THE HUMAN-CENTRED USER MODEL FOR ADAPTIVE DATA VISUALIZATIONS

D12

Report

IDEALVis Consortium

http://idealvis.inspirecenter.org/

European Union
European Regional Development Fund

Republic of Cyprus

Structural Funds
of the European Union in Cyprus

ΙΔΡΥΜΑ ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

Executive Summary

This deliverable aims to present the human-centred user model, one of the core components of IDEALVis. This human-centred user model is composed of four major dimensions: (i) the User characteristics, and the characteristics of the user's functioning context, such as the business environment, expressed in (ii) business tasks, (iii) data and (iv) visualizations. The central dimension (i.e., the user), considers human factors like perceptual preferences, cognitive capabilities in information processing, affective states, domain expertise, experience, etc.

To this end, this deliverable describes the selected human factors of the proposed user model, discussing related literature and arguing on the expected impact when end-users interact with business data visualizations. Ultimately, the goal of the user model is to serve as input to the IDEALVis adaptation engine to produce human-centred adaptive data visualizations that will facilitate explainable exploration and transparent analysis of complex and multivariate business processes and datasets and will support and enable more effective decision making on critical business tasks.

Our main objective is to provide an appropriate (i.e., adequately incorporating the main elements of each dimension) and flexible (i.e., extensible in the future with additional dimensions/characteristics) model that combines those human-centered characteristics together with the business contextual characteristics (e.g., role, purpose, requirements, tasks, business data and visualizations) to facilitate adaptive interventions and personalization conditions during the visual data exploration process. Additionally, this deliverable also studies the direct object of investigation, i.e., business tasks, data, and visualizations, that constitute the business environment i.e., the functioning context of a business data analyst user. In this respect, we present the results of a study with 59 business users (data analysts), to create a first understanding of the similarities and differences to current approaches and extract the requirements for adaptive and personalized interventions, and further support our choice of business tasks, data, and visualizations as important and distinct dimensions of the final user model.

Table of Contents

List of Figures

1 Introduction

Modern business intelligence and data analytics platforms use real time visual analytics to continuously monitor and analyze business transactions and historical data to facilitate real-time decision support. The result of the analysis is then exported into various standard format artifacts (e.g., tabular forms, graphs, etc.) offering customization options to end-users as means for (visual) data exploration for obtaining insights and unveiling complex patterns. However, according to IBM, every day we create 2.5 quintillion bytes of data – so much data that 90% of all the data in the world today has been created in the last two years alone (IBM, 2016). These data come from a variety of sources and in diverse formats, both structured and unstructured, creating a business ecosystem that brings new insights but also generates several complications and problems (e.g., delays in real-time processing, ineffective delivery of multi-purpose information). Consequently, this may disorient end-users that need to navigate and take decisions faster than ever when performing their daily business activities using data analytic solutions. Although such platforms may provide data visualizations that are considered more usable than others (Liu, et al., 2014), often their recipients (i.e., the decision makers) are overloaded from the vast amount of visual information, which in turn severely decreases their ability to efficiently assess situations and plan accordingly (Bonneau, et al., 2014) (Kinkeldey, et al., 2017). It is evident that current business data exploration and most of data visualizations are: (i) created based on task and / or data-driven models and methods; (ii) extracted based on data mining algorithms that do not consider any user needs and requirements or role-based specifications; and (iii) following an one-size-fits-all approach, presenting the same visualization type and content to all users irrespective of their needs, requirements, and unique characteristics.

In the context of the IDEALVis project, we argue that the complex nature of business processes, tasks, objectives, and many data visualizations make it indispensable to include human intelligence in the business data analysis and visualization process at an early stage. It is vital to enrich the current business analytic platforms with adaptation techniques and new possibilities for interactions that will bring the human-in-the-loop by considering the end-users' individual differences and their business context (e.g., role, purpose, requirements, tasks, business data and visualizations) in combination. With this in mind, the following sections demonstrate the human-centred user model (as the main component of the IDEALVis platform), that is composed of four main dimensions: User (in the centre), Tasks, Data and Visualizations (as the functioning context characteristics of the business data analyst user), and has the goal is to enable human-centred adaptive data visualizations that will facilitate explainable exploration and transparent analysis of complex and multivariate business processes and datasets, thus, enabling more effective decision making on critical business tasks.

This deliverable: (i) builds upon prior research on the impact of individual differences on data visualizations, for proposing an innovative human-centred user model in the business data analytics domain, and (ii) further studies the direct object of investigation, i.e., business tasks, data, and visualizations, that constitute the contextual frame of execution for a user. For that reason, we present the results of a study with 59 business users (data analysts), for (i) achieving a first understanding of the similarities and differences to current approaches, (ii) for extracting the requirements for adaptive and personalized interventions and finally for (iii) further supporting our choice of business tasks, data, and visualizations as important and distinct dimensions of the final user model.

2 Background Work and Motivation

Today, with the growing expectations of business end-users and the proliferation of heterogeneous business processes and datasets, traditional approaches for data interpretation and visualization often cannot keep pace with the continuous escalating demand, so there is the risk of delivering unsatisfactory and misleading results. Business data models and processes characterized by significant complexity, making the analysis, and understanding of data by managers, data analysts, business experts, etc., challenging, time consuming, if not many times impossible. In many cases, a single activity combines even custom-made developments (e.g., using Excel) for the subsequent execution of steps creating a dispersed, inconsistent, and error-prone reality. Hence, it is widely accepted that the increasingly large amount of data requires novel, seamless, transparent, and user-friendly solutions (Forrester, 2018). As such, handling, analysing, and gaining insights into these large multivariate processes and datasets through interactive visualizations is one of the major challenges of our days (Kerren, et al., 2008) (Kerren & Schreiber, 2012).

In recent years, many powerful computational and statistical tools have been developed by various organizations in the business sector, such as SAS Visual Analytics[1], IBM Analytics[2], Microsoft Power BI[3], SAP Busines Business Intelligence Platform[4], Tableau Business Intelligence and Analytics[5], Qlik Business Intelligence[6], etc., offering several solutions like interactive maps, charts, and infographics, visual business intelligence analysis, recommended actions, etc. These applications are currently designed to execute the same operations following a machine learning and artificial intelligence approach, which is solely based on data models and rigid tasks and objectives, and with power users (e.g., expert data analysts) in mind. They embrace the power of statistical methods to identify relevant patterns, typically without human intervention. Inevitably, the danger of modelling artifacts grows when end-user comprehension and control are not incorporated. To this end, although modern business intelligence and data analytics platforms offer vast repositories of data analysis tools and myriads of customizable visualizations; they have not kept up to the challenge when it comes to their dynamic adaptation and personalization depending on the role, experiences, intrinsic characteristics, or abilities of end-users and still follow a one-size-fits-all paradigm. This poses an issue as the effectiveness of a visualization in terms of usability and understanding differs amongst users (Liu, et al., 2014). The vast amount of visual uncertain information overwhelms the user’s perception, which in turn, severely decreases their ability to understand the data and make decisions (Bonneau, et al., 2014) (Kinkeldey, et al., 2017).

