The Role of Explicit Knowledge: A Conceptual Model of Knowledge-Assisted Visual Analytics
Supplement Material to Formalize the Model

Paolo Federico,[*‡] Markus Wagner,[†‡] Alexander Rind,[†] Albert Amor-Amorós,[*] Wolfgang Aigner,[†] Silvia Miksch[*]

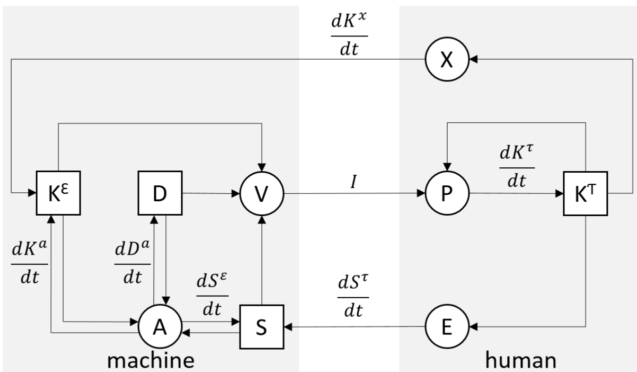

Figure 1: Conceptual Model of Knowledge-Assisted Visual Analytics. The model is divided into two spaces (machine and human) and describes knowledge generation, conversion, and exploitation within the Visual Analytics (VA) process, in terms of artifacts: explicit knowledge $\boxed{K^\varepsilon}$, data $\boxed{D}$, specification $\boxed{S}$, image I and tacit knowledge $\boxed{K^\tau}$; and processes: analysis A, visualization V, externalization X, perception/cognition P, and exploration E.

A The Mathematical Model Description

To further concretize our conceptual model we provide a formal, mathematical description as supplement material that substantiates the inner workings of the involved processes and components. At the same time, the formal description provides an additional perspective that focuses on the model's dynamics over time. The formal description itself is based on the notation used by Van Wijk [5], which are now extended to describe the novel 'Knowledge-assisted VA Model' from the mathematical point of view.

A.1 Definition of the Mathematical Elements

This section introduces the different elements used to describe the novel 'Knowledge-assisted VA Model' in combination with their definition and formal symbols (see Figure 1):

A **:= Automated Analysis:** Components used for automated data analysis based on different algorithms that can be used depending on the analysis problem.

D **:= Data:** Is used as the general term describing the two different types of data (D^r and D^a) which are included in the system.

D^r **:= Raw Data:** Specifies the raw input data of the system which are used as input for different automated analysis methods for example.

D^a **:= Pre-analyzed Data:** Refers to the output which is generated by one or more automated analysis methods (A).

E **:= Exploration:** This is based on the user's tacit knowledge K^τ to adjust the visualization V by the tacit specification S^τ.

I **:= Image:** Is the visual representation generated by the visualization V which is perceived P by the user.

K^τ **:= Tacit Knowledge:** Contains the users personal knowledge about the data and the insights gained during the perception P of the presented images I.

K^ε **:= Explicit Knowledge:** The computerized knowledge stored system internally, generated by the extraction X of the users tacit knowledge and automated analysis methods A.

K^x **:= Externalized Knowledge:** Specifies the externalized and computerized version of the users tacit knowledge K^τ.

K^a **:= Automated Analysis Knowledge:** Specifies the computerized knowledge which is generated by the use of one or more automated analysis methods A.

P **:= Perception:** The process how the user gains new insights to generate tacit knowledge K^τ.

S **:= Specification:** The combination of the specification S^τ based on the users exploration E by using tacit knowledge and the specification S^ε based on the explicit knowledge K^ε stored system internally.

S^τ **:= Specification by K^τ:** The specification part which is based on the exploration E of the users tacit knowledge K^τ.

S^ε **:= Specification by K^ε:** The specification part which is based on the explicit knowledge K^ε stored system internally.

t **:= Time:** Because data analysis is a interactive process, many components (e.g., K^τ, K^ε, S^τ, S^ε) are changing over time.

