Scalable Data Analytics,

Scalable Algorithms, Software Frameworks

and Visualization ICT-2013 4.2.a

Project	FP6-619435/SPEEDD
Deliverable	D5.3
Distribution	Public

http://speedd-project.eu

The Design of User Interfaces for the SPEEDD Prototype

3rd report

Chris Baber, Sandra Starke, Xiuli Chen, Natan Morar, Andrew Howes (University of Birmingham)

Status: Final

October 2016

Project	
Project Ref. no	FP7-619435
Project acronym	SPEEDD
Project full title	Scalable ProactivE Event-Driven Decision Making
Project site	http://speedd-project.eu/
Project start	February 2014
Project duration	3 years
EC Project Officer	Stefano Bertolo

Deliverable	
Deliverable type	Report
Distribution level	Public
Deliverable Number	D5.3
Deliverable Title	The Design of the User Interface for the SPEEDD Prototype – 3rd report
Contractual date of delivery	M34 (December 2016)
Actual date of delivery	October 2016
Relevant Task(s)	WP5/Tasks 5.1, 5.2, 5.3
Partner Responsible	UoB
Other contributors	
Number of pages	
Author(s)	C. Baber, S. Starke, X. Chen, A. Howes, N. Morar
Internal Reviewers	
Status & version	Final
Keywords	Cognitive Work Analysis, Visual Analytics, User Interface Design, Decision Modelling, Eye Tracking

0 Executive Summary

This report summarises the main work undertaken in Work Package 5 over the SPEEDD project. In particular, the focus on the report is on the relationship between Sociotechnical Systems concepts that informed the field work and discussions with Subject Matter Experts and the design process used to construct the dashboards. In section 4, this relationship is illustrated in a process that shifts focus between technical and 'human' (organisational and perceptual) constraints to produce a dashboard that can not only operate in the SPEEDD architecture but also reflects our understanding of the ways in which operators in the use case domains process information and make decisions.

In order to understand the relationship between information presentation and decision making, Work Package 5 has undertaken three interconnected activities. In the first, we have observed and interviewed (as far as practicable) Subject Matter Experts in their work domains, performing their normal working activities. From this, we have abstracted a core set of task demands which are used to design simulated tasks for laboratory experiments. The benefit of such tasks is that we can maintain control over independent variables in experimental design and minimize confounding variables in ways that are not possible in the operator work environments. This has been particularly beneficial in that it has allowed eye-tracking studies to be conducted in a controlled and rigorous manner (although we have reported eye-tracking studies in the DIRCE control room). In addition to supporting eye-tracking studies, the experimental tasks have allowed us to manipulate the reliability of the automated support. From a Human Factors perspective, we have been particularly interested in the question of how people respond to suboptimal decision support (e.g., at reliability of 25%) or to quite reliable support (e.g., at reliability of 81%). We have shown that, even when they have no indication of reliability, people are able to adapt their strategies and change their reliance on the support or their own decision rules.

Experiments and modelling work have focused on the manner in which people sample the information available to them. The modelling work, in particular, suggests that, for tasks in which there can be clear definition between validity of cues and a well-defined decision, people exhibit performance which agrees with an optimal decision model. Further experiments have suggested that, when the definition of cue validity of the definition of the problem is less clear, then people will sample more information (and, at times, this could lead to over-sampling which, while inefficient, did not seem to impair either decision time or accuracy). Thus, people are able to adapt their information sampling strategy to reflect their impressions of the reliability of automated decision support, their understanding of cue validity and the clarity of the decisions required. In other words, people will seek to make specific information from the situation space on to the decision space for their jobs.

We have used this understanding (of mapping situation and decision spaces) to inform the design of dashboards for the SPEEDD project. Evaluation studies, comparing different versions of the dashboard, suggest that we have produced reliable visualisations of the situation space for operators in both Road Traffic Management and Credit Card Fraud analysis.

Contents

Figures

1. Introduction

1.0 History of the document

Version	Date	Author	Change Description
0.1	07/10/2016	Chris Baber (UoB)	Set up the document and initial content
0.2	12/10/2016	Sandra Starke (UoB)	Descriptions of fraud analytics and experiments added
0.3	13/10/2016	Natan Morar (UoB)	Description of user interface design added
0.4	14/10/2016	Andrew Howes (UoB)	Discussion of decision modelling work added
1.0	15/10/2016	Chris Baber (UoB)	Comments incorporated and draft submitted for Consortium review
1.1	24/10/2016	Alex Artikis (NCSR)	Review of report
2.0	24/10/2016	Chris Baber (UoB)	Final version of report

1.1 Purpose and Scope of the Document

This report summarises the activity under Work Package 5 in terms of the objectives laid out in the Description of Work.

1.2 Work Package 5 Objectives in Description of Work

The SPEEDD project aims to make ‘Big Data’ accessible to human operators, so that recommendations offered by the machine learning components can be understood by operators. This is important, as decisions may have to be verified by human operators, either due to legal requirements (as in the road traffic control use case) or due to customer-service requirements (as in the fraud use case). A core question is, to what extent, the human operator will respond to 4Vs of Big Data (i.e., volume, variety, veracity, velocity). The main reason for introducing automation into Big Data systems is to cope with the volume of data and the speed (velocity) with which these data are presented for processing. Consequently, one would expect automation to provide some form of triaging of data prior to its presentation to the human operator. Hence, both SPEEDD use cases require the presentation of data that the automatic system has already processed. From this, the focus of the human is less on volume and velocity than on the other two Vs. In terms of veracity, the role of the operator could be to confirm the recommendations made by the automated system or to check the conclusions that such a system draws from the data, i.e., to provide checks and balances in terms of automated data processing. In such cases, a subset of data is presented to the decision maker in the expectation that this will provide a sufficiently detailed view of the situation space (Hall et al., 2007) to allow the decision maker to select an appropriate course of action (or, at least, to enable a clear understanding of the situation space). Consequently, there is an assumption that visualising the situation space should help the decision maker understand the decision space (Hall et al., 2007). In this context, the decision space can be defined as the set of possible decisions which could be made by the decision maker, given the information in the situation space. However, humans are known to be efficient when it comes to expending effort on information search (Pirolli and Card, 1999, Gigerenzer and Gaissmaier, 2011), and just making a large amount of data available does not mean that all of it would actually be used. In terms of variety, the human operators might have access to information that is not presented to the automated system. Such information could be presented verbally (through telephone or radio calls) or visually (through Close-Circuit Television systems), or it could be contained in systems that lie outside the scope of the automation, e.g., personal bank accounts which are not open to the automated analysis.

Work Package 5 sought to extend the state-of-the-art in the design, development and evaluation of visual analytics for Big Data applications in three ways:

- understanding expert decision-making in context,
- modelling expert decision-making,
- ecological interface design for visual analytics.

The primary focus of Work Package 5 was to develop novel approaches to visual analytics for the SPEEDD use case domains. In this context, we follow the definition of Visual Analytics offered by Thomas and Cook (2005), as the science of analytical reasoning supported by interactive graphical interfaces. This involves both the automated analytical reasoning (undertaken by our partners in SPEEDD and reported in WPs X & X) and the analytical reasoning made by the humans who interact with the visualisation. In this way, Keim et al. (2010) suggest that Visual Analytics supports *"the exploration of information presented by visualisations [in] a complex process of sense-making"* (p.116). This means that, in order to design and develop user interfaces which can support human decision making, we need to understand the processes of sense-making in the use case domains. Furthermore, in previous work we have shown that the manner in which people interact with Visual Analytics is not simply a function of what is displayed to them, but a matter of how this display corresponds to their existing knowledge (Ishack et al., 2015).

The interaction of human decision-makers with the SPEEDD prototypes, drawing on the observation studies conducted in the use case WPs 7 and 8, lead to:

- understanding the decisions involved in each SPEEDD use case and the information sources on which these decisions are based, in order to understand how decision-makers assimilate information from multiple sources in a socio-technical system;
- description of human decision-making as a rational, objective response to goals and changing context, thus understanding human decision-making in the context of SPEEDD and the manner in which this process can be learned and refined by both human and automated system through their interactions;
- development of novel, real-time visualization techniques for Big Data that support human decision-makers in their situation awareness, sense-making, decision-making and appreciation of recommendations made by the SPEEDD prototype.

From this perspective, Work Package 5 has involved Human Factors research into understanding how experts in the use-case domains make decisions in their day-to-day roles; how such decision making can be influenced by the ways in which the information available to them in sampled and utilised; how graphical user interfaces can be designed to best support this decision- and sense-making. Each of these areas is the subject specific tasks under this Work Package.

2. Modelling Decision-Making as a Socio-Technical Activity

2.1 Joint Optimization in Socio-Technical Systems

In any organisation, efficient operation involves a combination of Technical factors , relating to the technology, data and processes used by that organisation, and Social factors , relating to the behaviour of people in that organisation (Cherns, 1976). In an ideal state of affairs, there will be 'joint optimisation' (Cooper and Forster, 1971) of these factors so that they work seamlessly together. However, a defining aspect of any Socio-Technical System (STS) is that the relationship between these factors tends to be non-linear, so that optimization of one factor can lead to unpredicted consequences in the other. In terms of the SPEEDD project, the challenge is to balance the expectations and assumptions which inform the Technical solutions to the processing of Big Data in the two use cases, with the knowledge, skills and abilities of the people who will be using these solutions. In broad terms, this raises the concept of a 'whole task' from STS (Trist and Bamforth, 1951), in which the boundary between receiving information and performing an action is clearly marked and is meaningful for all parties involved in the work (both human and automation). The challenge for any user of Big Data analytics thus lies in how easy it is to understand the recommendations of the automated decision maker and how to incorporate these recommendations into one's own decision making. This also raises the question of how to manage the responsibilities for who (or what) will perform the action.

While the SPEEDD architecture incorporated assumptions about the division of processing between modules and the user interface, the manner in which the user might interact with the recommendations could be affected by either the understanding of the recommendation, or the perceived reliability of the automation. Consequently, in the Human-Automation Systems that SPEEDD is developing, the Allocation of Function, i.e., who does what, becomes an important issue and the ways in which such Allocation might shift in response to the reliability of automation is a question that has been addressed through a series of experiments in this work.

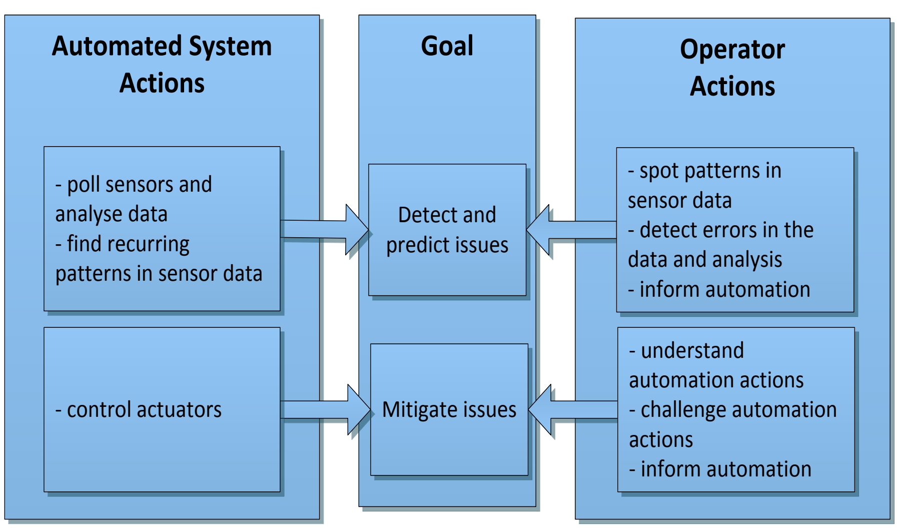


Figure 1: Contribution of System Components to System Goals [from Morar et al., 2015]

In the initial work on this question, Morar et al. (2015) note that there is a potential danger in seeing the user of the automated system simply in terms of the passive audience of the automated output: *"Not only does the relegation of the human to an acceptor of system recommendations miss the point of the Visual Analytics concept, but it also removes the human operator from the analysis loop."* In this case, the human could find it difficult to reason about the automation's recommendation (Bainbridge, 1987) or could have less trust in the automation than if the human had been involved in the discovery process (Keim et al., 2010).

Human intervention could, for example, arise when the automation's performance is below an acceptable level or when the human has information which is not available to the automation. In other words, human decision-makers will be drawing upon sources of information which need not be explicitly represented in the Big Data available to SPEEDD, e.g., hunches, rumours, guesstimates and other types of 'informal' data which could influence human decision-making. The question for our research is when, and why, people might introduce this material in their decision making, and what impact it might have on overall performance of the combined human-automation system. For example, in the experiments on identifying Fraud patterns [REPORT WP7.3?] participants reported attempting to scan the available information to develop a 'gut feeling' and then either reducing the number of cues to attend or to form a story, e.g., "[…] for example, has the person stayed out late at night or are they purchasing a one-off strange item. In these cases where there was a possible innocent explanation I generally let the transaction through". These 'stories' are not part of the presented data (nor necessarily even supported by the data) but illustrate how sense-making can involve adding information to better develop a frame in which to interpret the available data.

In terms of Road Traffic Management, a core information source for the control room operators is the images displayed on the CCTV screens in the control room. The content of these displays are not available to the SPEEDD decision algorithms (which rely on the data captured from sensors buried in Grenoble's ring road). However, the control room does employ software which scans the image content to recognise obstructions in the road, and then alert the operators. In D8.5, it was shown that operators did not spend much time monitoring the bank of CCTV screens on the back wall of the control room, but preferred to focus their attention of the desk-mounted display screens. Following our report to the control room manager, there has been a reconfiguration of the control room (as shown in figure 2).

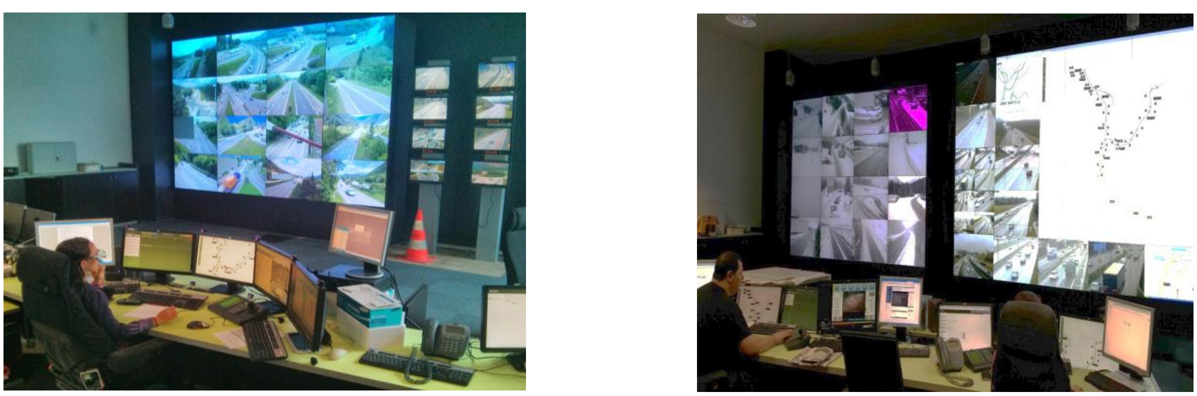


Figure 2: Comparing control room over time. On the left is the control room when we first visited it. On the right is the control room on our most recent visit.