On the other hand, the joint benefits of adaptation and personalization, and data visualizations and exploration that consider specific human factors in the core of their user models have been

[1] https://www.sas.com
[2] https://www.ibm.com/analytics/us/en/technology/products/cognos-analytics/
[3] https://www.powerbi.microsoft.com/en-us/
[4] https://www.sap.com/products/bi-platform.html
[5] https://www.tableau.com
[6] https://www.qlik.com

highlighted repeatedly in a variety of fields and applications, mostly in academia. Indicatively, research works have identified noteworthy associations of users' cognitive abilities like perceptual speed, in relation to performance, accuracy, and satisfaction when interacting with alternative data visualization (Lallé, et al., 2017) (Toker, et al., 2012); others focus on optimizing data visualizations based on the users' goal, behaviour, cognitive load and skills (Steichen, et al., 2013) (Toker, et al., 2017); investigate how human factors like personality and working memory affect user performance when interacting with visualizations (Green & Fisher, 2010) (Carenini, et al., 2014); how individuals' cognitive styles, like Field Dependent-Independent, impact interactions with various information visualizations and in relation to individual aid choices and preferences (Steichen & Fu, 2019); or how effective are emotion-triggered (e.g., boredom and frustration) adaptation methods for visualization systems (Cernea, et al., 2013). Hence, although significant effects have been shown in domains like public facing applications, educational and navigation contents, or health datasets, these ideas have rarely been applied, to our knowledge, to the business sector despite the encouraging results of prior studies (Poetzsch, et al., 2020). IDEALVis aims to address this research gap by highlighting the effect of a multi-dimensional human-centred user model in data visualizations and analytic applications that facilitates the execution of specific end-to-end business scenarios and tasks. The overarching innovation lies upon (a) the generation of knowledge and theory, rules, adaptive interventions, personalization conditions and explanations triggered by the joint influence of cognitive and affective characteristics on business data visualizations and exploration, and (b) the development of computational techniques, tools and methods that will put the user model into practice considering the requirements, constraints, and policies of real-life business settings.

3 The Proposed Human-centred User Model

This complex nature of information visualizations necessitates the development of a comprehensive user model that captures important factors, such as users' cognitive characteristics, affect, domain expertise and experience, as well as understanding of the end-user roles, objectives, context, and the characteristics of the data (Poetzsch, et al., 2020) (Mutlu, et al., 2021). Our main purpose is to employ those human aspects that together with the business contextual characteristics (e.g., role, purpose, requirements, tasks, business data and visualizations) will be able to jointly facilitate more appropriate adaptive interventions, personalization conditions and explanations during the visual data exploration process. In this respect, we propose ta user model, consisting of four main dimensions: User, Tasks, Data and Visualizations, as illustrated in Figure 1.

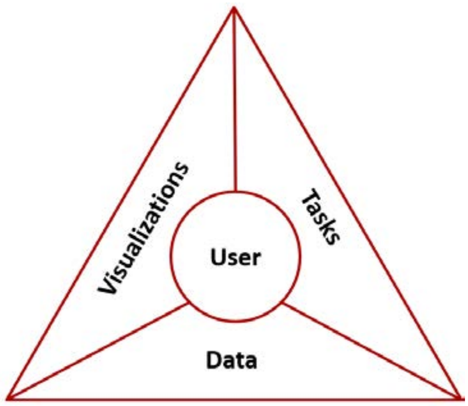

Figure 1 - Proposed Human-centred User Model

In the next sections, we provide details for each dimension, describing its internal characteristics and their interactions.

3.1 User Dimension

The business end-user is the focal point in the definition of the user model, referring on one hand to the understanding of the business roles, nature, and their contexts of functioning, and on the other hand to the identification of the intrinsic human factors that play the most significant role during their engagement with the data visualizations. Considering the various theories and models of individual differences in the literature, the following factors have been promoted as more applicable for the scope of this research work (in relation to specific business settings and actions): The **perceptual and cognitive processing characteristics**, are mainly distinguished in users': (a) high-level information processes, like cognitive styles (Witkin, et al., 1977) that have a direct impact on the user's ability to mentally extract shapes from their surroundings (i.e., field-dependent, -independent), or on the type (textual or imagery) of the content and may influence preferences and decision making in data visualization scenarios (Steichen & Fu, 2019), and (b) elementary cognitive processes (i.e., working memory, controlled attention and speed of processing), that have an effect on the complexity of the content regarding users' task performance, overall efficiency and cognitive control of visual information (Steichen, et al., 2013), or problem solving and comprehension during the interaction process. Regarding individual characteristics that affect the perception of visualizations, models that relate to graphical or visual (numeric) literacy or guidance (i.e., reading between and beyond data to understand abstract, data-driven associations - (Friel, et al., 2001)) have been qualified expecting that high levels of visual literacy will impact users' reasoning with visual representations making more elaborate inferences (extracting information from more complex visualizations) as opposed to those with low (Okan, et al., 2012). Furthermore, emphasis will be placed upon end-users' personality (Goldberg, 1990) (and Need for Cognition (Cacioppo & Petty, 1982)) as influential human traits of the perceptual process, motivation, and behaviour, expecting to affect users during the visual interaction process with respect to accuracy (including error rates), search and performance when executing tasks, problem-solving approaches, and skills (Green & Fisher, 2010) (Liu, et al., 2020).

The **affective processing** (or affective states) guides behaviour and emotions, as behavioural output of the process (Walla, 2018) and refers to a range of feelings that people experience, including discrete emotions, moods, and traits (such as positive and negative affectivity). It may be at some extent deduced into two basic constructs, i.e., Emotional Arousal and Emotion Regulation, influencing people's performance, judgement and decision-making process while interacting with data visualizations (Lekkas, et al., 2009). For example, users with a negative affective state require environmental enhancements to work more efficiently, as their emotional needs alter their behaviour and create different informational, and processing demands (Lekkas, et al., 2009).