V **:= Visualization:** The process generating an image I from the data based on the specification which is affected by the users input and the system itself.

X **:= Externalization:** The process how the tacit knowledge K^τ is computerized to be stored system internally as externalized knowledge K^x.

A.2 Formalizing the Model

To provide the comparability to the model by Van Wijk [5], which is used as conceptual grounding for the new 'Knowledge-assisted VA Model', a formal description is needed. Based on this, the reader gets supported in understanding the differences between these two models as well as the functionality of the new elements. Additionally, for the formal description of the new model, we followed the mathematical notations provided by Van Wijk to not confuse the reader (e.g., sometimes an addition symbol (+) is used to describe a union or combination of two sets).

However, in the 'Simple Model of Visualization' by Van Wijk [5] the input data D are seen as static and cannot change over time. Thus, the time t is the only dynamic variable of these model, which describes the changes of the included processes over time. From a general perspective, a visualization system gets raw data D^r as input data that can be transformed or restructured into a pre-analyzed

*Paolo Federico, Albert Amor-Amorós, and Silvia Miksch are with TU Wien, Austria. E-mail: {federico, amor, miksch}@ifs.tuwien.ac.at

†Markus Wagner*, Alexander Rind, and Wolfgang Aigner are with St. Poelten University of Applied Sciences, Austria and TU Wien, Austria. E-mail: {markus.wagner, alexander.rind, wolfgang.aigner}@fhstp.ac.at

‡Paolo Federico and Markus Wagner equally contributed to this paper and are both to be regarded as first authors.

Publication forthcoming in Proc. IEEE Conf. Visual Analytics Science and Technology (VAST 2017)

dataset D^a by automated analysis methods A if needed. For example, if the input data are temperature data measured every minute, the analysis A step calculates the mean value for each hour, day, and month to remove a seasonal component of the cycle length. Therefore, the analysis step uses the explicit knowledge K^ε (see Equation 1) which is generated by a combination of the externalized tacit knowledge K^t of the user (K^x) and the knowledge generated by automated analysis methods A defined as K^a.

$$\frac{dD^a}{dt} = A(D^r, K^\varepsilon, t) \tag{1}$$

whereby the generation of the pre-analyzed dataset D^a follows an integration over time t (see Equation 3), assuming that D_0^a is the initial pre-analyzed dataset containing the same data like D^r so that the initial dataset can be marked as D (see Equation 2) before the first analysis is carried out (or especially if no analysis is performed). Additionally, a new D^a is created by the combination of the current D_n^a and the new calculated D_{n+1}^a (see Equation 4) (i.e., a cascade of automated analysis steps):

$$D = D^r = D_0^a \tag{2}$$

$$D_{n+1}^a = \int_0^t A(D^r, K^\varepsilon, t)dt \tag{3}$$

$$D^a(t) = D_n^a + D_{n+1}^a \tag{4}$$

As described by Van Wijk [5], in the model, the visualization can be seen as the central process. The dataset D^a will be transformed into a time depending image $I(t)$ based on the specification S (see Equation 5):

$$I(t) = V(D^a, S, t) \tag{5}$$

Furthermore, it is also possible to directly send the explicit knowledge K^ε into the visualization process, to make it explorable for the user (see Equation 6):

$$I(t) = V(K^\varepsilon, S, t) \tag{6}$$

However, depending on the analysts needs and the systems requirements, a combination of the data D^a and the explicit knowledge K^ε can also be performed by combining these two visualization processes (see Equation 7) if needed:

$$I(t) = V(D^a, S, t) + V(K^\varepsilon, S, t) \tag{7}$$

This image I will be perceived by the user's perception P which results as an increase of the users tacit knowledge K^τ (see Equation 8):

$$\frac{dK^\tau}{dt} = P(I, K^\tau, t) \tag{8}$$

The current tacit knowledge K^τ of the user follows an integration over the time t, assuming that K_0^τ is the initial tacit knowledge at the time point t_0 (see Equation9):

$$K^\tau(t) = K_0^\tau + \int_0^t P(I, K^\tau, t)dt \tag{9}$$

A further important aspect is the exploration E described as $E(K^\tau)$. The user decided to adapt the specification S^τ (tacit part) of the visualization V based on the users current tacit knowledge K^τ. This happens through further exploration E (see Equation 12):

$$\frac{dS^\tau}{dt} = E(K^\tau, t) \tag{10}$$

whereby the current tacit specification S^τ follows an integration over time t, judging from S_0^τ as initial specification for the tacit knowledge K^τ (see Equation 13):.