2.2 Situation spaces, decision spaces, situation awareness

Complex decisions are rarely made by individuals; rather they are made by teams of people working as part of a socio-technical system. The primary socio-technical focus for this work is decision-making in the small groups formed by the SPEEDD prototype, its operators and their colleagues (figure 3).

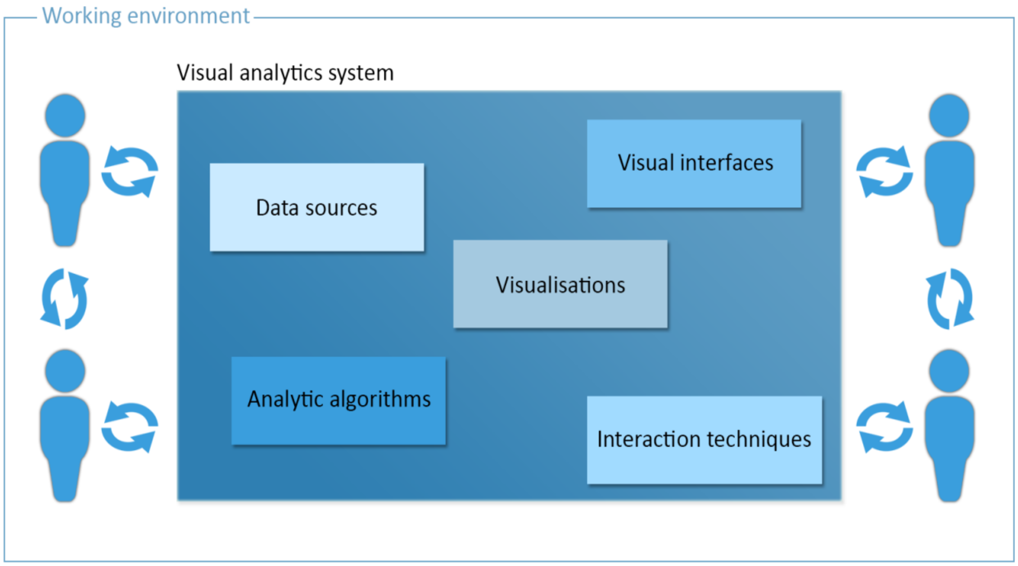


Figure 3: Visual Analytics in a Socio-Technical System [from Morar et al., 2015]

Considering the Road Traffic Management use case, figure 4 illustrates how functions (defined using Cognitive Work Analysis which is described in detail in D5.1 and D8.1) can be performed by different people inside and outside the control room. In any cell in which there is more than person, there is the assumption that some of communication could be required between these people in order to perform a given function in a given situation. For example, consider the function ‘Detect incident’ in the situation ‘Response to incidents’. In this cell, drivers, patrols, police / emergency, operator and the automated system can all contribute to performing the function. While the operator and automated system will communicate with each other, it is plausible that either patrols or police / emergency may also be in communication with the operator (via radio or telephone) and the operator will communicate with drivers using Variable Message Signs. This is not intended to be an exhaustive description of the possible communications which could arise in this situation, but does highlight that there will be flows of information that are beyond the behaviour of the automated system. This means that a critical role of the operator in any system involving automation is to provide a current and coherent awareness of the ongoing situation. This Situation Awareness (Endsley, 1995) involves perception of current state of the road network, comprehension of this information and projection (prediction) of possible future states and potential problems. For an event-driven decision support system that SPEEDD can offer, the projection could be enhanced by the automation.

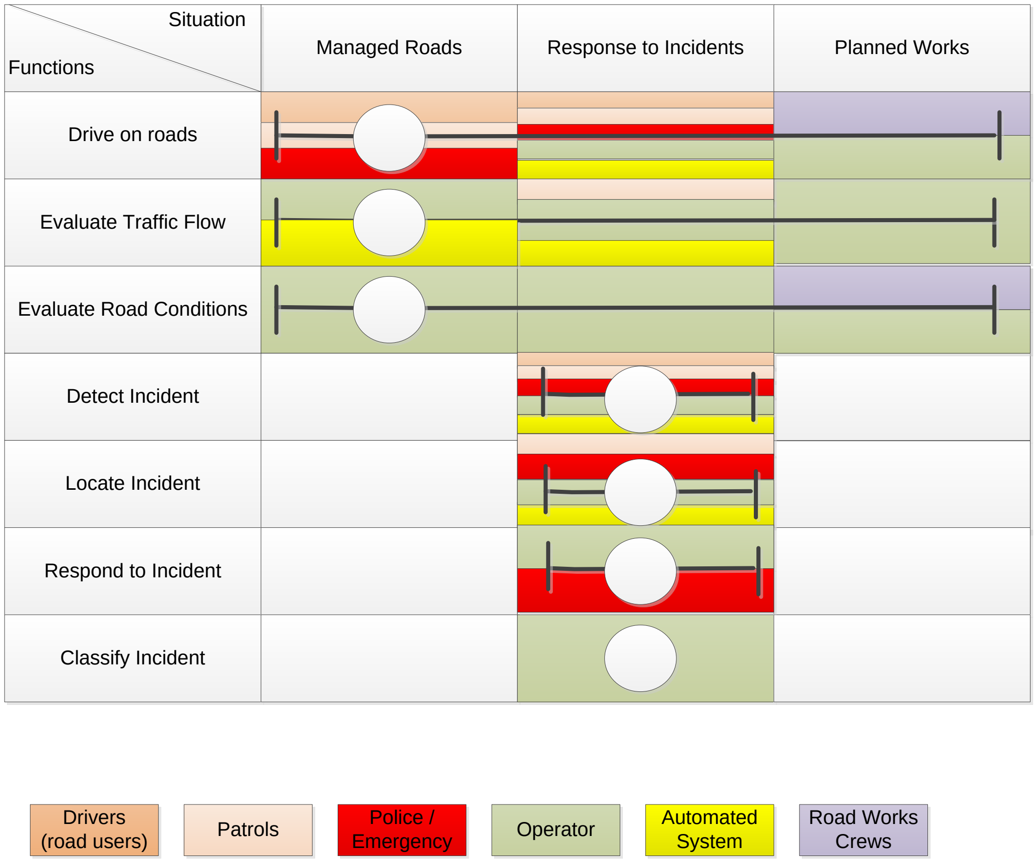


Figure 4: Illustrating the role and behaviour of different Actors in a Road Management network using CWA Social Organisation and Cooperation Analysis

2.3 Understanding the role of Fraud Analysts

Through discussion with Subject Matter Experts in Feedzai, FICO, UK Cards Association and high-street banks, we identified (at least) three different principal settings in which fraud analysts work with data related to credit card transactions and potential fraud: call centers, pattern analysts and technical investigative units. The roles, skills and responsibilities of staff vary largely between these settings and, depending on the role, staff work with different data.

Call center agents are usually the first point of contact after a transaction was blocked by an automated system: in credit card fraud detection, the transaction risk scoring is performed fully automatic through tailored software packages, such as Feedzai's Pulse system. The transactions at risk can be identified in near real time while a purchase is being validated. After automated risk assessment, the transaction is either allowed or blocked. Blocking a transaction can result in blocking the card from all further transaction, until the card owner has been spoken to. Call center agents are the staff who have to follow up these blocked transactions and speak to the customer. The exact way in which this is executed depends on the bank, and there are various approaches to make this point of contact more automatic. However, the normal case is that the call centre operative will get in touch with the customer via phone, enquiring about the questionable transaction. When contacting the card owner, call-center agents follow a clearly defined script. The aim is to determine whether the person answering the call is the genuine card holder and whether the card holder made the purchase legitimately or whether the transaction was a fraud which the card holder does not know anything about. When working in the role of call center fraud staff, agents have access to transaction and customer details, both past and present. In this role, each flag is dealt with reasonably fast; the total number of cases to be processed by an operator can be around 200 per day. This can be equated to almost 30 transactions per hour (assuming an 8 hour day with lunch break), or around 2 minutes to investigate each transaction. Hence, there is usually no deep level of investigation of the transaction specifics; the information is used to validate that the customer for example remembers some recent purchases (to confirm identity) and to explain to the customer which transaction was blocked, and why. Alerts are dealt with according to the risk level; alerts with the highest level of risk are usually worked on first (this is called 'priority mode') and the remaining alerts are cued. Given the rate at which the human operator might be expected to process these transactions, it might be expected that any design in the user interface that can result in marginal gains in operator time could be beneficial. Indeed, our final user interface for the SPEEDD prototype results in time to process a suspicious transaction of between 12.5seconds to 15 seconds.

Fraud analysis that is performed at a higher level, e.g. with the purpose of tuning automatic fraud detection system or exploring new trends that might be just emerging, is a process that requires more in-depth analysis of transaction behaviour at the level of individual accounts, but also e.g. across countries, institutions or time zones. Hence, this staff works with more data than call-center operators, often from multiple databases, and also requires the ability to aggregate data. The case load for specialist fraud analysts is substantially lower, with around 10 cases per day; training for this role is more substantial, and wages are considerably higher compared to call center operators. Analysts look at fraud patterns in broad terms, for example if regions-specific patterns emerge. They have access to past transactions, which includes the risk score assigned by an automated system and the outcome of the decision (as well as possibly a transcript or case summary) which the call center agent made after calling the customer following a blocked transaction. Further, agents may work with aggregated data (spending per day, transactions per hour etc.). As we described in section 3.3, we have produced a simulated fraud analysis task that allows us to explore how analysts search for information that can help them determine whether a transaction is likely to be fraudulent.

In addition to cardholder activity, merchant behaviour can be monitored: if three cards in a cue are confirmed fraud, all other cards in that cue may be blocked for reasons of caution. Similar to higher level fraud analysis, technical investigative units examine cases where fraud was due to exploitation of security gaps or technological interception. In our work we did not focus on setting, as it requires a different skill set and is driven more by software and hardware analysis.

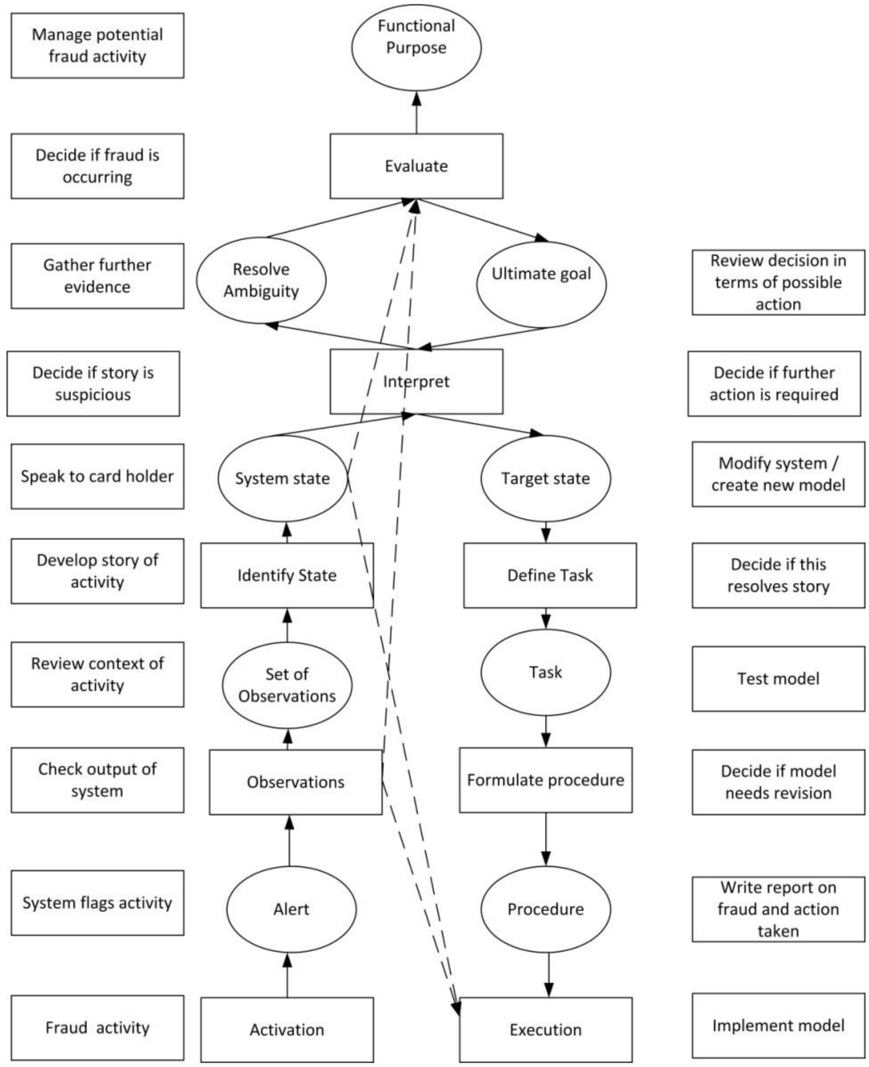


Figure 5: Decision Ladder (Control Task Analysis) for generic Credit Card Fraud Analyst Activity

From our initial consideration of fraud analysis, we developed a Control Task Analysis, or 'decision ladder' (figure 5). This shows the steps that we believe a generic credit card fraud analysis process might follow. Bearing in mind the challenges of speaking in detail about fraud detection practices, this figure is built from a variety of sources and is not intended to represent any specific organization. The central part of figure 5, i.e., the 'step-ladder' from Activation up to Functional Purpose and down to Execution, represents a flow of tasks that an operator or analyst might perform. The dotted lines between the two sides represent short-cuts that the experienced analyst might make in response to specific information. On either side of the decision ladder, are specific instances of actions that the analyst is likely to perform in order to complete the tasks.

2.4 Understanding Road Traffic Management

Road traffic control involves the monitoring of traffic, responding to incidents and influencing road user behaviour. Given that incidents can contribute to some 25% of the overall congestion levels on major roads (UK Highways Agency, 2009), it is important that any incident is resolved as quickly as possible. Regional Control Centers, such as the 'Direction Interdépartementale des Routes Centre-Est' (DIR-CE) in Grenoble, France, are the central focus of communications regarding major roads and will monitor traffic flow (through CCTV, through verbal reports or through sensor data from the roads or vehicles) and control the Variable Message Signs on these roads. In broad terms, the goals of such a center can be summarized as follows (Folds et al., 1993): i.) Maximize the available capacity of the roadway system; ii.) Minimize the impact of incidents (accidents, debris, etc.); iii.) Contribute to demand regulation; iv.) Assist in the provision of emergency services; and v.) Maintain public confidence in the control centre operations and information provision.