The **domain expertise** indicates how skilful a user is in the domain (s)he functions, and it is associated with graph understanding, accuracy, and performance (time spent) in relation to visual tasks complexity (e.g., less experienced individuals may spend more time in information retrieval and comparison of sub-stages) (Dreyfus & Dreyfus, 1980) (Allen, 2000). Also, it affects preference, satisfaction, and the capability of being familiarized or switching between graphs to obtain information, e.g., novice users have greater difficulties of using different visualization types (Toker, et al., 2017).

The **business role characteristics** refer to more "traditional" persona elements defined from a person's or an entity's business responsibilities, objectives, and tasks. It may include personal, professional, or technical information (Usability.gov, 2021), competencies, expectations, needs, feelings, painpoints, usually associated to specific activities that are tightly linked to one (or more) business processes within an organization. This user model builds on the premise that data visualizations should be coupled with the goals and requirements of each business role and consider the variability of tasks, level of knowledge, constraints, etc., for conveying the adequate information, when and how it is needed, and on the expected breadth and depth that could facilitate fast and accurate decision making (Amyrotos, et al., 2021).

3.2 Tasks, Data and Visualizations

Partially, the **business tasks** formulate the context of execution (sequence of project-specific actions) and interaction for an end-user, relating to situation-specific scenarios, requirements and constraints depending on the line of business. Tasks may be regarded as a solid point of reference for designing usable interactive data visualizations, but they usually comply with business data models and processes characterized by increased complexity and criticality, making the analysis and understanding of information by various non-power user (e.g., data analysts, business analysts) challenging, time consuming, costly, if not many times impossible. Usually, in real-life business scenarios the daily responsibilities of a role presuppose the engagement with more than one tools (and workflows, roles) in combination (e.g., native applications) to assign some meaning to data and extract useful knowledge for decision making. Hence, explainable, and transparent data visualizations and exploration that will support end-to-end tasks execution and facilitate understanding of the flow, dependencies, and contents of multi-variate information (usually generated by different business processes and data models) is of paramount importance in the business domain.

Such information resonates in various **data** sources found in different locations, are of different types, have different data characteristics (e.g., real-time, historical), and are connected to complex (customer-specific) data models and business processes. Hence, efficient semantic mapping among features is critical, so that integrated data analysis is possible and comparable through intuitive data visualizations. As such, structured learning and graphical models like probabilistic dependency networks, probabilistic decision trees, Bayesian networks and Markov Random Fields, are becoming popular business data mining tools helping to deal with open case-based data challenges like scalability, uncertainty and data quality, dynamicity, heterogeneity, etc. (Chen, et al., 2012). Therefore, it is widely accepted that the increasingly large amount of data requires novel, efficient, and user-friendly data visualization solutions. The mechanisms that will be considered at this stage will provide high quality business knowledge that will also determine (based on their properties) the significance of the data objects and the yielded adaptive data visualizations. This dimension will make sure that data integration is possible, by means of intelligent pre-processing and fusion of data; to render data from different locations or in different types to be comparable, and to create mappings among features so that integrated data analysis will be possible. Several unaddressed issues will be tackled in supporting data analysis of business datasets especially through the use of visualization, such as (a) very large, i.e., scalability, (b) dynamic, i.e., addressing the velocity aspect within the V's of Big Data, and (c) heterogeneous, i.e., consisting of different data types both in terms of acquisition method and representation (Chen, et al., 2012).

As such, handling, analysing, and gaining insights into these large multivariate datasets through interactive and explainable **data visualizations** is one of the major challenges of our days and this work. Main goal is to specify the properties and structure of the content of data visualizations and exploration support. Subsequently, a further identification and characterization of parameters that will enable the adaptation based on the human-centred user model will take place. Currently, there are different types of visualizations (e.g., bar, column, line and area charts, radar graphs, plots, tables) which communicate information and meaning out of data, always in relation to the scope and the needs of a task. For example, line charts represent a connection of data points in a Cartesian coordinate system which generates a sequence of values used to view trends and cycles over a period; while the possibility to generate multiple lines makes them also suitable for comparing values. Visualizations that have some common and comparable features, a recognizable impact of individual differences on them, and apply at a large extend in the business domain (as seen at the results of **RQ2** in section 4.3) will be qualified. Once data visualizations are defined, they and their sub-optimal counterparts will constitute several subsequent objects which will be enriched with metadata (semantic augmentation) enabling the filtering process according to the human-centred user model and the data attributes and structure. Thereupon, adaptation and personalization techniques will be crafted to offer: (a) Dynamic alteration of the content presentation and hierarchical structure of data visualization attributes (e.g., re-ordering, salience, size, saturation, texture, colour, orientation, shape, etc.); (b) provision of various navigation tools and support (e.g., visual prompts, explanations) for data/ visual exploration during end-to-end business tasks execution; (c) variable amount of user control (e.g., allowing further (deeper) data exploration); and (d) additional assistive tools (e.g., data properties and details), etc.

Given the users' diversified requirements, needs and perceptual preferences as well as the size, diversity, and processing overhead of big business data sets, it is expected that the proposed human-centred user model will yield flexible best-fit data visualizations and methods that will support the unique end-users during the end-to-end interaction process.

3.3 *Formalised User Model*

According to the abovementioned theoretical user model this sub-section provides the reader with the formalized definition of the user model. For further aiding the reader's understanding of the formalized user model, relevant mathematical definitions are described accordingly for each user model factor. Essentially, IDEALVis maintains a set of data analyst users ($U = \{u_1, u_2, \dots u_n\}$) and their respective user models, accessible for each user (e.g., user u_i) via the $um(u_i)$ function which takes the user as input. More specifically, the user model is composed of items coming from 2 categories, the demographics (category = d) and the psychometric characteristics (category = p). Moreover, the user model for each user is represented by a set of triplets of the form (ct, ch, val), where ct represents the user model item category (i.e., one of demographics or psychometric characteristics), ch represents an actual characteristic that belongs to the triplet's category (e.g., if category is demographics, ch can be age), and val which represents the respective value for the triplet's characteristic (e.g., 35 for age). A partial version of a user model for participant u^i could for example be:

$$um(u^i) = \{(d, age, 35), (p, wm, low), (p, fdi, fd)\}$$

denoting that u^i has an age of 35, a low Working Memory and is classified as field-dependent with regards to Field-Dependent Independent cognitive style.