$$S^\tau(t) = S_0^\tau + \int_0^t E(K^\tau, t)dt \tag{11}$$

Based on the definition of knowledge K by Chen et al. [1, p. 13], we differ between knowledge which is generated by the externalization of the users tacit knowledge K^x and the knowledge which is generated by automated analysis methods K^a. The combination of these two knowledge parts (K^x and K^a) will be referred as explicit knowledge K^ε in this work. At this point it is important to note that automated analysis methods A which are integrated in a system, do not necessarily need to generate knowledge (K^a) that can be stored.

To retain (parts of) the users tacit knowledge K^τ for further analysis support, it can be externalized X (extraction) and stored as externalized knowledge K^x in a computerized form (see Equation 12) whereby the knowledge extraction was also covered by Wang et al. [6] in a similar way:

$$\frac{dK^x}{dt} = X(K^\tau, t) \tag{12}$$

The externalized knowledge K^x also follows an integration over time t, assuming that K_0^x is the initial externalized knowledge, which will increase by further externalization of the users tacit knowledge K^t (see Equation 13):

$$K^x(t) = K_0^x + \int_0^t X(K^\tau, t)dt \tag{13}$$

Additionally, to retain (parts of) the knowledge generation by automated computerized analysis methods operating on dataset D^a which is based on the specification S, can be stored as analysis knowledge K^a in a computerized form (see Equation 14):

$$\frac{dK^a}{dt} = A(D^a, S, t) \tag{14}$$

Thus, the analysis knowledge K^a also follows an integration over the time t, assuming that K_0^a is the initial automated analysis knowledge which can increase by further automated analysis of the dataset D^a , based on the specification S (see Equation 15):

$$K^a(t) = K_0^a + \int_0^t A(D^a, S, t)dt \tag{15}$$

As former mentioned, the explicit knowledge K^ε can be seen as the sum or more precisely, the combination of the externalized knowledge K^x (generated from the tacit knowledge K^τ) and the automated analysis knowledge K^a (generated by automated analysis methods A) (see Equation 16):

$$\frac{dK^\varepsilon}{dt} = \frac{dK^x}{dt} + \frac{dK^a}{dt} \tag{16}$$

whereby the explicit knowledge K^ε (composed from the user's externalized knowledge K^x and the automated analysis knowledge K^a) follows an integration over time t assuming $K_0^\varepsilon = K_0^x + K_0^a$ as initial explicit knowledge K^ε (see Equations 17, 18 and 19) whereby K_0^ε can also contain knowledge, which was integrated during the system development:

$$K^\varepsilon(t) = K_0^x + \int_0^t X(K^\tau, t)dt + K_0^a + \int_0^t A(D^a, S, t)dt \tag{17}$$

$$K_0^\varepsilon = K_0^x + K_0^a \tag{18}$$

$$K^{\varepsilon}(t) = K_0^{\varepsilon} + \int_0^t (X(K^{\tau},t) + A(D^a,S,t))dt \tag{19}$$

In order to achieve a knowledge support, the explicit knowledge K^{ε} (stored computerized knowledge) is used for exploration and analysis support of the dataset D^a. this also described by the 'Visual Analytics Mantra': "Analyze first, show the important, zoom, filter and analyze further, details on demand" by Keim at al. [2]. Thereby, the explicit specification component S^{ε} is produced (see Equation 20):

$$\frac{dS^{\varepsilon}}{dt} = A(D^a, S^{\varepsilon}, t) \tag{20}$$

Wherein the current explicit specification S^{ε} follows an integration over time t, when starting from S_0^{ε} as initial specification for share explicit knowledge K^{ε} (see Equation 21):

$$S^{\varepsilon}(t) = S_0^{\varepsilon} + \int_0^t A(D^a, K^{\varepsilon}, t)dt \tag{21}$$