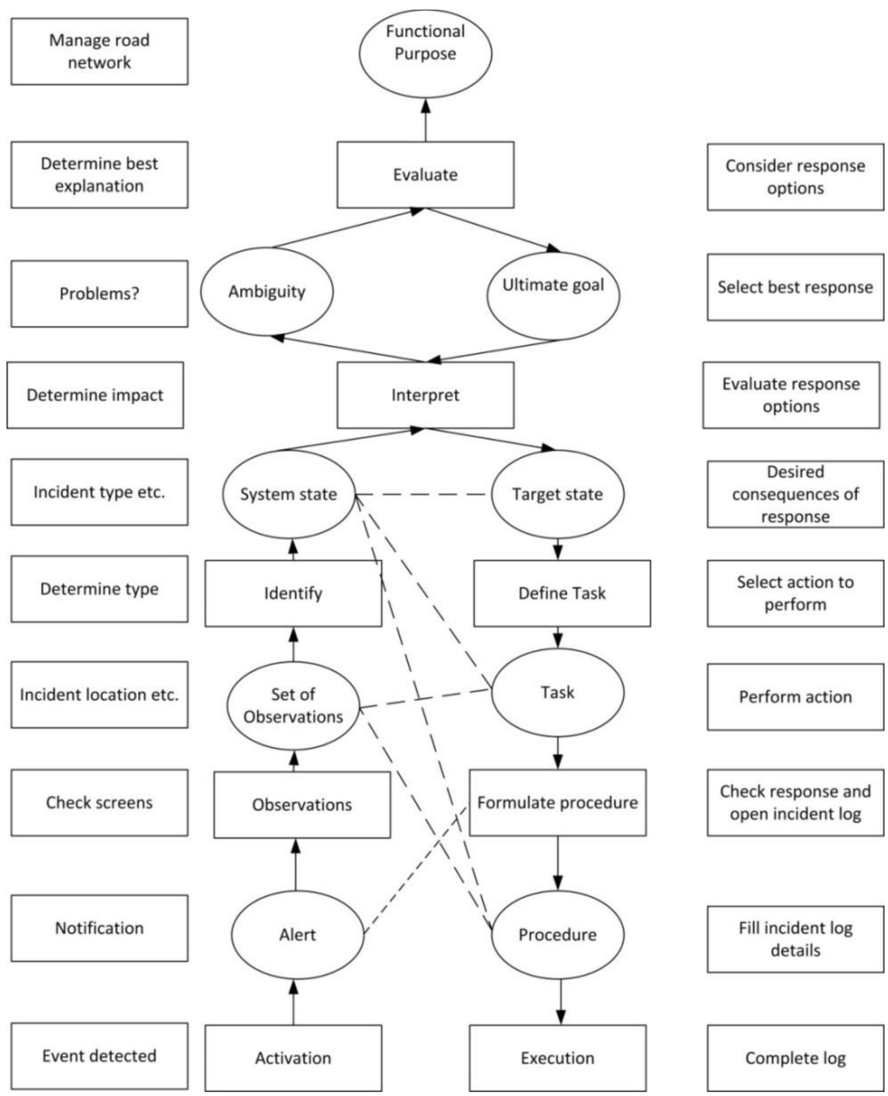


Figure 6: CTA for RTM

Studies involving eye-tracking of controllers in the DIRCE control room (detailed in D8.1 and D8.3 and reported in more detail in Starke et al., 2015, in press) shows how situation awareness involves two processes. The first relates to the perception of the situation and requires the ability to collate sufficient information to define the situation. As we discuss in D8.1, the definition of the situation is supported by the categorisation scheme used in the incident log. Indeed, operators must rely on the predefined categorisation (which has been designed to cover all eventualities that operators encounter). The second process which is relevant to perception of an event is the ability to recall previous, related examples. In our observations, this ability was supported mainly through discussion with colleagues (although there is also the likelihood that the operator would simply remember similar events). The operators are monitoring the situation and choosing an appropriate response to make. We believe that this is not a sequential process of perceive / comprehend and then respond, but rather an interleaving process in which both activities are performed in parallel, with one influencing the other.

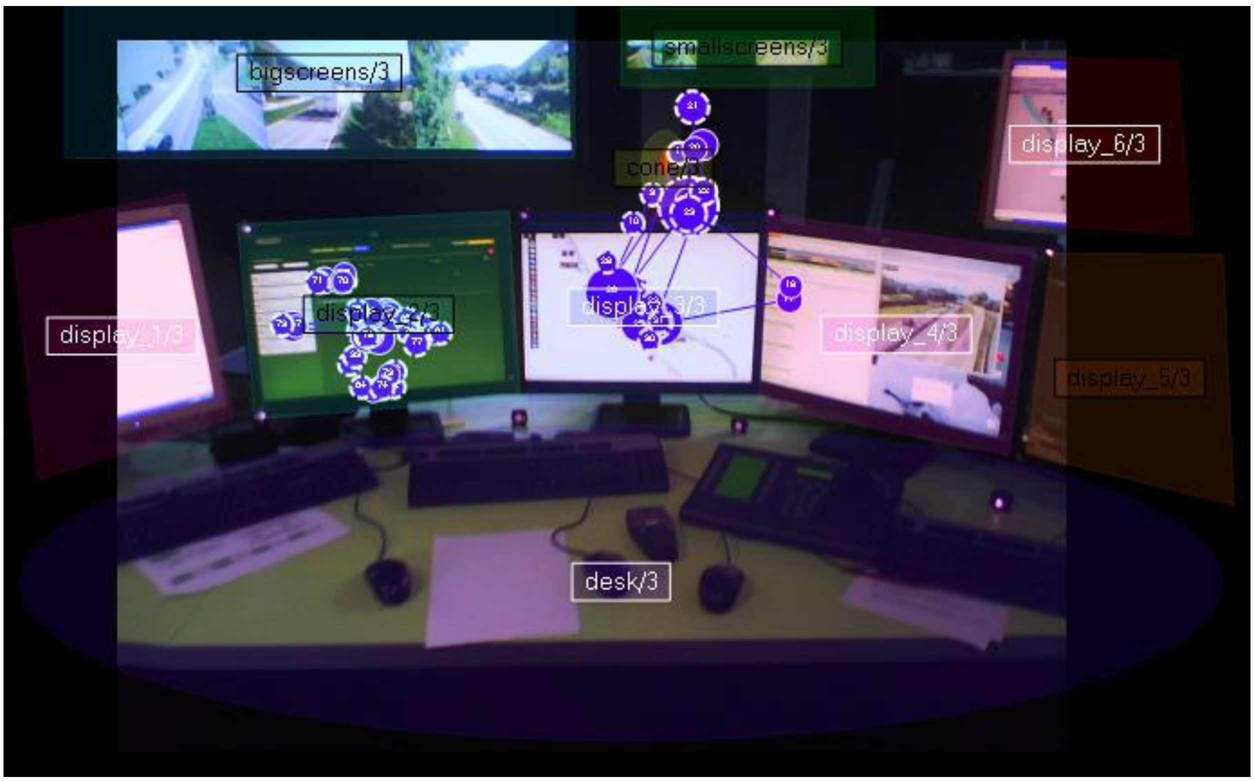


Figure 7: Summary of scan patterns in eye-tracking by one operator in DIRCE control room

3. Defining Objective Metrics for Evaluating Decision-Making

3.1 A Novel Approach to Modelling Human Decision Making

In Chen et al. (submitted) we proposed an integrative cognitive model of how the process of interaction with visualizations supports decision making. It is an attempt to explain how display design can contribute to reducing information access cost (in terms, for instance, of the perceptual and cognitive effort that is required to find and use information relevant to a specific action) and, therefore, efficient decision making. A key feature of the model is that the predicted user strategies are an emergent consequence of the cost structure of the task, rather than a model assumption. The model builds on a number of recent threads in Human-Computer Interaction and decision making research. These threads are described below.

3.1.1 Partially Observable Markov Decision Processes (POMDP).

A POMDP is a mathematical framework for modelling sequential decision problems in which (1) only incomplete observations of the state of the world can be made, and (2) the transitions between states are stochastic. These properties are important for modelling interactive decision making because of the partial and stochastic nature of human information gathering.

3.1.2 Active vision.

While vision that is centrally focused (foveal) around the retina is very accurate, peripheral vision is much less so (Kieras, 2014). As a consequence there is uncertainty as to which, and whether, objects in peripheral vision are detected. This uncertainty limits the extent to which visualizations in HCI can support direct perception. As a consequence multiple eye movements and fixations are often required to build a belief about what is on the display that is sufficient to guide action. Eye movements are, therefore, actively recruited in order to solve the user's decision task, not to build a complete model of the display. While we have used eye-tracking in several of our studies, from DIRCE control room to laboratory studies with the different user interface designs. Examples of these studies have been considered in the previous section. The data from these studies provide some of the fundamental measures used in the models developed below. In addition to eye-tracking data, recent work has shown how colour, shape, and size have different consequences for visual search of displays. In order to reflect these consequences, we have used the constraints proposed in Kieras (2014) to define the POMDP's observation function.

3.1.3 Machine learning.

Human eye movements can be predicted by solving a POMDP, bounded by a model of human vision, with machine learning (Butko, 2008, Chen, 2015). Machine learning is used to find a strategy for choosing actions given each belief about the state of the world, where beliefs about the world are built through repeated fixations. For any POMDP there are one or more optimal strategies that choose the actions that maximize the average utility of interaction. In interactive decision tasks, the problem is to find a strategy for moving the eyes so as to generate good estimates of the task-relevant displayed state and, therefore, best possible decisions. We might expect that good visualizations facilitate efficient strategies by increasing the rate or reliability with which the human eye gathers information.

3.1.4 Decision strategies.

Behavioural evidence suggests that people use simple rules (e.g., search rules, stopping rules and decision rules) in order to make good decisions quickly (Gigerenzer, 1999, Gigerenzer and Goldstein, 1996; Gigerenzer, 2011). The use of machine learning to derive strategies allows us to explore the extent to which these heuristics are an emergent consequence of adaptation to the decision task specified in the POMDP and, therefore, the limits of the specified cognitive-perceptual mechanisms (Howes, 2009; Lewis, 2014).

To test the model, data were collected using a laboratory study of human participants performing a laboratory variant of the credit card fraud detection task, requiring a decision about whether a credit card payment might be fraudulent (see section 3.3). In order to do this, participants were provided with information about various card transaction cues (e.g. transaction history, amount etc.) that were presented using one of two types of display. In one type of display, the information was presented numerically, and in the other, a colour-map was used. We hypothesized that the colour-map would enhance the human capacity to encode information using peripheral vision in accordance with the constraints investigated by Kieras (2014). Moreover, we used the model to predict the decision time, accuracy and number of fixations used for each interface. In addition, the study manipulated the cost of accessing each cue. In one condition, cues were always present on the display and in the other they were covered until individually revealed by a mouse click.

The contribution of the work is in providing a new model and empirical validation of how people make decisions through interaction with visualizations such as colour-maps. The model shows how decision making behaviour is an emergent consequence of adaptation to display design. Critically, the model does not make any priori assumptions about decision strategy, rather the strategy is calculated using a machine learning algorithm. In Chen et al. (submitted) comparisons between the model's behaviour and human performance metrics (including performance time and eye movements) are provided.

3.2 Task

We modelled the credit card fraud detection task. As noted in section 2.3, despite the use of automated detection algorithms, there continue to be key roles for people to play in the analysis process. These roles range from defining and tuning the algorithms that automatic systems deploy, to triaging and screening recommendations from such systems, to contacting customers (either to query a transaction or to explain a decision). For this task, we assume that an automated detection process is running and that this process has flagged a given transaction (or set of transactions) as suspicious and a user will engage in some form of investigation to decide how to respond to the flag.

In this study, we use a simplified version of fraud detection in which the task is to decide whether a transaction should be blocked (prevented from being authorized) or allowed. Participants are provided with 9 sources of information (cues) and these are presented using one of 4 display designs (visualizations). The cues differ in the reliability with which they determine whether or not a transaction is a fraud and the participants must discover these validities with experience and decide which cues are worth using to make a decision.

The visualizations presented in the experiment did not have the full richness and complexity of the dashboards proposed for the project (and presented in section 4). Our focus on colour is partly to enable simple experimental designs and also because colour is known to be more detectable with peripheral vision than shape.

The theoretical claim made by Chen et al. (submitted) is that decision strategies are an emergent consequence of both the statistical properties of the environment (the experienced cue validities, different time cost for extracting information) and of the constraints imposed by human perceptual mechanisms. For a visualization task, this theory can be written precisely by formulating the decision problem as a POMDP for active vision and solving this problem with machine learning to find an optimal decision strategy (emergent heuristics).

In the resulting model, eye movement strategies and stopping rules, are an emergent consequence of the visualization and the limits of human vision (Chen, 2015; Tseng, 2015; Butko, 2008). The assumption is that people choose which cues to look at and when to stop looking at cues informed by the reward/cost that they receive for the decisions they make. Better decisions will receive higher rewards, which will reinforce good eye movement strategies.

3.3 Problem Formulation

We assume that the problem faced by a decision analyst can be modelled as a Partially Observable Markov Decision Process (POMDP) (Kaelbling, 1998). On each time step, the environment is in one of the possible states that is not directly observed. Instead, by interacting with the environment (taking actions), observations and reward (cost if the value is negative) are received from it. That is, the environment is partially observable. This action-observation-reward sequence happens in cycles indexed by $t = 1,2,3,...$. Because the environment's state is not directly observed, actions are taken under uncertainty of the true environment state. On each time step, an action is chosen from the action space including both information gathering actions and decision actions. The action selection is dependent on the history of observations and actions.

The action-observation sequence is used to update the estimate of the underlying true state (called `belief state' below) using Bayesian inference. Q-learning was then used to learn which action to do next (e.g., to gather more information or to make a decision) given the current belief state, so as to maximize the expected future reward. It does so by learning the belief-action values through simulated experience. Belief-action values are updated incrementally (learned) as reward and cost feedback is received from the interaction during the simulated experience. For example, if the model looks at cue A in a certain belief state and subsequently makes an incorrect decision, then the value of the action (look at cue A) at this belief state will decrease. With enough simulation trials, the optimal strategy will emerge and the model will take the best actions given the beliefs.