3.3.1 COGNITIVE CHARACTERISTICS

Assume that a user u_i performs a task ts_j that belongs to a given cognitive ability / cognitive style test cs. The system captures the user's response and stores it as a quintuplet represented as $ts_j^{cs}(u_i) = (cs, j, u_i, val, t)$, where cs represents the cognitive test, j represents the task's number, u^i represents the user, val represents the correctness of the provided response ($true$ or $false$), and t represents the time in milliseconds taken by the user to provide the response. In the context of IDEALVis cs can be s (i.e., Speed of Processing), c (i.e., Control of Attention), vwm (i.e., Visual Working Memory) or fdi (i.e., Field-Dependent, -Independent).

The set of all tasks answered correctly by the user u^i for a specific cognitive test cs is denoted as:

$$T^{cs}(u_i) = \{ts_j^{cs}(u_i): ts.val = true, \forall\, j\}$$

The number of tasks answered correctly by the user u^i for a specific cognitive test cs is defined as:

$$cr^{cs}(u_i) = |T^{cs}(u_i)|$$

The average response time for a cognitive test cs for a user u_i is defined as:

$$rt^{cs}(u_i) = \frac{\sum_{\forall ts_j^{cs}(u_i) \in T^{cs}(u_i)} ts_j^{cs}.t}{cr^{cs}(u_i)}$$

The average response time for a cognitive test cs for all users is defined as:

$$RT^{cs} = \frac{\sum_{\forall u_i} rt^{cs}(u_i)}{|\{u_i: rt^{cs}(u_i) > 0\}|}$$

The deviation of the average response times for a cognitive test cs (applicable only to s and c) for all users is defined as:

$$dv^{cs} = \frac{RT^{cs} \times 10}{100}$$

The upward deviation of average response times for a cognitive test cs (applicable only to s and c) for all users is defined as:

$$dvu^{cs} = RT^{cs} + dv^{cs}$$

The downward deviation of average response times for a cognitive test cs (applicable only to s and c) for all users is defined as:

$$dvd^{cs} = RT^{cs} - dv^{cs}$$

Speed of Processing and Control of Attention: Cognitive test s measures the speed of processing level of a user. A user u_i is either classified as having a high level of processing speed, assuming an average response time $rt^s(u_i)$ lower than dvd^s; classified as having a low level of processing speed, assuming an average response time $rt^s(u_i)$ higher than or equal to dvu^s; or classified as having a medium level of processing speed, assuming an average response time $rt^s(u_i)$ that falls between dvu^s and dvd^s. Similarly, cognitive test c measures the control of attention level of a user. A user u_i is either classified as having a high level of attention control, assuming an average response time $rt^c(u_i)$ lower than dvd^c; classified as having a low level of attention control, assuming an average response $rt^c(u_i)$ higher than or equal to dvu^c; or classified as having a medium level of attention control, assuming an average response of $rt^c(u_i)$ that falls between dvu^c and dvd^c.

Visual Working Memory: Cognitive test vwm measures the visual working memory of a user. This cognitive test consists of 21 tasks ts which are broken down to 7 levels (3 questions per level). Each time the user answers three questions correctly their working memory level raises by 1. When the user makes a mistake, the test is ended, and their working memory level is the current level when they answered wrong.

The level of a user u_i for cognitive test vwm is defined as:

$$vwml = \frac{cr^{vwm}(u^i)}{21}$$

Moreover, a user is classified as having a low visual working memory if they have a $vwml$ of 1 or 2, medium visual working memory if they have a $vwml$ of 3, 4 or 5 and a high visual working memory if they have a $vwml$ of 6 or 7.

Field-Dependent Independent: Cognitive Style Test fdi consists of 18 tasks ts that user u_i must complete. The final score of a user u_i is denoted as $cr^{fdi}(u_i)$ and the level of the user (i.e., field-dependent, -independent, or intermediate) is calculated with two percentile values derived from the set of all users' scores $\{cr^{fdi}(u_i), \forall i\}$.

For the fdi cognitive test, the 50th and 75th percentiles of all users' scores are defined below respectively as $Plow$ and $Pmid$.

$$Plow = \left\lceil \frac{50}{100} \times \left|\{cr^{fdi}(u_i), \forall i\}\right| \right\rceil$$

$$Pmid = \left\lceil \frac{75}{100} \times \left|\{cr^{fdi}(u_i), \forall i\}\right| \right\rceil$$

Accordingly, a user is classified as being field-dependent if they have an fdi score less than $Plow$, field-independent if they have a score higher or equal to $Pmid$, and intermediate if they have a score higher than or equal to $Plow$ and lower than $Pmid$.

3.3.2 HUMAN FACTOR LIKERT-SACLE QUESTIONNAIRES

Assume that a user u_i answers a Likert-scale question q_j that belongs to a questionnaire lq. The system captures the user's answer and stores it as a quadruplet represented as $q_j^{lq} = (lq, j, u_i, val)$, where lq represents the questionnaire, j represents the question's number, u^i represents the user, and val represents user's response in a numerical format (maximum and minimum values of this variable are defined by the underlying questionnaire scale). In the context of IDEALVis lq can be epq (i.e., Eysenck Personality Questionnaire), $dmsi$ (i.e., Decision-making Style Inventory), $pssq$ (i.e., Problem-solving Style Questionnaire), erq (i.e., Emotional Regulation Questionnaire) or pet (i.e., Perceived Data Analysis Expertise Tool).

The responses for a given Likert-scale questionnaire lq provided by user $u_i$are defined as:

$$Q^{lq}(u_i) = \{q_j^{lq}(u_i), \forall\, j\}$$

The sum of all responses for a given Likert-scale questionnaire lg provided by user u_i is defined as:

$$qs^{lq}(u_i) = \sum_{j \in Q^{lq}(u_i)} j.val$$

Eysenck Personality[7]: Questionnaire epq measures aspects of personality for a given user and consists of 48 yes or no questions (stored as 1s and 0s respectively). Specifically, this questionnaire is used to measure four distinct personality factors $epqf$ including extraversion, neuroticism, lie-scale, and psychoticism denoted as ex,nu,ls and ps respectively. From the total

