Annual Conf. of the Cognitive Science Society, Boston, MA, USA (2011) 3244–3249

[Auer 14] Auer, E.: Result verification and uncertainty management in engineering applications. Habilitation thesis, University of Duisburg-Essen, Dr. Hut (2014)

[Auer 18] Auer, E., Luther, W.: Towards Human-Centered Paradigms in Verification and Validation Assessments, in Hajian, A., Luther, W., Han Vinck, A. J. (Eds.): Collaborative Technologies and Data Science in Smart City Applications. Logos Verlag Berlin (2018) 68-81

[Auer 19] Auer, E., Luther, W., Weyers, B.: Reliable Visual Analytics, a Prerequisite for Outcome Assessment of Engineering Systems, submitted to Acta Cybernetica

[Bakker 12] Bakker R, Bezgin G: Scalable Brain Atlas (2012) https://scalablebrainatlas.incf.org/

[Behrisch 18] Behrisch, M. Blumenschein, M., Kim, N. W., Shao, L., El-Assady, M., Fuchs, J., Seebacher, D., Diehl, A., Brandes, U., Pfister, H.,Schreck, T, and Weiskopf, D., Keim, D. A.: Quality Metrics for Information Visualization. EuroVis 2018 R. S. Laramee, G. E. Marai and M. Sedlmair (editors) 37(3) (2018) 38p. - STAR – State of The Art Report

[Biella 16] Biella, D., Pilz, T., Sacher, D., Weyers, B., Luther, W., Baloian, N., and Schreck, T.: Crowdsourcing and Co-curation in Virtual Museums: A Practice-driven Approach. Journal of Universal Computer Science, 22 (10) (2016) 1277–1297

[Bingue 14] Bingue, W. P. and Cook, D. A.: A Practical Approach to Verification and Validation, 2014. http://conferences.computer.org/stc/2014/papers/5034a024.pdf

[Blattnig 13] Blattnig, R. Luckring, J. M.,. Morrison, J. H., Sylvester, A. J., Tripathi, R K., and Zang, Th. A.: "NASA Standard for Models and Simulations: Philosophy and Requirements Overview", Journal of Aircraft, Vol. 50, No. 1 (2013) 20-28 https://doi.org/10.2514/1.C000303

[Blytt 13] Blytt, M.: Big Challenges for Visual Analytics - Assisting Sensemaking of Big Data with Visual Analytics. NTNU (2013) 13p

[Brodlie 12] Brodlie, K., Allendes Osorio, R., and Lopes, A.: A Review of Uncertainty in Data Visualization: Expanding the Frontiers of Visual Analytics and Visualization. Springer, London (2012) 81-109

[Brooke 96] Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability Evaluation in Industry 189 (194) (1996) 4-7

[Brügmann 14] Brügmann, J., Schreckenberg, M., and Luther, W.: A verifiable simulation model for real- world microscopic traffic simulations, Simulation Modelling Practice and Theory 48C (Nov. 2014) 58–92

[Buck 16] Buck, Chr.: Computerbasierte Modellierung und Analyse der Gestalt und Verteilung von Partikeln zur Optimierung des Reinheitsgrades von Stahlwerkstoffen – Konzeption und Implementierung eines Rahmenwerks (in German). PhD thesis, University of Duisburg-Essen (2016), http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=42570

[Cai 15] Cai, L., Zhu, Y.: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14 (2) (2015) 1-10

[de Weck 15] Olivier de Weck.O.: 16.842 Fundamentals of Systems Engineering. Fall 2015. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA

[Framework 99] Framework for Program Evaluation in Public Health, MMWR Recommendations and Reports, September 17, 1999 / 48(RR11) 1-40 https://www.cdc.gov/mmwr/preview/mmwrhtml/rr4811a1.htm

[Gewaltig 07] Gewaltig, M. O., & Diesmann, M.: Nest (neural simulation tool). Scholarpedia, 2(4) (2007) 1430

[Hines 97] Hines, M. L., & Carnevale, N. T.: The NEURON simulation environment. Neural computation, 9(6) (1997) 1179-1209

[IEEE 15] IEEE Standard for IA: http://standards.ieee.org/findstds/standard/1788-2015.html

[IEEE 17] IEEE Std 1012™-2016, approved 28 Sept. 2017, IEEE, New York, USA (2017) 260p.

[Jeong 15] Jeong, D. H., Ji, S.-Y., Suma, E. A., Yu, B., and Chang, R.: Designing a collaborative visual analytics system to support users' continuous analytical processes, Human-centric Computing and Information Sciences, 5 (5) (2015), Springer DOI 10.1186/s13673-015-0023-4

[Jirsa 10] Jirsa, V., Sporns, O., Breakspear, M., Deco, G., & McIntosh, A. R.: Towards the virtual brain: network modeling of the intact and the damaged brain. Archives Italiennes de Biologie, 148(3) (2010) 189-205

[Keim 08] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J. and Melançon, G.: Visual Analytics: Definition, Process, and Challenges. In Information Visualization, A. Kerren et al. eds., Berlin, Springer (2008) 154-175

[Kerren 07] Kerren, A., Ebert, A. and Meyer J. (Eds.): Human-centered Visualization Environments, volume 4417 of LNCS Tutorial, Springer, 2007 Challenges in Evaluation of Collaborative Visualization Environments