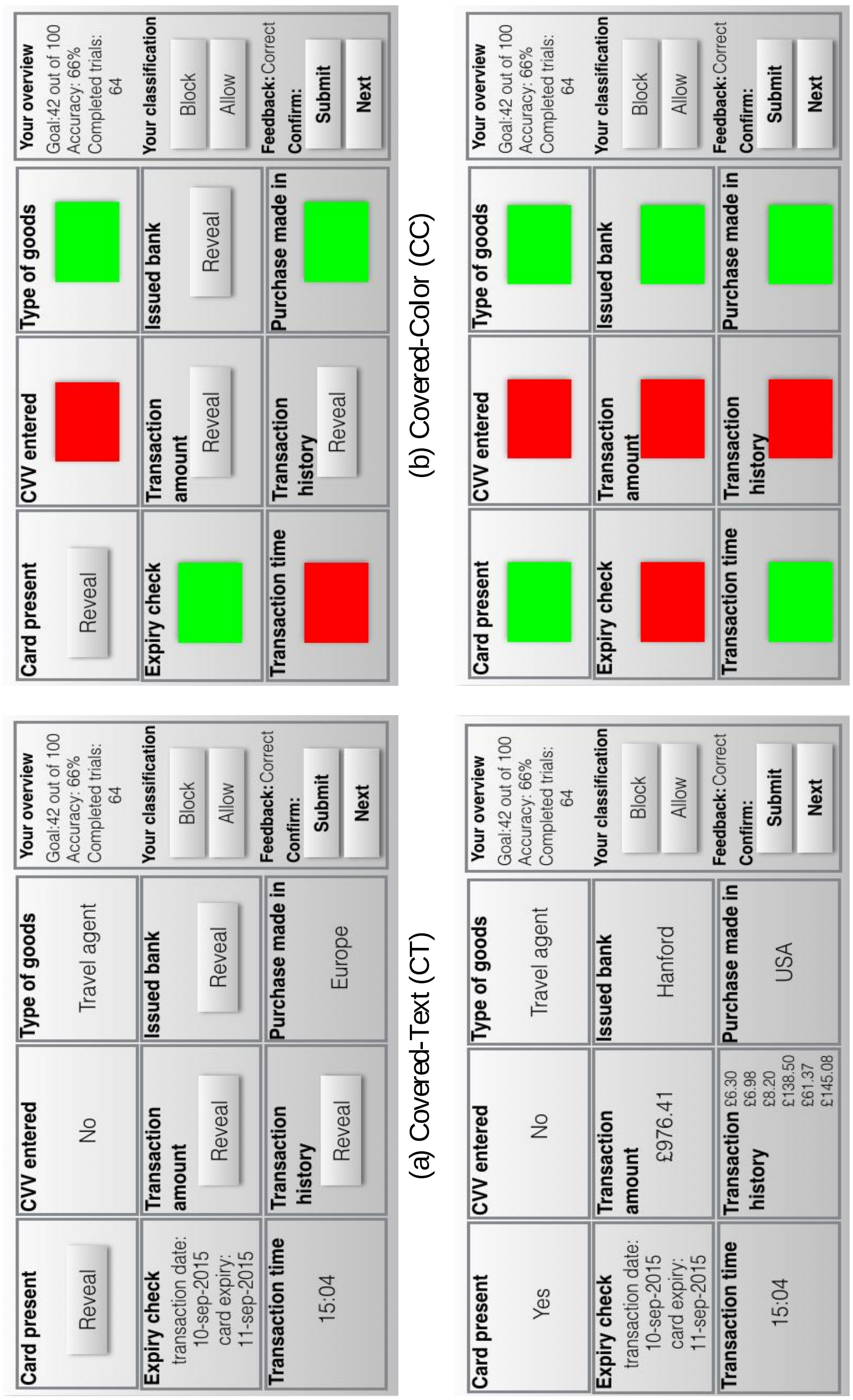
3.3.1 Experiment: Credit card fraud detection

Participants were asked to take on the role of a credit card fraud analyst at a bank. The task was to decide whether a given transaction should be blocked (prevented from being authorized) or allowed. As shown in each panel of the figure 8 information sources were laid out in a 3x3 grid. The derivation of these information sources are explained in section 3.5. For the description in this section, the key issue is the validities that were assigned to each cue. Each cue was presented in binary terms based on rules for fraudulent and non-fraudulent behaviour; rules were provided to the participants. The cues had validities [0.85, 0.70, 0.65, 0.60, 0.60, 0.60, 0.55, 0.55 and 0.55], where validity was defined as the probability that the cue indicated fraud given that the ground truth of the transaction is fraudulent. Validities were arbitrarily assigned to the nine cues and reflected the observation that high quality cues are relatively rare in many tasks. An operation panel was presented on the right side of the interface, including Block/Allow decision buttons and a feedback window. The location of each cue on the interface was assigned randomly for each participant and stayed constant across all trials. Participants were asked to complete 100 correct trials as quickly as possible. Trials in which an error was made (e.g. blocking a non-fraudulent transaction) did not reduce the number of correct trials required, so errors resulted in time costs.

The experiment was a 2x2 design. The independent factors were format and availability. (1) format had two levels: text vs. colour; (2) availability had two levels: visible vs. covered. The figure illustrates the user interface of each of the four experimental conditions.

- Covered-Text (CT) condition. The cue information was presented in covered text. In order to check each cue, the participants had to click on the associated button on each cue and wait for 1.5 seconds while a blank screen was shown.
- Covered-Colour (CC) condition: The cue information was presented by colour (green for possibly normal, red for possibly fraudulent). As with CT, the information was covered until clicked.
- Visible-Text (VT) condition: The cue information was presented in text. The information was visible immediately (no mouse-click was required).
- Visible-Colour (VC) condition: The cue information was presented in colour and no mouse-click was required to reveal it.

On each trial, participants were told to use as much information as they see fit (up to 9) by clicking the `reveal' buttons or by fixating depending on the condition they were assigned to. They then indicated their decision in the right panel by clicking the ‘Allow’ or ‘Block’ button. Following this, participants clicked on the cues that they based their decision on (for later cross-validation not reported here) and then confirmed their decision by pressing a ‘submit’ button. Participants then received feedback regarding the correctness of the decision. The next trial was shown after clicking the button ‘next transaction’. This button was placed in the top right corner of the display above the operational panel so as not to confound initial gaze data on the cues.



(a) Covered-Text (CT)

(b) Covered-Color (CC)

(c) Visible-Text (VT)

(d) Visible-Color (VC)

Figure 8: *Four interface variants for credit card fraud detection. The information cues are represented with text (left panels) or colour (right panels) and the information is either immediately available (bottom panels) or revealed by clicking the 'Reveal' buttons (top panels).*

3.4 Results

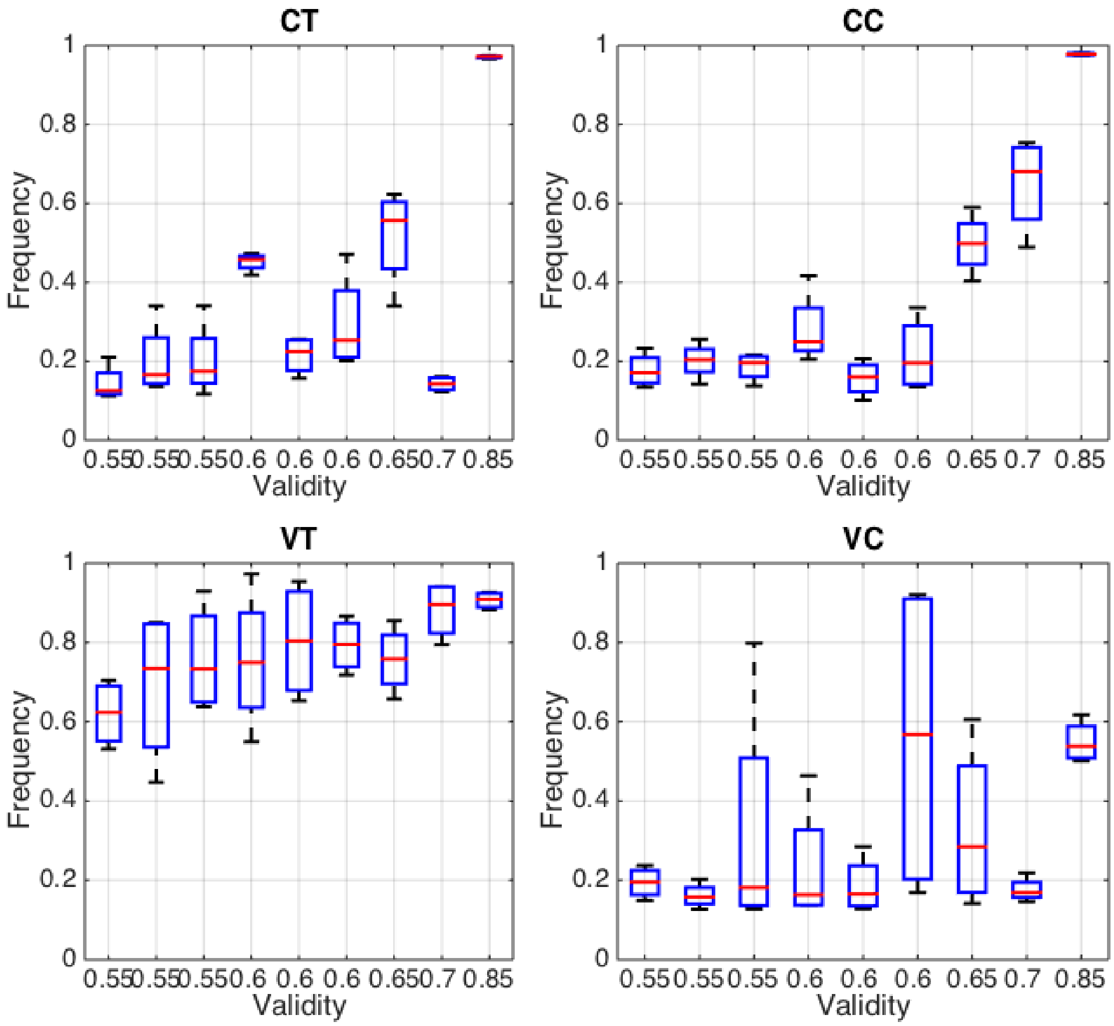


Figure 9: *The frequency that the model used each cue.*

Figure 9 shows the frequency with which the model used each cue. Cues varied in their validity and as expected, those with higher validity were used more frequently by the model. More interestingly, the model also predicts an effect of design. In the covered conditions there is a strong and consistent effect of validity, perhaps because of the high cost of information access. In the visible condition, it is the visualisation technology, rather than validity, that dominates the performance effect. With visible text there is an effect of validity but the major effect is that in this condition all cues are used more frequently than in other conditions. This is perhaps because information access cost is low and items must be directly foveated in order to be perceived. In contrast, in the visible colour condition very few items are directly foviated and there appears no effect of validity. Here, information can be gathered through peripheral vision, and so the frequency measure offers reduced sensitivity as a measure of information used.

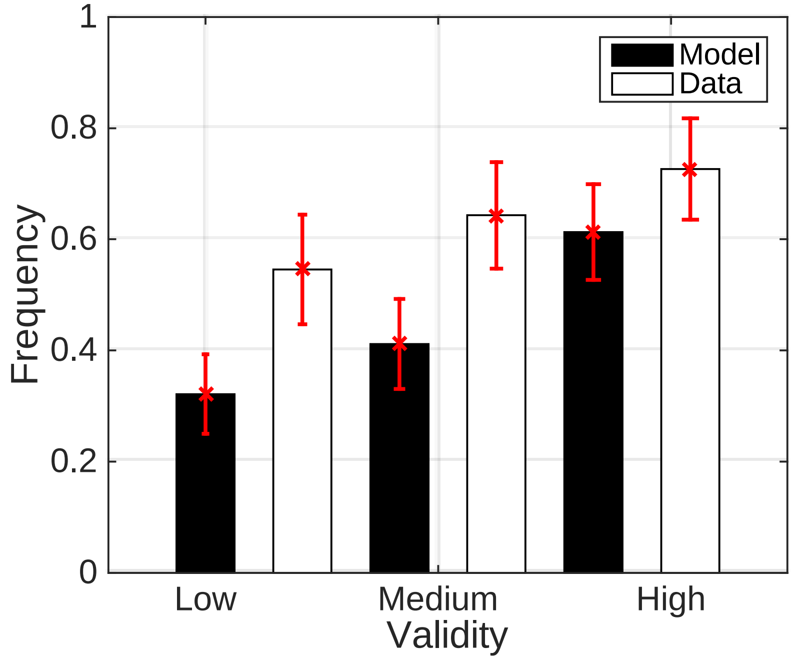


Figure 10: The frequency of cue use for three summative levels of validity. Results for model and data.

Figure 10 shows a good correspondence between the model's prediction of the number of cues that should be used and the human data. Recall that the model has not been fitted to these data and so it is offering a good prediction of the qualitative effect of validity on cue use. These results have implications for theories of compensatory heuristic decision making (Newell and Shanks, 2003) that we will explore in future work.

3.5 Discussion of Modelling Work

The fact that the model is able to predict how strategies for using information visualisations is a departure from models that are programmed with strategies so as to fit performance time. This is important because it suggests that the model might be easily developed in the future so as to rapidly evaluate the usability of a broader range of visualizations for a more extensive range of decision tasks. For example, in the near future we wish to consider multidimensional visualizations that not only make use of color, but also size, shape, grouping etc. It should be possible to increment the observation functions, for example with a shape detection capacity, and then use the learning algorithm to find new strategies for the new visualizations.

3.6 Experiments on Fraud Analysis

Data associated with credit card transactions are very rich, with often more than 70 fields of coded information associated with a transaction (Chung and Suh, 2009, Whitrow et al., 2009, Krivko, 2010), and more variables from derived and aggregated metrics. Many of the metrics are bank specific (and hence, confidential), although some can be considered as de facto standards. In order to understand decision making, we needed to reduce the amount of information in order to match comparable experimental paradigms from the cognitive psychology literature. After reviewing the literature and speaking to domain experts, we chose nine transaction attributes which are commonly examined during fraud detection and which also correlated with the case study fraud patterns chosen for SPEEDD. These nine attributes are:

- Transaction amount
- Transaction history
- Card present during purchase
- CVV entered
- Bank issuing the card
- Location where the purchase was made
- Check of the expiry date of the card
- Transaction time
- Type of goods purchased

After running several pilot experiments and analyses using eye tracking with trained fraud analysts (described in D7.3), our main laboratory studies focused on visual information use and information aggregation during decision making. We designed our main experiment around the benefits and trade-offs of data presentation, and a supplementary experiment to explore the interaction of users with highlighted information. The laboratory experiments simplified the scenario and user interface in order to generalise in context of other psychological studies, but equally we aimed to retain the general SPEEDD scenarios in order to make the work relevant to the case study domain. We hence simplified user interfaces to simple 3x3 grids, while still describing tasks to participants in context of banking and the domain of fraud analysis. Our studies were a hybrid between the task of call center staff and fraud analysts, although they were more akin to the work scenario of call center staff: participants were generally required to examine up to nine information sources relating to a transaction which – in the participant brief – was explained as having been flagged by an automated system. With this description we followed the normal first stage of the ‘blocked transaction’ workflow described above: an operator has to review the available information in order to get an understanding of the transaction specifics. However, instead of calling the customer, in our experiments the participant was then ask to make the decision whether or not the card should be blocked. This task was more akin to the work of fraud analysts, who have to integrate information to then for example predict fraud patterns or spot new trends. We chose this approach in order to gain an insight into fundamentals of decision making when working with complex data and the integration of multiple information sources.

In our main experiment (discussed in 3.2), we manipulated the data presentation approach, examining visual information sampling, correctness of decisions and decision times in four different conditions:

firstly, we were interested in differences in behaviour when presenting data as a simple colour scheme (green – possibly normal, red – possibly fraud) as compared to presenting the original data. Secondly, we were interested in how the direct availability of information affected user behaviour, comparing a condition in which information had to be revealed sequentially with a condition in which all information was in view straight away.

In our supplementary experiment, we used the same simplified 3x3 grid in which information was presented as available original data to examine how highlighting of potentially relevant information affected information sampling and integration. Here we pursued the research question whether information highlighting of relevant information sources, in context of SPEEDD for example driven by on-the-fly data mining, would change participant's sampling behaviour and improve their decisions. We designed this experiment after noting that – both in lab studies as well as in situ – participants tend to create stories and select cues accordingly, resulting in large variation in selected cues and cue weighting across participants.