[7]https://docs.google.com/document/d/155JQsvMnNfXLT9Nidi_BayWyQHYobiq3/edit?usp=sharing&ouid=114970685780107799249&rtpof=true&sd=true

of 48 questions, each factor $epqf$ has (i) a set of 12 related questions rq^{epqf} represented by their number j (see below),

$$rq^{ex} = \{ 3, 7, 11, 15, 19, 23, 32, 36, 44, 48,27, 41\}$$

$$rq^{nu} = \{1, 5, 9, 13, 17, 21, 25, 30, 34, 38, 42, 46\}$$

$$rq^{ls} = \{4, 16, 45,8, 12, 20, 24, 29, 33, 37, 40, 47\}$$

$$rq^{ex} = \{10, 14, 22, 31, 39,2, 6, 18, 26, 28, 35, 43\}$$

and (ii) a set cq^{epqf} that represents the correct responses for each q_j (see below):

$$cq^{ex} = \{1,1,1,1,1,1,1,1,1,1, 0, 0\}$$

$$cq^{nu} = \{1,1,1,1,1,1,1,1,1,1,1,1\}$$

$$cq^{ls} = \{0,0,0,1,1,1,1,1,1,1,1,1\}$$

$$cq^{ex} = \{1,1,1,1,1,0,0,0,0,0,0,0\}$$

A score for each personality factor $epqf$ is calculated for a user $u_i$according to the responses provided and the matching of those responses with cq^{epqf}. The scoring function for a given $epqf$ of user u_i is defined below:

$$SC^{epqf}(u_i) = \left|\{q_j^{epq}(u_i): q.val = cq_j^{epqf}, \forall\, j \in rq^{epqf}\}\right|$$

Given the score of a specific personality factor epqf the user can be further classified as having a low, medium, or high level for that epqf. Specifically, for any epqfusers with a score of 1 to 4 are classified as having a low level, 5 to 8 are classified as having a medium level and 9 to 12 are classified as having a high level.

Decision Making Style Inventory[8]: Questionnaire $dmsi$ measures decision making factors for a given user and consists of 25 Likert-scale questions (taking values from 1 to 5). Specifically, this questionnaire is used to measure four distinct decision-making factors $dmsif$ including rational, intuitive, dependent, avoidant, and spontaneous. From the total of 25 questions, each factor $dmsif$ has a set of 5 related questions rq^{dmsif} represented by their number j (see below),

$$rq^{rational} = \{ 1, 6, 11, 16, 21\}$$

$$rq^{intuitive} = \{ 2, 7, 12, 17, 22\}$$

$$rq^{dependent} = \{ 3, 8, 13, 18, 23\}$$

$$rq^{avoidant} = \{ 4, 9, 14, 19, 24\}$$

$$rq^{spontaneous} = \{ 5, 10, 15, 20, 25\}$$

A score for each decision-making factor $dmsif$ is calculated for a user u_i according to the sum of response values val provided for questions related to each decision-making factor $dmsif$. The lowest score a user can achieve for a factor in this questionnaire is 5 and the highest score is 25.

[8]https://docs.google.com/document/d/1MSCmNrfdzxisly09v-CBC9vgARMJfT5S/edit?usp=sharing&ouid=114970685780107799249&rtpof=true&sd=true

Specifically, for any $dmsif$ users with a score of 5 to 10 are classified as having a low level, 11 to 17 are classified as having a medium level and 18 to 25 are classified as having a high level.

Problem Solving Style[9]**:** Questionnaire $pssq$ measures problem solving factors for a given user and consists of 20 Likert-scale questions (taking values from 1 to 5). Specifically, this questionnaire is used to measure four distinct problem-solving factors $pssqf$ including sensing, intuitive, feeling and thinking. From the total of 20 questions, each factor $pssqf$ has a set of 5 related questions rq^{pssqf} represented by their number j (see below),

$$rq^{sensing} = \{4, 5, 10, 16, 19\}$$
$$rq^{intuitive} = \{3, 8, 11, 13, 18\}$$
$$rq^{feeling} = \{2, 6, 9, 14, 17\}$$
$$rq^{thinking} = \{1, 7, 12, 15, 20\}$$

A score for each problem-solving factor $pssqf$ is calculated for a user u_i according to the sum of response values val provided for questions related to each problem-solving factor $pssqf$. The lowest score a user can achieve for a factor in this questionnaire is 5 and the highest score is 25. Specifically, for any $pssqf$ users with a score of 5 to 10 are classified as having a low level, 11 to 17 are classified as having a medium level and 18 to 25 are classified as having a high level.

Emotion Regulation[10]**:** Questionnaire erq measures emotion regulation factors for a given user and consists of 36 Likert-scale questions (taking values from 1 to 5). Currently, only 10 questions from this questionnaire are used to measure two distinct emotional regulation factors $erqf$ including cognitive reappraisal and expressive suppression. From those of 10 questions, each a factor $erqf$ has a set of related questions rq^{erqf} represented by their number j (see below),

$$rq^{cogReappraisal} = \{27, 29, 31, 33, 34, 36\}$$
$$rq^{expSuppression} = \{28, 30, 32, 35\}$$

A score for the factor $cogReappraisal$ is calculated for a user u_i according to the sum of response values val provided for questions related to that factor. The lowest score a user can achieve for the $cogReappraisal$ factor in this questionnaire is 6 and the highest score is 30. Specifically, for $cogReappraisal$ users with a score of 6 to 13 are classified as having a low level, 14 to 21 are classified as having a medium level and 22 to 30 are classified as having a high level. A score for the factor $expSuppression$ is calculated for a user u_i according to the sum of response values val provided for questions related to that factor. The lowest score a user can achieve for the $expSuppression$ factor in this questionnaire is 4 and the highest score is 20. Specifically, for $expSuppression$ users with a score of 4 to 9 are classified as having a low level, 10 to 15 are classified as having a medium level and 16 to 20 are classified as having a high level.

Perceived Expertise: pet is a 10-item questionnaire that is used for measuring the perceived expertise of individuals in the data analytics domain (Germanakos, et al., 2021) and is one of the

---

[9]https://docs.google.com/document/d/1UkN1d2XeAnqsbBnPw98429DC-iHUDN59/edit?usp=sharing&ouid=114970685780107799249&rtpof=true&sd=true
[10]https://docs.google.com/document/d/1HQqERqvYZftKNZTPTuQyhG9bI9g-bmZj/edit?usp=sharing&ouid=114970685780107799249&rtpof=true&sd=true

most important findings/outcomes of the IDEALVis project. The total score a user u_i can acquire from this test is 50. The final score of a user u_i is denoted as $qs^{pet}(u_i)$ and the level of the user (i.e., low, high, or medium expertise) is calculated with three percentile values derived from the set of all users' scores $\{qs^{pet}(u_i), \forall i\}$.

For the pet questionnaire, the 25th, 50th and 75th percentiles of all users' scores are defined below respectively as $Plow$, $Pmid$ and $Phigh$:

$$Plow = \left\lceil \frac{25}{100} \times |\{qs^{pet}(u_i), \forall i\}| \right\rceil$$

$$Pmid = \left\lceil \frac{50}{100} \times |\{qs^{pet}(u_i), \forall i\}| \right\rceil$$

$$Phigh = \left\lceil \frac{75}{100} \times |\{qs^{pet}(u_i), \forall i\}| \right\rceil$$

Accordingly, a user is classified as having a low level of expertise in the data analytics domain if they have a pet score less than $Plow$, medium level of expertise if they have a score higher or equal to $Pmid$ and lower than $Phigh$, and high level of expertise if they have a score higher than or equal to $Phigh$.