In summary, the specification S can be seen as the combination of the tacit specification S^{τ} (depending on the tacit knowledge K^{τ}) and the explicit specification S^{ε} (depending on the explicit knowledge K^{ε}) (see Equation 22):

$$\frac{dS}{dt} = \frac{dS^{\tau}}{dt} + \frac{dS^{\varepsilon}}{dt} \tag{22}$$

whereby the specification S (composed from the tacit S^{τ} and explicit S^{ε} specification) follows an integration over time t assuming $S_0 = S_0^{\tau} + S_0^{\varepsilon}$ as initial specification for the combination of the tacit K^{τ} and explicit K^{ε} knowledge (see Equations 23, 24 and 25):

$$S(t) = S_0^{\tau} + \int_0^t E(K^{\tau},t)dt + S_0^{\varepsilon} + \int_0^t A(D^a, K^{\varepsilon}, t)dt \tag{23}$$

$$S_0 = S_0^{\tau} + S_0^{\varepsilon} \tag{24}$$

$$S(t) = S_0 + \int_0^t (E(K^{\tau},t) + A(D^a, K^{\varepsilon}, t))dt \tag{25}$$

Seen from an general perspective and extending the description by Van Wijk [5], visualization and the externalization of knowledge K (composed from tacit K^{τ} and explicit K^{ε} knowledge ($K = K^{\tau} + K^{\varepsilon}$) from the data D are objective processes in relation that the results do not depend on the person performing the analysis. Additionally, the analysis has to be repeatable by others and has to provide the same results under the same conditions [5]. However, visualization is not a well-defined process (always the same result relating to the same data). That means that the tacit knowledge K^{τ} does not change only based on the data D, it is also related to the specification S (e.g., given by hardware, parameter, algorithms and explicit knowledge K^{ε}), the perception P of the user and his or her tacit prior knowledge K_0^{τ} (see Equation 26):

$$\frac{dK}{dt} = P(V(D, E(K^{\tau},t) + A(D, K^{\varepsilon}, t), t)K^{\tau}, t) \tag{26}$$

A.3 Coverage of the Mathematical Description

As former described, we used the 'Simple Visualization Model' by Van Wijk [5] as conceptual grounding to generate the 'Knowledge-assisted VA Model' (see Figure 1). Therefore, we developed parts describing the externalization X of the users knowledge in a machine readable structure. Additionally, we also included a part describing the knowledge generation by automated pre-analysis A (described as "Analyze first" by Keim et al. [2]) in combination with the externalization of the users tacit knowledge $\boxed{K^{\tau}}$. It is important to note that the 'analyze first' criterion is only possible if one can apply automated analysis methods A to the dataset $\boxed{D}$ (especially to the dataset D^r to prepare a dataset D^a). This implies that knowledge-assisted visual analytics seemly requires a share of explicit knowledge $\boxed{K^{\varepsilon}}$ to be able to support and extend this preliminary analysis methods A. Thus, the explicit knowledge $\boxed{K^{\varepsilon}}$ is not necessarily worthless without corresponding data $\boxed{D}$ (e.g., a knowledge corresponding experiments) because also the explicit knowledge $\boxed{K^{\varepsilon}}$ alone can provide insights or helps to gain insights on corresponding datasets. On the contrary, it is important to note that it is not possible to fulfill the 'analyze first' step without automated analysis methods A which can be extended with 'integrate explicit knowledge' $\boxed{K^{\varepsilon}}$ to fulfill all the needs for a knowledge-assisted VA system.