Our results showed that in general, people like to look at as much information as possible. This was a trend we observed for both, experienced staff and novices. The presentation mode made hardly any difference to the performance with respect to the percentage of correctly classified transactions and learning of important cues. However, visualisations using colour often resulted in faster processing of cases. Highlighting information, even if this would lead to a 100% correct performance (unknown to the participant), was surprisingly often ignored; participants used this information to gain an insight into the 'opinion' of the automated system, but then preferred to rely on their own reasoning.

A further experiment was conducted in which the dashboards designed to support fraud analysts (described in section 4.3) were compared. The results showed that performance tends to be faster with the low-level detail presented to users, but that, in all cases, participants will (as in the previous studies) search for as much information as possible. The modeling work suggests that people will optimally search the information presented to them on the dashboard. We note that, if information is hidden (i.e., the user has to open a separate window to find this information) and the decision task is ambiguous (i.e., it is not clear which fraud pattern is definitely being shown), participants are more likely to search for the information. In other words, when the reliability of the automated decision system reduced, participants would search for more information, even when this information was available on the main display. Clearly, the ability to conduct efficient search in a well defined decision task could be compromised when the definition of this task becomes ambiguous.

3.7 Experiments on Road Traffic Management

Experiments on Road Traffic Management involved a version of the ramp-metering tasks that the SPEEDD prototype is designed to support. In these experiments, participants (either Road Traffic Controllers or students) are presented with a visual display of the road network and data concerning the activity of the ramp-metering system or the traffic on the road, and were required to determine which action would be most appropriate.

The theoretical position adopted in these studies followed the approach to joint decision making taken by Bahrami et al. (2010). In their studies, Bahrami et al. (2010) demonstrate the importance of information sharing and (more importantly) of weighting information by its reliability. In these experiments, participants were presented with a visual detection task (in which they had to spot a target against a background of distractors). For each decision, participants worked individually, then they shared the decision with another person, and then the two of them discussed the decision until they reached consensus. Their experiments, show that when two (human) decision makers have similar levels of reliability (or sensitivity) in a detection task, their combined performance is superior to that of either individual, providing they are able to communicate freely and can indicate their confidence and reliability in their own decisions. However, when either person has lower reliability, then performance is much worse than that of either individual. The model that Bahrami et al. (2010) propose assumes that the pair of decision makers are Bayes optimal and exchange (implicitly) their level of confidence in their detection decisions. This notion of optimality in decision making can be seen as a parallel to the assumptions made in the modeling work presented in sections 3.1 to 3.5). In a recent development of this approach, Koriat (2012) removed the requirement to discuss the decision, and replaced this with using the result of the most confident member of a pair makers (where confidence was measured using self-report). In this case, the initial findings of Bahrami et al. (2010) were replicated (i.e., relying on the performance of the most confident member of the pair leads to consistently superior performance), even in the absence of discussion. If the most confident member of the pair was, however, wrong, then performance deteriorates (because the least confident member accepts their partner's recommendation). This suggests that while Bahrami et al. (2010) saw their results, in part, as arising from the development of consensus through discussion, Koriat (2012) has demonstrated that the relationship between report of an answer and confidence of that person reporting the answer is key. In an interesting development of this work, Bang et al. (2014) show that the approach advocated by Koriat (2012) works well when participants are of 'nearly equal reliability' but when there are discrepancies then it is important to allow interaction. This seems to suggest that the approach taken needs to be adapted to suit differences in confidence and raises some interesting (and resolved) questions about how human participants are able to evaluate the credibility of each other's rating of confidence and how should they relate this to actual performance.

This raises interesting questions for Joint Cognitive Systems, particularly when the system is confronted with uncertain data. Assume that the pair consists of a human and an automated recommender system. If the automated system has high confidence (in this case we would use the notion of reliability as a proxy), then it is plausible to assume that the human will accept the answer it provides. Conversely, if the automated system has low reliability, then the human will rely on their own self-confidence rating and select their answer. However, we are not aware that the 'optimally interacting' research area, has considered what happens when one of a pair of decision makers is computer. If either the computer or the human partner in this decision making dyad exhibits different reliability to their partner, will joint performance deteriorate (as shown in the Bahrami et al. (2010) and the Koriat (2012) studies)? Will the human be able to notice discrepancies in reliability (either in terms of their own performance or in terms of the computer suggestions)? Will such discrepancies affect the trust that the person has in the computer, or lead to over-reliance of the person on the computer?

Experiment 1, reported in D8.3, compared performance of expert (Road Traffic Operators) and non-expert (students) on a ramp-metering task. The goal was to decide whether to agree or challenge the recommendation of the automated support. In this case, the ‘accuracy’ of the automated was set very low, i.e., only ¼ of trials contained correct information and the remainder had to be challenged. It was shown that, for this task, there was no significant difference in time to deal with an incident (typically around 5 seconds in all trials) but the expert group completed around 95% of trials correctly. For the student group, there were some students who were able to match the ‘expert’ level of performance, although the majority performed at around 75% reliability. Considering the visual sampling of displayed information (from eye-tracking recordings of participants performing the trials) we noted that there was a clear difference in strategy between expert and non-expert performance. The non-experts were not checking for display congruence, i.e., they were not looking at information which indicated whether the same ramp was being displayed in all windows on the display. In contrast, the expert group checked congruence with glances at the map and ramp displays, although also spent a large proportion of their time looking at the fundamental diagram. The suggestion was that, while all participants were presented with the same information, the non-experts were not able to judge the ‘worth’ of the displays for congruence checking and focused their attention on the automation checking aspect of the task.

It was not clear whether the problems experienced by the non-experts were solely due to lack of knowledge or due to some confusion in interpreting the rules for the task. A second experiment was conducted in which participants were required to define the rules that underlay their decision. In this experiment, reported in Morar et al. (submitted), we also compared performance when the (simulated) automated support was operating at reliability levels of 25% of 81%. We also noted that the ability to apply the rules helped participants, suggesting that it is potential beneficial to provide opportunities for humans to either read the rules that the automation uses or to reflect of the rules that they are using themselves. Results show that the reliability level had a clear impact on the user’s behavior: at low reliability, the humans would ignore any automation suggestion and rely on their own decision making, and at high reliability, the humans tended to accept the automation suggestion (even if this was incorrect).

In a third experiment, we compared performance on a ramp-metering task using the first and second SPEEDD prototypes. There were clear advantages to using the second SPEEDD prototype, in terms of decision time and accuracy, and (as in study 2), there was an effect of system reliability. When system reliability was high (80%) then average decision time was much faster than when it was either low (25%) or medium (50%). Participants were able to compensate for system reliability, i.e., they would be more inclined to disagree with the automated system when it had low reliability and more inclined to agree when it had high reliability. While this finding might appear self-evident, it is worth noting that participants had no indication as to the system’s reliability (which was changed over the course the experiment).

4. Real-Time Visualization for Human Decision-Making Visual Analytics and Ecological Interface Design

4.1 Introduction

Visual Analytics is not simply the visualization of the output from analysis processes, but the creation of insight in the decision-makers working with these visualizations. One approach to designing visual interfaces for decision-making is through the processes which underlie Ecological Interface Design. 'Ecological Interfaces' are designed to visualize the manner in which physical components of the system map on to the (more abstract) functions that the system performs. So, they are views of the process which are not simply maps of how physical components connect to each other but are abstractions which show how types of physical components affect particular functions. The purpose of such designs is to improve operator decision-making and diagnosis when dealing with faults relating to those specific functions. For SPEEDD, this means that the visualization will not only display the model's input and output, but also the relationships between elements in the decision space. Ecological Interface Design (EID) requires an understanding of the nature of the cognitive work performed by people in the domain under consideration. One element of EID is simply the reflection of the constraints in the work domain through constraints in the user interface. In this way, the 'ecology' of the work domain becomes reflected in the user interface through the definition and management of these constraints.

The user interfaces (dashboards) for the two use-cases underwent very similar development processes which involved a study of the work environments (when possible) and analysis of the tasks that are completed on a daily basis along with the procedures for completing them. The design of the dashboards also took into consideration the requirements and limitations of the underlying technical systems that support these tasks and the research on human visual perception and decision strategies in the modeling and experimental work. These three factors, more specifically, the organisational, technical and perceptual characteristics have guided the development of the interfaces and informed different aspects of it. Figure 11 illustrates how the social and technological environments constrain and define requirements for the content, format and form aspects of the interface. Content refers to the information that needs to be displayed on the screen, format relates to the way in which interaction with the information and control of the system in question is achieved, and finally, form is the manner in which the information is displayed to the user.

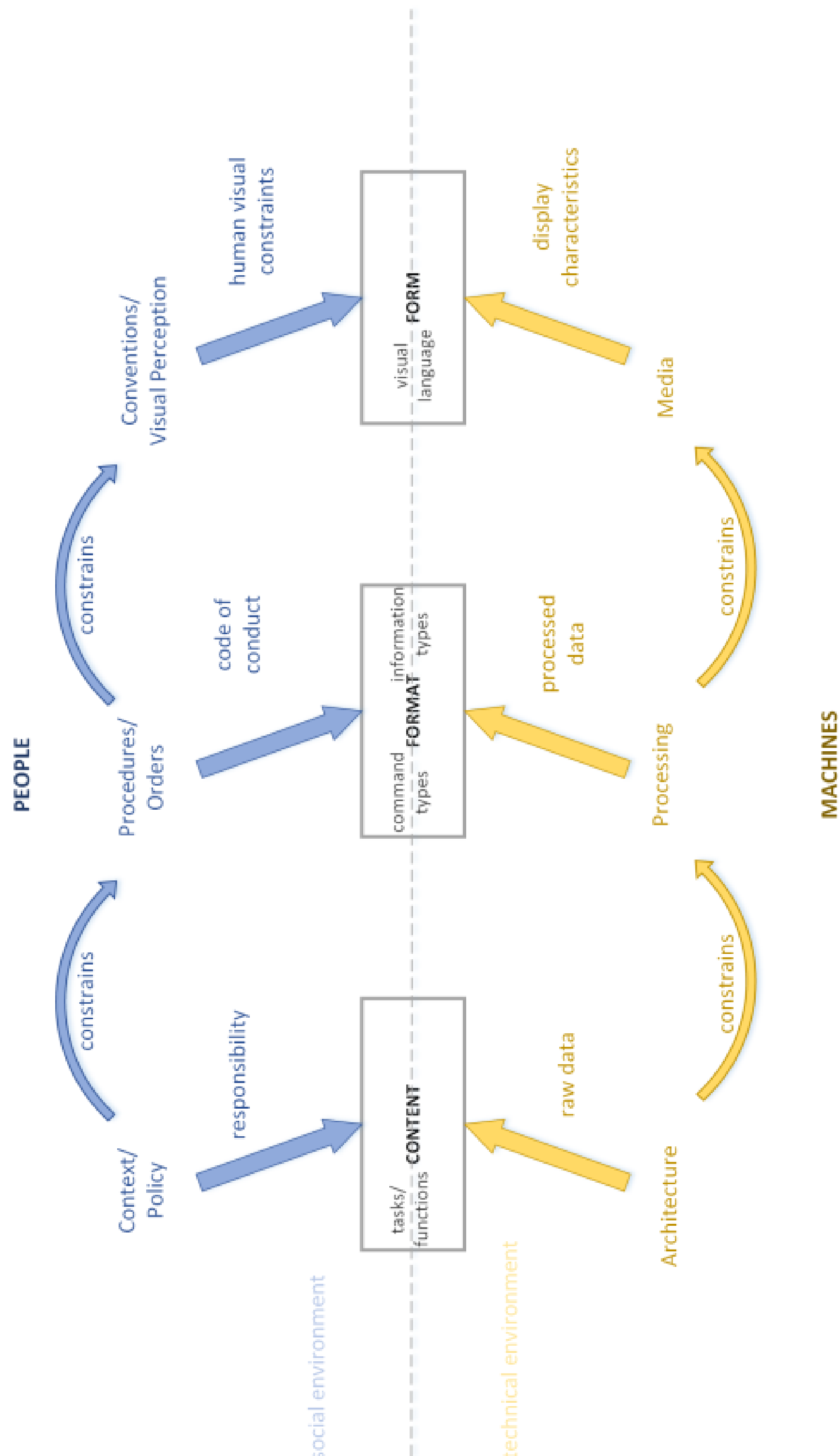

PEOPLE
Context/ Policy
constrains
Procedures/ Orders
constrains
Conventions/ Visual Perception
responsibility
code of conduct
human visual constraints
social environment
tasks/ functions
CONTENT
command types
FORMAT
information types
visual language
FORM
technical environment
raw data
processed data
display characteristics
Architecture
constrains
Processing
constrains
Media
MACHINES

Figure 11: Sociotechnical constraints on user interface design

Traditionally, the design of user interfaces is done by technical teams as a terminal to the automation developed for a specific user interface (Few, 2006). The danger with such an approach is that it assumes that the human operator will be sufficiently flexible and adaptable to be able to modify familiar and well-practiced patterns of working to cope with whatever the automation presents. However, this can result in a mismatch between the users' understanding of their work (e.g., in terms of the 'whole task' discussed in section 2.1), which in terms can either lead to confusion and frustration or to a reluctance (or refusal) to interact with the automation. The SPEEDD dashboards have undergone an incremental design process in which both, social and technological aspects of the work environment have been taken into consideration and have informed the final prototype designs, presented in this deliverable. The following subsections illustrates how this design process was achieved, presenting a history of the dashboards for both, Traffic Management and Credit Card Fraud Investigation, along with the corresponding inputs that guided and informed their design.

4.2 Road Traffic Management Dashboard Design

The design process began with a study of the work environment which provided us with an understanding of tasks traffic operators deal with on a daily basis, the available resources and usual procedures they follow, which determined the informational requirements of the visual display, or more specifically, its content. Deliverable D8.1 shows how this study was conducted and describes the data gathering process while D5.1 show how these data were used to gain an understanding of their informational requirements. Finally, a schematic user interface layout was produced (see Figure 12).

