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

Paolo Federico, Alexander Rind, Albert Amor-Amorós, Wolfgang Aigner, Silvia Miksch

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 (Df and Dp) which are included in the system.

K := Knowledge: Specifies the computerized version of the users tacit knowledge input and the system itself.

S := Specification: The combination of the specification Sf and Ss based on the specification Sf based on the explicit knowledge Kf stored system internally.

P := Perception: The process how the user gains new insights to generate tacit knowledge Kt.

E := Exploration: The specifcation part which is based on the exploration E by using tacit knowledge and the specification Ss based on the explicit knowledge Ks stored system internally.

t := Time: Because data analysis is an interactive process, many components (e.g., Kf, Ks, Sf, Ss) 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 Kt is computerized to be stored system internally as externalized knowledge Kx.

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 this model, which describes the changes of the included processes over time. From a general perspective, a visualization system gets raw data Df as input data that can be transformed or restructured into a pre-analyzed...
dataset $D^p$ 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 $K^e$ (see Equation 1) which is generated by a combination of the externalized tacit knowledge $K^x$ of the user ($K^x$) and the knowledge generated by automated analysis methods $A$ defined as $K^a$.

$$\frac{dD^p}{dt} = A(D^p, K^e, t) \tag{1}$$

whereby the generation of the pre-analyzed dataset $D^p$ follows an integration over time $t$ (see Equation 3), assuming that $D^p_0$ is the initial pre-analyzed dataset containing the same data like $D^p$ 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 $D^p$ is created by the combination of the current $D^p_n$ and the new calculated $D^p_{n+1}$ (see Equation 4) (i.e., a cascade of automated analysis steps):

$$D = D^p = D^p_0 \tag{2}$$

$$D^p_{n+1} = \int_0^t A(D^p, K^e, t)dt \tag{3}$$

$$D^p(t) = D^p_n + D^p_{n+1} \tag{4}$$

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

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

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

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

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

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

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

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

The current tacit knowledge $K^x$ of the user follows an integration over the time $t$, assuming that $K^x_0$ is the initial tacit knowledge at the time point $t_0$ (see Equation 9):

$$K^x(t) = K^x_0 + \int_0^t P(I, K^x, t)dt \tag{9}$$

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

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

whereby the current tacit specification $S^x$ follows an integration over time $t$, judging from $S^x_0$ as initial specification for the tacit knowledge $K^x$ (see Equation 13):

$$S^x(t) = S^x_0 + \int_0^t E(K^x, t)dt \tag{11}$$

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 $K^e$ 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 ($K^x$) that can be stored.

To retain (parts of) the users tacit knowledge $K^x$ 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:

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

The externalized knowledge $K^e$ also follows an integration over time $t$, assuming that $K^e_0$ is the initial externalized knowledge, which will increase by further externalization of the users tacit knowledge $K^x$ (see Equation 13):

$$K^e(t) = K^e_0 + \int_0^t X(K^x, t)dt \tag{13}$$

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

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

Thus, the analysis knowledge $K^a$ also follows an integration over the time $t$, assuming that $K^a_0$ is the initial automated analysis knowledge which can increase by further automated analysis of the dataset $D^p$, based on the specification $S$ (see Equation 15):

$$K^a(t) = K^a_0 + \int_0^t A(D^p, S, t)dt \tag{15}$$

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

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

whereby the explicit knowledge $K^e$ (composed from the user’s externalized knowledge $K^x$ and the automated analysis knowledge $K^a$) follows an integration over time $t$ assuming $K^e_0 = K^x_0 + K^a_0$ as initial explicit knowledge $K^e$ (see Equations 17, 18 and 19) whereby $K^x_0$ can also contain knowledge, which was integrated during the system development:

$$K^e(t) = K^e_0 + \int_0^t X(K^x, t)dt + K^a_0 + \int_0^t A(D^p, S, t)dt \tag{17}$$

$$K^x_0 = K^a_0 \tag{18}$$
\[ K^\varepsilon(t) = K^\varepsilon_0 + \int_0^t (X(K^\varepsilon, t) + A(D^\varepsilon, S, t))dt \] (19)