$$\textbf{System Types} := \begin{cases} |K^{\varepsilon}| = 0,\ A = 0,\ V > 0 \Rightarrow \textbf{VIS} \\ |K^{\varepsilon}| > 0,\ A = 0,\ V > 0 \Rightarrow \textbf{KAV} \\ |K^{\varepsilon}| = 0,\ A > 0,\ V > 0 \Rightarrow \textbf{VA} \\ |K^{\varepsilon}| > 0,\ A > 0,\ V > 0 \Rightarrow \textbf{KAVA} \\ |K^{\varepsilon}| = 0,\ A > 0,\ V = 0 \Rightarrow \textbf{AM} \\ |K^{\varepsilon}| > 0,\ A > 0,\ V = 0 \Rightarrow \textbf{KAAM} \end{cases} \tag{27}$$

Keim et al. [3] declared that VA can be characterized along two problem classes: "(1) Analytical Problems and (2) General Application Areas of IT" [3]. To solve these, they pointed out to "three methodological classes: a) Automatic Analysis, b) Visualization, and c) Visual Analytics" [3]. Based on [3] in combination with the novel 'Knowledge-assisted VA Model', it is now possible to distinguish between 4 different system types including visualization and two system types without visualization described in Equation 27.

The first time you use a visualization without explicit knowledge $\boxed{K^{\varepsilon}}$ and automated analysis methods A, the 'Visual Information Seeking Mantra': "Overview first, zoom and filter, then details-on-demand" [4] comes in use. If, after this, the user integrates step by step his tacit knowledge $\boxed{K^{\tau}}$ and/or knowledge generated by the integration of automated methods A in a machine-readable way, explicit knowledge $\boxed{K^{\varepsilon}}$ is integrated in the system as support for exploration and insight gaining. Based on this, the related systems can be defined as Knowledge-assisted Visualization (KAV) or Knowledge-assisted Visual Analytics (KAVA) depending on the integration of automated analysis A or not. If the system supports preliminary data analysis by automated analysis methods A without the integration or storing of explicit knowledge $\boxed{K^{\varepsilon}}$, further analysis will follow the 'Visual Analytics Seeking Mantra' by Keim et al. [2] and is can be described as VA system.

Additionally, automatic analysis methods also benefits from the integration and use of explicit knowledge. Assuming that there are systems available without containing a visualization V, the model also allows to describe Automated Analysis Method (AM) systems and Knowledge-assisted Automated Analysis Method (KAAM) systems. These systems can be seen as subtype of VA and KAVA systems but without including a visual interface for data representation.

An additional interesting aspect is that Van Wijk [5] expects that the data D did not change over time t, it seems that he considers this appears as a static entity throughout exploration / visualization. Thus, during the data exploration, no new datasets can be added to the system. Based on this assumption, the model could now be expanded by the integration of dynamic datasets or data sources $D(t)$ (e.g., different types of (time-oriented) streaming data) in the future.

References

[1] M. Chen, D. Ebert, H. Hagen, R. Laramee, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and

knowledge in visualization. *IEEE Computer Graphics and Applications*, 29(1):12–19, Jan. 2009. doi: 10.1109/MCG.2009.6

[2] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, eds. *Mastering the information age: solving problems with visual analytics*. Eurographics Association, Goslar, 2010.

[3] D. A. Keim, F. Mansmann, and J. Thomas. Visual Analytics: How Much Visualization and How Much Analytics? *SIGKDD Explor. Newsl.*, 11(2):5–8, 2010. doi: 10.1145/1809400.1809403

[4] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages. Proceedings*, pp. 336–343, Sept. 1996. doi: 10.1109/VL.1996.545307

[5] J. J. van Wijk. The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pp. 79–86, 2005. doi: 10.1109/VISUAL.2005.1532781

[6] X. Wang, D. H. Jeong, W. Dou, S.-W. Lee, W. Ribarsky, and R. Chang. Defining and applying knowledge conversion processes to a visual analytics system. *Computers & Graphics*, 33(5):616–623, Oct. 2009. doi: 10.1016/j.cag.2009.06.004