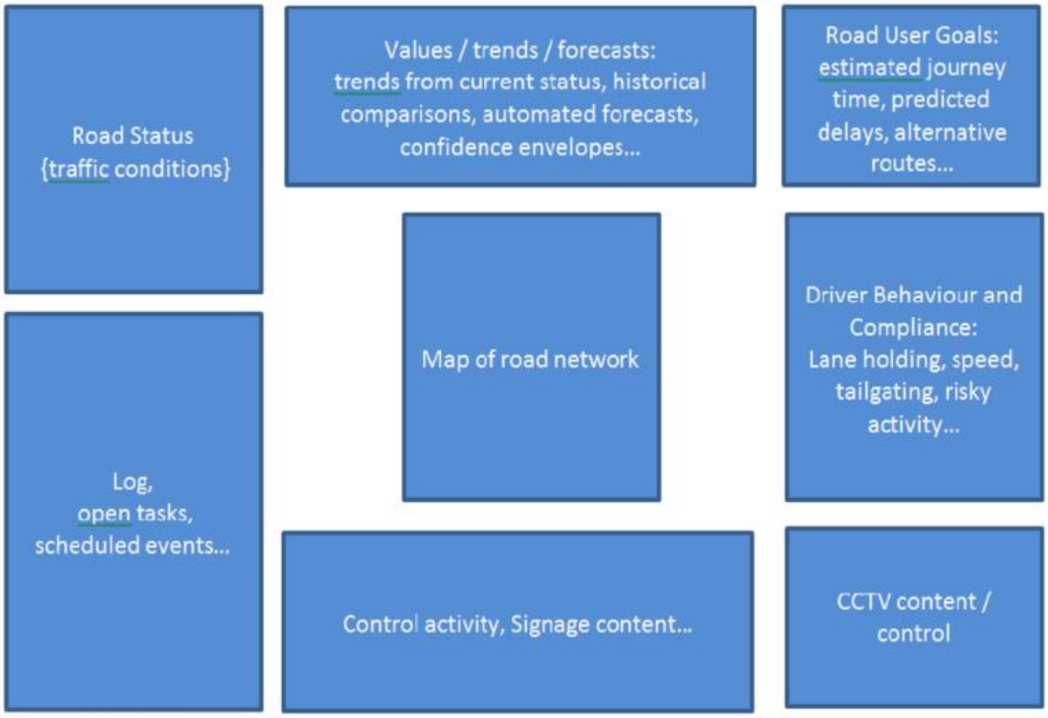


Figure 12: Schematic User Interface Layout

One would assume that transferring this conceptual design to an actual working prototype is straight-forward and implies the mere conversion of the individual boxes in figure 12 to visual components. After the analysis of the work domain is achieved, information requirements are defined and then a user interface is designed according to the identified requirements. However, for the actual implementation of such a user interface, its goals and functions need to be aligned with the technological aspects of this issue. Work needs to be done in order to ensure compatibility and congruity between the user interface elements and the underlying architecture that supports it.

In SPEEDD, the translation of the schematic UI to an initial design (figure 13) was guided by two important technical considerations: there was no data to support the Road User Goals window and the infrastructure did not provide a means to store the user's interaction with the system (open tasks, scheduled events, etc.). Moreover, the only prediction that the underlying system was able to make at this initial stage was whether a congestion was likely to occur. It was decided that this would be shown on the main road map along with the detected congestion event. The road status is indicated by a historical traffic flow vs density graph at each sensor location, selectable from the main road map. Also present on the main road map are the individual ramp metering, Variable Message Signs and lane closure control points. When one is selected, its current status and options for future control are shown in the control panel window. The video feed provides live CCTV footage at from a physical location selected on the map. Because there are already existing systems in Traffic control rooms, this is implemented in the SPEEDD prototype as a mock feed using, Google Street View API[1]. Finally, Driver Behaviour and Compliance window shows average speed of traffic participants and average distance between drivers. The reason for having only these two metrics was again driven by data availability, their source in this case being the sensors in the road. This shows how the technological environment shaped and constrained the content of the dashboard further from the initial requirements and constraints identified as a result of the work domain analysis.

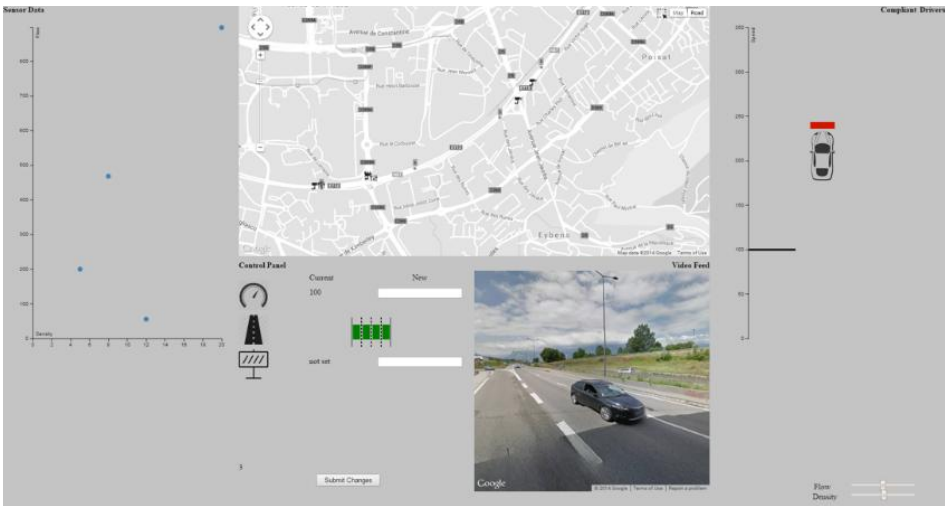


Figure 13:Initial version of the Traffic Management UI

The dashboard in figure 13 evolved into the version shown in figure 14 as a result of addressing accessibility issues regarding the selection of individual ramps and improvements of the Compliant Drivers visualisation by differentiating between north- and south-bound traffic behaviour. An additional technical requirement of displaying automated suggestions to the operator arose and is shown in the Suggested Activities window.

[1] https://developers.google.com/maps/documentation/javascript/streetview

Figure 14:Improved initial version

The design in figure 14 was presented the Traffic Control Room (DIRCE Grenoble) for the purpose of evaluation and to gather further input from work-domain experts. They identified two areas where improvements could be made: i) logging is a primary activity that they perform and the presence of a log is mandatory, and ii) their work does not involve the direct control of road users' behaviour, this being achieved by long-term governmental campaigns. Issue i) was addressed in the next iteration of the dashboard design by the addition of the Activity Log window, while ii) lead to the removal of the Compliant Drivers window. Both questions link to the content of the display.

Returning to the technical perspective, additional constraints on the content and format of the dashboard arose. First, SPEEDD deals primarily with the automatic control of ramp metering rates for the purpose of mitigating and potentially preventing traffic congestions. Therefore, the control of variable signage and lane closures is not something that is dealt with by the system and user interaction with automatic ramp metering control unit is expected to be minimum. This lead to a simplified control panel window where the user can set bounds for the ramp metering control unit and not absolute values. The content was modified by the exclusion lane closures and variable signage controls. Format was modified by the change in how the user interacts with the automatic ramp metering unit. The form of the window is kept the same, by preserving the text box inputs for setting the bounds.

Secondly, the outputs of the automated system are in the form of ramp metering levels to be applied at each particular ramp and not in the form of suggestions of what actions the user should perform. Therefore, the Suggested Actions window has been removed. This marks a further change in content.

An in-house usability analysis of the individual display components revealed the following characteristics. In terms of format, it has been identified that the raw sensor data is not spatially liked to the map. In order to see the raw data gathered by a specific ramp, the user needs to either know the physical location of the ramp in question and select it on the main map window, or know its number and select it from the ramp metering window. However, once a ramp is selected, there is no indication of its selection on the main map. This was solved by displaying the ramp numbers along with the icons at their corresponding physical locations on the map.

Another issue that relates to format was the fact that there is no indication of CCTV feed source once it was selected, therefore, it required the end-user to memorise its physical location. The video feed was not linked to the ramp metering in its vicinity either. This issue has been addressed by moving the video feed on the main map and displaying it on a ramp icon click.

The changes previously discussed are illustrated in the subsequent iteration of the display design and can be seen in figure 15. More information on the user evaluation of this and the previous prototype version can be found in D8.3.

Figure 5 – Second Version of the Traffic UI

The design shown in figure 15 has undergone further evaluation, both from the perspective of the underlying technical system and from the DIRCE traffic operators. From the discussions with traffic managers and the dashboard demonstrations that were run it has been concluded that the congestion event is too distracting. Initially when a congestion alert was triggered, a circle would appear on the main map and the map would automatically pan to its physical location. This resulted in a loss of the overall view of the system and was deemed very disruptive to the tasks that the operators could have been engaged in prior to the appearance of the event. Also mode of displaying congestion was designed under the hidden assumption that it would have the highest priority and this is not actually the case. The information that the operators provided us lead to the following change in display format: circles continued to be the means of displaying a congestion event, however, the map view would no longer pan to its location.

Traffic operators made an additional comment regarding the overview of the state of the road. While there was an indication of the ramp metering rates of all the inbound ramps (colours of the squares in the Ramp Metering window), there was no indication of ramp queue lengths. This lead to the addition of the Ramps-Quickview window, which displayed the values of both ramp rates and queue lengths. This change related to both, the content (the addition of queue lengths) and the form (the means of displaying ramp rates) of the display. The graph in the top-left of the display proved hard to read and a continuous view of the historical data was deemed unnecessary, the instantaneous view in the Ramps-Quickview window being sufficient for the task. This lead to a further change in display content, i.e. the removal of the Sensor Data window.

A complete redesign of the Log window was motivated by a new requirement from the technical environment to display more information for each automatic detection and control event. This marked a change in both the content and the form of the Log window, which was renamed Event List. It has been decided that the request to illustrate the performance of the automatic decision making unit has been addressed by the, now, dual function of the Ramps-Quickview window.

A more in-depth analysis of how the operators' jobs could be changed by the addition of an automatic ramp metering system revealed that users would not be expected to constantly correct and contribute to the computer's actions. Therefore, it was decided that the 'trimming' of ramp metering rate bounds should be moved from the main dashboard to a pop-up dialog, thus modifying the format of the Ramps window and leading to the removal of the Control Panel from the main dashboard.

The dashboard which resulted from applying the changes discussed above can be seen in figure 16.

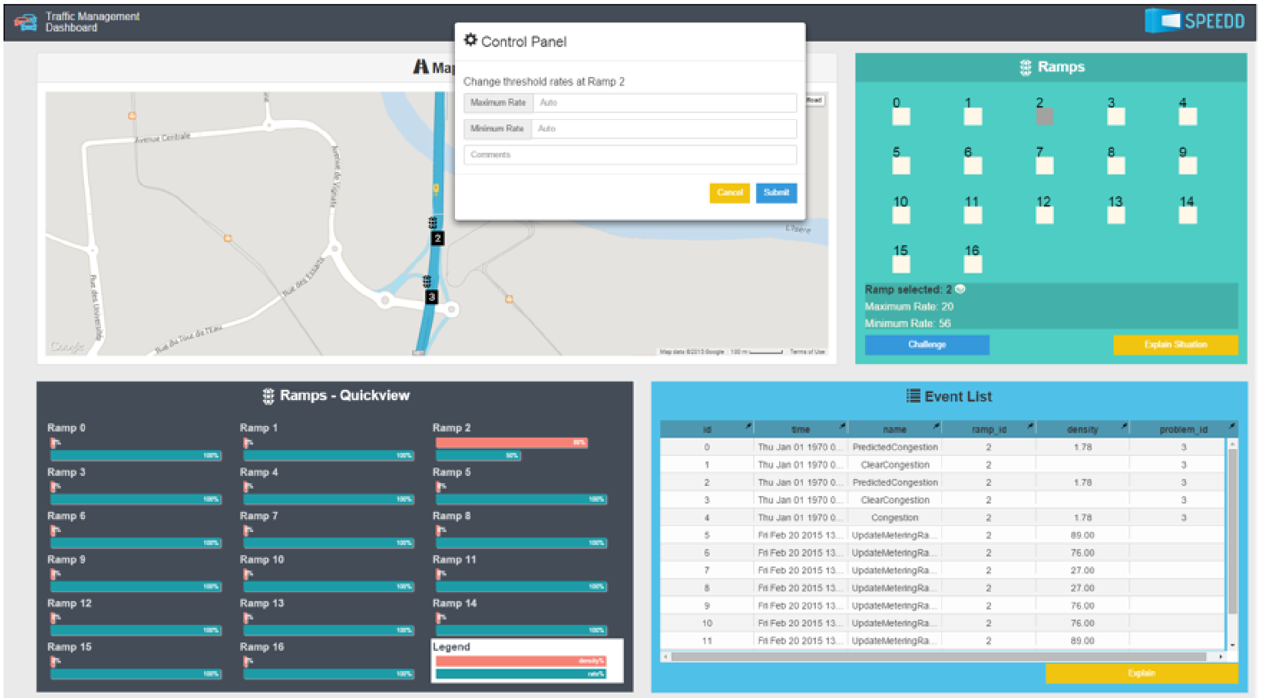


Figure 15: Third Version of the traffic UI

The user interface went through a major redesign process informed by principles of Ecological Interface Design and previous operator comments, while taking into account the technical aspects of the problem. The output of this design process can be seen in figure 16 and a detailed account of how this version was arrived to is shown in D5.2.

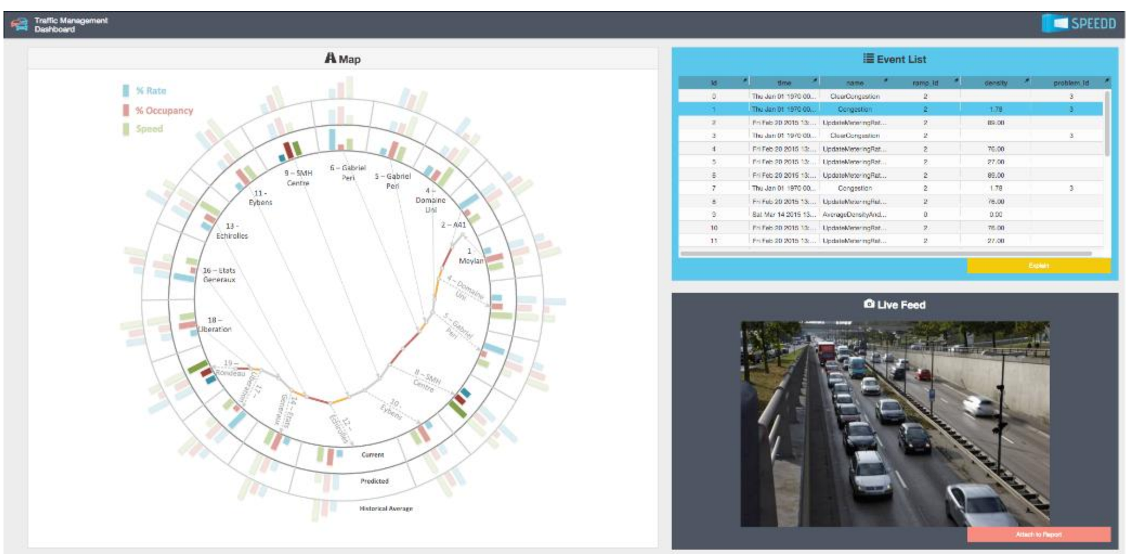


Figure 16: Fourth Version of the Traffic UI

One significant change from the previous design shown in figure 16 is the integration of multiple information sources into one view. The map along with the Ramps and Ramps-Quickview window have been replaced by the circular display shown in the left hand side of the figure. The schematic road is split into segments at the locations of inbound and outbound ramps. Ramps are represented by nodes (circles). Each node is linked by a thin arrow to a set of bars on the outer circle, the direction of the arrow indicating whether the node represents an inbound or outbound ramp.