4 A User Study for Exploring the Business Analytics Context

4.1 Motivation and Research Questions

Given the users' diversified requirements, needs and perceptual preferences as well as the size, diversity, and processing overhead of big business data sets the main challenge is to provide information in different modalities, navigation patterns and interaction logic. In this respect, the first step is to investigate the contextual building blocks of the business environment like tasks, data and visualization types (see section 3.2), to understand and model the expected adaptation and personalization specifications and to further support our choice of business tasks, data and visualizations as important and distinct dimensions of the final user model. We formulate the following research questions: **RQ1**: Which are the most common tasks of the data analyst in the business domain regarding data visualization and exploration, and how do those differ from tasks in other domains? **RQ2**: What kind of data, visualizations and methods are used for the defined tasks? **RQ3**: Which are the main challenges and needs of data analysts in the business domain?

4.2 Sampling and Procedure

For this exploration study, we involved business participants that have on average at least 2 years of experience in the field of data analytics, and their interaction with data visualizations is part of their daily job responsibilities. The recruitment was made possible with the support of the two partner organizations RAI Consultants and KPMG Cyprus; resulting in a total of 59 data analysts. The sample consisted of 28 Male and 31 Female participants, with their ages ranging from 22 to 56 years old (M = 32, SD = 7). For evaluating their proficiency and experience we analysed the reported educational status (all end-users had achieved higher education), their working experience (ranged from 1 to 25 years (M = 4.3, SD = 6.2)), as well as their Visual Literacy (M = 3.9, SD = 0.7 - captured using the Subjective Graphical Literacy Scale (Garcia-Retamero, et al., 2016)) and Self-Expertise (M = 3.1, SD = 1.3 - obtained through a single 5-point scale self-reporting measure of perceived expertise, i.e., "My level of expertise for the current business role is", where 1 is Novice and 5 is Expert.) Overall, the above findings suggest that the sample is within the initial expectations and goals of this study.

For the execution part, a Web-based environment was created including of a series of questionnaires/tests using several types of questions (e.g., multiple-choice, open-ended, and Likert-scale questions). The study ran in a controlled environment in two sessions with 36 participants in the first and 23 in the second. Each study session was hosted at the premises of each organization and was executed sequentially, with a group of 4 to 7 analysts completing the questionnaires at a time, depending on their availability. For every new group of participants, a researcher was presenting the overall study goals and an overview of the study tasks. At all times during the study the researcher was also in charge for guiding the participants and for answering any potential questions or even resolving any technical conflicts. The participation was voluntary, adhering to the GDPR rules and regulations (Protection, 2018), while each participant required on average 20-25 minutes for completing the questionnaire corpus. After the participants provided their demographic information, such as Gender, Age and Educational Status, they responded to a set of open-ended questions, aiming to collect information regarding **RQ1** with respect to typical business tasks they perform while using visualizations (e.g., Exploration, Correlation, Data Preparation) and their frequency, weekly data analysis requests and their working experience. For addressing **RQ2** participants were given: (a) a matrix of check boxes (19 visualisation types by 10 task actions) where they had to check a maximum of 3 visualization types that they preferred for completing each type of action e.g., Bar, Pie and Column chart used for performing Comparison, and (b) a number of visualization types where they had to report the complexity of each type on a Likert-scale. Finally, for **RQ3** participants had to state the challenges (i.e., pain points) they face during data exploration (including interaction with data visualizations) for accomplishing their business tasks and wishes for improving their daily operations.

4.3 Analysis and Discussion of the Results

Initially, the use of open-ended questions necessitates the extraction and coding of themes for each of the provided answers. Hence, our analysis adhered to the following process: (a) Clean textual responses by removing punctuation, stop words, single letters and unnecessary white space with custom string manipulation functions in Python; (b) generate a document-term matrix; (c) visualize the terms, i.e., words in a word-cloud; (d) manually read answers for formulating different themes and coding specific words into that theme, e.g., if answer contains the words “data” and “cleaning” then code this into a single new term named “Data Cleaning”; (e) repeat from step (c) until a list of themes and their frequencies for a question are formed. Accordingly, descriptive analyses such as frequency distributions and mean were obtained to characterize the derived data.

Thereupon, regarding **RQ1** (i.e., common business tasks), participants responded as follows: 71% Improve Data Quality, 13% Performance Analysis, 12% Correlation Analysis, 12% Comparison Analysis, 12% Drawing Conclusions and 10% Presentations. Other common answers included, pattern detection, correlation analysis, trend, or sales analysis and visualizing KPIs. During their business tasks participants reported that they use data visualizations for an average of 2.5 days per week ($M = 2.5$, $SD = 1.5$) and 2.5 hours per day ($M = 2.5$, $SD = 1.3$), while they handle an average of 3.5 data analysis requests ($M = 3.5$, $SD = 2.6$) on a weekly basis. When asked about the frequency of actions performed during their business tasks, participants responded with Data Preparation, Exploration and Data Communication as the most frequent actions, and with Correlation, Prediction and Classification as the least frequent actions.

The responses of the end-users **RQ2** (i.e., types and complexity of data visualizations, in relation to tasks) show that Pie Charts and Bar Charts (95%), Column Charts (86%) and Line Charts (71%) are considered as simple charts; Radar Charts (47%), Bubble Charts (37%), Rectangular Tree Diagrams and HeatMaps (30%) are considered as complicated charts; and Funnel Charts (44%), Frame Diagrams (42%), Gauge Charts (39%) and Rectangular Tree Diagrams (37%) are rated the highest for being “never used”. Our results for bar chart and radar graph partially agree with the findings of (Toker, et al., 2012) on visualization ease of use and comprehension, whereby the charts classified as simple are commonly used in various analytic systems and dashboards (Lee, et al., 2017) and thus people are more familiar with them. Figure 2 shows all findings regarding the reported visualization complexity. In addition, regarding the preferred types of visualizations for different types of task actions, the analysis revealed that for all actions (i.e., Comparison, Distribution, Contribution, Correlation, Deviation, Cycles, Composition, Trend and Relationship) participants tended to select visualizations that were considered as simple, with the bar chart to be the most preferred visualization. Figure 3 provides more detailed insights on the visualizations that received the highest preference for a specific task action. Some of the collected results are in line with previous findings (Saket, et al., 2019) e.g., using line charts for correlations.