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

\[ \frac{dS^\varepsilon}{dt} = A(D^\varepsilon, S^\varepsilon, t) \] (20)

Wherein the current explicit specification \( S^\varepsilon \) follows an integration over time \( t \), when starting from \( S^\varepsilon_0 \) as initial specification for share explicit knowledge \( K^\varepsilon \) (see Equation 21):

\[ S^\varepsilon(t) = S^\varepsilon_0 + \int_0^t A(D^\varepsilon, K^\varepsilon, t)dt \] (21)

In summary, the specification \( S \) can be seen as the combination of the tacit specification \( S^T \) (depending on the tacit knowledge \( K^T \)) and the explicit specification \( S^\varepsilon \) (depending on the explicit knowledge \( K^\varepsilon \)) (see Equation 22):

\[ \frac{dS}{dt} = \frac{dS^T}{dt} + \frac{dS^\varepsilon}{dt} \] (22)

whereby the specification \( S \) (composed from the tacit \( S^T \) and explicit \( S^\varepsilon \) specification) follows an integration over time \( t \) assuming \( S_0 = S^\varepsilon_0 + S^T_0 \) as initial specification for the combination of the tacit \( K^T \) and explicit \( K^\varepsilon \) knowledge (see Equations 23, 24 and 25):

\[ S(t) = S^\varepsilon_0 + \int_0^t E(K^\varepsilon, t)dt + S^T_0 + \int_0^t A(D^\varepsilon, K^\varepsilon, t)dt \] (23)

\[ S_0 = S^\varepsilon_0 + S^T_0 \] (24)

\[ S(t) = S_0 + \int_0^t (E(K^\varepsilon, t) + A(D^\varepsilon, K^\varepsilon, t))dt \] (25)

Seen from an general perspective and extending the description by Van Wijk [5], visualization and the externalization of knowledge \( K \) (composed from tacit \( K^T \) and explicit \( K^\varepsilon \) knowledge \( K = K^T + K^\varepsilon \)) 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 \( K^T \) 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 \( K^\varepsilon \)), the perception \( P \) of the user and his or her tacit prior knowledge \( S^T_0 \) (see Equation 26):

\[ \frac{dK^T}{dt} = P(V(D, E(K^\varepsilon, t) + A(D^\varepsilon, K^\varepsilon, t), t)K^T, t) \] (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 \( \bar{X} \) of the users knowledge in a machine readable structure. Additionally, we also included a part describing the knowledge generation by automated pre-analysis \( \bar{A} \) (described as “Analyze first” by Keim et al. [2]) in combination with the externalization of the users tacit knowledge \( \bar{K} \). It is important to note that the ‘analyze first’ criterion is only possible if one can apply automated analysis methods \( \bar{A} \) to the dataset \( D^\varepsilon \) (especially to the dataset \( D^\varepsilon \) to prepare a dataset \( D^\varepsilon(t) \)). This implies that knowledge-assisted visual analytics seemingly requires a share of explicit knowledge \( \bar{K} \) to be able to support and extend this preliminary analysis methods \( \bar{A} \). Thus, the explicit knowledge \( \bar{K} \) is not necessarily worthless without corresponding data \( \bar{D} \) (e.g., a knowledge corresponding experiments) because also the explicit knowledge \( \bar{K} \) 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 \( \bar{A} \) which can be extended with ‘integrate explicit knowledge’ \( \bar{K} \) to fulfill all the needs for a knowledge-assisted VA system.

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 four 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 \( \bar{K} \) and automated analysis methods \( \bar{A} \), 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 \( \bar{K} \) and/or knowledge generated by the integration of automated methods \( \bar{A} \) in a machine-readable way, explicit knowledge \( \bar{K} \) 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 \( \bar{A} \) or not. If the system supports preliminary data analysis by automated analysis methods \( \bar{A} \) without the integration or storing of explicit knowledge \( \bar{K} \), 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 \( \bar{V} \), the model also allows to describe Automated Analysis Method (\( \bar{AM} \)) 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. Lameekee, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and