In addition to these modifications of format, further changes of content and form can be identified. In terms of content, from our discussions with DIRCE operators and further analysis of their work it has become clear that they make extensive use of the CCTV panels for most of their tasks and, therefore, it was decided that their work would be better supported by the addition of a bespoke CCTV feed window on the main dashboard rather than having it in a separate window that requires additional actions to access. They have also pointed out that an indication of average traffic speed would complement the overview of the road status. This latter change was made possible by the fact that data regarding traffic speed was already being exchanged as part of the architecture messaging bus, thus being in line with the technological goals and requirements.

Several changes in format can be seen in this redesign of the Traffic Management dashboard. First, notifying the operator about a congestion can no longer be done by displaying a circle on the map as this is no longer a salient feature, its discernibility from other map element being reduced, therefore reducing its effectiveness as an alert. Road occupancy is now shown by the colour of the road segment, red signifying high density, yellow – medium, while grey showing normal to low levels of density. Secondly, ramp rates and occupancy bars are now linked to a physical location on the map, allowing for global patterns to be spotted. The text representing actual values of the measures were removed as operators have pointed out that they rarely need to know precise values and they are more interested in the state ramps are in (e.g. low throughput, high occupancy, etc.).

When bringing this design to traffic operators for the purpose of evaluation, they have pointed out that the circular layout and the radial placement of the ramp status bars made the visualisation very difficult read (see figure 17). Comparing this version of the User Interface with the previous one, we noticed that replacing the initial map with a schematic of the road results in some loss of context. To be more specific, the highways and city roads that connect to the Grenoble Ring road are no longer shown. It has been decided to the map on the background of the schematic. The issues of visualisation legibility and context were addressed in the next version which can be seen in figure 18. The form of the congestion event was changed, improving its salience by increasing segment thickness in addition to colouring it red.

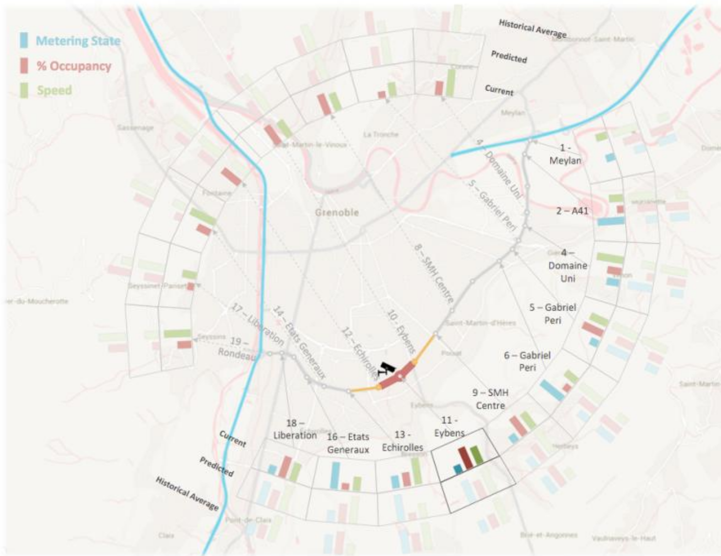


Figure 17:Final Map Visualisation

Figure 18 shows the final prototype for the Traffic Management dashboard. In addition to the changes discussed above, the content of the Live Feed window was changed by adding cycling views from other parts of the road network, thus contributing to an increase in the overall situation awareness of operators.

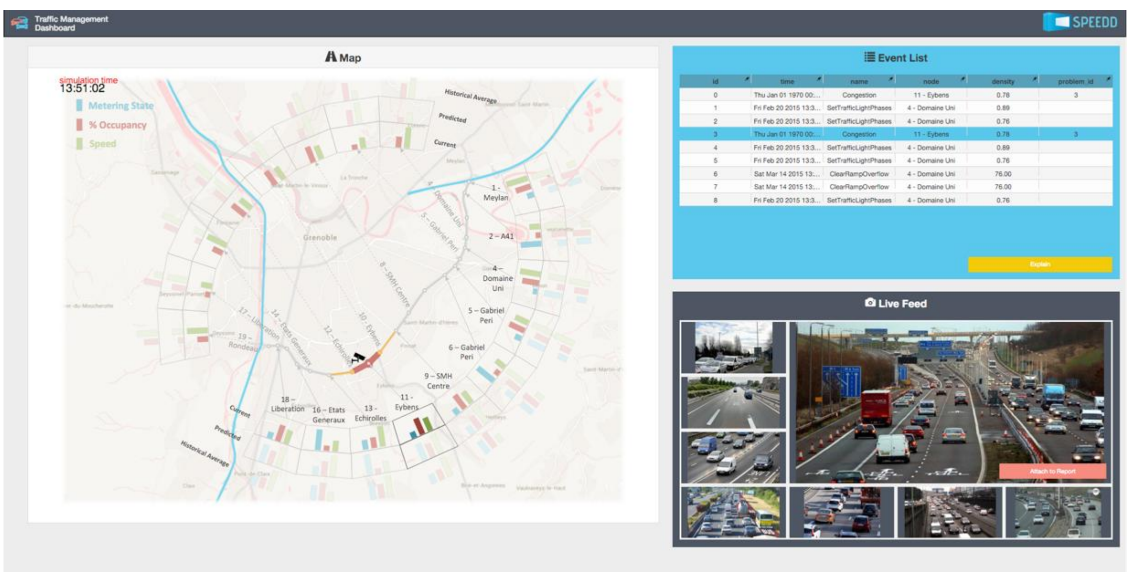


Figure 18: Final Dashboard for Traffic Management

As discussed in section 3.7, experimental evaluation comparing working prototype 1 (figure 15) and working prototype 2 (figure 18) showed significant gains in operator performance when using prototype 2 and a significantly higher usability rating. This evaluation is discussed in D8.5.

4.3 Credit Card Fraud Dashboard Design

A similar process was followed for the design of the Credit Card Fraud use-case dashboard. We began with the analysis of the work environment. Due to the high-security of the domain, however, this was limited to research into publicly available information about fraud and fraud detection and discussions with several financial institutions. The analysis was complemented by access to Feedzai experts which evaluated and tested the developed designs. Cognitive work analysis was performed and a list of information requirements along with an initial layout for the dashboard was produced as a result (see figure 19). The drives and requirements of both the technical and social (work-domain) aspects and their influence on the content, format and form of the display will be discussed.

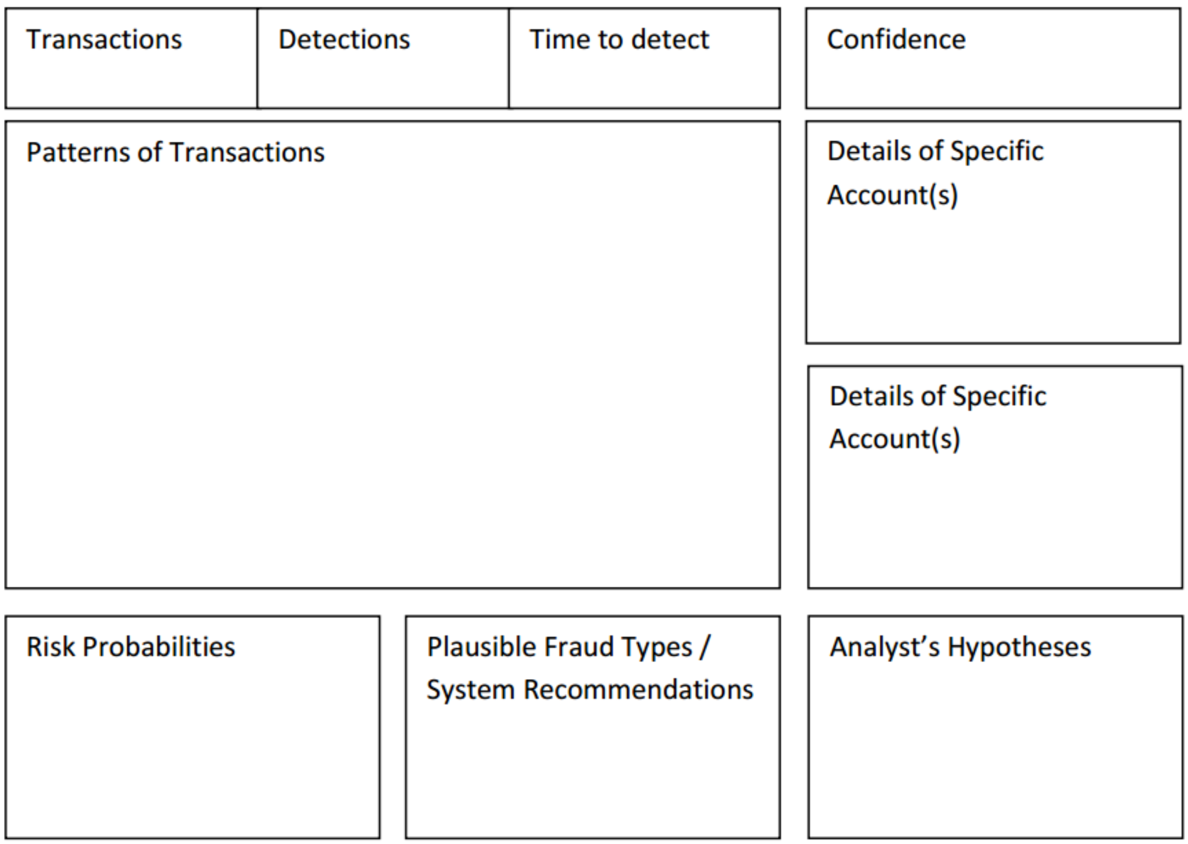


Figure 19: Initial Layout

The layout presented in figure 19 was directly translated to a first user interface design (figure 20) with the following changes in content, driven by the description of the technical architecture. First, time to detect refers to the time to spot a fraud pattern and this is an automatic process with very little variation within an interval. This variation is influenced by the rate of events, number of events in the pipeline and the type of fraud that is being investigated. Average time for analysts to detect a fraud pattern is not computed as part of the architecture and is therefore, left out. Second, the Risk Probabilities windows has also been removed. This is due to the fact that financial institutions which use fraud detection software differ in the way the calculate and manage risk. It is usually set at an organisational level and does not tend to vary within the organisation. Therefore, the window would have ended up displaying the rules or guidelines of dealing with potentially fraudulent transactions of the respective organisation. Traditionally these rules and guidelines are incorporated in the computation of the fraud score, or certainty associated with the pattern detected. The third content change is the removal of the account associated information windows. This is due to the fact data analysed in the project is anonymised and the event processing unit does not deal with individual accounts.

Figure 20: Initial Fraud UI

In figure 20, the top rectangles contain information on average transaction amounts and volumes, in addition to the total transactions investigated and total flagged transactions. When clicking on one of the metrics, the treemap below reacts, updating the area of each shape (which represents a country) in proportion to the selected metric. Analyst’s Hypothesis is renamed to Analysts’ Workspace. This window allows for analysts to submit new patterns for the automated system to investigate. New patterns are constructed using a logic builder model. The Other Info window at the bottom is meant to show external consumer statistics which could help the analyst make a more informed decision about the course of action to adopt in response to the flagged pattern.

Further analysis of this user interface version brought up a number of potential improvements to be made. Taking the technical perspective, the architecture is set up in such a way that each update of rules (fraud patterns to investigate) requires a redeployment of the runtime. Therefore, the addition of extra fraud rule or the modification of existing ones is not something that the architecture supports. Functionality such as that of the Analysts’ Workspace needs to be taken offline. This consideration resulted in the removal of the Analysts’ Workspace, which thus rendered the Patterns Investigated window (which lists the patterns currently investigated by the automation) redundant. These updates mark changes in display content.

Feedback from domain experts (reported in D7.2) led to a series of further changes to the display content, format and form. In terms of content, the discussions we had with financial institutions revealed that analysts use a device other than the main one (which displays the user interface) to access extra information, such as news, market statistics, etc. This lead to the removal of the Other Info window for the next design iteration. The Flagged Transactions Queue proved hard to read, the order being pre-set in terms of flagging time without the possibility of reordering patterns. The previous comment, coupled with the requirement of transparency of the automated flagging system led to a redesign of this area of the dashboard. The Flagged Transactions Queue was replaced with the event list. This is an ordered list which shows some of the more important information without the need of selecting a particular transaction. It allows the user to arrange the list in terms of any of the shown attributes. More information can be accessed by selecting a transaction and clicking on the Explain button. This brings up a modal window with all the attributes of the flagged pattern in question.

In terms of form, the treemap in figure 20 proved difficult to read and in the subsequent version, it was replaced with a contiguous cartogram. In this way the update model of the treemap (changes in area of the rectangles) is preserved by having the area of the countries get bigger and smaller depending on the selected statistics. A secondary change that resulted from this replacement was an improvement in geographical fidelity, making it easier for analysts to find a specific country. An additional change in content was the inclusion of an Explore button on the map, which brings up more detailed information about a selected region.

The changes discussed above can be seen in figure 21.

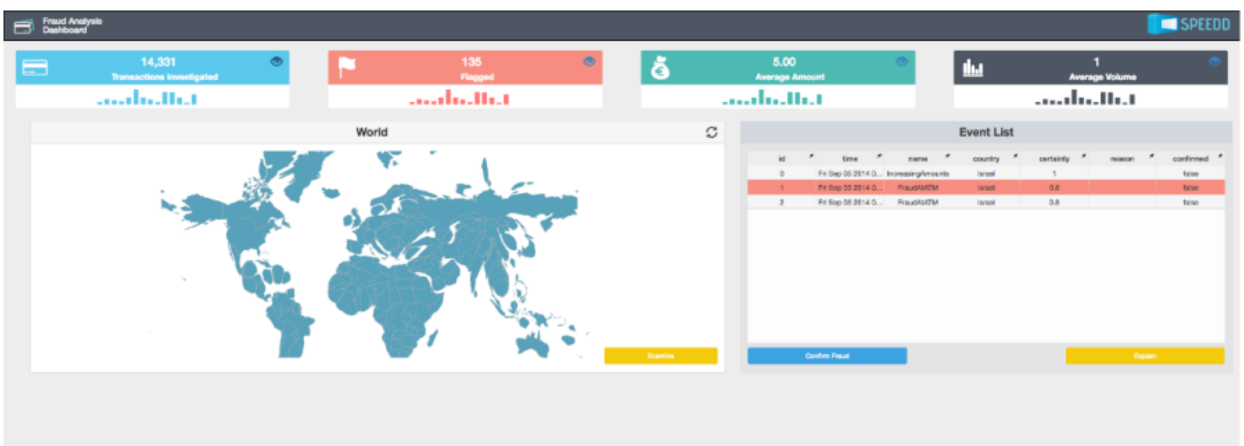


Figure 21: Second Version of the Fraud UI

In order to provide the analyst with a better understanding of how risk score or certainty is computed, in other words, what is the computer's 'rationale' for flagging a particular set of transactions as potentially fraudulent, we decided to find a way to display some of the automation reasoning alongside a flagged event. This was achieved by adding an Account History window and a visualisation of the contribution of transaction attributes to the final risk score. These changes mark refer to the content of the display and can be seen in figure 22. Colours from the account history relate to those in the Transaction – Score Contribution window, while the height of the graph (themed river) shows the risk level. The Transaction –

Score Contribution window is a graphical representation of the contribution to the total risk score of each attribute for the pattern selected in the Event List window.