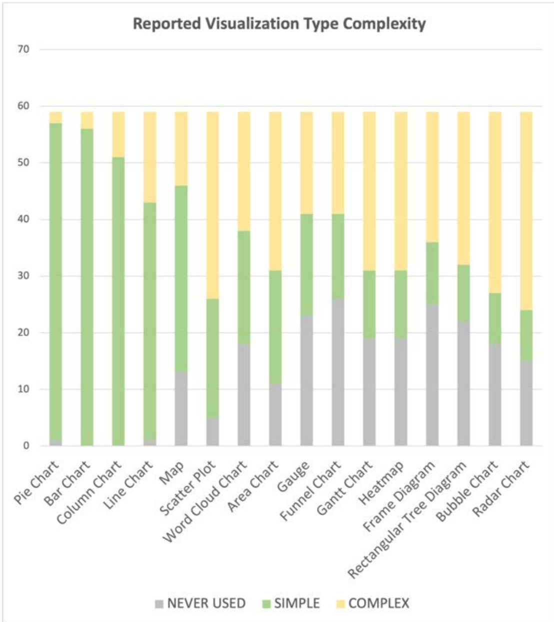

Figure 2 - Reported Visualization Complexity

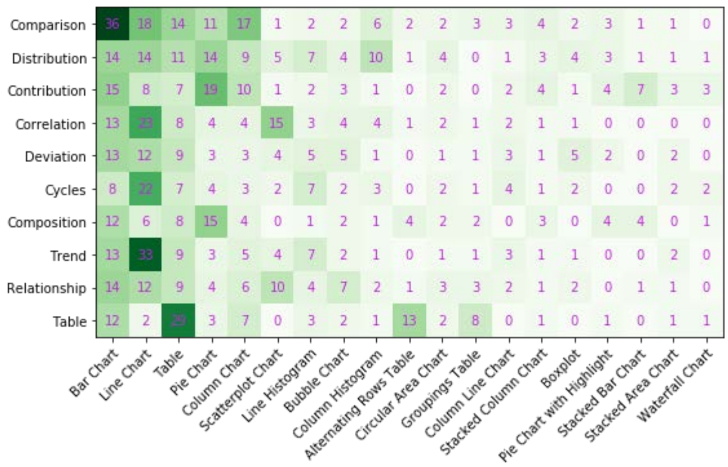

Figure 3 - Visualization Types for Task - Preference

Lastly, for understating the main challenges and needs of data analysts in the business domain **RQ3**, we analysed the main themes provided in end-users' responses about pain points and wishes. The major pain points reported were related to Time Consuming Processes (39%), data related issues such as bad quality of data (41%), data variability (13%), large data volumes (19%) and multiple data sources (7%), hardware speed (12%) and poor or not user-friendly visualizations (15%). On the other hand, participants' wishes were related to asking for better visualization (more automated) tools (17%), faster processes, i.e., better hardware (22%), reduction of analysis steps (8%), easier data integration (7%) and generally user-friendly tools (7%). In relation to **RQ2**, the above findings offer a preliminary input on the nature of data (i.e., large volume / dimensions, multiple data sources and dirty data) being used for the reported tasks in the business domain (also in alignment with the data characteristics in section 3.2).

Interpreting our findings with respect to adaptation and personalization requirements, at a first sight the business tasks could relate to more generic tasks' definitions and structures (Amar, et al., 2005), or specific data visualization types to be used for more commonly recognized actions (Carenini, et al., 2014) (Toker, et al., 2012) (Steichen, et al., 2013) (Saket, et al., 2019), applicable across domains. However, a closer look may reveal significant differences that focus primarily upon: (a) The process of data exploration in the business sector encapsulates a thought process (i.e., a sequence of tasks) that is composed of many subsequent tasks that need to be executed to satisfy a single goal. As opposed to other domains where single visualizations might reflect stand-alone tasks, in this case there is a purposeful workflow that needs to be satisfied, where information and consecutive actions are part of a bigger picture (goal) feeding other actions (from the same or different workflows / roles) until a produced logical result. Visual exploration needs to be flexible, conversational, cooperative, and interactive to be able to accommodate such composite requirements, triggered by process-driven (and not single task-driven) end-to-end scenarios; (b) in many cases, one simple business activity of users may be supported from custom-made developments (e.g., using Excel) for the successive execution of steps necessary towards the fulfilment of the primary objective. As a result, single data visualizations might refer to more than one tasks and need to be adjusted or integrated based on several diverse factors and tools; and (c) for a single objective a combined knowledge is required from end-users to accomplish a series of tasks, many times with hidden dependencies and implications driven by predefined business workflows. Accordingly, different datasets and descriptions may feed the same data visualizations, so transparent exploration and intuitive explanations need to capture the breadth, depth and inherent semantic dependencies generated by the data sources.

5 Expected Benefits and Impact

Given the users' diversified individual differences in cognitive processing, affect, perceptual preferences, role, requirements, needs, and expertise, as well as the size, diversity, and processing overhead of big business data sets, it is expected that this user model will yield flexible best-fit data visualizations and exploration methods that will support the unique end-users with the expected transparency and explainability during an end-to-end interaction.

The suggested adaptive interventions build on the premise that graphics and text have a complementary role in information presentation - while graphics can convey large amounts of data compactly and support discovery of trends and relationships, text is much more effective at pointing out and explaining key points about the data, in particular by focusing on specific temporal, causal and evaluative aspects. Crafting different modalities not only makes the presentation more engaging but could also better suit users with different cognitive abilities and affective states.

In a broader perspective, the results of this research work will have a wider business and economic impact by helping users to comprehend and familiarize themselves with usable data visualizations adjusted to their knowledge and abilities, enhancing their satisfaction and acceptability of related end-to-end business workflows and services. Main vision is that such practices, which provide human-cantered data visualizations and visual analytic services, will be incorporated in future tools and systems, increasing the support and effectiveness of decision making in critical tasks, enabling fast and inclusive action plans, and cutting down unnecessary iterations and costs.

6 Conclusion

While the influence and effect of human factors on visualisations has been widely explored and found significant in various application fields, the business sector, where IDEALVis focuses on, to date has failed to inclusively consider them in the modelling and implementation of data analytics solutions. To address this research gap, we proposed a user model with specific dimensions, including the user and the user's functioning context characteristics, detailing how it may extend prior research. We demonstrate preliminary results from a user study of 59 industry data analysts formulating an understanding of the business contextual characteristics (in terms of tasks, data, and visualizations) and the requirements for adaptation and personalization. Our results solidify our consideration of the business context characteristics as a distinctive facet / dimension of this application area / user model, revealing the complex nature of business tasks and data as well as the requirement for advanced usable visualization tools, i.e., built with the user in mind rather than solid one-size-fits all or data-driven approaches. We expect that the proposed human-centred user model and its dimensions will facilitate the data exploration journey by enabling flexible best-fit data visualizations and methods that will support the unique end-users during their end-to-end interactions.