[Kiel 14] Kiel, St.: UniVerMec - A Framework for Development, Assessment and Interoperability of Verified Technics. PhD thesis, University of Duisburg-Essen (2014)

[Isenberg 12] Isenberg, P., Bertini, E., Lam, H., Plaisant, C., and Carpendale, S.: Empirical studies in information visualization: Seven scenarios. IEEE Transactions on Visualization and Computer Graphics, 18 (2012) 1520-1536

[Laugwitz 08] Laugwitz, B., Held, T., & Schrepp, M.: Construction and evaluation of a user experience questionnaire. In Symposium of the Austrian HCI and Usability Engineering Group. Springer, Berlin Heidelberg (2008) 63-76

[MAUDE] Maude System: http://maude.cs.illinois.edu/w/index.php?title=The_Maude_System

[Munstermann 12] Munstermann, M., Stevens, T., Luther, W.: A Novel Human Autonomy Assessment System. Sensors 12 (2012) 7828-7854

[MacKenzie 92] MacKenzie, I.S.: Fitts' Law as a Research and Design Tool in Human-Computer Interaction, Human-Computer Interaction, 7:1 (1992) 91-139

[NASA GISS] https://www.giss.nasa.gov/tools/panoply/colorbars/

[NASA-STD] NASA-STD-7009(A): https://standards.nasa.gov/standard/nasa/nasa-std-7009

[Norman 88] Norman, D.: The Design of Everyday Things. New York, Doubleday/Currency Ed. (1988)

[Nowke 13] Nowke, C., Schmidt, M., van Albada, S.J., Eppler, J.M., Bakker, R., Diesrnann, M., Hentschel, B., & Kuhlen, T.: VisNEST—Interactive analysis of neural activity data. In IEEE Symposium on Biological Data Visualization (BioVis) IEEE (2013) 65-72.

[Rebner 15] Rebner, G., Sacher, D., Weyers, B., Luther, W.: Verified stochastic methods in geographic information system applications with uncertainty. Structural Safety, 52 (2015) 244-259

[Sacher 17] Sacher, D.: A generative approach to virtual museums using a new metadata format. A curators', visitors' and software engineers' perspective. PhD thesis, University of Duisburg-Essen. Logos, Berlin (2017) ISBN 978-3-8325-4627-4

[Sadri 11] Sadri, F.: Ambient Intelligence: A Survey. ACM Computing Surveys 43 (4) Article 36 (2011)

[Schlesinger 15] Schlesinger, S. et al.: Terminology for Model Credibility. Simulation, 32 (1979) 103-104

[Scholtz 14] Scholtz, J., Plaisant, C., Whiting, M. A., and Grinstein, G. G.: Evaluation of visual analytics environments: The road to the visual analytics science and technology challenge evaluation methodology. Information Visualization, 13 (4) (2014) 326-335

[Senk 18] Senk, J., Carde, C., Hagen, E., Kuhlen, T. W., Diesmann, M., & Weyers, B.: VIOLA-A multi-purpose and web-based visualization tool for neuronal-network simulation output. arXiv preprint arXiv:1803.10205 (2018)

[Sun 13] Sun, G. D., Wu, Y. C., Liang, R. H., Liu, S. X.: A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. Journal of Computer Science and Technology, 28 (5) (2013) 852–867

[Thurau 14] Thurau, M., Buck, Chr., and Luther, W.: IPFViewer: Incremental, approximate analysis of steel samples. Proc. SIGRAD (2014), Linköping University Electronic Press (2014) 1-8

[Thurau 14A] Thurau, M., Buck, Chr., Luther, W.: IPFViewer: A Visual Analysis System for Hierarchical Ensemble Data. Proc. 5th Int. Conf. IVAPP-VISIGRAPP Lisbon, Scitepress (2014) 259-266

[Weyers 12] Weyers, B., Luther, W., and Baloian, N.: Cooperative Reconfiguration of user interface models for learning cryptographic protocols. International Journal of Information Technology & Decision Making, 11 (6) (2012) 1127-1154

[Weyers 14] Weyers, B., Borisov, N., Luther, W.: Creation of Adaptive User Interfaces Through Reconfiguration of User Interface Models Using an Algorithmic Rule Generation Approach. International Journal on Advances in Intelligent Systems 7(1 & 2) (2014) 302-325

[Weyers 16] Weyers, B., Luther, W.: Risk Communication and Perception in Low- and High-Immersion Virtual Environments. APSSRA6, 28-30 May 2016, Shanghai, China, H.W. Huang, J. Li, J., Zhang & J.B. Chen (editors) (2016)

[Weyers 17] Weyers, B., Bowen, J., Dix, A., and Palanque, Ph.: The Handbook of Formal Methods in Human-Computer Interaction. Springer Publishing, (2017)

[Weyers 18] Weyers, B.: Toward Collaborative Virtual Environments for Reliable Visual Analytics, in Hajian, A., Luther, W., Han Vinck, A. J. (Eds.): Collaborative Technologies and Data Science in Smart City Applications. Logos Verlag Berlin (2018) 59-67

[Wise 93] Wise, J. A., Hopkin, V. D., Stager, P. (eds.): Verification and Validation of Complex Systems: Human Factors Issues. Springer New York (1993)