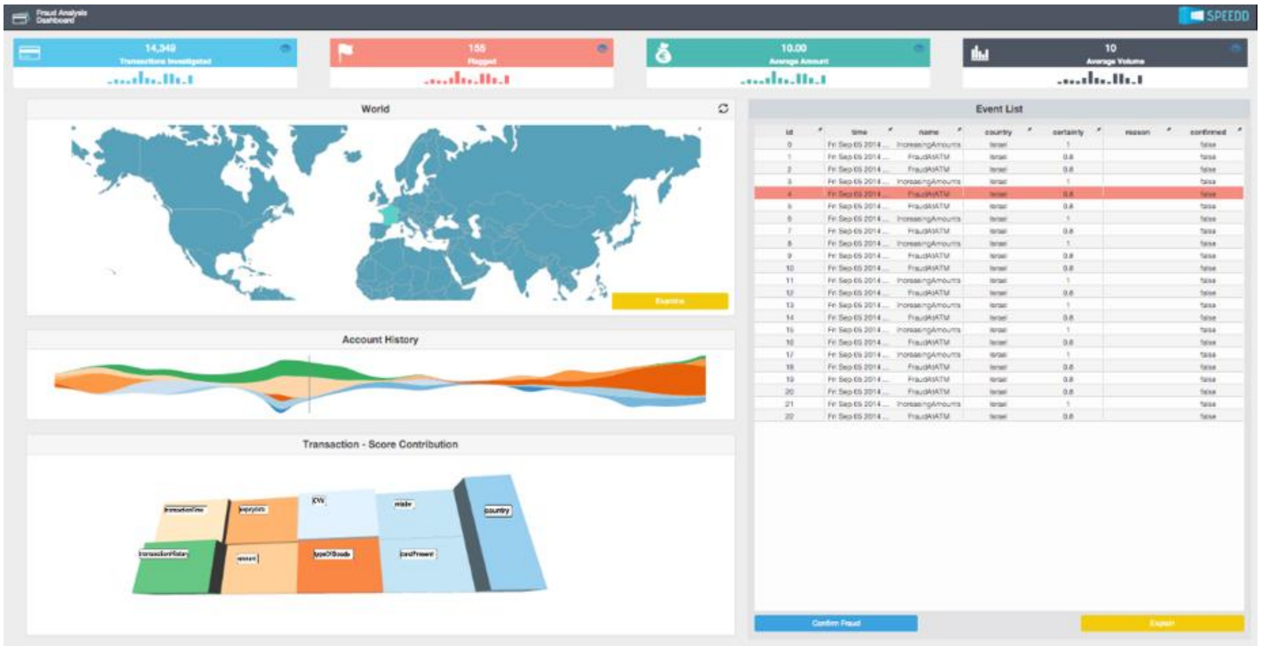


Figure 22: Third Version of the Fraud UI

While domain experts liked the idea of this information being made available in the process of decision-making, decoding the information from the themed river proved problematic due to the lack of reference levels (thresholds) on the graph. This view, along with the score contribution window have been abandoned due to technical constraints relating to data availability and due to the fact that in the way patterns were defined at the system level at that time implied a constant risk visualisation for each fraud pattern, rendering the visualisation no more useful than the pattern name. Another comment that analysts have made was that differences in country areas was very difficult to quantify, and the distortion of the map that resulted from the countries changing areas made it very difficult to recognise and differentiate between countries. It was decided to replace the contiguous cartogram with a choropleth map.

As discussed in section 2.3, there are multiple analyst roles within an organisation and the tasks they are responsible of differs. Two roles have been identified that could be supported using the SPEEDD dashboard: i) high-level supervisor, and ii) low-level call centre analyst, each requiring different support. The user interface in figure 23 presents an overview of all transactions analysed by the automatic risk scoring system and illustrates overall statistics for each geographical region. Individual flagged patterns are also shown with the possibility to further investigate them, bringing up additional available information. This user interface was designed for the supervisor role. The dashboard in figure 24 is a flagged pattern investigation tool designed with the call-centre analysts in mind. The window on the left shows the flagged patterns that have been assigned to the analyst, with the option to bring up more information. The window on the right provides the user with an integrated visualisation of the most relevant aspects of the pattern which triggered the flagging by the automated system.

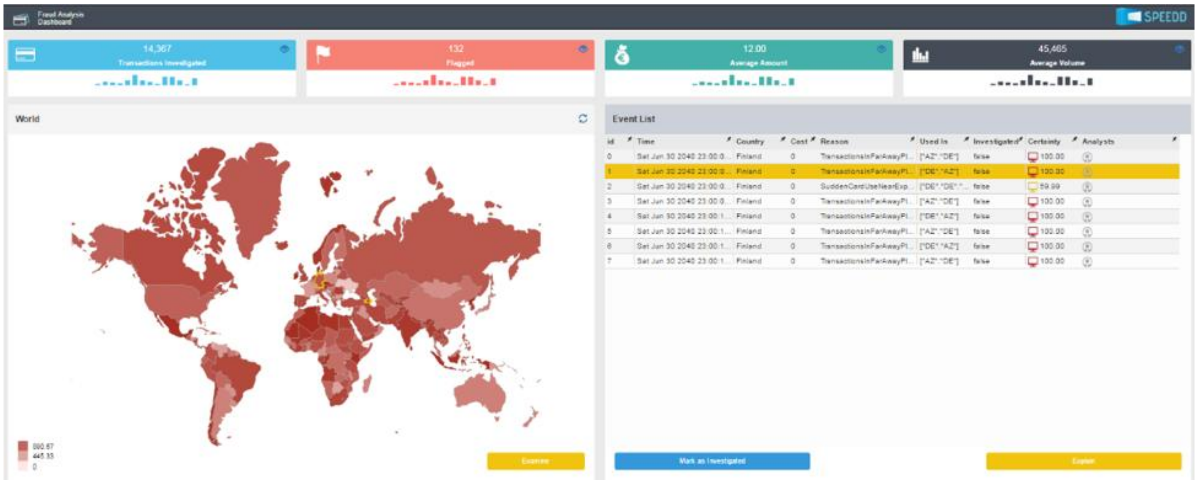


Figure 23: High-level fraud dashboard

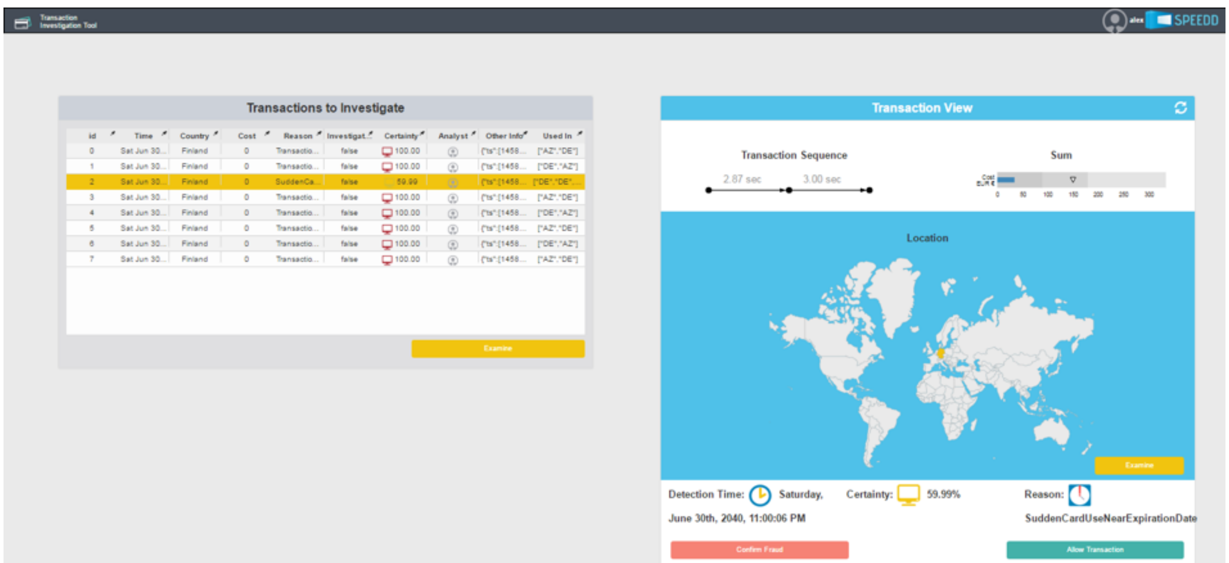


Figure 24: Low-level fraud dashboard

5. Conclusions

This report has summarized the work completed to date in Work Package 5. At present there are further experiments being conducted, using eye-tracking to explore search in the main fraud experimental task, and refinements to the model. The dashboards will undergo slight modification prior to the final review (to accommodate changes to the primary tasks in the use cases) but will otherwise retain the look and feel that they have now. Experimental evaluation of the two dashboards (comparing first and second prototype designs) have shown measurable improvements in the ability of users to make decisions based on the presentation of information.

Experimental studies and modeling have extended our understanding of our people use dashboards to make decisions. We have modified our designs on the basis of these results and have a set of papers being submitted to key journals in Human-Computer Interaction and Ergonomics as the project comes to a close.

5.1 Advancing State of the Art

T5.1: Modeling Decision-Making as a Socio-Technical Activity

The main objective of task 5.1 was to apply Cognitive Work Analysis (CWA) in the use-case domains to understand Human Decision-Making and then to *advance the state-of-the-art by* developing Cognitive Work Analysis for the study of human-automation collaboration in proactive decision-making. This has meant describing decision making terms of Joint Cognitive Systems and the insights from this description has underpinned the Sociotechnical design process for the design of dashboards in SPEEDD. As section 4 shows, the design process has not only been informed by our understanding of decision making but also the balance between social, technical and perceptual constraints. This work has been presented at the Human Factors and Ergonomics Society and EP for DM conferences.

T5.2: Defining Objective Metrics for Evaluating Decision-Making

In terms of defining metrics for decision making, Work Package 5 has used a combination of POMDP modelling, eye-tracking data and experiments to compare and contrast the effect of different approaches to information display on decision making. Together the results from these studies inform our understanding of the uses of visual analytics and, in particular, have allowed us to advance the state-of-the-art by developing an understanding of how human decision makers allocate their attention to different features in a decision problem and how these features contribute to satisficing in decision-making (both individually or in teams). The field-based eye-tracking work was presented at the Institute of Ergonomics and Human Factors conferences (where it was awarded best paper) and is currently under review for the journal Applied Ergonomics. The initial concept for the modelling work was presented at the ACM CHI conference and a follow-up paper has been submitted to the same conference.

T5.3: Real-Time Visualization for Human Decision-Making Visual Analytics and Ecological Interface Design

The primary aim of this task was the design and development of a visual analytics suite for real-time explanation of, and interaction with the Big Data in the use-cases. Here, we assume that the key challenge lies in displaying the situation space to the user in such a way that they can clearly discern the decision space available to them. In this way, the user interface is intended to support proactive decision support. For this task, the aim was to *advance the state-of-the-art by* applying concepts and principles from Ecological Interface Design to the development of visual analytics for decision support in Big Data applications. The primary goal of this task is to capture the constraints which influence the pursuit of the goals of human operators and to develop novel techniques for representing multiple information sources in terms of converged visualizations which will support decision-making and understanding of the proactive system recommendations. The design process is currently being written up for submission to the International Journal of Human-Computer Studies, and evaluations of the dashboards are being submitted to IEEE Transactions on Human-Machine Systems.

6. References

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G. and Frith, C. D. (2010) Optimally Interacting Minds, *Science*, *329*, 1081–1085.

Bainbridge, L., (1983) Ironies of automation, *Automatica*, *19*, 775-779

Bang, D., Fusaroli, R., Tylén, K. Olsen, K., Latham, P. E., Lau, J. Y. F., Roepstorff, A., Rees, G., Frith, C. D. and Bahrami, B. (2014) Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making, *Consciousness and Cognition*, *26*, 13–23.

Butko, N. J. and Movellan, J. R. (2008) I-POMDP: An infomax model of eye movement, In *2008 IEEE 7th International Conference on Development and Learning*, ICDL, 139–144.

Chen, X,. Starke, S., Baber, C., Howes, A. (submitted). A Cognitive Model of How People Make Decisions Through Interaction with Visual Displays.

Cherns, A. (1976) The principles of sociotechnical design, *Human Relations*, *29*, 783-792.

Chung, S.-H. and Y. Suh (2009). Estimating the utility value of individual credit card delinquents. *Expert Systems with Applications* 36(2, Part 2): 3975-3981.

Cooper, R., & Foster, M. (1971), Sociotechical systems, *American Psychologist, 26*, 467-474.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y-A., Waterschoot, S. and Bontempi, G. (2014) Learned lessons in credit card fraud detection from a practitioner perspective, *Expert systems with applications 41*, 4915–4928.

Few, S. (2006) *Information Dashboard Design: the effective visual communication of data*, London: O'Reilly.

Gigerenzer, G. and Gaissmaier, W. (2011) Heuristic decision making, *Annual Review of Psychology, 62*, 451–482.

Gigerenzer, G. and Goldstein, D. G. (1996) Reasoning the fast and frugal way: models of bounded rationality, *Psychological Review 103*, 650.

Gigerenzer, G. and Todd, P. M. (1999) *Fast and frugal heuristics: The adaptive toolbox*, Oxford University Press.

Hall, D.L., Hellar, B. and McNeese, M.D. (2007) Rethinking the data overload problem: Closing the gap between situation assessment and decision making. In: Proc. of the 2007 Symposium on Sensor and Data Fusion (NSSDF) Military Sensing Symposia (MSS), McLean, VA.

Kaelbling, L.P., Littman, M. L, and Cassandra, A. R. (1998) Planning and acting in partially observable stochastic domains, *Artificial intelligence 101*, 99–134.

Keim, D., Kohlhammer, J., Ellis, G. and Mansmann, F. eds, (2010*) Mastering the Information Age: Solving Problems with Visual Analytics*, Goslar, Germany: Eurographics Association.

Kieras, D. E. and Hornof, A. J. (2014) Towards accurate and practical predictive models of active-vision-based visual search, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3875–3884.

Koriat, A. (2012) When Are Two Heads Better than One and Why?, *Science*, *336*, 360–362.

Krivko, M. (2010). A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications* 37(8): 6070-6076.

Newell Ben R and Shanks David R. 2003. Take the best or look at the rest? Factors influencing" one-reason" decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition 29, 1 (2003), 53.

Thomas, J.J. and Cook, K.A. (2005) *Illuminiating the Path: the research and development agenda for visual analytics*, Richland, WA: National Visualization and Analytics Center, Pacific Northwest National Laboratory, http://vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf

Trist, E. and Bamforth, K. (1951). *Some social and psychological consequences of the longwall method of coal getting*, *Human Relations*, *4*,3-38.

Morar, N., Baber, C., Bak, P. and Duncan, A. (2015) What You See is What You Do: applying Ecological Interface Design to Visual Analytics, *paper presented at EP for DM 2015*.

Whitrow, C., D. J. Hand, P. Juszczak, D. Weston and N. M. Adams (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery 18*, 30-55.