7 Bibliography

IBM, 2016. *Bringing big data to the enterprise.* [Online]
[Accessed 1 1 2020].

Liu, S., Cui, W., Wu, Y. & Liu, M., 2014. A survey on information visualization: recent advances and challenges. *The Visual Computer: International Journal of Computer Graphics,* 30(12), p. 1373–1393.

Kinkeldey, C., MacEachren, A. M., Riveiro, M. & Schiewe, J., 2017. Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science,* 44(1), pp. 1-21.

Bonneau, G. et al., 2014. Overview and State-of-the-Art of Uncertainty Visualization. In: *Scientific Visualization.* London: Springer-Verlag, pp. 3-27.

Forrester, 2018. *Itelligencegroup - Forrester Research "Becoming an Insights-driven Company".* [Online]
Available at: https://us.itelligencegroup.com/hubfs/IB-NN-Analytics/Infographic-Forrester-Insights-Driven-Business.pdf?hsCtaTracking=19708524-effe-40f1-b888-
[Accessed 06 01 2020].

Kerren, A. & Schreiber, F., 2012. *Toward the role of interaction in visual analytics.* Berlin, Winter Simulation Conference.

Kerren, A., Stasko, J. T., Fekete, J.-D. & North, C., 2008. *Information Visualization - Human-Centered Issues and Perspectives.* s.l.:Springer, Berlin, Heidelberg.

Lallé, S., Conati, C. & Carenini, G., 2017. *Impact of Individual Differences on User Experience with a Visualization Interface for Public Engagement.* New York, Association for Computing Machinery.

Toker, D., Conati, C., Carenini, G. & Haraty, M., 2012. *Towards Adaptive Information Visualization: On the Influence of User Characteristics.* Berlin, Springer.

Shneiderman, B., 1996. *The eyes have it: a task by data type taxonomy for information visualizations.* Boulder, CO, USA, USA, IEEE.

Steichen, B., Carenini, G. & Conati, C., 2013. *User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities.* New York, Association for Computing Machinery.

Toker, D., Lallé, S. & Conati, C., 2017. *Pupillometry and Head Distance to the Screen to Predict Skill Acquisition During Information Visualization Tasks.* New York, Association for Computing Machinery.

Green, T. M. & Fisher, B., 2010. *Towards the Personal Equation of Interaction: The impact of personality factors on visual analytics interface interaction.* Salt Lake City, IEEE.

Carenini, G. et al., 2014. *Highlighting interventions and user differences: Informing adaptive information visualization support.* Toronto, ACM.

Steichen, B. & Fu, B., 2019. Towards Adaptive Information Visualization-A Study of Information Visualization Aids and the Role of User Cognitive Style. *Frontiers in Artificial Intelligence,* 2(22).

Cernea, D., Ebert, A. & Kerren, A., 2013. *A Study of Emotion-triggered Adaptation Methods for Interactive Visualization.* s.l., s.n.

Poetzsch, T., Germanakos, P. & Huestegge, L., 2020. Toward a Taxonomy for Adaptive Data Visualization in Analytics Applications. *Frontiers Artif. Intell,* Volume 3.

Mutlu, B., Gashi, M. & Sabol, V., 2021. *Towards a Task-based Guidance in Exploratory Visual Analytics.* s.l., ScholarSpace.

Chen, H., Chiang, R. H. & Storey, V. C., 2012. Business intelligence and analytics: From big data to big impact. *MIS quarterly,* pp. 1165-1188.

Witkin, H. A. et al., 1977. Role of the field-dependent and field-independent cognitive styles in academic evolution: a longitudinal study. *Journal of educational psychology,* 69(3), p. 197.

Friel, S. N., Curcio, F. R. & Bright, G. W., 2001. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in mathematics Education,* 32(2), pp. 124-158.

Okan, Y., Garcia-Retamero, R., Cokely, E. T. & Maldonado, A., 2012. Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. *Journal of Behavioral Decision Making,* 25(4), pp. 390-401.

Goldberg, L. R., 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology,* 59(6).

Cacioppo, J. T. & Petty, R. E., 1982. The need for cognition. *Journal of personality and social psychology,* 42(1).

Liu, Z., Crouser, J. & Ottley, A., 2020. Survey on Individual Differences in Visualization. 39(3).

Walla, P., 2018. *Affective processing guides behavior and emotions communicate feelings: Towards a guideline for the NeuroIS community.* s.l., Springer, pp. 141-150.

Lekkas, Z. et al., 2009. *The role of affect in personalized learning.* s.l., IEEE.

Dreyfus, S. E. & Dreyfus, H. L., 1980. *A five-stage model of the mental activities involved in directed skill acquisition,* s.l.: s.n.

Allen, B., 2000. Individual differences and the conundrums of user-centered design: Two experiments. *Journal of the american society for information science,* 51(6), pp. 508-520.

Usability.gov, 2021. *Personas.* s.l.:s.n.

Amyrotos, C., Andreou, P. & Germanakos, P., 2021. *Adaptive Business Data Visualizations and Exploration: A Human-centred Perspective.* s.l., s.n.

Garcia-Retamero, R., Cokely, E. T., Ghazal, S. & Joeris, A., 2016. Measuring graph literacy without a test: A brief subjective assessment. *Medical Decision Making,* 36(7), pp. 854-867.

Protection, D., 2018. *Rules for the Protection of Personal Data Inside and Outside the EU. 2018.*
[Online]
Available at: https://ec. europa. eu/info/law/law-topic/data-protection_en
[Accessed 03 10 2020].

Lee, S., Kim, S.-H. & Kwon, B. C., 2017. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics,* 23(1), p. 551–560.

Saket, B., Endert, A. & Demiralp, Ç., 2019. Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics,* 25(7), pp. 2505-2512.

Amar, R. A., Eagan, J. R. & Stasko, J., 2005. *Low-Level Components of Analytic Activity in Information Visualization.* s.l., IEEE Computer Society.

Germanakos, P., Lekkas, Z., Amyrotos, C. & Andreou, P., 2021. *Proposing a Perceived Expertise Tool in Business Data Analytics.* Utrecht, Association for Computing Machinery.